

Going Digital



EVOLUTIONARY AND REVOLUTIONARY
ASPECTS OF DIGITIZATION



NOBEL SYMPOSIA



Karl Grandin ed.

NOBEL SYMPOSIUM 147

Going Digital

EVOLUTIONARY AND REVOLUTIONARY ASPECTS OF DIGITIZATION

Karl Grandin

Editor

© The Nobel Foundation, 2011



CENTER FOR
HISTORY OF SCIENCE
THE ROYAL SWEDISH ACADEMY OF SCIENCES



NOBEL SYMPOSIA



ALL RIGHTS RESERVED. NO PART OF THIS PUBLICATION MAY BE REPRODUCED, STORED IN A RETRIEVAL SYSTEM, OR TRANSMITTED IN ANY FORM OR BY ANY MEANS, ELECTRONIC, MECHANICAL, PHOTOCOPYING, RECORDING, OR OTHERWISE WITHOUT THE PRIOR PERMISSION OF THE PUBLISHER.

PUBLISHED AND DISTRIBUTED BY THE CENTER FOR HISTORY OF SCIENCE AT THE ROYAL SWEDISH ACADEMY OF SCIENCES LILLA FRESCATIVÄGEN 4A, PO BOX 50005, SE-104 05 STOCKHOLM, SWEDEN

NOBEL SYMPOSIUM 147

BIDRAG TILL KUNGL. VETENSKAPSAKADEMIENS HISTORIA /
CONTRIBUTIONS TO THE HISTORY OF THE ROYAL SWEDISH ACADEMY OF SCIENCES NO 41

COPY EDITOR: VICTOR KAYFETZ
GRAPHIC DESIGN: ÅSA WALDTON, WALDTON DESIGN AB
LAYOUT: MACGUNNAR – INFORMATION & MEDIA
PRINT: MODINTRYCKOFFSET, STOCKHOLM, SWEDEN, 2011

ISSN 0081-9956
ISBN 978-91-7190-175-0

TABLE OF CONTENTS

PREFACE 8

THE PRE-DIGITAL WORLD: HOW DID ONE MANAGE? 11

1. Nicolas Barker, “Google, Gutenberg and Earlier Revolutions in Communication” 12

2. Adriaan van der Weel, “Explorations in the Libroverse” 32

DIGITIZATION:
WHY, HOW AND FOR WHOM? 47

3. Lisbet Rausing, “Do Libraries Dream of Electric Sheep?” 48

4. Peter Suber, “Open Access for Digitization Projects” 70

APPLICATIONS 1:
ADDING VALUE FOR EVERYONE 95

5. Alain Giffard, “Digital Reading and Industrial Readings” 96

6. Kevin Guthrie, “Opportunity and Challenge: Sustaining Platforms for Knowledge Dissemination in the Digital Age” 106

INFORMATION STRUCTURE,
INFRASTRUCTURE AND METADATA 127

7. Terje Hillesund, “Digital Humanities: Why Worry about Reading?” 128

8. Magdalena Gram, “Digitisation in Practice – an ALM Perspective” 162

APPLICATIONS 2:

ADDING VALUE FOR RESEARCH 171

9. Martin Bossenbroek, “Circulation of Knowledge: Reconstructing the 17th Century Dutch Republic of Letters” 172

10. Laurent Romary, “Stabilising Knowledge Through Standards: A Perspective for the Humanities” 188

11. Anke Lüdeling, “Corpora in Linguistics: Sampling and Annotation” 220

WHEN EVERYTHING IS DIGITIZED... THE FUTURE OF LIBRARIES, ARCHIVES AND MUSEUMS, AND ANALOGUE 245

12. Gregory Crane, “From Subjects to Citizens in a Global Republic of Letters” 246

13. Marco Beretta, “Digital History of Science: Integrating Texts with Artefacts” 266

14. Emma Rothschild, “The Future of History” 280

NOTES ON CONTRIBUTORS 295

Preface

This volume contains texts that originate from presentations given at the 147th Nobel Symposium in Stockholm June 23–26, 2009. The Nobel Foundation has since 1965 on a regular basis organised symposia that primarily deal with areas of science where breakthroughs are occurring. But these symposia might also deal with topics of cultural and social significance.

The title of the symposium is the same as the title of this volume: *Going Digital*. The reason for organising the symposia came from various directions. The ambition was to address broad questions like what is evolutionary and revolutionary when going digital today? An emphasis was however put on what it means for humanistic research.

The symposia was organised in six sessions that each addressed important aspects and problems in the new digital landscape. The first session, “The pre-digital world: How did one manage?” provided a historical perspective as well as addressing what aspects that could be considered evolutionary or revolutionary. The first day ended with a session on “Digitization: Why, How and For Whom?” Here important questions of access and the role various institutions should take in making the best use of the new digital means was raised. In the next session, the first of two, possible applications were discussed in “Adding value for everyone”. What does massive digitization entail for all of us, and how can it become stable information? The following session dealt with “Information structure, infrastructure and metadata”. Here examples were presented on how academics today do use the available digital resources in their everyday work. The final day we returned to applications once again in “Adding value for research”. Here ample evidence was given of various successful examples of digital academic work. The final session dealt with the pertinent question “When everything is digitized. . . The future of libraries, archives and museums, and analogue”. Here different opportunities for digital humanistic research was given. The symposium then ended in a fruitful final discussion.

Not all of the presentations made it into the present volume, but are still available as videos at www.center.kva.se. The symposia received quite some media attention. Interviews and discussions with several participants constituted three consecutive programs on national radio in Sweden for example, and not surprisingly some of the participants blogged live from the symposium. The symposium was also webcasted live and afterwards available as video on demand.

The initiative to organise the symposium originally came from professor Tore Frängsmyr at the Office for History of Science at Uppsala University. An international advisory board was constituted to assist in putting together the various themes and finding the best speakers. It consisted of professor Emma Rothschild, Harvard, USA, & Cambridge University, UK, professor Robert Darnton, Harvard, USA, Dr. Lisbet Rausing, Imperial College, UK, professor Marco Beretta, Bologna/Florence, Italy and director Martin Bossenbroek, National Library, The Netherlands. The local organisation was carried out by the Center for History of Science at the Royal Swedish Academy of Sciences in Stockholm in close collaboration with the Office for History of Science at Uppsala University and the National Library of Sweden. Responsible were professor Tore Frängsmyr, professor Karl Grandin, associate professor Maria Berggren, National Librarian Gunnar Sahlin and Deputy National Librarian Magdalena Gram.

On behalf of the organisers I would above all like to thank all the participants for contributing so positively to the symposium! The Nobel Symposium Committee generously provided the necessary funding, and our grateful thanks naturally go to them. Finally I would like to thank my colleagues among the local organisers for a most fruitful collaboration!

Stockholm, February 2011

Karl Grandin

The pre-digital world: How did one manage?

GOOGLE, GUTENBERG AND EARLIER REVOLUTIONS IN COMMUNICATION

Nicolas Barker

Memory, the art of remembering things, was perhaps the earliest and is now the most ancient human faculty. ‘Digital,’ in the sense that we are using it today, is a word first used in print in 1938, and coined not much earlier, even if its precise application only came with ENIAC in 1943. Memory has always depended on adventitious aids, and the digital computer is only the latest of a series of tools that the human mind has called in aid to support a neural system within each one of us about which we still know remarkably little. What we shall be discussing, therefore, over the next three days is the future of memory. As we look ahead, with a mass of experience and apparatus, chronologies, mnemonics, words and images (engraved on stone or wood or metal, written by hand or in print, captured on film), sounds (also capable of being recorded), and, most evocative of all, smells (for which we have no record), to help us recall the past, what part will a still new digital technology have in the future of memory?

The art of remembering words has a history far longer than the history of writing, a history paradoxically forgotten, since the whole aim of writing was to supplement, and thus supersede, memory. But if the verbal devices by which we recognise verse were originally memorial in function, they came to have their own discipline: matter followed form, as well as dictating it. Nor was verse the sole mnemonic system for recording words: rhetoric and religion had different needs, and developed different, though not distinct, ways of meeting them. To win an argument was an important but transitory need. A liturgy preserved itself by repetition, the oldest of memorial devices, and the one by which verse also was preserved and remembered. But verse had separated itself from all forms of prose long before the first preserved, written, records of either came into existence.

Traduttore, traditore: every translation is a betrayal. Reading verse, as opposed to speaking or listening to it, loses something. The mind’s ear is not so agile as the mind’s eye. Subtleties of assonance are less easily appreciated on the page than the beginnings and ends of lines, or rhyme-words, similarly (or, as striking, dissimilarly) spelt. Yet the movement of the poet’s eye, from line to line, column to column, or even across the page, with the reader in pursuit, has its own poetry. If sound is the ultimate criterion of verse, the visible pattern that it makes cannot be ignored. The loss and gain in translation from spoken to written verse is itself but a reflection of the more complex and longer process of transition involved in the need to create a visible record of speech.

The idea in the mind of the poet is itself imperfectly captured in words, as authorial change in draft, fair copy, proof, print, and any subsequent manifestation, all too often shows. An idea, an image here may have started somewhere else, beyond the author's control if not ken; it may end ages later, in the hands of an editor, translator, reader, where it may take on quite different adumbrations and resonances.

I said that reading verse loses something, and of course I meant reading to oneself, silently. But reading, etymologically, implies reading aloud, and it is important to remember not only that texts were composed and then written down to be read aloud over a far longer time than those destined for silent reading, but also that other texts had as long a life or longer before any means of recording them existed at all. What we now have to do is to go backwards: back, that is, not only towards the irretrievable notion first expressed in words that imperfectly rendered it even then, before they achieved a greater durability and a changed form in writing, but also to the roots of that form, the sources of the graphic expression that is all we have of the earliest surviving verse. This second quest is yet further complicated by the fact that these roots have themselves been recovered only in recent time, and in reverse order, chronologically speaking. The Rosetta Stone in Egypt led back through Greek and Coptic to demotic and hieroglyph. Similarly, when it came to cuneiform, the trilingual inscription at Behistun, moved from old Persian, to syllabic Elamite, back to the more ideographic 'third script,' then called 'Assyrian.' This proved to be the complex vehicle of two languages, Akkadian and the earlier Sumerian. The earliest pictograms of Uruk c. 3000 B.C. were expanded phonetically to absorb two language systems. As systematised in the scribal schools and copying centres of Assyria and Babylon, cuneiform script could be used to represent both languages, structurally and aurally quite different. Like Chinese characters, this gave it a flexibility that enabled it to become the vehicle of other languages.

Remember, none of these languages existed as a spoken tongue in modern time, and the actual sounds of consonants and phonemes have had to be derived from their descendants. Egyptian and Akkadian only exist in written form, which gives that form a special significance as the vehicle of the texts it conveys. The form is the cause of their preservation. Monumental words, in prose or verse, are only part of all that is signified by a monument. Papyrus rolls or clay tablets exercise their own formal influence on the messages that they preserve. This can make it both easier and harder to determine whether these are in prose or verse. Estate accounts and chronicles of battles may be factual, but are they

not perhaps hymns? A strong, regular rhythm is common to both. Is an invocation or prayer, a spell or incantation, a poem?

If definition may initially be a problem, the volume of verse, Egyptian and Akkadian, that has survived is a clear indication of the importance of such texts as a memorial device. The most famous Egyptian text, 'The Tale of Sinuhe,' a court official, became a classic. The early texts are written on papyrus continuously in columns with no stichic arrangement or indentation. But the text also survives in a Ramesside copy, an ostrakon fragment; here the lines are set out horizontally, right to left, with inserted punctuation in red, a device for improving delivery when read aloud. This first instance of 'rubrication,' as it came to be called, both to punctuate narrative and ease reading, also served as a memorial device. Another text, contrasting the lasting work of the scribe with the vain labour of other craftsmen, was clearly known to the author of Ecclesiasticus. This discovery was paralleled by the discovery in 1872 of a Babylonian 'Flood' narrative, while the 'Epic of Gilgamesh' was as striking a discovery to a generation newly attuned to Beowulf and the Nibelungenlied.

What is more startling is the realisation that most of these texts were composed and some written before the nucleus of the *Iliad* came into existence, even if there is as great a distance, in some cases, between the date of composition and the date of the earliest surviving text. But such distances are also a reminder that verse has a third – or is it fourth? – dimension, in performance. We must remember that all this, the ceremony with words spoken or chanted, the inscription on the unmoving stone, the copy on papyrus that might divulgate it, was part of a memorial process, which was itself the life-blood of society and the mainstay of the cosmic order. This was as true on the Tigris and Euphrates as on the Nile.

If there are similarities between the development of Akkadian and Egyptian script and texts, in the demand for memorial record, in diglossia (the simultaneous use of two languages),¹ in the movement from pictogram to phonetic system, in the use of 'classifiers' and 'determinants' to avoid ambiguity, and in its maintenance by a 'scholarly' class, there are also considerable differences. Egypt, despite its cultural complexity, was essentially a unitary society, while the nascent Mesopotamian cities of Assyria and Babylonia were like the Greek city-states later, with the same urgent need for commerce and its record, the same free association of temporal and spiritual power, tyrant and temple. The two languages of Egypt evolved over time, classical and demotic diverging in ways different from the scripts by which they were conveyed. Sumerian

and Akkadian had different roots before there was any script to record either; the evolution of cuneiform was an exercise in compromise that had far-reaching consequences.

If there are constant echoes in the subjects and treatment of texts in the two great riverine cultures, it is as well to remember that in one particular, where similarity might have been expected, script, there is none. What broke this deadlock was the alphabet, product of the collision of the two cultures. If the alphabet can be said to have a beginning, it lies somewhere between Egypt and Mesopotamia, in a Canaanite need to isolate consonantal sounds and give them graphic form, adapting both established forms of script, hieroglyph and cuneiform, to this end. Two distinct and evolving systems can be traced, the Northern Semitic and Southern Semitic (not to be equated with cuneiform and hieroglyph).² The invention of the alphabet was revolutionary in detaching the depiction of language from the immemorial pictographic systems that preceded it. It is a supreme irony that the two great systems of language and script then current should have combined to provide the means by which both were lost, until recovered in the 19th century A.D. In both cases, accident preserved the earliest documents of the languages and script that replaced them. Some of us can remember the excitement caused by the discovery of the Dead Sea Scrolls; our grandparents were as excited about the recovery of papyri, mostly in Greek, in Egypt, as theirs by the discoveries of Champollion and Layard. Many of the texts were ancient, and if few of the documents found were as old as their texts, they were a great deal nearer the source than anything else that had survived. For the first time, it became possible to construct a tangible line, however tenuous, back to the earliest known records, including poetry. Memory had taken a long leap back.

Poetry in both classical Egyptian and Akkadian regularly used dualism, both as visually presented formal structures and in composition: visually, in stanzas arranged in distichs and quatrains (pairs of distichs); and textually, in what Robert Lowth was the first to call 'parallelism',³ the repetition of an expression in the first half of a verse paraphrased in the second half. The Hebrew Psalms provide a familiar example of both types. In their earliest preserved form, the fragments of 27 distinct rolls found in caves round the Dead Sea, are in a script little different in essentials from the Hebrew familiar today 2,000 years later. It is only strange in lacking vowel points and cantillation signs, still omitted in Torah rolls. The texts are uniformly written in columns about 15.5–18 cm wide and 24–27 lines deep, either on skins or papyrus. The skins are

ruled individually with a blind point and sewn together with linen thread, through holes set diagonally. There are no headings, nor are the verses (as we think of them) marked as such; instead, pericopes are broken off as paragraphs, and pauses marked by spaces on the line, spaces that filled out to the end of the line for an open pericope, shorter for a closed pericope.

This system of dividing the text for oral delivery was not new, and the use of space to punctuate, as in the Psalms, continued in other languages, both Greek and Latin. But other means of punctuation existed, overlapping with this and each other. Few of the earliest manuscripts in either language bear witness to this, but the instructions for the division of texts for rhetorical or grammatical purposes (that is, as spoken or written), laid down for Latin by Quintilian (*sententiae, cola, commata*) or by Aristophanes of Byzantium (points, breathings and accents), were certainly known and sometimes applied.⁴ As the surviving texts are almost all in continuous unbroken strips of letters (*scriptio continua*), the need was and is conspicuous: the fact remains, however, that the use of punctuation in manuscripts is not found early.

Why? Well, obviously, the manuscripts that survive are only a fraction (how small a proportion we cannot know) of all that there were. Why they did, or did not, have punctuation is a fact that requires better knowledge of the circumstances of their manufacture than we possess. Nor were manuscripts the only vehicle in which verse is preserved. Epitaphs and other inscriptions cut on stone are sometimes in verse, and epigraphic conventions, in which the presence or absence of verse had little influence, dictated a layout that did, in fact, have an influence on the presentation of verse in other media. Inscriptions were something to be seen and then read, often at a distance, unlike manuscripts, held in the hand to supplement recitation. Words on stone started in two dimensions, painted with a brush as a guide to the carver whose engraving gave them a third dimension. Initially, the shape of the stone and the need to fill the available space determined the arrangement of the words on it, so that lines of verse were not necessarily set out as such. Distinguishing separate words with a medial point was an ancient inscriptional practice that may be Etruscan in origin, and was carried forward into manuscript practice. The earliest Latin verse text to survive in what was once book form remains the papyrus fragment of a poem on the battle of Actium (31 B.C.) that was found at Herculaneum, thus written before the eruption of Vesuvius in 79 A.D. It is written in a rustic script like the graffiti found at Herculaneum and Pompeii, each hexameter is set out on a separate line. As in the inscriptions, the words are separated by spaces

and punctuated with a medial point. Seneca noted the practice, but it seems to have died out at the end of the 1st century.⁵ That this was in no way unique is demonstrated by a contemporary *graffito* from Pompeii, a ‘vote for me’ poster with the slogan as an elegiac couplet. The script is signwriters’ rustic. The hexameter begins with a letter larger than the rest, a *littera notabilior*, and the pentameter is indented; the words are separated by the medial point.

All these practices and their purposes can only be derived from the surviving witnesses. The graffiti have survived, against all expectation. That they have survived may be due to accident, but now, long after their original purpose has passed, they have a new purpose and message. Inscriptions on stone are more durable. They were made specifically to preserve the text on them, but preservation was only one of their memorial purposes. An imposing appearance might strike even the illiterate. Paint might wear off, and stone decay. Bronze was more durable, yet Horace knew that his odes would last longer yet: ‘exegi monumentum aere perennius.’ So verse might well be transcribed without punctuation since, to a mind and ear familiar with text and sound, the structure of verse dictated speed, pauses, even intonation. Letters, marks to assist comprehension, even the substrate, were secondary. Similarly, liturgical use might require a certain layout to assist or determine performance, spoken or chanted, but might also, as with the omission of vowel points in the Dead Sea Scrolls, assume linguistic practice familiar to both reader and listener. The nature of administrative documents, whether relating to land-title, taxes, accounts of money paid or owing, tended to dictate layout. Rhetoric, as we shall see, was a very different matter, the more so as, like verse, its needs ranged from the ephemeral to the eternal, from the need to win an argument here and now to the permanent record in suitably imposing terms of royal or national achievement.

This difference between immediate and long-term memory is ancient, even if the two needs came to serve each other. The rhythmic devices by which the Congolese *muyum* memorised their chants began as an aid to recitation, but they came to be the stabilising force of the narrative, just as the narrative authenticated the regime that it celebrated. So the considerable apparatus of ancient rhetoric was first devoted to winning the argument, whether in court or in a larger political arena, but then, the argument won, was given permanent form. If gravestones, large or small, depending on means or taste, are the commonest of monuments, most of the rest, whether they record conquest or the erection of a fountain, represent an event deserving preservation beyond the lifetime of those

involved. The argument itself was important, which is why the works of the classic orators, Demosthenes and Cicero and their contemporaries, form so large a part of the literature of their time that has survived,⁶ and why speeches, however artificial, form an important part of the histories of Herodotus and Thucydides.

In this apparatus, memorial devices played an important part, and are described in passing by Aristotle, and more exactly in the later manuals of rhetoric of Quintilian, Cicero and the Auctor ad Herennium. So, besides the strictly rhetorical devices of voice projection, speech rhythm and the use of *clausulae*, comes equally practical advice on setting out and recalling matter. Here the emphasis is on visual images and their analogues. As far as possible the speaker should anchor his argument in the minds of his hearers by creating a picture. Its contents should be defined and set out in hierarchical order, long sets broken into convenient and thus memorable sub-sets. The links connecting each element should be naturally joined together by visual or emotional association, and the whole secured in the mind by repetition, the orator's natural discipline. The mnemonic devices involved either an architectural 'background,' the elements set out like the furnishing of rooms within a building, or arranged in a familiar graphic pattern like the signs of the zodiac, or linked textually, the first letter of the name of a visual object recalling the first letter of an element in the argument. These 'notions' were called *notae*, but writing them down was evidently not part of the process, nor were actual pictures made to support the imaginary background, at least in classical times.

It is interesting that 'learning by heart,' where orator and poet are at one, should not have suggested a link between the two arts, but not even Aristotle, who wrote at length on rhetoric and poetry, drew the parallel. Perhaps it was too obvious then; now, when the ability to recall and recite either poetry or prose at length is rare, this test of repetition seems more significant. But, underpinning the orator's performance were two other disciplines, as necessary also to the poet: logic and grammar. Argument had to be tested by moving from one proposition to the next to ensure their sequence. Language had to be defined in terms of the inflexion and construction of words. The evolution of logic and grammar to form the structure of verbal communication was dependent on memory; it is significant that some of the earliest literary texts, like the 'Tale of Sinuhe' mentioned earlier, should survive in 'school-text' form, copied as an example of the structure of language. Contrariwise, the disciplines that created the structure of language had themselves to be

memorised, and the oldest mnemonic technique, verse, was employed to drive their rules into childish – and older – heads. If the 4th-century *Ars Grammatica* of Donatus became the staple of Latin grammar for over a thousand years, the metrical treatise on scansion of Terentianus Maurus (not rediscovered until 1493) was probably written about AD 200, and the *Doctrinale* of Alexander de Villa Dei (c. 1175–1240), rivalled Donatus as the plague of medieval schoolboys.

So far, we have found little, apart from fragments, but texts far older than their exemplars. The contemporary reader is not much in evidence, the author still less so, and the clouds now become thicker. But it is from this time, too, that the earliest substantially complete manuscripts survive, the ‘Ambrosian Homer’ and the ‘Codex Augusteus’ of Virgil, and it is worth pausing to consider this paradox. These monumental books, codices, not rolls, are an indication of something new: the need for a more lasting record than the roll, whose be-all and end-all was the ephemeral task of reading aloud, in company or alone. The roll had only to answer the needs of the reader and the audience within the sound of his voice. The codex made it possible to communicate the text to an indefinite number of readers, spreading over time and space. Their needs, real or presumed, dictated new disciplines in terms of layout, made possible by the form of this new thing, the codex. It was about this time that St Augustine surprised, and was surprised by, St Ambrose reading silently.⁷ Neither was to know, though both may have surmised, the need that these books were going to serve. It is no surprise, however, that an even greater text, in their eyes, should have been similarly served. The earliest of these great bibles, the Vaticanus, has no contemporary punctuation or division; Sinaiticus signals new paragraphs by extending the line where the break comes into the left margin; Alexandrinus goes further and marks the line with its first letter (not the letter with which the new section actually begins) enlarged and in the margin.⁸ Sinaiticus is normally written with four narrow columns on the page: for the Psalms, the number is reduced to two. The reason for the change is clearly to get in each verse, wherever possible, within the line. Where longer verses run on, the second (and subsequent) lines are deeply indented (about 20 mm of a 128 mm line), the indent line vertically ruled like the columns. The first verse of each psalm is numbered in the left margin, each Greek numeral with a bar above and below it.

The reason for this complex arrangement was not only to make the text clearer to the reader: it also reminded him of intonation, familiar from the daily liturgy.⁹ Two needs had to be met, those of the reader

and of the performer. St Jerome, a pupil of Donatus, the grammarian, was very interested in this. When confronted with ‘difficult’ passages of the Hebrew text, such as Isaiah or Ezekiel, in which too he sensed the importance of intonation, he tried to make it easier by setting out his Latin *per cola et commata*, with a new line for each clause, like the speeches (he says) of Demosthenes or Cicero. Jerome had read Cicero well, and grasped the importance of *clausulae*, the orator’s need to provide aural aids to understanding. He was not alone in realising the importance of auditory aids to understanding. When the need to protect the learning of the past persuaded Cassiodorus to set up the Vivarium at Squillace in southernmost Italy, he taught his monks punctuation by pointing an existing (but presumably unmarked) copy of the psalter, knowing that the breaks in sense and rhythm were already familiar to them, so that they would grasp the purpose of his notation. St Benedict, setting up his monastery at Cassino, went further and drew up his *Rule*, in which every detail of the life of his community – liturgy, work, diet, time-table – should be written down and remembered.

But already, Latin was not the only language in which verse might be composed or written. All this time, like the susurrant of the distant sea that becomes a roar as you get nearer, verse in the vernacular is growing, filling people’s minds, and, ultimately, emptying their inkwells, as the impulse to record it becomes stronger. The earliest recorded rhyming verse, and verse that depended on stress and alliteration, shows that methods of setting out the Psalter and Latin verse affected the layout of poetry that was now given visual form for the first time. If features that recall an earlier oral existence remain, it is no coincidence that systems of lineation and the use of *litterae notabiliores*, developed to meet the needs of liturgy or metrical voice, were adapted for quite different purposes.

We are so used to the use of initials, a large letter to signal the beginning of a chapter or a new text, that it is hard to remember that this mnemonic device has a history. It goes back to Egypt where, not initials, but rubrication served the same purpose. Ignored in Greece and Rome, it survived in the Syriac and Coptic churches, where points between words also appear in inscriptions. The Coptic psalms also began with a large letter, followed by others diminishing in size. The same feature is found in contemporary North African and Spanish inscriptions. Did those who took the gospel thence to Ireland take these traits with them? They are found in the earliest Irish manuscripts, initial followed by letters declining in size, in or outlined in red. The Cathach was ‘written in Ireland, traditionally, but not certainly, by St Columba’¹⁰, therefore in

the second half of the sixth century. For all its humble size and appearance, it is revolutionary in introducing word-division, and has the ends of stanzas marked with a cross. Isidore of Seville was the contemporary authority on punctuation; this practice was adopted by the Irish, heirs to a long oral tradition, with the same ardour as Christianity, along with its language, Latin, whose sound was alien to their Erse ears and so could only be appreciated visually. If punctuation hallowed the process of conversion to a new religion for the Irish, a different graphic convention came to Northumbria straight from Rome. Where the Irish were obliged to innovate, at Jarrow the Roman uncial was copied with equal fervour and precision.

The conflict in scribal attitude, as well as doctrine, was fertile. Its product, 'Insular' script and scribal habits, was re-imported to the Continent by Boniface and others. Collision with Germanic conventions and a different oral tradition produced, in a new Christian empire, a new script, in which capital and minuscule scripts, each with a different function, were merged. From this developed the 'upper and lower case' alphabet familiar to us in print or on the screen.

Up to now, all texts and all means for commemorating them had been based on a single unit. Before Gutenberg, every form of communication destined to be read, whether engraved on stone or clay or written on papyrus, vellum or paper, was unique. Multiple copies might be made, the strictest uniformity might be enjoined, but the individual impulse of engraver or scribe would contrive some mark that distinguished each individual copy from another. The same text, whether the Bible or an imperial triumph, might be multiplied many times, but no demand for uniformity in presentation yet existed. From the 10th century onward, pressure in this direction grew. Monasticism and a more stable hierarchy imposed a new order in the transcription of the Bible and liturgy. St Benedict's Rule enjoined the copying of texts, and as other orders following his precepts spread, uniformity came both naturally and deliberately. Exact similarity, line by line, initial for initial, was enforced in the copies of the commentary on the Psalms of St Bruno of Wurzburg, founder of the Carthusians, made at his abbey of Reichenau and its daughter houses. The most important of all the texts copied was that of the Bible, for which increasingly elaborate glosses were required. The arrangement of the Bible text and the gloss on it in separate columns required exact planning, to facilitate comparison and to save space (vellum was expensive, and empty gaps in either text or gloss were to be avoided. As the extent of the glosses grew, so the structure had to be elastic. The great

Norman abbeys, Pontigny, Jumièges and others, spread the practice to England and also Paris, which became the acknowledged centre for the production of such books.¹¹ There, too, in the 13th and 14th century, small complete texts of the Bible were effectually mass-produced, in both text and decoration.

Abbeys were not the only institutions with an interest in uniformity. The growth of universities provided the foundation for scholasticism, reconciling the sum of human knowledge in Aristotle in the seven liberal arts, the trivium (grammar, logic and rhetoric) and quadrivium (arithmetic, geometry, astronomy and music), with Christian theology. At Paris, Bologna and Padua – the centres for teaching theology, logic, law and medicine – the need to communicate and retain an increased body of information brought about a new mode of transmission, which spread to other universities. Initially, the masters taught from their own copy of the texts on which they lectured, students making do with other copies, their own notes and memory. Soon the concept of the ‘set text’ took over. The approved version (called a *pecia*) was held by an appointed university stationer, who would lease it out, one section at a time, to students who would copy the section, return it, and draw the next. This ensured security (the stationer would not advance a second until the first was returned), and also speed – if the text had twenty sections, as many students could copy it simultaneously. The copies had to be as nearly identical as possible for the system to work. This had the further advantage that the master could refer to passages by page and place on it, knowing that the students could find it. Examples of *peciae* and copies made from them exist (in one remarkable instance both coexist), and more remain to be found (they are humble and inconspicuous books). The *pecia* system made a remarkable change in educational practice, since it reduced the need for memory and accelerated the absorption of knowledge. It lasted for a century and a half, from about 1200, and was one of the many casualties of the Black Death. If it had survived, it might have anticipated the invention of printing, whose produce it so closely simulated.

The traditional methods of copying still existed, individual, in monasteries, and, increasingly, in professional scribal shops. Demand did not slacken. It is possible to discern in this growing obsession with the word as read a decline in the practice of memory, or rather its transfer from the mind and its ways to a different form on the page. A new interest in the totality of human knowledge, evinced in the *Summa totius theologiae* of Thomas Aquinas, seemed to separate it from the contents of the individual human mind, while making it more attainable. The

idea of memory formalised in the graphic aids advocated by classical writers reappears in the *Summa*, but Aquinas also (following Aristotle) distinguished memory as a faculty, both neural as part of the body but also of the soul. Cultivating this was part of the practice of the virtue of Prudence, that specially Thomine virtue.

The art of memory also found a new outlet in the medieval theory of illumination by miniatures. Pictures as illustration, an art immemorially old, and united with texts in classical times, as witness the Ambrosian Homer, the Romanus of Virgil or the Vienna Genesis, now took on a new function as a guide to devotion. As well as the subject as a whole, each detail came to have a particular purpose in stimulating response in the viewer through memory of its signification. As well as pictures there were diagrams, such as those demonstrating the logical propositions that established the basis of faith through reason. More complex semi-realistic designs, combining words with images, returned the function of memory to the similitude of art. The ambiguous attitude of the schoolmen to the Arabic philosophers, who combined better access to Aristotelian texts with religious opposition to realism in art, also influenced memorial techniques. From the Islamic world came what we call 'arabic numerals' but which in fact were Indian signs for the nine basic numbers, to which zero was added. The 9th-century treatise on them of Al-Khwarizmi added 'algorism' to Western language. The teaching of Llull, poised in Majorca, half way between Paris and Tunis, combined Aristotelian logic with contemporary Arabic science (in particular the *zairja*, a calculating device of two concentric discs, whose sectors could combine different facts and attributes) in his *Ars generalis ultima*. It is no coincidence that his preferred system of *notae* was the signs of the zodiac, rather than the classic architectural model. Its idealism combined with mechanistic logic found new echoes in Neoplatonism; as the ancestor of Leibniz's 'ars combinatoria' it anticipates the principle of a computing 'engine.'

Where, now, in the history of memory, was verse? The new fashion for rhyme, still in Latin, could be traced back to the Sequences, so familiar that they could be set out continuously, not lineated, with (or without) punctuation to point the rhyme-words.¹² Indeed, it is hardly possible to think of rhyming *lines*: rhyme and assonance follow so fast on each other. But increasingly, the way that verse was written reflected the rhyme scheme, just as earlier scansion changed metrical verse from *scriptio continua* to reflect the shape of metre. Thus, the 'Cambridge Songs' were set out in two columns, the first letter of each song with a large initial, each four-line stanza with its own smaller initial, offset to the

margin, and the lines, *aabb*, run on but punctuated.¹³ The growth of verse in the vernacular demanded innovation; from ‘The Owl and the Nightingale’ to the *Divina Commedia*, new kinds of poetry, new types of texts, found an equal variety of forms. The reader can confront the poet directly. What we read before was mediated through the pen of a scribe, we can see Petrarch composing sonnets on the very sheet, changing his mind and meaning as he copies it out fair. Nor was the use of poetry for memorial purposes restricted to the Christian West; in Isfahan, Avicenna wrote philosophy and medicine in verse, so that it might be more easily remembered. Besides his metrical grammar, Alexander de Villa Dei wrote his *Carmen de Algorismo*, putting the decimal system of calculation into verse.¹⁴

Verse in manuscript, increasingly lineated and set out in stanzas, created an appearance that was part of its mnemonic function. A tall narrow format was preferred for the texts of classical poets, Virgil and Horace, used for educational purposes, and increasingly for didactic verse of all sorts in all languages. It is verse of this sort that forms the bridge between all the verse written before the invention of printing and all that came after. The art of memory was related to the layout of all texts, not least poetical; preservation and use equally demanded explanation, in parallel with or encircling the text expounded. School-room use, the need for a group of students to be able to refer to the same line or phrase in an identical place on the page, demanded the closest approach to identicality. Those who were taught in this way can always remember where on the page a particular theme or phrase appeared, and sometimes even the number of the page.¹⁵

The popularity of the versified Latin grammar of Alexander de Villa Dei was one of the most widely divulged of all medieval texts.¹⁶ It lasted into the early years of printing, and fragments, some on vellum, of lost editions exist, as they do of the older and still more popular Donatus. Like all grammarians, Alexander was interested as interested in exceptions as in the rules that they broke, and this links his work with a poet, not much of a poet, but the first author to see his own works in print. Johann Borne owed this distinction to his employment, from about 1460, as editor and proof-corrector at the press of Johann Fust and Peter Schoeffer. But he has a greater claim to fame as the first person to celebrate the new invention that was radically to change the purpose as well as method of communication. He did this in verse that appeared in successive works printed by Schoeffer. The words that he chose (several invented by himself, as no current words existed for facts and things so

new) are not easy to understand, and his Latin hexameters limp rather than run. But his sense of wonder at the potential of the great new invention, providential in its power to further the word and work of God, is unmistakable.

His other work was a metrical Latin grammar. Like Alexander, he was fascinated by exceptions, and his rules abound with lists of words (including some of his own inventions, to which they provide a useful guide). But there his work is, under his own name (suitably obscured in allusive verses). For the first time a poet is confronted with a new and transforming process for his work, its translation into print and through print to a far wider readership than any manuscript could reach. But, primitive as it still was, it was not quite adequate for Borne's work. The sections and conjunctions into which it was divided had to be signalled, and a handful of leaden letters in only one size, without the other devices the scribe had at his disposal, were insufficient. So, like the writer of the Egyptian papyrus with which we began, he dipped his pen in red ink to add the necessary words and point them in the right direction.¹⁷

If Johannes Borne was the first living author to see his work in print, that is his sole claim to distinction. But the problems that he experienced in the transfer of his work from manuscript, his own or a copyist's, into print are the beginning of a whole new series of compromises between the words as they first took form in the author's head and the words on the printed page that finally conveyed them to the reader. An intervening listener was now a diminished factor in the process; other figures, editors, copy-preparers, compositors, publishers, came to fill the gap. The manuscript, however, was not diminished, becoming an alternative to print, as it once was to the voice. But print, its disciplines and artifices, limitations and vast new power to gain and retain readers, had come to add a new element to the apparatus of verse, to translate it into a new medium. The idea in the mind of the poet is itself imperfectly captured in words, as authorial change, in draft, fair copy, proof, print, and any subsequent manifestation, all too often shows. An idea, an image, here may have started somewhere else, beyond the author's control if not ken; it may end ages later, in the hands of an editor, translator, reader, when it may take on quite different adumbrations and resonances. This elasticity was threatened by the immobility of print, but, paradoxically, it was this that set a new value on the poet's original composition, *fons et origo* of all the forms that might follow it. The freezing of the text in the printed image of inflexible metal led to a new interest in the fleeting image of the poet's mind and hand in the act of composition, to be caught, as we

can see, between the rough draft and fair copy of Petrarch's sonnets. But even now, ancient habits, the result of seeing as well as hearing verse, persisted, influencing the way poets now wrote down their words, visible signs of a process in the mind's ear of which no record could be kept.

This revolution, so apparently sudden and yet far-reaching, in terms of its effect on the transmission of verse in Europe over the next two centuries, is only part of a much larger and longer lasting change in the appearance and structure of books. This change was accelerated, but not initiated, by the invention of printing, which however canonised it. Armando Petrucci has shown how the 'modern book,' a small (in printing terms octavo) narrow oblong page, evolved from the combination of the university text, the humanistic revival of the classics, and the popular 'vernacular' book designed to be carried in a satchel.¹⁸ The need for simplicity, uniformity and speed in copying of texts for students, the preference of the early Italian humanists for plain unglossed texts of the classics, and the practical concerns of carriage, all led in the same direction. But it was printing, at first imitative and then normative, that came to co-ordinate these tendencies. Paradoxically, too, it canonised letterforms that were not new: with so much at stake, the inventors chose letters that had lasted well already. This 'backward look' characterises each technical innovation.¹⁹

The existence and availability of multiple virtually identical copies brought into being all sorts of mnemonic devices hitherto unimagined or unnecessary. Numbering leaves or pages had been optional; it now became vital for reference. This in turn led to the paginated index at the end of a book, first of words beginning with the same letter, then listed in exact 'alphabetic order,' an idea created by this need. A list of contents became a regular feature of the beginning, and footnotes, keyed to the relevant passage, took the place of meandering glosses. All these changed the business of reading from a visual to a mental activity, making the remembering of a text a mechanical rather than an imaginative process. Distinguishing quotations and numbering lines added to the apparatus of consultation.²⁰

A new mechanical system of mnemonics did not discourage older notions. Giordano Bruno may have been given more credit than he deserves for extending Llull's combinatorial circles through the 'ideas' of Neoplatonism to create a memorial sum that was greater than its parts, but the new mathematics was already extending numbers beyond human computation. As the boundaries of human knowledge, whether of the earth, the universe, or *omne scibile*, seemed to be expanding exponentially,

so the need to systematise the methods by which all this novelty was recorded grew. If verse as song or ballad retained its original memorial function, it acquired a new one when, transferred to the printed page, it became part of a much larger system. A further physical change took place. The book in the form standardised over the last millennium was a stable, immobile thing, as its name, codex, implies. Large size and solid structure made books weighty, both literally and metaphorically; small books were exceptional. This now changed, with the invention of the Aldine pocket classic, Petrucci's 'modern book' in printed form. The combination of the small page, the exact same size as Petrarch's manuscripts (perhaps by no coincidence, its proportions are those of the 'Golden Section'), the use of the new italic type, above all its simplicity, made the series of classics projected by Aldus Manutius an instant publishing success. It had many imitators and followers. A pocketable text, whether of classical poetry, the liturgy, or such novelties as a dictionary, gazetteer or *materia medica*, reduced what had to be held in the human memory. What could be kept in the pocket did not have to be held in the mind.

So, if the sixteenth century had teemed with eager theorists, each with a system allied to the fingers, the planets or other adventitious aids, a serious intellectual assault on the subject could be made in the seventeenth. It began in 1613 with Sacchini's short *De ratione libros cum profectu legendi*, continuing with Drexel's *Aurifodina artium et scientiarum omnium* (1638), and ending with Locke, whose *Méthode Nouvelle de Dresser des Recueils*, published by Jean Le Clerc in 1686, linked extracts to systematic record in the common-place books that were his special legacy. Its apogee was reached in Vincent Placcius *De arte excipiendi vom gelehrten Buchhalten* (1689), on the art of extracting the meat from learned works. Once you have done that and put it where you can find it again, you can safely forget it, he says, and to make it plain he added a plate depicting a thumb-index book. Alberto Cevolini, the modern historian of the mechanisation of memory, summarises these experiments, adding Moser's *Vortheile vor Canzleyverwandte und Gelehrte in Absicht auf Akten-Verzeichnis* (1773) and ending with *Kommunikation mit Zettelkästen* by the great modern theorist, Niklas Luhmann, whose *Die Gesellschaft der Gesellschaft* (1997) he frequently quotes. If the index-card is the essential tool, preservation devices, from the spike (certainly medieval and probably older) to the card-cabinet (a mid-seventeenth century invention) are as essential. The purpose of a catalogue thus constructed is to act as an interface between two closed systems, source and object, and its purpose, in Pomian's coinage, 'futurocentric'²¹: it aims to anticipate an as yet unrealised need; for

an extract thus to work it must be brief. As new modes of communication emerge, bent on creating as well as anticipating intellectual needs as yet undreamed of, Cevolini's record of what the human hand and brain could achieve in past time has an undiminished interest. Can our age, in his piquant phrase, safely 'learn to forget'?²²

The next revolution in the means of memorialisation affected not the physical apparatus, the shape of books or the filing systems imagined by Jesuit philosophers, but a return to that most abstract element, the shape of letters themselves. Scribes writing on papyrus formed letters with a diagonal stress, at an angle to the grain of the substrate, to prevent the pen catching on the fibres. Sign-writers preparing inscriptions on stone followed suit, as did punch-cutters imitating script for type. But when in 1683 a committee of the Académie des Sciences set up the Commission under the Abbé Bignon to design a new letterform to celebrate the achievements of Louis XIV, the result changed this. The letters, drawn and engraved on a grid of 2,304 squares (anticipating today's bit-maps), had a vertical stress. This was followed by the punch-cutter, Philippe Grandjean, in creating the *romain du roi* in 1696. Protected from imitation, it was only in the last quarter of the next century that Giovanni Battista Bodoni in Parma and François-Ambroise and Pierre-François Didot made it universally popular, answering the new taste for Neoclassicism. The new 'Modern Face' type was more elastic than its now 'Old Face' predecessors. It could be made thinner or fatter, and otherwise distorted in many ways; as many optical devices to assist or distract the human mind and eye, from timetables to pyrotechnic playbills, followed. These changes accompanied an industrial revolution that saw paper-making and printing mechanised and accelerated by the Fourdrinier, Koenig & Bauer's steam press and Earl Stanhope's stereotype process.

These changes were not without effect on the form of poetry. Up to now, the process by which words in the poet's mind took visible form, written first on paper, and thence transmuted into print, can be seen as a progress through tension, even contention. Not much had changed since lineation had been standardised, canonised by a comma, unrequired by sense, at the end of lines. Putting things down on paper might be a step forward, preserving what would not last long in human memory, but it was also a constriction, silencing the aural component. Print, as Tennyson found, cuts the line into sections, stilling its motion and fixing its form at that point. At the same time it creates, through multiple copies of that form, a lateral movement that increases its audience, whose perception and needs may react with the poet's aims and carry the theme forward

in a different direction. Mechanisation of the press and improved communications increased this pressure: the printed page, directed to a less defined body of readers, becomes impersonal, a void that the poet can fill. Unconstrained by the old link with printed letters, the poet's script became another form of self-expression, as well as the words written.

Increasingly, the printed word followed the written more and more faithfully. Old trends were again followed when the Monotype and Linotype companies, combining the composing and casting of type, came to apply the same logic to their predecessors as the first printers to the manuscript sources that they copied. At first content to imitate the types presently used by the printers who, they hoped, would buy the new machines, they moved on backwards to imitate the best – that is, most influential – designs of the remoter past. Allusive typographic design offered a new mnemonic aid to comprehension. Quite apart from the ability to replicate and commemorate earlier letterforms, the mechanics of letter-generation, first using technology of the Jacquard loom in punched paper tape form and then the same information digitally encoded in electronic form, detached the text from its image on the page.

It all started when the first text preserved by memory alone was entrusted to its first written surrogate. Over the next 5,000 years, the page, a rectangle enclosing lines of size and length convenient to eye and mind, on stone, clay, metal, papyrus, hide or paper, has been the unit of consultation, record and preservation. This shape, however recalled, lies at the root of human memory. Now, some five millennia later, the text, free of the constraints and opportunities of graphic preservation, faces the advantages and perils of a record in an entirely new medium.

The ideas put forward here grew out of 'Things Not Revealed' the Panizzi Lectures given at the British Library in 2001 and the Rosenbach lectures given at the University of Pennsylvania in 2002.

References

- 1 R. Parkinson, *Cracking Codes* (London, 1999), 49–50; J. Bottéro, *Mesopotamia* (Chicago, 1992), 90–4.
- 2 D. Diringer, *The Alphabet*, 3rd ed. (London, 1968), I. 148ff, 158ff.
- 3 In *De Sacra Poesi Hebraeorum* (1753); see J. Kugel, *The Idea of Biblical Poetry* (New Haven, 1981), 12ff.
- 4 On Latin *praelectio*, B. Townend, ‘Some problems of punctuation in the Latin hexameter,’ *Classic Quarterly* 14 (1969) 33off; cit. Parkes, *infra*.
- 5 M. Parkes, *Pause and Effect* (Aldershot, 1992), 10, citing R. Seider, *Paläographie der lateinischen Papyri*, ii/1 (Stuttgart, 1978), nos 6 and 10, and G. Nicolau, ‘L’Écriture et l’Histoire du Mot Phonétique en Latin,’ *Revue des Études Indo-européennes*, I (1938), 85–8.
- 6 Similar texts are found earlier in Mesopotamia and Egypt.
- 7 St Augustine, *Confessiones*, VI.3.
- 8 Codices Vaticanus, MS.gr.1209, Sinaititicus, B.L.Add.43723, Alexandrinus, B.L. Royal MS. I.D.v–viii, (SMP&S, fig. 19–21).
- 9 B.L. pap. 230; Milne 207, inv.230.
- 10 Royal Irish Academy, s.n., C.L.A., II.266; S. Morison, *Politics and Script* (Oxford, 1972), 152.
- 11 C. de Hamel, *Glossed Books of the Bible and the Origins of the Paris Book Trade* (Woodbridge, Boydell & Brewer, 1984).
- 12 As in Bodley, Digby 19.f.74 (Regular Sequence).
- 13 Cambridge, University Library, MS.Gg.5.35.
- 14 Charles Burnett, ‘From Abacus to Algorism: the Discovery of Arabic Mathematics in the West’ (Manchester, 30 April 2007).
- 15 T. H. White, *The Sword in the Stone*, (London, 1938), 8.
- 16 B.L., Harley 3583 is a fine early example.
- 17 N. Barker, ‘*Grammatica Rhythmica*: Copy, Text and Layout, 1466–1468,’ *Hellingsa Festschrift/Feestbundel/Mélanges* (Amsterdam, 1980), 43–57; ‘A Contemporary Panegyrist of the Invention of Printing: the Author of the *Grammatica Rhythmica*,’ *Incunabula: Studies in Fifteenth-Century Printed Books Presented to Lotte Hellinga*, ed. M. Davies (London, 1999), 187–214.
- 18 A. Petrucci, ‘Alle origini del libro moderno: libri da banco, libri di bisaccia, libretti da mano,’ *Italia Medioevale e Umanistica* XII (1969), 295–313.
- 19 H. Carter, *A View of Early Typography* (Oxford, 1969).
- 20 S. H. Steinberg, *Five Centuries of Printing* (2nd ed., 1961), 28–30.
- 21 K. Pomian, ‘Les Archives: du Trésor des Chartes au Coran,’ in *Les Lieux de Mémoire*, ed. P. Nora (Paris, 1997), 4058.

- 22 Alberto Cevolini, *De Arte Exciendi: Imparare a dimenticare nella modernità* (Florence, 2007).

EXPLORATIONS IN THE LIBROVERSE

Adriaan van der Weel

One of the *délices de Leyde* is the Bibliotheca Thysiana at the city's central canal, the Rapenburg, a few hundred metres away from the old university building. This beautiful library, built in 1655 in the stern Dutch classicist style by Arent van 's-Gravesande, forms the perfect embodiment of the ideal of scholarship and learning. What greets the visitor inside is a massive (certainly by seventeenth-century standards) collection of knowledge on a wide spectrum of subjects, resulting from centuries' worth of scholarship. Laid down in books neatly arranged on their shelves, this knowledge sits there in permanent readiness to be consulted and turned into further knowledge: *pabulum mentis* for its founder (Joannes Thysius) and his friends.

Thysius' well-ordered collection of seventeenth-century learning may serve for the purpose of this essay as a symbol of what I would like to call the 'knowledge system'. The knowledge system is characterised by such properties as order (as reflected in shelving and cataloguing systems); closure (each book being a finished account of a certain area of knowledge); stability (enabling readers at other times and places to consult the same knowledge; and indeed Thysius collection, which he left to the university, is still with us three and a half centuries later); registration (the books' title pages almost without exception attributing the texts to known authors), and authority (the texts having been deemed of sufficient quality to have been enshrined in print).

One of the eye-catching objects in Thysius' library is a seventeenth-century wooden book wheel. If the physical library with its rows upon rows of knowledge incarnate can symbolise the knowledge system, this book wheel, being circular, is an apt illustration of the way in which it was used. This was in an iterative process which may be termed the 'knowledge cycle,' which is at the core of the knowledge system. In the knowledge cycle readers, consuming and digesting existing knowledge, become writers, producing books containing new knowledge, which can be read by new readers in turn. This process depends centrally on the properties listed earlier. The system of attribution and authority, for example, has in the course of time become very transparent. It can be epitomised by the footnote as it has evolved over time.¹ This system permits the reliable identification of, and attribution to, prior authoritative sources used in the knowledge creation process at large. The book wheel is one of the most ingenious machines ever devised to accommodate this universal scholarly practice.

The origins of this system are obviously very much older than the printed book, going back to the invention of writing, which first permitted the creation of a lasting material

record of human achievement. But if I take the printed book as my point of departure it is not only because of the fantastic human achievement it represents but, more to the point, because it was the printed book that afforded the means for this system to come truly into its own, and because it still dominates our thinking about the production, dissemination, consumption and preservation of knowledge today. This way of thinking is often referred to as ‘print culture,’ though I prefer the term ‘Order of the Book’.² The Bibliotheca Thysiana can be said to symbolise the Order of the Book as a repository of collective human knowledge in the form of an orderly and stable textual record. The knowledge system serves to create, disseminate, consult and preserve knowledge by recording, multiplying and distributing it, describing it bibliographically and collecting it in physical (or as we say in a world that is rapidly becoming dominated by the virtual, ‘brick-and-mortar’) depositories such as libraries, in the stable material form of print.

It is to this way of thinking about knowledge, dominated by the material book as a knowledge machine, that the term *libroverse* from my title refers. Admittedly, it is an ugly coinage. However, as a counterpart to the equally ugly but more established term *docuverse* (coined by Theodore Nelson, to whom the world also owes *hypertext*), it nicely epitomises the two-way application of mental (pre)conceptions about our knowledge environment, from the analogue to the digital and from the digital to the analogue. The comparison implied in this juxtaposition offers a good way to study some of the effects of the recent digital deluge on the book-based knowledge system.

My suggestion is that the docuverse as a universe of linked digital documents is a much more revolutionary departure from existing textual practices in the libroverse than even the most die-hard preacher of the digital gospel has proclaimed so far. So can we fit digital knowledge production, dissemination, consumption and preservation into this age-old, familiar, tried and tested analogue system which I have called the Order of the Book, or is Order of the Book at the end of its use-by date and does a revolutionary new (digital) knowledge system need to be built from scratch? Indeed, can we have a system at all? To help answer these questions, and to underpin my suggestion of the revolutionary nature of the docuverse, in this essay I would like to contrast two perspectives, one of continuity (the evolutionary perspective), and the other of discontinuity (the revolutionary perspective).

1. *Continuity/Evolution*

It is often suggested that ours is the ‘information age,’ resulting from a veritable information explosion, and leading all of us to the brink of acute information overload. Explosions being the stuff revolutions are made of, we are apt to think that a revolution is what has hit us. However, we should remind ourselves that we are by no means the first generation to feel thus affected. There is perhaps more continuity than the term ‘revolution’ might suggest.

Many since Gutenberg’s time have lamented the proliferation of knowledge laid down in print, but solutions were not long in coming. These solutions tended overwhelmingly to use the very medium that caused the problem also to solve it: printed books, including annotated bibliographies, extensive lists of abstracts, and no end of encyclopaedias. One of the most familiar attempts at subduing the chaos resulting from the unbridled proliferation of printed knowledge is no doubt Conrad Gessner’s impossibly ambitious project of the *Bibliotheca universalis*, begun in 1545 and – to us unsurprisingly – never completed.

From the early twentieth century, solutions became more technological in nature. In the Mundaneum, proposed and actually begun by the Belgian Paul Otlet (1868–1944), the world’s combined knowledge was to be brought together and to be made accessible through a telecommunications system that encompassed both the documents themselves and the bibliographical system that described them.³

To try and manage an excess of information was also the prime motive for Vannevar Bush to invent his ‘memex’ in 1945.⁴ Slightly less advanced than Otlet’s vision of a World Wide Web *avant la lettre* this ingenious device is perhaps best understood as a twentieth-century makeover of the book wheel. Based on microfilm technology, it was designed to deal with vastly greater quantities of information than the book wheel, but crucially, it too was intended to enable the user to make annotations and link ideas to their sources. The linking system of the memex, mechanically connecting microfilmed texts, has been hailed as the precursor of hypertext.

Since such referencing of authoritative sources, through footnotes, bibliographies and so on, remains the foundation of all scholarly work, it is not surprising that the WWW continues to support it. Indeed it improves on it in various ways. The WWW is no longer restricted to the individual scholar’s workplace but instead, as Otlet had already imagined, allows the user to tap into an unprecedented wealth of linked resources, and in return to share his own resources with other users. Furthermore,

there is no restriction to the extent and nature of the links. Apart from the usual bibliographical references, hypertext notes may also contain direct links to the referenced materials themselves, be they primary or secondary sources. The footnote – the epitome of scholarship – could in other words be said to find its ultimate expression on the WWW: the hyperlink makes it actionable and gives instant access to the material that is being referenced. (Talking about continuity, in retrospect the book wheel is also vivid proof that the much-vaunted non-linearity of hypertext that is supposed to have released us at last from the bonds of the linear book is not all that new.)

Besides accommodating, and indeed improving on the age-old practices of annotation and referencing, there are many more ways in which the digital medium continues well-established analogue practices. To manage identification and access of digital items, metadata substitute for catalogue cards. The mark-up of digital texts with forests of labels enclosed in angular brackets – for example according to the guidelines of the Text Encoding Initiative (TEI)⁵ – may look forbidding, but the system that underlies that forbidding appearance is firmly based on the same principles of hierarchical structure and order that underlie typographic texts, including the attribution to named individuals of every last character of the transcribed text and its critical interpretation. This means that for the benefit of humans it is easy to conjure up – with the help of a style sheet – a view of the marked-up data that conforms fully to the same typographic conventions that have been used for centuries in the world of books. But this way of encoding texts with mark-up enables not only humans but also the computer to decode their meaning. And the computer can be instructed to create all the usual props of scholarship that used to be made laboriously by hand, such as indexes or tables of contents, at the proverbial touch of a virtual button. Again, the computer is thus used to improve existing practices.

In the digital world the sum of human knowledge grows at an even more inordinate rate than in the analogue world, but just as in the case of books, the same technology that causes that growth can again be employed to keep control over it. Metadata and free-text search engines give access to anything and everything on the Web. A major contribution to this growth is the mass digitisation of legacy documents – for example of books by Google and other parties. This brings closer than ever that elusive ideal of collecting the record of all knowledge ever produced in the world, an ideal that goes back at least as far as the Alexandrian Library.

There is conceptual continuity too, borne out by ample linguistic

evidence: we tend to discuss the docuverse in terms of the libroverse. The typographic vocabulary of the internet, for example, is essentially the same as that of the book. Web metaphors, too, show their heritage in the world of the book, with Web *pages* being *published*, *scrolled*, *browsed*, and *bookmarked*. Even the word *library* is still current, even though libraries are being turned into ‘information resource centres’ at a furious rate. Outside of Special Collections, library holdings are increasingly digital and have less and less to do with (paper) books, rendering the term more and more metaphoric.

The history of knowledge machines (which offer ways to process the sum of human knowledge) began with writing, continued with the printed book, and has now reached the computer. That history has always been, and appears to continue to be, based on such order and structure that enable the full identification of the sources of our knowledge, and their attribution to known individuals and materials. From the original wooden contraption designed by Agostino Ramelli to Bush’s micro-film reader to the WWW, the book wheel may be said to epitomise the knowledge system, thus also becoming a symbol of the continuity between the old and the new.

2. *Discontinuity/Revolution*

The case for evolution certainly seems strong, and the continuity perspective is, moreover, a soothing one. However, the very fact that a Nobel symposium on the subject of digitisation is being organised (and so many symposiums, conferences and colloquiums like it the world over) suggests that we are not altogether easy about it. In fact, the appearance of continuity may well be deceptive. How much discontinuity might be hovering under the relatively placid surface just described? In the perspective I would now like to take I will begin by simply identifying some of the discontinuities that suggest themselves to a critical observer. Next I would like to suggest why these discontinuities matter: what their effects on the knowledge system are.

In fact, this discontinuities perspective has a longer history than the continuities one. It goes back to such early hypertext theorists as George P. Landow, Jay David Bolter and Richard A. Lanham: all proclaimants of a textual revolution.⁶ (Paradoxically the unbridled zeal with which the likes of Landow, Bolter and Lanham proclaimed the revolution may well have led many to underestimate the revolutionary potential of digital textuality.) In the discontinuity perspective, I join them in proclaiming a revolution, although not necessarily for the same reasons. This is largely

a matter of reflection afforded by hindsight. These authors wrote when the real revolution, that of computers linked in a global communication network, had simply not yet happened. The World Wide Web was only launched in 1991 and its potential had yet to be recognised.

With the benefit of the hindsight that we have since acquired, I would like to single out the following three chief reasons why digital textuality, first by itself, and then augmented by the fact that this digital textuality no longer takes place in a standalone environment but on the Internet, and more particularly the World Wide Web, is really revolutionary and offers greater discontinuity than continuity with analogue text and the *libroverse*.

1. Digital text has its being in an environment that is not exclusively textual: the computer is now also a new medium for 'texts' in other modalities. Initially this was just still images and computer-generated graphs. Now it includes anything that can be medially transmitted, such as moving images, speech and music.

2. Digital texts remain computable. This translates into a number of effects, of which I will single out two:

(2a) The first is what I like to refer to as instability. By that I mean that texts may simply vanish, or be made to vanish, instantly and, more pervasively and potentially more disruptively, intrinsically lack closure. Both form and content are subject to continuous change. Lack of closure and its companion non-linearity (which I suggested was less of a departure from existing reading habits than the theorists claimed) are tropes familiar from hypertext theory, which hailed them as a liberation from the confining textual space provided by the Order of the Book.

(2b) The second effect of having texts in a digital rather than analogue form may be equally obvious, but bears contemplating nevertheless. That is that digital textuality is a hybrid phenomenon. It is a replacement for writing and printing – that is to say, a medium capable of replacing many functions of those much more familiar media – and at the same time digital text, because of its inherent instability, can be subjected to all of the programming capabilities of the computer as a Universal Machine. These capabilities take digital textuality far beyond the medial functions of the book and make it into an extraordinarily versatile research instrument – or, better, sociotechnical research envi-

*ronment. Medium and research environment are, moreover, so seamlessly integrated in the same computing environment as to be virtually indistinguishable.*⁷

To illustrate what I mean by this continuing computability take the (deliberately very simple) example of full-text searching. Digital texts can of course be found as whole entities, i.e., no different from the integral physical items library catalogue records refer to. Such items (let's call them books) may contain an index created by the author, or an editor or publisher, accessible upon retrieval of the item from the library shelves. However, in a full-text search of the docuverse a user 'computes' the docuverse of texts on the internet in order to look inside all of them at once, bypassing the conventional means of access through library catalogues that only lead to the entities as a whole, and indexes that only become available upon retrieval of each individual item.

3. In combination with the fluidity of computable text the two-way server-client architecture of the computer-in-a-network has enabled a distributed digital knowledge creation and exchange environment (i.e., what in the world of scholarship is now called e-science, cyberscience, etc.). The so-called collaboratory is perhaps the most familiar concrete example.

Having identified what I regard a number of fundamental discontinuities as such, I would like to single out some salient *consequences* they might carry for the way the knowledge system functions. In the digital information environment *more* things can be done than in its analogue counterpart, but I would like to emphasise that even when we appear to be talking about doing the *same* things, in crucial ways these are done *differently*. Why do these differences – the discontinuities I have just identified – matter?

1. The multimodal nature of the digital environment, seamlessly integrating text, still images, graphs, moving images, speech and music, enables very different forms of knowledge inscription than the preponderantly textual one that has reigned supreme for so many centuries. Ultimately this might spell the end of the privileged position of text as the basis for the dissemination of knowledge.

2. Full text searching. That digitisation (and a fortiori mass digitisation) creates new ways of access, with full text searching taking the place

*of indexing can, as I suggested, easily be constructed as a continuity. However, its consequences are potentially revolutionary, for example because to a large extent it bypasses the experts that conventionally create and guard access to textual resources: bibliographers and librarians. That the experts are up in arms about the lack of quality and integrity both of the digitised data themselves and of the alternative access offered to them is certainly not deterring users.*⁸

That the same computing environment in which digital texts always remain computable functions both as a medium and as a workspace where texts can be manipulated in any way, *and* that this environment encompasses the entire internet has some truly revolutionary consequences:

*3. The multiplicity of authorship. The Web's collaborative means of knowledge production favours multiple authorship, leading to what may be called a 'granularity issue' in attributing individual contributions (which may be very small) to named actors. There are – at least practical – limits to the scope of such attributions. This obviously threatens conventional notions of both authorship and ownership on which the libroverse has come to depend.*⁹

4. The porosity of the boundary between actors inside and outside the traditional knowledge system. In Web 2.0 the process of knowledge creation has spread to include all web users, also inviting 'amateur scholarship'. Commenting and tagging of web resources by general users regardless of formal qualifications are good examples of such 'democratic' knowledge production. The mixed provenance from inside and outside the traditional knowledge system leads, for example, to questions about the assessment of quality.

5. The porosity of the boundary between knowledge that is the result of a process of 'digestion' (rational deliberation, or however one may wish to define scholarship) and knowledge that is left implicit in the raw materials presented. Digital editions proffer a good example. These are often presented as archival collections of primary data that remain to be 'computed'. That is to say, there remain choices to be made by the user, who needs to establish the relative value of the various raw materials. The rawness of these 'scholarly semi-manufactures' leads to what I have called elsewhere the 'deferral of the interpretative burden',¹⁰ i.e., from the editor to the reader.

6. *The porosity of the boundary between object and tool. As a result of the continuing computability of the digital text, the digital knowledge environment is focused strongly on the process, both of knowledge creation and of knowledge dissemination, rather than on the outcome (the frozen contents, or ‘monuments’ of knowledge familiar from the Order of the Book).*¹¹

7. *The rate of growth of digital data. Both absolutely, as the population grows, and relatively, as a percentage of the population, the number of people thronging to gain access to the net continues to grow, and with it the power to add to or change the digital textual record.*¹² *This growth is hardly hampered by any physical constraints: the cost of storage memory continues to drop at about the same rate as computing power is growing. If one adds to this the versioning challenge that results from the fact that these data remain in a permanent state of flux, one can readily imagine the Sisyphean nature of the task of maintaining some semblance of bibliographic order.*

8. *The rate of change of the computing environment. Computers being universal machines, new digital technologies will keep being invented. The rate of such inventions will only accelerate. Already there is a continual influx of new technologies before there is even time to understand properly the nature and implications of existing ones. Any future is unforeseeable, but the problem is that the unforeseeable future is no longer experienced as being far ahead – or even in the future at all: it is constantly with us now. This means that, whether as scholars, publishers, librarians or archivists, however hard we are trying to dam the tide of innovation and change, through such aids as metadata, controlled vocabularies, ontologies and preservation schemes, we are bound always to lag behind.*

Not a conclusion

The concept of a conclusion being rather alien to the nature of the docuverse, which is characterised by lack of closure, I won't attempt one. However, I *would* like to suggest that after five centuries and a half, we might need to give up the book as our chief model for the production and dissemination of knowledge. Incidentally, I don't mean this at all as an apocalyptic view; I like to think that I am rather matter of fact about it.

Conditioned by centuries of print, Western man has evolved to be *homo typographicus*, and the apparent continuities between analogue and

digital text (expressed at the most basic level in that both are character based) are apt to mislead him. In reality the universe of digital texts is more alien to our sensibilities than we have been ready to admit – or have even begun to recognise. The discontinuities are significant and the very different nature of the digital textual medium represents a major challenge to our knowledge system, which is firmly based on the analogue practices of a pre-digital information architecture shaped by the particular nature of the printed book.

I think we need to recognise the prominent role of technology in causing major discontinuities in our knowledge system.¹³ Our very concept of knowledge has been shaped by the Order of the Book, that is to say, by properties that resulted from the technological affordances of print that I listed at the beginning of this essay. It would be naive to think that the advent of the digital medium would not affect it, in the various ways I have suggested, and others. For example, the digital media are eating away at the strong time-honoured connection between what is regarded as knowledge and the Order of the Book's system of attribution and authority.¹⁴ This means that one of the effects of digital textuality is to question the very concept of what knowledge is.

We live in a transitional era, between an orientation towards the past (the Order of the Book) and a digital future of some sort. In this transitional era the analogue knowledge system, which by dint of its long history continues so far to dominate our thinking, is straining to accommodate the products of the digital textual medium. For this, however, it is not well suited. Our instinct continues to be to create order and control in the spirit of the Order of the Book. The more the digital information environment resists this, the more our ingrained typographical habits are in danger of becoming like a straitjacket or, worse, like a Procrustean bed.

At the same time there is (as yet) no digital knowledge system. In fact, I fear that it may simply prove impossible even to design a digital knowledge *system* – at least one bearing any resemblance to a system as we conceive it. The technological nature of the digital medium, as well as the use we make of it socially, are too much at odds with the knowledge paradigm of the book. So perhaps the wheel needs to be reinvented after all – and it may not much resemble a book wheel – or any other wheel that we know.

In the meantime, the knowledge system – or those who continue to believe in it – may be fighting a rearguard battle in trying to assert its continuing value and usefulness. The *coup de grace* as I see it is the strong pressure that is building up even from within the traditional knowledge

system to open the floodgates to methods and products from outside the knowledge system.¹⁵ If this pressure was not ultimately caused by the nature of the digital medium itself we might perhaps conspire to ostracise the upstart authors. As it is, the very properties of the digital knowledge environment predict the likeliness of the system yielding to this pressure.

It was always one of the key features of the traditional knowledge system that it tried to defend itself against impostors and interlopers. However, a new knowledge production paradigm may need to set aside traditions of gate-keeping, selection and control, the checks and balances of the analogue knowledge system, based on hierarchy and experts. Instead it might be left to the end user to analyse and make sense of huge quantities of undifferentiated data.¹⁶

We are closer than ever to bringing together the record of all human knowledge. However, we are rapidly losing our control over it. So I would like to make a case for adopting a much messier concept of knowledge.¹⁷ Stephen J. Gould advocated an understanding of the descent of man that substitutes for the image of a tree (a hierarchical image if ever there was one) that of a bush. Just as he maintained that humans are but a 'tiny, late-arising twig on life's enormously arborescent bush – a small bud that would almost surely not appear a second time if we could replant the bush from seed and let it grow again,'¹⁸ we need to revise our hierarchical view of knowledge. We might have to replace it with a more rhizomatic view, lacking all order or system. This new knowledge paradigm is a 'free for all,' where the 'system' is replaced by the network. In negotiating this massive network of information – this 'world brain' – we may need to learn entirely new ways to extract the knowledge that we seek. Rather than rely on ready-made knowledge attributed to individual minds, we may need to explore our own pathways through the collective hoard of facts and opinions, for example using hosts of tiny bots and other assorted artificial intelligences to 'make sense'.

The problem with this vision is of course that we will gradually have to let go of our innate desire for control. We have to abandon a system that has constructed our view of knowledge, our view of the world, and ultimately our view of ourselves for a very long time, but is now becoming antiquated. This may not be easy, but if an eighteenth-century Anglican clergyman could do it, so can we. For Laurence Sterne's *Tristram Shandy* must serve us as a remarkably early and very instructive deconstruction of the book as a knowledge machine:

Thus – thus, my fellow-labourers and associates in this great harvest of our learning, now ripening before our eyes; thus it is, by slow steps of casual increase, that our knowledge physical, metaphysical, physiological, polemical, nautical, mathematical, aenigmatical, technical, biographical, romantical, chemical, and obstetrical, with fifty other branches of it (most of 'em ending as these do, in -ical), have, for these two last centuries and more, gradually been creeping upwards towards that 'Akun' of their perfections, from which, if we may form a conjecture from the advances of these last seven years, we cannot possibly be far off.

When that happens, it is to be hoped, it will put an end to all kind of writings whatsoever; – the want of all kind of writing will put an end to all kind of reading; – and that in time, – As war begets poverty; poverty, peace, – must, in course, put an end to all kind of knowledge, – and then – we shall have all to begin over again; or, in other words, be exactly where we started. (Laurence Sterne, Tristram Shandy, Chap. 21)

And if we have any doubt whether we are capable of willingly destroying our carefully constructed knowledge machinery, let's face it, the new generation is already doing it. They appear very happy to forego the guidance of their elders and betters, and to be comfortable in finding their own way through the jungle.

References

- 1 See Anthony Grafton, *The footnote: A curious history* (London, 1997).
- 2 A phrase inspired by the title of Roger Chartier's *L'Ordre des livres* of 1992 (translated into English by Lydia Cochrane as *The order of books: Readers, authors, and libraries in Europe between the fourteenth and eighteenth centuries*, Stanford, 1994).
- 3 See, for example, Françoise Levie, *L'homme qui voulait classer le monde: Paul Otlet et le Mundaneum* (Bruxelles, 2006).
- 4 Vannevar Bush, 'As we may think,' *Atlantic Monthly*, July 1945, pp. 101–8.
- 5 The latest edition is *TEI P5: Guidelines for electronic text encoding and interchange*, ed. Lou Burnard and Syd Bauman (Oxford, Providence, Charlottesville and Nancy, 2008).
- 6 George P. Landow, *Hypertext: The convergence of contemporary critical theory and technology* (Baltimore and London, 1992) (rev. edns 1997, 2006); Jay David Bolter, *Writing space: The computer, hypertext, and the history of writing* (Hillsdale, N.J., Hove and London, 1991) (rev. edn., 2001); Richard A. Lanham, *The electronic word: Democracy, technology, and the arts* (Chicago and London, 1993). Theodore Holm Nelson's *Literary machines* of 1981 preceded them by a long stretch, but Holmes was too visionary to be taken seriously by more than a handful of followers. It was the invention of the World Wide Web by Tim Berners Lee in 1990 that served as the foremost catalyst for his ideas.
- 7 See Adriaan van der Weel, 'New mediums: New perspectives on knowledge production,' in Wido van Peursen et al. (eds), *Text comparison and digital creativity: The production of presence and meaning in digital text scholarship* (Leiden: Brill, 2010), pp. 253–68.
- 8 As regards searching and finding it is worth noting, incidentally, that the multimodality of the digital environment, where all modalities mix freely, is creating an imbalance in this respect. Where search capabilities are concerned, some modalities turn out to be more equal than others. Time-based media, which are on the increase also in the scholarly sphere, require a different way of access than text, and they cannot (yet) be searched in the same way as text-based ones. (See, for example, the *Journal of visualized experiments*, <http://www.jove.com/>, which was established as a new tool in life science publication and communication, to visualise the 'intricacies of life science research': 'Visualisation greatly facilitates the understanding and efficient reproduction of both basic and complex experimental techniques, thereby addressing two of the biggest challenges faced by today's life science

research community: i) low transparency and poor reproducibility of biological experiments and ii) time and labour-intensive nature of learning new experimental techniques.' I am indebted to Ernst Thoutenhoofd for drawing my attention to this phenomenon.)

- 9 Recently on TEI-L, the Internet discussion list of the Text Encoding Initiative, under the heading 'question about translators in biblFull' a discussion took place on the relative significance of contributions varying in nature and extent to the intellectual content of a work. The example was given of 'crowd-sourced translations' in a social networking site, where many individuals contribute units of information too small to document (TEI-L, 27–29 May 2009).
- 10 Adriaan van der Weel, *op. cit.*, p. 263.
- 11 This is also implied in the view of Clifford Lynch: 'The role of libraries, he argued, will shift from primarily acquiring published scholarship to a broader role of managing scholarship in collaboration with the researchers that develop and draw upon it.' (Diane Goldenberg-Hart, 'Libraries and changing research practices: A report of the ARL/CNI Forum on e-Research and Cyberinfrastructure,' *ARL: A bimonthly report on research library issues and actions from ARL, CNI, and SPARC*, 237 (December 2004), pp. 1–5, at p. 1).
- 12 *Time* magazine characterised this development well in 2006 by electing as their 'person of the year:' *You* (*Time*, 13 December 2006).
- 13 Clifford Lynch stated a few years ago that 'massive changes in scholarly practice are occurring across all disciplines.' He argued that 'new practices, products, and modes of documenting and communicating research will have far-reaching implications for all organizations involved in managing the scholarly record and supporting the ongoing enterprise of scholarship' (Diane Goldenberg-Hart, *op. cit.*, p. 1).
- 14 Referred to as 'reputational control mechanisms' by Virtual Knowledge Studio (Paul Wouters, Katie Vann, Andrea Scharnhorst, Matt Ratto, Iina Hellsten, Jenny Fry and Anne Beaulieu), 'Messy shapes of knowledge: STS explores informatization, new media, and academic work,' in *The handbook of science and technology studies*, eds. E. J. Hackett, O. Amsterdamska, M. Lynch and J. Wajcman (Cambridge, MA), pp. 319–35, at p. 333.
- 15 Cf. James Surowiecki, *The wisdom of crowds*, New York, 2004. Cf. also Cass R. Sunstein, *Infotopia: How many minds produce knowledge* (OUP, 2006) and David Weinberger, *Everything is miscellaneous: The power of the new digital disorder* (New York, 2008). Also compare the fascinating experiments conducted by Luis Von Ahn as described in

his lecture 'Human computation' (Google TechTalk, 26 July 2006, <http://video.google.com/videoplay?docid=-8246463980976635143>).

- 16 Cf. what is already happening in the case of digital text editions, as I have observed above.
- 17 I am indebted for that phrase to Ernst Thoutenhoofd, who alerted me to one of the Amsterdam Virtual Knowledge Studio's research projects, called 'Messy shapes of knowledge.' See, for example, Virtual Knowledge Studio, *op. cit.*
- 18 Stephen J. Gould, 'The evolution of life on earth,' *Scientific American*, 271 (October 1994): pp. 85–91 (also at <http://brembs.net/gould.html>; I am indebted to Ruud Hisgen for pointing out this parallel). That it 'would almost surely not appear a second time if we could replant the bush from seed and let it grow again' offers a nice parallel with the fact that certain knowledge creation processes, too, are becoming so complex as to be to all intents and purposes unrepeatable, which was one important rationale for the foundation of the *Journal of visualized experiments*.

Digitization: Why, How and For Whom?

DO LIBRARIES DREAM OF ELECTRIC SHEEP?¹

Lisbet Rausing

Imagine a New Alexandria. Imagine a library that contains all the natural and social sciences of the West – our source-critical, referenced, peer-reviewed data – as well as the cultural and literary heritage of the world civilisations, and many of the world's most significant archives and specialist collections. Imagine that this library is electronic and in the public domain: sustainable, stable, linked and searchable through universal, semantic catalogue standards. Imagine that it has open source ware, allowing legacy digital resources and new digital knowledge to be integrated in real time. Imagine that its Second Web capabilities allowed universal researches of the bibliome.

Well, why not imagine this library? Remarkable electronic libraries are being assembled. Google Books aims for about 16 million books.² The non-profit Internet Archive already has some 1 million volumes. Public expectations run ahead even of these efforts. To do research, only one in a hundred US college students turn first to their university catalogue. Over 80% turn first to Google. But we cannot rely on Google to build the New Alexandria. Google has pioneered technical knowledge. Over the medium term, it might (perhaps) be relied on to provide stable funding as well. But a New Alexandria demands a centuries-long ethos of digital conservation, scholarship and public access. By definition, even, and certainly in practice, for-profit companies do not have that ethos. It is found mostly – or most safely – in those non-profit institutions that have lived on for decades or centuries.

If New Alexandria is to be built, it needs to be built for the long term, with an unwavering commitment to archival preservation and the public good. A true public good, it probably needs to be largely governmentally funded. And, while a global and co-operative venture, it needs to be hosted by one organisation that is reputable, long-standing, non-profit and in a stable jurisdiction. The Library of Congress, the flagship institution of the world's only surviving Enlightenment republic, comes to mind. There might be other possibilities, such as the New York Public Library, or the British Library, or a consortium of the world's leading university libraries – UCLA, Harvard, Cambridge University and so on.

Building New Alexandria is not a question of technology. We live in the age of electronic reproduction. The technological future is certain. It is being magically created, at dazzling speed, by scholars, entrepreneurs and amateurs, in a Schumpeterian process of creative destruction. Even our most traditional products of learning (monographs,

academic articles and so on) are disintermediating at an ever-faster pace. As marginal costs of replicas approach zero, what constitutes an archive, publisher, bookseller, or even consumer is all put in question. Things fly apart at the seams.

This challenges those of us on the other side of the divide – the state-builders, gatekeepers, academics, politicians, civil servants, librarians and non-profit workers. Do we have the imagination and generosity to collaborate? Can we build legal, organisational and financial structures that will preserve, and order, and also share and disseminate, the learning and cultures of the world? Scholars have traditionally gated and protected knowledge, but also shared and distributed it, in libraries, schools and universities. Time and again they have stood for a republic of learning that is wider than the ivory tower. Now is the time to do so again. We must stand up, as the Swedes say, for *folkbildningsidealet*, that profoundly democratic vision of universal learning and education.

The question for scholars and gatekeepers is not whether change is coming. It is whether they will be among the change-makers. And if not they, then who? Who else will ensure long-term conservation, and search abilities that are compatible across the bibliome and over time? Who else will ensure equality of access? Again, this is not a challenge of technology, finances, or ultimately even laws, difficult though they are. It is a challenge of will and imagination. This article discusses three such challenges that libraries face: the cult of the artefact, the problem of abundance and the question of the audience.

Let me begin with the cult of the artefact, which is really a story of our imaginary horizons. Our iconic library stories are romances of destruction, decay and amnesia. We tell tales of time, fire, and barbarians, and of heroic rescues of fragments of lost and esoteric knowledges. We still mourn Alexandria. We revere St Catherine's Monastery, the Vatican archives, and the Dead Sea Scrolls. We grieve over the Christians closing the academy of Athens, and we listen in horror to the tales of the fall of Constantinople, where in desperation the last Grecian scholars lit the cannons with their manuscripts. Who among us has not lamented, with Aenaeas Silvius (Pius II), that Homer and Plato have now "died a second death?" Boethius, the monks of Iona, the fleeing Byzantine humanists – these are our heroes and role models.

Look only at the annual reports of the important libraries of the West. Why else do they concern themselves so centrally with the single and exceptional object – antique maps, sheet music, letters penned by the

hand of the cultic genius? Why else are the purchases of scholarly databases hidden away in sub-clauses and footnotes? Should we rejoice when dwindling library acquisition budgets are spent on “rare books,” “rare,” admittedly, but not in a meaningful sense threatened or endangered?

Should we perpetuate the cult of the book as fetish? What is the value – to the scholarly community, to the public at large – of the proximity and materiality of the object? And how do we price that intangible “there”ness? Does it matter that incunabula are held at Rice University, rather than Oxford University? Why? And what do we Europeans mean, when we code our purchase of a manuscript into having “saved” it – saved it, most often, from being cared for by a well-endowed and well-intentioned US institution? What does that code of “rescue” denote, except a Benjaminian cult of the physical artefact?

Throughout history, libraries have depended on destruction. Today, in an era of electronic abundance, they still operate within an increasingly imaginary economy of scarcity – fragments, incunabula, manuscripts, rare books. Libraries were set up to house pricey collectors’ items, painstakingly recorded sets of symbols, crafted sometimes by hand, sometimes in block print, and sometimes in movable type, but only very recently (remember the last printers’ strike of Britain) conjured up from the bowels of computers. Once, books were chained to the wall. Today, print is an afterthought: “Do you want a receipt with that?”

In today’s era of electronic abundance, how can libraries archive the dreams and experiences of humankind? What do we discard? And if a library can no longer be understood as a warehouse of treasures, a primitively accumulated *Schatzkammer*, what is it?

Take Harvard, the world’s largest university library, as an example. Its 16 million volumes rival those planned for Google Books. One took nearly four hundred years to achieve. The other, less than a decade. Such histories matter. Harvard’s institutional memories are those of 1638, and 400 books. As late as the mid-nineteenth century, it was a stated duty of its Overseers to count its volumes. In those days, Harvard’s books were threatened by fire and water (as London’s booksellers’ wares crossed the Atlantic on clippers and schooners). Today, our sailing – or counting – skills mean little.

All good research libraries are hierarchically organised. The core/reference section, with bibliographies, dictionaries, encyclopaedias, library catalogues and so on, is rapidly dematerialising, as it moves into “the cloud.”³ So is much of the record of scholarship, especially in the

natural sciences, and at least some of the record of the human experience.

But what about grey data, such as laboratory notebooks, lectures, conference proceedings, dissertations, data sets, course works? Is it not the task of libraries to preserve the processes as much as the products of knowledge? How else can we test it? Or indeed comprehend it historically? The papers of Newton, Darwin, Einstein, and Bohr can be (and indeed are being) produced in toto. But what about “big science”? The ATLAS detector of the Large Hadron Collider at CERN – that 27-kilometre underground ring tunnel searching for the Higgs boson – takes 90 million measurements 600 million times a second, and these are analysed by some 6,000 high-energy physicists. Worldwide, scientific data files are approaching a petabyte⁴ and they double every year. Even artisanal lab skills, once handed down by lore and practice, are now recorded on wikis.⁵ What is to be preserved? By whom? For how long? How do we process, calibrate, re-organise, analyse and store our data?⁶ What do we do, when our software, let alone our brains, cannot keep up? What do we do when bits degrade, software and hardware go extinct, and cyberspace turns out to be a decaying maze?

Scholars rightly argue that we cannot meaningfully analyse our peer-reviewed knowledge without also archiving its primary sources. But today’s knowledge quest is universal: our primary sources encompass all the knowledge, hopes and dreams of humanity. Our Alexandria was not burnt, our Byzantium still stands, and our Athenian academies are blossoming. And next to the near-infinitude of our scholarly endeavours and their materials, we want to preserve that which we have not yet incorporated into our learned canons: the near-extinct and the barely remembered, the oral traditions and the dying languages, the esoteric and the sacred, the reviled, even, and the persecuted. We want the Nazi state papers *and* the Lodz ghetto archives, the Soviet encyclopaedias *and* the samizdat literature, the Maimonides commentaries *and* the Genizah fragments, the Ethiopians’ church songs *and* their memories of the recent famines.

Next to the rare, well-studied, cultic artefact – the letter by Jefferson, the Magna Carta – we want ephemera: pamphlet literature, theatre bills, immigrant broad sheets and poetry workshops. And we are right to want ephemera. We have belatedly realised that humankind understands only poorly what will last through the ages. Think of John Clare, Emily Dickinson or Barbara Pym. Or think of Isaac Bashevis Singer. A Nobel Prize winner in 1978, yet what is arguably his masterpiece, *Shadows on the Hudson*, was only translated in 1998, more than 50 years after it was

serialised in a Yiddish magazine, *Vorverts*. And the premise of the novel is the self-evident fact, as seen by Singer and his community just after the War, that the Holocaust would soon be forgotten. Perhaps it could have been. Raul Hillberg's *Destruction of the European Jews*, the seminal scholarly work on the Holocaust, was begun in 1948, but it was only published in 1961, and by an obscure press. Reputable university presses turned it down as being a topic unlikely to be of general interest, and it was translated into German only in 1982.

What if our next "peasant poet," as John Clare was known, twitters?⁷ What if he writes a blog or a shojo manga? What if he publishes via a desktop, or a vanity publisher? Will his output count as part of legal deposit material? What if there is a masterpiece being filmed in Bollywood? What if one among many Nigerian novelettes, which typically address a young heroine's agonised choice between a village boy and a "big man," turns out to be written by a Jane Austen? And even if none are, don't we want to preserve them all regardless, so that one day we can run larger studies on them, studies perhaps as yet unimaginable, because they depend on computer uses not yet invented? Moreover, to investigate very large datasets – whether texts, numbers, or images – is a job for consortia, and beyond the capacity of any one library or university, especially if the data to be mined is raw (unorganised) – digital satellite imagery, census data, survey responses, and so on. Moreover, such studies might engage, not only university-affiliated scholars, but also the community.

What is the library, when the totality of experience approaches that which can be remembered? When we no longer preserve only those fragments that time, fire and barbarians have left us? When we no longer are able to safeguard only remnants of our discourses on thought, memory and images, but thought, memory and images themselves and complete? When we shall have, not only the *Lives of the Most Excellent Painters, Sculptors, and Architects*, but also Vasari's blog, Wiki, Twitter, texts, emails, chatroom, Facebook, radio interviews, TV appearances, and electronic notebooks?⁸ What happens when people think the "canon" means an online strategy war game, or a shojo manga? Or – the Wikipedia debates on this open up a fascinating alternative world of scholarship – when the "canon" means emergent rules integrating new materials – including "fanon," or materials authored by fans – into ever-expanding, fictional universes, such as Star Wars and Star Trek? The question is not: what is the library in the era of the internet (1974), the web (1991), or Google (1998)? It is: what is the library in its Second Life?⁹ When we begin to

see multiplying alternative realities such as SciLands? Today we bracket them as isolated trivia – Grand Theft Auto, Sims or the United Confederation of Interstellar Planets. But soon, alongside Vasari’s blog, they will be, not only California, but our earth, dreaming.

In 2008, the web’s founder, Tim Berners-Lee, reflecting on his topsyturvy child, noted that the web’s vast emergent properties are perhaps best modelled by biological concepts, such as plasticity, population dynamics, food chains and ecosystems.¹⁰ But how do we understand the web, when this also means grasping its quasi-biological whole? Do libraries dream of electric sheep? For that matter do electric sheep dream of libraries? Who will preserve? Who will be preserved? How will we tell the difference? Will Simfrog 2.0 be conscious? Can we live at once in nature, Second Nature, Second Nature 2 and Second Nature 3?¹¹ Will Second Life take on life – and if so, what will be its – and our – library?

I started this article by asserting that we could build a New Alexandria. As the open web movement has it, an old tradition and a new technology have converged to make possible an unprecedented public good.¹² The “new technology” means that the marginal costs of electronic replicas are now nearly zero, triggering a gloriously chaotic disintermediation. Think only of Kindle/Amazon, Google Books, the Espresso Book Machine, or Mills & Boon’s e-books.¹³ But the “old tradition” is less often discussed. Scholars publish without direct pay, for the sake of knowledge, and with peer recognition and social utility as their reward. In practice, peer recognition reigns foremost. Most scholars are only mildly interested in widening their audiences. This matters, for scholars run archives and libraries, and they run them according to their lights. These institutions do a fine job collecting, but their guardians mostly grant access to, well, fellow scholars.

When speeches are given, universities describe their mission as “producing, preserving, and propagating knowledge.” But in local governance parlance, the purpose of university libraries is to serve their faculty and students, and, when feasible, scholars at peer institutions. In other words, university libraries typically define their constituencies as those scholars formally associated with their universities. Not even alumni are mentioned. These narrowly defined constituencies are worth stressing, because many people think that the great university libraries set out to serve the public. They do not, at least not directly.

This matters, because the public today is not the public of fifty years ago. Okies, hillbillies, sharecroppers and mill workers may not have had

the energy or learning to engage with scholars. But today's public is educated and engaged. Indeed, it has proven its disinterested spirit of scholarship by participating in the collective projects of knowledge the technological rupture has enabled. The World Community Grid signs up volunteer computers. Other projects such as Wikipedia and SETI¹⁴ turn to volunteers via their computers. Through Folding@home, some 40,000 PlayStation 3 volunteers help Stanford scientists fold proteins.¹⁵ In foldReCAPTCHA, amateurs help digitise the *New York Times* back catalogue. In the ESP project, the public has labelled some 50 million photographs to train computers to think. In GalaxyZoo, some 160,000 people help astronomers at Johns Hopkins University and elsewhere to classify galaxies, and in Africa@home, volunteers study satellite images of Africa, to help the University of Geneva create useful modern maps. Conservation biology, a whole academic field, depends on amateur surveys, both outdoors and in historical collections. At Herbaria@home, for example, volunteers decipher herbaria held in British museums.

Much of this crowdsourcing, or mass voluntary participation, is “grunt work”: basic, lab assistant type work that often deals with image recognition.¹⁶ Scholars engage less with the “hive mind,” or the public, in more complex or interpretative work. But here, too, there are exceptions. For example, in Israel, the Rothschild family and others are pioneering a project to put the Dead Sea Scroll fragments into a public domain website, thereby engaging with religious communities that have unparalleled language skills. But by and large, the scholarly community has not made available to the public its “core” research material, such as, to choose a few examples, the House of Commons Parliamentary Papers, Historical Statistics of the United States Online, BMJ Clinical Evidence, Early English Literature Online, ehRAF Collection of Ethnography, Index of Christian Art, ProQuest Dissertations and Theses, Index Islamicus, Frantext, Oxford Music Online, ARTstor and Aluka.¹⁷ Try accessing these databases via Google instead of through your university account. It is a thought-provoking experience. Many make clear indeed that they are commercially owned and thus debarred to all except for those able to pay eye-watering fees. And even university-controlled collections are expensive. Take the “Index of Christian Art,” assembled by Princeton since 1917. There are vast, learned – and poor – Christian communities worldwide. Should this magnificent assemblage of digitised photographs be limited to those able to pay \$500 annual fees? It is free for Dumbarton Oaks Fellows, but even their spouses are only allowed to see the electronic database if they have “appropriate academic qualifications in his or her

own right". So much for familial economies of scholarship, and the rights of that generation of women who left college to get married, yet engage with their husbands' work. So much, too, for modern families – why make a gesture towards spouses, but not partners?

My examples of closed academic databases are random. I do not mean to single out anyone for special blame. But nor do I want to absolve anyone. The wider point is this: few academic databases and research tools are in the public domain, even though the public has paid for them – through research grants, tax breaks and donations. Nor is the higher-order academic commentary available to the public. It is arguably especially problematic that PhDs and M.A. theses are not in the public domain, given that these *Meisterstuecke* delineate those supposedly "appropriate" boundaries of access. In other words, the gate-openers remain hidden from those debarred from accessing that to which they open the gate. It is equally problematic that JSTOR, the splendid 1997 database of most twentieth-century scholarly articles in the social sciences and humanities, is off-limits for the public¹⁸ (although in fairness JSTOR's hands are largely tied, since it and indeed other academic knowledge managers face near-impossible copyright laws).

The academic databases have at least entered the digital realm. Public access – the right to roam – is a press-of-the-button away. But academic monographs, although produced by digitised means, are then, in what is arguably an act of collective academic madness, turned into non-searchable paper products. Moreover, both academic articles and monographs are kept from the public domain for the author's lifetime plus seventy years. My own PhD dissertation,¹⁹ published in 1999, will come into the public domain in about 110 years, around 2120.²⁰ And no matter what Congress might claim, I do not think my royalty earnings will be a big income for my grandchildren. I would rather reach out to fellow scholars and enthusiasts.

In any case, grandchildren's rights are not the issue here. If they were, Congress would not have applied the same, centuries-long lock-up periods to out-of-print works, where copyright holders and publishing presses can no longer be found. The public do not even have *allmansrätt*, to use the Swedish medieval term for the right to roam, on those vast thought-lands that lie fallow and abandoned. Because of copyright, few dare to adopt these orphaned works into the public domain, no matter how central they are for scholarship, or how interesting to the general public. Few dare to re-issue them even in paper format. Additionally, restrictive fair use rules mean that libraries that own a copy do not dare

digitise it for the public domain or even for their own constituencies. In the age of electronic reproduction, many books are legally enjoined to remain as few and as rare as Gutenberg Bibles.

As things stand, scholars sign over their copyright to for-profit academic presses and journals. Sometimes, in violation of their contracts, they also put it on their own websites. Publishers are not suing yet. It is a “don’t ask, don’t tell” stand-off. But that is hardly ideal. It means that free public access to scholarship, as far as it exists in fragments here and there, is based on a wholesale violation of copyright. And in any case, self-archiving is inherently unstable and transient. The legal profession rightly worries about judgements based on since-vanished references, and those of us who work in twentieth century history or the social sciences know the difficulties of citing ever-changing websites. Thus New Alexandria falters, most immediately on copyright legislation and market failure. The academic publishing market has bifurcated into a fragmented paper market for monographs and an oligopolistic electronic market, or cartel, for journals. The inflation rate for scholarly monographs is bad enough (and more academic books are published every year). But prices are hyper-inflating for commercial academic journals. Three firms control 85% of the periodicals market. Karl Marx and Adam Smith, both experts on the natural evil of monopoly, would nod knowingly on learning that an annual subscription for a scholarly journal can cost up to \$25,000, and that the price per page for commercial journals is up to 12 times more than for non-profit ones. And this is not because the for-profit journals are better. In the field of economics at least, the cost per citation is sixteen times higher for commercial journals than for those published by scholarly societies.²¹ And this is only counting subscription fees. Additionally, a *higher* proportion of closed access journals than open domain journals charge publication fees, and at the high end, they charge *more* than the most expensive open-access journal, PLoS, Public Library of Science.

After all, there are no substitute goods, and the purchasers of the journals (university libraries, but ultimately university administrators) are not the consumers (the professors and students). Thus publicly funded institutions first give away and then buy back their own research, research that they paid for in the first place. To add insult to injury, these for-profit journals are produced by non-paid, volunteer editors and peer reviewers. Here too, labour is donated for free, by those same scholars who also sign over their copyrights for free. It is, shall we say, an unusual business model. The producer gives away a product that he then buys

back after having helped the intermediary package it. It is no wonder that private equity companies circle these publishing companies. It is no wonder, either, that these publishers work hard to ensure regulatory capture. Congress is the academic publishers' most natural client and constituency, and – thanks to their alliance with Hollywood and the music industry – their success in locking up and rendering irrelevant the output of academic research has been nothing less than astonishing.

Robert Darnton, head librarian of Harvard and a renowned scholar, has rightly warned that what happened to journals will happen to books. The 2008 settlement between Google and the Book Rights Registry, after all, explicitly states its purpose is “to maximise revenues”. And while the US research libraries that participate in the Google-digitising project nominally retain a digital copy, they are banned from making this copy available even to their own members, let alone members of other participating libraries, or the general public.²² A recent *Financial Times* article agrees with Robert Darnton, warning that by means of the Books Rights Registry, Google and the publishing industry have created “an effective cartel,” with “significant barriers to entry.” New competitors are de facto barred from scanning books, and even if they were not, “Google’s effective most-favoured-provider status” would stifle competition. An “effective monopoly provider” always eventually charges monopolistic and discriminatory prices, the *Financial Times* notes, adding: “just as happened with academic journals in the past.”²³

Of course, there are signs of hope. Around 10% of Anglophone academic journals are now open access, and the “gold” ones are edited and peer-reviewed. Even scholars only seeking peer recognition are well advised to publish in them since, with prestige-factors equalised, citation rates are significantly higher from open access articles.²⁴ As Kevin Guthrie of Ithaca has noted, however, as long as journal and university press brands continue to be used as a proxy for quality in tenure committees, the stranglehold of commerce will remain. Yet this is unnecessary. After all, tenure committees read candidates’ work and canvass outside experts: the proxy is not really needed.

Other worthwhile initiatives to open up scholarship to the public are emerging too. Thanks to the pioneering efforts of Robert Darnton – efforts at times opposed by fellow giants in the field such as Tony Grafton – the Faculty of Arts and Sciences at Harvard has begun to put its members’ forthcoming scholarly articles on a public domain website, managed by the newly established University Library’s Office for Scholarly Commu-

nications. The Association of College and Research Libraries is searching for solutions to the periodicals crisis. The National Institutes of Health, which direct some \$29 billion a year for biomedical research, stipulate that their 325,000 or so grantees must publish their NIH-supported research in PubMedCentral. The UK's largest biomedical research charity, the Wellcome Trust, encourages open access, and the seven UK Research Councils are "committed to the guiding principles that publicly funded research must be made available to the public and remain accessible for future generations".²⁵ Dutch universities are pioneers in this field, not least in how they co-operate with each other. Physicists have run an open access pre-print archive for years, first at Los Alamos and now at Cornell. There is the Public Library of Science, the Open Knowledge Commons, the OpenCourseWare, the Open Content Alliance, the Internet Archive, Creative Commons, the Budapest Open Access Initiative and so on.

The great libraries of the West understand they can no longer compete against each other as to who can warehouse the most treasures. But if the collectivities of libraries are to remain the guardians of our patrimony, as they must, how do they divide that task between themselves? Increasingly (and encouragingly), they agree that stewardship must be joint, cross-unit and complementary – a mash-up, even. Innovations and ideas abound, such as joint rather than parallel collecting of duplicative materials, strengthening the Centre for Research Libraries and other membership organisations, interlibrary loan services, "joint-view" union catalogues, common licences and joint negotiations for e-resources, co-ordinated collection developments and storage protocols, etc.²⁶ These are matters of electronic knowledge management, and form contested sites, with uneasy and shifting alliances between IT support and library staff. And critical questions of governance remain. How does one manage outsourcing, leases and rents while still ensuring permanent access to permanent content? In a mash-up, who takes what responsibility for materials being captured, curated, preserved, ordered and delivered? Who plants the flag, asserting that we are here for centuries to come?

Yes, there are worthwhile initiatives to make scholarship public: but wider and deeper collective action is needed. We need a greater sense of urgency. We need more alliances, outreach and advocacy work. We need to embrace the neo-Gutenbergian shift, this disaggregating and democratising rupture of time and space, whose profound cultural significance and depth none of us have yet fully grasped.

Why not a legal nudge – a presumption of open access along the lines of presumed organ donor intent? Could copyright be revocable – a

lease, rather than a sale? Could copyright be deemed to automatically lapse when it stops generating income? At the very least, should not copyright have to be asserted and renewed, in order to remain in force? A more public-minded policy at the university presses would make a great difference, too. The presses could, for example, release their backlists into the public domain. Could university libraries be more imaginative? Could we make alumni lifelong members? Could the materials held by the Open Universities in England and Israel become, well, open? Could we develop pay-per-view portals into scholarly resources that are invoiced monthly and electronically? And in doing so could we, ahem, lower prices? *The Journal of Interdisciplinary History*, for example, optimistically charges \$10 for a book review,²⁷ and the average price for a JSTOR article – if you are lucky enough to find one the publisher is willing to sell – is approximately \$17.²⁸ Compare that to iTunes! Could we digitise out-of-copyright books on demand and for a small fee, so that members of the public could “liberate” their chosen books? Could university catalogues be turned into blogs? That is to say, could university members – or the public – add commentaries and hyperlinks? After all, views could be switched from catalogue-only, university-affiliate-commentary-only, and open commentary. And today’s filters remove defamatory or offensive comments. At the very least, if libraries are to continue in their traditional role, as reliable repositories of our cultural memories and collective knowledge, that is, if libraries are to become the spiders in the internet, their catalogues need to provide reliable URLs, backed by long-term maintenance policies and institutional guarantees. The alternative is to rely on Google’s search-engine algorithms, which is to say, on ephemeral beauty contests.

And can we not lobby better? Many in the open access movement were disheartened by the British Library’s response to the 2006 Gowers Review of Intellectual Property (by the Treasury). The British Library pleaded for unpublished works to have “only” a copyright of life plus 70 years.²⁹ It asked for permission to copy old sound and film recordings, since the then proposed extension of the 50-year music copyright to 95 years otherwise ensured the certain destruction of most of the British Library Sound Archive.³⁰

Could we be tougher? Could we name-and-shame, tagging out-of-print works with a “Congress/the EU/Parliament has banned this work from coming into the public domain”? Could academics put their own house in order? University teachers may not be able to put course materials into the public domain. But they can issue reading lists, and they can

YouTube the lectures as well as summarising them – or ask students to summarise them – on Wikipedia. Each one of us, in our own station, can help to open up scholarship to the public.

And we need to. Projects to open up scholarship to the public mostly pertain to the natural sciences, and mostly concern present scholarship. Twentieth century scholarship in the humanities and the social sciences is lacking. Authored by academics hoping, not for monetary gains, but renown among their peers and influence over the public, and financed by means of taxes and charitable gifts, this incomparable treasure trove is locked away from society by “The Sonny Bono Copyright Term Extension Act of 1998” (also known as the “Micky Mouse Protection Act”).³¹ It is an ironic fate – a second death if you will – for the great refugee scholars of Europe. Think only of Erwin Panovsky, Gershom Scholem, Kurt Gödel, Marc Bloch, Ludwik Fleck or Simone Weil.

Look at JSTOR (if you can). There you find the evidence-based, source-critical foundations of sociology, anthropology, geography, history, philosophy, classics, Oriental studies, theology, musicology, history of science and so on. They are all closed to the public. It is wonderful, of course, that high-energy physics and string theory are open to all. But is it not ironic that we have opened the gates only to that scholarship which few professors, let alone members of the public, have the cognitive capacity and appropriate training to grasp?

The opportunity costs for society are self-evident. But what about the opportunity cost for scholars? For example, the public has set itself the task to rewrite knowledge for the public domain through Wikipedia and the like. Should not these sites be hyperlinked with JSTOR?³² By excluding the public from their scholarly literature, academics make it impossible for amateurs to use sound research methodologies, critically examining evidence by cross-referencing and source analysis. Scholars then critique the public’s output for not being sufficiently academic. Academics commonly refer to the occasionally wobbly scholarly standards of Wikipedia as proof the public does not wish to pursue scholarship. Might it not instead prove that they do not let them?

Forget for the moment about the morality of thus adding insult to injury. Consider instead the downside for the universities. Does not the professoriate take a reputational risk? After all, the web tech community is working on how to verify information on the web, or as they put it, “engineering layers of trust and provenance.” In the longer term, the question is not whether the web will be scholarly in some perfectly mean-

ingful sense. It is whether traditional twentieth century scholarship in the humanities and the social sciences will be integrated into that emerging, increasingly cross-referenced and even more scholarly world of the web.³³ Or will what James Boyle has nicely termed our cultural agoraphobia – our undue scepticism of open networks – lead the universities to become bystanders in the new worlds of open access knowledge?³⁴

If scholars continue to hide away and lock up their knowledge, do they not risk their own irrelevance? Public domain access is often cast as debates over Africa and other out-of-the-way places. A more immediately important debate, I think, is to be had over how academics fail to engage with their natural constituency (and former students): journalists, business leaders, lawyers, entrepreneurs, politicians and civil servants. These people are the ruling classes, if you would like. They are the ones who house and feed professors. Is it really in academics' long-term interest to not let these well-educated and well-intentioned people as much as glance at, say, the Index of Christian Art? Is it really in their interest not to show the public their scholarly articles and academic monographs? What does this tell the public about who academics think is clubbable? And how will that affect how the public think about, say, federal research grants, or top up fees?³⁵

Half a millennium ago, at the dawn of the age of mechanical reproduction, German town folk were dazzled by the thought that, thanks to their new-fangled printing presses, God's word might now be put in the hands of the laity. There would be no need for intermediaries. God's word would speak, not through the clergy, but to each soul, no matter how humble his station on earth. Of course the intermediaries struck back – the Counter-Reformation was arguably just that, a rebellion of intermediaries. Indeed, Ireland retained a Catholic censorship until its belated modernity, a few decades ago. But the technological rupture of the printing press was such that the disintermediation was inevitable over the *longue durée*. We became – and look closely at the word – Protestants.

Today, at the dawn of the age of electronic reproduction, the intermediaries are again striking back. The publishers are the most blatant and crude, of course. But academics are also intermediaries. And while they may not think of it this way, arguably they too are striking back. Then as now, obstacles are imagined – and created. University libraries are closed shops, JSTOR remains blocked, theses are inaccessible, and academic monographs are available, if at all, only on paper and at prohibitive prices. For this sorry state of affairs, we should blame not only Hollywood and

the music industry. The obstacles to a true and electronic Reformation are real, but perhaps to be found also in the continuation of “business as usual,” perhaps ultimately founded in older folk’s mental difficulties in imaginatively re-drawing work practices, and organisational and legal “silos.” Remember Henry Ford’s comment: “If I had asked my customers what they wanted, they would have asked for a better horse carriage.”

However the research done in my field, the history of science, offers comfort in the morbid but accurate observation, ultimately traceable to Kuhnian theory, that “science marches ahead one funeral at a time.” Obstacles can delay, but not stop, a technological rupture of this magnitude. Excepting the odd Wykehamist or yeshiva boy, our children – always on, multi-tasking, mobile – will not engage with a body of scholarship their elders have incomprehensibly surrounded by barbed wire. But they will remain engaged in learning. The question is not whether there will be future scholars. It is how these future scholars will remember and integrate previous scholarship. And in pondering that, which means pondering our own scholarly legacy, it is worth remembering that “the generational war is the one war whose outcome is certain.”

References

- 1 I derive the title from Philip K Dick's futuristic novel *Do Androids Dream of Electric Sheep* (New York: Doubleday, 1968), which also formed the basis of the 1982 film *Blade Runner*.
- 2 Harvard has nearly 16 million items but about half of those are periodicals. About 7 million are books, and of those, three-quarters come from outside the US, although serious collecting abroad only started from the 1860s.
- 3 "The cloud," a dematerialised and outsourced network, consists of huge data centres with software applications used by millions of people at the same time. Yahoo, Wikipedia, YouTube, Twitter, Amazon, and so on are all built on such centres. Indeed, Amazon is transforming itself from a book seller to a cloud-space renter, in Amazon Web Services, which already uses more bandwidth than its retail side. Its Simple Storage Service has some 52 billion virtual objects. In manufacturing, a parallel to "the cloud" might be "outsourcing". A more homely example might be how your music experience moves from CDs, to JPEGs on your hard drive to Pandora, which is situated in a cloud. Feature length films are of course next. What household would not appreciate instantaneous rental films from "the long tail" (the entire backlist) of Hollywood, or for that matter all the other film industries of the world? *Herald Tribune*, 15 June 2009.
- 4 10 to the fifteenth power, or quadrillion.
- 5 Take for example the 2005 wiki OpenWetWare, started by biological engineers at MIT, which unexpectedly morphed into a vast manual of lab techniques, alongside its original function as a collection of laboratory notebooks. Mitchell Waldrop, "Science 2.0," *Scientific American*, May 2008, pp. 47–51.
- 6 Alex Szalay, "Preserving digital data for the future of eScience," *Science News*, 20 August 2008, p. 32. Szalay, a cosmologist at John Hopkins University, especially discusses the recently completed Sloan Digital Sky Survey, the world's most detailed map of the northern sky.
- 7 At present, Google worries about how its Twitter searches are indexed when they are a few minutes old, rather than in real time. *Herald Tribune*, 15 June 2009.
- 8 Facebook alone contains more than 15 billion photos, uploaded by more than 200 million users. *Herald Tribune*, 15 June 2009. By mid-2008, some 112 million blogs existed – although possibly this did not account for the 72 million or so in China. Nigel Shadbolt and Tim Berners-Lee, "Web Science Emerges," *Scientific American*, October 2008.

- 9 Second Life refers to an online virtual world where scientists have begun conducting real research projects, essentially on Darwinist theory, while taking on digital alter egos. It was founded in 2003 by Linden Lab as an open-ended platform where users (avatars) can create their own environment. It is, if you will, an open-ended SIM world, and at its height it had some 13.5 million users in mid-2008. By that date, its SciLands had grown into a mini-continent of some 45 islands – of, admittedly, nearly a thousand in all – inhabited not only by individual scientists but also by more of 300 universities and museums as well as by organisations such as NASA. Recently it has begun declining. *Science News*, 24 May 2008, p. 20–23.
- 10 Nigel Shadbolt and Tim Berners-Lee, “Web Science Emerges,” *Scientific American*, October 2008.
- 11 All part of Second Life.
- 12 Freedom for IP, “Budapest Open Access Initiative,” 19 November 2007: <http://freedomforip.org/category/open-access/>.
- 13 Harlequin sells about 200 million Mills and Boon romances annually. Its e-books, especially its romantic digital mangas in Japan, are selling very well. *The Week*, 2 May 2009.
- 14 For SETI (the Search for Extra Terrestrial Intelligence), with three million people donating spare computer time to seek for narrow bandwidth radio signals in space.
- 15 In 2007 it entered the record books as the most powerful distributed-computer network on Earth. *The Economist Technology Quarterly*, 8 December 2007, p. 16–18.
- 16 Crowdsourcing is also of course a tool for political activists. It is used to demonstrate corruption (by tracing the flights of Tunisia’s presidential jet), to find war criminals (in Darfur) or to advocate changes in the Catholic Church. *The Economist Technology Quarterly*, 6 September 2008, 8–10.
- 17 Cf.

The House of Commons Parliamentary Papers (Chadwyck-Healey) <http://parlipapers.chadwyck.co.uk/marketing/index.jsp> Username and password required. Available at a fee as a perpetual-access web service for universities, higher education institutions, public and specialist libraries.

Historical Statistics of the United States Online: <http://hsus.cambridge.org/HSUSWeb/toc/hsusHome.do> Username and password required. Users must pay for tables or essays.

BMJ Clinical Evidence: <http://clinicalevidence.bmj.com/ceweb/>

index.jsp. Username and password required. 12 month subscription (£142), 30 day season ticket (£25) and 48 hours (£5–£15) pay per view access to individual reviews are all available. Discounts offered for BMI members and first year medical students receive free access.

Early English Literature Online (Chadwyck-Healey): <http://eebo.chadwyck.com/marketing/trial.htm>. Username and password required. Available for a fee for universities, higher education institutions, public and specialist libraries.

ehRAF Collection of Ethnography: <http://ehrafworldcultures.yale.edu/ehrafe/> Restricted access. The server ehrafworldcultures.yale.edu at eHRAF requires a username and password before even allowing entry to the website. There is no link to subscribe.

Index of Christian Art: <http://ica.princeton.edu/> Institutional access is provided at a cost of \$1,500 per annum. Individual subscribers using their own non-institutional I.P. address are charged \$500 per annum.

ProQuest Dissertations and Theses: <http://www.proquest.co.uk/en-UK/catalogs/databases/detail/pqdt.shtml>. The style of this website is heavily commercial. It promotes a variety of services which can be bought by libraries and institutions. Visitors (libraries and institutions only – not individuals) are directed to contact an account rep.

Index Islamicus: www.indexislamicus.com: Username and password required. A free 30-day free trial (institutions only), consortia deals and other pricing options are available by sending an email to brillonline@brill.nl.

Frantext: <http://www.lib.uchicago.edu/efts/ARTFL/databases/TLF/> IP address access. Access restricted to ARTFL subscribing institutions using 'PhiloLogic'. Email address is provided on the website to arrange access for individuals within qualifying institutions.

Oxford Music Online: www.oxfordmusiconline.com: Username and password required. Log-in either through subscriber log-in or library barcode log-in. Online form needs to be completed for yearly subscriptions or free trial.

ARTstor: www.artstor.org: Username and password required. There is a registration link on the homepage (fees apply) and a link to list of participating organisations.

Aluka: www.aluka.org: Username and password required for some areas of site. Access was free of charge during the 'Preview period' until June 2008. It is possible to contribute and some articles still seem to be accessible without password, but the website says that licensing access and participation fees now apply.

- 18 JSTOR is said to have hundreds of millions of referrals from Google a year, the vast majority of which are refusals. There is considerable internet rage over JSTOR being closed.
- 19 Lisbet Rausing, *Nature and Nation in Linnaean Travel* (Harvard University Press, 1999).
- 20 A few years ago, the head librarian of Cambridge University said to me that they would not allow digitisation of post-1865 material for “copyright” reasons. It made perfect sense legally.
- 21 “Improving Scholarly Publishing Practice at Harvard: Report on the Provost’s Committee on Scholarly Publishing,” Harvard, p 5. The report notes that commercial publisher’s profits for scholarly journals are estimated at around 40%, an astonishingly high figure for any industry.
- 22 Robert Darnton, letter to the *New York Review of Books*, 26 March 2009. The sop within the agreement that each public library have one data terminal with free access to the Google database is meaningless. It allows for the use for a few of those lucky enough to have the resources and energy to reach a public library.
- 23 *Financial Times* (London edition), 19 June 2009.
- 24 A now-classic *Nature* article of 2001, “Online or invisible?” (Vol. 411, nr 6837) analysed c. 120,000 articles in computer science from 1989 to 2000. It found that, standardised for age-cohort, public domain articles had 4.5 times more citations. The correlation also held for top-end articles, from prestigious conferences.
- 25 Research Councils UK, “Access to Research Outputs:” <http://www.rcuk.ac.uk/research/outputs/access/default.htm>
- 26 Yale’s YuFind library catalogue tolerates natural language searches, integrates with external networked contact and has stable URLs allowing bookmarking.
- 27 The prices for book reviews in PDF Plus format came from the Pay Per Article feature of Volume 40, Issue 1 – Summer 2009 edition of the *Journal of Interdisciplinary History*, which is available at the journal’s website: <http://www.mitpressjournals.org/toc/jinh/40/1>
- 28 Ithaca, internal document written for Arcadia Trust, 23 April 2007.
- 29 A 2006 European Union directive stipulates copyright protection for life plus 70 years for authors of literary, artistic, cinematographic and audiovisual works. EUR-Lex, “Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the term of protection of copyright and certain related rights (codified version)”: <http://eur-lex.europa.eu/LexUriServ.do?>

uri=CELEX:32006L0116:EN:NOT. Literary works mean more than just fiction. The EU directive refers to “the rights of an author of a literary or artistic work within the meaning of Article 2 of the Berne Convention”. Article 2 of the Berne Convention states, “The expression ‘literary and artistic works’ shall include every production in the literary, scientific and artistic domain, whatever may be the mode of form of its expression, such as books, pamphlets and other writings; lectures, addresses, sermons and other works of the same nature; dramatic or dramatico-musical works; choreographic works and entertainments in dumb show; musical compositions with or without words; cinematographic works to which are assimilated works expressed by a process analogous to cinematography; works of drawing, painting, architecture, sculpture, engraving and lithography; photographic works to which are assimilated works expressed by a process analogous to photography; works of applied art; illustrations, maps, plans, sketches and three-dimensional works relative to geography, topography, architecture or science.” World Intellectual Property Organization.” Berne Convention for the Protection of Literary and Artistic Works: http://www.wip.int/treaties/en/ip/berne/trtdocs_wo001.html#P85_10661. EU copyright protection is automatic and does not need to be formally registered. BUYUSA.GOV, U.S. Commercial Service, “Copyright Protection in the European Union:” http://www.buyusa.gov/europeanunion/ipr_copyright.html.

- 30 Undated pamphlet: picked up at the British Library, “Intellectual Property: A Balance: The British Library Manifesto”.
- 31 To be compared to the first British copyright statute, the Statute of Anne in 1710, which set copyright at fourteen years, renewable only once. The need to renew copyright was removed in the US in 1992, and additionally copyright has become an assumed (rather than to be asserted) right.
- 32 Look up, for example, the eminent Natalie Zemon Davies in Wikipedia. The bibliography is good. But few of the entries are blue (linkable). Then look up, say, typhus, or any other major illness. You can hyperlink to recent medical research – but only rarely to the history of medicine references.
- 33 The Research Blogging website, started by Seed Media Group, aggregates and indexes posts on peer-reviewed data, and allows them to be tagged with metadata enabling priority of publication (*The Economist*, 20 September 2008, p. 96). The Transparent Accountable Datamining Initiative is at MIT and has a wide remit. The DBpedia project

was started at the Free University (Berlin) and Leipzig University. It semantically queries the infobox templates embedded in Wikipedia's (English) articles (2.3 million of them, as of late 2008). Nigel Shadbolt and Tim Berners-Lee, "Web Science Emerges," *Scientific American*, October 2008, p. 65 and passim.

- 34 I borrow the concept of cultural agoraphobia from James Boyle at Duke University, and from a lecture he gave at Cambridge University Library on 12 March 2009, entitled "Cultural Agoraphobia and the Future of the Library".
- 35 I count among my English friends a newspaper editor who is writing an authorised biography of a major political figure; a City man who is editing his letters from the Korean War; a CEO who is summarising his investment philosophy; and an MP who is pondering new legislation. None use, indeed none have ever heard of, JSTOR or other academic databases.

OPEN ACCESS FOR DIGITIZATION PROJECTS*

Peter Suber

When should digitization projects commit to open access (OA)?

I want to focus this question on public policy, not law or utility. If it were a question about law, the answer would be easy. As far as I know, there is no legal obligation in any country to make the results of any kind of digitization project OA. If it were a question about utility, the answer would also be easy, though the reverse. The results of a digitization project would always be more useful if they were OA.

Yet there may be good policy reasons to make some digitization projects OA even when not legally required, and there may be good reasons to change the law. Likewise, there may be good policy reasons to allow some access decisions to be made by stakeholders who will not choose OA.

Worldwide, more than 30 public funding agencies now operate on the principle that the results of publicly-funded research should be OA.¹

I started this essay to see how far I could defend the analogous principle that the results of publicly-funded digitization projects should be OA. The presence of public funding supports an OA argument in both domains. But digitization projects differ in OA-relevant respects more often than public funding agencies do, and even when they seem to be similar in all relevant respects, they frequently differ in their access policies. There is very little discernible pattern, and no matter what perspective we take, some of the policy divergence will be justified and some will not. This is a good reason to step back and think about the principles that ought to guide access policies for digitization projects.

Let me start with two relatively simple cases.

Case 1. When a digitization project uses public funds, and digitizes works in the public domain (PD), then the results should be OA.

For example, when Ontario digitizes the print editions of its historical statutes, it should provide OA to the digital editions.²

Case 2. When a project uses private funds, and digitizes works under copyright, then it should follow the wishes of the copyright holder. The results needn't be OA.

For example, when a private journal uses its own money to digitize recent back issues,

* Paper revised August 2009.

still under copyright, it needn't make them OA. It may put them online behind a paywall and sell access to them. Or it may keep them offline for its own private research purposes.

When *The Atlantic* digitized all 151 years of its backfile at its own expenses, chose to provide OA only to the most recent 15 years' worth, and toll access (TA) to the rest, then both the OA and TA parts of its project were entirely within its prerogatives.³

I'm sure you already see the supporting arguments for these two outcomes, but let me sketch them anyway. The principles behind them will help us navigate the issues in the more complicated cases.

The first case depends on the principle that public funds should be spent in the public interest. OA provides public access, and anything less than OA, or any access and usage restrictions, would compromise the public interest. The use of public funds obliges us to serve the public interest, and when we are digitizing PD works we encounter no barrier in the form of a copyright holder demanding access or usage restrictions. Taxpayers should not have to pay again for access to the digital editions. They should not pay to create an asset for the private enrichment of one citizen, one group, or one corporation, especially at the expense of the general public. Nor should they pay to create a digital asset which can only be accessed offline by the lucky few who are able to travel to a certain physical library or archive.

The second case depends on copyright law. Copyright holders have enforceable rights to their works, even if those rights are limited and temporary. Whatever the limits happen to be at a particular place and time, copyright holders should be free to exercise their rights up to the edge of those limits. They may waive or transfer their rights, of course, and it will be important that they might be asked to do so in order to enter a certain contract or use someone else's funds, especially to use public funds. But when copyright holders are using their own funds or the funds of a willing partner to digitize their own works, they should be free to offer the digital editions on any terms they please. The copyrighted backfiles of a journal might be more useful if they were OA. But I do not want to defend the idea that everything useful should be free, which would entail the abolition of copyright.

The principle of the first case leads us to applaud Ontario for providing OA to its digitized statutes, which are all in the public domain.⁴

Likewise, it leads us to criticize Oregon for falsely claiming copyright in the digital edition of its statutes and threatening to sue anyone who copied them. (This was Oregon's position until challenged by Carl Malamud in June 2008.)⁵

It leads us to criticize Pakistan for making the digital edition of its statutes freely accessible only to the country's lawyers rather than OA to all users.⁶

The British Library Digitisation Strategy 2008–2011 tells us that the BL plans to use public funds to digitize a mixed collection of PD and copyrighted works. Some of the digital editions will be OA and some will not. We can praise the library if the plan is to provide OA to the PD works. In Case 3 we will ask whether the use of public funds is enough to require OA even for works under copyright.⁷

JISC used public funds to digitize the backfiles of Oxford journals, which had already been supported by Oxford's own public funds. Whether JISC and Oxford should provide OA to issues still under copyright will be explored in Case 3. But under the principle of our first case, they should at least provide OA to any issues old enough to have passed into the PD. However, Oxford provides OA to none of the digitized backfiles – as opposed to more recent back issues which may have been OA from birth. (More below.)⁸

The principle of our second case leads us to conclude that *The Atlantic* didn't have to provide OA to any of its backfile, not even the oldest part which had passed into the PD. Its decision to provide OA to the most recent 15 years' worth is beyond the call, even if based on self-interest. Its decision to provide TA to the rest, especially to the PD issues, may prove difficult to enforce. (At least in the U.S., users may lawfully treat any copies which escape the paywall as works in the PD.) But as long as the journal avoids copyfraud, or the false claim of copyright, it should be free to try.⁹

The Dutch medical journal, *Nederlands Tijdschrift voor Geneeskunde*, is like *The Atlantic* except that it chose to provide OA to the oldest issues rather than the newest, using a five-year moving wall. Like *The Atlantic*, it paid for the digitization of its 150+ year backfile with its own funds (as far as I can tell). Like *The Atlantic*, it did not have to provide OA to any of it. Unlike *The Atlantic*, it does not have to try to restrict access to PD digital editions of PD back issues, which once online, may be copied and redistributed at will.¹⁰

Precisely because public funding pulls toward OA and private funding pulls toward the interests of the private funders, participants at the October 2007 LIBER/EBLIDA meeting in Copenhagen noted (in Recommendation 23) that "Private funding of digitisation activity may restrict access by the user, which is not in the interests of the European citizen. LIBER and EBLIDA do not see this as the preferred route for funding the digitisation of content."¹¹

Someone might object that some publicly-funded agencies should follow a cost-recovery model. The agencies have a mission to serve the public (the objection would continue), but they can best serve the public by charging for access, recovering their costs, and making their budgets go further. For example, this is the model of the Ordnance Survey, the UK mapping agency.¹²

In reply we can point out that several independent empirical studies conclude that OA stimulates significant economic activity, and that governments can generate much more revenue through taxes on that economic activity than through access fees on public data. In the case of research, this has been well-documented in several studies by John Houghton.¹³

For example, Houghton's first major study concluded that "With the United Kingdom's GERD [Gross Expenditure on Research and Development] at USD 33.7 billion and assuming social returns to R&D of 50%, a 5% increase in access and efficiency [Houghton's conservative estimate] would have been worth USD 1.7 billion; and... With the United States' GERD at USD 312.5 billion and assuming social returns to R&D of 50%, a 5% increase in access and efficiency would have been worth USD 16 billion."¹⁴

In the case of public data of the sort collected and sold back to the public by the Ordnance Survey, the UK Office of Fair Trading concluded that the cost-recovery model "cost the UK economy £500 million [per year] in lost opportunities."¹⁵

Even if Cases 1 and 2 are not themselves very simple or non-controversial, I want to use them to mark the two poles of a spectrum of cases which are even less simple. Here are three of those less simple cases.

Case 3. All the funds are public, but all the works to be digitized are under copyright.

In this case, the use of public funds pulls in favor of OA. But the copyright pulls in favor of the copyright holder. Should one side have its way at the expense of the other? If not, what compromise should we seek?

This case arises, for example, when a public agency like the US National Library of Medicine (NLM) or the UK Joint Information System Committee (JISC) funds the digitization of a journal's backfile, including issues still under copyright. When the NLM funded the digitization of the *BMJ* backfile, *BMJ* was willing to make the backfile OA without delay. The entire *BMJ* backfile to 1840 has been OA since May 2009.¹⁶

When JISC funded the digitization of the Oxford journal backfiles,

Oxford was not willing to make them OA, apparently even with a delay, although JISC did buy a license to the Oxford backfile for UK citizens. (The license will expire in July 2011. I cannot tell whether UK taxpayers, through JISC, paid once for the digitization and then paid again for the national license. Of course if the license is renewed in 2011, taxpayers will pay yet again; and if it is not renewed, they will lose their access.)¹⁷

For now let's focus on the case of a journal seeking a grant from a public funder, hoping to use the grant to digitize its copyrighted backfile and hoping to sell access to the online digital edition. It is the Oxford case, but artificially tidied up to eliminate the national license (close to OA for UK residents), TA for those outside the UK (one way in which the license falls short of OA), the limited duration of the license (another way in which it falls short of OA), and the possibility of multiple payments from the public funder.

We can imagine many kinds of compromise between the public and the rights-holding publisher. For example, we could make the works free of charge but not free for any sorts of use or reuse beyond fair use or fair dealing (see Case 5). We could make the OA copies low-res and the TA copies high-res. We could put ads on the OA copies. I mention these in order to stimulate the imagination. Over time the stakeholders may find many acceptable ways to strike the compromise, even if they also find many unacceptable ways to do it.

Here I want to focus on a compromise suggested by the analogy to publicly-funded research.¹⁸

In the case of publicly-funded research, the US National Institutes of Health (NIH) pioneered a compromise later followed by all other funding agencies with OA policies: a period of temporary exclusivity for the publisher followed by OA for the public. When NIH grantees publish articles based on NIH funding, they must deposit the peer-reviewed manuscripts in the NIH's OA repository (PubMed Central) as soon as they are accepted for publication. But the manuscripts are not made OA until after an embargo period of up to 12 months.

The delay is a compromise with the public interest, just as it is a compromise with the publisher's private interest. Because the embargo exists, publishers have a period in which to sell access to their priced editions without competition from OA editions. Because it is temporary, the public eventually gets public access to publicly-funded research.

Publishers who believe the NIH policy is not a fair compromise should seek a different compromise, for example by tweaking the embargo period, rather than demand a no-compromise position which could

deprive the public of OA for the full duration of copyright. While publishers have their reasons to lengthen the embargo, many other groups have reason to shorten it, among them researchers, practicing physicians, patients, non-profit organizations, and for-profit manufacturers. If both sides acknowledge the need for compromise, then their engagement on the length of the embargo, or on the precise terms of the compromise, is much more likely to be fruitful and constructive.¹⁹

The analogy of publicly-funded research and publicly-funded digitization should not leave the impression that the embargo compromise works the same way in both domains. We must note an often-overlooked aspect of the NIH policy. The NIH requires grantees to retain the right to authorize OA through PubMed Central. Hence, grantees are not in a position to transfer the full bundle of copyright to publishers. Publishers never acquire the right to deny permission for OA or claim infringement, and therefore cannot be called “the copyright holders” without qualification. Publishers who oppose the NIH policy understand the incompleteness of the transferred bundle of rights very well, and protest it. Nevertheless, in their lobbying rhetoric they call themselves “the copyright holders” without qualification, misleading many observers and policy-makers.²⁰

By contrast, in a digitization project we are often dealing with the full copyright holders. Nevertheless, the embargo compromise can be extended naturally to publicly-funded digitization projects.

Suppose a private journal applies to a public funder for funds to digitize its back run, and suppose that the entire back run is still under copyright. The funder would be justified in awarding the grant. At least the fact that the journal is private and under copyright need not stop it. The funder would also be justified in putting an OA condition on the grant. The grant need not require immediate OA and could allow the publisher a temporary period in which it could charge for access to the digital edition without competition from an OA edition.

More importantly, the public funder would *not* be justified in awarding the grant *without* the OA condition, or in using public funds to create a privately-owned asset which would exclude the public. Similarly, Oxford may use public funds to digitize the backfiles of Oxford journals, and it may sell access to the copyrighted issues for a temporary period. But after that the backfile must become OA.

How long should the embargo be? That should be decided by public debate and negotiation. But I have two rough criteria: First, the deal should give us OA sooner than we would otherwise have it. The public-

ly-funded digitization and OA condition will accelerate OA, while the embargo period will delay it. These conflicting trends should net out in favor of the public. If we could get OA faster some other way, then there is no reason to spend public money on the project.

Second, the longer the proposed embargo, the lower the project falls on the priority list for public funds. If the funder had to choose between two projects, one requesting a one-year embargo and another requesting a two-year embargo, then (other things being equal) it should pick the one with the shorter embargo. It might even ask the applicant proposing the shorter embargo to cut it even further to qualify for public funds. The US National Endowment for the Humanities follows the rule that, other things being equal, it will favor funding applications that promise (immediate) OA over those that do not promise OA at all.²¹

If the funder thinks a journal's proposed embargo period is too long, the journal might argue that it will still provide OA sooner than otherwise. For example, if the oldest articles it wanted to digitize would remain under copyright for another 50 years, and then it might argue that publicly-funded digitization with a 49 year embargo would give the public OA sooner than otherwise. As the copyright holder, it is in a position to insist that in the absence of public funding it will not allow OA until the expiration of copyright. The public funder need not deny the publisher's prediction or its good faith. It need only reply that it has better uses for its limited public funds than to create a 49 year monopoly for a private interest at the expense of the public.

The journal might object: "You can't require OA to our copyrighted articles!" The public funder would have several responses. "We can put conditions on our grant. You need not apply for publicly-funded grant. You can call this is an 'OA requirement' if you like, but it is really just a condition on a voluntary contract. Moreover, of course, we are a public agency and must spend our money to benefit the public."

A government would not be justified in making an *unconditional* requirement that journals provide OA to their backfiles, or at least not until it was ready to abolish copyright law. But it is fully justified in telling those who seek public funds for digitization projects, "If you take public money for this project, then you must provide OA to the results. If you don't like that, then don't take public money."

A member of the public might object: "You can't allow toll access to a publicly-funded work of digitization!" Again, the public funder would have several responses. "It is temporary. Moreover, we only funded the digitization, not the original work, and the original work is still under

copyright. But above all, in our best judgment, the public investment will make the work OA sooner than otherwise.”

Someone might object that under this rule many journals will not seek public money to digitize their copyrighted backfiles. Yes, that might happen. But it is no calamity, especially when the unpursued projects would have used public funds while excluding the public from access to the results. There is no reason why public funds should be spent on private interests unwilling to provide even delayed OA.

On the other side, for what it is worth, the prediction that many journals would rather reject both the compromise and public funds than accept both seems less likely than the opposite prediction. Allowing the private grantee a temporary period of exclusivity will invite many journals to seek public funds when an uncompromising OA principle would have scared them off.

Someone might object that I have not been consistent. I have said that copyright holders should be free to exercise their rights up to their limits (Case 2). But here I am recommending that copyright holders waive one of their rights in order to benefit from public funds.

The two positions are entirely consistent. I am not arguing that copyright holders don't have the right to insist on TA, or that they could not exercise the right if they wanted to. I am saying that they might choose to waive that right in exchange for the benefit of public funds. If they don't think it is a good deal, they don't have to take it. The deal doesn't limit their freedom; it merely offers something of value which they might or might not find worth the price of waiving their right to block delayed OA. Publishers themselves should understand this situation very well. It is exactly the kind of deal they offer to authors: give up some set of your rights in exchange for the benefit of publishing in our journal.

To obtain this kind of waiver, the public funder must deal directly with the rights-holder. The case gets more complicated when the rights-holder is not the one desiring digitization or applying for funds. For example, consider the microfiche digitization project of the publicly-funded US Education Resources Information Center (ERIC). ERIC wanted to digitize and provide OA to about 340,000 microfiche documents, some of them up to 40 years old. The documents were written by hundreds of thousands of different authors and might have hundreds of thousands of different rights-holders. Some of the documents might, after diligent inquiry, turn out to be orphans, and some might not. ERIC undertook the enormous job of trying to hunt down each copyright holder. In the end it was able to clear permissions for about 55% or 192,000 of the

documents. The rest may never be OA, despite the willingness of the U.S. Department of Education to spend public funds on their digitization.²²

Even if we adopt OA-friendly rules for orphan works, we must first go to the trouble of trying to locate the copyright holders. Otherwise we will not know whether or not the works are orphans. For more, see Case 10 in the appendix.

Case 4. The funds are provided by a public-private partnership, and all the targeted works are in the public domain.

First consider a much easier related case. If Penguin Books digitizes an early PD edition of *Pride and Prejudice* with its own funds, it should be free to sell it. It need not give it away just because the original was PD. If you agree, then it seems that public funding is a more critical variable than PD status.

The difficult case here is when we pay for the digitization of a PD work with a mix of public and private funds, a common practice. Many public funders are unable to pay for a certain project on their own, or try to stretch their budgets by recruiting private partners. The use of public funds pulls the project toward OA, and the use of private funds pulls the project toward the wishes of the private funders, which may be TA.

Consider the Digitizing American Imprints program, which is using public funds from the Library of Congress and private funds from the Sloan Foundation to digitize 100,000 PD books.²³

Another example is the Medical Journals Backfiles Digitization Project, co-sponsored by the Wellcome Library (private), JISC (public), and the US National Library of Medicine (public). The project includes some copyrighted but orphan works, which it promises to remove if the copyright holder steps forward and asks it to; more in Case 10.²⁴

A third is the World Digital Library, with public partners like 13 national libraries and UNESCO, and private partners like the Brown University Library, Yale University Library, and the Wellcome Trust Library.²⁵

The private partners in these three projects want OA as much as the public partners. That is good for the public and good for the working harmony of the partnership.

But what if the private partners oppose OA and want to sell access to the digital editions without competition from OA editions? In that case, we can use the embargo compromise that we used in the previous case. The private funder could erect a temporary toll gate on access to the digital editions.

If members of the public object that the digital editions are temporarily TA, we answer as we did in the previous case. A private funder made an essential contribution to the project and without its contribution OA would be delayed even further.

If the private funder objects that its period of exclusivity is only temporary, our replies are variations on the theme of our replies in the previous case. First, the public made an essential contribution to the project and must benefit as well. Second, the partnership is voluntary and the private partner did not have to join.

But beyond these, we have two additional replies we could not have used in the previous case. First, the private partner has no rights in these works, which we have stipulated are PD. Second, if the embargo period never expired, then for a fraction of the cost of digitization we would allow a private company to buy permanent exclusive rights to works in the PD (not the PD originals but the PD copies produced by the project).

If the private partner objects that the embargo period is not long enough to recoup its investment, and that it can't afford to take the risk of not recouping its investment, then it need not participate. If it has enough money to do the digitization by itself, without public partners, then it can proceed on its own and follow its own rules, turning this into the related, simpler *Pride and Prejudice* case. If it doesn't, then it should understand the need to allow all the investment partners to get something out of the deal.

In setting the length of the embargo, we must remember that it is a compromise with the public interest. The purpose is to give the private partners something, not everything, just as the public partners are only getting something, not everything. The compromise gives the private partners a chance to recoup their investment, not a guarantee. To give them all the time they need to recoup their investment could require a permanent embargo and eviscerate the very idea of compromise.

Someone might object that under this policy we could lose the contributions of profit-seeking private companies willing to invest in digitization projects. Yes, we could. But as before, it is no calamity to lose the chance to spend public funds on a project which excludes the public, or to lose the chance to spend public funds collaborating with those unwilling to provide even delayed OA.

Nevertheless, if governments wanted to do more to encourage the participation of private partners, without giving up on timely OA for the public, they could combine a fixed deadline on the embargo with a tax deduction for any part of the private partner's investment not recouped during the embargo period.

This could part of a larger plan to use tax deductions to get private companies to open up access to their research and data.²⁶

Public institutions taking on private partners for digitization projects should advertise their needs openly. If they accept a secret, no-bid, or unsolicited offer from a private company, they might end up agreeing to a longer embargo than necessary. Before accepting any private partners or at least any private partners who will resist OA and require a compromise, public agencies should undertake a transparent process of public consultation and competitive bidding. The rationale is simply that it is bad public policy to compromise the public interest more than necessary.

All three of the projects mentioned earlier, the Digitizing American Imprints program, Medical Journals Backfiles Digitization Project, and the World Digital Library, provide OA without any embargo at all. The private partners in all three cases came to the projects with the same purposes as the public partners, making compromise unnecessary. That is worth noting for two reasons. First, it shows that the principle here is that embargoes are permissible, not mandatory. The embargo is a compromise and is only necessary when a compromise is necessary. Second, it reminds us that the private partners in public-private partnerships don't always oppose OA.

(Conversely, public funders don't always support OA, as we have already seen in Case 1, on cost recovery, and will see again in Case 5, on the database right and sweat-of-the-brow doctrine.)

But some private partners do oppose OA. In January 2007 the U.S. National Archives and Records Administration (NARA) announced a partnership with Footnote.com. Under the deal, Footnote would digitize millions of pages of PD documents from the National Archives, including the papers of the Continental Congress and Matthew Brady's Civil War photographs. The deal gave Footnote non-exclusive rights to sell access to the digital editions for five years. During that time, the digital editions could be viewed without charge from terminals in NARA reading rooms in 16 states. After five years, the digital copies would be OA at the NARA web site.²⁷

During the five-year embargo period, Footnote's online access fees are \$1.99 per page or \$100 per year.²⁸

The non-exclusivity of the deal meant that other companies could sell access to their own digital editions, if they could make their own digital editions. But NARA is only willing to deal with Footnote. Moreover, the Footnote deal was not publicly announced until the contract was already signed and Footnote had already digitized 4.5 million PD documents.

There are several problems here. One is the length of the embargo period. Five years is very long. Footnote might argue that the combination of its money and embargo will speed up OA more than slow it down. But that seems unlikely in light of NARA's February 2006 deal with Google. Under that deal, Google funded the digitization of 101 PD films from NARA and provided immediate free online access to all of them.²⁹

In any case, the long-embargo problem is inseparable from the secret, no-bid contract problem. We will never know whether other private partners would have done the work with a shorter embargo period, lower access fees, or both.

In July 2007, NARA made an even worse deal with CustomFlix, a division of Amazon. The deal allowed CustomFlix to digitize films from the National Archives and sell DVD editions through Amazon. Members of the public who visit the NARA facility in College Park, Maryland, could copy the films without charge. In contrast to the Footnote deal, nothing in the CustomFlix contract or press release mentions an embargo period, suggesting an effectively permanent embargo.³⁰

The NARA-CustomFlix contract was secret until Rick Prelinger forced its disclosure with an FOIA request in August 2007. The contract gave Amazon perpetual non-exclusive license to sell the digital editions and gave NARA its own copies of the digital files and the right to use them in any lawful manner. Hence, it allowed NARA to provide OA at any point. But in striking contrast to the Footnote deal, NARA never promised to provide OA, on any timetable.³¹

In May 2008, NARA released a set of principles to guide its future digitization projects. Interestingly, it requires public comments on proposed private partnerships and highlights the importance of minimizing embargo periods. It seems that NARA heard the public criticism of the Footnote and CustomFlix deals and resolved to fix at least some of the problems.³²

Case 5. All the funds are public and all the works PD. So far, this is Case 1. But suppose that the host or funder wants to restrict the use of the digital editions.

Let's say that a work is gratis OA when it is digital, online, and free of charge, even if it is still subject to copyright or licensing restrictions. A work is libre OA when it is gratis OA and also free of at least some copyright and licensing restrictions. Gratis OA allows no uses beyond fair use (or fair dealing etc.), and libre OA allows at least some. Gratis OA removes price barriers and no permission barriers, while libre OA removes price barriers and at least some permission barriers.³³

Using these terms, we can restate Case 5 more succinctly: All the funds are public, all the works are PD, but the funder wants to make the digital editions gratis OA, not libre OA. The Oregon statutes from Case 1 fall under this description, and the issues raised by Oregon-type cases deserve a closer look.

Legally, the least complicated way for a digitizer to restrict use of a digital work is to keep it offline. Fair use and the public domain give us the right to use certain works in certain ways, but they don't give us the right to enter buildings where copies may be under lock and key.

But if the digitizer puts the digital edition online and still wishes to restrict usage, then the requested restrictions might have any of these four grounds:

1. *Copyright*. The work might be under copyright; but if so, we have dealt with the major issues in Case 3.
2. *Sui generis or database right*. The European Database Directive creates a kind of legal protection, outside copyright law, for databases that require substantial investment but lack the originality required for copyright. Ordinary digitization lacks the originality required for copyright, and individual examples lack the database element required for the sui generis right. But when many digitized files are brought together into a database, and when the effort requires substantial investment, then the collection may qualify for the sui generis or database right. If so, however, then Case 3 covers the major issues. If we can put an OA condition on public funds for holders of strong copyright, then we can do the same for holders the weaker sui generis right.
3. *Unenforceable request*. The online host might acknowledge that it has no legally enforceable right to restrict usage. But it might make an admittedly unenforceable request, appealing to courtesy or respect rather than law. For example, in the downloaded copies (but not the online copies) of Google-scanned PD books, Google asks users to retain attribution and avoid commercial use and automatic querying.³⁴
4. *Copyfraud*. The host might falsely claim copyright and attempt to ground its requested restrictions in copyright law.³⁵

Consider The European Library (TEL). This is an online collection of exhibits digitized from the national libraries of Europe. TEL didn't do the digitizing or set the copyright and licensing terms for the individual exhibits. It coordinates the separate efforts of the separate contributing libraries. In most cases, it doesn't even host the exhibits but links to digital editions hosted by the separate libraries.³⁶

It appears that all of the works on display through TEL were digitized with public funds, and that some of the digital editions are under copyright, some under the sui generis right, and some fully PD.

TEL provides no item-level rights or licensing information. See for example the image-scan of a handwritten letter from Napoleon I to Joachim Murat, King of Naples, from October 7, 1813,³⁷ or the image-scan of the Heiberg translation of the Marseillaise into Danish, published in Copenhagen in 1793.³⁸

TEL does provide item-level metadata, even if they don't include rights or licensing information. But the deep links to individual exhibits (which I used above) don't include the metadata. To find the metadata for the Napoleon letter or Heiberg translation, you have to locate the exhibits within this larger exhibition, click on them, and read the metadata off an unlinkable pop-up window.³⁹

But since that method doesn't tell us about rights or licensing, we can only learn the status of the Napoleon letter or Heiberg translation by consulting the TEL "terms of service", which tell us that:⁴⁰

The Conference of European National Librarians and its licensors hold the copyright for all material and all content in this site, including site layout, design, images, programs, text and other information (collectively, the "Content") held in The European Library. No material may be resold or published elsewhere without the Conference of European National Librarians written consent, unless authorised by a licence with the Conference of European National Librarians or to the extent required by the applicable law.

Even on the most charitable reading, this statement is false for many or most exhibits in the TEL. For the PD exhibits, it is entirely false. For the exhibits under the sui generis right, it falsely states the rights are based on copyright instead. (This matters, among other reasons, because copyright lasts more than five times longer than the sui generis right.) The attempted restriction on the sale and publication of the exhibits is groundless for the PD content, even if lawful for the other two categories. But TEL says that all the contents are under copyright, and none merely under the sui generis right and none in the PD. If it is true for some exhibits, it is copyfraud for others.

TEL might have intended the copyright statement to apply to the web site's apparatus, not to the exhibits themselves. But nothing in the statement suggests that distinction, and the clear language of the statement

(“all content in this site...”) suggests the opposite. Moreover, the absence of item-level rights and licensing information on individual exhibits forces us to turn to the general terms of service for that information. The statement might apply only to TEL-hosted content, rather than to content at the separate national libraries to which TEL merely links. But even the TEL-hosted content seems to fall into all three categories, not just the category of copyright, and in any case TEL points to the same terms of service for TEL-hosted exhibits and for library-hosted exhibits.

TEL might have intended the statement to be part of a clickwrap license, under which visitors agree to waive their rights to use and reuse any of the contents which happen to be PD. But the site does not ask users to click their assent to any licensing terms before viewing exhibits, and the terms of service claim to base the reuse restrictions on copyright, not contract. In any case, even if TEL used a clickwrap license to create a contract with the user, and even if the contract was enforceable, users who redistributed files that are actually PD would be making them available to people who were not bound by the contract.

This copyfraud creates several problems. First, for the PD content, the claimed restrictions are unenforceable. Anyone selling or publishing the digital edition of a PD work would be exercising protected rights under copyright law. Second, for content under the *sui generis* right, the copyright claim implies rights for the full term of copyright rather than the much shorter (15 year) term of the *sui generis* right.

Third and most important, the false claim of copyright might deceive or intimidate some users into giving up rights they are entitled to exercise. It inhibits the lawful and legitimate use of this valuable historical content.

Even the onerous NARA-CustomFlix contract acknowledged that “Content obtained by researchers through public access [via a NARA reading room] is in the public domain” and its uses could not be restricted.⁴¹

TEL should drop the false claim of copyright. It should acknowledge that much of its content is PD, and that users may use and reuse the PD content without restriction. If any of the exhibits are under the weaker *sui generis* right, rather than copyright, it should acknowledge that as well.

I don’t want to underestimate the difficulty of adding item-level rights information to each exhibit in a large collection. It can be one of the larger costs in a large digitization project. But if TEL can’t add accurate item-level rights information, it should at least stop using inaccurate site-level information in its place.⁴²

The TEL can’t do much more than that, since it didn’t digitize the

works in the collection. But the national libraries of Europe which participate in TEL can do more. They are using public funds to digitize PD works. Even if the EC Database Directive allows them to claim a *sui generis* right in the digital editions, they need not take advantage of the option. On the contrary, there are good policy reasons why they should not. It is hard to imagine how their purpose in trying to restrict usage could outweigh their mission to serve the public, promote access to the historical materials in their collections, and foster research, scholarship, art, education, and cultural development. (We know that the libraries' purpose is not cost-recovery, since they are already consenting to gratis OA.)

Finally, they should understand that libre OA facilitates preservation, among other forms of use and reuse. Long-term preservation requires making copies and migrating them to new media and formats to keep them readable as technology changes. Copyright and the *sui generis* right both raise the barrier to those useful copies, either by blocking them altogether or by requiring the expense or delay of seeking permission.⁴³

Until recently, Cornell University took a position roughly similar to TEL's for the PD books digitized from its library. It posted the works online, without a clickthrough license, but required users to seek permission for any commercial use. In May 2009, however, it reversed course. It acknowledged that the books are PD, stopped trying to restrict usage, and explained why in an exemplary public statement. In the statement, Cornell said it did not wish to "limit the good uses" of these works. On the contrary, it "decided it was more important to encourage the use of the public domain materials in our holdings than to impose roadblocks." Moreover, Cornell recognized that claiming the right to restrict usage was copyfraud, and that the criticism of copyfraud was justified.⁴⁴

Cornell would have been within its rights to put the digital editions behind a password, require users to assent to a clickthrough license, and then charge for access or impose usage restrictions. Likewise, it could have put the works online without a clickthrough license and made an admittedly unenforceable request to restrict usage. But in May it chose not to do either of these things, and not to rest on copyfraud either. The Cornell solution is especially commendable because Cornell is a private university. Either it used its own, private funds for the digitization or it used Google's. (Cornell has been a partner in the Google Library Project since August 2007.)

The U.S. doesn't recognize the *sui generis* database right and Cornell could not have relied on it. But even institutions in countries which do recognize the right can use the Cornell solution. They simply have to

decline to use the right available to them, and (in Cornell's words) decide to put "good uses" ahead of "roadblocks".

Cornell is a private university, but its solution is compelling even for public institutions. Indeed, if a private institution can drop copyfraud and support the full use and reuse of PD works, then public institutions using public funds should be able to do so as well.

Appendix

Here is a quick summary of the five cases I have discussed:

Case 1. All the funds are public, and all the works to be digitized are PD.

Case 2. All the funds are private, and all the works to be digitized are under copyright.

Case 3. All the funds are public, and all the works to be digitized are under copyright.

Case 4. The funds are provided by a public-private partnership, and all the works to be digitized are PD.

Case 5. All the funds are public, all the works PD, but the funder only wants to allow gratis OA, not libre OA.

Here are five more hard cases that will have to wait for another day:

Case 6. All the funds are private and all the works to be digitized are PD. So far this is the easy *Pride and Prejudice* case. But now add that the targeted works are rare, unique, or fragile.

The *Pride and Prejudice* case is easy in part because it's easy to get a copy of the print book for digitizing. If one digitization project offers the digital edition on onerous terms, then others can digitize the same book and offer their editions on more liberal terms. But the realistic odds of re-digitization plummet when the original is rare, unique, or fragile.

Consider the *Codex Leicester*, a volume of Leonardo da Vinci's handwritten journal which Bill Gates bought from Armand Hammer in 1994 for \$30.8 million. It is the only original da Vinci now in private hands.⁴⁵

Gates has been generous with its display: the original is on loan to a different museum every year; high-res photos of every page have been published in a book (a priced, printed book, not an OA book); and OA thumbnails are available online at Corbis. But as far as I know, he has not allowed OA to high-res images.⁴⁶

Is there a strong policy argument for asking a private individual like Gates to provide OA to this kind of unique PD work? If not, does the

argument become stronger if the owner is a private university like Cornell?

What if the digitization of the Dead Sea Scrolls is funded by private donors?⁴⁷

Case 7. All the funds are private and all the targeted works are PD. So far this is either the easy *Pride and Prejudice* case or the hard *Codex Leicester* case. Now add that all the targeted works will be provided to the project by a public institution, which acquired and curated them with public funds for public benefit.

A typical example is Google's project to digitize PD books from public university libraries, such as the University of Michigan library. Should Michigan put an OA condition on its collaboration with Google?

Case 8. All the funds are private, all the targeted works are PD, and all the works will be provided by an institution which has acquired and curated them at some expense. So far this is Google-Michigan case. But instead of a public institution using public funds, let it be a private institution acting for non-commercial purposes and with public subsidies through untaxed property and tax deductible contributions.

A typical example is Google's project to digitize the PD books from private university libraries, such as the Harvard and Cornell libraries. Should Harvard and Cornell put OA conditions on their collaboration with Google?

Similar issues arise when a PD digitization project is funded by private philanthropy, such as the Mellon Foundation, with no public partner.

Do the policy arguments for OA that apply to public funders also apply to all institutions with non-commercial purposes and tax breaks, even if private?

Case 9. The funds are from a public-private partnership, and the works to be digitized are PD. So far, this is Case 4. But instead of mere digitization, the project extends to editorial work and copyrighted commentary. The plan is to integrate the PD texts and the copyrighted commentary. The private partners and copyright holders want to publish the results in print books or TA web sites and oppose any attempt to make them OA, even after an embargo period.

See the NARA plan for a digital edition of papers of US Founding Fathers:

- The case for OA.⁴⁸

- The case for TA (though misleadingly called “open access”).⁴⁹
- The case for a middle position (including an embargo period followed by libre OA).⁵⁰

Case 10. Take any of the variations above in which the works to be digitized are still under copyright (for example, Cases 2 and 3). Now add the variable that they are orphan works.

Should the digitizer follow the Wellcome Library and make the digital editions OA, promising to take them down if the copyright holder steps forward and objects?⁵¹

Should it follow the Google book settlement and sell access?⁵²

For a middle position, see Peter Eckersley’s argument that all Google-digitized books, and especially the orphan works, whether based on originals from public or private institutions, should become OA after an embargo period.⁵³

If we diligently look for the copyright holders, fail to find them, and responsibly conclude that we are dealing with orphan works, then should we assume the lack of permission for OA until we have explicit consent from the copyright holders or national legislature? Or should we assume permission for OA until we have explicit dissent? Even after responsibly concluding that we are dealing with orphan works, should we adopt a compromise like an embargo period?

Or should we start to rethink the very idea of *permission* in cases like this? Normally, medical care without consent is battery, just as full-text copying of copyrighted texts without permission is infringement. But when an unconscious person is wheeled into an emergency room, and we are unable to get an explicit “yes” or “no”, then we start to talk about “implied consent” to receive care and “privilege” to render care. When diligent effort fails to turn up a copyright holder, and we are equally unable to get an explicit “yes” or “no”, then should we also start talking about implied consent and privilege? The stakes are not the same, but the consent quandary is the same. Do we only want to solve the consent quandary in matters of life and death, or might we also want solve it in matters of scholarship, research, art, culture, and education?

An earlier version of this article was published in the *SPARC Open Access Newsletter*, July 2, 2009.⁵⁴ For helpful feedback on that edition I thank Frode Bakken, Klaus Graf, Sebastian Krujatz, and Sanford Thatcher.

References

- 1 <http://www.eprints.org/openaccess/policysignup/>
- 2 <http://www.earlham.edu/~peters/fos/2009/04/oa-to-historic-ontario-legislation.html>
- 3 <http://www.earlham.edu/~peters/fos/2008/01/free-online-access-to-12-years-of.html>
- 4 <http://www.earlham.edu/~peters/fos/2009/04/oa-to-historic-ontario-legislation.html>
- 5 <http://public.resource.org/oregon.gov/>
- 6 http://www.dailytimes.com.pk/default.asp?page=2006\12\09\story_9-12-2006_pg11_4
<http://www.earlham.edu/~peters/fos/2006/12/free-access-to-pakistani-law-for.html>
- 7 <http://www.bl.uk/aboutus/stratpolprog/digi/digitisation/digistrategy/index.html>
<http://www.earlham.edu/~peters/fos/2008/08/british-library-digitization-plans-some.html>
- 8 http://www.oxfordjournals.org/access_purchase/archives.html
- 9 <http://www.earlham.edu/~peters/fos/2008/01/free-online-access-to-12-years-of.html>
- 10 <http://www.earlham.edu/~peters/fos/2007/11/dutch-medical-journal-opens-backfile.html>
- 11 Paul Ayris, "LIBER-EBLIDA Digitisation Workshop," *Liber Quarterly*, April 2008, p. 17.
<http://liber.library.uu.nl/publish/articles/000222/article.pdf>
- 12 <http://ur1.ca/5yms>
- 13 <http://www.cfes.com/EI-ASPM/>
- 14 <http://www.cfes.com/documents/wp23.pdf>
<http://www.earlham.edu/~peters/fos/2006/08/oa-increases-return-on-investment-in.html>
- 15 <http://www.freeourdata.org.uk/blog/?p=86>
- 16 <http://dx.doi.org/10.1136/bmj.b1744>
<http://www.earlham.edu/~peters/fos/2009/05/169-years-of-bmj-now-oa.html>
- 17 http://www.oxfordjournals.org/access_purchase/archives.html
- 18 I am not the first to suggest this compromise. See the High Level Expert Group on Digital Libraries, Sub-group on Public Private Partnerships, *Final Report on Public Private Partnerships for the Digitisation and Online Accessibility of Europe's Cultural Heritage*, 12010 European Digital Libraries Initiative, May 2008.

http://ec.europa.eu/information_society/activities/digital_libraries/doc/hleg/reports/ppp/ppp_final.pdf

In Recommendation #9 at p. 22, “The sub-group recommends that public domain content in the analogue world should remain in the public domain in the digital environment. If restrictions to user’s access and use are necessary in order to make the digital content available at all, these restrictions should only apply for a time-limited period.”

- 19 <http://www.earlham.edu/~peters/fos/newsletter/08-02-07.htm#embargo>
- 20 <http://www.earlham.edu/~peters/fos/newsletter/10-02-08.htm#nih>
- 21 <http://www.neh.gov/grants/guidelines/editions.html>
- 22 http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&_pageLabel=Digitization
<http://www.earlham.edu/~peters/fos/2009/05/eric-microfiche-digitization-project.html>
- 23 <http://www.loc.gov/today/pr/2009/09-10.html>
<http://www.earlham.edu/~peters/fos/2009/01/loc-digitization-program-crosses-25000.html>
<http://www.earlham.edu/~peters/fos/2009/03/oa-book-scanning-program-from-library.html>
- 24 http://library.wellcome.ac.uk/doc_WTD037630.html
http://library.wellcome.ac.uk/doc_wtx043243.html
http://library.wellcome.ac.uk/doc_wtdv025956.html
- 25 <http://www.worlddigitallibrary.org/project/english/index.html>
<http://project.wdl.org/project/english/partners/>
- 26 <http://weblog.infoworld.com/udell/2006/06/12.html>
<http://www.earlham.edu/~peters/fos/2006/06/tax-breaks-to-support-oa.html>
<http://www.earlham.edu/~peters/fos/newsletter/07-02-06.htm#nih>
- 27 <http://archives.gov/press/press-releases/2007/nr07-41.html>
- 28 http://www.dancohen.org/blog/posts/national_archives_footnote_agreement
<http://www.earlham.edu/~peters/fos/2007/01/digitized-public-domain-docs-from-us.html>
- 29 <http://www.archives.gov/press/press-releases/2006/nr06-64.html>
<http://www.archives.gov/digitization/google-agreement.pdf>
- 30 <http://www.archives.gov/press/press-releases/2007/nr07-122.html>
- 31 <http://blackoystercatcher.blogspot.com/2007/08/national-archivesamazon-agreement.html>
http://www.panix.com/~footage/NARA_Amazon.pdf

- 32 <http://www.archives.gov/digitization/strategy.html>
- 33 <http://www.earlham.edu/~peters/fos/newsletter/08-02-08.htm#gratis-libre>
- 34 Here's an example, though the link may only work from IP addresses in the United States:
http://books.google.com.ph/books/download/Pride_and_Prejudice.pdf?id=2_S8xAws2G4C&hl=en&output=pdf&sig=ACfU3U36yL_-URJJlcMilNPjhX7LOkUgfA
- 35 http://papers.ssrn.com/sol3/papers.cfm?abstract_id=787244
- 36 <http://www.europeanlibrary.org/>
- 37 http://www.theeuropeanlibrary.org/exhibition/Napoleonic_wars/images/texts/_large/ennw22_l.jpg
- 38 <http://img.kb.dk/ma/nybrev/12-05/spalvor-mars.pdf>
- 39 http://www.theeuropeanlibrary.org/exhibition/Napoleonic_wars/
- 40 http://www.theeuropeanlibrary.org/portal/organisation/footer/termsofservice_en.html
- 41 <http://blackoystercatcher.blogspot.com/2007/08/national-archivesamazon-agreement.html>
http://www.panix.com/~footage/NARA_Amazon.pdf
- 42 Disclosure: I'm on the advisory board for TEL and have made my objections known. I'm still hoping to resolve the problem.
- 43 <http://www.ippr.org.uk/publicationsandreports/publication.asp?id=464>
<http://www.clir.org/pubs/reports/pub135/sec6.html>
<http://www.pocket-lint.com/news/news.phtml?newsId=3566>
- 44 <http://ur1.ca/6lew>
<http://www.earlham.edu/~peters/fos/2009/05/cornell-allows-unrestricted-use-of-its.html>
- 45 http://en.wikipedia.org/wiki/Codex_Leicester
- 46 http://www.amazon.com/Leonardo-Da-Vinci-Leicester-Notebook-Genius/dp/1863170812/ref=sr_1_2
[http://pro.corbis.com/Search/SearchResults.aspx?q=%20" Codex%20Leicester%20by%20Leonardo%20da%20Vinci"](http://pro.corbis.com/Search/SearchResults.aspx?q=%20)
- 47 <http://www.earlham.edu/~peters/fos/2007/11/digitizing-dead-sea-scrolls.html>
http://en.wikipedia.org/wiki/Dead_Sea_Scrolls#Digital_copies
- 48 <http://blog.librarylaw.com/librarylaw/2008/06/free-the-foundi.html>
- 49 <http://www.archives.gov/nhprc/publications/founders-report.pdf>
- 50 <http://www.earlham.edu/~peters/fos/newsletter/08-02-09.htm#update>

- 51 <http://wellcomelibrary.blogspot.com/2009/06/orphan-works.html>
- 52 <http://books.google.com/booksrightsholders/>
- 53 <http://www.eff.org/deeplinks/2009/06/should-google-have-s>
<http://www.earlham.edu/~peters/fos/2009/06/another-idea-for-building-oa-into.html>
- 54 <http://www.earlham.edu/~peters/fos/newsletter/07-02-09.htm#digitization>

Applications 1: Adding value for everyone

DIGITAL READING AND INDUSTRIAL READINGS

Alain Giffard

Although the symposium *Going Digital* mostly concerns the dilemmas of digitisation in the field of humanities, scholarship and research, one must not forget that digital reading is a public question, a question of the public. People, parents and teachers ask: will the screen (the computer, the web) replace the book? Is digital reading true reading? Will it succeed “classical” reading? Will children continue to read, as their parents and their school try to teach them? And if they read differently, how will literacy and literature be transmitted?

These questions are arising in a context where many countries face a dramatic decrease in classical reading. Not only are the numbers of books and papers being read declining, but performance in reading itself is diminishing, as the PISA surveys show, for example. Digitisation is neither the origin nor the cause of this decrease, but the way the first trend coincides with the second one is a major issue in education, literacy and culture.

My research is generally about digital culture and the digitisation of culture. More precisely is it about reading, the future of reading and the potential for reading. The background of this research is both practical and theoretical.

As the Computer and Information Technologies Director of the “TGB,” or Very Large Library, almost twenty years ago, I had to produce the design for the information system, and also of the digital library, of the Bibliothèque nationale de France, which is known today as Gallica. In this programme, I prepared, with the philosopher Bernard Stiegler, a software system called the “Computer-assisted reading environment”. When the French prime minister asked me to organise something to help reduce the digital gap, I also confronted the question of digital literacy.

More recently, in 2007, I wrote a study for the Ministry of Culture about reading as a digital cultural practice, which has been published under the title “Des lectures industrielles” (‘About industrial readings’).

This article examines two points. In the first part, I will try to propose a diagnosis of “digital reading.” For this approach, according to the lesson of the historians, reading is seen as a practice, not a simple consequence of digital text, digital media or computerised devices for reading.

The second part of the article focuses on the notion of “industrial readings” that I put forward. Industrial readings not only indicate the economic environment of digital reading. They are also the main explanation as to why the technology of digital reading and the associated practice are what they are and not something else.

Digital reading

So let us begin with this very simple and elementary question: is digital reading at least a certain form of reading? I think the answer is yes. Digital reading is reading.

Reading on a screen and digital reading

There is an important qualitative difference between reading before the web and reading on the web. So I suggest we speak of “reading on a screen” for the first one, and “digital reading” for the second one.

Before the web existed, in the practice of reading on a screen, the text is not the objective of the reader. Rather is it control reading, a way of deciphering and monitoring the information and operations of a computer. And reading is subordinate to another activity that is the real goal. Credit cards, word processors and phototypesetting are examples of such readings on a screen. Umberto Eco said “word processor e una machina molto spirituale” but the reading functionalities of a word processor are not spiritual at all.

Reading on a screen has become a reality, a social practice with the invention of the personal computer, which means that the computer has become a medium. It has become easier, more “user-friendly” according to the marketing jargon, but has stayed more difficult than reading on paper, which remained its reference, as the “Wysiwyg” formula, “What You See Is What You Get” shows. Basically, text was lacking in this first form of reading with a computer: a certain quantity, and if possible a certain quality of texts to create a true reading environment.

The web has done it. HyperTextMark-upLanguage, the language of the web, has made digital texts exist on the basis of the Internet. The web produces the conditions of a textual environment on the screen consistent enough to arouse a regular and massive practice of reading. This point seems evident.

Less evident is the fact that digital reading is not only a consequence of digital texts, but also a condition for the development of the web, and the basis of the emergence of a reading technology. There are a lot of examples, as the uses of hyperlinks by research engines. I also like to quote the origin of the word “blog” from “web-log,” reading log on the web, to “we-blog.”

The technology of digital reading is a technology by default

The traditional technology of reading is human. It may be a method for the reader, as in the Didascalion of Hugh of Saint Victor (12th century),

or it may be applied to the text itself or to the medium, as it is the case, for example, with the invention of word separation in the British Isles in the 7th century, studied by Paul Saenger precisely as a new way of reading.

In comparison, the technology of digital reading is paradoxical. If you search for the word “reader” on Google, you will find a lot of machines, software and languages. It seems more difficult to find a reader who is a man or woman. And in fact, looking at browsers, readers, e-books and search engines themselves – the Google robot is a reading machine – there are lots of techniques of reading.

But, on the other hand, there is no global technology for digital reading. The paradox is everywhere. We can have a text in more and more visible form, approaching a printed page; or we can have a medium with all the potential of digital text. But we cannot have the two together in a simple way, I mean during the act of reading. We can have access to a great number of texts on the web; we can also annotate and create personal paths of reading. But we cannot practice this sort of active reading on the whole web.

Only on a mere functional level, the technology of digital reading lacks unification, fullness and integration. The act of digital reading is complicated and difficult. Difficulties, which notably have been emphasised by cognitive psychologists, are everywhere: from the visibility of the screen, typography and layout to the absence of unity, which prevents the reader from projecting his model of understanding. The reader has a certain idea of the text. He must replace it during each manipulation, but the launching and the execution of this new operation tend to make him forget the first version, the first idea of the text. The path of reading, “le fil de lecture”, is cut. The heaviness and difficulty of manipulation create a supplementary problem of attention. Cognitive overflow syndrome does not only come from the overloading of information in the text or about the text; it is fundamentally an operative overflow.

What I am trying to summarise with this idea of a technology by default is simply the fact that digital reading technology is mostly on the “reading machine” side and not on the “reading man” side.

Digital reading and reflection

The most important point in the balance of digital reading is the type of effective reading it allows. And there is no way here to avoid a comparison with classical reading and reference to the history and philosophy of reading. The general meaning of reading practices is the key for their evaluation and critic as cultural practices.

I submit my hypothesis with caution. It seems that there is a risk, and sometimes more than a risk, of a convergence between the achievement of the act of reading, the sort of attention of the reader, and the content of reading.

About the practice of browsing, we may speak of pre-reading, a notion that comes from Romans as *praelectio*. There are many reasons that make the preparation of reading necessary: think about the difficulty of reading a papyrus roll, or an un-separated word text. Browsing is an activity that produces the text to be read, and is a sort of digital pre-reading. More generally, the act of digital reading seems in an unachieved state, as if different manipulations were to prepare a reading that doesn't come.

Many commentators of digital practices focus on the question of attention. Katherine Hayles has contrasted hyper-attention, attention that needs to be frequently activated, with the deep attention that she sees as characteristic of classical reading. Even if we do not agree with the idea of different generational cognitive styles, it seems quite reasonable to acknowledge that the digital reading environment does not favour deep attention and that it multiplies the opportunities to lose one's concentration. Yet the difference is not between continuity and discontinuity, but between different types of continuity, and different types of discontinuity.

An American researcher, Ziming Liu, has tried to explain the frequent junctions from screen to paper. He has underlined the difficulty, with digital reading, of moving from a scanning towards sustained reading, that is the difficulty of passing from informational reading to studying reading.

Now "studying reading" has been constructed, in Western Culture, on the foundation of silent reading, and based on the link imagined by Augustine, and systematically established by Hugh of Saint Victor, between reading and reflexion, *lectio* and *meditatio*. Studying, which implies deep attention, complete and sustained reading, is what leads to reflexion either on the text, or from the text, or from the subjective situation of the reader. Such reading, which I call "studying reading", is not more or less intensive, more or less active, more or less serious than informational reading. It has another finality. And here we must mention Michel Foucault's concept of reading as "the technology of the self" and Brian Stock's concept of reading as an intellectual or spiritual exercise, *askêsis*.

It seems that the convergence between hyper-attention rather than deep attention, pre-reading rather than complete reading and informational reading rather than studying reading may create a situation where

the link between reading and reflection, *lectio* and *meditatio* is no longer established.

Simulation of reading is not literacy

My last point is about simulation. We have seen that the technology of digital reading was mostly on the machine side. But this machine is based on the functionalities of human activity. For example, if someone uses an automatic translator, he simulates the activity of a professional who is translating texts, by using software. But simulation is not knowledge transfer and simulation of reading is not literacy. If one uses a search engine, he needs to qualify the results by using three different skills: knowledge about the subject in general, knowledge about the subject as it is digitised and knowledge about the search engine. Of course his expertise is not total – which is why he submits a search request. But if his expertise is not enough, the simulation is reversed – it becomes a simulacrum – a pretence of reading. In the end, the whole activity of reading may be simulated this way, as writing may be simulated by copy-and-paste.

Industrial readings

What is reading becoming? An *industry*. The notion of industrial readings is of course economic. But it is also political, as part of the question of the new public space that Robert Darnton has called the “digital republic;” and it is also cultural, particularly about knowing-how-to-read: literacy. It has appeared to me that this notion was an answer to two different questions: how can we improve the technology? And how can the reading public cope with this situation?

Reading industries

Reading industries are a true novelty. History knows reading technologies, mostly human technologies, and if they were outside the man, they were inscribed in the medium or in the text itself. History also knows literary industries, the name Tocqueville gave to the book publishers in “Democracy in America.” And Adorno, in his essay on the Culture Industry describes Beethoven throwing away a book by Walter Scott and saying, “This bloke writes for money.” But until now, reading industries were unknown.

One characteristic of the digital era is what Jeremy Rifkin has called the “access economy,” and the prominent place of access industries rather than content – a word I don’t like too much – industries. From this point

of view, if publishing industries are content industries, reading industries are access industries.

Reading industries, of which Google is the most remarkable and brilliant example, appear at the meeting point of the information industry, the cultural industry and the marketing industry.

They have three activities or three sectors of activities. The first one is computerisation of the means of reading, software, hardware, and also digitising texts on the model of an access service. For example, Google's robot is a reading machine: crawling is automatic reading and indexation is a traditional activity of reading. Another activity is the production of reading acts and reading texts: that is exactly what Google does when replying to a search request. A reading text is a meta-text, which has no significance if separated from a first one. It was the case for glosses that became a separated text after the production of a glossary. A Google search result is such a meta-text of reading produced by the robot. The third activity – in fact the basis of the business model of most of the reading industries – is the trading of readings and *trading of readers* for marketing purposes.

Reading industries transform the relationship between reader and text into what Edward Bernays, the inventor of marketing and Sigmund Freud's nephew, has called public relations. The word "publicity" here has the two meanings: principle of publication of the text, and language of the economy through marketing. Reading is decentralised in a public space, but a commercial public space, and readers become consumers.

Statistics and calculation are an obsession of the actors on the web. Each reading act is considered as a "hit." The association of statistical information to the recording of reading operations is at the heart of reading industries. I call its products: industrial readings.

Reading industries perform polarisation and separation between communication and reflection. Their technology is adapted only to the first purpose and is far away from reading considered as a technology of the self.

Digital readers as a public

The space of industrial readings deeply differs from the previous combination of public and private space, as mentioned for example by Kant around the principle of publicity, *Öffentlichkeit*. To be effective, Kant's space, or Habermas' public space, requires a public of readers; it requires school, transmission of literacy and, in modern states, it has been one of the major accomplishments of public authorities. But almost everywhere

these authorities have disengaged themselves from their responsibilities for digital literacy and digital reading knowledge.

So the industrial readings space looks like a face-to-face encounter between reading industries and the reading public. And this public of digital readers assumes a lot of big and, I should say, unusual responsibilities.

- Responsibility for the technology: making this technology by default become effective and achieve a sort of provisional consistency.
- Responsibility for the text: la *clôture du texte*, the closure of text. The reader closes the text after browsing, assuming the traditional work of author and publisher.
- Responsibility for one's reading: how to shift to studying reading, and put simply, when to decide to read off the screen.
- Responsibility about one's own training and the formation of a public, what is known as social networks.

Time to debate

So, to conclude, one of the main problems of digital reading and industrial readings, which is also a general problem of the so-called information society or knowledge society, is: are the different readers prepared for such responsibilities?

Readers who possess classical literacy have no problem individually if they agree to use a computer. They do not confuse informational reading with studying reading; they know how to associate reading and reflexion; if necessary, they turn to paper; they don't have the "Google-and-copy-and-paste syndrome."

The situation is very different with young people who are referred to as "digital natives". In 2010 the competition between book and other media among the youth is not a new thing. Almost everybody has known something about this competition in one's life. But even to be a simple consumer, one has to know how to read, and, until now, classical reading – reading of books and papers – was still the reference. Today, media that compete with books and writing – the generational media – reflect another technology, another practice, another space of reading. And this happens at an age in individual lives when the link between reading and reflexion has not yet been constructed.

I am not pessimistic. I do not believe in cultural apocalypse. But the scenario of a gap between the two types of reading – coming back to the title of this symposium: the evolutionary type and the revolutionary type – is the most plausible.

The evolutionary type: a unified way of reading books and digital texts and media; a new technology of digital reading; the preservation of the link between reading and reflexion. The revolutionary type: a contrast between digital reading and classical reading; no link between reading and reflexion; and all sorts of things we will see in young people who are now five to ten years old, before our societies shake themselves out of their torpor.

It is time to debate.

References

- Robert Darnton, "Google & the Future of Books," *The New York Review of Books* (12 February 2009).
- Umberto Eco, Preface to Claudio Pozzoli, *Come scrivere una tesi di laurea con il personal computer*, (Rizzoli, 1986).
- Michel Foucault, *L'Herméneutique du sujet*, Hautes Études (Gallimard/Seuil, 2001).
- Alain Giffard, "Des lectures industrielles" in Bernard Stiegler, Alain Giffard & Christian Fauré, *Pour en finir avec la décroissance* (Flammarion, 2009).
- Alain Giffard, "La lecture numérique à la Bibliothèque de France," in *L'Édition du manuscrit*, Aurèle Crasson ed. (Academia Bruylant, 2008).
- Katherine Hayles, "Hyper and deep-attention: The generational divide in cognitive modes," www.nlajournal.org (2007).
- Ziming Liu, "Reading behavior in the digital environment," *Journal of Documentation*, vol. LXI, no. 6, (2006).
- David S. Miall & Teresa Dobson, "Reading hypertext and the experience of literature," *Journal of Digital Information*, 2-1, (2001).
- Jeremy Rifkin, *The Age of Access* (Putnam, 2000).
- Bernard Stiegler, "Machines à écrire et matières à penser," *Genesis, Revue internationale de critique génétique*, no. 5, (1994).
- Brian Stock, *Augustine the Reader: Meditation, Self-Knowledge, and the Ethics of Interpretation* (Harvard University Press, 1996).
- Jacques Virbel, "Reading and managing texts on the BnF station," in *The Digital Word*, P. Delany, G. Landow eds. (The MIT Press, 1993).

OPPORTUNITY AND CHALLENGE: SUSTAINING PLATFORMS FOR KNOWLEDGE DISSEMINATION IN THE DIGITAL AGE

Kevin Guthrie

Introduction

There is no longer any doubt that the Internet is profoundly and fundamentally changing the nature of knowledge production and dissemination in higher education. What is less clear is how to characterize that change. On the one hand, there have clearly been dramatic and exciting advances in access to information and in the ways digital content is increasingly integrated into teaching and research. For some disciplines, especially in the sciences, the impact is beginning to have a transformative effect. At the opposite end of the spectrum, others see the scholarly enterprise as increasingly threatened by the explosive growth of the web. These observers fear that the traditional strengths and values of the research university are being undermined. But the changes wrought by the Internet are not just about the habits and preferences of scholars – accompanying them are real threats to the institutions and practices that for centuries have formed the bedrock of scholarly publishing and dissemination. A new economy for knowledge production in the digital age is beginning to emerge, but the precise shape of that economy is not yet known.

This paper will argue that the current “messy stage” of the evolving new economy, as new players jockey for position and compete for resources alongside traditional institutions, provides an opportunity for the scholarly community to ensure that its values and priorities are protected, while maximizing the potential of the Internet to transform scholarship in powerful new ways. While the focus of this paper is on higher education, the implications of the transformation that is currently underway have much broader ramifications. The universities of the world are where knowledge is created and where future generations grow the capacity to turn that knowledge into creative forces for change. In one way or another, the health and vitality of universities, and the knowledge economy that they feed, has a ripple effect that touches every corner of society.

The Evolving Information Environment

The Shift Away from Paper

For centuries, the currency of the realm in the world of scholarly publishing and academic discourse was paper. Research was conducted the old-fashioned way: observations

recorded with pen and paper, field notes logged in notebooks, notes scribbled in margins and on index cards, manuscripts hand-written, typed, edited, re-typed, typeset, printed, shipped in cardboard boxes, ending finally on a shelf in a library or faculty office. Meanwhile, armies of catalogers and indexers meticulously and laboriously indexed and organized everything, first on card catalogs, then in MARC records, and finally on bar-codes downloaded with the click of a mouse and safely glued on the back cover. In the library itself, miles of shelves silently safeguarded the output of scholarship, from one generation to the next. Even as computers and electronic databases emerged in the 1970s and 1980s, paper's dominance was not threatened. Computers simply made it more efficient to find and use the paper scholars and students needed.

All of this began to change in the mid-1990s as the World Wide Web grew in importance. At first, the academic sector was the cutting edge of new advances in digital technology. Innovations such as email, hypertext, web-browsers, and search engines all originated at universities. Scholars and librarians were among the first to realize the advantages of the web for accessing and searching information. It was far more efficient and convenient to access the library catalog and journal databases online than it was to browse the library stacks, where materials were often missing or checked out. Many faculty quickly saw the advantages of posting syllabi and other course materials online. Others experimented with ways to link materials in new and creative ways and began to imagine completely new types of scholarly resources. An early example in the mid-1990s was the "Valley of the Shadow" project on the Civil War, followed by other innovative work at the Institute for Advanced Technology in the Humanities at the University of Virginia. At the University of Michigan and Cornell University, the "Making of America" project, launched in 1995, was another showcase of the potential of this powerful new medium.¹ Before long, current journal issues became available online, dramatically improving ease of access and discoverability.

In hindsight, this seemingly revolutionary first wave of changes was not so revolutionary after all. Although digital technologies have increased productivity significantly, the changes have still been based upon the old paper paradigms. Successful digital initiatives in the scholarly domain have generally replicated existing analog activities, although making them much faster or more productive. Thus journal literature has migrated from print to digital formats while leaving much of the framework of paper journals in place (pages, articles, issues, journals). Such efforts allow scholars and students to do what they have always

done – find and read an article – much, much more conveniently, but they have not fundamentally changed the way research, teaching, and knowledge dissemination take place. There has been much progress – and there remains much more to be done in digitizing more legacy content, further improving search techniques, and making links ever more valuable and intelligent – but the point to be highlighted here is that recent and transformative changes in how people communicate and collaborate on the commercial web have not yet deeply penetrated the environment of authoritative scholarly communication, libraries, and publishing.

JSTOR, the online archive of scholarly journals, is a telling example of this phenomenon. It was one of the first online resources to be widely adopted by faculty and students, to the point where many now consider it indispensable for their work. As described above, one of the primary reasons for JSTOR's success was that it allowed people to do the same thing they had always done much more efficiently and conveniently. A resource like JSTOR presumes that researchers are using the same process they have always used to find, retrieve, and engage intellectually with the information. A full-text or keyword search improves upon searching the card catalogue or article index. Users can browse by issue or title, the equivalent of manually scanning a series of articles in the stacks. The MyJSTOR functionality of saving citations is equivalent to keeping a personal bibliography or notes, and reference linking saves another search in the card catalogue, or another visit to the library.

While innovative in their presentation and ability to link resources, even resources like the Valley of the Shadow built on familiar and long-established paradigms of how primary sources are selected, curated, and used in teaching. An online encyclopedia is essentially the same as its print counterpart, only richer and easier to update. By contrast, resources that enable faculty to do things in completely new ways and that maximize the capabilities of the new technologies have been much slower to take off. Nor have new paradigms of electronic resources been accepted as equivalent to traditional forms of publishing for establishing academic credentials on the path to tenure.

Two Worlds Collide

Against this backdrop, the evolution of the web in the past three to five years is beginning to pose profound new challenges for the academy. Consensus has been growing in recent years about the existence of a growing gap between the information needs and preferences of scholars and researchers and the development trajectory of the online informa-

tion environment. Non-academic entities like Google and Facebook are increasingly shaping user needs and expectations as scholars and students expect to see the same functionality in the resources they use for scholarly research. The academic community is using tools that were designed to support a consumer-focused and advertising revenue-driven search business to do serious academic research. The same tools used to find driving directions or a good local restaurant are used to launch an investigation of Faulkner's narrative technique or the role of the Philadelphia chromosome in a leukemia diagnosis.² Students and researchers are drawn to these tools by their desktop convenience, ease of use, and broad reach. The situation is compounded by the emergence of a new class of "digital natives:" those (often of a certain generation) who intuitively turn to the Internet for all types of information, have high demands of efficiency and convenience when searching online (and who may occasionally forgo accuracy or thoroughness because of this), and who are building extensive online identities through social networking sites. Today's college freshmen were literally born about the same time as the World Wide Web, and have integrated it into their lives like no other generation before them.

In effect, two formerly separate worlds are colliding as the traditional barriers between the academic and commercial sectors begin to crumble. The existence of the network means that the exchange of scholarly information is no longer segregated from commercial activity. In the past, the academic sector existed as its own separate ecosystem. The supply and demand for the products of scholarship were largely insulated from the commercial marketplace. Scholarly publishers, libraries, and universities all co-existed in a mutually supporting and self-sustaining environment that served them well for decades. Not surprisingly, the services provided by the commercial web are not tuned to the needs of the academy. Scholars, researchers, and faculty members wish to marry the scholarly values of authority, quality, peer review, and preservation with the "digital values" of access, interactivity, and speed. However, they do not find that the scholarly resources they rely on, the professional interactions they crave, and the digital tools and resources they wish to use are easy to integrate in the library, publishing, research, and information environments currently available to them.

A second reason the academic sector is making slower progress in the transition to the digital era has to do with scale and centralization. Scale matters in the digital age. To the casual end-user, this point is often counter-intuitive – putting something online seems to cost next

to nothing. However, the underlying infrastructure necessary to provide the services expected by today's digital natives requires substantial investments of capital for hardware and software, as well as significant ongoing costs to maintain and upgrade. Although marginal costs to support the *n*th use are very small, they are not zero; when enterprises reach massive scale, a tiny cost multiplied by millions of uses results in large costs. In order to be cost effective, the costs must be spread out over large numbers of users. Moreover, the value of the network lies in its ability to aggregate and link information with users – the more users and sources of information the greater the value. Paradoxically, even the most decentralized web-based services, such as eBay, YouTube, Facebook, and Wikipedia, require a centralized coordinating enterprise or catalyst. They are overcoming the challenge of coordinating infrastructure and end-user services. Commercial entities are able to accumulate the scale and market power to thrive in this environment, but the academic sector, for reasons described below, faces structural issues that work against achieving the kind of scale and centralization that enables change at the pace of the digital age.

The world of Facebook, MySpace, and Twitter represents something radically different from the insulated world of higher education. These entities, springing up seemingly overnight, require a scale, pace, and degree of centralization that is unprecedented in the academic community. Hotmail had 8.5 million subscribers in 18 months and, after it was bought out by Microsoft, had 30 million users a year later.³ MySpace grew from 2 million users to 100 million users in just over two years.⁴ Launched in December 2005, YouTube was sold to Google for \$1.65 billion one year later. And, as the experience of Netscape illustrates, the fall can be as rapid as the rise. No organization in this economy has the luxury of sitting still for long; today's new innovation and value-added service is quickly adopted by its competitors and becomes tomorrow's commodity.

The challenges posed by comparative lack of scale are not only evident in these large social networking enterprises. The long-term trend towards consolidation in scholarly publishing over the last two decades has resulted in a tremendous concentration of power in a relatively small number of large commercial publishers, which in turn has resulted in steep increases in the cost of essential scholarly publications. The predominance of these large commercial concerns, and their ability to bundle journals into "big deal" packages that are more economical than purchasing the same journals separately, dramatically reduces the ability of smaller university-based publishers to compete for a share of library

budgets on a level playing field. At the same time as they face these marketing hurdles, the smaller publishers also lack the economies of scale necessary to invest in building the technological platforms and tools necessary to compete with these large commercial concerns.

Realignment is Hard

These new realities – the collision with the formerly separate world of commerce and the necessity of scale – have profound implications for the higher education sector. The entire enterprise – libraries, publishers, societies, and universities – is being forced to realign itself. Libraries are struggling to define their role as scholars become more self-sufficient in finding resources for their work online. University presses are finding themselves increasingly marginalized.⁵ Not-for-profit publishers and societies are scrambling for new economic models. For a variety of reasons, the necessary realignment is extraordinarily difficult to achieve.

First is the simple fact that changing course is time-consuming and difficult for anyone, much less the large, high-profile, traditional institutions that characterize the system of higher education in most countries. There is a tendency to assume that existing practices will persist indefinitely, that “the way we do things” is essential and intrinsic to the identity of the institution. Indeed, universities and libraries typically pride themselves on resisting forces of change, adopting a self-conscious long-term perspective on preserving traditional values and practices. It takes time for leaders of such institutions to even recognize the necessity for change – and courage to be willing to abandon ways of doing things that have “stood the test of time.” In this environment – especially the rapidly evolving world of technology – new entrants have a significant advantage. They can innovate and move quickly without the burden of legacy and history. The clash of cultures and values can be seen in common responses to the pressures brought to bear by the commercial web. As new innovations sweep across the online environment, the academic community often assumes that the traditional ways of doing things will withstand the external pressure. Yet those traditional “fortresses” often turn out to be sand walls that are quickly breached. For example, it was widely assumed only a few years ago that Google would never be used to find curated, high-quality resources – but then Google Scholar and Google Books entered the scene. Similarly, it has long been assumed that eBooks would never be a substitute for a printed book – yet now innovations like the Kindle, iPad and other book reading devices are calling those assumptions into question as well. Even Wikipedia, long-scorned

as a low-quality, unreliable resource, is growing in credibility and use simply because of its enormous scale and ubiquity.

Second, nearly every aspect of the existing infrastructure in the academy – both capital and services – has been tuned to support the paper-based exchange of information. The academic enterprise is facing a new level and form of external and disruptive competition. Transformative change will require a substantial reallocation of resources, and established institutions are understandably reluctant to make changes on this scale until they are confident they understand how the digital economy works. Inertia rules the day. Again, new entrants do not have to deal with these existing cost structures and their level of risk is much lower.

Third, there are structural factors that make it difficult for dominant players to emerge in the academic sector, at least in the United States. One such factor is the highly specialized nature of the academy, where disciplines – and sub-disciplines – tend to work within vertical stovepipes, making the development of large-scale interdisciplinary tools and resources difficult. Even services that perform a similar function – for example, platforms for preprints or working papers – are operating separately in physics, economics, and political science, all duplicating functionality but not allowing for integrated search or tagging. A second factor is that academic culture favors building custom solutions in lieu of compromising in order to permit shared endeavors. There is a propensity for the members of the academic community to want to “do it themselves.” This has been possible during the first years of the current massive move to the web, in part because there have been relatively abundant resources in higher education.

Finally, there are strong competitive dynamics that pit universities against one another in ways that work against deep and broad collaboration. Many colleges and universities have had access to sufficient resources to support local and custom initiatives, which are appealing because they remain under local control and reflect campus distinctiveness, even if they do not represent either the best or the most economically efficient solutions. Without collaboration, the capital and other resources necessary to build a common scholarly infrastructure, and the number of institutions that rely on it, will be limited.

Despite the powerful market forces putting pressure on the academic sector – to a large extent beyond the control of any individual institution or set of actors – the higher education community is unlikely to change overnight. It is an enormously diverse and complex system, with multiple actors and sub-sectors, all reacting to their own particular circumstances.

The current economic downturn only adds to the uncertainty. The next few years are thus likely to continue to be a “messy stage” in the transformation process, with a variety of experiments and new models, successes and failures, all playing out in unpredictable ways. Some initiatives are focusing on aggregating materials according to content type or format, i.e. journals, books, archives, images, etc. The Google books project is a commercial example of this on a huge scale, with 10 million books already scanned. ARTstor is a non-profit example focusing on art images. Other initiatives are building partnerships at the institutional level. Hathi Trust, a non-profit repository of scholarly resources, exemplifies this approach by building partnerships among universities to leverage the resources invested in a wide range of digitization initiatives, including Google Books. Yet another approach is to aggregate content by discipline. Two leading examples are arXiv, an open-source, decentralized, community-run repository focusing on the sciences, and SSRN, a for-profit community-built repository of social science content. Commercial publishers and aggregators, such as EBSCO online, ProQuest, and Thompson Gale, are pursuing their own models, attempting to bridge disciplines and formats by building increasingly sophisticated platforms that provide a range of services to scholarly content holders and authors. It remains to be seen which of these models, or which combination of models, will eventually emerge to serve the new and evolving needs of scholars.

What do Scholars Want in the Digital Era?

Meanwhile, as all these various initiatives are appearing on the scene, each hoping to find their niche, the limitations of the current state of the scholarly infrastructure are becoming increasingly clear. Some of the problems that scholars experience illustrate the difficulties that arise when they attempt to pursue research in environments designed for other purposes. Faculty users must sift through immense amounts of online materials to find what is relevant, and they often must navigate through a variety of disconnected resources with widely differing interfaces, data standards, levels of authority, and quality controls to follow the threads of scholarship. The ready availability of a broad pool of information can serve to suggest new avenues for exploration, it is true, but it also tends to discourage deeper searches. Those who wish to pursue deeper searches find that many of the most important sources (primary source materials, datasets, secondary materials) are not yet available online. As

data and published material migrate to the online world, there are fewer assurances of long-term preservation, and even over the short-term, some materials that are available today become dead links tomorrow. Students face a similar set of challenges, and are also generally less well-equipped to assess the credibility of the resources they find online or to locate the most authoritative or important work in a particular field. They need more guideposts and filters to help them land in the right places, as a librarian would have helped them do for print resources. Google and commercial search are therefore inadequate tools for serious scholarship.

What is needed, many believe, is a new scholarly “cyberinfrastructure” to meet these needs: to create an information environment in which scholarly virtues and digital virtues are integrated and mutually supportive.⁶ This infrastructure should be flexible and facilitate the easy access to and preservation of the valuable inputs to knowledge creation. It should offer the tools necessary to turn those inputs into useful knowledge. In sum, it should offer an environment that is specifically suited to a new and evolving process of digital scholarship, supporting the work of both the knowledge user and the knowledge producer.

More specifically, the scholarly community needs an infrastructure to support coordinated repositories that are able to hold all kinds of content at all stages of the process of knowledge production (data, drafts, pre-prints, published articles), as well as guarantee the long-term preservation of the important content. The tools should make it easy to discover content, contribute content, integrate and organize materials, share and collaborate around the content, evaluate and rate (peer review) content for the purposes of quality screening, validation, and certification, and ascertain how content has been used and evaluated by others. This type of comprehensive infrastructure does not currently exist, although numerous examples and small-scale pilot projects have demonstrated the potential. Yet the range of obstacles and structural factors described above are so far preventing the emergence of a widely accepted solution.

Challenges to the Sustainability of Content Businesses – Lessons from Newspapers

While the end point of the transition that is now underway is not yet known, it is certain that change is coming and that many existing models may not survive. A key factor in a successful transition depends on the economics that support new initiatives and approaches. As a cautionary

tale, one need only look to the experience of commercial newspapers to see the kind of disruption that might lie in store for scholarly publishing. Like many academic publishers, the newspaper industry has a long and stable history. Newspapers built their brands on strong editorial values and high quality authorship, and competed in relatively small geographical niches, similar in some respects to restricted domains of specific academic disciplines.

In the past, newspapers were to a large extent “protected” from external competition as a result of specializing in local content not available elsewhere, as well as the physical format of papers that kept distribution to a fairly close range, and the limited other choices for readers in their region. For advertisers wanting to reach those readers, the newspaper was deemed the most reliable method in most markets. Newspapers did not really compete with other broadcast media – they reached a different audience. Most newspapers pursued a blended economic model that relied on both subscriptions and advertising. An important component of revenue was classified advertising, for which there were few truly competitive options in local markets. In a similar way, academic materials have traditionally inhabited a relatively protected market space due to their specialization and academic credentialing.

Initially, as digital technologies were introduced, the trends looked quite promising for newspapers. New production technologies brought cost savings while ad revenues continued to grow. During the 1990s profits nearly doubled, from 14 percent to 27 percent for the newspaper sector as a whole.⁷ The only troubling sign, largely ignored at the time, was a decline in circulation for the first time in decades. In a similar way, scholarly publishing seemed to be on solid ground during the 1990s as digital technologies were introduced. Similar improvements in supply-side infrastructure supported profits as the first transition of journals moved electronic. Yet cost efficiencies, the emergence of “big deal” offerings to consortia that aggregated large numbers of journal titles, and a robust economic environment were masking a developing underlying crisis associated with an increasing supply of new journals that would soon exceed available resources in the face of flat library budgets.

In the past two to three years a sharply different picture has emerged. The newspaper industry is at a precipice. Newspapers’ average profit margins have been in decline since 2002, and their stock prices fell 83 percent in 2008 after a drop of 42 percent from 2005 to 2007.⁸ Print ad revenues are dropping, with 2008 seeing the worst decline in 50 years, a fall of 17.7 percent from 2007.⁹ Even online advertising is showing signs

of a slowdown.¹⁰ The number of newspapers is shrinking, and some papers are cutting back on their print editions on lighter advertising days. Even leading brands like the *New York Times* and *Los Angeles Times* are facing severe revenue shortfalls as subscriptions dwindle and ad revenues decline.¹¹ Staff cuts over the past few months have included 70 positions at the *Los Angeles Times* in February 2009, where total staff has declined to 570 from 900 at the start of 2008. In 2008 the *Washington Post* reduced newsroom staff by 100, and the *Baltimore Sun* laid off 60 newsroom employees in April 2009. The Gannett Company, the largest U.S. newspaper chain, cut approximately 3,000 workers in 2008, about 10 percent of its workforce.¹² *The Guardian* offered media training to all journalists, saying it will not cut those willing to “make the transition to online.”¹³ Business leaders both inside and outside the industry have sounded the alarm. Warren Buffet pronounced that newspapers are “a business of permanent decline” and the editor-in-chief of the *San Francisco Chronicle* concluded: “The business is broken, and no one knows how to fix it. . . and if any other papers say they do, they’re lying.”¹⁴

Initially, the main impact was in staff cuts, but a more recently there has also been a wave of bankruptcies. The Tribune Company, publisher of the *Chicago Tribune* and the *Los Angeles Times*, declared bankruptcy in December 2008, \$13 billion in debt. It was followed by the *Minneapolis Star Tribune* in January 2009, the *Philadelphia Newspapers* and the Journal Register Company in February 2009 (owner of 20 daily newspapers and 159 other newspapers), and the Sun-Times Media Group, publisher of the *Chicago Sun-Times*, in April 2009.¹⁵

What is behind the threat to the traditional newspaper business model? First, newspapers face new competition for their audience. Newspapers now compete online for readership with other newspapers from all over the world. They also compete with producers of all traditional news media, not just newspapers (the BBC, CNN, National Public Radio, etc.). Although readers have gone online in droves, they can get news from many other sources aside from newspapers’ websites (blogs, newsletters, or hosted sites like Wikipedia for fast-moving news). Second, newspapers face new competition for ad revenue. They now compete with all other online news sites for ad dollars. Classified ads face tough competition from disruptive new entrants (eBay, craigslist.org, etc.). Online advertising, though growing, only accounts for a small part of newspaper revenues – under 10 percent on average – partly because of the way online ads are valued. Some papers, in an effort to attract more visitors (and more ad dollars), have chosen to become free online, dropping subscrip-

tion revenues. Yet online advertising revenue is not projected to meet the mix of subscription and advertising revenue that existed in the print environment. Recent announcements by the *Financial Times* and rumors from Rupert Murdoch suggest that these newspapers will try to return to a subscription model to support journalism.

The experience of the newspaper industry serves as a cautionary tale for academic publishers, and indeed for any traditional publisher. It demonstrates how the rush to the web and the promise of free content can have disruptive unintended consequences. In the case of newspapers, one possible outcome is the development of mega-newspapers with huge audiences. They may be the only ones to do professionalized investigative journalism on national and global issues. Will the regional and other newspapers in the middle be able to compete? How many newspapers will go out of business? What will become of their content? Who will cover local issues – perhaps “super-local” newspapers? Concerns have been raised that without subscription revenue, newspapers will be increasingly pressured to shape their editorial content to their advertisers. Whatever the outcome, the digital age is ushering in a dramatically different competitive landscape for the newspaper industry and some existing players will not survive the transition.

Academic libraries have reason to be concerned as well. In a much-shared presentation in January 2008, the executive director of the Digital Library Federation vividly demonstrated the impact on libraries of the shift away from paper. Between 1996 and 2004, in-person visits to Library of Congress reading rooms dropped by over 40 percent. The Association of Research Libraries tracks trends in reference queries and library circulation at its member libraries. Between 1995 and 2003 reference queries dropped by nearly half and circulation declined by about 30 percent (both statistics are indexed to the number of full-time students).¹⁶

While alarming and headline-grabbing, not all observers believe these reports about the imminent demise of the newspaper industry are as dire as they seem. Many newspapers are finding their own niches and adapting in creative ways. For starters, the impact of the web is not nearly as great on newspapers in countries where fewer people have high-speed Internet access, or on smaller local newspapers with little competition. Newspapers in some international markets (e.g. Singapore and Australia), where the industry has already undergone significant consolidation and where papers tend to focus primarily on local news, are doing reasonably well.

In addition, publications with strong “niche” audience, such as the

Wall Street Journal and Bloomberg, demonstrate that subscription models are still viable. The movement toward aggregation and coordination also helps provide scale to attract more advertisers (Yahoo! and the Quadrant One ad networks being cases in point). For some observers, the current situation is more of a wake-up call than a crisis:

If you'd ask me to describe the state of the newspaper industry based on the scary coverage about it alone, I would have guessed that it had fallen by half and that we were back to 1970s levels. Instead, it's a \$45 billion business, which is twice as big as Google and Yahoo combined. ... The truth is that the newspaper business is still a huge industry and will be around in one form or another for the rest of my life. That is not to dismiss the declines, but only to note that there's still a lot of money there and what is required is strategic change, not giving up the ghost.¹⁷

Similarly, the major takeaway for academic publishers is not that their situation is hopeless, but that they need to think deeply about their value proposition, adopt a flexible, entrepreneurial mindset, and be prepared to make strategic changes and try new models. Their greatest danger may lie in sitting still.

Sustainability Models Being Pursued to Support Academic Online Content

If newspapers have struggled so in the transition to digital, what are the prospects for the academic sector? The past decade has witnessed a surge in new academic content coming online, not only from traditional academic publishers and libraries, but from a wide assortment of new players – ranging from academic departments to scholarly societies, cultural institutions, government agencies, and other not-for-profit organizations. Millions of dollars have been invested by foundations, state agencies, universities, and cultural organizations to create these resources. Many of these projects have now been in existence long enough that they are well past the start-up phase, and a number have become essential resources for certain academic audiences. How have they fared, and are there any broad lessons or trends emerging that can shed light on how the transition to digital may take shape in the years ahead?

One clear trend is that external grant funding for new digital projects is beginning to dry up, especially in light of the recent economic down-

turn, and there is increasing pressure for digital projects to demonstrate financial sustainability. Many foundations, having seen projects launched with great promise only to disappear a few years later, are insisting that grant proposals include strategies and business plans to ensure that digital content lasts beyond the start-up phase. Over the past two years, Ithaka's Strategy and Research group has examined the question of sustainability through a series of reports and workshops, and, most recently, through in-depth case studies of 12 organizations representing a range of sustainability strategies and revenue models. The objective was to illuminate the experiences of real-world projects, both positive and negative, and provide examples from which other projects can learn. While a major focus was on understanding revenue models, we also explored the evolution of strategy and the decision-making process of those managing the projects.

One of our first challenges was to come up with a working definition of sustainability, which was not as straightforward as we had anticipated. We began with a narrow definition that focused on economic resources, but after interviewing project leaders came to a somewhat broader conception: *sustainability is the ability to generate or gain access to the resources – financial or otherwise – needed to protect and increase the value of the content or service for those who use it.*¹⁸ Sufficient economic resources are required to support the personnel and non-personnel costs associated with project management and administration, content development, technological infrastructure, business planning, and revenue generation. But by recognizing that not all resources are financial in nature, our new definition acknowledges the fact that many projects rely on a wide range of resources such as volunteer labor, partnerships, and in-kind contributions. In addition, one clear lesson from the case studies was that sustaining the value of the resource requires more than just “keeping the lights on.” As new technologies develop and user expectations shift and grow, a resource risks fading slowly into irrelevance if it does not constantly grow and innovate in ways that continue to benefit its constituents. Project leaders must devote time and resources towards ensuring that the resource remains vital and valuable to its users over time.

The 12 organizations profiled in the case studies represented a range of sectors, revenue models, organizational models, and sustainability strategies: five projects were from the UK, including resources such as parliamentary papers, correspondence from leading figures in the Enlightenment, and images from the Victoria and Albert Museum, four were from the US, ranging from Greek literature and philosophy to amateur bird sightings, and one each was from Egypt (open access journals),

France (audiovisual content), and Germany (scholarly journal backfiles).¹⁹ The revenue generation strategies employed by the 12 projects included: institutional subscriptions, content licensing, advertising and sponsorships, endowment income, grants, government subsidy, open access (both author-pays and cross-subsidy models), premium features combined with “free” content, and a wide range of hybrid models combining two or more of these elements.

At a superficial level, most of the projects we studied appeared to be sustainable, in that they were generating enough revenue to cover their direct costs. But a closer look revealed that a significant factor in doing so was keeping their apparent direct costs low by shifting many expenses to the parent or host institution. This in-kind support took a variety of forms, such as free rent and utilities, technical support or server space, contributed time and effort of staff, or non-billed overhead costs for financial management, human resources, and legal services. In one way or another, this was true for every one of the not-for-profit projects studied (one, the Hindawi Publishing Corporation, was a for-profit and did not have a host institution or parent company). In addition to in-kind support, several projects received direct financial support from their host institutions through contractual arrangements of one kind or another. Two of the projects were still in start-up phases relying on grant funding and were either still grappling with long-term sustainability strategies or it was too early to know the success of their plans. Nor was it clear that all of the projects had adequate plans in place to fund the upgrades and enhancements that will be needed to maintain value to their users in the medium to longer run.

Reliance on a third party such as a parent university is inevitably riskier than a strategy built on revenue tied more directly to the value of the content, as this institutional support is subject to local environmental factors – such as the budgetary needs of other departments, or a new provost’s priorities – that have nothing to do with the project itself. To the extent that institutional support is hidden or not accurately quantified, leaders may incorrectly assume their project is relatively inexpensive to operate.²⁰ As projects face greater scrutiny during times of budget austerity, the full costs may become evident and the institution may choose to reduce its support, leaving the project in a highly vulnerable position. Not many leaders had articulated a strong value proposition to their host institution, or developed contingency plans in case institutional support dries up. We also observed that a number of the project teams lacked business expertise and entrepreneurial skills, or were heavily dependent

on one or two key leaders – factors that further increase project risk. Furthermore, as noted in the first section of this paper, academic projects tend to grow in relative isolation, and are generally not interoperable with other important resources in their field. This tends to decrease their value in the eyes of senior management and the broader community of users and funders. Although none of the projects we looked at appear to be in any imminent danger of disappearing, our research does raise questions about how sustainable they – and the hundreds of projects like them scattered around the higher education landscape – really are in the long run. More fundamentally, the case studies graphically illustrate the sustainability challenges that affect the sector as a whole. When all the individual hidden costs, cross-subsidies, and inefficiencies are added up, how much financial burden do these types of projects place on the entire higher education system? How long can institutions afford to shoulder these costs, without even realizing the full potential of the networked environment since so many individual projects are under-resourced, disconnected, and not really meeting the needs of scholars and students? And, as resources become more tightly constrained, can universities continue to justify expending resources to “serve the world” and not just their own constituents? As budgets become tighter, there is a risk that institutions will only retreat further into their own silos.

Conclusion

In order for traditional universities and cultural institutions to survive the transition to the digital era, they will have to reallocate resources in a massive way. The paper era is passing away, and infrastructure needs to be reshaped, from top to bottom, to serve digital dissemination instead of physical dissemination. New competitors are marshalling the resources and showing the scale and agility needed to take advantage of the new networked environment. In theory, the academic sector has the resources and depth of talent to meet the challenge, but will universities be able to move fast enough, or will they endlessly be running to catch up, reacting to forces beyond their control rather than leading the way? Experience has shown that reallocating resources within universities is not an easy task. Political challenges and turf battles are inevitable in decentralized institutions where every budget line item has its champion. Savings in one budget are not easily transferred to another. For example, can subscriptions to high-profile journals be redirected to support university-based

publishing options? Will faculty accept such changes, given the current incentives for tenure? It will take leadership and courage from senior management to make the difficult choices that are needed.

If reallocation within institutions is so difficult, one wonders if transformative reallocation of resources across institutions – across the entire sector – will even be possible. One thing that seems certain is that new forms of collaboration will be needed to accomplish the transition and break down the silos and competitive pressures that are at odds with a fast-moving networked environment. It is not that collaboration is not already a part of academic culture – it is, but it tends to be organized around membership models in which all participants are equal and change is at the margins. Such collaborations have rarely resulted in systemic, radical change, at least not at the pace and scale of entities like Google, Facebook, and Twitter.

Ten plus years into the digital transition, we know what the challenges are and we know what faculty and students are looking for in a scholarly cyberinfrastructure. The challenge now is actually moving forward and putting in place the system-wide changes that are needed. Unfortunately, the academic community is not optimally organized for this kind of change. It may be that, in the digital era, more independent catalysts are needed – organizations that share the values of the academy but are not weighed down by past investments and can mobilize institutions to work together so that all benefit.

As we contemplate the current state of affairs, there are both risks and opportunities for the academic community. The risk is that, by not acting on their own behalf, and ceding control over much of their infrastructure to the commercial sector, the values and mission of the academy will be compromised. Innovation will continue to occur, but it will be fragmented and incremental in nature, and unlikely to live up to its full potential. Rather than participating in building a new infrastructure that takes advantage of the network to maximize the benefits for higher education, the academic sector will be left to make the best of solutions dictated by the commercial web.

The opportunity facing the academic community is to recognize the inevitability of the transformations that are coming and to work together to create solutions that will support its mission and values. The university sector represents a vast reservoir of innovation and creative energy. At the same time, it is sobering to realize that in recent years the most significant changes have come through brand new enterprises, not through established institutions and traditional approaches. The challenge will be

to overcome the structural issues that have often limited collaboration in the past, and establish new structures and approaches that, while not necessarily aspiring to compete head-to-head with commercial entities, can at least generate the resources and critical mass needed to be an effective counterweight to best serve the academic sector.

References

- 1 For more information on the Valley of the Shadow project, see <http://valley.vcdh.virginia.edu/>. Information on the Making of America project is available at <http://quod.lib.umich.edu/m/moagrp/> and <http://cdl.library.cornell.edu/moa/>
- 2 In 2006, almost two-thirds of faculty surveyed by Ithaka said that they used Google often or occasionally to find information in journals. We believe that scholars' reliance on Google has increased significantly since then. In 2007, 75 percent of JSTOR's inbound links came from Google URLs; these accounted for 23 percent of total article views in the archive.
- 3 "Hotmail, Microsoft talk deals," *CNET News*, December 5, 1997, <http://news.cnet.com/2100-1023-206039.html>. The Microsoft figures are reported in a Microsoft press release, "MSN Hotmail: From Zero to 30 Million Members in 30 Months," *Microsoft PressPass*, February 8, 1999, <http://www.microsoft.com/presspass/features/1999/02-08hotmail.msp>
- 4 Pete Cashmore, "MySpace Hits 100 Million Accounts," August 9, 2006, *Mashable: The Social Media Guide*, <http://mashable.com/2006/08/09/myspace-hits-100-million-accounts/>
- 5 Laura Brown, Rebecca Griffiths, and Matthew Rascoff, "University Publishing in a Digital Age: An Ithaka Report," July 26, 2007, *Ithaka*, <http://www.ithaka.org/ithaka-s-r/strategy/Ithaka%20University%20Publishing%20Report.pdf>
- 6 See "Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences" (American Council of Learned Societies, 2006), http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf, and "Cyberinfrastructure Vision for 21st Century Discovery" (National Science Foundation Cyberinfrastructure Council, March 2007), <http://www.nsf.gov/pubs/2007/nsfo728/nsfo728.pdf>
- 7 Journalism.org, "The State of the News Media 2004: An Annual Report on American Journalism," under "News Investment," <http://www.stateofthemedias.org/2004/>
- 8 Journalism.org, "The State of the News Media 2009: An Annual Report on American Journalism," under "Newspapers," <http://www.stateofthemedias.org/2009/>
- 9 Newspaper Association of America, "Trends and Numbers, Adver-

- tising Expenditures,” <http://www.naa.org/TrendsandNumbers/Advertising-Expenditures.aspx>
- 10 “The State of the News Media 2009” under “Economics.”
 - 11 On March 17, 2009, after no buyer could be found, Seattle’s second largest newspaper, the *Seattle Post Intelligencer*, printed its last edition after 146 years of operations. Rather than closing completely, it is one of the first major papers to experiment with an online-only version.
 - 12 Figures for the *Los Angeles Times*, *Washington Post*, and Gannett Company are from “The State of the News Media 2009,” under “News Investment.” The report on the *Baltimore Sun* is from the Baltimore Business Journal, “Baltimore Sun Newsroom Cuts Include Top Editors,” April 29, 2009, <http://www.bizjournals.com/baltimore/stories/2009/04/27/daily36.html>
 - 13 Editors Weblog, “UK: *The Guardian* will not cut older staff for newer media skills, provides multimedia training” (interview with Chris Elliot, Managing Editor of the Guardian), March 28, 2008, http://www.editorsweblog.org/newsrooms_and_journalism/2008/03/uk_the_guardian_will_not_cut_older_staff.php
 - 14 Phil Bronstein, editor-in-chief, *San Francisco Chronicle*, as reported in O’Reilly Radar, “SF Chronicle in Trouble?” March 23, 2007, <http://radar.oreilly.com/2007/03/sf-chronicle-in-trouble.html>
 - 15 “The State of the News Media 2009,” under “Ownership.”
 - 16 Peter Brantley, “What Rupert might tell the DLF: Why libraries are like newspapers, and how we avoid lining the pet cage,” January 2008, <http://www.slideshare.net/naypinya/what-rupert-would-tell-the-dlf>. Peter Brantley is currently the Director of the Internet Archive.
 - 17 Chris Anderson, March 29, 2008, www.thelongtail.com (italics added).
 - 18 Nancy L. Maron, K. Kirby Smith, and Mathew Loy, “Sustaining Digital Resources: An On-the-Ground View of Projects Today,” July 2009, p. 11. The earlier definition for sustainability, from Ithaka’s May 2008 report, was: “having a mechanism in place for generating, or gaining access to, the economic resources necessary to keep the intellectual property or the service available on an ongoing basis” (Kevin Guthrie, Rebecca Griffiths, and Nancy Maron, “Sustainability and Revenue Models for Online Academic Resources: An Ithaka Report,” May 2008, p. 10). Both reports are available on the Ithaka website at <http://www.ithaka.org/publications>.
 - 19 The 12 projects include: the BOPCRIS Digitisation Centre at the Uni-

versity of Southampton, the Centre for Computing in the Humanities at King's College London, DigiZeitschriften, based at the University of Göttingen, the eBird project at the Cornell University Lab of Ornithology, the Electronic Enlightenment at Oxford University, the Hindawi Publishing Corporation in Cairo, L'Institut national de l'audiovisuel in France, the National Science Digital Library Middle School Portal based at Ohio State University, the UK National Archives Licensed Internet Associates Programme, the Stanford Encyclopedia of Philosophy, the Thesaurus Linguae Graecae at the University of California, Irvine, and V&A Images at the Victoria and Albert Museum in London. The full case studies are available at <http://www.ithaka.org/ithaka-s-r/strategy/ithaka-case-studies-in-sustainability>

- 20 One of the major challenges in the case studies project was getting a complete picture of project finances that could be accurately compared across organizations and projects. In part to address this problem and ensure accountability for increasingly scarce public funds, in the U.K., since 2006, all funding proposals are required to use TRAC (Transparent Approach to Costing) as the standard method for costing projects. TRAC is based on principles of full economic costs and is designed to capture these sorts of hidden costs or in-kind contributions.

Information structure, infrastructure and metadata

DIGITAL HUMANITIES: WHY WORRY ABOUT READING?

Terje Hillesund

Introduction

To state that “reading is at the heart of the humanities” may seem a perfect truism. However, if reading so indisputably constitutes a core activity in philosophy, literature and other fields in the humanities, it might have been expected that reading presented itself as a central area of research in the digital humanities, in which computers are extensively used for analysing and reading text. Curiously, the emergence of a new digital reading paradigm is not an issue in the digital humanities.

On the contrary, critical interest in digital reading seems to have diminished. In the 1990s the focus was on digital remediation of text¹ and hypertext² and, according to Claire Warwick, “many humanities scholars expressed excitement about the possibilities of electronic text, predicting that the experience of reading would change fundamentally.”³ Despite predictions in the first half of the 1990s of great changes in scholarly reading, she observes that in “the last fifteen years critical interest within humanities circles with respect to reading has waned and little progress has been made in understanding how electronic textuality may affect reading practices, both of academic and non-academic readers.”

While one might suspect Warwick of slight exaggeration, or underestimation, an examination of the Digital Humanities conference programmes over the last years shows that very few papers specifically deal with reading, let alone the diversification of reading resulting from the evolution of digital texts. Nor does *A Companion to Digital Humanities* have chapters specifically on the actual uses of humanities Web sites.⁴ The *Companion* deals with many aspects of humanities computing, including text encoding, electronic scholarly editing, textual analysis, archiving and preservation, in archaeology, history, linguistics, literary studies and music, but the focus is clearly on production rather than dissemination and reception. A little better off is *A Companion to Digital Literary Studies*, with several theoretical articles on reading, mostly of hypertext literature.⁵ In *Scholarship in the Digital Age*, which includes critical editing, digital libraries and scholarly publishing, Christine L. Borgman uses very little space on end-users and their reading of digital scholarly publications. On the issue of online scholarship, Borgman maintains that: “An initial question in constructing a [digital] scholarly infrastructure is what to build when we know so little about how it will be used.”⁶

In a self-reflective paper, Peter Robinson – on behalf of digital editorial philology

– ponders the slow changeover from printed critical editions to digital text editions,⁷ and similarly Patric Juola examines how “the emerging discipline of ‘digital humanities’ has been plagued by a perceived neglect on the part of the broader humanities community.”⁸ As John Bradley points out, only a small percentage of humanist scholars “go beyond general purpose information technology and use digital resources and more complex digital tools in their scholarship.”⁹

For this situation to improve, Robinson and Juola call for new tools and killer applications; Robinson for collation tools and Juola tools for index generation, resource exploration and collaborative research. However, their analysis is primarily done from the editors’ perspective and, as Lisa R. Schiff emphasises, in addition to addressing the problems involved in producing digital critical editions, a major area of concern must be the feasibility of such works on the part of the general scholar and lay researcher.¹⁰ In line with Schiff’s view, Warwick and her colleagues claim it is incumbent for producers of digital resources to understand the working practices of the scholars for whom they design, and Bradley suggests that tool builders in the digital humanities would have greater success persuading their non-digital colleagues if the digital tools were more in line with the ways in which humanities scholarship is generally conducted.¹¹

In different ways, Schiff, Warwick, Bradley and Borgman call for more “use-inspired basic research.”¹² This paper will present results from such a user-oriented study. By focusing on text materiality and physical aspects of reading, the study examines working habits among humanist scholars, trying to discern challenges in transferring sustained reading of long-form text on to a digital platform. First, however, a short presentation of theories on reading will indicate that much research relevant to user issues is actually taking place. Considering the request for user studies, it is likely that perspectives from this research will soon pervade discussions in the digital humanities.

Research on reading

In Western research, the historical and geographical multiplicity of writing systems (clay, stones, parchment, Chinese characters, Korean hangul etc.) is often recognised. Nonetheless, reading is usually treated as interpretation of alphabetic text, including pattern recognition, decoding of letters and words, and – against the background of cultural understand-

ing – construing of lower level interpretations into meanings at increasingly higher levels.¹³ Even if reading for literate people is automated, it is a very complex activity, comprising the physical text and its context, as well as the use of hands, eyes and brain.

In neuroscience, much research is done in order to determine what parts of the brain are involved and how it works when we are reading.¹⁴ The partly linear and saccadic eye movements during reading have been thoroughly traced,¹⁵ and combined with centuries of typographical experience this has given a good idea of how to ensure readability and how to present text in order to support fluent reading of lengthy texts.¹⁶ In literacy studies, scholars have described the variety of text media, the history and sociology of reading, and the global spread of literacy.¹⁷ Literary studies, especially post-structuralism and reader-response theory, have contributed multiple theories of text interpretation, focusing on the active role of the reader.¹⁸ Hypertext theorists have called attention to the distinctiveness of multi-linear and interactive reading¹⁹ and, as a recent development; social semiotic researchers have started examining the reception and interpretation of multimodal texts.²⁰ In the history of books, researchers have shown growing interest in the correlations between the materiality of texts and distinct ways of reading.²¹

Current research on reading is immense, and the number of studies on professional journal reading alone is probably in the thousands.²² Practically unnoticed by humanist scholars, large amounts of empirical research have also been done specifically on digital reading, often comparing onscreen reading to paper reading. Mainstream research on digital reading is measuring legibility and comprehension of texts presented on stationary computer displays,²³ on handhelds²⁴ or on electronic paper,²⁵ and preferences regarding paper and screen reading,²⁶ reading of e-books²⁷ and user behaviour in digital environments.²⁸

As this account should indicate, reading is a phenomenon studied within several rather separated fields. Thus, in reading research there is a profusion of perspectives and a multitude of models, with long-lasting oppositions between a constructivist and a dense processing view²⁹ and between reader-oriented and text-oriented perspectives.³⁰ Despite differences, in most traditional research there has been a tendency to treat reading in a rather abstract way, as if all reading were more or less the same – a silent and inwardly individual act of interpretation. Similarly, text is often abstractly assessed, presumed to be a plain long-form text.

Reading, body and the materiality of text

It is probably fair to say that reading research has a strong focus on visual perception and the cognitive aspects of reading, that is, on reading as a mental activity, sometimes with an emphasis on sociocultural bearings on this activity. Until recently, little attention has been paid to bodily aspects of reading, or to the fact that reading involves the physical handling of books, computers and other reading media. However, a growing number of theorists refer to Roger Chartier, who has emphasised the significance of text materiality and corporal aspects of reading. In his treatment of text interpretation and meaning making, Chartier argues that the same text apprehended through a different mechanism of representation is no longer the same.³¹

Chartier points to several historical developments that are pivotal for reading practices and experiences.³² First, and extremely important, was the implementation of the codex in the second and third centuries. Gradually replacing the scroll, codices gave new and more effective ways of organising, accessing and navigating text. However, the early reading of heavy and densely written parchment codices (with many abbreviations and no word spacing) required the service of more or less the whole body, including the voice. Then, during the Middle Ages, the introduction of smaller books and new text features, such as word spacing, punctuation and paragraphs, made reading easier and more fluent.³³ This development led to a consequential shift from oral reading, indispensable for comprehension, to reading that could be visual, silent and fast.³⁴ The printing press provided new ways of reproducing text, made books even more legible and accelerated the spread of literacy, having a profound cultural and social significance.

In his theoretical outline of a history of reading, Robert Darnton underscores the significance of text materiality for reading practices,³⁵ and Alberto Manuel's copious *A History of Reading* is very much a description of codices, printed books and other technologies related to reading.³⁶ In the field of material studies, Mary and Richard Rouse and Paul Saenger have shown that the advancement of the codex to an easily navigable book was a very long process indeed, a process that only accelerated in the 13th century when manuscripts were provided with pagination, indices and concordances.³⁷ For Peter Stallybrass, the art of printing is primarily a culmination in the development of the navigable book.³⁸ In printed books, accessibility was further improved by a more systematic provision of titles, chapters, tables of contents, page numbering, headers

and footers, assisting fingers and eyes in browsing and navigation. Stallybrass shows that printed Bibles in 16th century England were designed to support discontinuous reading, with indices and concordance lists supporting Protestant interpretations of the scripture. Through studies of contemporary book annotations and diaries, Stallybrass documents that the Bible was in fact read discontinuously. That reading at a later stage, in the 18th and 19th centuries, was to be dominated by silent and continuous reading of books, especially novels, can according to Stallybrass be seen as a return to an earlier form of reading: “To imagine continuous reading as the norm in reading a book is radically reactionary: it is to read the book as if it was a scroll.”³⁹

In his analysis, Stallybrass emphasises the continuation and amplification of codex features in the computer and on the Web. In a digital environment many typographical features are the same as before; access to pages is simple, book marking is easy, and advanced search functions make it possible for readers to follow up on themes in a discontinuous reading process, jumping from page to page and site to site. However, while Stallybrass accentuates continuity, Chartier focuses on the break, claiming that the new “immaterial” materiality of digital text inevitably requires new ways of reading.⁴⁰

Among researchers studying current changes, semioticians are particularly preoccupied with materiality of semiotic resources. Since the 1970s and 80s, desktop publishing and offset printing techniques have dominated composition and printing, making use of photo and graphic illustrations far less complicated. Today newspapers, magazines, textbooks and trade books are often sophisticated publications in which much of the information is provided by visual means. Researchers such as Gunther Kress and Theo van Leeuwen,⁴¹ have described the visual grammar of multimodal texts, and they have suggested that multimodal reading is not primarily a continuous or discontinuous reading of verbal text, but rather a composite kind of reading in which attention jumps back and forth between illustrations and text. Researchers encounter great challenges in trying to explain how meaning is construed in multimodal reading.

Among hypertext researchers, George Landow and Jay Bolter are technologically very explicit.⁴² They have examined how the computer alters the materiality of text, the physical handling of text and ultimately the reading of text. Having little of the tactile materiality of printed text, digital text is a volatile virtual image of an electronically stored text. Thus, digital texts are highly editable, extremely moveable and through the linking system of the Web, globally accessible. The links on the

Web and in hypertext literature provide readers with multiple choices; by clicking the mouse, users can choose their own reading paths. Digital reading is thus multi-linear and discontinuous. In a Web environment, text boundaries are not obvious; in a sense the entire Web is one enormous, interconnected text.

While hypertext theorists celebrate a new-won freedom for readers (and writers), others claim that the current shaping of the Web induces a new form of constraint; a psychological urge to click. Asserting that different kinds of materiality impact our reading, Anne Mangen examines the impact of the intangibility and volatility of digital text on the reading process, which she claims is dominated by shallow forms of reading, such as scanning and skimming.⁴³ Taking as a basis Merleau-Ponty's phenomenological concept of humans as bodies-in-the-world and William Thorngate's psychological theories of attention, Mangen argues the near impossibility of getting immersed in hypertext or online reading in the same way we get lost in a book. Most websites provide an abundance of attention-switching possibilities and promise new stimuli in the form of links, pictures and videos. As a rule, then, when we have options to easily rekindle our attention through outside stimuli, we are – psychobiologically as well as phenomenologically – inclined to resort to them, rather than try to resist such distractions by attempting to structure consciousness from within, and thus continue reading.⁴⁴ In front of the computer screen – and especially online – we are tuned in to change; we expect something to happen, and are thus compelled by a powerful urge to click.

Many studies support Mangen's description of online reader behaviour as dominated by shallow reading. Ziming Liu has reviewed the research on digital reading and, backed up by own research, concludes that: "Screen-based reading behaviour is characterised by more time spent on browsing and scanning, keyword spotting, one-time reading, non-linear reading, and reading more selectively, while less time is spent on in-depth reading and concentrated reading."⁴⁵ Liu notes decreased sustained attention, and his results have been confirmed by research at University College London.⁴⁶ In a series of articles, David Nicholas, Ian Rowlands and associates describe viewing and reading habits among academics visiting digital journal libraries. Even if the picture is varied, most of the library viewing is cursory in nature, described as "bouncing;" a pattern of behaviour whereby a high proportion of users view only a few Web pages from the vast numbers available and then never return to the site. Some click forward from lists and abstracts to full-text view-

ing of articles, but as the studies show: two thirds of article views lasted less than three minutes and 40 per cent were completed in a minute or less.⁴⁷ This viewing and bouncing behaviour is called “squirreling” – an energetic search for treasures that are downloaded for later consumption. Scholars seldom read the same text online for a long period of time: from undergraduates to professors, the pattern of behaviour in digital libraries is characterised by Nicholas and Rowlands as “power browsing.”

The UCL projects do not tell us what scholars eventually do with downloaded papers and articles. Many are no doubt forgotten, but some are obviously read. According to Liu, research confirms the general belief that a majority of users often print out electronic documents for reading, and that the traditional habit of highlighting and annotating texts has not migrated to the digital environment.⁴⁸ The latter was also a major finding in the studies of Abigail Sellen and Richard Harper, who in an extensive research project studied paper and computers and their different uses in various organisations.⁴⁹

Sellen and Harper’s studies show that the computer system is superior in the actual making and remaking of documents, in storing, accessing and retrieving documents, and in facilitating the distribution of documents.⁵⁰ Paper, on the other hand, is used in many creative tasks such as editing, commenting and collaboration on text, and in tasks that require certain levels of sustained concentration, such as reading, in which annotation, quick navigation and spatial layout of documents allow readers to deepen their understanding of the text and create a plan for their own writing. Thus, even if the computer is the main tool in writing, paper reading is an important part of most writing processes, resulting in piles of books and print-outs surrounding the writer and the computer. Some of these books and papers are “hot” and spread out within reach for immediate use; others are “warm” and piled up for later or potential use. At some point, documents become “cold” and are removed from the desk.

In her article, Mangen underscores the sensor-motor dominance of the tactile in reading and examines the interplay between the body and the materiality of text. Sellen and Harper’s research richly illustrates different ways in which reading actively brings the body into play and how reading is inscribed in space and time. They also clearly demonstrate how deeply embedded most reading is in practice; reading is an integral and essential part of a multitude of tasks and activities, including – obviously – scholarly study and research.

Research questions and method

Mangen's and Sellen and Harper's perspectives are exactly the focus of a study on reading which I carried out among humanist scholars in the spring of 2009. In 14 semi-structured qualitative interviews 10 participants (all established academics) were asked questions on their reading: how and where is it done, how they sit when reading and how they use their hands and fingers. The participants were asked about their note-taking and underlining, and whether their reading was continuous or done in parts, following links or linear, embedded in scholarly practices or part of ongoing communicative acts. In short: the corporal, material and contextual aspects of their reading.

A more elaborate account of method, research problems and results is provided elsewhere.⁵¹ However, as in much hermeneutic phenomenological research, data in the study include information gathered from participants through interviews, from results of other researchers, as well as the researcher's personal reflections based on the very important first-person experience.⁵² Thus, in addition to being built on findings from the interviews, the following description of reading is to some extent influenced by my own experience and to a large extent by the experience of other researchers, many of which are mentioned above.

Aspects of scholarly reading

Concepts

In order to trace and describe corporal and material aspects of reading, both data and concepts are needed. In this study, the concepts used to analyse reading are not entirely theoretical, but rather moulded and shaped by empirical findings in an ongoing hermeneutical process. The aim of the study is not to provide statistical evidence on academic reading behaviour, but to disclose and analytically describe basic features of this reading. Hopefully, the discussion will clarify some of the many fuzzy concepts characterising the debate on digital reading. What do we actually mean by deep reading, sustained or discontinuous reading?

The discontinuousness of academic reading

It is a fact that text is laid out in space and read in time, and that text always deals with some kind of subject matter. On the basis of these fundamental characteristics, reading can be described by degrees of con-

tinuousness, including temporal and spatial continuousness, and thematic connectedness.

Time spent on an act of reading varies dramatically, from very short acts of reading text messages or e-mails to lengthy periods of reading magazines and books, which can go on – more or less undisturbed – for hours. There are obviously all kinds of reading spans in between, but usually the uninterrupted reading of a magazine or journal article is considered a long read and an example of continuous reading – in a temporal sense. Reading that is repeatedly interrupted by other engagements is thus discontinuous.

However, the concepts of continuous and discontinuous reading are ambiguous and often used in a spatial rather than temporal meaning. Spatial continuous reading is reading that follows the linear and sequential order in which a text is presented, such as the normal way of reading a novel. Usually it starts at the beginning, and the reading has to cover a certain amount of text (last for a minimum of time) to be labelled continuous. Discontinuous reading is reading out of order, so to speak, in which the reader jumps back and forth in a text, reading parts of a book or an article without following the linear and sequential order of its presentation. Reading a succession of very short texts or parts of texts is also a form of discontinuous reading, common when browsing the Web or scanning printed newspapers. Discontinuous reading of multiple texts of disconnected subject matter is called fragmented reading.

A combination of discontinuous and continuous reading is frequent when flicking through newspapers or browsing the Web, intermittently slowing down to continuously read an article or two. Combined, as well as purely discontinuous reading, can go on for a long time and, in a temporal sense, be continuous. However, to call the same reading session both continuous and discontinuous is awkward; in the following I will call all kinds of lengthy acts of reading sustained reading.

Sustained discontinuous reading seems to be characteristic of scholarly reading. One of the participants in the study, “George,” says that when he receives presumably interesting material, such as a printed book, he starts by carefully reading the table of contents and then part of the introduction to see if the book interests him. If it does, he flicks through the pages scanning for keywords and skimming small bits of text, trying to get a better feeling of the overall structure and style. He then studies the bibliography and, using the index, he finds two or three places that seem particularly promising. If these passages are of interest, he reads the adjacent subchapters or chapters, and usually this is what he reads in

a book. However, sometimes he expands the reading and, in rare cases, he reads the whole book.

“Jane” tells a very similar story, only that when she starts reading at a promising spot, she takes notes on a piece of paper. Using subtitles, she makes a rough outline of the chapter in question, and when reading she takes notes in order to discern the line of argument and the ways of combining ideas. However, both “George” and “Jane” are willing to abandon reading at every point in the process. “Jane” says she stops reading if the text is too unfamiliar or difficult, or if it in any way is irrelevant for her present research. However, if she feels compelled to do so, she sometimes intensifies the note-taking, using all kinds of resources in her office to complement the reading; reference books, scholarly works, and online resources.

Stories of article reading are not dissimilar to those of reading books. “Carl” says he usually skips the abstract of an article and goes straight to the introduction, reading some paragraphs to get an idea of the research problem and theoretical standpoints. If it is interesting, he jumps to the conclusion and then he studies the empirical findings, flicking back and forth. This, he says, gives him enough understanding and usually renders reading the rest of the article unnecessary, at least for the moment. However, some articles he starts re-reading from beginning to end, always underlining and making notes. Sometimes he reads an article or a book several times, and “Susan,” during her interview, showed a heavily underlined and annotated paper she had read several times in order to use theoretical points in her writing. However, the annotations stopped mid-way in the article, and asking her why, she said that the last part presented empirical findings, which did not interest her. Asking her if she had read it, she said no.

Discontinuous and notoriously treacherous reading thus constitutes a distinctive scholarly reading characteristic. Yet, as indicated, continuous reading of book chapters and articles is not uncommon as part of the overall study of a subject. But continuous reading of whole books also occurs, often as complementary reading on subjects adjacent to the main focus of interest. “Carl” speaks of “scholarly reading for pleasure” which is continuous reading of non-fiction books rather similar to the continuous reading of novels. In the study, the participants unanimously said that they read novels continuously, if not always to the end.

All the reading commented on so far has been reading of printed books or print-outs. While research confirms that reading of scholarly articles and books is still dominated by paper,⁵³ some research may indicate that sustained screen reading is increasing.⁵⁴ Several of the participants in

the study reported that they sometimes read articles onscreen and even online, although their onscreen reading is not particularly conscientious and is usually for the sake of overview, typically without note-taking and often terminated before the end. “George” commented that the more relevant the article, the higher its chances of being read on paper.

Online scholarly reading is definitely discontinuous and often fragmented in character. Participants reported that they often browse the Web and scanned pages in search of information and updating, some frequently re-visiting favourite sites and blogs. Some also browse the cultural sections of online newspapers such as *Le Monde* or the *New York Times*. Occasionally, browsing leads to continuous reading of one or two articles, and more or less always to the following of links in a process that, according to one participant, “can go on forever.” When following links in a random way, informers often get led astray; but then fortuitous surfing also leads to discoveries of new and interesting sites.

The Web is also habitually used to search for specific authors or subjects, using search facilities such as Google, Google Scholar and sometimes databases offered by the university, mostly the national library search catalogue and international portals, such as Ebsco and FirstSearch. Participants also use online dictionaries and encyclopaedias, such as the Oxford English Dictionary and Wikipedia. The Web is sometimes used in collecting data from statistic databases, manuscript databases and newspaper collections. Obviously, the way these resources are used varies considerably among the participants. The point here is that all these scholarly uses of the Web are associated with discontinuous reading aimed at finding, scanning and delivering text.

Immersive reading

When analysing the power browsing behaviour of academics in digital libraries, the UCL researchers characterise the reading as shallow, presumably in contrast to a more deep involvement with text.⁵⁵ The shallow-deep contrast, however, is not always a very clarifying dichotomy. When “Adam” searches for full-text literature in journal portals and gets interesting hits, he examines many potentially relevant articles, skimming abstracts, looking at keywords, studying reference lists and reading introductory parts of the articles. Many articles are discarded as irrelevant. Some are downloaded and stored. Of these, a few are printed out on paper and read, either immediately or at a later stage. Adam’s way of reading online may seem superficial, but is actually the first step in an exhaustive in-depth reading of selected articles.

Even if it is sometimes difficult to determine what acts of reading are shallow or deep, it is nevertheless a common experience that reading does have different levels of involvement, indicated by the many words for reading: to leaf, flick or thumb through, to look over, to browse and skim, to study, scrutinise and peruse, to decipher and interpret. Marie-Laure Ryan has examined the many metaphors used in describing engaged reading, usually associated with movement, saturation and depth, often in combinations.⁵⁶ We talk of being carried away or lost in reading, being deeply involved or immersed, deeply absorbed or engrossed in reading. Based on Ryan and participant stories, it would seem that engaged humanist scholarly reading can be described as comprising two major types: imaginary and reflected reading.

In immersive imaginary reading, readers get involved in a story, conjuring up vivid images of persons and places; living through situations, empathising with characters. Readers are carried away into imagined worlds, anxious to know what is going to happen. Despite being absorbed in a book, “Carl” claims that simultaneously he has an eye and an ear for literary qualities. For him, becoming immersed in reading is emotionally satisfying, and “Adam” said that he consciously uses novel reading as a form of escape. “Silvio,” a literary critic, claimed that absorbed literary reading is an ultimate goal he seldom experiences, due to the analytic and evaluative character of his interpretations. Immersive imaginary reading is often associated with narratives and requires reading to be fairly fluent.

In immersive reflective reading, readers get involved in argumentative texts, eager to understand, interpret and learn, to see connections and consequences, and to widen their understanding. “Carl” compares the satisfaction offered by reflective and imaginary immersion, saying both experiences produce a place where he wants to be: in absorbed reflective reading he is away in a theoretical world, but recognisable insights makes him feel at home. “Jane” says that to be really engaged in a scholarly text, it has to be familiar, yet challenging. As with imaginary immersion, reflective immersion requires reading to be fluent. As indicated by neurologists, fluency has automated basic reading operations in the brain, adding time for inferences, thus facilitating thinking that goes beyond the text.⁵⁷

All readers feel that text involvement is relative, that immersion is a matter of degree. Sometimes readers get tired or hungry and drop out; at other times readers find a text difficult, uninteresting or downright boring. Moreover, the imaginary-reflective distinction is not particularly clear. Rather than inducing imaginary immersion, much fictional

literature, both classical and modern, requires a high degree of reflection in order to make sense and arouse interest. Regardless of this, both imaginary and reflective engagement is sometimes called hermeneutic immersion, a term based on Don Ihde's concept of a hermeneutic relation between users and technology, in this case between reader and book.⁵⁸ According to Ihde, there are three basic human-technology relations.⁵⁹ In embodiment relations, the tool functions as an extension of body and senses, such as an axe or a pair of glasses. In a hermeneutic relation, the instrument tells the user of conditions somewhere else, such as a map or a book describing the coming of *Homo sapiens*. In the hermeneutic relation, attention is towards technology as an object of perception; however, through interpretation of signs, the user becomes conscious of circumstances elsewhere, and this virtual world becomes the primary focus of attention. In alterity relations, the user's focus is very much on technology itself, often as an object of fascination.

A requirement for hermeneutic immersion is that the technology offers minimal disturbances on the part of the user; that it becomes more or less transparent. Over the centuries, typographers have refined the design of books, making the printed book an effective reading technology. In books, facing pages are important composing units, and typographers' knowledge of fonts, lines, whitespace and margins has been extremely beneficial for readers. Many books and printed publications are very well suited both for continuous and discontinuous reading, and for imaginary as well as reflective immersion.

Over the last decades, much traditional typographical knowledge has been transferred into word processors and desktop publishing software, making pre-press production of printed publications digital. Software applications also make it relatively easy for laymen to produce highly readable print-outs. Dedicated reading software, such as "Mobipocket" and "Microsoft Reader," also utilises time-honoured typographical features to some extent. The ordinary computer screen, however, is far less suited than paper to creating optimal reading conditions. This is partly due to technological limitations, such as low resolution, tiring backlight, widescreen formats and the stationary position of the screen, which in sum makes reading tiring. In addition, most reading software is designed with toolbars, side panels and icons, and the applications are often placed within the interface of a web browser or an operating system, with their own toolbars and icons. Thus, with numerous eye-catching elements within the visual field, all offering actions to the user, there is a considerable potential for fluency disruptions intruding on text immersion. In the

study, “George” said he frequently reads e-books using Adobe Reader, however, when asked further, it boiled down to reading of two classics on the laptop at home in his sofa, an implement he held to be quite all right. Without actually being able to explain why, participants in the study generally complained at the laborious experience of sustained on-screen reading, even with specialised software.

Quite different then, is the new generation of devices based on electronic paper, such as the Amazon Kindle and Sony Reader, specifically designed for reading. Even if they introduce new obstacles, such as slow paging, in these devices backlight, resolution and immobility are no longer issues. Weight, format and typography make them well suited for continuous reading of fiction and non-fiction. “Adam,” who owns an Amazon Kindle and a Cybook and reads e-books on a daily basis, claims that electronic paper is very good for reading novels. After a very short time, he claims, the device becomes transparent and the story springs forth. His enthusiasm notwithstanding, “Adam” describes at length the lack of annotation possibilities and poor navigational aids, arguing strongly that e-paper devices, as yet, are not suited for the exhaustive ways of reading in study and research.

In the study, “Adam” was the only participant with e-paper experience, but his account is in close accordance with my own experience, and also in line with research. Ergonomic evaluations reveal that, from a technical point of view, in readability, current e-paper displays are good enough to compete with ordinary office paper; they only need slightly greater illumination.⁶⁰ When the groups in a Texas A&M study had used the Amazon Kindle for a month, around half the participants were still, to various degrees, conscious of the physical device when reading; some were distracted by the clicking of the next page button and the time lag in paging. The other half had become accustomed to the device, saying that it eventually faded into the background. Nevertheless, many participants wished for improved navigation and better bookmarking and annotation features.⁶¹ A Princeton report⁶² and several other studies confirm that students regard the Kindle device as better suited for leisure reading than reading for study and research.⁶³

Among digital formats, Web browsers are probably the least suited for immersive reading. As pointed to, the Web has its strength in reading activities related to searching and browsing, uses of dictionaries, encyclopaedias and databases. The Internet and the Web are unrivalled in disseminating and accessing information and have facilitated a range of new communicative forms such as e-mail, discussion groups, chat

and net communities, all of which entail reading, usually in a rather fragmented form. In addition, of course, the Web is effective in disseminating and sharing music, videos and movies. According to Ryen and Mangen,⁶⁴ computers have their own specific form of immersion. It is very common to be caught up in front of the computer screen, deeply engrossed in browsing, chatting, socialising or gaming, as some participants reported. Nevertheless, online immersion is very different from the hermeneutic immersion of reading. In imaginary and reflective reading, the text is fixed and the signs arbitrary; meaning is created by internal processes in the user's mind. By contrast, online immersion is the result of external stimuli and the user's manipulation of the computer (using mouse, keyboard or joystick) in response to the flow of pictures, animations, videos, and text snippets. Much of the fascination lies in the ability to affect the outcome of the "game."

Obviously, computer technology has facilitated a wealth of possibilities in the calculation, manipulation and presentation of data, in receiving immediate input and in cyber socialising. However, technological features that are powerful in some domains may be weaknesses in others. As I will examine further, from the point of view of reading it seems that hypertext, multimodality and the multi-functionality of the Web are incompatible with sustained reflective and imaginary reading, and thus with hermeneutic immersion.

Multimodal reading, hypertext and activity: the urge to click

Multimodality is not a new phenomenon. Illuminated manuscripts and illustrated books have a long history and, as Kress and van Leeuwen point out, the use of graphs, diagrams, maps, models, drawings and photographs often increases the informational and aesthetic value of print publications.⁶⁵ In addition, a heavily illustrated magazine or textbook offers the user several choices. The reader can look at pictures and the accompanying captions and titles and form a good idea of what the article is about. Parallel to this, the background information and explanations of the main text can be read to get the full story. Either way, due to inclinations in our perception and the salience of pictures, the eyes will jump back and forth between text and illustrations. Direct visual perceptions will complement or replace the mental images usually produced during reading. In a spatial sense, strictly verbal reading will thus be discontinuous. Multimodal reading, on the other hand, will in a temporal sense go on uninterrupted; the reader will construe visual-verbal meaning units not reducible to any of the two modalities. As the

use of illustrations increases, however, a visual logic will eventually take precedence and dominate, as is the case with many modern magazines and text-books. In publications of this kind, verbal text plays an auxiliary or reciprocal role, anchoring and contextualising pictures. For readers, the meaning is derived from self-sufficient visual-verbal entities dominated by images, and the process of reading inevitably changes as the reader starts looking and flicking.

By contrast, non-illustrated printed texts offer no option but to read. Moreover, many text genres imply an intended reading order. In the eyes of hypertext theoreticians, especially Landow, these features represent limitations and impose constraints on the reader, keeping the author in control.⁶⁶ However, with particular reference to reading of scholarly articles and books, this study shows the opposite to be true. In academic genres, the strict structural order of texts seems to offer readers a high degree of freedom in choosing their own reading paths: Always indicating where you are, a tight structure allows for jumping, skimming and discontinuous reading, still making good sense of the text. On the other hand, hypertext theoreticians obviously have a point; hypertext literature and the Web do offer the user a wealth of opportunities. Yet, from the point of view of sustained reading, this hypertext freedom seems to come at a high price.

As David Miall and Theresa Dobson report, evidence from empirical studies suggests that certain aspects of hypertext, such as links and image hotspots, may disrupt reading.⁶⁷ In a study of readers who read either a simulated literary hypertext or the same text in linear form, they found a range of significant differences suggesting that “hypertext discourages the absorbed and reflective mode that characterises literary reading.” The Web, since its introduction in the early 1990s, has developed from being a system for linking plain text documents, to be an exceedingly multimodal hypertext, including graphics, pictures, sound, video and animations, as well as interactive and communicative features, such as games and chat, and net societies, more or less combining it all. Needless to say, in this highly audiovisual and communicative universe, solitary sustained reading of long-form text is on the defensive. On the Web, most users are tuned in to everything but reading.

Even rather straightforward Web pages have moved far away from time-tested typographical principles applied in publications meant for sustained reading. When “Adam” illustrates how he reads a Norwegian research site, he at first claims to be very good at focusing his attention on the content section of the page, turning a blind eye to the surrounding

columns with their ads and links. However, when asked why he initially scrolled down a bit, he said it was to get away from the top advertisements and the title head, both being a nuisance. And while he is at it, he says that he sometimes scrolls sideways to get away from the vertical ads, and he ends up with a long anti-ad harangue. When referring to one of the ads depicting Charles Darwin in one of the right columns and asking if he had noticed it, he said that it been there for weeks and that he had actually clicked on it a couple of times to see what it was.

In the interview, “Adam” continued by browsing one of the quality Norwegian online newspapers. When asked to say what he was looking at, he said he went straight for the titles and text and was less interested in the pictures, which he hardly noticed. However, when changing to *The Guardian*, there was a long silence as he studied the front Web page. “Here it takes more time,” he said. “The news area is plain text and I have to start reading to detect what the articles are about. *The Guardian* is different from the Norwegian papers in which the pictures give a pretty good idea of what the stories are all about,” he said. When asked if he actually did look at the pictures, he said: “Well, yes, I suppose I do.”

As mentioned earlier, a Web page is presented within a browser within an operating system, all with their tool bars and icons. The pages usually have headers and footers and several columns and both graphics and pictures. Salient visual elements necessarily make the reading multimodal and qualitatively different from linear reading of plain text. In addition, pictures and animations are often part of the linking system of the Web pages, making stirring hotspots adding to the psychological and bodily urge to click, as described by Mangen.⁶⁸

As for the urge to click, “Adam’s” story indicates that it is psychologically very hard to fight off distractions and alluring links even when reading interesting subject matter. Still on *The Guardian* site, he illustrates how he sometimes barely skims the first part of an article and only starts reading seriously when he has scrolled past the links and ads to the page area where text is the sole element. However, at this point in the interview he picks up his iPhone, saying that he actually prefers to read lengthy news articles and comments on that device. Even if the display is rather small, it is clear and when reading, text is the only thing occupying the visual area – there are no distractions. Unsolicited, he enthusiastically starts relating how he can sit at home, feet high, enjoying reading the long, informative articles in *New York Times* on his iPhone. He also says he reads e-books on the iPhone, and that e-books on iPhone and Amazon Kindle are somehow synchronised. Related to Mangen’s reasoning on

text materiality and reading, “Adam’s” stories suggest that it is not the intangibility of digital text as such that prevents hermeneutic immersion, but rather the prevalent hypertextual and multimodal manner of its presentation.

Adam’s story about home reading also points to a very important contextual factor. For academics, the computer is a multi-functional working tool and, as some participants pointed to, extremely demanding. Participants said they often felt obliged to start the day checking their e-mails, answering some of the messages or responding to others, for instance by starting to prepare an upcoming meeting. While checking, one might also open the learning management system to see if more students have submitted their obligatory papers and then suddenly remember those unfinished Power Points. Several participants had data or texts that needed further analysis – and then there is the writing. All participants were working on at least one unfinished manuscript for a journal article or a book. Thus, with its strategic location, the computer is a constant reminder of things undone. Taken together, all these factors – ergonomics, multi-functionality, multimodality and hypertext – make it extremely hard to maintain an act of digital reading for a substantial period of time.

Bodily aspects of reading

In the study, the participants were very conscious of the obligations and allurements of the computer and, preferring paper, all had in different ways developed strategies to avoid being distracted or tempted by the screen while reading, usually positioning their body so as not to stare directly into the beckoning display. Some participants simply turned their back on the computer, using another part of the desk. “Carl” had cleared a well-lit corner of his office couch, and “Eric” said he sometimes found a quiet spot in the canteen to get things read. All said they often read at home.

While reading, the participants use their hands very actively to hold the book or print-out in the visual focal area, flicking back and forth in a discontinuous way of reading, as previously described. In addition, especially with print-outs, the participants hold a pen, pencil or highlighter in their hand. Using rather different systems, they underline, highlight and make carets or exclamation marks, lines or squiggles, notes or comments, in the margins or around the text. “Carl” said he felt uneasy without a pencil in his hand, and “Susan” said she always operated a highlighter, using it like a weapon to help her concentrate and hunt for important pas-

sages. Among the participant, several said that the use of hands, fingers and pen or pencil was an indispensable part of their scholarly reading.

The annotation and marking tools were mainly used by the participants for two interrelated purposes: to improve comprehension of the text and to render visible relevant connections to their own writing. The highlighting and note-taking helps slow down the pace of reading and often leads to re-reading of passages and recording of points. From the point of view of cognitive psychology, the annotation habit is probably a way of processing information, giving it time to fit into schemas in the long-term memory and also providing time and space for reflection and the discovery of inferences. In addition to using words describing comprehension, such as understanding, insight and cohesion, the participants ended up saying that they very often related the underlining and annotating to their writing. Annotation is accordingly described as an aid to help find again important points or citations for use in their own articles and books. “Eric” is probably the participant most conscious of this instrumental aspect of reading, saying he never reads anything without connecting it to coming lectures or writings. For him, the obligatory annotation tool is used not so much for underling as for comment and relating observations in the text to thoughts of his own. “Eric” says that the most stimulating texts are those shedding new light on his own research problems and thus reaching beyond the text at hand.

As noted earlier, reading is a bodily and mental process evolving in time. When “Carl” meditates on the pleasures of reading, he recognises glimpses of insight as inherent in the process of reading, happening while reading and then quickly fading. To him, underlining and note-taking is a somewhat futile attempt to externalise these insights, making them visible and lasting, an endeavour which is only completed when his own writing is successful. “Carl” also tells an anecdote that lucidly illustrates the relations between body, materiality of text and ways of reading, especially through his use of metaphors. In the story he tries to uncover why he prefers paper to computer when engaging in an act of reading. Sometimes, he said, when working away from printers at his cottage, he urgently needs to read an article on his laptop. At one level this is unproblematic; he can read through the text, understanding every sentence and paragraph. At the same time, he often gets a feeling of not getting fully to grips with the text, that he somehow loses oversight and is unable to fence the article in, which is frustrating. When asked to expand on these utterances, he says that in order to see connections and make inferences when reading, he often needs to have several text passages or

ideas present simultaneously, and in printed versions the passages are physically there on the sheets of paper. He can flick back and forth, stick his finger in between sheets and keep several ideas in his memory at the same time: comparing, relating and thinking. On the laptop he is not able to do this; on screen, the text and the author slip through his fingers.

In the study, all participants reported digital scholarly reading to be more superficial than paper reading, with more skimming and less annotating. In paper reading, the participants very actively utilise the tangibility and physical shape of books and print-outs. The bundle of sheets, the solidified text and the over-writable paper make it easy to flick back and forth and use a pen in the typical discontinuous and annotating way of concentrated scholarly reading. By contrast, the digital nature of text and the use of the mouse and keyboard make computers very well suited for browsing, searching, accessing, downloading and skimming text, which is also what participants use the computer for. The participants' answers and "Carl's" story make a very good case for Mangen's claim that the intangibility and volatility of digital text make it hard to combine with immersive reading, at least with the scholarly form of immersive reflective reading.

Writing in reading

The briefly mentioned relationship between reading and writing points to another important dimension of reading: the degree to which reading is integrated in more comprehensive tasks. Sometimes, of course, reading is a stand-alone activity, done for its own sake, such as leisurely reading of novels. Often, however, reading is part of a wider activity, such as filling in a form, receiving an instruction or doing all kinds of administrative work. For scholars, reading text is an integral part of their work; they read to prepare lessons, to evaluate, to give assessments, to review, to correct and to comment – all activities with their own specific way of relating to the text. In the participants' research, texts are mainly read for three purposes: texts are the object of analysis, such as works of philosophy and literature; documents are the source of information and data, as in history and linguistics; and journal articles and monographs present theories and data relevant for the scholar's own research and writing.

An example of the first is "Johanna," who is writing a monograph on a minimalist author, studying the intertextuality in several novels. Even though she knew the author well, during research she repeatedly re-read the novels, first in a rather continuous way, mostly underlining, then more discontinuously and analytically, hunting for intertextual refer-

ences, citations and allusions, extensively taking notes and making comments. During the third re-read, she also followed up on the allusions and references, reading the referenced works along with scholarly literature on their authors, again taking notes and comparing her own findings with those of other researchers. All notes are handwritten; however, the specific writing of “Johanna’s” book is done in Microsoft Word, and the notes, being spread around the computer, are constantly being consulted in the process of writing, the notes in fact being the groundwork upon which the monograph is built.

“Silvio” illustrates the second purpose. When “Silvio” was writing his author biography he travelled to libraries all over Scandinavia and Germany reading documents, old letters, journals and microfilm newspapers, transcribing, taking notes and sometimes (when allowed) copying documents. When circumstances forced him to write by hand, the notes were later entered into his laptop. These collected notes and data, the result of his reading, were indispensable in his later writing.

The third purpose is exemplified by the following. As pointed out, participants relate most of their academic reading directly to their writing and in this respect, the method of “George” is revealing. When writing a document, he creates a parallel document containing the outline of his work. When reading scholarly books and journal articles, he starts by underlining and making annotations in the margins and then later writes comments, citations and references into his parallel document for potential use in his own writing process.

Writing and the activity of research are thus very much ingrained in scholarly reading. Relevance for writing is what scholars look for when searching for literature, it decides what they download and it determines whether the text is actually printed out and studied. Their own research questions form the perspective used in interpreting and extracting meaning from the text, and future writing, finally, is what guides the hand, fingers and pen in highlighting and annotating while reading. In the end, the notes are used when writing.

Reading in writing

As writing is part of reading, so is reading part of writing. In the study, all participants were writing on a computer using a keyboard and screen, often producing text from notes and sometimes from handwritten rough drafts. When writing, the participants were all reading, yet in different ways. “Johanna” said she was reading the text as she was writing, more or less simultaneously, correcting errors and usage on screen as she went

along. “Silvio,” on the other hand, was also simultaneously reading what he was writing, but he could write several paragraphs or subchapters before going back, re-reading and correcting the text.

In the initial phase, while still actively writing, participants’ reading and correcting are done on screen. At some point, however, the text is always printed out on paper. “Silvio” said that he once sent a whole chapter to his publisher without having it locally printed at all. However, this was an exception, and “Silvio,” who assiduously worked on the computer, frequently used paper when correcting and editing his own text. However, the print-out frequency varied considerably among the participants, some printing out their text constantly, while others wrote several subchapters before printing. In the subsequent reading and editing process, they made active use of a pen or pencil to correct misspellings, change words and survey the structure of the text. Sometimes this editing leads to major changes, and often it leads to discoveries of deficiencies in their own text, which at times leads to very purposeful reading of scholarly literature on the subject.

During writing, a piece of text is thus read many times. Adding to the number is the fact that other people also read the unfinished text: colleagues, editors, peer reviewers and proofreaders. In fact, a great deal of scholars’ reading is of unfinished and unpublished texts, both their own and others’. When asked how many times a randomly chosen sentence in their finished book had been read before publishing, the usual participant reaction was laughter followed by exclamations saying it is hard to tell, but it is certainly many.

Paper, pen and computer

Thus, reading is inside writing, and in much the same way printed paper resides within the computer. In this study, it has become clear that humanist scholars combine rather different text technologies in their literary activities. At the core of the activity is reading, and when reading in the reflective mode, the participants predominantly use printed paper in the form of books and print-outs. When writing, they use software designed to facilitate paper text production, such as Microsoft Word, thus writing with paper genres and paper reading in mind. As an outer shell, the computer system is used for the definitive writing of text, and for storing, disseminating, accessing, skimming and downloading of text. In addition, the computer provides a number of scholarly aids, such as dictionaries and encyclopaedias. Today, one can safely say that humanist scholars have some fingers in a paper based text cycle and the rest

in a digital text cycle. Moreover, by focusing on the physical aspects of reading, the study reveals that indeed three historical systems are in use in scholarly literacy events: the modern computer system, traditional printing and the ancient system of handwriting and annotation.

Discussion

The ongoing transformation of the long-standing written culture has many sides to it, and Chartier warns that “the transfer of a written heritage from one medium to another, from the codex to the screen, would create immeasurable possibilities, but it would also do violence to the texts by separating them from the original physical forms in which they appeared and which helped to constitute their historical significance.”⁶⁹ Clearly, changes have taken place throughout history, and written discourse is far from static. Editorial philology illustrates how supposedly stable written texts are by no means fixed entities, neither in the manuscript tradition nor within the print culture. Texts change and, as the history of the book reveals, so do their material and typographical representation. Moreover, book history studies give evidence of great variations in reading practices over time.

Nonetheless, since the introduction of the codex, all these changes have taken place without altering the fundamental structure of the book, with its bound collection of written or printed sheets of parchment or paper. In contrast to previous modifications, the digital transformation fundamentally changes the physical form of text. In computers, the written text is no longer physically tied to the surface of a medium that simultaneously stores and represents the text. In the computer system, storing and representation are separated making digital text extremely malleable, moveable, and – through systems of links and search facilities – globally accessible and essentially borderless.⁷⁰ According to Chartier, such new material features will inevitably and imperatively require new ways of reading, new relationships to the written word and new intellectual techniques.⁷¹

As for intellectual techniques, digital characteristics open up new ways of studying classical as well as contemporary scientific, philosophical and literary texts. In the digital humanities, tools have been developed for text corpora search, collation and collocation, comparison of text, study of genealogy, as well as statistical tools for frequency and distribution studies. Even if the potential of digital analytic tools is generally underexploited,

average humanist scholars have indeed changed their reading and working habits. Obviously, the current study does not examine variation over time. However, personal computers and Web technologies are relatively new, and the study shows that scholars habitually use the Web for searching and accessing literature, and that the computer is used for writing, storing and dissemination of text. Furthermore, Carol Tenopir shows that the average number of journal articles read by scholars has risen over the last decennia, whereas time spent on each article has decreased, which might indicate that discontinuous reading has increased.⁷² What has not changed significantly are the material and corporal aspects of sustained reading of long-form texts, which is still done on paper actively using hands and fingers, and sometimes a pen.

From the point of view of tradition, these enduring reading habits may be reassuring, as they secure ties to long-lasting written discourses in literature, philosophy and science, in which complex narratives, detailed descriptions and long arguments are important. However, from the point of view of cultural movement, in which written discourse is increasingly dominated by computers and networks, and probably more so among the young, the dependency on paper may be alarming, and before long a critical question could be how to successfully transfer sustained imaginary and reflective reading on to a digital platform.

For digital humanities these issues are important for two reasons – the first being the limited use of critical digital resources, a situation that some researchers argue can only be improved by designing applications that resemble the studying and reading habits of average humanist scholars. The second issue is more profound and relates to questions of cultural heritage. For digital humanists, a main goal is to bring authenticated text from the cultural heritage on to a digital platform, both to provide wide dissemination of high quality texts and to facilitate analytic activity. New digital tools are already being used in studies of texts and language, their origin, development and meaning. However, if these new analytic tools are accompanied by radical new ways of reading and interpreting text, the studied texts are no longer the same, as Chartier has pointed out.⁷³ Without being able to transfer basic characteristics of traditional reading into the new technologies of text, digital scholars may end up studying a significantly modified cultural heritage, remote from the long-standing tradition of written and printed discourse.

Conclusion

With a focus on text materiality and physical aspects of reading, the current study suggests there are two major obstacles and challenges to long-form text transference. The first is to replicate conditions for continuous reading and imaginary immersion, and the second to create favourable conditions for sustained discontinuous and reflective reading. Whatever the solutions, digital texts will under no circumstances be the same as printed texts and, in relation to reading and studying experiences, it will never be more than a question of proximity. The following reflections take as a premise the use of transformable text formats (such as XML) in transcriptional and editorial practice. Utilising the separation of storage and representation in computer technology, transformable text formats make possible multiple and variable presentations of texts, both on displays and on paper.

With regard to the first challenge of continuous reading, it seems clear that the stationary displays of PCs and laptops are unfit for imaginary immersive reading, for reasons thoroughly accounted for in the study. However, as indicated, there seems to be a relatively easy solution to this particular challenge that nevertheless would require a radical change of attitude for many scholars. Handheld devices, especially dedicated e-readers, seem to be able to fairly well approximate reading experiences of printed books, such as novels, and at the moment e-paper devices seem the most promising. Such devices fit easily into the hand and let users position the body for reading. While the user cannot flick through the pages in the ordinary way, the devices engage the fingers in paging by clicking buttons. They are generally highly readable, easy on the eye, and some devices indicate where the readers are within the overall text. Thus, current e-paper devices create good conditions for transparency and provide an efficient hermeneutic relation between user and technology. In the future, if the navigational and annotating capacities are enhanced, e-paper may prove a viable alternative for reflective study of text. Already, e-paper ought to be considered an alternative by digital humanist sites delivering long-form texts for continuous reading.

As discussed in this paper, reading is influenced by the design of current multipurpose personal computers, with separated displays and keyboards, mouse, widescreens and backlight. Inevitably, computer design will change over time, as shown by experiments with tablets, micro-computers, book-like two-screen solutions, virtual keyboards, touch screens and multimedia reading devices, such as the iPad. New designs

all introduce novel ways of using hands, fingers and eyes, and how these designs affect reading would be interesting issues in future research. However, it is a fact that people can sit in front of computer screens for hours, engaged in all kinds of tasks. In principle, therefore, it should not be impossible to perform onscreen reading activities for substantial periods of time. This study suggests that it is not the intangibility of digital text as such that impedes reading, but rather the distractive way text is usually presented.

The second challenge, to create good conditions for reflective reading, is demanding and will require considerable intellectual and technical ingenuity. Humanist study is often a combination of continuous and discontinuous reading and – as shown in the study – discontinuous ways of reading involve very active use of hands in flicking, underlining and annotating, all within the physical unity of a text. Nor for this purpose is sole reliance on Web browsers fruitful. In the following, based on experiences in the study, I will tentatively present ways of meeting the challenge.

First of all, humanist Web sites should use Web browsers for what they are good at: presenting overviews and accessing information through links and search facilities. They may also be used for highly visual and multimodal presentations, which is important in many branches of the digital humanities. For the rest, Web browsers should be a deliverer of adapted applications, in a layered sandwich structure. In a layered structure, browsers may be the interface for analytic tools, such as collation, corpora search and the like; however, as soon as long-form text and sustained reading is required, readers should be able to use dedicated reading software. In this software, all the distracting visual elements of web browsers and operating systems should be cleared away and all unnecessary links and graphics removed. Drawing on typographical knowledge, in reading application the focus should be on reading, preferably in a two-layer arrangement with one mode for continuous reading, stripped of everything unnecessary, and one for discontinuous reading, with navigation, highlighting and annotating tools, if possible with links to word processors. In addition to delivering adapted segments of text to dedicated reading applications, it would be beneficial to offer compounds of text prepared for handheld reading devices, such as e-paper and touch screen readers. Versions for print-outs would also be preferable, and in some cases even providing for printed books.

With such a layered design, literary, philosophical or theological texts would be accessible in formats fit for their actual use, which is very often analytic, reflective or imaginary reading, preceding or entwined in ana-

lytic tasks. However, establishing the particular needs of specific groups requires further research. In such research, it would be informative to closely examine how tool-related ways of handling text affect text interpretation, memorisation and reflections on text, as well as the ways different implements integrate reading processes in the overall intellectual work flow. In my view, user accommodation of tools, particularly reading tools, ought to be a main research problem and development task in the digital humanities. Not only could such studies be instrumental in bringing more mainstream humanist scholars to the Web sites, it might even bring texts from the cultural heritage into the digital domain in a fashion that preserves some of the links to the very long tradition of written and printed discourse.

References

- 1 J. Bolter, *Writing Space: The Computer, Hypertext, and the History of Writing* (Hillsdale, N.J.: Lawrence Erlbaum, 1991); idem, *Writing Space: The Computer, Hypertext, and the History of Writing*, 2nd ed. (Mahwah, N.J.: Lawrence Erlbaum, 2001) & J. Bolter & R. Grusin, *Remediation: Understanding New Media* (Cambridge, Mass.: MIT Press, 1999).
- 2 G.P. Landow, *Hypertext: The Convergence of Contemporary Critical Theory and Technology* (Baltimore, Md.: Johns Hopkins University Press, 1992) & idem, *Hypertext 2.0* (Baltimore, Md.: Johns Hopkins University Press, 1997).
- 3 Claire Warwick, "Paper 2: Humanities scholars, research and reading, in physical and digital environments" abstract in *Book of Abstracts Digital Humanities 2008*, University of Oulu, June 24–29, p. 32.
<http://www.ekl.oulu.fi/dh2008/Digital%20Humanities%202008%20Book%20of%20Abstracts.pdf> accessed 11 August 2010.
- 4 S. Schreibman, S. Siemens & J. Unsworth, eds., *A Companion to Digital Humanities* (Oxford: Blackwell Publishing Ltd, 2004).
- 5 R. Siemens & S. Schreibman, *A Companion to Digital Literary Studies* (Malden, MA & Oxford: Blackwell Publishing Ltd, 2007).
- 6 Christine Borgman, *Scholarship in the digital age: information, infrastructure, and the Internet* (Cambridge, Massachusetts & London: The MIT Press, 2007), p. 33.
- 7 Peter Robinson, "Current issues in making digital editions of medieval texts—or, do electronic scholarly editions have a future?" *Digital Medievalist* 1.1 (Spring 2005).
<http://www.digitalmedievalist.org/journal/1.1/robinson/>, accessed 11 august 2010.
- 8 Patrick Juola, "Killer Applications in Digital Humanities," *Literary and Linguistic Computing* (2008) Vol. 23, No. 1.
- 9 John Bradley, "Pliny: A model for digital support of scholarship," *Journal of Digital Information* (2008) Vol. 9, Nr. 26.
- 10 Lisa R. Schiff, "Advancing Scholarship through Digital Critical Editions: Mark Twain Project," *Online Proceedings ELPUB 2008 Conference on Electronic Publishing* – Toronto, Canada – June 2008.
- 11 C. Warwick, M. Terras, P. Huntington and N. Pappa, "If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data" *Literary and Linguistic Computing* (2008) Vol. 23, No. 1; C. Warwick, I. Galina, M. Terras, P. Huntington and N. Pappa, "The Master Builders: LAIRAH Research on Good

- Practice in the Construction of Digital Humanities Projects,” *Literary and Linguistic Computing* (2008) Vol. 23, No.3; Bradley, “Pliny.”
- 12 Borgman, *Scholarship in the digital age*.
 - 13 Stephen B. Kucer, *Dimensions of Literacy: A Conceptual Base for Teaching Reading and Writing in School Settings*. 2nd ed. (Mahwah, New Jersey & London: Lawrence Erlbaum Associates, 2005).
 - 14 Maryanne Wolf, *Proust and the Squid: The Story and Science of the Reading Brain* (New York: HarperCollins, 2007).
 - 15 K. Rayner, “Understanding eye movements in reading,” *Scientific Studies of Reading* (1997) Vol. 1, No. 4.
 - 16 Maria dos Santos Lonsdale, Mary C. Dyson & Linda Reynolds, “Reading in examination-type situations: the effects of text layout on performance,” *Journal of Research in Reading* (2006) Vol. 29, Issue 4.
 - 17 David Barton, *Literacy: An introduction to the ecology of written language* (Oxford: Blackwell Publishing, 1994); Randal Holme, *Literacy: An introduction* (Edinburgh: Edinburgh University Press, 2004).
 - 18 Wolfgang Iser, *The Act of Reading: A Theory of Aesthetic Response* (London & Henley: Routledge & Kegan Paul, 1978); Jane P. Tompkins, *Reader-response criticism: From Formalism to Post-Structuralism* (Baltimore & London: The John Hopkins University Press, 1980).
 - 19 Bolter, *Writing Space 1st & 2nd eds.*; Landow, *Hypertext 2.0.*; idem, *Hypertext 3.0: Critical Theory and New Media in an Era of Globalization* (Baltimore, Md.: The Johns Hopkins University Press, 2006).
 - 20 G. Kress & T. Van Leeuwen, *Multimodal Discourse: The Modes and Media of Contemporary Communication*, (London: Routledge. 2001); idem, *Reading images: The grammar of visual design*. 2nd. ed. (London: Routledge, 2006); M. Walsh, “The ‘textual shift:’ Examining the reading process with print, visual and multimodal texts,” *Australian Journal of Language and Literacy* (2006) Vol. 29, Nr. 1; Radan Martinec & Theo van Leeuwen, *The Language of New Media Design: Theory and Practice* (London & New York: Routledge, 2009).
 - 21 Robert Darnton, *The Kiss of Lamourette: Reflections in Cultural History* (New York & London: W.W. Norton & Company, 1990); Roger Chartier, *Forms and meanings: Texts, performances, and audiences from codex to computer* (Philadelphia: University of Pennsylvania Press, 1995); idem, “The End of the Reign of the Book,” *SubStance* (1997) volume 26, number 1, issue 82 (Special Issue: Metamorphoses of the Book), pp. 9–11.
 - 22 Donald W. King, “Measuring Total Reading of Journal Articles,” *D-Lib Magazine* (2006) Vol. 12 Nr. 10

- <http://www.dlib.org/dlib/october06/king/1oking.html>, accessed 11 August 2010.
- 23 Andrew Dillon, "Reading from paper versus screens: a critical review of the empirical literature," *Ergonomics* (1992) Vol. 35, No. 10; Mary C. Dyson & Gary J. Kipping, "Exploring the Effect of Layout on Reading from Screen," in Roger D. Hersch, Jacques André & Heather Brown, *Electronic publishing, artistic imaging, and digital typography: 7th International Conference on Electronic Publishing: Proceedings* (Berlin: Springer, 1998); Mary C. Dyson, "How physical text layout affects reading from screen," *Behaviour & Information Technology* (2004) Vol. 23, No. 6; Anuj A. Nanavati & Randolph G. Bias, "Optimal Line Length in Reading: A Literature Review," *Visible Language* (2004) Vol. 39, No. 2; Lonsdale, Dyson & Reynolds.
 - 24 Ian Darroch, et al., "The Effect of Age and Font Size on Reading Text on Handheld Computers," in M.F. Costabile, & F. Patemnd, eds. *INTERACT 2005, LNCS 3585*.
 - 25 Der-Song Lee, et al., "Effect of character size and lighting on legibility of electronic papers" *Displays* 29 (2007); Dennis T. Clark, et al., "A qualitative assessment of the Kindle e-book reader: Results from initial focus groups," *Performance Measurement and Metrics* (2008) Vol. 9, Nr. 2.
 - 26 Eve Berne, et al., *Reading on screen*, The United Kingdom Literacy Association (2007); Jan M. Noyes & Kate J. Garland, "Computer- vs. paper-based tasks: Are they equivalent?" *Ergonomics* (2008) Vol. 51, No. 9.
 - 27 I. Rowlands, D. Nicholas, H.R. Jamali & P. Huntington, "What do faculty and students really think about e-books?" *Aslib Proceedings* (2007) Vol. 59 No. 6, pp. 489–511.
 - 28 Ziming Liu, "Reading behaviour in the digital environment: Changes in reading behaviour over the past ten years," *Journal of Documentation* (2005) Vol. 61, No. 6; D. Nicholas et al., "Viewing and reading behaviour in a virtual environment: The full-text download and what can be read into it," *Aslib Proceedings* (2008) Vol. 60 No. 3, pp. 185–98; idem et al., "UK scholarly e-book usage: a landmark survey," *Aslib Proceedings* (2008) Vol. 60 No. 4, pp. 311–334.
 - 29 Stephen B. Kucer, *Dimensions of Literacy: A Conceptual Base for Teaching Reading and Writing in School Settings*. 2nd ed. (Mahwah, New Jersey & London: Lawrence Erlbaum Associates, 2007).
 - 30 C. S. Ross, L. E. F. McKechnie & P. M. Rothbauer, *Reading matters: What the research reveals about reading, libraries, and community* (Libraries Unlimited: Westport, Connecticut & London, 2006).

- 31 Chartier, *Forms and meanings*.
- 32 Ibid.
- 33 Paul Saenger, *Space Between Words: The Origins of Silent Reading* (Stanford, California: Stanford University Press, 1997).
- 34 Chartier, *Forms and meanings*.
- 35 Darnton, *The Kiss of Lamourette*.
- 36 Alberto Manuel, *A History of Reading* (London: Harper Collins, 1996).
- 37 Mary A. Rouse & Richard H. Rouse, *Authentic Witness: Approaches to Medieval Texts and Manuscripts* (Notre Dame, Indiana: University of Notre Dame Press, 1991); Saenger, *Space Between Words*.
- 38 Peter Stallybrass, "Chapter 2: Books and Scrolls: Navigating the Bible," in Jenifer Andersen & Elisabeth Sauer, *Books and readers in early modern England: Material studies* (Philadelphia: University of Pennsylvania Press, 2002).
- 39 Stallybrass, p. 48.
- 40 Chartier, *Forms and meanings*.
- 41 Kress & van Leeuwen, *Multimodal Discourse*; idem, *Reading images*; G. Kress, *Literacy in the new media age* (London: Routledge 2003).
- 42 Landow, *Hypertext*; ibid., *Hypertext 2.0*; Bolter, *Writing Space*; idem, *Writing Space 2nd ed.*
- 43 Anne Mangen, "Hypertext Fiction Reading: Haptics and Immersion," *Journal of Research in Reading* (2008) Vol. 31, Issue 4.
- 44 Ibid.
- 45 Liu, "Reading behaviour in the digital environment."
- 46 Rowlands, et al., "What do faculty and students really think about e-books?"; Nicholas et al., "Viewing and reading behaviour in a virtual environment;" idem et al., "UK scholarly e-book usage."
- 47 Nicholas et al., "Viewing and reading behaviour in a virtual environment;" idem et al., "UK scholarly e-book usage."
- 48 Liu, "Reading behaviour in the digital environment."
- 49 Abigail Sellen & Richard Harper, "Paper as an Analytic Resource for the Design of New Technologies," *CHI 97 Electronic Publications: Papers* (New York: ACM, 1997); idem, *The myth of the paperless office* (Cambridge, Mass.: MIT Press, 2002); Kenton O'Hara & Abigail Sellen, "A Comparison of Reading Paper and on Line Documents," *CHI 97 Electronic Publications: Papers* ACM (1997).
- 50 Sellen & Harper, *The myth of the paperless office*.
- 51 Terje Hillesund, "Digital Reading Spaces: How Expert Readers Handle Books, the Web and Electronic Paper," *First Monday*, Volume 15, Number 4-5 April 2010.

- <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2762/2504>, accessed 11 August 2010.
- 52 Don Ihde, *Experimental Phenomenology* (State University of New York Press: Albany, 1986); Donald E. Polkinghorne, *Narrative Knowing and the Human Sciences* (Albany: State University of New York Press, 1998).
 - 53 Liu, "Reading behaviour in the digital environment;" Carol Tenopir, et al., "Electronic journals and changes in scholarly article seeking and reading patterns," *Aslib Proceedings: New Information Perspectives* (2009) Vol. 61, No. 1.
 - 54 Nicholas et al., "UK scholarly e-book usage."
 - 55 Rowlands, et al., "What do faculty and students really think about e-books?"; Nicholas et al., "Viewing and reading behaviour in a virtual environment;" idem et al., "UK scholarly e-book usage."
 - 56 Marie-Laure Ryan, *Narrative as Virtual Reality: Immersion and Interactivity in Literature and Electronic Media* (Baltimore & London: The John Hopkins University Press, 2001).
 - 57 Wolf, *Proust and the Squid*.
 - 58 Mangan, "Hypertext Fiction Reading."
 - 59 Ihde, *Experimental Phenomenology*.
 - 60 Lee, et al., "Effect of character size and lighting on legibility of electronic papers."
 - 61 Clark, et al., "A qualitative assessment of the Kindle e-book reader."
 - 62 Princeton University (2009): *The e-reader Pilot at Princeton*.
http://people.ischool.berkeley.edu/~janani/files/214_kindle_presentation_final.pdf accessed 11 August 2010.
 - 63 Jessica Santana et al., "Amazon Kindle Usability Assessment," *INFO 214 – Presentation*, UC Berkeley, School of Information (2008),
http://people.ischool.berkeley.edu/~janani/files/214_kindle_presentation_final.pdf, accessed 11 August 2010; M. Cristina Patuelli & Debbie Rabina, "Forms, effects, function: LIS students' attitudes towards portable e-book readers," *Aslib Proceedings* (2010) Vol. 62 (3), pp. 228–244.
 - 64 Ryan, *Narrative as Virtual Reality*; Mangan, "Hypertext Fiction Reading."
 - 65 Kress & van Leeuwen, *Multimodal Discourse*; idem, *Reading images* 2nd.
 - 66 Landow, *Hypertext* & idem, *Hypertext 2.0*.
 - 67 David S. Miall & Teresa Dobson, "Reading Hypertext and the Experience of Literature" *Journal of Digital Information*, (2001) Vol. 2, No. 1.

- 68 Manguerra, “Hypertext Fiction Reading.”
- 69 Chartier, *Forms and meanings*, p. 22.
- 70 Terje Hillesund, “Digital Text Cycles: From Medieval Manuscripts to Modern Markup,” *Journal of Digital Information* (2005) Vol. 6, Issue 1, <http://journals.tdl.org/jodi/article/view/62/65>, accessed 11 August 2010.
- 71 Chartier, *Forms and meanings*.
- 72 Tenopir, et al., “Electronic journals and changes in scholarly article seeking and reading patterns.”
- 73 Chartier, *Forms and meanings*.

DIGITISATION IN PRACTICE – AN ALM PERSPECTIVE

Magdalena Gram

ALM perspectives

The abbreviation ALM stands for Archives, Libraries and Museums and refers to an idea which came to mind about 15 years ago and related to co-operation and co-ordination between these sectors, above all within the field of digitisation. Being aware of the different traditions within knowledge organisation is in fact an excellent starting point for an ALM strategy. This strategy should aim at minimising the differences and describing how to make various resources accessible in a borderless virtual context.

Interoperability and seamless access to various databases demands that all units involved in the structure follow established standards. However, as the differences between traditions and commissions within archives, libraries and museums still influence the achievement of an ideal solution I would like to give you some examples of these differences. Photography and metadata are two relevant fields.

Photography

The various perspectives with respect to photography within archives, libraries and museums are notable in several contexts, for example in relation to collecting, organisation of knowledge and digitisation. At the same time, digital technology and digitisation have resulted in the insight that all sectors collect photographs and that co-operation is necessary to rationally achieve the goal of making photographic archives and collections virtually accessible.

Archives and libraries

While regional and local museum archives realised quite early on that photographs form an important part of the archives, for many years most central national archives have considered photographs to be a complex and low-priority material. In several respects, photographic material has also been hidden from scholars. Since a decade ago a change has been noticeable, and today photographs are viewed as an important source within several academic disciplines.

Libraries are generally not associated with photographic collections. The important exception is the national libraries, which collect photographs due to an earlier tradition of collecting graphic materials. In France, photography is part of the legal deposit system and in Denmark the national photo museum and the national library are integrated.

The National Library of Sweden has extensive photographic collections, in particular of documentary character. They are, however, not a result of the legal deposit law and they have primarily been used for reproductions. Photography as such has been regarded as something outside the real commission, which is collecting, preserving, describing and providing access to printed material. Today an investigation of the photographic collections is ongoing and the future outcome of this could result in a changed appraisal.

Photographers employed at national archives and libraries have traditionally been primarily working with reproductions and microfilming. Digital technology has changed this situation. Today professional photographers are engaged in producing materials intended for publication, while scanner operators are responsible for producing reproductions used as working copies and for mass digitisation.

Museums

For a long time many museums have had their own skilful photographers who have documented their unique collections. This was for security and economic as well as aesthetic reasons. Museum photographers are knowledgeable about the collections and know how to handle them. If an object is damaged or lost, the photographic reproduction is important documentation. Exclusive studio photographs are also used for more spectacular illustrations in museum publications as well as for commercial publications. In the last-mentioned case, reproductions are also an important source of income.

Even if the museum photographer is a respected professional, there is a difference in status between artistic studio photography and documentary reproduction of objects. Nowadays, when digital cameras are used by almost everybody, professional museum photographers sometimes work exclusively with artistic studio photographs, while employees working in storage areas use the camera for documentary purposes.

From a historical point of view, photography has been ascribed an aesthetic value relatively recently. Today photographs collected by art museums often form a category of their own among other art objects or, as in the case of so called photo-based art, an integral part of collections of paintings or graphic art. Photographs and photo-based art are also published on the web.

Photographic practices at art museums and ethnological museums differ. At ethnological museums, a camera is not only a tool for the professional photographer but also for the curator. The objects docu-

mented by the curator are not accessed museum objects but artefacts and documentation of everyday life.

Metadata

Archives

Archives have mostly organised their material in accordance with the international provenance principle. Their method of describing documents has been comprehensive, and studies using archival materials can therefore be very time consuming. In recent decades, increased demands of users for better access have resulted in efforts to provide more thorough descriptions and also to harmonise metadata with library standards. Standardised names and terms are used, and digitised archival materials are more often described as individual objects. When it comes to classifying text documents by subject, archivists – unlike librarians – are fundamentally resistant to guiding the user in a specific direction.

Manuscript collections held by important libraries have been handled in different ways. Ancient and medieval manuscripts are often thoroughly described in printed catalogues, sometimes with illustrations in facsimile. Collections of letters and private archives have been described comprehensively. Today, when many earlier hidden collections are catalogued and sometimes also digitised, thorough and standardised descriptions have become more customary. In spite of their ambition to harmonise with library standards, manuscript archivists often prefer a combination of the archival tradition of categorising materials and free text searching to subject classification and controlled indexing. Scholars, however, often want more metadata and better searchability.

Libraries

Libraries have a remarkable advantage because of their long and purposeful efforts to make their collections more easily available. This awareness has resulted in national and international co-operation including joint agreements about formats, cataloguing rules, classification, indexing and metadata standards. Library catalogues such as LIBRIS, built up by Sweden's National Library in co-operation with other Swedish research libraries, also contain huge amounts of structured and high quality metadata.

Library standards for classification and subject indexing have been developed for the description of published text documents. Nowadays, when earlier hidden special collections are digitised, this way of organis-

ing objects appears to be inadequate. For example, when describing drawings, prints and photography libraries can make use of museum standards.

Museums

When transferring handwritten information in inventories to printed catalogues, and later to databases, museum curators have gradually been obliged to systemise and standardise the different information elements. As museums generally have several different categories of objects in their collections and these objects are unique, they must handle comparatively detailed and specific metadata. Art museums have been more successful than ethnological museums in their efforts to develop standards and to join international agreements. This difference could partially be explained by the great importance which art museums have attached to securing correct information about the creator and provenance of works of art. The fact that objects of art generally have a more international character than ethnological objects has probably also stimulated international co-operation among art museums.

As the names of the creators of ethnological objects are often unknown, other names are of greater importance for their identification, for example names of artefacts, professions and methods of handicraft. Ethnological museums also have hidden resources of non-refined metadata, for example names of known craftsmen and photographers. In the future, this metadata could be refined and made accessible for archives and museums.

Whilst knowledge organisation is more complex within the museum sector than in libraries, the situation is even more complex for other types of museums. Museums of natural history, for example, have to use many different taxonomies and thesauruses when classifying their objects.

ALM

What efforts do archives, libraries and museums make in order to facilitate access to materials which have been placed in an institution within a specific sector, sometimes more at random than systematically? Today “interoperability” and “seamless access” are frequent keywords in discussions about accessibility. Authorised names and controlled vocabularies are important ingredients in this strategy, as well as cataloguing standards which are independent of a specific format. Following technical standards and, as far as possible, sharing technical platforms are also important conditions.

Even if technological development today supports interoperability, much remains to be done within the field of digitisation and accessibility.

The costs of equipment, labour and long-term preservation are enormous and, as our governments probably cannot find the money for the resources needed, many “memory institutions” – including archives, libraries and museums as well as other kinds of cultural heritage institutions – try to co-operate with commercial partners. As part of this development, new business models and new attractive services are being discussed. The agreement between Google and some of the most important European research libraries concerning mass digitisation of their collections illustrates the present balance of power, both between Google and memory institutions and between the interests of copyright holders and the users.

The National Library of Sweden

The ongoing work at the Swedish National Library within the field of digitisation follows an international pattern, where there is a shift from digitising single objects towards mass digitisation of complete collections. In contrast to the situation in some other countries, the Swedish government has not supported continuous mass digitisation, but from time to time it has allocated millions of kronor to temporary labour market programmes aimed at making collections accessible through digitisation.

There are several disadvantages to this temporary support. Today the Swedish National Library has a relatively restricted volume of digitised material. There is also digitised material which – due to insufficient metadata and an inadequate technical platform – cannot be displayed. Great efforts are being made, however, to set up goals and strategies which could improve the situation.

A history of trial and error

When digitising started during the 1990s, posters were high-priority material. The reason for this was both that this category of ephemera was printed on fragile paper, but also that many of the illustrated posters were well-suited for being exposed on the web. Thousands of posters were prepared and captured as pictures, but only a few of them were described with relevant metadata. This experience made many people realise that capturing the objects as pictures was the smallest part of the work, while preparing, cataloguing and long-term preservation took most of the time.

A production flow was developed where representatives of different professionals co-operated, primarily within the fields of preservation, cataloguing, photography and IT. As a result the various professionals

were obliged to devise a terminology which could be understood by all individuals involved in digitisation. Concepts such as “picture,” “image,” “original” and “copy” were discussed and, after defining the concepts, specific terms were stipulated. The library’s standards for classifying and indexing were also analysed, especially as tools for describing the content of pictures. After a first internal report in 2000, the National Library initiated an ALM-oriented project in which significant problems were identified and analysed in dialogue with external professionals, above all with representatives of central memory institutions. The report from this project was presented in 2003.

Suecia antiqua

Memory institutions with huge collections and long traditions often need time to advance from theoretical investigations to the practical work of digitisation. At the same time financing and copyright issues must be addressed before work can start.

The National Library’s digitisation of the most important single source of visual information about 17th century imperial Sweden, *Suecia antiqua et hodierna* (Ancient and Modern Sweden), can be considered as a step from theoretical discussions and frameworks to practical action. *Suecia antiqua* is an imposing illustrated work, with hundreds of engravings that display the most important places and buildings in Sweden during its 17th century Great Power Period. The National Library of Sweden has several copies of this work, and the plates are often used for reproductions.

The idea of digitising the plates was logical, since this frequent use caused wear and tear, both on the material and the staff who were to deliver the orders. Prospective users were scholars and an interested public, both in Sweden and abroad. A web application that would provide a presentation of the work and also serve as a tool for submitting orders was one of the project’s goals. Copyright was no problem. The goals of this project were achieved, and although the interface was rather primitive it was usable.

Development of methods

After considerable efforts to co-operate with archives and museums and to harmonise different solutions, it was again time for the National Library of Sweden to focus on internal consolidation and decisions on best practices. In a report from 2005 on digitisation and coordination within the National Library, a process-oriented method for digitisation was presented.

At about the same time, a pilot project started aimed at developing methods for the digitisation of the National Library's sensitive collections of weekly magazines. The final report, *Veckopress i nytt format* (Weekly press in a new format) from 2007 provides support for the conviction that digitisation is a method for substituting physical objects threatened by serious destruction. In spite of this optimistic attitude, some problems related to access still remained, especially questions concerning copyright.

Before this conflict between goodwill and potential was solved, the Swedish Parliament decided to spend millions of kronor on a digitisation project to stimulate the labour market. The goal of this project, called Access, was to improve the accessibility of the collections of memory institutions. The National Library now started planning for digitisation on a larger scale. A decision was taken on a thematic approach, and with a starting point from the working title "travelling in different times" a selection of items began. Books, magazines, journals, ephemera, prints, drawings and photographs were all included. The project group soon realised that this differentiated approach was both time-consuming and complicated with respect to production flow, technical equipment, meta-data and presentation on the web.

Partly in parallel with this Access project, digitisation of one of the National Library's most spectacular manuscripts was under way. The medieval *Codex Gigas*, a war trophy conquered by Sweden from Czechoslovakia during the Thirty Year's War, is known as the largest manuscript in the world and also as a Czech national treasure. The digitisation process was once again complicated and resource consuming. In 2007–2008 the physical manuscript was exhibited in Prague, and when the exhibition was closed the National Library of Sweden donated a high-resolution digital copy to the Czech National Library. The Codex Gigas project had been a public relations success, and much was also learned about the importance of realistic project planning.

Preparing for mass digitisation

In 2007 the Swedish National Library was reorganised. At the same time a Digital Library was established as a special department oriented towards technical development, digitisation, e-publishing and long-term preservation. The creation of the Digital Library made it necessary for other departments, especially the Department of Collections, to redefine their tasks and roles from being active partners in a process-oriented digitisation organisation to being partners preparing for digitisation from a more technical perspective.

Rational solutions to problems of mass digitisation were now given priority, and a pilot project for the mass digitisation of newspapers began. Meanwhile new methods were tested for cataloguing and displaying digitised items, especially manuscripts and pictures.

These different development projects had not been finished when the Swedish Government decided in 2009 to integrate the National Library of Sweden and the Swedish National Archives for Sounds and Moving Images. This integration strongly affected the organisation of IT and digitisation. The Digital Library was dissolved and a new IT department, including support, management, production and development, was created. At the same time, employees working with collections once again became more involved in the practical work of digitisation.

Today's reality

There are many questions on digitisation still to be answered, but some clear tendencies can today be observed at the National Library of Sweden:

- For 2009, the Government had allocated SEK 8.3 million for the digitisation of items which otherwise would be destroyed. The library has already planned mass digitisation of radio and television broadcasts and newspapers. Mass digitisation of newspapers will be outsourced and in-house digitisation will concentrate on ad hoc orders from users and digitisation on a smaller scale.
- Both potential sponsors and scholars are requesting co-operation within large-scale projects. The National Library of Sweden will therefore intensify its direct dialogue with scholars about projects which should be carried out, sometimes on a co-operative basis. When co-operation is on the agenda, the National Library will be responsible for developing the technical infrastructure and for knowledge of the collections, while the scholars will be responsible for related research.
- Many institutions do not know how to organise the digitisation process efficiently – locally, regionally, nationally and internationally. The National Library of Sweden has initiated a national digitisation plan, which aims at serving both as a theoretical framework and as a practical tool. The framework should include principles, standards and best practices concerning selection criteria, metadata, technical platform and access. In developing the plan, the National Library is co-operating with a national expert group for digitisation.

- The digitisation programme should be followed by practical planning of materials assigned high priority for digitisation, both within the National Library and from a national perspective.
- The metadata in LIBRIS are used for digitised objects. However, many huge special collections still remain to be catalogued.
- Today the National Library supports several different technical platforms and systems for the registration of different categories of library materials. The intention is, however, to build a broad platform usable for all categories. This solution will also allow seamless access to different databases.
- ALM co-operation is steadily developing, and this year the National Library and the National Archives unveiled Sondera, a service which makes it possible to search simultaneously in three different databases for archival, audiovisual media and traditional library materials.
- Most users are not interested in knowing whether a digital document is a result of digitisation or if it was born digitally. Today a new legal deposit law, complementing the present law on the legal deposit of physical objects and regulating deliveries of electronic resources published in electronic networks, is under preparation. The new law will enable access to certain *completed* electronic documents but will not include the dynamic web as such. As a consequence, the National Library needs to develop its previously launched web-harvesting methods and technical platform.

Utopias, dystopias and new strategies

During the past 15 years, concepts such as “access” and “accessibility” have been used as political mantras in the discussions about the future of memory institutions. Meanwhile the success of the Internet has dramatically increased user expectations for fast and efficient access. When archivists, librarians and museum curators realise the enormous theoretical and practical challenges and the costs of digitising, web harvesting and long-term preservation, such expectations – which are also part of a political vision – are sometimes regarded as unrealistic utopias.

Future success will probably depend on new technical solutions and new models for co-operation and partnerships. The National Library of Sweden has the ambition to save digital information for future generations and find new ways to solve the problems that we can identify today.

Applications 2: Adding value for research

CIRCULATION OF KNOWLEDGE: RECONSTRUCTING THE 17TH CENTURY DUTCH REPUBLIC OF LETTERS¹

Martin Bossenbroek

In 1670 Baruch de Spinoza published his *Tractatus Theologico-Politicus*. The treatise instantly stirred up virulent discussions all over Europe: some scholars were as impressed by its content as others were horrified. Although it cannot be assumed that the book circulated freely in the Dutch Republic, it was not officially forbidden by the authorities until four years later, to be precise, in July 1674. In the meantime, Spinoza had finished an even more daring work, the *Ethica* ('*The Ethics*'). In 1675, he took it to his publisher Jan Rieuwertsz in Amsterdam. However, discouraged by rumors about the book's forthcoming unfavorable reception, Spinoza refrained from the planned publication. He then took the manuscript, presumably a fair-copy, back to The Hague, where he had been living for the previous couple of years, and set it aside for future publication.

When he died, in February 1677, the work still hadn't been published. Immediately after his death, his landlord arranged for his writing desk and manuscripts, including the *Ethica*, to be taken to Jan Rieuwertsz. The publisher not only secured the manuscripts, he also enabled some friends to publish Spinoza's *Opera Posthuma*, soon to be followed by a Dutch translation, the *Nagelate Schriften van B. d. S.* Rieuwertsz must have sold the first copies of both editions by the end of the year, in Amsterdam, but also in other Dutch cities like Leiden and The Hague.

As Spinoza himself had feared, the book's content alarmed the authorities. The most explicit parts were read out loud during city council meetings to denounce it and obtain a ban. On 25 June 1678 the *Ethica* was forbidden by the States of Holland and Friesland because it contained 'a lot of profane, blasphemous and atheistic assertions'. Nevertheless, the scholarly world continued to show great interest in Spinoza's latest work. As a matter of fact, the *Ethica* spread all over Europe by means of a secret 'scholarly black market'. We know this by direct evidence from letters or other documents, for instance of Dutch scholars like Jan Swammerdam and Constantijn Huygens sending copies to friends in Paris. And there is also plenty of indirect evidence. Gottfried Wilhelm Leibniz for example had asked Georg Hermann Schuller to keep him informed about works published in the Dutch Republic and especially those by Spinoza. Letters to colleagues prove that Leibniz later must have received a copy of the book in Hannover.²

In this case, Spinoza is the protagonist. We can get a glimpse of his personality and his ideas through the works he left us. Letters and documents of other scholars add

valuable extra information about his work and how it was perceived and judged by contemporaries. Naturally, the network of persons who had connections with Spinoza is far more complex than described here, and it goes without saying that there were numerous other relevant relationships, e.g. the role of Ehrenfried Walther von Tschirnhaus and the reception of Spinoza's work by members of the Royal Society in London. Even with the documents at hand, there is space for many interpretations about what really happened, but taken together they give us an impression of the relations between scholars, for example the close collaboration of Spinoza and his publisher Rieuwertsz, Leibniz' interest in Spinoza and his work and – last but not least – the contradictory role Schuller might have fulfilled in the relationship between the two.³

The case of Spinoza's *Ethica* is just one example of the way the network of the 17th century scholarly community can be made visible to contemporary researchers. Letters in particular can provide us with new knowledge, especially when these letters become contextualized.



Figure 1. On this map of the 'Belgii Foederati' you can actually see the Dutch Republic as it was documented in 1658 by one of the most renowned Amsterdam cartographers of his time, Johannes Janssonius. The Dutch Republic had been proclaimed 70 years before this map was printed, in 1588. The small Republic's Seven United Provinces roughly coincide with what is the Netherlands today. Its independency as a republic was going to last until the French invasion of 1794/95.

The more documents can be aggregated in one place and the better they can be made minable in a meaningful way, the more successful 21st century scholars will become in deriving new insights from known documents. That is what the challenging project *Circulation of Knowledge* is all about: exploring the consequences and threats but most of all the thrilling chances of digitization in a modern information environment. Before ‘going digital’ in a more specific way, first let me introduce the historical scenery.

During the first half of its existence, approximately during the 17th century, the Republic gained a leading position in an international trade network and became one of Europe’s wealthiest regions. The Dutch had the most powerful fleet in the world. It was kept operable by a class of extraordinarily rich merchants, who maintained far-reaching trade relations with the East and the West. Therefore, the Republic’s wealth was one of the two important factors in attracting inhabitants from neighboring regions, above all from the former Southern provinces.

The other reason was its relative open-mindedness and tolerance. As was shown by the ban on Spinoza’s *Tractatus Theologico-Politicus* and *Ethica*, even here authorities set limits to expressing ideas and convictions freely. Nevertheless, many people who had experienced problems in their home countries found a safe refuge in the Republic. Among these were many outstanding craftsmen, printers, artists and scholars. They contributed to the impressive cultural and scholarly production of the Dutch ‘Golden Age’.

Refugees preferably moved to large cities: Amsterdam, Leiden, Rotterdam, Haarlem, Middelburg and Utrecht. The Hague, the formal capital where the National Library was to be founded in 1798 (yes, the French invasion had some positive outcomes, too), was much smaller and did not even have city rights. To give an impression of the extent of urbanization and migration in the Republic, counting about 2.5 million inhabitants in total in those days, the city of Amsterdam, located in the densely populated province of Holland, grew to over 200,000 inhabitants during the 17th century. One third of them spoke the Antwerp dialect because they originated from that area in the Southern provinces. In Leiden (counting about 65,000 inhabitants) the percentage of foreigners was even higher at certain times, due to its work facilities in textile production and – more importantly in this context – because it had developed into an important centre for scholarship after its university was founded in 1575. Since then, four more universities were founded in the Republic, namely in Franeker, Groningen, Utrecht and Harderwijk – where by the way, Carolus Lin-

naeus, spent seven days, just enough to receive his Ph.D. on 23 June 1735. Next to that, there were ‘illustre scholen’ in Amsterdam and Deventer.⁴

Dutch cities were linked by a sound infrastructure, via overland routes and, more important, by canals that led throughout the country. After 1657, one could comfortably travel between Amsterdam and Leiden (via Haarlem) by ‘trekschuit’ (a boat moved by horse- or manpower). This took about seven hours. This Dutch infrastructure facilitated travel and a smoothly functioning postal system that enabled the rapid exchange of information, in print or hand-written. Letters and pamphlets were an efficient and effective way to circulate news, observations and ideas between individuals.⁵

Scholars made extensive use of this medium, especially before the advent of the first scientific journals in 1665, starting with the *Journal des savants* in Paris and the *Philosophical Transactions* of the Royal Society in London. Scientific journals came close to letters and pamphlets in circulation speed, especially when compared to books that took much longer to



Figure 2. Pieter Corneliszoon Hooft, Hugo Grotius, Caspar Barlaeus, René Descartes, Constantijn Huygens and his son Christiaan, Baruch de Spinoza, Antoni van Leeuwenhoek (the only one of the scholars shown here who had not received any university education) and Jan Swammerdam (this is actually a ‘fake’ portrait – there is no ‘certified’ contemporary image of Swammerdam). The other five depicted scholars did *not* live in the Dutch Republic but maintained a close relationship with their colleagues who did, for example the Frenchmen Nicolas Fabri de Peiresc and Marin Mersenne (he truly deserves the title ‘information broker’ due to his central position in the scholarly information network), Henry Oldenburg, Blaise Pascal and their German colleague Leibniz, just to name a few. The intensive contacts between these scholars can be deduced from thousands of letters that have survived.⁷

prepare, publish and disseminate. In the early days, journals contained reports in the form of scholarly letters. It was only later, that the form of the contributions changed. Journal publications had a larger outreach than letters because they could be read by many. Journals were to become an effective and efficient distribution channel for scholarly information.⁶

Now we have come a bit closer to the Dutch Republic of Letters, where the power of the word reshaped the view on the world of contemporary and future generations. Among the scholars who lived and worked in the Dutch Republic during the 17th century were famous personalities, arranged according to their year of birth on the previous page.

These letters are the rich sources the project *Circulation of Knowledge* is going to explore and to exploit. The main objective is to gain insight in the transfer of knowledge between 17th century scholars via their correspondence – in other words, to reveal social networks and paths of information exchange in the scholarly world of the Dutch Golden Age. The project aims to offer new possibilities for access to historical content and to work with it through the application of custom-made information and communication technologies, thereby supporting researchers in answering novel research questions in an unprecedented way. The most ‘tangible’ result of the *Circulation of Knowledge* project will be the creation of a web-based laboratory for researchers, a so-called ‘collaboratory’.

Six partners work together in the project: the University of Utrecht is represented by the Descartes Centre for History and Philosophy of the Sciences and the Humanities; the University of Amsterdam is a partner with its Department of Historical Dutch Literature; the Huygens Institute, one of the institutes of the Royal Netherlands Academy of Arts and Sciences (KNAW), supports the project as well as the Koninklijke Bibliotheek (KB) and two other organizations linked to the KNAW: Data Archiving and Networked Services and the Virtual Knowledge Studio for the Humanities and Social Sciences. Through these partners, the project can benefit from other initiatives and networks that are being deployed at this moment. In addition, the project is supported by other parties, for example CLARIN (the acronym of ‘Common Language Resources and Technology Infrastructure’), a European initiative that has a strong backbone in the Netherlands. The project is funded by the Netherlands Organization for Scientific Research (NWO). Having started in November 2008, it will be completed after four years, in October 2012.⁸

Being the Director of Collections and Services of the KB, I permit myself the liberty of singling out the involvement of my own organization.⁹ The KB in The Hague is the National Library of the Netherlands,

and as such, focuses on preserving and managing Dutch cultural heritage as well as making and keeping it accessible to the public. The KB is responsible for the collection of printed and electronic publications from the Netherlands and administers and updates the national bibliography. It offers researchers in the humanities a rich collection of resources in a wide variety of fields – for example, highly relevant in this particular case, the correspondence of Constantijn Huygens. The revolutionary changes in our modern information environment have been a strong motivation for the KB to support research projects with respect to standardization, digitization and digital preservation. And of course, these efforts would not be complete without enlarging and improving the portfolio of web-sites that make the KB – and other institutions’ – collections available online. To achieve all this, the KB contributes to many initiatives which explore novel ways to provide content and to produce, enrich and contextualize it.¹⁰

One of the most challenging initiatives is this *Circulation of Knowledge* project. The resulting collaboratory will be developed according to the needs of researchers in the humanities, but in fact everybody will be able to make use of the facilities it offers, if he or she has internet access. So what can you actually *do* on this collaboratory website in the making?

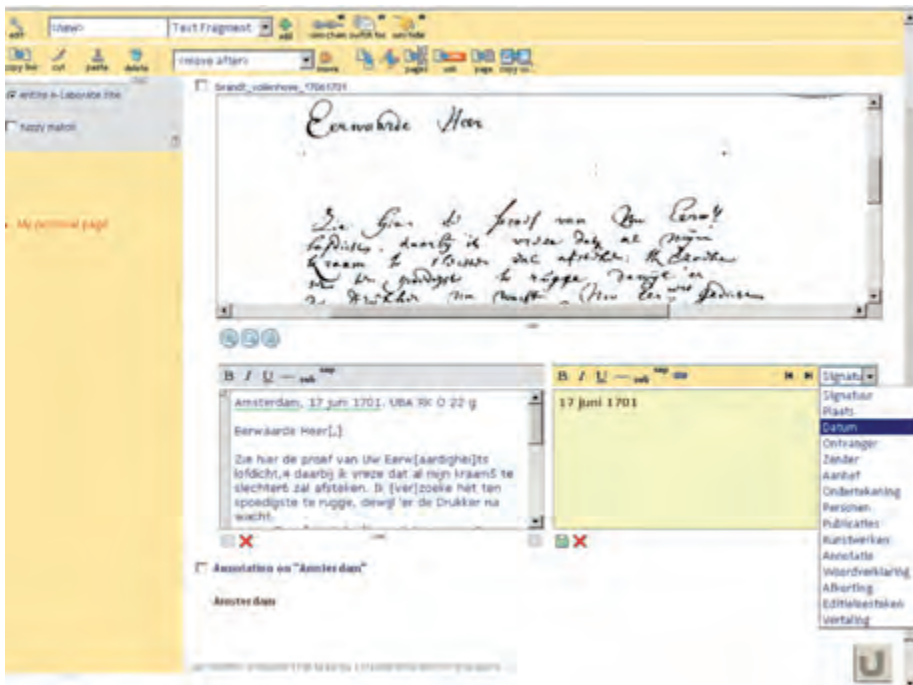


Figure 3. Collaboratory website of Circulation of Knowledge project.

First of all, it will be possible to enter and manipulate the aggregated data. As mentioned above, the core material consists of letters by 17th century scholars with a strong relationship to the Dutch Republic. The starting point of the project is a collection of transcriptions of those letters, in other words electronic text. The transcriptions can be annotated. They are complemented with metadata (which is obligatory) and scans (which is optional). New transcriptions can be entered once a scan from an original document is uploaded to the system. The system should be able to ingest large quantities of data and therefore the aim is to develop a solution that is scalable. On the previous page is a screenshot of such an uploaded scan of a letter, with room underneath for the transcription and for annotations, linked to a certain passage of the transcription. In the toolbar at the top, one can execute all kinds of operations, such as upload, arrange, edit and search documents.

Secondly, visitors to the website will be able to work with the content. The system will offer a powerful search facility within the transcriptions and the metadata, tools for analyzing content *and* for visualizing results, be it semantically, chronologically or geographically. Standard functions

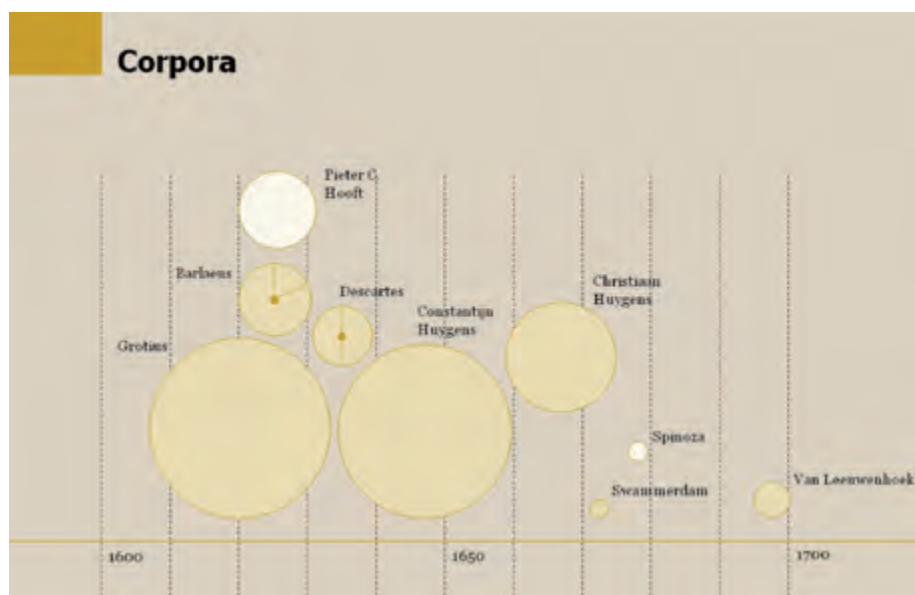


Figure 4. From left to right one can distinguish the correspondence of Hugo Grotius, Caspar Barlaeus and René Descartes. The Grotius correspondence, consisting of 7,725 transcriptions of letters exchanged with nearly 400 correspondents, is available in annotated XML/TEI. René Descartes' letters come to us in a similar format; nearly half of the 750 letters were transcribed before the start of the project. Of Caspar Barlaeus' correspondence about one sixth is available in transcription. The rest of the 1,257 letters still needs to be processed. All of them still have to be converted from Word documents into XML/TEI.

to save or output results in a personalized environment are provided as well. Next to that, the core content can be contextualized, discussed and reused. It will be possible to establish links to other resources, such as biographies, bibliographies and catalogues, or other web environments. The idea is to offer professional users support and a platform where they can exchange ideas or work together. Next to the content which will be made available, all of the tools to be developed in this project will be open source software that can be reused in other environments.

A selection of transcriptions has been made to work with during the project. To give an impression of the quantity of letters, the correspondence is displayed on the previous page in a graphical representation. Each circle's size represents the quantity of letters; each circle's centre is positioned right in the middle of the period a particular correspondence took place.

Two more large sets of transcriptions will be part of the collaboratory, namely the correspondences of Constantijn Huygens and his son Christiaan. Huygens senior and junior will form the backbone of the system, not only because of the large number of their letters (7,295 and 2,900 respectively) but above all because of their central position in the scholarly network, from a social as well as a geographic and a chronological angle. Father Huygens, for instance, can be linked to no less than 1,307 correspondents.

Furthermore there is the smaller correspondence by Jan Swammerdam (about 80 letters) and the collection of letters written by Van Leeuwenhoek (nearly 400 letters). The latter is exceptional because it is unilateral: Van Leeuwenhoek used the letter as a medium to publish; most of his letters were directed to the Royal Society or its members in London and meant to be printed in its *Philosophical Transactions*. Some other corpora might be enclosed in the system as well, if intellectual property rights allow this, in particular the correspondence of Pieter Corneliszoon Hooft and of Spinoza. Most of their letters will have to be converted to meet the requirements of the collaboratory. In total, it will contain about 20,000 letters.¹¹

As can be deduced from these data, there are a number of problems to be tackled. The content is highly heterogeneous. The number of letters per correspondent differs hugely, and especially a unilateral correspondence might pose a problem whenever overlap in correspondence is necessary to answer a particular research question. Then there is the varying level of enrichment; whereas some of the transcriptions are heavily annotated, others lack any annotation at all. Transcriptions have been prepared on the basis of different transcription criteria; therefore

they diverge enormously in quality. The edition of Constantijn Huygens' letters, for instance, requires further improvement. Some of these letters were transcribed a hundred years ago, and more than once the editor limited himself to a summary of a letter or even skipped it completely.

Another problem to be solved is the multiple languages used in the letters. Most of them are written in Latin and French, but there are also letters in Dutch, English, German and Italian. A search term entered in one language should retrieve the relevant term in all documents, regardless of their language. Next to dealing with a variety of languages, the system has to cope with all kinds of spelling variations. Named entities should be recognized in all possible variations. The same applies to synonyms. An even greater challenge is the recognition of – for example – evolving ideas. Although a concept might be articulated in a completely different way, software should be able to detect a correlation between documents dealing with the same concept. And of course, the formal representation should be standardized. Documents and metadata have to be aligned sufficiently in order to apply advanced software tools.

This is a very brief and by no means exhaustive list of problems that will have to be solved in the course of the project. Some approaches for solving these problems include a clear definition of baseline requirements, and the use of technology to enrich texts automatically. For the challenge of multilingualism a solution has to be found in modern language resources and technology. Texts should be tagged and categorized semi-automatically, with manual work being restricted as much as possible. Metadata should be extracted automatically. The problem of named entity recognition can be tackled by existing resources and technologies, made suitable for the project's specific needs. And to improve interoperability, existing standards will be adopted whenever possible.

This last-mentioned aim underlines the fact that this is not a completely unique project. There are a couple of similar projects, which are all wonderful initiatives that have given users new research opportunities. All of these projects aggregate transcriptions of high quality, sometimes accompanied by annotations and scans of the source material. The transcriptions have sometimes been prepared in joint efforts, but usually not in online collaboration. Various possibilities are offered to enrich and mine the transcriptions, but usually no tools are offered – at least not yet – for statistical analysis and visualization. The systems aim at distinct target groups and online access is not always free of charge. There are not many projects that are designed to be scaled up in the future – this depends on the source material ingested.

Similar Projects		Statenbijbel	Van Gogh Letters Project	Hartlib Papers	Electronic Enlightenment	Circulation of Knowledge
Large corpus of transcriptions (+ annotations)				■	■	■
Scans provided	■	■	■	■	■	■
(Simultaneous) collaboration	■					■
Text enrichment tailored to researchers' needs			■	■	■	■
Searchability (full text, metadata, languages)	■	■	■	■	■	■
Tools for analysis and visualization						■
Links to other sources		■			■	■
Scalability					■	■
Free online access	■	■				■

Figure 5. Comparison between projects.

The projects I would like to refer to are: the Dutch collaborative transcription of the *Statenbijbel* of 1637, the Van Gogh Letters Project, the Hartlib Papers project and last but not least: the wonderful Electronic Enlightenment, of which Robert Darnton was the great instigator. The *Statenbijbel* project has succeeded in collecting the input of more than hundred volunteers to transcribe the text. It is freely available on more than one website and every website offers its own interface with particular facilities, e.g. for searching the transcription or for consulting the scans.¹² The Van Gogh Letters Project provides the visitors of the website with annotated transcriptions, translations and scans of the originals. The content is interlinked with other resources, above all with image material that is of particular relevance when consulting Van Gogh's correspondence, consisting of approximately 900 letters.¹³ The Hartlib Papers Project may be the oldest initiative, having started with a publication of transcriptions of the papers of Samuel Hartlib on CD-ROM in 1996, including the scans. The correspondence's publication was transferred to hriOnline in 2002. The number of pages is large, namely 25,000 folios of manuscript.¹⁴ The Electronic Enlightenment comprises more than 54,000 letters of over 6,000 correspondents at this moment, in transcription with annotations and links to external sources. This collection is extremely rich and will be even richer in the future. Subscription is necessary to access resources that can be cross-searched via a

user-friendly interface, both in full-text and metadata search. Navigation is simple yet flexible. Transcriptions can be used alongside biographies and other resources that researchers might want to consult.¹⁵

The *Circulation of Knowledge* project aims to achieve something similar like the Electronic Enlightenment, however with the focus on a different time interval. It will furthermore deploy tools for text mining and visualization that will add to the user's understanding of the content, which can be analyzed both qualitatively and quantitatively. It will offer users a growing number of transcriptions, which can be created by a large group of contributors due to the online collaboration environment. The workflow will be organized in a way that allows for quality control. The result of the project will be an infrastructure that enables future growth, not only in content, but also in tools that can be added and applied to the content. The collaboratory will become more and more valuable as a research environment.

To illustrate the added value the *Circulation of Knowledge* project aims to establish, I would like to end with two historical examples, both featuring the French philosopher and mathematician René Descartes, who

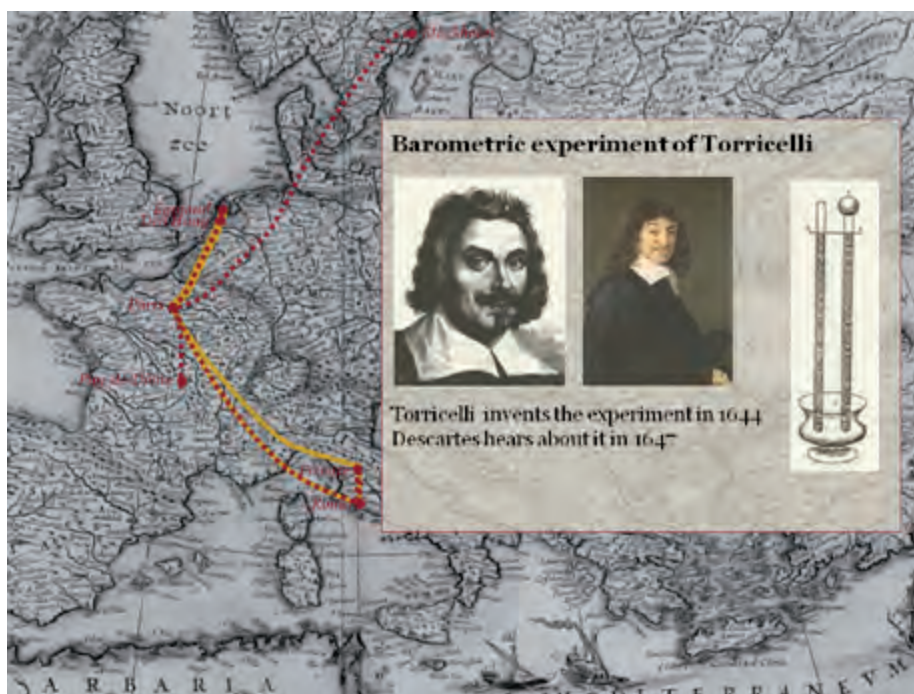


Figure 6. This map of Europe, taken from Joan Blaeu's *Atlas Maior* (made in 1664), serves as a background to visualize the consequences of an event that took place in 1644, namely the invention of the barometer, which led to the discovery of the vacuum.

lived and worked in the Dutch Republic from 1628 till 1648 – and both linked to Stockholm. Of course, the pictures displayed on the previous page are far inferior to the (animated) visualizations the collaboratory will facilitate.

It all started in 1644 in Firenze, where Evangelista Torricelli experimented with a glass tube filled with mercury, set vertically into a basin with the same liquid. He noticed that the mercury in the tube dropped down to a certain extent and concluded – rightly – that the space that resulted above it was a true vacuum. At the time, that was a breathtaking view, and Torricelli's discovery unleashed agitated discussion. Many contemporaries considered a true vacuum to be impossible, and the mere idea of empty space provoked strong reactions. The experiment brought about a whole chain of similar experiments performed by various European scholars during the next few years. The interesting aspect in relation to the *Circulation of Knowledge* collaboratory is how relevant information was passed on and – perhaps even more interesting – which parts of information were *not* transmitted. It is crucial for understanding 17th century science, or better, natural philosophy, to carefully observe how the conclusions of individual scholars complemented existing knowledge.

For instance, Descartes, in this case acting as the 'Dutch connection,' only heard about the experiment in 1647, but came up right away with his own explanation for its outcome. In fierce debates with other scholars like Blaise Pascal in Paris and Pierre Chanut in Stockholm, and referring to several new experiments, among others at the Puy de Dome, Descartes kept insisting that the empty space in the tube was not a vacuum but contained, as he called it, *subtle matter*. In fact, he wrote in a letter to Constantijn Huygens that 'the only true vacuum exists between Pascal's ears.'¹⁶

But instructive it is, and the same goes for the second visualization on the next page, this time fully derived from Descartes' correspondence. It is a simplified mock-up of his letters and his travels during the last three years of his life, and as such gives a clear idea how complex data may be visualized. His travels are indicated with yellow lines, his letters with black (dotted) ones. The names of Descartes' correspondents are in grey, their whereabouts – and of Descartes himself – in red. Of course, this presentation is not meant to go into details, just to give an idea of the overall picture. The only detail singled out would be his last letter, written on 10 February 1650 to his brothers Pierre and Joachim. By then he was on invitation here in Stockholm, teaching Queen Christina. The next day, 11 February, he died from pneumonia.

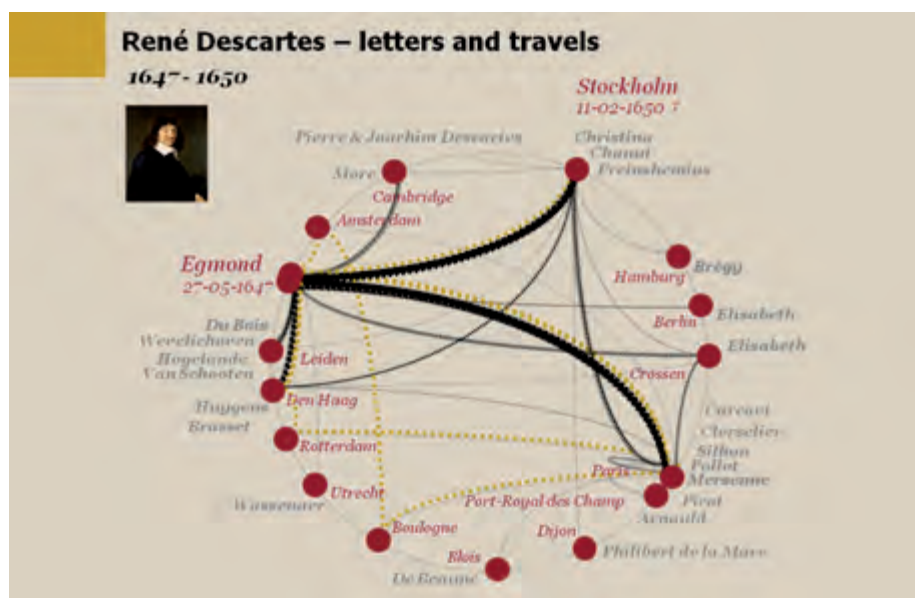


Figure 7. René Descartes – letters and travels.

So we're back in Stockholm again, where I'd like to conclude by summing up the most important ways in which ICT can add value to research in the humanities:

- content aggregation, together with free availability and exchange of data are basic ingredients for facilitating research in the humanities;
- adopting existing standards in the development of open source software will give the scholarly community powerful tools at hand;
- deployment of text mining on various levels offers the chance to pose and answer novel research questions;
- tools for data analysis and visualization will enhance the possibilities to derive new knowledge from complex data;
- application of ICT not only saves the researcher time but also opens up new ways to look at 'old data';
- researchers are enabled to collaborate and exchange their ideas and results.

References

- 1 The full name of the project is 'Circulation of Knowledge and Learned Practices in the 17th century Dutch Republic. A Web-based Humanities Collaboratory on Correspondence.' I'd like to thank Erik-Jan Bos, Guido Gerritsen, Erik Jorink, Dirk van Miert and Huib Zuidervaart, all members of the project team, for their valuable contributions to the presentation and the article. I owe special thanks to Brit Hopmann for preparing the basic text.
- 2 Jonathan Israel, 'The Banning of Spinoza's Works in the Dutch Republic (1670–1678),' in: Wiep van Bunge and Wim Klever eds., *Disguised and Overt Spinozism around 1700* (Leiden, 1996) 3–14. Piet Steenbakkers, *Spinoza's Ethica from Manuscript to Print* (Assen, 1994).
- 3 Wim Klever, *Mannen rond Spinoza 1650–1700* (Hilversum, 1997).
- 4 Dirk van Miert, *Humanism in an Age of Science. The Amsterdam Athenaeum in the Golden Age, 1632–1704* (Leiden, 2009).
- 5 Maarten Prak, *The Dutch Republic in the Seventeenth Century* (Cambridge 2005). H. Bots and F. Waquet (eds.), *Commercium Litterarium. La Communication dans la République des Lettres. Forms of Communication in the Republic of Letters, 1600–1750* (Amsterdam-Maarssen, 1994).
- 6 Hiram Caton, *The Politics of Progress. The Origins and Development of the Commercial Republic 1600–1835* (Gainesville, 1988). Hans Bots ed., *La diffusion et la lecture des journaux de langue française sous l'ancien régime: actes du colloque international, Nimègue, 3–5 Juin 1987* (Amsterdam, 1988). I. Avramov, 'Henry Oldenburg and the Early Years of *Philosophical Transactions* 1665–1677,' in: V. Bouzov and Veliko Turnovo, *Europe at the Early Modern Time Period: Universities, Science, Humanism* (2009) 175–188.
- 7 Henk Nellen, *Hugo de Groot: een leven in strijd om de vrede, 1583–1645* (Amsterdam, 2007). Eric Jorink, *Reading the Book of Nature in the Dutch Golden Age* (forthcoming publication in *Brill's Studies in Intellectual History*). Klaas van Berkel, Albert van Helden en Lodewijk Palm eds., *A History of Science in the Netherlands* (Leiden, 1999).
- 8 Cf. Dirk Roorda, Erik-Jan Bos and Charles van den Heuvel, 'Letters, Ideas and Information Technology: Using Digital Corpora of Letters to Disclose the Circulation of Knowledge in the 17th Century,' International Conference Digital Humanities 2010, King's College London, <http://dh2010.cch.kcl.ac.uk>.
- 9 On 1 November 2010 I rejoined the active labour force of the scholarly community, as an Associate Professor at the History Department of Utrecht University.

- 10 See for the variety of projects and other activities the KB's website: <http://www.kb.nl>.
- 11 Since the Nobel Symposium of 24–26 June 2009 considerable progress can be reported. I will just sum up some highlights. My presentation in Stockholm led to an invitation by Professor Pietro Corsi for the Day of Data of the Cultures of Knowledge (CofK) project at St Anne's College, University of Oxford, 10 July 2009. Contacts between the two projects since then have been intensified, as with the Academic Computing/Humanities Center at Stanford University. The collaboration focuses on the standardization of personal and geographical names, and on defining a minimal metadata set, in order to be able to share information. Other (international) contacts relate to visualization expertise, especially on Topic Modeling tools (with specialists of Indiana University in Bloomington), and to using the Associative Neural Network for the Humanities (ANNH) as a comparison system (with researchers of the Philosophy Department of Utrecht University). As a result of a combined workshop with a CLARIN-EU delegation in September 2009 in Lancaster, plans were developed to create an analysis pipeline based on linguistic and statistic tools and a demonstrator, focused on keyword extraction. CLARIN-NL granted additional financial support for this subproject. By September 2010, the total research corpus consists of more than 15,000 letters (Grotius, and father and son Huygens). The correspondence of Barlaeus, Descartes, Leeuwenhoek and Swammerdam are next in line.
- 12 <http://www.inl.nl/images/stories/bijbels/statenvertaling1637>
- 13 <http://www.vangoghletters.org/vg>
- 14 <http://www.sheffield.ac.uk/library/special/hartlib.html>
- 15 <http://www.e-enlightenment.com>
- 16 Erik-Jan Bos and Theo Verbeek, 'The Role of Observation and Experiment in Descartes' Correspondence,' in: D. van Miert ed., *Observations in Early Modern Letters, 1500–1650* (London, 2011) [forthcoming].

STABILISING KNOWLEDGE THROUGH STANDARDS: A PERSPECTIVE FOR THE HUMANITIES

Laurent Romary

Introduction

This paper is all about a paradox. How indeed can we associate the notion of standardisation, which is all about fixing rules for a given field of knowledge, with scholarly work, which, on the contrary, is basically about departing from existing knowledge and the discovery of new concepts? It may be even more paradoxical to address the issue of standardisation in the humanities, which does not seem to rely on specific technological environments to fulfil its research endeavours.

Still, this rhetorical association is not so much a paradox if we consider that science¹ is all about sharing information between researchers, thus requiring that minimal joint practices are actually agreed upon among them. Indeed, bibliographical references for instance, have been since ages the subject of standardisation activities, so that one can easily retrieve the exact source of a given citation.² Further, the increasing role of data in science has already encouraged some scientific communities, such as astronomers, to define coherent protocols and formats for sharing their information.

All in all, the humanities are not that far behind the other scholarly fields, but efforts remain to be made to widen the acceptance of pioneer works in digital humanities or language technology. In this context, we attempt to provide a picture of what an involvement of scholars in standardisation could look like, by eliciting both the possible conceptual and technical backgrounds. Our main objective is to contribute to extend the awareness of the scholarly community on the role of standardised data and possibly encourage more scholars to consider participating in the definition of such standards as a natural component of their mission.

Dealing with data in science

Just as in all aspects of the human society, the quick and intense evolution of information technologies has had a very strong impact on the scholarly world. One can now observe that a new generation of digital scholars has emerged, who are now carrying out most of their research activities online, and fears are already expressed that it may become difficult to deal with the forthcoming deluge of data.³

We knew already that science is essentially a matter of information. For more than two centuries, the wide dissemination of scientific journals has allowed research results to be quickly and widely known by the corresponding scientific community. In domains like mathematics, this has become the major record of science, whereas other branches of science have identified the importance of research data (protocols as well as results proper) as an essential element of accumulating knowledge. In the human sciences, primary sources have always been the basis for the grounding and comparison of scholarly statements.

Lexica as a representative example in the humanities

Still, humanities cannot be taken exactly in the same manner as other scholarly fields. Indeed, the main sources of information, *primary* sources, are part of a continuum of textual documents that makes observations and conclusions pretty much of the same nature. A simplified view on humanities research could indeed make it boil down to producing commentaries on existing textual sources, where these commentaries in turn become sources for further scholarly work.

However, the situation is by far more complex than this simplified picture. Over the years, humanities scholars have based their research on a variety of information sources, most of them becoming digital over the years. Archaeologists, as the first example that would come to mind, have always been forced to maintain huge collections of place and object descriptions to be able to identify precisely the correlations between discoveries. Historians of art have had a similar need to record the characteristics of artistic objects or buildings. Even more, step by step they have put together large prosopographic databases describing persons and places related to the artefacts under study. Similar trends⁴ have touched most of the domains of historical research, which could hardly work without widely available well-maintained prosopographic sources.

Scholars working specifically on language have also long since developed digital databases and methodologies to deal with the huge amount of information that is required for meaningful research in such domains as general linguistics, field linguistic or philology. Such works as that of Fr. Roberto Busa,⁵ producing a fully lemmatised version of the works of Thomas Aquinas, have been seminal in putting together the core methodological principles of what is now known as computational linguistics.

In this context, lexica have always played central role as major information structures for linguistic observations. Whether these are used as sources or as the main research output of a study, structured lexical data can be seen as prototypical methodological objects in the humanities. They encompass a wide range of forms, from simple word lists to complex encyclopaedic data. They intend to reach many different communities, layman or scholars. And of course, at a very early stage in the digital era, they have been the objects of specific attention so that they could be used directly on a computer.

Observing ongoing projects working on digital dictionaries or lexical databases reflects these various aspects and the corresponding complexity in providing a coherent view of the various type of corresponding data formats. Without aiming at being exhaustive, which would indeed be impossible given the intense activity in this domain, let us go through a few illustrative examples.

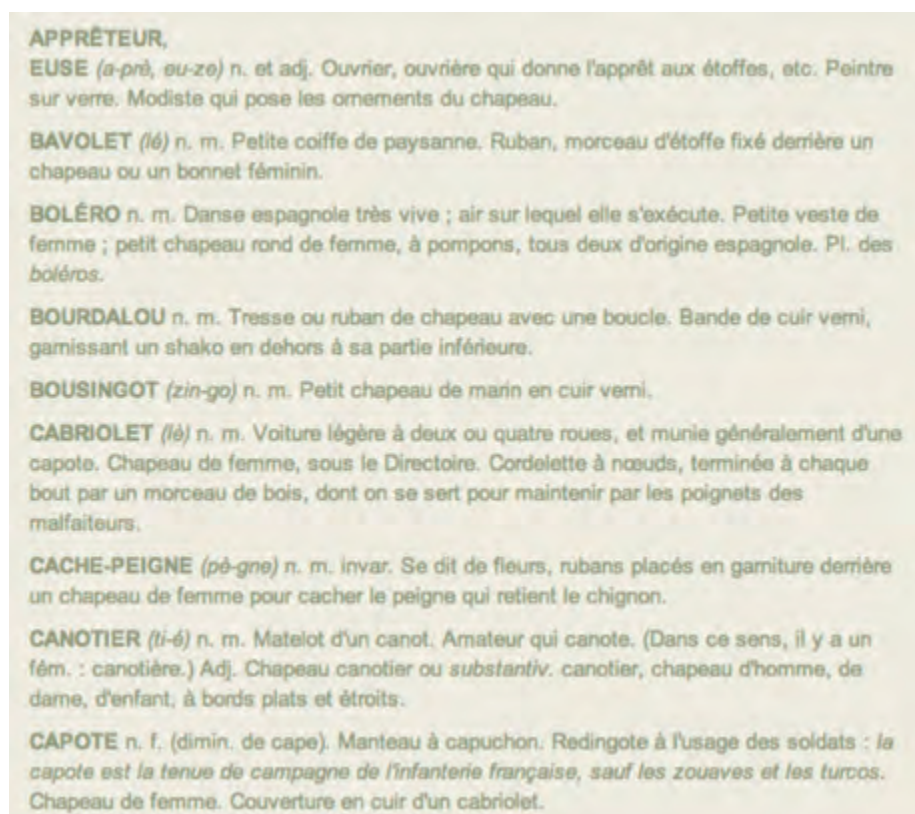


Figure 1. Beginning of the results obtained in a search for "chapeau" in all definitions of the *Petit Larousse 1905*, source: LDI.

As in many other scholarly fields of interest in the humanities, the usual transition out of traditional book-based scholarly work is to digitise reference sources in order to increase their dissemination and usability. In this respect, many scholarly groups have focused on providing full-text transcriptions of dictionaries that were either their focus of research as lexicographers, or important reference sources for wider linguistic studies. For instance, the LDI laboratory⁶ has digitised several generations of the *Petit Larousse*, with the purpose of providing augmented views, field-based query facilities (see Figure 1) on all fields and cross-links with other dictionary sources (e.g. the *Dictionnaire de l'Académie*). The generic research question⁷ here is to observe the evolution of structure and content across the years. As can be seen, the overall structure remains straightforward since dictionaries at this period had already acquired a good level of structural stability.

On the contrary, when similar attempts are based on more ancient form of dictionaries, the achievement of a fully digitised version bumps into several kinds of obstacles. Whether because of the quality of paper, the nature of the font, the orthographic variations or the actual blurred transitions between the various fields of the dictionary, it can appear to be quite complex to organise the initial material as a real structured object which can be precisely queried and cross-linked. As can be seen in Figure 2, the various levels of complexity are made even more complex because of all the implicit segments (e.g. for expressing morphological variations) that appear in even a very simple entry. As illustrated by the exemplary work done within the Textgrid project,⁸ the digitisation of such an ancient dictionary can actually lead to the identification of generic guidelines for a wider range of dictionary type.

Beside such projects, which take a printed primary source as basis for creating the digital resource, there are more and more lexicographic initiatives that are born-digital. Some are actual dictionary projects in the

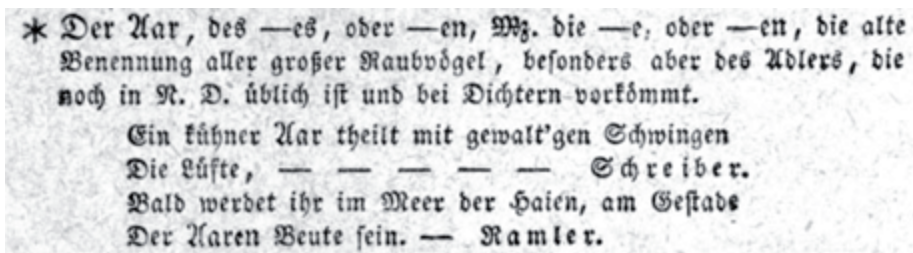


Figure 2. Entry for 'Aar' in Joachim Heinrich Campe „Wörterbuch der deutschen Sprache,” 5 volumes, Braunschweig 1807–1811.

traditional sense. The Franqus⁹ dictionary project, for instance, aims at describing the French language as spoken in Québec and has put together a workflow and methodological principles that are very close to similar large-coverage dictionary projects.

Still, the recent decades have seen an increase of activity in creating lexical resources, which, instead of being intended for human usage, are conceived as information resource for the automatic processing of text and speech. Such resources do not actually contain any definitions, or illustrative examples, but gather specific morphological (inflected forms) or syntactic (possible constructs associated with a word) information, which can then be used to tag the various surface forms encountered in a text. Such lexica thus bear a very flat structure, so that for a given form such as “ferme” in French, the following record can be made:

ferme ip1s ip3s sp1s sp3s im2s,fermer v

which indicates the possible morphological features associated with it (ip1s: indicative, present tense, singular, first person), the infinitive form (“fermer”) and the part of speech (‘v’ for verb).

Such lexica are usually joint deliverables with fully annotated corpora,¹⁰ and aim at having as wide a linguistic coverage as possible. Larger initiatives such as the Multext-East¹¹ consortium, have even worked in parallel on multilingual lexica covering most languages of Eastern Europe, creating the need for joint descriptive methodologies of lexical content across languages.

Finally, we would not have a representative sample of lexical initiatives here if we did not mention the importance of lexical description for the record of endangered languages and more widely for the documentation of languages and dialects around the world. In this case, the objective is no longer to provide a resource which is targeted at a specific usage, but to gather as much evidence as possible about a language for which there are not in general any other records, in particular any written documents or literature. As exemplified by the variety of projects within the DOBES initiative, field researchers are using lexical resources as a means to combine various levels of linguistic description, ranging from the simple usage of the word in concrete utterances to the combination of multimedia illustration of the actual meaning of the word in situational contexts. As an example, Figure 1 shows how the entry for the word “tpile wee,” which represents both an insect and by analogy a sung and dance performance depicting it, includes a video showing such a performance.

Why standardise lexical structures?

One could argue of course, that for the creator of a lexical resource – a lexicographer or linguist – a specific combination of descriptive information reflects a trade-off between the available information (e.g. field observation, extraction of lexical information from a textual corpus) and the intention to provide comprehensive coverage of a given subset of linguistic features. Depending on the actual objectives, the designer of a lexical resource may, for instance, focus essentially on grammatical descriptions of words, on the precise identification of senses or on the provision of testimonial examples for a corpus of literary work. By defini-

tion, such a combination of descriptive features is biased towards a certain aim, but also reflects the fact that it actually makes no sense to think of a lexical resource which would cover all aspects of linguistic description. A dictionary is the mirror of the knowledge a person, a school of thought or more widely a society has on a given language.

For this very reason, however, a given lexical resource will share with other similar, or less similar, resources some or many descriptive features, which allow one to relate and compare the available information across them. In the simplest case, where dictionary structures associate linguistic descriptions to lexical entries characterised by headwords, for instance to describe the various senses of a word, one can typically compare the actual orthographic forms provided for each entry, the possible grammatical features associated to each word or more specifically the domains of usage of a word. Such features, shared across several dictionaries, may allow one to see these as a continuum of lexical descriptions, which one may decide to traverse according to his or her linguistic interests.

From a technical point view, this requires us to be able to provide some level of *interoperability* between the electronic representations (the computer formats) associated with these various forms of lexical description. By providing coherent lexical formats, one may indeed intend to achieve three complementary objectives:

- To allow the export and exchange of data to third parties. In this context, interoperable data reflects the capacity to query a lexical resource and above all to parse the information provided by an external resource;
- To favour the pooling of lexical information coming from different sources, in such a way that similar pieces of knowledge can actually be mapped or at least compared;
- To limit the duplication of software development works by fostering the reusability of lexical management and consultation tools.

From a scholarly point of view, an increase in interoperability across lexical resources may also have a great impact on the capacity that scholars can have to provide fair and accurate comparisons of their results. There are indeed many situations where the creation of a lexical resource is an important part of the actual scholarly work of a linguist. In such cases, assessing the quality of the work requires having the capacity to understand the organisational choices of a lexical resource and relate these to similar choices taken by others. Such a comparison may actually take place at two complementary levels. First, one must be capable of

assessing the general organisation of the lexical database, as well as the coherence and comprehensiveness of the structure of each lexical entry. Once this is achieved, the assessment can then focus on the relevance of the elementary linguistic descriptions associated with a given entry. Without some elementary interoperability principles between lexical representations, such an assessment of lexicographic work can only be performed manually.

But is it at all feasible?

In the most ambitious sense, it is clear that achieving full interoperability across lexical resources is just an impossible goal. This may be understood intuitively from the variety of lexical configurations that we identified so far, but also, when one takes a more precise look at it, from potential variations at various levels of representation of a lexical resource.

First, there are various possibilities in organising lexical content that may not necessarily be compatible with one another. Seen from the greatest possible distance, the main conceptual divide in lexicographic work reflects whether the word or the meaning is the actual origin of the linguistic description. A. Zauner named this difference for the first time in 1902¹² when he identified two modes of lexicographic description:

- a semasiological view of a lexicon, that associates meanings to words, as can be encountered in most customary print dictionaries. Such an organisation is usually intended for wide-coverage lexica, whether monolingual or multilingual.
- an onomasiological view of the lexicon, whereby words or expressions are associated with meanings, which usually correspond to a list of concepts relevant for a specific field. This approach has been the core of the 20th century school of terminology initiated by E. Wüster,¹³ and is particularly suited to the description of the lexicon in specific technical fields.

Even if we remain in the semasiological tradition of dictionary making, lexical entries themselves can vary a great deal in the way they are actually structured. Just to take a simple example, a lexicographer may consider a variation in lexical category (e.g. the word “cut” in English seen as a noun or verb) as strong enough a marker to justify the creation of separate entries, whereas others may just use such information to indicate more precise sense or usage variation within an entry. The same elementary

piece of information (what we will refer to later as a *data category*) may thus be considered part of the organisational principles of a dictionary or simply a qualifier for a specific descriptive level.

Going deeper in the organisation of a lexical entry, we can also observe the variety of possible content values associated to specific descriptors, which can vary extensively in precision and complexity. Grammatical constructs associated with a verb can thus range from very simple opposition (transitive vs. intransitive verbs) in basic dictionaries down to complex representations of syntactic structures, for instance motivated by teaching purposes¹⁴ or adherence to a linguistic theory.¹⁵ This may also be true for very simple features such as pronunciation, where the lexicographer may want to associate precise rhythmic and prosodic information as opposed to a basic phonetic representation.

Finally, one major issue that may provide further food for thought on the complexity of lexical standardisation is again the scholarly perspective. By definition, scholarly work cannot just be based on an existing fixed set of structures and categories. It should be possible at any time for a linguist to contemplate the introduction of a new category whenever he has the feeling that he has observed a linguistic phenomenon that does not match the doxa of lexicographic principles.

Basic principles that should lead standardisation in the lexical domain

In the preceding sections, we have seen the reasons for standardising lexical structures, as well as the underlying principles that should provide the basis for such standardisation activity. From a little distance, we can also identify the conditions under which we would want to see standards to emerge within the humanities community.

First, from a purely scientific and technical point of view, we need to acknowledge that standardisation is a process that should fulfil two opposite and at times contradictory purposes:

- it should be a compendium of stabilised knowledge, which documents existing practices, so that future users recognise the existing culture. In this respect, it should cover the different variations in lexical structures that we have tried to identify in this paper;
- it should also look ahead for potential roadblocks so that future, and still unforeseen, forms of lexical structures could be taken into account by means of the appropriate generalisation.

Finding the best compromise between these two constraints, as well as putting together the best experts who would both be able to have in-depth knowledge of existing practices and a vision of where the future of digital lexica could be, are probably the greatest challenges that we could face in our endeavour to design optimal standards.

From an organisational point of view, we must also consider the basic constraints that affect all standards in all fields. In this domain, identifying “real” standards boils down to the following three aspects:

- as anticipated, a standard should be the result of a *consensus* building process. As a result, a standard proposal can by no means be self-claimed by an individual or a group without a prior check that it enjoys acceptance from a wider community. In general, in particular in a mature organisation such as ISO or the TEI, standards are iteratively designed in order to fulfil this criterion;
- a standard should also be *available* in the long term, so that a stable reference to it can always be made and consequently provide a stable background for many possible applications;
- finally, a standard should be *maintained*, so that technology changes or progress within a community can be systematically incorporated into new versions of the document. This maintenance is usually based upon either a systematic review (as carried out within ISO) or a feedback mechanism, reporting bugs or requiring new features for the standardisation body (as in the TEI).

These three constraints comprise a core characterisation of the kind of services a standardisation body should offer and *a contrario* what proprietary initiatives should be aimed at when wishing to disseminate a given specification as an international standard.

Can scientists adopt standards?

Even though we have managed to provide quite an extensive account on why standards are essential for scholarly work in (at least) the domain of linguistics, tackling at times the actual difficulties that such an endeavour may be facing, it may also be interesting to identify why scholars can be reluctant to adopt standards as a core component of their lexicographic work. In the following section, which has to be read with the appropriate distance, we try to identify good or less good reasons in this respect.

Freezing knowledge

Standards are basically associated with the notion of stability. By definition, a standard is a fixed set of constraints that one has to follow in order to be compliant with it. As such, it is understandable that standardising could be seen as an activity that is contradictory to the notion of scientific discovery, especially if the topics at hand are closely related to domains where scholarly work is very active. In such cases, we need to find ways to either show the complementarities between what is being standardised and what is the object of research, or even, provide means to continuously integrate new concepts in the process of updating standards.

Wasting time

Complying with a standard, not to mention contributing to standards development, may be regarded as a waste of a precious time that would be otherwise dedicated to research proper. When, in the course of a specific research project, one has to produce some digital primary sources, or an electronic version of a lexical resource, it is often seen as much more efficient to just improvise a self-defined computer format, then spend time in reading a documentation that may, all in all, not even match the precise needs of the research. After all, scholars know exactly what they need, don't they?

Forcing those with divergent views to agree

One of the most complex issues in standard development is to find an optimal consensus between technical views that may initially differ a great deal. This is all the more sensible in the humanities, where in particular the names and precise definitions given to concepts are strongly related to specific schools of thought. For instance, there is hardly any agreement on how to name the arguments of a verb, since it is strongly related to the understanding that a scholar may have of the corresponding semantic relations.¹⁶ This aspect of standardisation is all the more problematic when a specific scholar or school of thought that feels excluded from the standardisation process may, as a matter of principle, reject the resulting proposal. Anyone coordinating a standardisation activity within the humanities should definitely keep this in mind and make sure that a) the

group of experts involved in the definition of the standard is representative of the corresponding community and b) there are mechanisms to account for various schools of thoughts within the standard.

Forcing people to make data accessible to others

Even if the argument is not explicitly made, experience proves that non-compliance with standards is often seen as a way to ensure that scholarly data can hardly be used without involving the initial producer. There is always a fear that giving up to others may lead someone to publish results even faster than those who have provided these data. On the contrary, standardisation can be seen as a facilitating factor to get data samples easily from a given producer to another party interested in it. In the humanities, as in any other kind of science, this is an essential factor behind progress.

A crash-course in XML

Before we go any further with this paper, it is necessary to provide some background technical knowledge to our reader. Indeed, we have been speaking a lot in the preceding sections about lexical structures and formats without actually hinting at the way such structures could concretely be represented in a computer. Besides, it just happens that in the current web-based technological context a representation language has taken the lead for all kind of data interchange on the Internet, namely XML, the eXtended Markup Language, and this language has been the basis for the definition of most standards that are applicable in the linguistic domain. As a result, we will allow ourselves to convey here some minimal knowledge about XML, so that the reader may feel, if not at ease, at least minimally acquainted with the vocabulary we will be using later on in the paper.

XML is about trees

XML was designed to account for the presentation of tree structures. From a computational point of view such a tree structure (cf. a simple example in figure 4) is characterised by one single *root* node (here

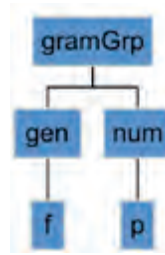


Figure 4. A simple tree structure (grammatical information in TEI).

gramGrp) and each node other than the root node has one and only one *parent node*.

XML is about angle brackets

The very simple structure delineated above can be represented (or *serialised*) in XML as a sequence of characters organised as *elements*, which can be isomorphically (i.e. one-to-one) mapped onto nodes of a tree. An element is identified by means of an *opening tag* (e.g. *<gen>*) and a *closing tag* (e.g. *</gen>*). To reflect on the hierarchical structure of a tree, element may be embedded one within the other (the element *gen* is embedded in the element *gramGrp*) but can in no way overlap. The following XML excerpt serialises the tree presented in the preceding section:

```

<gramGrp>
  <gen>f</gen>
  <num>p</num>
</gramGrp>

```

Issues in using XML for data modelling

In the XML context, modelling data structures in general resides in carrying out two complementary activities:

- Defining the vocabulary of tags that are allowed to describe such a structure, as well as the rules for combining them, i.e. which element is allowed within which context and how many times;
- Providing the actual semantics associated with each element, in order to ascertain for what purpose a given element may be used and under which conditions.

For instance, in the above example, one should be able to express that the <gen> element should occur within the <gramGrp> at most once and that indeed, <gen> indicates the grammatical gender within a group (<gramGrp>) of grammatical features attached to a lexical entry.

It is essential that such syntactic and semantic constraints, even if instantiated within the specific technical framework of the XML language, be accessible to scholars, since at the end of the day, they should precisely reflect the conceptual structures that the scholar has identified as relevant for his field of knowledge.

From theory to practice – standardisation in the open sea

The theory so far is nice and actually covers all aspects that anyone might have in mind in a similar context. However, all this would probably be completely useless if we did not have the prospect of actually having real standards being designed. Since there are indeed some, we will now see how such standards are actually designed in the light of their specific standardisation context, since we will address, in the coming sections, the work done by two complementary institutions:

- The Text Encoding Initiative (TEI), which has been put together within the humanities themselves and attempts to provide ready-to-use formats for the representation of digital texts;
- The International Organisation for Standardisation (ISO), which is an international consortium covering all technical domains and which has recently been pursuing work on the standardisation of language resources, in its committee TC 37/SC 4.

By describing how both initiatives have managed to deal with the lexical issue, we will try to both identify which kind of purpose they thus aim at and see how much, though being complementary in nature, they should both be seen as modelling tools for scholars.

Modelling lexical structures with the TEI

The Text Encoding Initiative was initiated in 1987 when a group of textual databases worldwide decided to join forces and define guidelines for the representation and interchange of textual documents. As the year of their first meeting immediately followed that of the publication of

the ISO SGML standards (which was to become XML 15 years later), the consortium unanimously agreed to adopt it as background for the definition of their own recommendations.¹⁷

One of the main characteristic of the TEI infrastructure is that, while being an XML application, it is not intended to be used on-the-shelf as a monolithic group of XML objects. Indeed, the TEI is by design an environment that a user needs to adapt (customise) to his/her own needs when using the TEI for a specific type of documents.

The customisability of the TEI results from its general information architecture that is based on the following core concepts:

- the TEI guidelines are based on a series of *modules*, each representing either generic technical components (core elements, header, encoding of names) or domain-specific subsets (for the encoding of specific genres or document types, e.g. drama, manuscripts or the transcription of spoken data). The module for “print dictionaries” is one of the latter type;
- most of the elements are attached to classes which group together those with similar semantics or having the same structural behaviour (e.g. appearing as children of the same elements).

The organisation of the TEI ontology in modules and classes is particularly important to allow the appropriate selection of descriptive elements that may fit a particular purpose. In the following section, we will see some of the concrete possibilities through the presentation of a simple example.

Matching the TEI “standard” with scholarly needs

Let us consider a scholar who would want to describe a simple dictionary using the TEI infrastructure. He/she will first select the basic modules of the TEI, allowing him to have the generic structure of a TEI document comprising metadata (the TEI header), together with the “dictionary” module, that will provide him with the various elements he/she may need for representing a lexical entry. By doing so, he/she will immediately be able to set up an environment where he can write constructs such as the one represented in Figure 5. What the TEI provides him/her with at this stage is the following:

- a group of elements to represent lexical features within a dictionary entry, for instance <pos> for representing the grammatical category (“part of speech”) of a word;

```

<entry>
  <form>
    <orth>table</orth>
  </form>
  <gramGrp>
    <pos>n.</pos>
    <gen>f.</gen>
  </gramGrp>
  <def>Pièce de mobilier...</def>
  <cit>
    <quote>Une table de cuisine</quote>
  </cit>
</entry>

```

Figure 5. A simple (constructed) dictionary entry represented according to the TEI guidelines.

- the syntactic constraints bearing upon these XML elements, by means of an XML schema (DTD, RelaxNG or W3C schema);
- the corresponding semantics, expressed by means of a precise online documentation, combining a comprehensive description of the way dictionaries may be represented,¹⁸ and a specific documentation for each element.¹⁹

Once such a first representation is achieved, our scholar will then want to refine his/her representation in various ways and make sure that these constraints are reflected in the schemas and documentation he/she will give to his/her students to produce further lexical entries. Let us see how the TEI mechanisms would allow him/her to actually implement three types of constraints.

Selecting appropriate descriptors

As seen in the various examples that we presented earlier in this paper, the actual combination of possible features for the various components of a lexical entry is an essential design aspect of a lexical database, in particular when such features are not at all relevant for a given language. Typically, a human-oriented dictionary will limit grammatical information to a set of very basic features related to the provision of the grammatical category (<pos> – *part of speech* element in the TEI)

and possibly grammatical gender (<gen> – *gender* element) for nouns and transitivity (<subc> – *subcategorisation*) for verbs. Through its class system, such de-selection is a core mechanism of the TEI infrastructure, since each element is individually connected to a class. For instance, all grammatical descriptors are members of a single class (named “model.gramPart”) and a schema tuned for a specific project could actually limit the content of this class to <pos>, <gen> and <subc>, whereas other elements available in the TEI vocabulary (e.g. <mood>, for expressing the grammatical mood of verbs) would be disallowed.

Adding one’s own categories

In addition to this first customisation possibility, the TEI allows one to define its own extension to the existing elements. This is particularly needed in a generic framework such as the TEI to account for descriptive features that a scholar may want to express, but which are not consensual enough – or specific to a group of languages that have not been represented in the standardisation process – to be integrated into the standard. For instance, a lexicographer who might want to represent a language that bears politeness markers on inflected forms of verbs (e.g. Japanese or Korean), could document an additional element named <politeness> and make it a member of the class of elements forming part of a grammatical description (<gramGrp>).

Constraining possible values

In many projects, it is important to set constraints on the possible values of a descriptive element, so that, for instance, the various editors involved in a dictionary project do not provide a grammatical gender as a random set of strings such as: “f,” “fem,” “feminine,” etc. To this end, the TEI infrastructure offers a “change” mode in the specification of a customised schema, so that an element such as <gen> (*gender*) keeps all its characteristics, except what the user has explicitly modified, for instance setting the possible values to the set {m,f}.

Overview

The TEI infrastructure is set in such a way that by default a user has a usable infrastructure and can, with little or no extra effort, use it directly to encode its digital data. Still, the specificity of a dictionary project can be taken into account by customising the infrastructure and in this respect, the TEI offers simple mechanism not only to express constraints on the available XML objects, but also to benefit from a comprehensive documentation for this new schema. This in turn contributes to a better exchange of information about practices between scholars.

For some scholars, however, it may be seen as an oversight not to have a more abstract modelling tool at hand that would not be necessarily tied to a fixed XML vocabulary. We shall see in the following sections in what way ISO provides such an environment.

Lexical descriptions in the ISO context

The International Organisation for Standardisation (ISO)²⁰ is one of the major standardisation bodies worldwide. By and large, its coverage encompasses most technical and scientific domains and it can be seen from two complementary perspectives:

- from an administrative point of view, ISO is an association of national standardisation bodies each having an equal right to initiate, comment and approve a standard project;
- when looking at the technical content proper, ISO is organised in technical committees and sub-committees, which group together experts within a dedicated field where the standardisation work is to take place.

Following a whole series of internationally funded projects²¹ in the 1990s aiming at providing (pre-normative) guidelines for the encoding of language resources, various experts worldwide deemed it necessary to go a step beyond and initiated the creation of a new ISO committee (ISO/TC 37/SC 4) dedicated to language management issues. This committee, founded in 2002, quickly achieved success and put together a whole portfolio of standards and standard proposals covering most domains needed for achieving interoperable language resources.²²

The ISO “Lego” model of lexical structures

ISO committee TC 37/SC 4 started to work in 2003 on offering a standardisation of lexical structure in the context of the LMF (Lexical Markup Framework) project. Published in 2008, ISO standard 24613 exemplifies perfectly the modelling strategy developed for language resource modelling at large.^{23, 24}

The modelling framework adopted by ISO committee TC 37/SC 4 is inspired by information modelling principles developed within object oriented languages and implemented by the OMG in the UML specification language. The framework requires one to describe an informational structure on the basis of two complementary elements:

- a *metamodel* which represents the abstract organisation of information and informs current practice for a given type of information. Such a meta-model is described as a combination of *components*, representing elementary units of information;
- a selection of *data categories* that can be used to qualify each component of the metamodel, and which form the basis for instantiating a metamodel.

In its simplest form, the LMF metamodel can be depicted as in Figure 6, where a lexical database is seen as a combination of some metadata (*Global Information*) and a series of lexical entries. According to the semasiological view, each lexical entry is characterised by means of a form, which in turn may be associated with one or several meanings. As can be seen, meanings can be recursively embedded to form a full semantic description of a word. The diagram also shows that a lexical model can be further designed by adding lexical extensions to the core metamodel so that specific representations dedicated to syntactic or multilingual de-

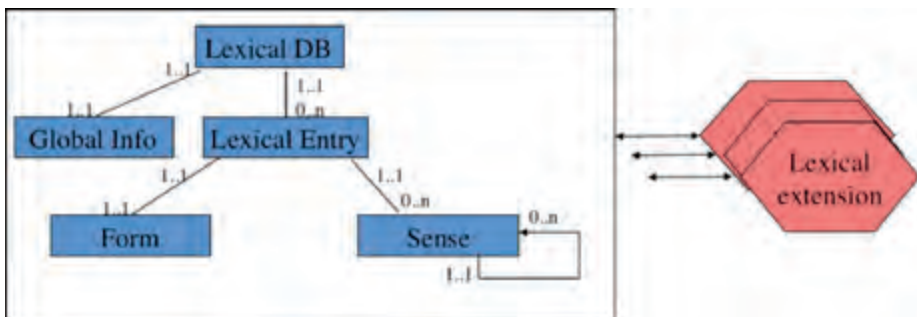


Figure 6. The LMF (ISO) metamodel.

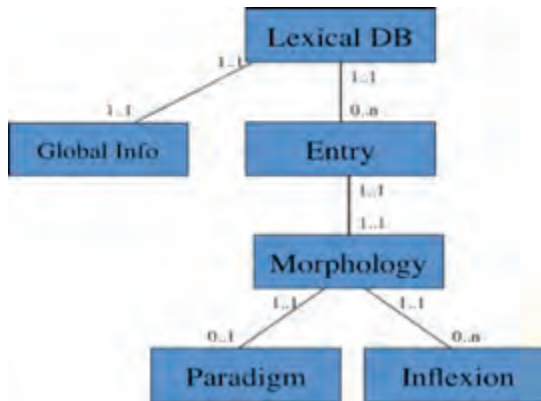


Figure 7. An LMF metamodel for full-form lexica.

scriptions can be added as needed to shape a more complex lexical metamodel. In order to have a full lexical model, such a metamodel must be combined with a selection of descriptors, also known as data categories.

To illustrate this modelling process, we can outline a possible model of full-form lexica that may be used to describe the inflected forms of a given language. To this purpose, we consider that each lexical entry groups together all possible inflected variations within a specific extension, whose entry point is the component *Morphology*. This component contains in turn an optional *Paradigm* component to characterise the inflection class of the lexical entry (for instance, first group of French verbs with “-er” ending) and as an *Inflexion* component, which can be iterated as many times as needed. The combination of these components with the core LMF components leads to a structure depicted in Figure 7.

This general structure, extending the core LMF metamodel, is in turn a metamodel for any language for which such a description could be applied. Still, if we actually want to make it a complete model for a full-form lexicon, we now need to “decorate” it by means of data categories corresponding to the characteristic of the language to be represented. This is illustrated in Figure 8, where a simple model has been outlined based on three groups of data categories, namely:

- the lexical entry is characterised by a lemma and a grammatical category (*part of speech* the paradigm is simply associated to a paradigm identifier;
- each inflected form is described by means of an actual written form (*word form*) and a series of grammatical characteristics for gender, number, person or tense.

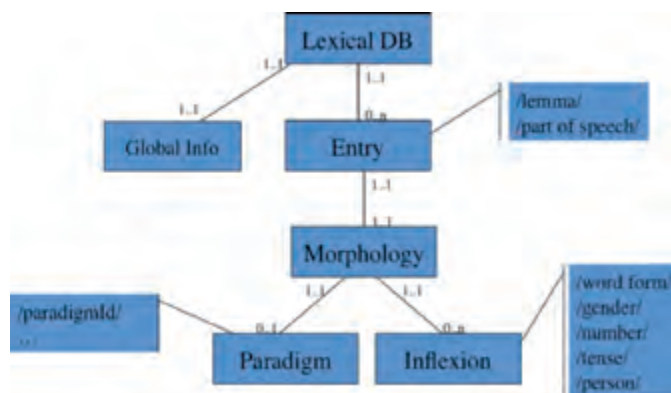


Figure 8. A completed LMF-compliant model for full-form lexica.

Even if this model is particularly simple, it is already sufficient to describe full-form lexica in quite a number of Western languages. According to the LMF principle such representations can be expressed in XML using any kind of vocabulary, under the condition that the XML abstract structure be isomorphic to the model outlined in Figure 8. For instance, Figure 9 shows the entry that would correspond to the word “chat” (*cat*) in French.

```

<lexicalEntry>
  <lemma>chat</lemma>
  <grammaticalCategory>noun</grammaticalCategory>
  <morphology>
    <paradigm>
      <paradigmIdentifier>fr-s-plural</paradigmIdentifier>
    </paradigm>
    <inflexion>
      <wordForm>chat</wordForm>
      <number>singular</number>
    </inflexion>
    <inflexion>
      <wordForm>chats</wordForm>
      <number>plural</number>
    </inflexion>
    ...
  </morphology>
</lexicalEntry>

```

Figure 9. Sample entry for a full-form lexicon.

Overview

As can be seen, ISO standard 24613 (LMF) provides a very powerful tool for anyone to design all kinds of lexical structures as needed in scholarly or commercial contexts. Its flexibility could indeed also be seen as a handicap since it allows one to create combinations of lexical information that may not necessarily make sense from a lexicographic point of view. Still, in the same way as the TEI does, the elicitation of core concepts and the corresponding methodology creates a joint culture across lexical projects so that they can further work together on joint (LMF-compliant) guidelines.

Another issue to consider here is, of course, that LMF does not provide a fixed XML serialisation, which prevents it from being used as a real interoperability standard. Each application or group of applications must identify what serialisation optimally represents the corresponding model – clearly, one could assume, in the perspective of further convergence across standardisation initiatives, in order to assess how much the TEI and its customisation facilities could play a role to this end.

Finally, so far we have been using the concept of data category without providing a precise account of their nature and role in the standardisation process. This will be the theme of the last part of this chapter.

Data categories as a conceptual market place

Data categories are elementary descriptors seen as abstractions upon the various possible implementation as a database field, an XML object, or whatever human- or machine-readable representations. For instance, a field linguist may define a series of simplified codes while transcribing and annotating some recordings on paper and in parallel associate such codes with reference data categories. This will allow him/her to document his/her data and make sure he can compare them with the work of others, or even be able to work again on his own observations in the future. The same applies if he has recorded part of his data in a computer file, whether word processor, spreadsheet or more elaborated database.

In order to fulfil the needs of describing basic feature-value representations, data categories may be of two basic sorts:

- *complex data categories*, corresponding to placeholders for a specific descriptors, such as /part of speech/, /grammatical gender/ or /author/;
- *simple data categories*, which represent elementary values such as /feminine/, /plural/, or /ablative/.

Complex data categories may naturally be constrained by providing either a generic data type (e.g. date, number, string, etc.) or a list of allowed simple data categories.

As such, data categories play two complementary and closely related roles:

- they are a tool to record and document the semantics of the concepts used by a scholar in the course of his data description activities;
- they provide means for an accurate specification of data formats, which are in keeping with the actual scientific concepts that the scholar has in mind with regards to his data.

This last point is particularly important to bear in mind, since it represents a fundamental basis for the actual interface between a scholar and the information technologist who may be in charge of providing him with the appropriate tools to fulfil his data-based research. A remarkable example of such a configuration is indeed the work done at the Max Planck Institute for Psycholinguistics in Nijmegen, where the Lexus tool allows the linguist to specify (in an LMF-compliant way) his lexical structure, before he actually starts entering his data in this lexical management environment.

To formalise further the notion of data category, we can refer to the standardisation work carried out within the database community to record the semantics of fields in a database model.²⁵ As depicted in Figure 10, ISO 11179 formalises the notions that we outlined at the beginning of this section as a two-level model:

- The first, abstract level organises information objects at *data element concepts*, which in turn may be characterised by means of a conceptual domain;
- The second level corresponds to the concrete instantiation of the first level within an information structure and is based upon *data elements*, which in turn may have values taken out of a *value domain*.



Figure 10. The ISO 11179 organisation.

The one difference between the ISO 11179 model and the notion of data categories described here is that *simple data categories* do not have real equivalents in ISO 11179. This is indeed a major addition, since elementary values such as /singular/ are core members of a linguistic ontology.

The limits of standardisation – the case of gender

Providing such an abstract model as the one presented above may not suffice to satisfy the representational needs of scholars working on languages. Indeed, it is probably the right time for us to stay away for a while from technological aspects and ponder upon the potential for standardising a given linguistic concept at all. As can be anticipated, there is already considerable debate, which we can only describe superficially here by providing some general lines of conflict as well as possible grounds for compromise in the context of a specific example.

The data category we will consider here is that of *grammatical gender*, which is probably one of the most widely used descriptors in lexicographic works. For instance, it is a core component of any *tagset*, that is, the list of descriptive tags used by computational linguists to annotate a text at the word level. As such, many implementers of such natural language processing systems have considered for many years²⁶ that providing a reference description for such a notion would be essential.

At the opposite, one could assert – the view prevailing among typologists – that no two single concepts are shared between two given languages, even within one language between two observations. The idea is that a descriptive feature may always be seen as the specific link between a language sample and the analysis of a phenomenon, thus forbidding any kind of generalisation.

To better situate where the problem is, let us consider the issue of /gender/ as a possible candidate for standardisation. The category has many interesting features as an illustrative example. It is intuitive enough for most languages, is widely used in many lexicographic applications and – as we shall see – has the appropriate complexity to help us understand where standardisation can find its place.

As defined by G. Corbett,²⁷ grammatical gender is a purely morphosyntactic concept seen as “a classification of nominals, as shown by agreement.” Such an agreement is usually related to the association of determiners with nouns, as can be observed in the distinction in German between “die Katze” (*the cat*; feminine) and “der Hund” (*the dog*;

masculine). Such an agreement can then propagate within a sentence to various linguistic components such as adjectives, numerals or verbs.

At least one thing is clear at this stage: the notion of gender does not match at all that of natural sex, and any further analysis should be strictly based on linguistic grounds and observations. This point definitely rules out, when contemplating standardisation, any further use of generic database-oriented standards such as ISO 5218,²⁸ which is intended to represent human sex in computer applications.

The definition of gender in relation to agreement can also be expressed over wider distances of texts when separate pronouns²⁹ may be used anaphorically in relation to a preceding noun phrase of a given gender. This may lead to elaborate pronoun systems as illustrated in Table 1 for Rif Berber, where gender distinction apply to both second and third person pronouns both with singular and plural genders.

1sg	nəʃ	1pl	nəʃnin
2sg.m	ʃək	2pl.m	kəniw
2sg.f	ʃəm	2pl.f	kənint
3sg.m	nətta	3pl.m	nitnin
3sg.f	nəttæθ	3pl.f	nitənti

Table 1. Personal pronoun system in Rif Berber [McClelland 2000: 27].

Another issue is that both the notion of gender itself, as well as the number of possible values for gender, varies dramatically from a language to another. Whereas gender agreement does not exist at all in some languages like Japanese (and within a very limited scope in English), some languages appear to have a complex gender system, sometimes with more than 20 values.³⁰ Such an observation in itself could jeopardise any attempt to standardise the notion of grammatical gender at all, but before giving up, we may want to consider the issue further.

Indeed, we would like to make two complementary considerations as to what “standardising gender” would mean. First, whether or not it applies to a specific language, gender as a grammatical concept is widely shared among scholars or engineers working on languages at large, whether they describe particular languages or implement tools for analysing them. As such, it may be important to provide them with a fixed point to which they could systematically refer, when they want to make sure that colleagues, or other software systems, will understand their data in the same way. For instance, this would facilitate cross-language studies, the definition of generic query systems for linguistic corpora

or the design of similar presentation modes for online language learning environments. From this point of view, “grammatical gender” as such, but also probably elementary values for gender, could be given standardised identifiers, together with some generic, if possible language-independent, definitions. Moreover, the variations across languages as to how and under which conditions grammatical gender applies, could and probably should be recorded in order to a) refine the generic definitions provided for the concept at large and b) to provide further constraint, through the precise listing of applicable values. Of course, this second level leads to a potentially complex attempt at precisely describing the languages of the world and one must provide the means to see this as a long-term endeavour.

To summarise, the example of grammatical gender can be seen as a prototypical case of the complexity of standardising linguistic concepts in general. Still, we do think that the endeavour is manageable if we offer a standardisation framework implementing a good compromise between linguistic genericity and linguistic accuracy. Such a framework must allow both engineers and scholars to feel at ease by finding there both stable points of reference and trustful linguistic content.

The ISO data category model

In the context of its work on modelling various types of language resources, ISO technical committee 37 has designed a specific framework for recording, documenting and standardising data categories. Among the design principles of the *data category registry*,³¹ we can mention here the following guiding ideas:

- providing an open space of reference concepts allowing linguists and developers to relate his own practice with standardised definitions and identifiers;
- providing extensive multilingual support so that both the variation of the semantics of a category across languages and the recording of terms used to refer to them would be precisely taken into account;
- complying with existing standards and practices such as ISO 11179 for metadata registries (used in the field of computer databases) or the W3C OWL recommendation (for representing ontologies).

In order to provide the best compromise to these constraints, a two-level model was introduced, as depicted through the specific example of

<i>Entry Identifier:</i>	grammatical gender		
<i>Profile:</i>	morpho-syntax		
<i>Definition (fr):</i>	Catégorie grammaticale reposant, selon les langues et les systèmes, sur la distinction naturelle entre les sexes ou sur des critères formels (Source: TLFi)		
<i>Definition (en):</i>	Grammatical category... (Source: TLFi (Trad.))		
<i>Conceptual Domain:</i>	{/feminine/, /masculine/, /neuter/}		
<i>Object Language: fr</i>	<i>Object Language: en</i>	<i>Object Language: de</i>	
<i>Name: genre</i>	<i>Name: gender</i>	<i>Name: Geschlecht</i>	
<i>Conceptual Domain: {/feminine/, /masculine/}</i>	<i>Name: grammatical gender</i>	<i>Name: Genus</i>	
		<i>Conceptual Domain: {/feminine/, /masculine/, /neuter/}</i>	

Figure 11. The grammatical gender data category (constructed example).

/grammatical gender/ in Figure 11. The first level of representation, which is also the entry point for the category, provides a unique identifier for persistent reference, a generic definition (in one or several languages) for this category, one or several profiles (the possible domains of application), and, when applicable, a list of possible values for the data categories. These values, recorded in the so-called *conceptual domain*³² of the data category, are the set of all recorded values, independently of language-specific constraints.

The second level of representation is dedicated to language-specific information and may be iterated as many times as there is a need for it. Two essential types of information may be recorded there: first, it is the place where any refinement of the semantics of the data category for the corresponding language can be traced, for instance by means of a more precise definition, or as is the case in the example, by indicating a subset of the possible values applicable for that language (grammatical gender in French can only take two values, namely, masculine and feminine). Second, the actual terms that may be used in the language at hand to refer to the data category are listed, to facilitate search and/or display of the entry.

As can be seen, the model designed by ISO (and implemented within the ISOCat platform)³³ is inherently semasiological, since it places the data category, as concept, at the centre of information recorded by the

registry. Even more, it goes beyond the traditional notion of semasiology, which is limited to lexical description, to encompass the provision of multilingual semantic constraints for a concept.

Standardisation as a component of the scholarly process

As could be induced from the evangelising tone of this paper, we are only at the start of a process where standardisation becomes a natural dimension for scholarly work in the humanities. Whereas specific sub-communities have since long integrated this component, typically for annotated text representation within the Text Encoding Initiative, it remains a non-consensual issue for many scholars whose perspective is not to immediately be able to interchange data with colleagues. Still, as depicted in Figure 13, we want to conclude by arguing that the standardisation process belongs to the knowledge discovery endeavour. Thereby, at some specific stages, a scientific community identifies and references what in its everyday practices may be seen as stable knowledge that could be widely disseminated as standardised concepts. Such standardised concepts, as they are extensively used and implemented outside the initiating scholarly context, may be put to the test as well as used within the scholarly world itself as comparison points for further analyses or experiments.



Figure 13. Relating standards creation with the scholarly process.

Acknowledgements

The author wishes to address his warmest thanks to the various colleagues who have provided him with additional illustrative examples or have taken the time to comment on this paper. Specific thoughts are expressed to Werner Wegstein, Hélène Manuelian and Marc Kemps-Snijders.

References

- 1 Here we are using the word science in a broad sense, i.e. encompassing natural as well as social and human sciences.
- 2 See, for instance, the guidelines of the Modern Language Association (<http://www.mla.org>)
- 3 See “Riding the Wave: How Europe can gain from the raising tide of scientific data,” Report of the High-Level Group on Scientific Data, John Wood (ed.).
- 4 See, for instance, the prosopography portal http://prosopography.modhist.ox.ac.uk/_at at the University of Oxford.
- 5 An online access to the corresponding index is available under <http://www.corpusthomisticum.org/it/index.age>
- 6 LDI (Lexiques, Dictionnaires, Informatique) UMR 7187, <http://www-ldi.univ-paris13.fr/>
- 7 H. Manuélian, “Une nouvelle ressource lexicographique en ligne: le Petit Larousse Illustré de 1905,” *Euralex 2010, Leeuwarden, Pays Bas*.
- 8 Werner Wegstein, Mirjam Blümm, Dietmar Seipel, Christian Schneiker: Digitalisierung von Primärquellen für die TextGrid-Umgebung: Modellfall Campe-Wörterbuch. TextGrid-Report, 4.1, Version: 12. Oktober 2009 – 1.0 http://www.textgrid.de/fileadmin/TextGrid/reports/TextGrid_R4_1.pdf
- 9 <http://franqus.ca/projet/>
- 10 A *corpus* is a collection of written or spoken documents that have been chosen or identified as representative of a linguistic behaviour and which is used to study this behaviour.
- 11 <http://nl.ijs.si/ME/>
- 12 Adolf Zauner, “Die romanischen Namen der Körperteile: Eine onomasiologische Studie,” *Zeitschrift für romanische Philologie (ZrP)* (1902) Vol. 27, Issue 527, pp 40–182.
- 13 Eugen Wüster, *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie* (Würzburg: Ergon, 1991).
- 14 Collins COBUILD English Language Dictionary, under the editorship of John Sinclair at the University of Birmingham.
- 15 Cf. the adoption of Mel’cuk linguistic theories in: Jean Binon, Serge Verlinde, Jan Van Dyck, & Ann Bertels, *Dictionnaire d’apprentissage du français des affaires* (Paris: Editions Didier, 2000).
- 16 Cf. ISO WD 24617-4, Language resource management – Semantic annotation framework – Part 5: Semantic Roles (SemAF-SRL), edited by Martha Palmer, University of Colorado, Boulder.

- 17 The work carried out in defining the TEI guidelines actually contributed to the emergence of XML, with TEI experts such as Michael Sperberg-McQueen, Steve DeRose or Henry Thompson taking the lead in defining many of its core concepts. [Lou, Seb Alan Renear]
- 18 <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html>
- 19 For instance, the documentation for the <pos> element under:
<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-pos.html>
- 20 <http://www.iso.org>
- 21 In particular EAGLES <http://www.ilc.cnr.it/EAGLES/home.html> initiated by Prof. Antonio Zampolli
- 22 See http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_tc_browse.htm?commid=297592&published=on&development=on for an overview of the programme.
- 23 L. Romary, S. Salmon-Alt, G. Francopoulo, “Standards going concrete: from LMF to Morphalou,” 20th International Conference on Computational Linguistics – COLING 2004 (2004) – <http://hal.inria.fr/inria-00121489/fr/>
- 24 L. Romary, “Standardization of the formal representation of lexical information for NLP,” *Dictionaries: An International Encyclopedia of Lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*, Mouton de Gruyter (ed.) (2010) <http://hal.archives-ouvertes.fr/hal-00436328/fr/>
- 25 In ISO/IEC joint committee JTC1/SC 32
- 26 See for instance the work done in the context of the Multext-East project: <http://nl.ijs.si/ME/>
- 27 Greville G. Corbett, “Number of genders,” *wals.info*
- 28 ISO 5218, Information technology – Codes for the representation of human sexes. This standard identifies four values, namely : 0 for ‘not known’, 1 for ‘male’, 2 for ‘female’ and 9 for ‘not applicable’
- 29 Anna Siewierska , “Gender Distinctions in Independent Personal Pronouns,” *wals.info* <http://wals.info/feature/44>
- 30 Greville G. Corbett, “Number of genders,” *wals.info*.
- 31 Ide Nancy & Laurent Romary, “A Registry of Standard Data Categories for Linguistic Annotation,” 4th *International Conference on Language Resources and Evaluation – LREC’04* (Lisbon, 2004), pp. 135-138.
- 32 As defined in ISO standard 11179 part 3.
- 33 <http://www.isocat.org>

CORPORA IN LINGUISTICS: SAMPLING AND ANNOTATION¹

Anke Lüdeling

1. Introduction

Gregory Crane said at the beginning of his talk at the Nobel symposium ‘Going Digital’ that in linguistics, electronic corpora can be used in three ways: (a) they can help answer old research questions faster and more accurately, (b) they make it possible to formulate and answer new research questions through new ways of looking at and analysing textual data, and (c) they make the data widely available. This paper deals with two aspects of corpora that pertain to all three issues: sampling and annotation. My focus is methodological – I want to describe *how* electronic corpora – if designed and annotated transparently – have contributed to and changed linguistic research.² I want to start with the following statement by Hermann Moisl:

Data is ontologically different from the world. The world is as it is; data is an interpretation of it for the purpose of scientific study. The weather is not the meteorologist’s Data – measurements of such things as air temperature are. A text corpus is not the linguist’s data – measurements of such things as average sentence length are.³

What Moisl says is that ‘the world’ must be interpreted. The ‘world’ for many linguists is a given linguistic variety, be it Modern German, Old English or the language of teenagers in a certain suburb of Stockholm that they want to analyse. One possible way to interpret the ‘world’ – here the linguistic variety under question – is to collect a corpus of that variety, according to specific design criteria – which then is a *sample* of the world. This corpus can then be interpreted further: Segments of the corpus such as words or phrases or sentences can be classified and analysed. The classes could be grammatical (in Moisl’s statement, for example, the strings <data>, <world>, <world>, <data>, <purpose>, <study>, ... could be classified as nouns), rhetorical (in Moisl’s statement we could classify the third sentence as exemplifying sentence; in a conversation we can classify the utterances as statements, questions, answers etc.) or anything else one might want to study. If such classification is coded directly in the corpus we speak of *annotation*.

Only if we take both sampling and annotation seriously and make all the decisions involved in them explicit we can perform rigorous experiments with reproducible results. Here lies the real advantage of electronic corpora: In electronic corpora we can docu-

ment every step of interpretation so that every user is able to see exactly how we arrived at the empirical base for our analyses. We can then use corpus data for exploratory studies in which we look for connections between variables, such as sentence type or part-of-speech category, or find categories, such as subordinate clause or verb,⁴ as well as for testing hypotheses that come from theoretical considerations.

This paper is organised into three sections. First I want to show how pre-electronic corpora were used and where the empirical problems lay. The following sections then correspond to the two steps of interpretation mentioned above, sampling (Section 4) and annotation (Section 5). Although this is mainly an overview paper, I shall illustrate each general point with specific results from a deeply annotated special corpus (a learner corpus called Falko) which is briefly introduced in Section 3.

2. Corpus linguistics: research questions and methods

Many of the research questions and methods in corpus linguistics today date back to pre-electronic times. Long before there were electronic corpora, people collected real-life examples (in contrast to made-up examples) to develop and illustrate grammars, define lexical entries, argue about different readings of a word etc. (this has often been described; for an overview see e.g. Meyer, 2008). In some sense it can be said that (at least part of) linguistics developed to preserve an earlier language stage (the Sanskrit Vedas, as described in the works of Panini in India or Homer's writings in Greece, cf. the 'Ungrammatical Words' by Aris-tonicus of Alexandria) that was passed on as a body of text – a corpus. A typical use of pre-electronic corpora is illustrated in example (1), from the German Neogrammarian Hermann Paul. An example taken from a real text (a quote, a sentence, a word in context) is used to illustrate a grammatical or lexicographic fact. Paul here writes about a word formation phenomenon that he considers marginal. (Don't worry: The phenomenon itself – the addition of the 'linking element' *s* between the two stems in certain compounds – is not crucial here; I am only concerned with the way the evidence is presented.)

Sometimes the linking element *s* appears after feminine nouns, although this is not yet common in written language, compare e.g. *Gemeindsversammlung* 'council + *s* + meeting' Hebel 452,24, *Huldszeichen* 'benevolence + *s* + sign' Heine 2, 111, *über Naturs*

Größe ‘about nature + s + greatness,’ *Sprachsverbesserer* ‘language + s + improver’ Leibniz, *Unvorgreifl. Ged.* 67,3, *Vernunftswahrheiten* ‘sensibility + s+ truths’ Le(ssing) 12, 434, 32...⁵

Here, as in many other similar grammars, examples from literary works are used to provide evidence (and thereby lend authority) for the claims made by the grammarian. I do not intend to demean the enormous achievements of grammarians like Paul, but in the light of current empirical standards in linguistics, quotations like these are methodologically problematic: Paul does not attempt to explain why in these cases (contrary to what he considers the correct written language) the linker *s* is being used. He does not give any context for his examples. He does not tell us whether the cited authors use the same words (or at least the same non-head words) without the linking element *s*. He does not even give us a definition of what he considers a word or tell us why *Naturs Größe* (which could be considered a phrase) is included in the list. In addition there are two problems with regard to the selection of the cited authors. First, Paul, like most of the other grammarians before (and often after) structuralism, cites only well-known ‘authoritative’ authors. He therefore describes a written literary register without acknowledging this. Second, Paul wants to write a grammar for German at the beginning of the 20th century. The lives of the cited authors span two hundred years (from Leibniz’ birth in 1646 to Heine’s death in 1856). Language has certainly changed in those 200 years (and the roughly 70 years between Heine’s death and Paul’s grammar). If we think back to the desiderata formulated in Section 1 (making sampling and annotation explicit) we see that Paul fails in both respects.

The issues are very similar for other areas of linguistics, for example for pre-electronic lexicography.⁶ Typically, lexicographers manually collected quotations that consisted of the word to be described and its context. The quotations were then sorted and analysed (there are impressive collections of paper slips in all lexicographic institutions). While it is amazing how much data some of the scholars processed and kept in mind, we cannot say that either sampling method or search was fully systematic.⁷ Inherently corpus-based areas like historical linguistics or language acquisition studies had their own issues with respect to availability, access and coverage, but also in these areas pre-electronic texts could not be systematically searched (manual searches often took months or years and additional research questions meant that everything had to be read again) and results could not easily be reproduced. The method –

producing a concordance of a key word with some context – is, however, still the research method most widely used in corpus studies today (only now the searches can be systematic).

The first electronic corpus was made in the 1940s⁸ but it took a long time – essentially until the 1980s or 1990s – until corpora became freely and widely available for linguistic research (computational linguistics used them earlier). For many languages and varieties there are still no electronic corpora. Many influential linguists in the 20th century did not use corpus data, and in many departments corpus construction and analysis were not on the agenda. This has changed considerably. More computing power, internet access, search tools and standardisation have also helped. In recent years we have seen both a tendency towards the development of increasingly larger corpora for linguistic and NLP purposes⁹ and a tendency towards many more small dedicated corpora (dialect corpora, historical corpora, spoken corpora, learner corpora, etc.) designed for very specific research questions.

3. The Falko Corpus and Overuse/Underuse Statistics

I want to illustrate my general points using one of these special corpora, the learner corpus Falko.¹⁰ Falko contains written essays from advanced learners of German as a foreign language as well as native speaker control data.¹¹ Learner data is studied in order to find acquisition patterns.¹² Learner data often contains errors and these can be indicative of the learners' hypothesis of their target language. The learner in Example (1), for instance, uses the wrong preposition, the verb *vorbereiten* 'to prepare' needs *auf* instead of *für*. Single errors such as this one are not interesting in themselves – they are merely anecdotal. But if many students make the same kind of error again and again this could tell us something about acquisition.

- (1) [...] *dass ihr Studium sie nicht für* (should be: *auf*) *die wirkliche Welt und ihre berufliche Zukunft vorbereitet* (fk_006_2006_08)
 '[...] that their course of studies does not prepare them for the real world and their future job.'

In a learner corpus we could now see whether the verb *vorbereiten* is often used with the wrong preposition. We could then see whether this is dependent on the learner's native language (if it is English there could be

a transfer from *prepare for*) or whether generally the polysemy of prepositions leads to these problems etc. What is important for us here is that we see the learner language that we collected as a sample for the learner language of a given population and that errors need to be found and classified. Many learner corpus studies concentrate on finding and classifying errors (error analysis, EA). Other studies compare the learner corpus quantitatively with another corpus – this can be a native speaker corpus or a different learner corpus (contrastive interlanguage analysis, CIA). In the following I will speak about a combination of EA and CIA.

Before I introduce the method used here, I need to say a few words about learner language.¹³ The language produced by learners is not random. It follows an internal production grammar, just like the language of native speakers does.¹⁴ This internal grammar is, of course, different from native speaker grammar. The corpus data is an expression of the internal grammar of the learners and can be used to discover the internal grammar of the learners. One source of evidence that points to the interlanguage is the kinds of errors that learners make – I will say more on this in Section 5.2.

One other source of evidence is quantitative. Everybody knows the phenomenon: Learner varieties – such as, for example, scientific papers written in English by non-English speakers like myself – are often recognisable as non-native even if they do not contain any outright errors – they are too ‘formal’ in some situations or too ‘informal’ or just ‘strange’. This perception is based on quantitative differences between learner varieties and native varieties; the learners use certain words, phrases, sentence types or other constructions not in the same distribution that native speakers would in that situation: Our learner English does not contain enough variety in sentence types or perhaps too many passive sentences or too few adverbs etc. This perception implies an implicit quantitative comparison – it is not ungrammatical at any point in the article to use a passive sentence – native speakers would just not do it as often.

There are many quantitative methods for comparing learner language and native language (or any other two corpora, see Section 4). In this paper I want to concentrate on just one method, namely overuse/underuse statistics. The idea behind this is that learners of a language tend to (consciously or unconsciously) avoid certain constructions which, for some reason, are difficult for them – here we find an *underuse* when we compare the learner corpus to the native corpus – and compensate by using alternative constructions too often – here we find an *overuse*.

Overuse and underuse can show different things: It might be the case that the learner is not aware of the ‘native’ distribution of the construction under question because he or she has not had enough exposure to the variety. It might also be the case that a certain construction is underused because it is ‘difficult’.

Table 1 shows how underuse and overuse can be visualised.¹⁵ The table shows underuse and overuse of the most frequent words in the corpus. The German data (deu) is used as a reference. Falko contains data from learners with 48 different native languages (L1s) but sometimes the number of texts written by learners of a given language is too small to produce significant results. Here, we therefore only look at the five largest learner groups (with L1s Danish, English, French, Russian and Uzbek). The colours are an easy way to see whether a given item is underused or overused. Cold colours signal underuse, warm colours signal overuse. The intensity of the colours signals the strength of the over- or under-use¹⁶. We see, for example, that the definite article *die* is overused by all learners and the conjunction *und* is underused by all learners. The other words do not behave uniformly across the learner groups.

word	deu	dan	eng	fra	rus	uzb
die (form of the definite article)	0,028296	0,03748	0,037161	0,040397	0,03881	0,035178
und "and"	0,023288	0,022804	0,022409	0,019756	0,020837	0,020638
der (form of the definite article)	0,019667	0,019585	0,021245	0,022803	0,024495	0,023218
es "it"	0,013827	0,010638	0,013025	0,011991	0,01352	0,004925
nicht (negation)	0,013695	0,013857	0,014414	0,012679	0,012725	0,015478
zu "to"	0,013608	0,012166	0,015352	0,015038	0,013679	0,009146
ist "is"	0,012615	0,011075	0,014189	0,014645	0,015588	0,011726
in "in"	0,012308	0,014239	0,014489	0,01553	0,015429	0,011257

Table 1. Visualisation of overuse and underuse for the most frequent words in the Falko corpus. The reference numbers are in the native German (deu) column; the other languages are Danish (dan), English (eng), French (fra), Russian (rus), and Uzbek (uzb). Overuse is signalled by warm colours, underuse by cold colours.

After this short introduction of the corpus and the overuse/underuse method, I will now come to the first general issue: sampling. Some of the general issues are then illustrated using the Falko data in Section 4.2.

4. Sampling – corpus design

Ideally, the research question determines the data to be used, so that all variables can be controlled. A corpus is almost always a sample of the linguistic variety to be analysed. Only in very few cases – when the entire population is finite, small and accessible – is it possible to work without sampling. If, for example, one wanted to analyse all of Nobel's letters to his mistress it could, in principle, be possible to take all of them (if they were available). But usually the linguistic variety to be analysed ('Modern English,' 'Modern British English,' 'the English of boys between 10 and 15 who live in Manchester,' ...) is too large to be completely analysed and one has to take a sample. Sampling, of course, depends on many parameters. One has to decide on a sampling algorithm, size etc. All of this plays a role in determining the possibility of generalising from observed data (the corpus) to unseen data (the variety in question). Here I can only speak about variation as one of the issues involved, but see e.g. Biber or Hunston for further thoughts on sampling.¹⁷

4.1 *Variation*

A speaker can express the same thought in many different ways. He can use active or passive voice, different lexical items, be more or less specific etc. It has long been recognised that external factors (such as modality, age/sex/education of producer, text type and probably many more that we have not yet identified) influence language production on all linguistic levels. There are also language-internal factors such as information structure or previous discourse. While some differences seem obvious (but might really be due to faulty observation)¹⁸ many such differences can only be observed and really tested by using corpus material (often together with elicited material).¹⁹ The idea behind all of these studies is that a speaker has several options (variants) to realise a variable. While there might be cases of genuine free variation, many areas of variation can be explained (or at least one can find co-varying variables), and these are the areas we want to detect.²⁰

Although there might be categorial differences between varieties (feature A appears in texts of type Y but never in texts of type Z), most of the differences are quantitative (feature A is more frequent in texts of type Y than in texts of type Z).²¹ There are, of course, many different ways to design experiments that test variation. Biber & Jones distinguish between two types of studies:²² Type A studies test whether and how the distributions of variants of the same variable in the same corpus might

be explained. Type B studies test differences between two corpora. There are simple descriptive methods but these are limited in scope. “Statistical inference is necessary because any sample from a language is subject to random variation.”²³ Statistical modelling and testing is necessary, especially because corpus studies have shown that the distributions of different categories differ widely and thus that not each statistical model can be applied to each type of corpus data.²⁴ But how much variation is there, and is it really important to study variation? Some people argue that variation is not interesting because some ‘core’ or ‘standard’ grammar always applies and it is only important to find this core grammar. However, without understanding variation it is not possible to understand language acquisition, language use and language change. The fact that speakers predictably produce different varieties shows that this awareness is part of the general knowledge of language. The differences are subtle but can reliably be detected. This has been shown in numerous studies, for example in the register studies by Douglas Biber and colleagues.²⁵ They use a multidimensional method where they annotate the different corpora on many linguistic levels. This creates a multidimensional space, which is very difficult to interpret. The dimensions are then reduced by calculating co-occurrences²⁶ and then interpreted functionally. Using this method, Biber is able to find linguistic differences within scientific articles – introductions can be reliably distinguished from the rest of the article, for example. There are many similar results:²⁷ people vary their phonology,²⁸ their word formation behaviour,²⁹ syntactic expressions,³⁰ etc. Fields that research variation include dialectology, sociolinguistics, pragmatics, historical linguistics, language acquisition and many more. All of these use corpus data in some way or other.

4.2 *Variation in Falko*

In the following I want to employ the overuse/underuse statistics introduced in Section 3 to show which effects variation can have in our learner corpus. Consider the Uzbek row in Table 2: all three cells show underuse. If there were no other corpora one could conclude that Uzbek speakers have problems with respect to the complementiser *dass* ‘that’ as well as with respect to the reflexive (and indeed there are many studies that use exactly this one-to-one comparison).³¹ However, when we look at the cells for the other languages we see that the situation is completely different for the two cases. The complementiser is overused in all other subcorpora, the reflexive is underused in all other corpora as well. For the complementiser, we could assume that the L1 of the learners influences their acquisition

behaviour.³² For the reflexive we could assume that there is a problem in the target language that is independent of the L1 of the learner (as indeed the different L1s of the learners use the reflexive in completely different ways). The underuse of *Frauen* ‘women’ has yet a different reason. The learners were free to choose between four topics for their essays, one of them about feminism, and the Uzbekian learners happened to like the other topics better and simply didn’t write about women.

word	deu	dan	eng	fra	rus	uzb
sich [reflexive]	0,012294	0,005892	0,005255	0,006389	0,004613	0,00469
dass “that”	0,007709	0,012602	0,009797	0,008748	0,011293	0,004221
Frauen “women”	0,003051	0,006438	0,008408	0,005111	0,006362	0,001173

Table 2. Overuse/underuse of *sich*, *dass* and *Frauen* for five Falko subcorpora.

The overuse/underuse tables can be used as a diagnostic to find interesting cases. The real linguistic analysis still has to follow.

While it is not surprising that the native language of a learner influences the interlanguage, it might come as a surprise that the gender of a learner influences his/her linguistic behaviour. Table 3 shows the differ-

Word	L1_f	L1_m		L2_f	L2_m
die [form of definite article]	0,02847173	0,02773356		0,03873803	0,03793173
ist “is”	0,01309508	0,01108118		0,01296901	0,01558681
sich [reflexive]	0,01263493	0,01120362		0,00552113	0,00629799
das [form of definite article]	0,01090937	0,01138729		0,00895775	0,00963391
für “for”	0,00718983	0,00887719		0,00704225	0,00701694
Frauen “women”	0,0036045	0,00128566		0,00802254	0,00382481
wir “we”	0,00105451	0,00140811		0,00376338	0,00316337
um “in order to”	0,00375788	0,00177544		0,00307606	0,00350847
was “which, what”	0,00370037	0,00281621		0,00345915	0,00370977
aus “out (of)”	0,00325939	0,00367332		0,00161127	0,00135162
nach “after”	0,00262668	0,00330599		0,00305352	0,00238691

Table 3. Overuse/underuse comparison of some frequent words between men and women in Falko. L1_f: native German speakers female, L1_m: native German speakers male, L2_f: learners female, L2_m: learners male.

ences between native (L1) men and women as well as between L2 men and women. Because we have more essays written by women we chose the women as the ‘standard’ (this is, in essence, arbitrary). Again, there are clear differences in content words which, of course, reflect the topic and might be gender specific (unsurprisingly again *Frauen* ‘women’ is underused by all men), but more interesting are the grammatical words like prepositions etc. that are fairly independent of the topic matter. Again, two areas of comparison are interesting. First, one could look for those cases where all men differ from all women (like *die, das* ‘the’ or *Frauen*). Then one could look at those cases where learners differ from native speakers (most of the other cases in the table).

Again, it would be necessary to investigate further. What I wanted to show is that gender – just like native language – is a relevant parameter for this kind of corpus, even for seemingly content-independent grammatical items. The Falko L2 corpus has more than 2/3 female authors. So whatever we say about ‘the learners’ might be just a fact about female learners.

4.3 *Metadata and Subcorpora*

This poses a question: If so much variation exists, what is the status of a given corpus datum? The answer depends, of course, on the research question one wants to answer. But very much research in linguistics does not take variation into account as much as it should – there are still many opportunistic corpora or corpora unsuitable for a given research question. Often we see analyses and models without having access to the data. But how do we then know whether a given analysis is valid? If it seems plausible we assume that it is if it doesn’t seem plausible we assume that it isn’t. Those cases are in essence like the situation in Hermann Paul’s time.

Two points become very clear: (1) the data should always be provided with any analysis³³ and (2) the use of (standardised) metadata is important so that subcorpora can be constructed at any point in the analysis. Both issues rely on electronic corpora, common coding standards and strategies for data preservation (see Romary in this volume) and powerful search tools.

5. Annotation – interpretation of the data

Section 4 showed that the sample we chose for a linguistic study crucially influences the results. This section is concerned with further interpretation of the sample. It is impossible to use data without interpreting it in

some way. Even deciding which surface forms are to be analysed is a way of interpretation. Interpretation is category-building or abstraction and abstraction entails a necessary loss of information: One does not want to look at individual cases but at classes of cases.

5.1 *Necessary Loss of Information*

In corpus studies, interpretation happens at two levels – at the level of the primary data, and at the level of annotating the primary data. First, one has to decide on a reading or an analysis of the primary data which can be ambiguous syntactically, semantically etc. In many cases, linguists concur in the interpretation of a datum and do not notice the ambiguity.³⁴ This might have implications for an analysis. The issue is even more interesting for data that is not covered by established grammatical analyses (learner data, historical data, spoken data etc., see Section 5.2).

Here I want to focus on the second level of interpretation, namely annotation. In Hermann Paul's time – and very often today – a linguist used categories (e.g. feminine nouns with linker *s*) for his or her analysis without assigning them directly to the data – the researcher works in his own spreadsheet, file-card system or whatever. If it were completely clear which categories there are and how these were to be assigned, this would be unproblematic. But linguists almost never agree on a category. There are scores of articles on how to assign even the most basic part-of-speech categories (just try to decide on a definition of 'adverb'). This is where electronic corpora are crucial: In multi-layer corpus architectures (see Section 5.3) it is possible to assign any item to a pre-defined category so that the categories can be searched just like the primary data. This could be part-of-speech categories, syntactic information, co-reference information, semantic information, information about rhetorical structure, narrative structure or whatever else one wants to study. There are two steps involved: First one needs to decide on the relevant categories (the set of relevant categories is often called a *tagset*) and then one needs to decide on the guidelines for assigning each category. All the decisions involved in the assignment can be made explicit, so that a user of the corpus knows exactly what is in the corpus. And if she disagrees with the categories or the guidelines she can open up a new annotation layer and provide her own interpretation of the data.

There has been a lot of discussion about the status of annotation. Many corpus linguists want to work in a corpus-driven paradigm (following Sinclair, 1991) and some argue against annotation. But *every* qualitative analysis that goes beyond a single instance and every quantitative

endeavour rely on some kind of categorisation (be it only word forms) and some kind of linguistically motivated question. In 1991 multi-layer annotation was not possible and very often the only annotation layer was perceived as some kind of ‘true’ interpretation of the data. I believe that we can take Sinclair’s arguments against such annotation seriously but rather than *not* annotating I think we need to have possibly conflicting independent layers of annotation and thereby make our assumptions visible.³⁵

I want to exemplify why this is important again using the Falko data. I will come back to multi-layer corpus architectures in Section 5.3.

5.2 Falko: Target Hypotheses

In Section 3 I said that one possible way of studying learner’s inter-language is error analysis: Errors in the learner output are identified, classified and analysed. From these errors, one can hypothesise what kind of a ‘rule’ or ‘regularity’ the learner might have about the target language. This sounds straightforward but is extremely difficult to implement because it is unclear what an error might be.³⁶ I want to illustrate this with an example. The learner utterance in (2) contains orthographic, word formation, case and number errors – all of these are at first glance uncontroversial.

(2) *Die politiker die in Korruption aktivitäten sind wird im Gefängnis gehen.* (kne21_2006_09)

‘≈ The politicians that are involved in corrupt activities will go to prison.’

But even with the seemingly uncontroversial example we have to note that an error analysis is not possible without at least implicitly assuming a correct version of the sentence. And if we look more closely there are different possible ‘correct’ versions (which we call target hypotheses). This is illustrated in Table 4. Target hypothesis 1 is possible and very close to the learner utterance but target hypothesis 2 is probably a more idiomatic German sentence. It becomes immediately clear that the target hypothesis determines which errors are found. There are a number of errors common to both target hypotheses (capitalisation of *politiker* ‘politician,’ commas around the relative clause, compounding of *Korruptionsaktivitäten* ‘corrupt activities’) but there are also errors that concern only one of the target hypotheses (addition of verb in the relative clause and using a more idiomatic verb in the main clause for target

x utterance	gloss	target hypo-thesis 1	gloss of th1	errors th1	target hypo-thesis 2	gloss of th2	errors th2
Die	the	Die	The		Die	the	
politiker	politicians	Politiker	politicians	orth	Politiker	politicians	orth
		,	,	orth	,	,	orth
die	that	die	That		die	that	
in	in	in	In		in	in	
Korruption	corruption	Korruptions-aktivitäten	corrupt activities	word formation	Korruptions-aktivitäten	corrupt activities	word formation
aktivitäten	activities						
					verwickelt	involved	lexicon
sind	are	sind	Are		sind	are	
		,	,	orth	,	,	orth
wird	will (sing.)	werden	Will	number	werden	will	number
im	in+the (dative)	ins	into+the (accusative)	case	im	in+the (dative)	
Gefängnis	prison	Gefängnis	Prison		Gefängnis	prison	
gehen	go	gehen	Go		landen	end	lexicon

Table 4. A learner utterance and two different target hypotheses. The common errors are marked in blue, the errors that only refer to one of the hypotheses are marked in red.

hypothesis 2). These problems arise in almost every sentence and can have a potentially huge effect on the analysis (see Lüdeling, 2008 for an experiment on the effects of target hypotheses).

If the target hypothesis is so important for all further analysis and if there are potentially always several possible target hypotheses, it becomes clear that the target hypothesis should be coded as an annotation layer in the corpus. There can be no single “correct” version; every target hypothesis is an interpretation of the data. (This is true for every error analysis, independent of whether the target hypothesis is explicitly coded in the data or not.) Only if the hypotheses are annotated *in* the corpus does one always know what the basis for a given analysis is (and one can choose to define a different target hypothesis if necessary).

The Falko corpus is annotated with two different target hypotheses (the decisions behind them are described in Reznicek et al.).³⁷ The first one is minimal and only corrects grammatical errors even if the sentence does not make sense or does not fit into the narrative. The second target hypothesis looks at the sentence within the text and also corrects unidiomatic expressions or anaphoric errors.³⁸ This means that Falko can be used as an example in which we can show what effects different interpretations may have on the analysis. I want to exemplify this with an experimental question where we can use the same overuse/underuse method that we used before. Only now we do not calculate overuse or underuse in the primary data but in the annotations.

Above we saw that learner language differs quantitatively from native language. One reason is the errors: If, for example, many learners use incorrect prepositions (as in Example 1), there could be an underuse of more difficult or rarer prepositions. Since both target hypotheses are ‘correct’ German, this effect should be gone. But the target hypotheses are still close to the learner utterances. We could use the annotated data to see how close the target hypotheses are to the native speaker data. In other words: Are the target hypotheses more similar to the L1 corpus than the learner data? How strongly does the learner data ‘shine through’?³⁹ We could, of course, look at each word in the corpus separately. But for questions like these it is more interesting to look at classes of words. Table 5 shows a comparison of different inflectional categories of main verbs. Main verbs constitute an open class, which means that it is not easily possible to look at all items separately. We can therefore use the part-of-speech annotation. The ‘standard’ against which everything is compared is the German (deu) column. The ‘paler’ the other cells are the more similar to the original they are. If target hypothesis 2 is closer

pos	deu	dan	eng	fra	rus	uzb
VVFIN	0,045553	0,044081	0,046582	0,03735	0,048513	0,054409
VVINF	0,028383	0,025259	0,028415	0,029487	0,03038	0,033537
VVIZU	0,001898	0,000927	0,001051	0,001376	0,000954	0,000469
VVPP	0,020543	0,01713	0,015915	0,015333	0,013361	0,008443

Table 5a. Overuse/underuse table of main verbs in Falko, original learner utterances.

pos	deu	dan	eng	fra	rus	uzb
VVFIN	0,045553	0,045485	0,042596	0,0354	0,042345	0,026316
VVINF	0,028383	0,024555	0,029251	0,029326	0,025733	0,031579
VVIZU	0,001898	0,001154	0,001379	0,001466	0,001954	0,002632
VVPP	0,020543	0,017029	0,016398	0,017386	0,014984	0,015789

Table 5b. Overuse/underuse table of main verbs in Falko, target hypothesis 1.

pos	deu	dan	eng	fra	rus	uzb
VVFIN	0,045553	0,045829	0,042658	0,03617	0,043803	0,031008
VVINF	0,028383	0,024453	0,029999	0,03042	0,026282	0,036176
VVIZU	0,001898	0,001319	0,001527	0,001464	0,001622	0,005168
VVPP	0,020543	0,016266	0,016452	0,018085	0,016223	0,015504

Table 5c. Overuse/underuse table of main verbs in Falko, target hypothesis 2.

pos: part-of-speech category,⁴⁰ VVFIN: finite main verbs, VVINF: infinite main verbs, VVIZU: infinite main verbs with particle *zu*, VVPP: participles

to the L1 data than target hypothesis 1 and target hypothesis 1 is closer to the L1 data than the original data Table 5c should be paler than Table 5b and Table 5b should be paler than Table 5a.

The expectation is borne out, except for the Uzbek data. Looking at the Danish, English, French and Russian corpora we see that all main verb categories are underused. The underuse is less strong in the target hypotheses. The Uzbek corpus differs markedly. We would now have to analyse where these differences come from.

This study shows that different interpretations (here the different target hypotheses) lead to different results and that annotations strongly depend on the interpretation. Since it is not possible to choose the one ‘correct’ interpretation the only way out is to provide as much transparency as possible. This can only be done through annotation.

5.3 *Multi-layer Annotation*

I argued that one of the crucial improvements electronic corpora made for empirical work in linguistics is transparency. In this section I want to say just a few words about the technical background – multi-layer corpus architectures and standardised exchange formats – that makes this possible.

As mentioned above, many linguists do their analysis away from the corpus in separate files or spreadsheets instead of directly in the corpus. There are several reasons for this. First, many linguists are not aware of the usefulness of direct annotation. This will hopefully change with time. Second, many corpora are not available in a format which can be annotated by users, but only via e.g. a Web interface with limited context. This is an obstacle to research, many people have realised that and the situation is changing (again: provided that the legal and ethical issues can be solved).⁴¹ The third obstacle used to be technical – for a long time it was technically not possible to annotate the same primary data with different kinds of annotation, such as token-based annotation, syntactic annotation, pointing relations etc. But in the past ten years or so multi-layer annotation has developed⁴² and today there are free and accessible tools to annotate a file in many different ways. Many of these tools output a standardised and well-described format (usually in XML) and in recent years several integrative frameworks have been developed (see Romary in this volume). Using these, it is possible for researchers that are far from each other and have different theoretical notions and different research questions to annotate the same data. The annotation layers can be integrated and each annotation layer can be searched separately or annotation layers can be combined for the search.

6. Summary: Transparency and Availability

Corpus data is one of the empirical bases for linguistic research. It was used long before the electronic era, and many of the research questions and methods were there before computers came along. In this paper I illustrated how the possibilities for systematic sampling and transparent interpretation that electronic corpora provide have changed the understanding of linguistic work and empirical methods:

Corpora are always samples of a given variety. Through more systematic search and analysis tools – especially for quantitative

analysis – we are beginning to understand how much variation there is between different varieties of a language and how many external factors influence language use. Careful sampling techniques and especially the use of metadata allows us now to see exactly which factors are relevant.

Every linguistic analysis is an interpretation of the data. In electronic corpora it is possible to explicitly store the interpretation with the primary data so that it is visible to the user. If the data is stored in an appropriate format, many scholars can work in parallel on the same data.

Corpora are just data. The research questions must be formulated within a given theory or model using linguistic knowledge. The results of the analysis must be integrated into the theory or model. In between question and results, we need to choose the appropriate empirical bases and suitable research methods. Corpora are one type of data with possibilities and limitations, and many interesting qualitative and quantitative methods for corpus analysis have been developed. We have learned a lot about language from using corpora and there is still so much more that we do not yet understand. We also learned a lot about the empirical work and – again – there is so much more that we do not yet understand. But electronic corpora enable us to make every step of interpretation and analysis transparent and reproducible.

References

- 1 I want to thank the members of the corpus linguistics group at Humboldt University for many heated and fruitful discussions about corpus linguistics and no less for more direct contributions to this paper: Amir Zeldes and Florian Zipser gave valuable input to specific points, Hagen Hirschmann and Marc Reznicek read the pre-final draft and their constructive criticism and suggestions made this paper more coherent and accessible. Amir Zeldes and Marc Reznicek helped with the Falko data. I also want to thank the organisers and colleagues at the Nobel Symposium – especially Karl Grandin – for one of the most stimulating and interesting conferences I have ever attended.
- 2 For many years the dominant issue here was *whether* corpus data should be used to answer ‘interesting’ (often understood as ‘generative’) questions (cf. Charles Fillmore, “‘Corpus linguistics’ vs. ‘Computer-aided Armchair Linguistics,’” in *Directions in Corpus Linguistics*. Nobel Symposium 82, 4–8 August 1991. Berlin Mouton de Gruyter, 1992), pp. 35–60; Fred Karlsson, “Early generative linguistics and empirical methodology,” in *Corpus Linguistics: An International Handbook*. Vol. 1, Anke Lüdeling & Merja Kytö, eds. (Berlin: Mouton de Gruyter, 2008), pp. 14–32, and many other papers). Most linguists do not find this issue interesting anymore; it seems uncontroversial meanwhile that corpus data is relevant for many qualitative and quantitative research questions in many frameworks – in recent years the focus of the debate has shifted to the issue of *how* corpus data can be used in linguistic research (cf. among many others for example the Corpus Linguistics and Linguistic Theory Special Issue on Grammar and Grammaticality (2007) or the articles in Stefan Kepser & Marga Reis, *Linguistic Evidence. Empirical, Theoretical and Computational Perspectives* (Berlin: Mouton de Gruyter, 2005). There are, of course, many linguistics research questions which cannot be answered by using corpus data. In this article, I focus only on those questions for which corpus data is helpful.
- 3 Hermann Moisl, “Exploratory Multivariate Analysis,” in *Corpus Linguistics: An International Handbook*. Vol. 2, Anke Lüdeling & Merja Kytö, eds. (Berlin: Mouton de Gruyter, 2009), pp. 874–899, on p. 876.
- 4 Moisl, “Exploratory Multivariate Analysis,” section 4.
- 5 My translation of: “Gelegentlich erscheint auch sonst ein s in der Kompositionsfuge nach Femininum, ohne daß es in die Schriftsprache durchgedrungen ist, vgl. z.B. *Gemeindsversammlung* Hebel 452, 24, *Huldszeichen* Heine 2, 111, *über Naturs Größe* Le. 11, 209,

- 5, *Sprachsverbesserer*, Leibniz, Unvorgreifl. Ged. 67,3, *Vernunftswahrheiten* Le. 12, 434, 32...” Hermann Paul (1959, Band V, 13).
- 6 See e.g. Ulrich Heid, “Corpus Linguistics and Lexicography,” in *Corpus Linguistics* Vol. 1, pp. 131–153.
 - 7 This is not problematic if the purpose is simply to find any example to illustrate a statement. It is problematic if the purpose (as it was explicitly stated for the Oxford English Dictionary) is to cover the range of constructions or meanings and also if the purpose is to discover co-variation. There are a few pre-electronic dictionaries based on a systematic and quantitative study of a corpus (see for example the manually accumulated frequency dictionary by Friedrich Wilhelm Kädling, *Häufigkeitswörterbuch der deutschen Sprache* (Berlin: Privately published, 1897)). There are also a few pre-electronic quantitative studies on language data (for an overview see e.g. Reinhard Köhler, “Gegenstand und Arbeitsweise der quantitativen Linguistik,” in Reinhard Köhler, Gabriel Altmann, & Rajmund G. Piotrowski, (eds.), *Quantitative Linguistics/Quantitative Linguistik: An International handbook/Ein internationales Handbuch* (Berlin: Mouton de Gruyter, 2005), pp. 1–16.), some of them extremely far reaching and interesting.
 - 8 Roberto Busa, *Index Thomisticus* (Stuttgart: Frommann-Holzboog, 1974); idem, “The Annals of Humanities Computing: The Index Thomisticus,” *Computers and the Humanities* 14 (1980), pp. 83–90.
 - 9 One extreme is Web-based corpora (see e.g. Marco Baroni, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta, “The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-crawled Corpora,” *Journal of Language Resources and Evaluation* 43 (3) (2009), pp. 209–226 or Jan Pomikalek, Pavel Rychly & Adam Kilgariff, “Scaling to Billion-plus Word Corpora,” *Advances in Computational Linguistics. Special Issue of Research in Computing Science Vol 41* (Mexico City, 2009)).
 - 10 Anke Lüdeling, Seanna Doolittle, Hagen Hirschmann, Karin Schmidt & Maik Walter, “Das Lernerkorpus Falko,” *Deutsch als Fremdsprache* 2 (2008), pp. 67–73.
 - 11 Falko is freely available at <http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko/standardseite-en>
It contains several subcorpora of learner data and comparable native speaker data in two genres. Here I only use the essay data (122778 tokens L2 and 68491 tokens L1, version 2.0; ‘L1’ is used for first

- (native) language and ‘L2’ is used for second/foreign language). The learners represented in the Falko corpus are advanced adult learners (university students) who learned German in a classroom setting.
- 12 For an overview, see e.g. Sylviane Granger, “Learner Corpora,” in *Corpus Linguistics* Vol. 1, pp. 259–275.
 - 13 When I speak about learner language, I refer to foreign (tutored/classroom setting) and second (non-tutored) language learners (the differences are not relevant for the purpose of this paper, but see e.g. Rainer Dietrich, “Zweitsprache – Fremdsprache,” in Ulrich Ammon, Norbert Dittmar, Klaus Mattheier & Peter Trudgill, (eds.), *Soziolinguistik: Ein Internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft*. 2nd ed. (Berlin: Walter de Gruyter, 2004), pp. 311–313 or Judith F. Kroll & Gretchen Sunderman, (2003) “Cognitive Processes in Second Language Learners and Bilinguals: The Development of Lexical and Conceptual Representations,” in Catherine J. Doughty & Michael H. Long, (eds.), *The Handbook of Second Language Acquisition* (Oxford: Blackwell, 2003), pp. 104–129). There are, of course, also numerous corpora and corpus-based studies of first language learners (see e.g. Brian MacWhinney, “The CHILDES system,” *American Journal of Speech-Language Pathology* 5 (1996), pp. 5–14; Heike Behrens (ed.), *Corpora in Language Acquisition Research: History, Methods, Perspectives* (Amsterdam: John Benjamins, 2008); Holger Diessel, “Corpus Linguistics and Language Acquisition,” *Corpus Linguistics* Vol. 2, pp. 1197–1212).
 - 14 This is called interlanguage, following Larry Selinker, “Interlanguage,” *IRAL* 10/1972, 209–231, reprinted in Richards, J. C. (ed.) *Error Analysis. Perspectives on Second Language Acquisition* (London: Longman, 1972), pp. 31–54.
 - 15 Overuse and underuse can be computed for any two corpora. The visualization Add In for Excel was written by Amir Zeldes and is freely available at <http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/amir/>
Anke Lüdeling, Hagen Hirschmann & Amir Zelds, “Variationism and Underuse Statistics in the Analysis of Relative Clauses in German,” *Corpus Analysis and Diachronic Linguistics*, eds. Yuji Kawaguchi, Makoto Minegishi & Wolfgang Viereck (Amsterdam: John Benjamins, in press) show how the same method can be used in detecting language change.
 - 16 All corpus counts are normalised. All cells that are coloured in this

and the following tables show a statistically significant difference to the German data.

- 17 Douglas Biber, "Representativeness in Corpus Design," *Literary and Linguistic Computing* 8 (1993), pp. 243–257; Susan Hunston, "Collection strategies and design decisions," *Corpus Linguistics* Vol. 1, pp. 154–168.
- 18 Cf. Christian Mair, "Corpora and the Study of Recent Change in Language," *Corpus Linguistics* Vol. 2, pp. 1109–1125.
- 19 In many cases it is necessary to complement corpus data with other data (elicited data, data from psycholinguistic or sociolinguistic experiments etc.). Corpus linguistics as an empirical method uses many of the same methods used in sociolinguistics and psycholinguistics (see e.g. Terttu Nevalainen & Helena Raumolin-Brunberg, *Historical Sociolinguistics: Language Change in Tudor and Stuart England* (London Pearson Education, 2003) for historical sociolinguistics or Suzanne Romaine, "Corpus Linguistics and Sociolinguistics," *Corpus Linguistics* Vol. 1, pp. 96–111 for an overview of corpora in sociolinguistics, or Gaëtanelle Gilquin & Stefan Th. Gries, "Corpora and experimental methods: A state-of-the-art review," *Corpus Linguistics and Linguistic Theory* 5(1) (2009), pp. 1–26 for an overview of the combination of corpus methods and experimental methods).
- 20 For example, see William Labov, *Principles of Linguistic Change. Social Factors* (Oxford: Blackwell, 2001); idem, "Quantitative Analysis of Linguistic Variation," Ulrich Ammon, Norbert Dittmar, Klaus Mattheier & Peter Trudgill, (eds.) *Sociolinguistics/Soziolinguistik*. Vol 1 (Berlin: de Gruyter, 2008), pp. 6–21.
- 21 See e.g. Douglas Biber, "Multi-dimensional approaches," *Corpus Linguistics* Vol. 2, pp. 822–855.
- 22 Douglas Biber & James K. Jones, "Quantitative Methods in Corpus Linguistics," *Corpus Linguistics* Vol. 2, pp. 1286–1304.
- 23 Marco Baroni & Stefan Evert, "Statistical Methods for Corpus Exploitation," *Corpus Linguistics* Vol. 2, pp. 777–803, on p. 778.
- 24 Just briefly: Words are distributed in a Zipfian way, which means that some words are very frequent but many words are rare. In essence this shows even large corpora have not yet come close to sampling all the words in a language (this is due to morphological productivity, borrowing, creativity and other factors, which lead to the fact that the vocabulary of a language is, for all intents and purposes, infinite). For Zipfian (or LNRE) distributions only certain statistical models can be used (see R. Harald Baayen, *Word Frequency Distributions* (Dor-

- drecht: Kluwer, 2001)). Closed class categories such as part-of-speech categories often show a normal distribution and can be analysed using ‘regular’ statistical tests (Marco Baroni, “Distributions in Text,” *Corpus Linguistics* Vol. 1 (2008), pp. 803–822). It is necessary to know which kind of distribution one is looking at. For more on these issues see also Stefan Evert, “How Random is a Corpus? The Library Metaphor,” *Zeitschrift für Anglistik und Amerikanistik* 54(2) (2006), pp. 177–190 and Adam Kilgariff, “Language is never ever ever random,” *Corpus Linguistics and Linguistic Theory* 1 (2) (2005), pp. 263–276.
- 25 For an overview cf. Douglas Biber, “Multi-dimensional approaches,” *Corpus Linguistics* Vol. 2, pp. 822–855.
 - 26 There are several ways of doing this, for example a Principal Components Analysis or a Factor Analysis (see e.g. R. Harald Baayen, *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R* (Cambridge: Cambridge University Press, 2008)).
 - 27 There is so much work in each of these areas that I can only point to a few examples.
 - 28 William Labov, *The Social Stratification of English in New York City* (Washington The Center for Applied Linguistics, 1996) 2nd ed. (Cambridge: Cambridge University Press, 2006); Norma Mendoza Denton, *Homegirls: Language and Cultural Practice among Latina Youth Gangs* (London: Wiley/Blackwell, 2008).
 - 29 Ingo Plag, Christiane Dalton-Puffer & R. Harald Baayen, “Morphological productivity across speech and writing,” *English Language and Linguistics* 3 (1999), pp. 209–228.
 - 30 Christopher D. Manning, “Probabilistic Syntax,” in Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic Linguistics* (Cambridge: MA, MIT Press, 2003), pp. 289–341; Marilyn Ford & Joan Bresnan, “Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English,” *Language* 86(1) (2010), pp. 186–213.
 - 31 See Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson, (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (Amsterdam John Benjamins, 2002).
 - 32 This is called transfer (after Larry Selinker, “Language Transfer,” *General Linguistics* 9(2) (1969), pp. 67–92.). There are many studies that show that transfer happens on all linguistic levels, can be either positive or negative, and is much more complex and difficult than initially thought (Rod Ellis, *The Study of Second Language Acquisition* (Oxford: Oxford University Press, 2008)).

- 33 There are well-known legal and (less well-known) ethical issues involved sometimes. See Suber and Rausing in this volume.
- 34 See Thomas Wasow, Amy Perfors & David Beaver, (2005) “The Puzzle of Ambiguity,” in C. Orhan Orgun & Peter Sells, (eds.), *Morphology and The Web of Grammar: Essays in Memory of Steven G. Lapointe* (Stanford: CSLI Publications, 2005), on different aspects on ambiguity in natural language.
- 35 Although it is possible to have several annotation layers that code the same linguistic level (say: part-of speech), I am aware of only very few projects that have tried to do this. One prominent example is the AMALGAM project that compared part-of-speech tags and grammatical coding schemes for English
<http://www.comp.leeds.ac.uk/amalgam/amalgam/amalghome.htm>
- 36 The status of learner errors in language learning and teaching has been extensively and controversially discussed (among many others see Stephen Pit Corder, *Error Analysis and Interlanguage* (Oxford: Oxford University Press, 1981); Ellis, *The Study of Second Language Acquisition*). Here I cannot recapitulate this discussion but want to limit myself to problems of error detection and annotation.
- 37 Mark Reznicek, Maik Walter, Karin Schmid, Anke Lüdeling, Hagen Hirschmann, Cedric Krummes, *Das Falko-Handbuch. Korpusaufbau und Annotationen* (2010) Online at
<http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/standardseite>
- 38 The learner text and both target hypotheses are automatically annotated with part-of-speech categories (using the TreeTagger, Helmut Schmid, “Probabilistic Part-of-Speech Tagging Using Decision Trees,” *Proceedings of International Conference on New Methods in Language Processing, September 1994.*) Online at <http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf>
 In addition there is an automatic error coding.
- 39 The term ‘shining through’ is used in translation studies, cf. Teich (2003).
- 40 Part-of-speech tagging is usually done automatically. Most taggers use a mixture of lexicon-based and statistical techniques to assign a part-of-speech category to each word in the corpus. A tagset determines which part-of-speech tags can be used. Here we used the Stuttgart-Tübingen tagset which can be found at <http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html>. As you can see, this tagset combines genuine part-of-speech information (VV –

verb) with morphological information (FIN – finite, INF – infinite etc.).

- 41 Many linguists are organized in large infrastructure projects like CLARIN (<http://www.clarin.eu/external/>) and DARIAH (<http://www.dariah.eu/>), which promote common coding standards and accessibility. A well-known and widely used standard is formulated by the Text Encoding Initiative (<http://www.tei-c.org/index.xml>).
- 42 Multi-layer annotation was first developed for multimodal corpora, which comprise audio data, video data, and text (see Peter Wittenburg, (2009) “Preprocessing Multimodal Corpora,” in *Corpus Linguistics* Vol. 2, pp. 664–685 for an overview). There are several models for standoff multi-layer corpora today (Jean Carletta, Jonathan Kilgour, Tim O'Donnell, Stefan Evert & Holger Voormann, “The NITE Object Model Library for Handling Structured Linguistic Annotation on Multimodal Data Sets,” in *Proceedings of the EACL Workshop on Language Technology and the Semantic Web (3rd Workshop on NLP and XML, NLPXML-2003)*; Christian Chiarcos, Stefanie Dipper, Michael Götze, Ulf Leser, Anke Lüdeling, Julia Ritz & Manfred Stede, “A Flexible Framework for Integrating Annotations from Different Tools and Tagsets,” in *Traitement Automatique des Langues* 49(2) (2009), pp. 271–291) and suggestions for converting annotated data out of and into several formats (see e.g. Florian Zipser & Laurent Romary, “A Model-Oriented Approach to the Mapping of Annotation Formats using Standards,” in *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010. Malta* (2010), pp. 7–18. Online at <http://www.lrec-conf.org/proceedings/lrec2010/workshops/W4.pdf>) The Falko corpus is stored in such a format and can be searched using the freely available search tool Annis2 (Amir Zeldes, Julia Ritz, Anke Lüdeling & Christian Chiarcos, “ANNIS: A Search Tool for Multi-Layer Annotated Corpora,” in *Proceedings of Corpus Linguistics 2009*, July 20–23, Liverpool, UK).

When everything is digitized...
The future of libraries, archives and
museums, and analogue

FROM SUBJECTS TO CITIZENS IN A GLOBAL REPUBLIC OF LETTERS

Gregory Crane

“We alone regard a man who takes no interest in public affairs, not as a harmless, but as a useless character; and if few of us are originators, we are all sound judges of a policy. The great impediment to action is, in our opinion, not discussion, but the want of that knowledge which is gained by discussion preparatory to action.” – Pericles on Athenian Democracy, as he is quoted in Thucydides’ *History of Peloponnesian War* (Thuc. 2.40, tr. Jowett).

The Greek historian Thucydides is arguably the true inventor of historical inquiry within the continuous tradition of European literature, but we have no way of gauging the relationship between the speeches that he includes and what was actually said – the *New York Times*, alas, was not able to publish transcripts. But the Funeral Oration that Thucydides attributes to Pericles provides a broad analysis that prescribes the features of a successful democracy as much as it describes the Athens of the fifth-century BCE. The passage quoted above already identifies two elements that are central to democratic thought and, as we ponder how the digital present may evolve into the future, we must explore two of the same questions. On the one hand, we must provide to the broadest possible audience that knowledge and understanding upon which sound judgments must depend. On the other hand, we have a chance to expand the opportunities that human beings have to contribute, each in their own way, something new and distinctive to the world.

The first of these goals is far easier to understand, both for us and for Pericles. If we can provide good information and then forums for rational discussion, we can, at least in theory, improve collective decision-making and the world that those collective decisions make – whether the collective decisions are overtly political or affect day-to-day decisions (e.g., about how we conserve and consume energy). Obviously, there are problems with this simplistic proposition but it is easy enough to see how we can advance this general goal: the World Wide Web has provided unprecedented new instruments whereby we can transmit ideas and information around the globe. More members of the human race have access to more and more detailed information about their world than was not only possible, but even practicably conceivable, a generation ago. Humanists, in particular, have preached for years that we teach students critical thinking and prepare them to evaluate ideas on their own.¹ In a world where fewer intermediaries

filter information available to us, those skills are more important than ever. And while the noise of false information has risen, it is at the same time much more difficult for powerful interests to repress information. The better we do our job educating the next generation to think critically, the stronger our societies around the world will be.

There are huge challenges to rational, public discourse – those of us who live in the United States of 2010 certainly see these challenges on a daily basis. But most of us may be, like the Pericles that we see above, implicitly more pessimistic about who can originate, as opposed to critique, ideas. Nevertheless, Wikipedia, for example, has demonstrated a new mode of intellectual production, one that draws upon members of the general public, most of whom would have had no opportunity to contribute in the print and mass media cultures of the twentieth century. Many of us in academia thought the model of distributed, lightly mediated production was absurd – many of us still publicly dismiss resources such as Wikipedia, even as some of us draw regularly upon these and similar resources for information.²

But even in the conservative world of academia we can see changes at work. On an anecdotal level, a colleague of this paper's author spoke to undergraduates about his experience on a U.S. National Science Foundation panel that awarded fellowships for graduate study. He reported that grade point average and standardized test scores could, if they were strikingly poor, pose barriers for students seeking internships, but also noted that it was more important to have demonstrated research potential. The fellowship committee looked for, and found, candidates who had gone beyond earning high grades and begun to conduct independent research as undergraduates. A colleague who had earned a tenure stream position in Computer Science and who had received his B.A. in 1997 expressed his amazement at the degree to which undergraduates have begun to contribute as researchers in their own right.³

The culture of the humanities has not progressed so far as in the sciences. In the fall of 2009, Tufts University advertised a tenure track job in Classics – one of the few posted in an economically stressed year. The job listing thus attracted almost 200 candidates from virtually every major Ph.D. program in North America and the U.K. (as well as many beyond). The brief job listing concluded with the explicit statement: "We especially welcome candidates who can support contributions to and original research by undergraduates as well as M.A. students within the field of Classics." This statement took up about one third of the job notice and was a clear indication of what the department wanted.

As we read through the applications, almost none of the candidates alluded to the contributions and original research by undergraduates and M.A. students. When the committee conducted face-to-face interviews with eighteen candidates, most tried to avoid the topic, staring at their feet, shuffling in their seats and looking uncomfortable as they tried to imagine what undergraduates might possibly be able to do other than Xerox articles. Some of the rising scholars were able to talk about undergraduate and M.A.-level contributions and research but they had developed these ideas on their own. Classics Ph.D. programs of 2009 supported a range of intellectual approaches but they trained all of their students to think in terms of single-author publications, produced for 19th century journals and 20th century presses, written by and for specialists.

In Classical Studies and in most of the Humanities, even our students have been subjects rather than citizens. Before going to graduate school and training to become professional scholars, they have the opportunity to contribute little but numbers to departmental enrollments. Where our colleagues in the sciences often select from their classes the most promising students for their labs and research projects, humanists generally view the end of the semester as an end of distractions and an opportunity to work undistracted by students on their own projects.

The digital world has transformed the landscape. Scholars have been studying the Greco-Roman world for thousands of years – our understanding of Greek and Latin builds upon an immense and cumulative structure of ideas. In the digital environment, this cumulative body provides a critical foundation and places well-studied subjects such as Greek and Latin at an advantage. But the shift from print to digital goes beyond photographing print pages and we face a transformation in intellectual practice and culture as profound as the rise of writing and then of print.⁴ Drawing upon what we have inherited from previous generations, we must build an intellectual space in which we can study not only Greco-Roman culture but also see Greco-Roman culture as one element within the full cultural heritage of humanity.⁵ There is a lot to do – far too much for a handful of professional scholars.

The Transformation of the Linguistic Record

The linguistic record of humanity is undergoing three profound transformations. The first transformation, which involves the creation of semantically encoded, machine actionable editions,⁶ resembles a shift from black and white to color: we see the same objects as we did in the past, but we are able now to perceive elements within them that were previously

invisible. This first shift, when carried to its logical conclusions, includes a wealth of annotation that produces something qualitatively different and allows us to see patterns at a finer resolution than with the naked eye – this transformation resembles the development of microscopes and the revelation of inner space. At the other extreme, within the flood of publicly accessible digitized sources the amount of source material has exploded in quantity. We now need instruments with which to find patterns in much larger spaces than we had ever imagined working with in the past.⁷ For this we need instruments that are analogous to telescopes so that we can identify distant objects in a rapidly expanding universe. Each of these transformations opens up new fields of intellectual activity.

1. The Transformation of the Familiar

Publishing images of printed pages obviously constitutes only a superficial form of digitization – the printed form simply becomes available in a digital space, even if we can generate searchable text and we can within those print pages detect previously invisible patterns circulating across time, culture, and language. If, however, we encode the underlying structures implicit in print publications in a machine-actionable form and we make our publications available for analysis by many different services and researchers, then we can in fact create digital publications that represent a qualitative advance over their print predecessors.⁸ Where the editions, lexica, encyclopedias, commentaries and other components of our print infrastructure had reached maturity, were published in a fixed form and updated rarely, sometimes not for decades or even centuries, their digital analogues now must draw not only upon scarce expert labor but also upon automated methods⁹ and contributions from a wide range of students and members of the general public.¹⁰ If we are to study the past, we must exploit the possibilities of the wiki-world in which we now live.

Figure 1 illustrates a Web page for part of a Greek edition of Plutarch's *Moralia*. The data is based upon a paper edition and uses a format based on printed editions. The icons for XML and CC (Creative Commons)¹¹ indicate two different elements that differ fundamentally from print practice and possibilities.

The Creative Commons license allows anyone to download the source text for free and to create derivative works. Attribution simply asserts that those who use this text must credit the source (in this case, a digital

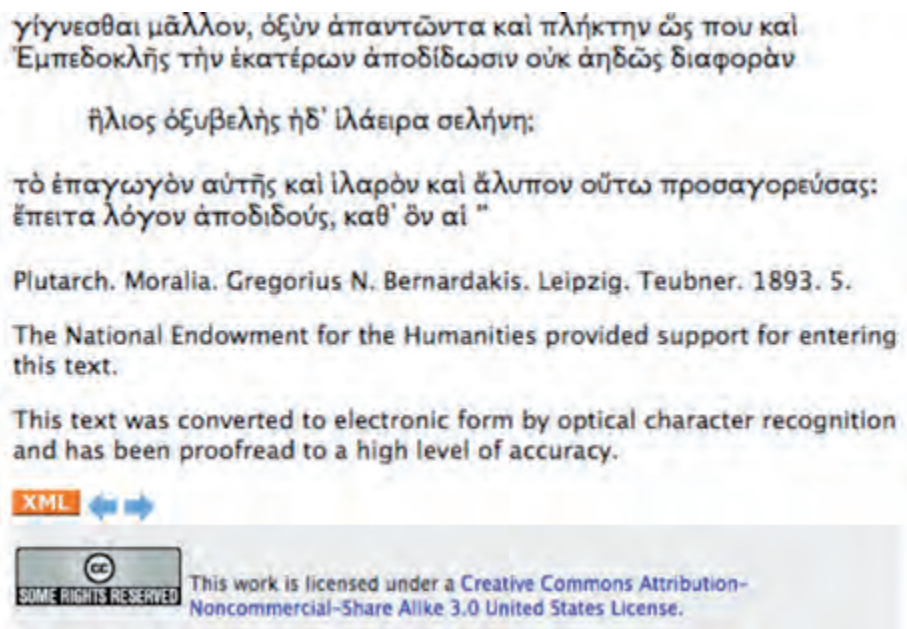


Figure 1. Like a book but more: a page from a Greek edition of Plutarch's *Moralia*.

edition from the Perseus Digital Library based upon the Bernardakis edition of Plutarch's *Moralia*). The non-commercial restriction is designed to prevent a commercial entity from incorporating this work into their products without permission and especially from adding some marginal improvements to create a proprietary version. The "Share Alike" attribute is probably enough on its own by specifying that any new versions of this edition be distributed under the same Creative Commons license – subsequent versions of this digital edition will drop the non-commercial restriction.

The Creative Commons license not only confers a standard set of rights but is also available in a machine-actionable format. This means that researchers can aggregate millions of Creative Commons sources from thousands of collections and know, precisely, what they can do with the data that they have collected. Research in archaeology, history, and any cultural framework increasingly depends upon our ability to collect disparate materials automatically.¹² Signing and faxing slightly different rights agreements for each collection does not scale up to the needs of research – such compartmentalized collections are the 21st century equivalent of early railroads, each with a slightly different gauge, that forced passengers and freight to shift from train to train every few miles.

New forms of scholarship increasingly include datasets to which re-



Figure 2. A passage from Plutarch's *Moralia* encoded in XML with the TEI tagset.

searchers have added structures and other annotations. These annotations may be produced automatically and affect billions of words. Researchers cannot conduct this work unless they have the right to reproduce the original source data upon which their work was based.

The XML icon indicates first that the HTML text that is the source for what the reader can see is derived from an XML source file and that this source file is also available for download. The XML source files use the Text Encoding Initiative (TEI)¹³ guidelines to encode semantic information about the content of the text whereas the HTML encoding focuses upon page layout. The line of Greek that appears in the middle of Figure 1, for example, and begins with the word *hēlios* is the beginning of a quotation from another author. The XML text includes markup indicating that this is a QUOTE and should be rendered as a block quote.

Traditional editions assume considerable expertise from their readers, so much expertise as to minimize the success, if not defeat the purpose, of the editor's work. Scholars on the advisory board of one established editorial project who were literary critics and not professional editors would refer to the textual notes at the bottom of the page, with their abbreviated accounts of variants and other issues, as the "zone of fear." The scholars for whom the edition was produced simply ignore the textual notes.

As we encode textual data in a machine-actionable form, we have new methods to represent textual relations in ways that are not only flexible but also are clearer and more precise than was possible. We can, for exam-

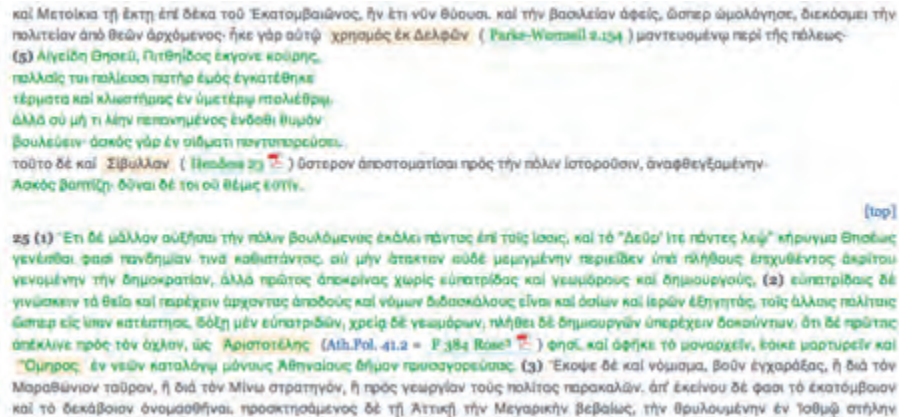


Figure 3. Plutarch's *Life of Theseus*, with passages that are thought to paraphrase particular sources highlighted in green <http://www.fragmentarytexts.org>.

ple, highlight the differences between various manuscript sources¹⁴ and illustrate much more complex textual traditions, and do so in a fashion that makes more information comprehensible to a wider audience than was ever feasible before.

One venerable genre of scholarly edition illustrates the more general transformation. The vast majority of Greek and Latin writings have been lost, and most ancient authors are known only because the few surviving authors mention, quote, or paraphrase them. Scholars have labored for centuries to reconstruct what we can learn from these snippets and have created editions of fragmentary authors, which are essentially annotated excerpts from surviving texts.¹⁵

Figure 3 illustrates one model that inverts the traditional representation of fragmentary authors. A traditional edition of a fragmentary author collects excerpts from dozens or hundreds of authors. Readers need to find the original sources from which these excerpts are drawn if they are to understand the context for the quotations or evaluate the relationship between the paraphrase and its lost source. In Figure 3, the classicist Monica Berti has begun to mark those passages within Plutarch that scholars have included in their fragmentary editions. These passages can, of course, be extracted automatically and presented as a series of excerpts – but the resulting anthology can also retain its links back into the source texts, allowing readers to go directly to see the excerpts in context. Such quantitative changes in the time of access can have a qualitative impact if fragmentary editions have links to hundreds of original sources, many of which are digital versions of books that were once available only as paper objects in a handful of research libraries.

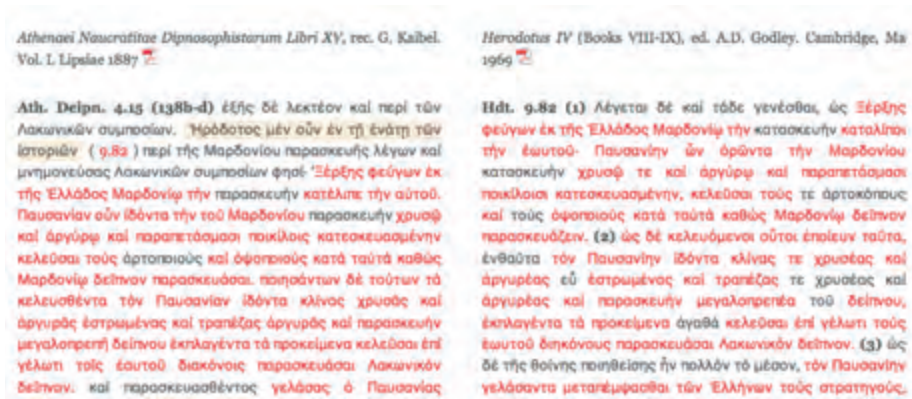


Figure 4. Athenaeus using Herodotus. Words that Athenaeus quoted directly from Herodotus are in red. The un-highlighted words represent places where the texts diverge <http://www.fragmentarytexts.org>.

Figure 4 illustrates another direction with which to place an ancient problem in a new light. When one author paraphrases another, we need to decide how close we think that paraphrase is to the original.¹⁶ The best way to do this is to study how authors paraphrase their predecessors when the paraphrased original is available for comparison. Figure 4 depicts a visualization that allows us to see how the Greek author Athenaeus nearly (but not quite) quotes Herodotus word-for-word. By analyzing such passages we can begin to develop models about how Athenaeus (and other authors) use their sources and then apply these models to those passages where the paraphrased original does not survive. The goal is not to arrive at final conclusions but instead to make explicit and systematic reasoning that before had been based solely on our instincts and to advance our understanding of the problem.

Editions of surviving authors and of fragmentary texts remain critical instruments of study in a digital space, but their form and function also undergo profound changes and yield something that is qualitatively different from its print predecessors. We now need to reorganize *all* of our textual sources – certainly every author who warranted a critical edition in print now needs to be re-edited in a digital space. Simply typing in a transcript of the printed sources is not enough, nor are reformatting tools sufficiently intelligent to capture the logical structures implicit in the print sources.

Other core resources such as dictionaries, encyclopedias, and commentaries are undergoing similar transformations – every resource that evolved in print culture has a digital analogue, but each digital analogue must not only build as much as possible on printed knowledge but also be

redesigned and rebuilt from the ground up.¹⁷ We have an immense amount of work to do, much of it well within reach of our students and others who are not professional scholars. The tiny cadre of professional scholars simply does not have the time to do all the work that needs to be done. We must reorganize our field and our work, opening up contributions to a much broader base and rethinking our relationship to society as a whole.¹⁸

2. Discovering the inner space of a linguistic world

The field of corpus linguistics uses scientifically designed collections of data to study linguistic patterns. Corpus linguists have developed not only an increasing body of research questions about language, but also increasingly sophisticated methods whereby they can study various phenomena. At the core of these methods stands the act of annotating textual corpora and then studying the resulting patterns.¹⁹ Much of philology is, of course, inherently an exercise in corpus linguistics because students of historical languages work, by definition, with linguistic corpora and must exploit every method available to learn as much as they can from the limited sources that survive from the past.

Syntactic analysis is arguably the most important category of annotation for a corpus. Figure 5 illustrates the opening line of the *Iliad*, analyzed according to a dependency grammar and stored as part of the Ancient Greek Dependency Treebank developed by the Perseus Project.²⁰

The Greek and Latin Treebanks, developed by Perseus, are among



Figure 5. Dependency tree of *μήνιν ᾄειδε θεὰ Πηληϊάδεω Ἀχιλῆος* ("Sing, goddess, of the rage of Achilles, the son of Peleus"), Homer, *Iliad* 1.1. Arcs are drawn from heads to their dependents. <http://www.fragmentarytexts.org>

Play	Total	Nominal	θ/θn		% nominal
			general	3rd p. pres.	
Persae	75	26	31	9	45.61
Sept.	103	52	26	7	66.67
Suppl.	116	49	35	14	58.33
Agam.	140	68	49	10	58.12
Cho.	97	46	34	14	57.50
Eum.	103	31	49	14	38.75
P. V.	116	70	34	11	67.31
TOT	750	342	258	79	57.00

Type	Nominals		Copula		Significance
	tot	%	tot	%	
adj. in -toc	23	54.76	19	45.24	
adj. comp. with εἰς	21	67.74	10	32.26	*
adj. comp. with ἐκ	11	78.57	5	21.53	*
adj. comp. with ἀ-	18	38.30	29	61.70	*
complex adj.	16	55.17	13	44.83	
adj. comp. with preposition	22	61.11	14	38.89	
adj. + complement	37	64.91	20	35.09	*
adj. + verb	11	68.75	5	31.25	*

Figure 6. Tables from Francesco Mambrini, "A Treebank of Aeschylus: A corpus-based approach to Ancient Greek poetic syntax," forthcoming.

efforts that have begun to apply emerging methods from corpus linguistics to the study of historical languages – an area of study underway for millennia but now poised for a transformation. Both the Treebanks themselves and the process of their production constitute material steps towards a new synthesis of traditional scholarly goals and data-driven digital analysis, both in terms of fundamental linguistic research and in providing intellectual access to texts.

Corpus linguists developed resources such as Treebanks to study general linguistic trends. Students of historical languages are able now to place their understanding of languages that have been studied for hundreds or even thousands of years on a fundamentally new foundation and to begin asking questions that were not imaginable with print tools. Figure 6 illustrates calculations extracted from the first reasonably complete author corpus available in our existing Ancient Greek Dependency Treebank, the seven surviving plays of the Greek tragedian Aeschylus. These illustrate an initial step in a vast re-analysis of Greek, Latin and every historical language.



Figure 7. Three interpretations of the same sentence from Aeschylus' *Agamemnon* (lines 176–178) encoded in a dependency grammar and visualized as trees. The interpretations from left to right are by Denniston-Page, Fraenkel, and Bollack (Bamman, Mambrini and Crane, 2009).

Corpus linguists focus on general linguistic trends, but philologists can apply the same methods to compare differing interpretations of the same passage. Figure 7 illustrates three scholarly interpretations of the same sentence in the *Agamemnon* of Aeschylus. Each interpretation was published in print form but the Classicist Francesco Mambrini has represented them under the formalism of a dependency grammar.

It would be hard to overemphasize the potential importance of these new methods from corpus linguistics for the study of historical languages. At the same time, annotation is a relatively new task and many categories of annotation (such as syntactic analysis) do not lend themselves to automated methods. Francesco Mambrini created the Treebank for Aeschylus because the text of Aeschylus is problematic and heavily studied. Many texts, however, are less problematic and in these cases we draw upon contributions from students and other non-professionals to augment what professional scholars can do.

Figure 8 illustrates two sentences from the Treebank for the Homeric *Iliad*. Each sentence contains credits: Michael Kinney (mpkinn10) and Molly Miller (millermo), then (respectively) an undergraduate at Holy Cross and an M.A. student at Tufts University, each independently analyzed these sentences and thus served as the primary annotators. Jack Mitchell (nicantor), a Homerist now on the faculty at Dalhousie University, served as the secondary annotator. He compared their analyses and produced the final interpretation now published as part of the Treebank.

```
- <sentence id="3044" document_id="Perseus:text:1999.01.0133" subdoc="book=6;card=1" span="pa/ntas0:4">
  <primary>mpkinn10</primary>
  <primary>millermo</primary>
  <secondary>nicantor</secondary>
  <word id="1" form="pa/ntas" lemma="pa=s1" postag="a-p---ma-" head="3" relation="OBJ"/>
  <word id="2" form="ga/v" lemma="ga/v1" postag="g-----" head="3" relation="AuxY"/>
  <word id="3" form="file/esken" lemma="file/w1" postag="v3siaa---" head="0" relation="PRED"/>
  <word id="4" form="o/dw=l" lemma="o/do/s1" postag="n-s---md-" head="5" relation="ADV"/>
  <word id="5" form="e/pi" lemma="e/pi1" postag="t-----" head="7" relation="AuxP"/>
  <word id="6" form="oi/ki/a" lemma="oi/ki/on1" postag="n-p---na-" head="7" relation="OBJ"/>
  <word id="7" form="nai/wn" lemma="nai/w2" postag="t-appamn-" head="3" relation="ADV"/>
  <word id="8" form="." lemma="period1" postag="u-----" head="0" relation="AuxK"/>
</sentence>
- <sentence id="3045" document_id="Perseus:text:1999.01.0133" subdoc="book=6;card=1" span="a)lla0:2">
  <primary>mpkinn10</primary>
  <primary>millermo</primary>
  <secondary>nicantor</secondary>
  <word id="1" form="a)lla" lemma="a)lla/1" postag="d-----" head="14" relation="AuxY"/>
  <word id="2" form="oi/" lemma="oi/" postag="p-s---md-" head="8" relation="OBJ"/>
  <word id="3" form="ou)" lemma="ou)" postag="d-----" head="8" relation="AuxZ"/>
  <word id="4" form="tis" lemma="tis1" postag="p-s---mn-" head="8" relation="SBJ"/>
  <word id="5" form="tw=n" lemma="o(1" postag="j-----" head="4" relation="ATR"/>
  <word id="6" form="ge" lemma="ge1" postag="g-----" head="5" relation="AuxZ"/>
  <word id="7" form="to/" lemma="to/le1" postag="d-----" head="8" relation="ADV"/>
  <word id="8" form="h/rkese" lemma="a)rke/w1" postag="v3siaa---" head="14" relation="PRED_CO"/>
  <word id="9" form="lugro/n" lemma="lugro/s1" postag="a-s---ma-" head="10" relation="ATR"/>
  <word id="10" form="o/leqroo" lemma="o/leqroo1" postag="n-s---ma-" head="8" relation="OBJ"/>
```

Figure 8. Two sentences from the Treebank for the Homeric *Iliad*, represented in an XML format.

The official Treebank thus records the contributions of professor and student alike in this new knowledge source.

New forms of analysis depend upon the ability to draw upon our students and members of the public – here again, there is simply too much for a handful of scholars to do. Tasks such as contributing to a Treebank are non-trivial and demand advanced knowledge of languages such as Greek and Latin, as well as an understanding of sophisticated formalisms such as the dependency grammars used in the Greek and Latin Treebanks. Scholarship can only advance in this way if we treat our students and members of the public as potential collaborators.

3. The explosive growth in the corpus of public record

The corpus of public record for subjects such as Greek and Latin has been vanishingly small. A few large research libraries have maintained comprehensive collections of primary sources in Greek, Latin and other historical languages. Physically accessible to few, these sources were intellectually accessible to even fewer, for most of those students and scholars who work in such libraries have not studied languages such as Greek and Latin. In a digital age, all the major primary sources for Greek and Latin are available in digital form, many of them in open access collections such as the Perseus Digital Library. Moreover, a growing array of services such as automated dictionary lookup and word/phrase translation have enabled students of these languages to do more with their knowledge than was possible with static print documents. Moreover, the judicious use of these tools along with source texts linked to translations has allowed researchers to work with concepts in languages in which they have no training at all.

The corpus of public record has exploded. Not only are well-studied canonical works increasingly available, but also a much larger body of material is now finding its way to the general public. The vast majority of surviving Latin, for example, was produced after the Classical period. Perhaps 50 million words survive from the period before 500 CE. In 2010, researchers at Tufts and the University of Massachusetts at Amherst extracted 27,000 books from the open source library of digital books available at the Internet Archive.²¹ After analyzing roughly one third of this collection, researchers discovered that they had about 380 million words of Latin and that the final total would probably exceed 1 billion words.²²

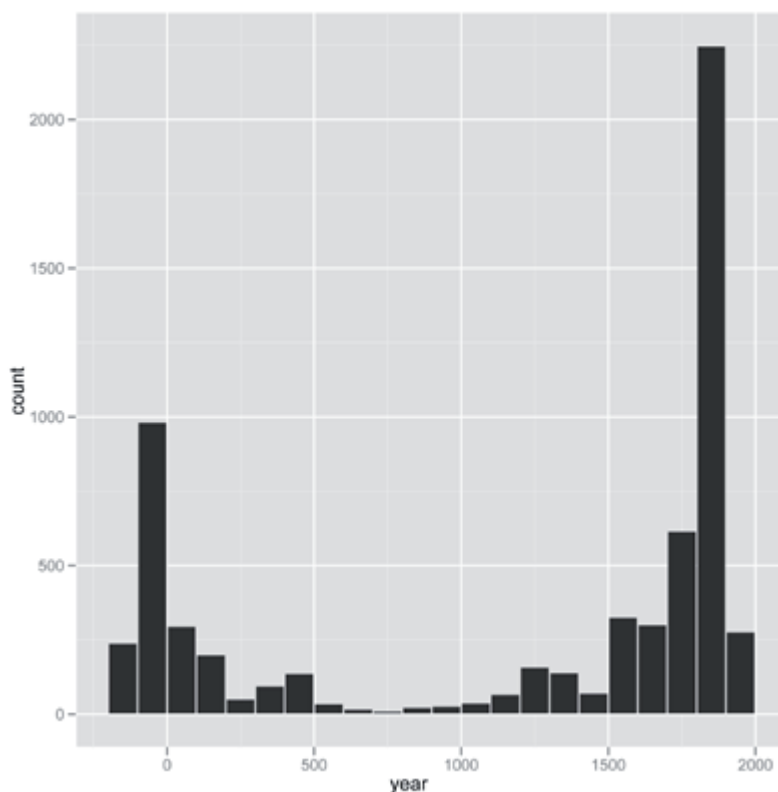


Figure 9. Distribution of Latin from the 9,000 books (out of 27,000 books labeled as being in Latin and downloaded from the Internet Archive) that were properly dated and analyzed as of September 2010.

The billion words of Latin available today was extracted from a library of 1.2 million books and represents only a small subset of what survives – Google has already digitized 15 million books and much Latin survives in manuscripts and elsewhere outside of print culture. We face a vast expansion of an already vastly expanded corpus of public record for Latin. Historical languages such as Greek, Classical Arabic, Classical Chinese, and Sanskrit face similar transformations as these enormous corpora find their way in publicly accessible collections.

Generations of students will be able to select documents from these emerging corpora for analysis, producing linguistic annotation, commentary, contextualizing introductions and translations. The best of these can be placed in university archives that can provide open access to this work and preserve these contributions for generations. Efforts such as Wiki editions can augment what our students can do and provide the materials that humanity will need in order to understand the record of the past.

Conclusion

The shift from a print to a digital world has transformed not only the questions that researchers can pose, but also the ability – and indeed the necessity – of opening up every field in the humanities. The learned elites that developed our resources for the past in print culture simply cannot do the job alone.

Four basic shifts are visible. First, we need to reinvent our basic instruments of study such as editions, lexica, and grammars. Second, fields such as corpus linguistics allow us to see within existing collections at a resolution and with a precision not previously possible. Third, fields such as computational linguistics and information retrieval have allowed us to work with the vastly expanded corpora of public record now available in digital form. Fourth, we now see not only a transformation within fields such as Classics but a need to integrate the study of topics such as Greco-Roman culture alongside topics such as the civilizations of the ancient near east (with which the Greco-Roman world interacted), ancient India and China. We need to think of Classics as a global network of cultures, whose interactions we can trace over four thousand years of the linguistic record. The resulting republic of letters must draw more fully upon a wider range of scholarly communities (e.g., scholars of Greco-Roman culture who publish in Croatian or in Arabic) and foster new communities that connect previously separate fields such as Classical Greek and Classical Chinese. We need, in short, a republic of letters that is global in scope and that encourages people from around the world to contribute as citizens.

References

- 1 For example, the importance of teaching critical thinking to students and how to digitally model this process within a shareable learning object has been explored by Eleanor OKell, Dejan Ljubovic & Cary MacMahon, "Creating a Generative Learning Object (GLO): Working in an 'Ill-Structured' Environment and Getting Students to Think," *Digital Research in the Study of Classical Antiquity* (eds.) Gabriel Bodard & Simon Mahony (Burlington, VT: Ashgate Publishing, 2010), pp. 151–170.
- 2 One recent study explores this dichotomy between how faculty use Wikipedia in their own research and what they tell their students regarding its use for their own research. Patricia L. Dooley, "Wikipedia and the Two-Faced Professoriate," *WikiSym '10: Proceedings of the 6th International Symposium on Wikis and Open Collaboration* (New York, NY, USA: ACM, 2010), pp. 1–2.
- 3 The increasing importance of undergraduate research and the various approaches and attitudes of faculty towards implementing undergraduate research projects across the disciplines has been explored by a variety of research. For a recent cross-disciplinary study, see Chris Craney, Tara McKay, April Mazzeo, Janet Morris, Cheryl Prigodich & Robert de Groot, *Cross-Discipline Perceptions of the Undergraduate Research Experience*. Report from the California Center for Publication Research at the University of California-Los Angeles, December 2009. <http://papers.ccpr.ucla.edu/papers/PWP-CCPR-2009-054/PWP-CCPR-2009-054.pdf> for broad perspectives on undergraduate research within the humanities, see Sheila T. Cavanagh, "Bringing Our Brains to the Humanities: Increasing the Value of Our Classes while Supporting Our Futures," *Pedagogy*, 10 (2010), pp. 131–142 and Reed Wilson, "Researching Undergraduate Research in the Humanities," *Modern Language Studies*, 33 (2003), pp. 74–79, and for specific details on undergraduate research projects within classics and archaeology, see Chris Blackwell & Thomas R. Martin, "Technology, Collaboration & Undergraduate Research," *Digital Humanities Quarterly*, 3 (January 2009) <http://www.digitalhumanities.org/dhq/vol/3/1/000024.html> and Arne R. Flaten, "The Ashes2Art Project: Digital Models of Fourth-Century BCE Delphi, Greece," *Visual Resources: An International Journal of Documentation*, 25 (December 2009), pp. 345–362 respectively.
- 4 The sheer breadth of digital sources in classics, for example, and the new types of study it has made possible are illustrated by the diver-

sity of topics covered in a recent book that presents groundbreaking research in digital classics: Gabriel Bodard & Simon Mahony (eds.), *Digital Research in the Study of Classical Antiquity* (Burlington, VT: Ashgate Publishing Company, 2010). Earlier research has also examined the impact of information technology on the study of classical antiquity, including Jay D. Bolter, "The Computer, Hypertext & Classical Studies," *The American Journal of Philology*, 112 (1991), pp. 541–545 and Lorna Hardwick, "Electrifying the Canon: The Impact of Computing on Classical Studies," *Computers and the Humanities*, 34 (August 2000), pp. 279–295.

- 5 The need for an intellectual space or a cyberinfrastructure that supports a more holistic and integrated approach to the study of cultural heritage and the humanities has also been explored by Martin Doerr & Dolores Iorizzo, "The Dream of a Global Knowledge Network – A New Approach," *Journal on Computing and Cultural Heritage*, 1 (June 2008), pp. 1–23 and Tobias Blanke, "From Tools and Services to e-Infrastructure for the Arts and Humanities," *Production Grids in Asia* (eds.) Simon C. Lin & Eric Yen (Boston, MA: Springer US, 2010), chapter 10, pp. 117–127.
- 6 The need to create "true" digital editions that include semantically encoded information that is both distributed and machine-actionable has also been explored by Peter Robinson, "Editing Without Walls," *Literature Compass*, 7 (2010), pp. 57–61 and Gabriel Bodard & Juan Garces, "Open Source Critical Editions: A Rationale," in *Text Editing, Print and the Digital World* (Burlington, Ashgate Publishing, 2009), pp. 83–98.
- 7 The need for computational tools that can help scholars both mine and make sense of larger patterns within massive text collections was noted by the ACLS report on cyberinfrastructure in the humanities: American Council of Learned Societies, *Our Cultural Commonwealth: The Final Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities & Social Sciences* (2006). <http://www.acls.org/cyberinfrastructure/> and research in both data and text mining within digitized humanities text collections is an active area of research, see for example Brad Pasanek & D. Sculley, "Mining Millions of Metaphors," *Literary and Linguistic Computing*, 23 (September 2008), pp. 345–360 and Matthew Kirschenbaum, "The Remaking of Reading: Data Mining and the Digital Humanities," *NGDM 07: National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation*. October 2007.

<http://www.cs.umbc.edu/~hillol/NGDM07/abstracts/talks/MKirschenbaum.pdf>

- 8 The need to encode the implicit knowledge structures found within traditional printed sources when they are digitized, particularly reference works, so that they can become machine-actionable and used for new types of research has also been articulated by Timothy Stinson, "Codicological Descriptions in the Digital Age," *Kodikologie und Paläographie im digitalen Zeitalter-Codicology and Palaeography in the Digital Age* (2009), pp. 35–52.

<http://kups.ub.uni-koeln.de/volltexte/2009/2959/> in terms of codicological descriptions, and Russell Horton, Robert Morrissey, Mark Olsen, Glenn Roe & Robert Voyer, "Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclopédie," *Digital Humanities Quarterly*, 3 (June 2009).

<http://www.digitalhumanities.org/dhq/vol/3/2/000044.html> in terms of encyclopedias.

- 9 The use of automated methods such as information extraction to help create new historical knowledge bases as well as enhance and extend digital analogues of printed knowledge sources has a growing body of literature. For some specific examples, see Paul D. Clough, Neil Ireson & Jennifer Marlow, "Extending Domain-Specific Resources to Enable Semantic Access to Cultural Heritage Data," *Journal of Digital Information*, 10 (2009).

<http://journals.tdl.org/jodi/article/view/698> and Ryan Shaw, Michael Buckland & Ray Larson, "Integrating Tools for Synthesis into Digital Libraries," *Proceedings of JCDL 2009 Workshop: Integrating Digital Library Content with Computational Tools and Services* (June 2009).

- 10 A variety of libraries and museums already allow users to contribute notes and metadata to historical digital image collections through the Flickr Commons (see for example, Michael Zarro & Robert Allen, "User-Contributed Descriptive Metadata for Libraries and Cultural Institutions," *Research and Advanced Technology for Digital Libraries* (2010), pp. 46–54 on the Library of Congress experience). User contributed annotations (e.g. through the use of Amazon's Mechanical Turk): see Chris C. Burch, "Fast, Cheap & Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk," *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, August 2009, pp. 286–295.

<http://www.aclweb.org/anthology-new/D/D09/D09-1030.pdf>) are also becoming increasingly important in many natural language

- processing tasks: see Rion Snow, Brendan O'Connor, Daniel Jurafsky & Andrew Ng, "Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks," *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (Honolulu, Hawaii: Association for Computational Linguistics, October 2008), pp. 254–263.
<http://www.aclweb.org/anthology-new/D/D08/D08-1027.pdf>.
- 11 <http://creativecommons.org/>
 - 12 The importance of open data that can be copied, automatically collected and stored in distributed locations, and reused freely has also been explored by Eric Kansa, Sarah Kansa, Margie Burton & Cindy Stankowski. "Googling the Grey: Open Data, Web Services & Semantics," *Archaeologies*, 6 (August 2010), pp. 301–326 in terms of archaeology, and more generally in terms of classics by Hugh A. Cayless, "Ktêma es aiei: Digital Permanence from an Ancient Perspective," *Digital Research in the Study of Classical Antiquity* (eds.) Gabriel Bodard & Simon Mahony (Burlington, VT: Ashgate Publishing, 2010), pp. 139–150.
 - 13 <http://www.tei-c.org/index.xml>
 - 14 The best way in which to semantically encode the variants and differences found between various manuscript sources has received a great deal of discussion that is beyond the scope of this paper, but see Desmond Schmidt & Robert Colomb, "A Data Structure for Representing Multi-Version Texts Online," *Int. J. Hum.-Comput. Stud.* 67 (June 2009), pp. 497–514 for extensive bibliography and an overview of various approaches (e.g. standoff markup, directed graphs).
 - 15 For a brief discussion of the traditional features of fragmentary editions, see Monica Berti, Matteo Romanello, Alison Babau & Gregory Crane, "Collecting Fragmentary Authors in a Digital Library," *JCDL '09: Proceedings of the 2009 Joint International Conference on Digital Libraries* (New York, NY, USA: ACM, 2009), pp. 259–262.
http://www.perseus.tufts.edu/publications/JCDL09_sp.pdf
 - 16 Computational approaches to measuring text reuse and paraphrase of earlier sources by later authors have been presented by Marco Buehler & Annette Geßner, "Citation Detection and Textual Reuse on Ancient Greek texts," *DHCS 2009-Chicago Colloquium on Digital Humanities and Computer Science*. November 2009.
<http://lingcog.iit.edu/%7Eargamon/DHCS09-Abstracts/Buechler-Gessner.pdf> and John Lee, "A Computational Model of Text Reuse in Ancient Literary Texts," *Proceedings of the 45th Annual Meeting of*

- the Association of Computational Linguistics* (Prague, Czech Republic: Association for Computational Linguistics, 2007), pp. 472–479.
<http://acl.ldc.upenn.edu/P/P07/P07-1060.pdf>
- 17 For further exploration of how traditional reference works must be redesigned for a digital environment, see Gregory Crane, Brent Seales & Melissa Terras. “Cyberinfrastructure for Classical Philology,” *Digital Humanities Quarterly*, 3, (January 2009).
<http://www.digitalhumanities.org/dhq/vol/3/1/000023.html#>
 - 18 A recent article by Lisbet Rausing, “Toward a New Alexandria: Imagining the Future of Libraries,” *The New Republic* (March 2010).
<http://www.tnr.com/print/article/books-and-arts/toward-new-alexandria> explores the closed nature of academic scholarship and how for both libraries and the humanities to thrive they must re-engage with the public and make scholarship openly available. A slightly edited version is found in this volume.
 - 19 For a recent overview of corpus linguistics and the creation of annotated corpora, see Anke Lüdeling & Amir Zeldes, “Three Views on Corpora: Corpus Linguistics, Literary Computing & Computational Linguistics,” *Jahrbuch für Computerphilologie*, 9 (2007), pp. 149–178.
<http://computerphilologie.tu-darmstadt.de/jg07/luedzeldes.html>
 - 20 A list of relevant publications regarding the development of the treebanks and downloadable versions of the treebanks are available at
<http://nlp.perseus.tufts.edu/syntax/treebank/>
 - 21 <http://www.archive.org>
 - 22 The results of this work will appear in David Bamman & David Smith, “Extracting Two Thousand Years of Latin from a Million Book Library,” *Journal of Computing and Cultural Heritage*, to appear 2011.

DIGITAL HISTORY OF SCIENCE: INTEGRATING TEXTS WITH ARTEFACTS

Marco Beretta

The digital revolution has given an extraordinary impulse to the formation of a comprehensive reassessment of the historiography of science and the treatment of its sources. Until a few years ago the historian of science primarily dealt with written records but the recent emphasis on the importance of the experimental practice, its instruments and visual aids have revealed the central role of the material culture as a source of inspiration for new creative ideas. However, this growing awareness faces the difficulty of relating objects, which are often hosted in different institutions. Thanks to the possibilities provided by the application of information technologies, the sources of the history of science can be investigated within a wider perspective, which both encompasses and enhances its exceptional variety. Through the presentation of different projects developed at the Museo Galileo in Florence, I shall illustrate the main direction of research we have undertaken to achieve the integration between the literary and the material records of science.

The discovery and examination of primary sources still constitute two of the most distinctive features of our understanding of the past. In the history of science, however, such features are rendered problematic by the fact that the classes of primary sources are as varied as those of the natural sciences. This means that, in contrast to other kinds of history which are mostly centred on literary and manuscript sources, the history of science also has to deal, or rather should deal, with many varied classes of three-dimensional sources such as instruments, specimens, tools and, to some extent, milieus. The material realm of the scientific endeavour is particularly apparent in contemporary science, where discoveries can hardly be disentangled from the laboratory and instruments from which they stem. As recent studies have shown, this is also true of the past.¹ Yet as much as recent historians emphasise the central importance of instruments and material tools, they have rarely dealt with the physical objects, relying almost exclusively upon the sole authority of their literary descriptions. In painting such a picture, though, we should not be too quick to blame historians of science for being superficial in their research: the situation is objectively difficult and the relationships, for example between instruments and texts, are not as obvious as they should be.

For a long time, and to an even greater extent today, scientific heritage was not regarded as a primary source in its own right, but as a constellation of fragmented classes of objects that, over the course of time, have been scattered and dispersed in very differ-

ent institutions. The handling of scientific collections represents a typical demonstration of this approach. When a celebrated scientist died and left his collection to the State, it was thought to be senseless to preserve this scientist's papers, books, instruments and natural history collection in one exclusive institution, as was customary during the sixteenth and seventeenth centuries.² Thus when it came to deciding what to do with a scientific collection, the decision was, and often still is, the following: the manuscripts went to an archive or to a library; the instruments went to a museum or a university if any of these were willing to accept them (in most cases the reply was negative); and the natural history collection was absorbed by museums without any concerns as to the objects' origins or owner. This dispersal is aggravated by the different methodologies and standards of classifications that have been adopted at different institutions. Even at libraries, which are the oldest institutions charged with the preservation of our historical heritage, the standards of cataloguing and classifying books are different to those used to catalogue and classify manuscripts. And the fact that a collection of books and manuscripts belonged one owner does not hinder its dismemberment. As scientists owned, in addition to manuscripts and books, also instruments and natural collections, the classification of their belonging was further complicated by the inclusion of more standards of cataloguing, such as those devoted to minerals, botanical specimens, photographs, scientific instruments and so on. It is worth pointing out that each and every one of these standards of cataloguing has its own well-defined disciplinary boundaries, rigid international protocols of input data and a highly specialised educational curriculum. Within this context, it is obvious that the assessment of a collection or a donation as a whole is not tenable. It is not surprising, then, that for historians of science the spread of collections has generated an artificial hierarchy of sources, with the literary ones on the top, which has made difficult, if not impossible, a rigorous and comprehensive historical reconstruction of the scientific endeavour. The recent development of information technologies has represented a powerful means to challenge this state of things and it proved to be particularly useful in changing the prevailing attention to theories and ideas that for such a long time have characterised the history of science. With the development of the history of science and the refinement of historiographic approaches, it has become clear that the reconstruction of the development of scientific theories and technologies should be carried out precisely by attempting to rebuild those connections and material contexts which are no longer recognisable through the traditional approach.

In 1997, together with Andrea Scotti and Daniele Nuzzo, I developed *Pinakes*, a prototype of a relational database, with the requisites we were looking for. *Pinakes*³ has been designed as a new model for documentary classification within the discipline of the history of science. The main task in building such a database was to create an application designed to implement different existing methods of classification as well as to create a standard description of non-uniform objects (texts, manuscripts, instruments, specimens, etc.).

A system based on this model, resulting from an analysis of heterogeneous sources, must focus on ensuring that the descriptive methods adopted to the task implicitly share the same logical procedure, and that they all have the same logical target.

For these reasons, in order to create *Pinakes* we had to develop a so-called *generalist model*, able to deal with information which was neither qualitatively nor structurally homogeneous. In other words, we wanted to build a structuring tool capable of managing and indexing data, independently from the fact that the datum might be a scientific instrument, a painting, a text or a mineral. At the same time, we needed to store and qualify the relationships existing between all these objects and their descriptions.

Pinakes also had to satisfy a dual requirement: to be a tool for very specialised documentary investigations and, at the same time, to serve a large-scale classification campaign in archives, libraries, museums, archaeological sites etc. In some cases, like in *Panopticon Lavoisier*, this tool reflected both requirements, providing both an accurate and a very specialised classification of each kind of documentary source connected with the life and work of Antoine-Laurent Lavoisier and a large-scale classification campaign. In fact, Lavoisier's case proved to be an excellent example of the variety of sources a historian of science has to deal with. For reasons that I have explored elsewhere⁴ most of Lavoisier's collections of instruments, books, manuscripts and natural specimens have survived and were kept together until the end of World War II. However, in the early 1950s, just about at the time when the history of science was becoming an independent discipline, Lavoisier's collections were dispersed to the following institutions:

The collection, which had once existed in the French chemist's laboratory and which – in its entirety – traced the build-up to the first great revolution in chemistry was now annihilated. The instruments, without the manuscripts describing their operation, rapidly became nothing more than mysterious and silent objects and rhetorical symbols of an obscure form of experimentation.

Institution	Items
Archives de l'Académie des Sciences – Paris	5,000 manuscripts
Musée Lecoq – Clermont-Ferrand	4,000 minerals
Kroch Library (Cornell University), Ithaca, New York	Approx. 600 items, mostly books from Lavoisier's library, numerous manuscripts and some memorabilia
Musée des arts et métiers – Paris	558 items (mostly instruments and some natural specimens and furniture)
Bibliothèque Municipale et Interuniversitaire de Clermont Ferrand	Approx. 250 manuscripts
Bibliothèque de l'Institut – Paris	89 books and manuscripts
Bibliothèque Interuniversitaire de Bordeaux	64 books from Lavoisier's library
Private collectors and Lavoisier's heirs	An undetermined number of books, manuscripts and iconographic material is still in the hand of the heirs of private collectors (mostly French)

Table 1. Lavoisier collections.

The manuscripts, which classified collections of natural history, fossils and minerals, became devoid of meaning and the minerals and the library inevitably suffered a similar fate. What can a historian do in the face of such dispersion, what may any visitor to a museum glean from an instrument which has been removed from its epistemological and material context?

Pinakes was designed to address these difficulties and help recreate the context of production of Lavoisier's scientific work. Within this aim, texts were no longer the most important sources to deal with, and all the surviving items of Lavoisier's collection have been classified without any given hierarchy. The project *Panopticon Lavoisier* began in 1997 and



Figure 1. Panopticon Lavoisier.

since its first release on the web in 1999⁵ it has been constantly updated and it is now close to completion.

Between 1997 and today, I have filled the database with the following types of data, which are all available via guided consultation through a menu (Fig. 1)

- **Chronology.** The chronology offers a comprehensive survey of Lavoisier's life and scientific career (1743–1794). Around 1,000 events are arranged in chronological order, most of them with a link to the relevant digital resources available in the database.
- **The manuscript collection.** The inventory, fully integrated in the database, aims at recording all of Lavoisier's manuscripts. While the 5,585 manuscript documents kept at the *Archives de l'Académie des sciences* and other host institutions have been included, their digital acquisition is in progress and thus far *Panopticon* has published 13 *Registres de laboratoire* and all of Lavoisier's *carnets de voyage*. The manuscripts are arranged by shelf mark, but they can also be viewed as a gallery of images.⁶
- **The instrument collection.** In the guided catalogue of Lavoisier's instruments, principally kept at the *Musée des arts et métiers – CNAM* (Paris), the objects have been arranged by inventory number. The list includes also those items that, since their donation in 1954, have been lost. Since not all the items in the Lavoisier collection kept at the *Musée* are instruments, in order to search for a complete list of the collection it is necessary to click on the institution's name. Although they have not yet been located, a few instruments were donated by Lavoisier's heirs to the University of Clermont-Ferrand, and a few



Figure 2. Pinakes gallery.

have remained with George Dupont since 1954. The consultation of this section offers two alternative options: a Gallery (an overview of all the available digital photos of instruments), and a Catalogue (the complete list of the instruments arranged by inventory number). (Fig. 2)

- **The library.** The catalogue of Lavoisier's library was published in 1995 by me (*Bibliotheca Lavoisieriana. The Catalogue of the Library of Antoine Laurent Lavoisier*, Firenze: Leo S. Olschki). In 1998, a new collection of books bearing Lavoisier's bookplates were found at the *Bibliothèque de l'Institut*. The digital catalogue of Lavoisier's library includes all the new findings and, in a few cases, the digitisation of a few titles. The total number of titles includes 2,610 bibliographic records. In the guided visit to the database, the books are arranged alphabetically by author as one would expect by consulting the catalogue of a library.
- **The collected works.** The whole of the national edition of collected works by Lavoisier (*Oeuvres*, Paris, 1862–1893, 6 vols.) has been scanned and the volumes may be easily browsed. The digital acquisition of the text is also available. The digital edition of Lavoisier's posthumous *Mémoires de physique et de chimique*, which includes some important memoirs not published in the *Oeuvres*, is also available. Each memoir and work has been linked to its original manuscript version, its first edition, and its bibliography, if any.
- **Lavoisier's bibliography.** The bibliography of Lavoisier's own published writings, arranged in chronological order, is the result of an editing of Denis I. Duveen's and Herbert S. Klickstein's *A Bibliography of the Works of Antoine Laurent Lavoisier* (London: Dawsons, 1954) with Supplement (London: Dawsons, 1965). The bibliography is constantly updated.
- **Bibliography on Lavoisier (1794–).** A complete bibliography of the secondary literature on Lavoisier has never been published. The present survey includes the selections made by Denis I. Duveen, William A. Smeaton, Patrice Bret and Keiko Kawashima (altogether some 750 titles) and published more than 2,000 items. This list of articles, essays and books on Lavoisier are listed by authors in alphabetical order.
- **The iconography.** The iconography of the Lavoisiers includes the known paintings, sculptures, painting, engravings, etchings etc. portraying the French chemist and his wife, Marie-Anne Pierrette née Paulze. The consultation of this section offers two alternative options: either a Gallery (an overview of all the available digital illustrations),



Figure 3. Lavoisier's iconography.

or a Catalogue (a complete list of the iconographic items arranged by author). (Fig. 3)

- **Digital library.** This menu offers all the printed material (books, memoirs, articles, pamphlets, etc.) that is available in a digital version. The titles are arranged by authors in alphabetical order and the list is constantly updated. This menu also offers all the available textual resources by and related to Lavoisier. The texts may belong to the following classes: 1. Texts from the *Oeuvres*; 2. Transcription from manuscripts (either by or related to Lavoisier); 3. Texts from Lavoisier's library; or 4. Texts from bibliographic references. The texts are arranged by author and by title.
- **Experiments.** The list offers a chronological list of Lavoisier's main experiments (220 entries). Each experiment is linked to its literary and bibliographic sources within the database.
- **The mineral collection.** The *Muséum d'histoire naturelle Henri-Lecoq* of Clermont-Ferrand (Puy-de-Dôme, France) preserves Lavoisier's collection of minerals. The catalogue of the collection, under the direction of Stephane Pelucchi, will be soon completed, and it has been achieved by using the SN-BBASE database. The migration of the data into *Panopticon* will follow after approximately a year of delay. The *Muséum* also preserves three manuscripts describing the collection: one by Lavoisier (incomplete), and two made after the donation, during 1856–57 and 1881 respectively. The collection consists of specimens partly preserved in 993 blown-glass receivers. These have been inventoried without any specific information other than their size. They contain not only minerals and fossils (of which there are



Figure 4. Lavoisier's minerals.

909), but also vegetable specimens (73), animal specimens (5) and archaeological items (5). (See Fig. 4)

All of these collections are integrated within the database, and one may explore the possibilities of searching by opting for an open advanced search rather than the guided menu. Through this option, the user can search by text, shelf mark, material (of the objects), subject/s, person, profession, date, periods, date ranges, city, institution, etc.

By aligning these findings with the three-dimensional sources described in *Panopticon*, we were able to achieve a very refined degree of analysis and no doubt explore Lavoisier's works beyond what traditional means of historical research would allow.

Panopticon Lavoisier emphasises not only the extraordinary relevance of the application of information technologies to analytic research, but also – and especially – the importance of these technologies in enhancing our understanding of scientific heritage objects and in helping historians to better appreciate the multifaceted structure of the scientific endeavour. We have applied Pinakes to other projects such as the collections of the Renaissance naturalist Ulisse Aldrovandi,⁷ (Fig. 5) mostly preserved in various institutions, libraries and museums of the University of Bologna and, though to a limited extent, to the Waller Collection of History of Science and Medicine of Uppsala University.⁸

However useful these resources were and, to some extent still are, their aim is mostly confined to specialised research within monographic themes and collections. They are stand-alone applications designed to respond to the needs of users. One may say that their main limitation,



Figure 5. Aldrovandi

if this is the right word, is of being a tool created by users for the benefit of other users.

While an open source version of *Pinakes* is under preparation and will still be applied to monographic themes and collections, we became aware of the fact that it was impossible to go beyond this limited scope of application without confronting the issue related to the different existing standards of classification adopted for the cataloguing of cultural heritage in general and scientific heritage in particular. As I have argued earlier, it is difficult if not impossible to expect the adoption of a unified standard that, as we have advocated, could meet the need of the community of users and not those of the professionals involved in the strategies of preservation. This being the state of the art, we have recently decided to adopt the logic behind *Pinakes* and to establish one recognised standard as the rule for the classification of objects which are often classified with other systems of classification. This was also the case with the collections held at the Museo Galileo in Florence.⁹ The museum was founded in 1930 after the successful *Esposizione Nazionale di Storia della Scienza* which was held in Florence in 1929 and exhibited some 10,000 items related to the history of Italian science. With its Renaissance mathematical instruments, the Galileo and Cimento collections of astronomical and meteorological instruments and the Lorraine collections of mechanical and electrical machines, the permanent exhibition of the Florentine museum is now one of the most important in its kind. The Museum and the institute attached to it hold the following collections (both physical and digital):

Instruments	3,834
Books	120,000
Iconographic resources	80,000
Digital library	10,000
Bibliography of the Italian history of science	60,000 records
International Bibliography of Galileo	20,000 records
Manuscripts	8,000 records
Minor collections (coins, medals, paintings, sculptures)	1,500

Table 2. Resources of the Museo Galileo.

In the last 20 years, the Florentine museum has invested resources to promote the organisation of temporary exhibitions, the results of which have been made less temporary with virtual exhibits.¹⁰ Until recently all these classes of items have been catalogued and classified according to different cataloguing standards, and remote users are consequently forced to consult each of them separately. The *Pinakes* experience however, showed us that integration was possible. We strongly believed that its usefulness was not restricted to researchers and scholars in the history of science, but also to any user interested in exploring a collection without having to learn about the reasons which have generated barriers between different classes of objects.

In order to be identified, an object – logical or physical – requires information which, in its essential features, is homogeneous. In most cases an object has an author, a textual description, a date, a physical description and so forth and each of these characteristics is common to most of the classificatory systems in use. As the largest archive of the Museo Galileo is the catalogue of the library three years ago, we began a project leading to the integration all our collections into the OPAC. Regardless of the database we used for this purpose, the main aim of the project is to use the existing standard for the bibliographic description MARC 21 and to extend its function to manuscripts, instruments, medals, photographs, virtual objects such as exhibitions, digital books and external sites. Although we are still far from the conclusion of this project, to date we have been able to successfully integrate books, instruments, photographs, manuscripts and digital resources.

If, for instance, a user searches the OPAC¹¹ browser by entering “Galileo lente”¹² (Galileo lens) the result of her/his query will produce 122 records of which 73 are texts, 9 are instruments or objects, 17 are

manuscripts and 23 are iconographic material. Among the items 12 are provided in full text and 21 are not owned as they are either described in the above mentioned bibliographies or listed in online exhibitions.

The result of the query shows that the user will have the opportunity to study the history of the objective lens of Galileo's telescope by perusing different classes of items and not, as it is customary in most digital libraries, by circumscribing the research the exclusive authority of the texts.

Once this project is concluded, all the resources of our museum will be integrated in a database, the structure of which will respond to the widely diffused and easily exportable standard of Marc 21. By this strategy we will ensure the professionals involved in the input data that their work will be not perishable; at the same time we will ensure users the right to have fully integrated access to resources which are historically, logically and at times physically related. This kind of access will not only change the way people view the collections of our museum but also the perspective by which the scientific endeavour is usually perceived by the public. By making scientific instruments part of a literary and cultural story, the user will hopefully feel less powerless in front of the technical difficulties entailed in the appraisal of this kind of artefacts. IT therefore offers the instrument to go beyond both traditional museology and history of science and bridge the gap between highly specialised disciplines and public curiosity surrounding science and its past. We are aware that more work has to be done, but we feel confident that the path we have taken will lead us to consider the preservation and classification of scientific heritage in a new, more comprehensive perspective.

I wish to thank Stefano Casati and Alessandra Lenzi for providing me useful information concerning the collections of the Museo Galileo. Parts of this paper have been presented on earlier occasions and in articles (one co-authored with Andrea Scotti) cited in the references.

References

- 1 For a useful synthesis of the topic, see Jim Bennett, “A Role For Collection in the Research. Agenda of the History of Science?” in Görel Cavalli-Björkman, Svante Lindqvist (eds.), *Research and Museums*, (Stockholm: The Nobel Museum, 2008), pp. 193–209; see also my essay, “Towards an Archaeology of Scientific Knowledge: Museums and the History of Science” in *Research and Museums*, Görel Cavalli-Björkman and Svante Lindqvist ed. (Stockholm: National Museum of Fine Arts and Nobel Museum, 2008), pp. 116–134.
- 2 Ulisse Aldrovandi’s extraordinary collection of manuscripts, books, xilographies, natural specimens bequeathed to the University of Bologna was kept together for over two centuries after the death of the Bolognese naturalist, and it was not until the first decades of the twentieth century that it was scattered into various departments of the University. The Accademia del Cimento’s collection of instruments faced a similar fate, so much so that by the end of the 18th century, more than one hundred years after its closure, the Lorraine sovereigns made their priority to keep it within their most valuable collections.
- 3 For information concerning the authors of *Pinakes* and the projects that have been managed with this application, see <http://www.pinakes.org>. *Pinakes* has been promoted by the Institute & Museum for Science History of Florence in co-operation with the following European institutions: Blanceflor Foundation (Stockholm), CNRS (Paris); Office for History of Science, Uppsala University (Uppsala) and Uppsala University Library (Uppsala).
- 4 I have reconstructed the vicissitudes of this collections in the following essays: “Lavoisier’s Collection of Instruments: A Checkered History,” in Marco Beretta, Paolo Galluzzi and Carlo Triarico (eds.), *Musa Musaei: Studies on Scientific Instruments and Collections in Honour of Mara Miniati*, (Firenze: Leo S. Olschki, 2003), 313–334; Marco Beretta, “Sur les traces du laboratoire de Lavoisier,” *La revue. Musée des arts et métiers*, 41 (Mai 2004) :14–23 ; and “Collected, Analyzed, Displayed: Lavoisier and Minerals” in *From Private to Public. Natural Collections and Museums*, Marco Beretta ed. (Sagamore Beach Mass.: Science History Publications/USA, 2005), pp. 113–140.
- 5 <http://moro.imss.fi.it/lavoisier/>
- 6 The gallery presents only the first thumbnails of the manuscripts, the whole of which are available in the individual records describing them.

- 7 Of Aldrovandi's museum two versions are currently available: one which was published in 2000 <http://www.filosofia.unibo.it/aldrovandi/> and one which is still in progress <http://moro.imss.fi.it/aldrovandi/>
- 8 <http://moro.imss.fi.it/waller/> The Waller project, based on the Pinakes experience, has been continued at Uppsala University Library <http://www.ub.uu.se/arv/waller/eindex.cfm>
- 9 www.museogalileo.it
- 10 A select list of the exhibitions of which a website was produced is available at the address www.museogalileo.it
- 11 <http://biblioteca.imss.fi.it/>
- 12 <http://colombo.imss.fi.it/IMSS/?uilang=en&q=galileo+lente&Submit=Go!>

THE FUTURE OF HISTORY¹

Emma Rothschild

History, which is a long encounter with oblivion, changes continuously over time. It also changes discontinuously with the technologies of listening, seeing, saving, and remembering. The early twenty-first century is not the first, or the most spectacular, of the epochs of sudden change in the technologies of history. The clarity of new documents pierced the darkness of the monarchy, “like a white electric light,” Jules Michelet wrote in his mid-nineteenth-century history of France. The personalities of the wars of religion dissolved in an “intense, implacable and terrible light;” the “lightning genius of modern times multiplies its miracles every hour, every minute, with steam and daguerreotypes, railways, the electric telegraph (where will soon be found the conscience of the globe).”² Electricity was a simile, for the romantic historians of the new age of global connections of the mid-nineteenth century, and it was also a technology of understanding.

But the twenty-first century is itself a time of intense and implacable change in the possibilities of historical scholarship, as the discussions at the Nobel symposium have made clear. What I would like to outline, in these concluding remarks, is some observations about three of the ways in which history seems to me to be changing, in relation to a period of the past which is strikingly evocative in our modern times, or the late eighteenth century in Europe; the period about which Robert Darnton talked so eloquently at the outset of the symposium, and which was in the understanding of many contemporaries a new age of light and insight, or of searching for enlightenment, as in the title of Tore Frängsmyr’s wonderful book. It was also a time, in Adam Smith’s description of 1763, when the “English nation think they have a right to information in political affairs by books and pamphlets;” “this age of information,” in the expression, in 1779, of the English essayist Vicesimus Knox.³

My observations will be rather practical and idiosyncratic, and I do think of history as an “art of doing,” or a collection of practices, “concrete forms of research and writing.”⁴ But I will conclude with some remarks about what I take to be the larger consequences of the changes which are now taking place: about the opportunities and the risks of the new technologies of history, especially in relation to the period of contemporary history in which oblivion is a very close neighbor; and about some possible improvements to be considered.

I start with the eighteenth-century revolution in printing and publishing, and the

technologies of intellectual and cultural history. In the last twenty years, almost everything has changed in intellectual history; the places in which historians work, the time of day (or night) in which we do research, or “searches,” the relationship to paper, and the relationship to scholarship, or learning. Let me take the year 1774, which was the year of the decision of the British House of Lords in the case known as “The Question of Literary Property.” The cause was widely considered at the time, as it has been by many recent scholars, to be an event of momentous importance in the history of copyright, bookselling, publishing, and intellectual property. It was also one of the great set-pieces of eighteenth-century legal, political and literary oratory. “Every man has a right to his ideas. Most certainly, every man who thinks has a right to his thoughts,” Lord Camden said in support of the reform of perpetual copyright; “but what if he speaks, and lets them fly out in private or public discourse? Will he claim the breath, the air, the words in which his thoughts are clothed? Where does this fanciful property begin, or end, or continue?”⁵

A few years ago, there were no books, pamphlets or newspapers published in 1774 that were available online. Today, there are 9,849 books published in that year which are available in full text on Google Books, accessed in Sweden in June 2009; 4,645 of them in French, 2,865 in English, and 111 in Swedish. There are 1,733 articles published in newspapers in 1774 which include the word “information,” and which are available online in the principal proprietary database of British newspapers.⁶ Even JSTOR has 37 articles which were published in 1774, and 30,382 articles which include the date, or the number “1774.” One of the Google Books volumes in Swedish is the proceedings for the year 1774 of the Royal Swedish Academy of Sciences, which includes a thrilling account of something described as a Perspective Drawing Machine. (The thrill is slightly mitigated by the classification assigned to the volume, which is for some reason “Juvenile Nonfiction.”)⁷

It is obvious that the new technologies have changed the conditions of scholarship in intellectual and cultural history. The activities of intellectual history over the past generation – to look at ideas in context, “to read an historical text... as an historical product,” to reconstruct “the actual intentions of the author,” or at least the texts that the author was reading, to describe how concepts changed over time – are simply more productive.⁸ As in the assembly line in *Modern Times*, this sort of history has speeded up. It is not so much a matter of historians being able to read faster, or of engines (“search engines”) reading more efficiently on

their behalf. It is rather that the ratio of reading to looking for books or negotiating access to rare books libraries has increased prodigiously. It is possible to read more (even without reading faster); to read rare books in the middle of the night, in one's nightdress; to read differently.

I have no doubt that this change in the technologies of intellectual and cultural history is a "democratization" of historical understanding, to use the expression of a great, almost Swedish archivist, to whom I will return, and who described the very large scale microfilming and duplication of manuscripts of the early 1950s as "a democratization of the archival reference service," to be made "readily and equally available to the scholars of all countries;" "a final break with the archivist's proprietary attitude toward his records."⁹ It is also an equalization, or a reduction in inequality, in historical scholarship, in that it sharply reduces the (relative) advantages of scholars who are in proximity to, and have privileged access to great libraries. There are literally millions of historians, or students of history, who can now study the European age of information. As Lord Camden also said, in 1774, "Knowledge and science are not things to be bound in such cobweb chains; when once the bird is out of the cage – *volat irrevocabile*..."¹⁰

There are costs, certainly, of the new technologies in relation to this sort of history. Erudition is not what it used to be. Browsing is a very different activity (although agreeable.) There are very serious dangers to a world in which it is only a canon of important works, as determined by the funders of digitization projects, which are studied. The fullest report of the debates over the 1774 literary property decision happens to be missing from the relevant volume (because of a nineteenth-century binder's error) in the only one of the expensive proprietary databases in which it is available.¹¹ The online searches flatten literary history, against the spirit of the history of the book and of reading, or of seeing.¹² They provide a better history of the "high" thought of theorists, than of the "medium" thought of publicists, or the "low" thought of everyone else.¹³

Let me turn, then, to a second objective of historical inquiry, or the history of the ideas and sentiments of people who were not intellectuals; to how it really was for the "other people" who are the subject matter of cultural history, or of the social history of ideas. I start, again, with the 1770s, and the consequences of digitization for archival research; and with my own research. Ideas exist in the minds of individual men and women, and not in a collective or national mind. So the old question of micro-history, and its relationship to large or macro-questions – of

representativity, or illustrativeness – is of central importance. I have been working for the last few years on trying to describe the ideas and information – including ideas about long-distance or global connections – of individuals in the age of enlightenment who were not themselves intellectuals. This has in part been a study of a large, obscure Scottish family, of four sisters and seven brothers, and of their servants and slaves. In the U.K. alone, I have been able to use, among publicly-funded archives, the national archives of England and Scotland, the county record offices of Clackmannanshire, Cumbria, Hertfordshire, Norfolk, Perth, and Yorkshire, the “A2A” or “Access to Archives” catalogue of more than 400 English archives, the online collection of digitized images of all wills proved in the Prerogative Court of Canterbury between 1384 and 1858, the East India Company records in the British Library, the collection of mortgages on slaves preserved in the Bristol University Library, and the records of 197,745 criminal trials in London between 1674 and 1913, available at the Old Bailey Online.¹⁴

I have also been working on a woman, Marie Aymard, who lived much further from the world of printing and publishing, in the town of Angoulême, a riverine port in west central France. Marie Aymard was the wife of a carpenter, who indentured himself, in 1753, to go to the island of Grenada.¹⁵ In 1764, she appeared before one of the notaries in Angoulême, to tell the following story. She had learned that her husband had been able to buy “a certain quantity of negroes and some mules,” had made a “small fortune,” and had set off to return to his family. But he had died in a charity hospital in Martinique on the way home, “after having left his fortune in the hands of M. Vandax a shipowner or merchant living on the harbour promenade of Martinique;” she wished to give her power of attorney to a sub-lieutenant in the merchant navy who was going to the West Indies, because “her indigence, and the distance forced her to defer, until present, her researches” (“ses recherches”). Now Marie Aymard could not read or write, as she informed the notary. But she had information about distant places. “These are the facts of which the deponent has been instructed at different times by certain people in the town of Angoulême,” she said.¹⁶ We can also get some sort of sense of the information she herself provided; the cloud of information, or misinformation in which she lived. There were her five children, and their husbands and wives. There were her creditors; a different notarial record lists five of them, as well as her property (six tin spoons, and ten “old bad chairs”). And there were her family and friends; when her daughter

married a tailor, also in 1764, there were eighty signatures on the marriage contract.¹⁷ So this seemed to me to be the multiplier effect of Marie Aymard's connections to the outside world; her own information society.

I do not really believe, with Marc Bloch, that the details of the technology of research are of gripping interest.¹⁸ But let me say a few words about the technology of looking for Marie Aymard. I have worked in the notarial archives in Angoulême since the mid 1990s, and they are impossible; a documentary ocean, in which a lifetime would not suffice.¹⁹ But something happened in 2006 which was for me like a light going on in the archives. This was the availability online of a digitized copy of the quarto inventory of the Charente notarial archives, published in 1906; an electric light amidst the bundles or "liasses."²⁰ It became possible to look for individuals, or families, or inventories, or transactions which involved Saint-Domingue. It also becomes possible to find lives in different archives, to find out what really happened with M. Vandax.

These new possibilities are related, in important ways, to the digitization of history. They are not, in general, a matter of the digitization of archives, but rather of the digitization of access to archives. Carlo Poni and Carlo Ginzburg, in their manifesto of micro-history of 1979, spoke of the name as the red thread, "Ariadne's thread" in the labyrinth of the archives.²¹ The technology of looking for individuals by their names has changed beyond recognition in the past thirty years. There are anomalies, of course. Some of the French departmental archives have digitized their own inventories. The catalogue in which I found Marie Aymard was digitized by Google, and is not, because of different copyright and fair-use regimes, accessible in France. Earlier this week there were 3,478 books with the words "inventaire" and "archives" in the title on Google Books accessed in the U.S., of which 2,278 were available in full-text; and 3,513 on Google accessed in the U.K., of which only 2 were available in full-text.

But I am convinced that the digitization of access to archives is not only an extraordinary valorization of the work of late nineteenth century provincial archivists and antiquaries. It is also a change in the character of archival scholarship. It takes just as long to read notarial documents as it ever did, even though I can now take copies home with me, and puzzle over the eighty signatures on Marie Aymard's daughter's marriage contract in the middle of the night. But it takes far less long to find the documents, and to browse in different ways. It has become possible, I believe, to explore different sorts of relationships between micro- and

macro-histories; not of representativity, but of the individuals' own connections.

My third illustration is different again. It has to do with the economic history of the late eighteenth century. Digitization, which is by definition quantification – a transformation of the world into numbers – can be thought of as having a particular affinity to the history of the economy. It is certainly the case that the increase in computing power has made possible quite new uses of the statistical material that has been left to us by the past. The digitization of printed and archival material has made it possible to identify and locate “lost” surveys made by the enthusiastic enumerators of what J.W. Goethe described as “our statistical times.”²² The new technology also makes it possible to generate new series out of old materials; I am thinking of the project to scan a part of the Paris notarial archives, and to extract information about credit transactions, work which has transformed understanding of the dynamism of the late *ancien régime* economy.²³

But there are, again, both risks and opportunities. One risk has to do with the growing distance between the users of easily downloadable statistical series, and the archival and other historians who understand the circumstances in which the series were arrived at. The uses and misuses of historical financial data have been much discussed in the context of the models (not exactly business models, but models of businesses) which failed so spectacularly in the financial crisis of 2008; when did the data start? Did they go back far enough? Did they measure the right things? There is a different risk, for historians, of extracting numbers from documents, such as notarial records, and leaving behind names and lives; a risk well evoked in the French expression for using, or “dépouiller” an archival series: literally, skinning, or skinning alive.

I would like to suggest, even, that digitization may present qualitative as much or more than quantitative opportunities; opportunities, that is to say, for an economic history, or a history of economic life, which is pluralistic in its use of sources and its conception of the lives of individuals in the past. Since the 1920s, the history of economic thought has been sharply distinguished from economic history, and the history of “high” or canonical economic thought has been distinguished from the thought of entrepreneurs, officials, salesmen, and “ordinary people.” Economic history has itself been identified increasingly by the techniques that it uses: techniques, to use the Nobel expression, of the “economic sciences.” It seems to me that the new technologies of history, which flatten so much

else, can also flatten these distinctions. One can, for example, study the ephemeral literature, the almanacs for merchants of the eighteenth century, and find a new geography of the small industrial towns of the center of France. One can locate business archives, uses of economic terms, and references to transactions in particular commodities. One can study tax codes, and parliamentary debates over copyright laws.

The benefits are potentially very large indeed, for a reinvention, without the paraphernalia of nineteenth-century historicist nationalism, of the “economic-cultural history” which was so ardently aspired to in the 1860s.²⁴ This would be a history that transgresses the assumption, imported from economics into history, that individuals in their economic lives are “rational fools,” or “strong, silent men who never speak.”²⁵ It would be a post-materialist history, not only in the familiar sense of going beyond the vulgar Marxian presumption of ideas as determined by material circumstances, but also in the sense of going beyond the other vulgar nineteenth-century presumption, of rational economic man. The press release for the only economic historians to have won a Nobel prize in economic sciences referred to their having “contributed to the elimination of irrelevant theories.”²⁶ The digital world may encourage a conception of the history of economic life as a “human science.”

To return, now, to the future. I believe that the new technologies of history of which I have tried to provide a glimpse pose several rather deep questions. The first has to do with the relationship between the historian, her technologies, and the archivists and librarians who are the intermediaries among them. The technologies I have been talking about are not “our” technologies, or historians’ technologies. They are changing extraordinarily rapidly, and they are very big business indeed. I do not think that we should be excessively romantic about a lost golden age of history, which was a very manly golden age, in which the historian-philosopher was also the historian-librarian and the historian-archivist. This was the epoch of David Hume, for example, who when he left public office, requested from King George III, and was granted, “the Liberty... of inspecting all the public Records and all the Papers in the Paper-Office.” Or of Barthold Niebuhr, with his vivid descriptions of the English bombardment of Copenhagen in 1801 – “I got dressed [and] packed up the archive... It is a battle that must be compared to Thermopylae” – and of the chance discovery of a manuscript of Ulpian in the Cathedral in Verona, in need of conservation, “I was obliged hastily to prepare for myself an infusion of gall-nuts.” Or of Jules Michelet, who was

head of the historical section of the Archives de France, where he worked for twenty-two years, traveling around the provincial archives, including the archives of the administrative region in which Marie Aymard lived, where he found that a substantial pile of records had been burned in the revolutionary wars, having previously been made into a effigy of William Pitt.²⁷ But I do think that the distance between historians, archivists and librarians is alarmingly great today, at a time, when the technologies of digitization are changing the archivists' and librarians' own worlds so rapidly.

The second deep question has to do with the relationship between the historian-scholar and other kinds of history: public history, which is also the history that surrounds the students of the historian-scholar or the historian-teacher; and family history, or genealogy. The digitization of archival records is inspired, to a great extent, by public and family history. The "canon" is determined by public history – consider, for example, a future in which the only wills available online in England were the 102 of which the U.K. National Archives write that "we have separated out the famous original wills" – at a time when historians are themselves interested in an extraordinary range of private ephemera.²⁸ "The dignity of source materials of history," the almost-Swedish archivist, Ernst Posner, wrote in 1941, had in the course of the twentieth century been "extended to an ever widening scope of documents," and the extension is continuing, now, in ways that have an odd and interesting relationship to public and popular history.²⁹ Yet the websites of public and private family historians are almost entirely unconnected (too unconnected, perhaps) to the scholarship of historians; a "family history" which is a genealogy of living individuals, conceived in the diachronic metaphors of a tree or a root ("Connecting Generations," or "Where Generations Meet.")³⁰

The third consequence is the deepest. All the new kinds of history I have talked about have been in different ways an engagement with fragments; pieces of the past. They have been incomplete, or an attempt to tell a true story about the past, on the basis of very disparate kinds of evidence and very partial series. They are also an involvement with a universe of evidence which changes continuously over time. I am sure, following Robert Darnton, that being a historian is rather like being a police reporter.³¹ But it is also a bit like being a private detective. In my history of the Scottish family, I realized that I was doing something – doing exactly the same thing every few days, or searching for the same name – that had a distinct resemblance to a Hegelian bad infinity. (I did

eventually find out who the youngest brother married, when a Liverpool newspaper of the 1770s was digitized, and I found out something, too, about his father-in-law's connection to the slave trade.)

I am very conscious, in this respect, of how far we have come from the Rankean ideal of archival research; the ideal of going to an archive, finding a series, and reading from beginning to end. Now historians know – Ranke certainly knew – how much artifice goes into the creation of a series, or a “fonds.” So it is right that we should be conscious of the incompleteness of historical evidence. But it seems to me that this sense of incompleteness is now a very important part of the historian's present and future. In philosophical terms – eighteenth-century terms, in particular – it challenges us to think even more probabilistically than in the past about our material. It puts new demands on our criteria of scholarly responsibility. Let me try a post-post-modern simile, inspired by references to progressive jpegs or variable resolutions. It is from the description of the Seadragon software as a “multi-resolution experience” of “super-smooth zooming over billions of pixels,” a “seamless web of images and information, allowing you to browse a virtual universe of interconnected scenes that constantly evolves and changes over time.”³² Or a simile from the great legal historian F.W. Maitland, who was one of the inspirations of the early *Annales* historians; of history as a process of “inference and generalization,” in which “some of the best because the truest history books are those which are professedly fragmentary,” and “by their every page impress upon the reader that he has only got before him a small part of the whole tale.”³³

Let me turn, in conclusion, to the risks of the new world of history and digitization, in reverse order of seriousness. The first is of dangers internal to scholarship, including dependence on a canon, loss of critical understanding of archives and libraries, and changing views of scholarly standards. These dangers are not trivial, but they are in essential respects within our capacity to affect, as a scholarly community.

The second problem is of inequality, in the world of democratization which Ernest Posner described. We have talked at this symposium about the inequality of access to proprietary resources, and I too, like Lisbet Rausing, am deeply traumatized by my intermittent encounters, even as a citizen of the richest scholarly republic in the world (the quasi-republic of Harvard) with the notice “you do not have access to this resource at your current location.” The one thing we have not talked enough about is the universe outside the Atlantic world. India is the digital and digitization

centre of the world economy, and optical character recognition in Hindi and other Indian languages is progressing very fast. But the resources available to scholars in India, and working on Indian history, are very different indeed from the resources we have been discussing. I think there is also a deeper question. Even if a million historians in India will, fairly soon, have access on-line to hundreds of thousands of eighteenth-century European works – which I believe will be the case – will they also be able, if they decide to become scholars of France in 1770s, to visit the Royal Swedish library, and to hold in their hands the volumes we held yesterday? Will there be two kinds of historians, virtual historians and historians who are in a position to “touch the real world”; an inequality by place and institution, which is also an inequality of beauty?³⁴

The third risk is of loss of access to online resources. I cannot imagine working, now, without Google Books. But this access is subject to monopoly or oligopoly power. I do not think that any of us who listened to the discussion at this symposium could feel confident that what was outlined as the “core value proposition” of Google, for example, is one that excludes a very substantial change, in the future, in the way access is priced, or costed.

The fourth and most serious risk is not of loss of access but of loss. In 2009 a large part of one of the most important early modern archives in Europe, in Cologne, collapsed into mud and rubble.³⁵ The National Mission for Manuscripts in India is engaged in the conservation, in Niebuhr’s sense, of manuscripts and printed works which fly away into dust at a breath; the heritage of the economic super-power of the next century.³⁶ Google’s own digital collections could be lost, even if they are not stored on a solar barge at an undisclosed location in the Pacific Ocean.³⁷ There is above all the possibility of the loss of an entire period of our own contemporary history, or of the digital age since the mid-1990s, which is so much closer to oblivion than the eighteenth century, and which is, or may become a black hole in history. Almost all national archives are anxious about the preservation of digital records; the records, for example, of the regulation of the present economic and financial crisis. The situation of the archives of international institutions is even more difficult. So is the situation of business enterprises; what will be the records of Google, for example, for economic and business historians fifty years from now? Even the archivists of banks, with statutory requirements in relation to records, are anxious; all this history could be gone.

My practical suggestions, then, are modest. They are, first, that uni-

versities and archives should as a matter of urgency seek ways to bring young historians back into connection with libraries and archives; a requirement, for example, in advanced training, not only for languages but also for an internship in an archive or library. Second, and in relation to the inequality of beauty, that library and archive digitization projects should be accompanied by programs to provide funding for young scholars, and especially scholars in other countries and in unopulent institutions, to visit the archives and libraries where the original works are. Third, that the endangered archives projects, which Lisbet Rausing has pioneered so admirably, should be enhanced. Fourth, that in a time of extreme economic difficulty, including for young scholars, our institutions should think of engaging young scholars in the great collective exercise of understanding the problems of physical and digital preservation.

Ernst Posner, to whom I have referred several times, was a Prussian archivist and editor of *Acta Borussica*, an expert on the international politics of archives, who was imprisoned in the Sachsenhausen concentration camp in 1938, and then came to Sweden. He was unable to stay in Sweden, but he was supported, before he went on to the U.S., by the staff of the Swedish National Archives, who invited him to give a series of lectures on international archives.³⁸ One of them, in May 1939, was about the "Survey of Federal Archives," organized by President Roosevelt's Works Progress Administration for "the so-called 'white-collar unemployed,'" about the collections saved from destruction, and about the new "archives consciousness among the clerks."³⁹ Is this not a time, again, for a survey of national and international archives in the digital world? Finally, I return to 1774, and London booksellers, and the question of literary property. Is it not time for a new history, and a new future, of copyright and of fair use? For the bird is out of the cage; *volat irrevocabile*.

References

- 1 I am very grateful for helpful discussions with Robert Darnton, Anthony Grafton, and Lisbet Rausing, and with Sunil Amrith, Caitlin Anderson, D'Maris Coffman, Leigh Denault, Philipp Lehmann, Amy Price, Julia Stephens, David Todd, Robert Watson and other participants in the Digitization of History project at the Centre for History and Economics, and in the Mellon project on Exchanges of Economic and Political Ideas since 1750.
- 2 Jules Michelet, *Histoire de France au dix huitième siècle: Louis XV et Louis XVI* (Paris, 1867), p. ii; *Histoire de France au seizième siècle: les guerres de religion* (Paris, 1836), p. vii; *Les femmes de la révolution* (Paris, 1855), p. 319.
- 3 Tore Frängsmyr, *Sökandet efter Upplysningen* (Uppsala, 1993); Adam Smith, *Lectures on Jurisprudence*, ed. R.L. Meek, D.D. Raphael and P.G. Stein (Oxford, 1978), p. 303; "On Preaching and Sermon Writers," in Vicesimus Knox, *Essays Moral and Literary*, vol. 2 (London, 1779), p. 157.
- 4 Jacques Revel, "Présentation," in Revel, *Un parcours critique: Douze exercices d'histoire sociale* (Paris, 2006), pp. 10, 21.
- 5 Speech of Lord Camden of February 22 1774, *The Parliamentary History of England, from the Earliest Period to the Year 1803* (London, 1813), vol. 17, col. 997. On the importance of the 1774 decision, see Mark Rose, *Authors and Owners: The Invention of Copyright* (Cambridge, MA, 1993), and William St. Clair, *The Reading Nation in the Romantic Period* (Cambridge, 2004).
- 6 The 17th and 18th Century Burney Collection of some 700 bound volumes of newspapers, available through Gale Cengage.
- 7 *Kongl. Vetenskaps Akademiens Handlingar För År 1774*, vol. 35 (Stockholm, 1774), "Perspectiv-Ritnings-Machine," p. 15; accessed in Sweden on Google Books, June 24 2009.
- 8 Richard Tuck, "History of Political Thought," in *New Perspectives on Historical Writing*, ed. Peter Burke (Cambridge, 1991), 193–205, p. 194.
- 9 Ernst Posner, "The National Archives and the Archival Theorist" (1954), in *Archives and the Public Interest: Selected Essays by Ernst Posner*, ed. Ken Munden, with a new introduction by Angelika Menne-Haritz (Washington, DC, 1967; repr. Chicago, IL, 2006), p. 135; Solon J. Buck, "The Archivist's 'One World,'" *The American Archivist*, vol. 10, no. 1 (1947), 9–24, p. 19.
- 10 Speech of Lord Camden of February 22 1774, *Parliamentary History*, vol. 17, cols. 1000–1001.

- 11 On the circumstances of the error, see Emma Rothschild, "A Note on Literary Property," August 21, 2007, available at http://www-histecon.kings.cam.ac.uk/research/digitization/note_literary_property.html.
- 12 On the history of the book, see Robert Darnton, *The business of enlightenment: a publishing history of the Encyclopédie, 1775–1800* (Cambridge, MA, 1979) and Roger Chartier, *L'ordre des livres: lecteurs, auteurs, bibliothèques en Europe entre XIV^e et XVIII^e siècle* (Aix-en-Provence, 1992).
- 13 On "high," "medium" and "low" thought, see Emma Rothschild, "Arcs of Ideas: International History and Intellectual History," in *Transnationale Geschichte: Themen, Tendenzen und Theorien*, ed. Gunilla Budde, Sebastian Conrad and Oliver Janz (Göttingen, 2006), 217–226.
- 14 See <http://www.nationalarchives.gov.uk/documentsonline/wills-deathduties.as>; <http://www.oldbaileyonline.org/>; and <http://www.nationalarchives.gov.uk/a2a>.
- 15 "Marché d'engagement de Ferrand et Delorière à M. Cazeau de Roumillac," December 15, 1753, Caillaud, notary (November–December, 1753), Archives Départementales de la Charente, Angoulême (ADC), 2 E 259.
- 16 "Procuration par Marie Aymard," October 16, 1764, Jean Bernard, notary (July–December, 1764), ADC, 2 E 153.
- 17 "Vente de meubles par Marie Aymard à Gabriel Ferrand son fils," January 10, 1764, Bernard, notary (January–June 1764), ADC, 2 E 153; "Contrat de mariage de Estienne Allemand et François Fermand," November 9, 1764, Bernard, notary (July–December, 1764), ADC, 2 E 153.
- 18 Marc Bloch, *Apologie pour l'histoire ou métier d'historien*, ed. Étienne Bloch (Paris, 2007), p. 82.
- 19 On notarial archives in relation to overseas histories, see Gabriel Debien, *Les engagés pour les Antilles (1634–1715)* (Paris, 1952) and Pierre Goubert, "Une belle enquête: Saint-Domingue au XVII^e siècle," *Annales Histoire, Sciences Sociales*, vol. 7, no. 3 (1952), 329–331.
- 20 *Inventaire sommaire des archives départementales antérieures à 1790, Charente, archives civiles – série E (Art. 1736–3040)*, ed P. De Fleury and J. De La Martinière (Angoulême, 1906).
- 21 Carlo Ginzburg and Carlo Poni, "Il nome e il come: scambio ineguale e mercato storiografico," *Quaderni Storici*, no. 40 (January–April 1979), 181–190, p. 185.
- 22 Observation of September 11 1786 on the Bolzano fair, in J.W. Goethe, *Italianische Reise*, ed. Christoph Michel (Frankfurt, 1976), vol. 1, p. 35.

- 23 See Philip T. Hoffman and Jean-Laurent Rosenthal, "New Work in French Economic History," *French Historical Studies*, vol. 23, no. 3 (2000), 439–453; Philip T. Hoffman, Gilles Postel-Vinay and Jean-Laurent Rosenthal, *Priceless Markets: The Political Economy of Credit in Paris, 1660–1870* (Chicago, 2000); and on the "ARNO Bases" of the "principal information contained" in all the notarial minutes of the Paris notaries for four illustrative years, 1551, 1751, 1761 et 1851, see <http://www.archivesnationales.culture.gouv.fr/arn/>.
- 24 The expression *nationalökonomische Culturgeschichte* was used by the historical economist Bruno Hildebrand, in 1863, to describe a new historical political economy, "in the context of the history of the total political and legal development of nations and statistics." Bruno Hildebrand, "Vorwort," in *Jahrbücher für Nationalökonomie und Statistik*, vol. 1 (1863), p. 3.
- 25 Amartya Sen, *Choice, Welfare and Measurement* (Oxford, 1982), pp. 9, 84.
- 26 "Robert W. Fogel and Douglass C. North," available at http://nobel-prize.org/nobel_prizes/economics/laureates/1993/press.html.
- 27 On David Hume, see letters from Hume to Andrew Millar of July 17 1767, and to the Comtesse de Boufflers on November 27 1767, in *The Letters of David Hume*, ed. J.Y.T. Greig (Oxford, 1969), vol. 2, p. 151, p. 172; "the King himself has been pleased to order that all the records and public offices shall be open to me, and has even sent for some papers from Hanover, which he thought would be useful." On Barthold Niebuhr, see letters from Niebuhr to Dore Hensler of April 3–4 1801, in *Die Briefe Barthold Georg Niebuhrs*, ed. Dietrich Gerhard and William Norvin (Berlin, 1926), vol. 1, pp. 258–260; and to Karl von Savigny of September 4 1816, in *The Life and Letters of Barthold Georg Niebuhr*, ed. and transl. Susanna Winkworth (London, 1852), vol. 2, pp. 52–53. On Michelet in Limoges, see Jules Michelet, *Journal (1828–1848)* (Paris, 1959), p. 209.
- 28 See <http://209.85.229.132/search?q=cache:jiCeeIV34goJ:https://www.nationalarchives.gov.uk/documentsonline/prob1wills.asp>.
- 29 Posner, "Effects of Changes of Sovereignty on Archives" (1941), in *Archives and the Public Interest*, p. 174.
- 30 See www.scotlandspeople.gov.uk and www.familysearch.org.
- 31 Interview with Robert Darnton, August 31 2003, available at <http://www.booknotes.org/Transcript/?ProgramID=1743>.
- 32 See <http://labs.live.com/photosynth>, accessed on August 10, 2007. I

am grateful to Professor Anthony Grafton for conversations about Seadragon and Blaise Aguera y Arcas.

- 33 "The Body Politic," in *The Collected Papers of Frederic William Maitland*, ed. H.A.L. Fisher, vol. 3 (Cambridge, 1911), 285, 294.
- 34 On touching the real world, see Arlette Farge, *Le Goût de l'Archive* (Paris, 1989), p. 18.
- 35 See <http://www.stadt-koeln.de/5/kulturstadt/historisches-archiv/>.
- 36 On the National Mission for Manuscripts, established by the Government of India in 2003, see <http://www.namami.org/>.
- 37 I am grateful to Alexandra Berry, Leigh Denault and Robert Watson for discussions of the conservation of digital records.
- 38 On Posner, see Paul Lewinson, "Introduction: The Two Lives of Ernst Posner," in *Archives and the Public Interest*, 9–21; Wolfgang A. Mommsen, "Ernst Posner, Mittler zwischen deutschem und amerikanischem Archivwesen Zu seinem 75. Geburtstag," *Der Archivar* (1967), 217–230; *An Interrupted Past: German-speaking Refugee Historians in the United States after 1933*, ed. Hartmut Lehmann and James J. Sheehan (Cambridge, 1991.)
- 39 Posner, "Archival Administration in the United States" (1939), in *Archives in the Public Interest*, p. 128.

Notes on contributors

Nicolas Barker is editor of *The Book Collector*. He has written many books, among them *Stanley Morison* (1972), *Bibliotheca Lindesiana* (1977), and *Aldus Manutius and the Development of Greek Script and Type in the 15th Century* (2nd edn 1992). He is the editor of recent editions of John Carter's classic *ABC for Book Collectors*.

Marco Beretta is Professor of History of Science at the University of Bologna and Vice-director of the Museo Galileo. He is the editor of *Nuncius – Journal of the Material and Visual History of Science* (Brill, Leiden). He has published extensively on the history of chemistry and chemical technology. He has also been editor of several websites devoted to the sources of the history of science and medicine and editor, together with Iolanda Rolfo, of the site www.museogalileo.it

Martin Bossenbroek is a historian and has written various books and numerous articles on the nineteenth and twentieth century history of the Netherlands and its former colonies. He has been director of Collections and Services and acting Director-General of the Koninklijke Bibliotheek from 2004 till 2010. In November 2010 he picked up his academic career, as an Associate Professor of Modern History at Utrecht University.

Gregory Crane is Professor and Chair of the Department of Classics, Winnick Professor Technology and Entrepreneurship, Adjunct Professor of Computer Science, and a struggling student of Arabic at Tufts University. An author of books on Homer and the Greek historian Thucydides, he is also the Editor-in-Chief of the Perseus Digital Library, a set of digital collections under development for since 1987.

Alain Giffard is Directeur du Groupement d'intérêt scientifique "Culture-Médias&Numérique", Ministère de la culture et de la communication.

Magdalena Gram is Assistant National Librarian at the National Library of Sweden. She has a Masters in Library Information and Sciences and a Ph.D. in Art History. She has vast leadership experience in the museum sector (The Art Library, Nationalmuseum; Department of Archives, Library and Pictures, Nordiska museet) and has been engaged in numerous projects and inquiries regarding cross-domain collaboration amongst archives, libraries and museums.

Karl Grandin is Professor and Director of the Center for History of Science at the Royal Swedish Academy of Sciences in Stockholm. He has an MSc in solid-state physics and a PhD in history of science both from Uppsala University. His research has dealt with history of physics in the 20th century. Since 2006 he is the editor of the Nobel Foundation yearbook *Les Prix Nobel*.

Kevin M. Guthrie is an executive and entrepreneur with expertise in technology and not-for-profit management. Kevin holds a BSE in Civil Engineering from Princeton University and a Masters in Business Administration from Columbia University. Kevin was the founding president of JSTOR (1995) and Ithaka (2004). JSTOR and Ithaka merged in January 2010 to form a new organization (ITHAKA) with a mission to help the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways. Kevin is a frequent speaker on the impact of the transition from analog to digital technologies on higher education and scholarship and has authored numerous articles on these topics. Prior to launching JSTOR, he authored a book exploring the challenges of managing non-profit organizations entitled: *The New-York Historical Society: Lessons from One Nonprofit's Long Struggle for Survival* (Jossey Bass).

Terje Hillesund is Associate Professor at the University of Stavanger, Norway, where he teaches media and communication theory. His main research interests have been in printed media – newspapers, journals, and books – and their transformation in the age of computers, the Internet and handheld communication technologies. From 2000 to 2002, Dr. Hillesund led a project for Arts Council Norway in which a research team studied the potential impact of e-books on the Norwegian book industry. The results were presented in three reports in 2002. Dr. Hillesund has studied and published scholarly articles on subjects such as e-books, XML, digital text editions, open access and digital reading.

Anke Lüdeling is Professor for Corpus Linguistics and Morphology at Humboldt-Universität zu Berlin. With a background in linguistics as well as computer science she is working on using digitally available tools and quantitative methods to explore linguistic questions. Her main interests all have to do with linguistic variation: language change, foreign language acquisition, and register.

Lisbet Rausing is a Senior Research Fellow at Imperial College's Centre for the History of Science, Technology and Medicine. She was educated at the University of California Berkeley and Harvard University, where she also taught for eight years. She has written two academic monographs as well as numerous scholarly articles. She serves on various boards and committees including the Harvard Board of Overseers and Yad Hanadiv. She holds honorary doctorates from Uppsala University, Imperial College and SOAS, is an elected member of St Catherine's College, Cambridge, an Honorary Fellow of the British Academy, a Fellow of the Linnean Society and a Fellow of the Royal Historical Society.

Laurent Romary is Directeur de Recherche INRIA, France and guest scientist at Humboldt University in Berlin, Germany where he carries out research on the modelling of semi-structured documents, with a specific emphasis on texts and linguistic resources. He is the chairman of ISO committee TC 37/SC 4 on Language Resource Management, as well as member (2001–2007) then chair (2008–) of the TEI (Text Encoding Initiative) council. In the recent years, he led the Scientific Information directorate at CNRS (2005–2006) and established the Max-Planck Digital Library (sep. 2006–dec. 2008). He also contributes to the establishment and coordination of the European Dariah infrastructure and has recently been appointed as Dariah transitory director.

Emma Rothschild is Professor of History at Harvard University and Honorary Professor of History and Economics at the University of Cambridge. She is Director of the Joint Center for History and Economics and a Fellow of Magdalene College, Cambridge. She is the author of *The Inner Life of Empires: An Eighteenth-Century History* (2011) and *Economic Sentiments: Adam Smith, Condorcet and the Enlightenment* (2001).

Peter Suber is a Faculty Fellow at the Berkman Center for Internet & Society at Harvard University, a Research Professor of Philosophy at Earlham College, and a Senior Researcher at the Scholarly Publishing and Academic Resources Coalition (SPARC).

Adriaan van der Weel is educated at the Universities of Leiden, Dublin (Trinity College) and Sydney, and now he is Bohn Professor of Modern Dutch Book History at the University of Leiden, and lecturer in Book and Digital Media Studies, and English Literature. His research interests centre on the role of the medium in the transmission of knowledge and culture. His latest book is *Changing our textual minds: Towards a digital order of knowledge* (Manchester UP, 2011). Currently Van der Weel is European articles editor of *Digital Humanities Quarterly*; editor-in-chief of *Logos: Journal of the World Book Community*; co-founder–editor of the Scholarly Communication book series (Brill); and he serves on the committee of the Dr. P.A. Tiele-Stichting for the promotion of the study of the book.

ISSN 0081-9956
ISBN 978-91-7190-175-0



**CENTER FOR
HISTORY OF SCIENCE**
THE ROYAL SWEDISH ACADEMY OF SCIENCES