Information technology,
warden of our past?

by Ludo Milis

'*A surfboard with Internet capability, cars that drive themselves, gadgets that recognize forgotten acquaintances, the extinction of humans by super-computers. All are possible…*'[1] These are the challenging ideas launched by a panel of Silicon Valley innovators last February. We know they are right. '*What are the things that have really transformed our digital future?*' John Hennessy asked, an ICT specialist and president of Stanford University. '*It really was the personal computer, cellular wireless communication, and, of course, the Internet.*'

In the abstract of the aforesaid meeting on the internet, the reviewer Chelsea A. Young quoted some other important statements, more important for us, historians of the very first years of the new millennium. Over time, one of the panel members said, he had been struck by the general '*willingness to share information*'. I think that in our field, and in scholarly behaviour in general, this attitude may be confirmed. The academic community seems to be prepared to follow this path, in other words to honestly generalize their research topics and their results.

Yet another hot item had been addressed, it is '*scaling—the ability to accommodate the vast and increasing demand for information storage space on the Internet*'. Technically speaking no problem arises. Computer memories will always be able to store the increasing information people exponentially publish on the Internet.

Yet one is more and more aware of another issue, if not a problem. Indeed, how long will it be and remain possible to continuously supervise the overall contents? In fact, we have just begun developing information and computer technology. Only seventy years passed before the first computer was built, twenty five years before the personal computer appeared in people's home, and less before the Internet appeared. Browsers help us to search. Google and Yahoo are tremendous tools for

---

[1] *Stanford Report, March 5, 2009*: Ch. A. Young, *Digital revolutionaries discuss past, future of technology, by* (http://news.stanford.edu/news/2009/march11/technology-google-microsoft-hennessy-031109.html).

all of us, but very often it already by now turns out to be an extremely burdensome and tedious task to check the addresses we are dealing with. A lot of the URLs turn out to be out of date, or to have lost their importance, or do not reach any intellectually quality level, or simply does not meet the criteria we hoped for. What will it be in a near future: an overwhelming mass in which no one is able to find its way? And then, I leave a far future out of consideration. Today Google says it makes already eight billion sites searchable.[2] What will the number be decades from now?

Another major issue, if not a problem, is what one of the conference participants *'nicknamed the 'bit rot' problem, which has to do with the longevity of information on the Internet.'* This *'bit rot'* is linked to what I have just said, the abundance of information. The speed with which technology develops hardware as well as software, makes information obsolete after a very short time. In half a century, all in our lifetime, one has seen the introduction of the punched cards, then tapes arrived, and floppies of successively different dimensions and storage capacities, limited to a 1,44MB maximum. After fifteen years of success these got superseded by CDs, DVDs, USB-sticks, and external hard drives. It is difficult if not impossible to buy the necessary hardware to read the old floppies or, even more problematic, to recover their contents and save them on newer types of memory devices. My first computer at home, bought in 1983, had a memory of 64KB, corresponding to some twenty typewritten pages in the classical system. It was the Commodore 64, successful and affordable at that time. Now computers, PCs as well as notebooks, just for our personal use, reach gigantic storage capacity, due especially to the development and the commercial success of memory devouring files, such as video, games, and music. As a consequence of this development, memory and price setting proportionally grew in the opposite way. Last week I got an ad in my mailbox suggesting to buy at a very reasonable price a desktop with a maximum storage capacity of 2 TB (terabytes), in other words 2000 billion bytes.

Bit rot means that within undefined, but apparently short periods electronic information is not readable anymore, or only if transferred in an up to date system, and such an act should be repeated every time a new one comes into use. One of the participants at the recent meeting theorized: *'It's the year 3000. You've just done a Google search and you turned up a 1997 PowerPoint file. You're running Windows 3000. The question is: Does it know how to interpret a thousand-year-old PowerPoint file? And the answer is probably no.'* It is a nightmare for archivists! It is a nightmare in the institution that hosts us today. How can archivists possibly deal with the perennial goal of perennial document preservation? The whole evolution in ICT is going on so fast, even in the limited time span of some decades, that preservation is already by now a structural problem. What then in a

---

2 http://www.google.be/intl/nl/options/

future further ahead? We are historians, some of us are medievalists, and we find it normal and obvious that dozens of thousands of charters are preserved, thousands and thousands of manuscripts, and even under such circumstances we complain that our information it too scanty. We lament that whole segments of medieval people's lives cannot be reconstructed. We are even unable to scrutinize whole segments of their overall life style. Often we entirely fail to detect their mental concepts. It is a strange statement that the first writings were on stone, followed then by clay, papyrus, parchment, strong hand made paper, vulnerable industrial paper, up to the signals of telegraph and telephone, and the even more ethereal digital ways of communication. More and more information and communication facilities implied and imply increasing vulnerability in the preservation chain. And once again, we are just at the beginning of this tremendous technological revolution, speeding up its innovations to unimaginable dimensions. Some sort of 'writing' needs to be and will be invented and worked out to preserve information and to keep it searchable. I put 'writing' between quotation marks. Obviously, it will not be any traditional writing, using a pen, a typewriter, a keyboard, nor forming characters, words, and phrases. The support will not be paper or anything alike, not even disks or servers. I imagine it will be a new type of conventions, and I guess it will be something to be conceived through the brains, or directly in the brains. The more the functioning of human intelligence processes will be explored, the more parallel computer functioning will be invented and applied. But who am I to make a serious projection of the future avoiding that it would become a mere prediction, and in someone's eyes just fantasy?

I do not think that our meeting of today has to discuss these items. We do not need to find solutions for the big challenges that link the production and collection of actual information to what Marc Bloch called the '*métier d'historien*'.[3] Nevertheless, we need to consider all these ideas, directions, and evolutions, sometimes changing fuzzily.

The themes we have to address are more immediate and need to be framed within the limits of the *Porta Historica* project. What does its charter say? '*The Network shall bring together, at the international level, institutions that play a national or international role in facilitating historical research by editing and publishing historical sources in a broad sense. The objective… shall be to promote cooperation among these institutions and to increase their expertise.*'[4]

But the successive members meetings have shown how different their funds are, as well as their staff, their scope, and their actual achievements. It

---

[3] I refer to the title of one of the most stimulating books on historical methodology ever written by one of the most gifted historians of the twentieth century, Marc Bloch: *Apologie pour l'histoire ou métier d'historien.* Paris, Cahiers des Annales, 3. An electronic version of the (second) 1952 edition: http://classiques.uqac.ca/classiques/bloch_marc/apologie_histoire/apologie_histoire.html.

[4] Complete text of the charter: http://www.portahistorica.eu/charter.

makes me think of sport events, when athletes run the 3000 metres hurdles. Some arrive in the front row without overthrowing any hurdle, other risk not to arrive after overthrowing them all. Fortunately cooperation is an explicit and major asset in the charter. And as an example, we as members of the Royal Historical Commission, are grateful for the assistance the Dutch '*Instituut voor Nederlandse Geschiedenis*' is willing to render.

It is impossible to survey all the achievements of ICT in our study field. When I prepared this introduction I planned to provide you with a list of these achievements. I intended to split them up according to a series of different criteria, the first being the sites presenting primary sources (I mean documents), published directly in an electronic format. Meanwhile the *Instituut voor Nederlandse Geschiedenis*,[5] the *Institute for Historical Research*,[6] and the commercial enterprise *Brepolis*[7] lead the field. For secondary sources (modern historians' books and articles), I would have liked to stress the importance of the *Directory of Open Access Journals*,[8] offering free access to over 1400 revues in all scientific fields. The electronic libraries of for instance *Google*[9] and *Openlibrary*,[10] storing gigantic collections of some millions of books are definitely of a tremendous help. These are just images, yet searchable, from texts printed in former times. I got a lot of help from such sites as the *Medieval Manuscripts in Dutch Collections*,[11] and the achievements of the *Instituut voor Nederlandse Lexicologie*,[12] that composes dictionaries, for both the historical and current Dutch language. Electronic availability completely revolutionized how we, historians, work. Moreover, I could have made divisions too according to the format the documents are presented in: doc-files, PDF, JMEG, etc. I could have divided them according to their publisher. On the one hand, there are private editors such as Brepols with its excellent products, like the *Thesaurus diplomaticus*,[13] but very expensive and thus of a too limited use. On the other hand, there are the public initiatives: commissions working individually or cooperating, for instance in establishing the *Monasterium* database. An institution like the *Monumenta Germaniae Historica*, a joint initiative of Munich and Vienna, follows a double track. It works with Brepols (*eMGH*) and on its own (*dMGH*).[14] I could have stressed the difference in focused public: scholars finding out

---

[5] http://www.inghist.nl

[6] http://www.ihr.org/

[7] http:// www.brepolis.net

[8] http://www.doaj.org/

[9] http://books.google.com

[10] http://openlibrary.org/advanced

[11] http://www.mmdc.nl/static/site

[12] http://www.inl.nl/index.php?option=com_content&task=blogcategory&id=92&Itemid =523

[13] A CD produced by *Brepols* and the Belgian *Koninklijke Commissie voor Geschiedenis / Commission royale d'histoire.*

[14] http://bsbdmgh.bsb.lrz-muenchen.de/dmgh_new/

sources for an in-depth research, or students getting acquainted with the whole methodological context of documents, mainly enhancing the auxiliary sciences approach. I then think of the *École des Chartes* [15] or the initiatives in American universities, such as Fordham, [16] to provide the students with anthologies, the Latin texts being often translated into English.

The ICT achievements are overwhelming, and as a result even the biggest libraries tend to increasingly and rapidly become a repository for static and ageing knowledge, a type of frozen memory.

Earlier in this paper I addressed the bit rot problem. A partially linked item requires our consideration, viz. the problem of sustainability. These billions of websites existing today and even much more billions of obsolete versions that would exponentially become billions and more in the future, where do they end up? And I mean it as well for their extended use as documents for the institutions or persons who created them, but also for us historians. As a medievalist I am glad that not that much has been produced in writing in the millennium long Middle Ages and that time has made a fuzzy selection of what survived. How could an historian ever be able to deal with whatever theme if all documents would still be available. However, I consider it a social duty that institutions, mainly public ones, take care of a long lasting conservation and consultation. I know that the *Institute for Historical Research* is very much interested in this problem, and our friend Jane Winters in particular. Among others, the DANS-project of the *Dutch Royal Academy of Sciences*, with its special 'seal of approval' in sustainability are leaders in the field. [17] The *Council of European Social Science Data Archives* locate, access, and share data of 25.000 datasets. [18] Another interesting example as repository for datasets is the British *Arts and Humanities Data Service*, but its survival seems at risk due to financial cutbacks. [19]

When one evaluates all the achievements so far, one major observation, and a critical one, apparently imposes itself. Historians have scanned texts and made them accessible through search. Others have, one step further, produced texts immediately published on the internet with or without a parallel printed version. Auxiliary tools are numerous, powerful, and updated regularly if not in real time. But none of these performing realisations freed itself from the manual intervention of the historian himself.

We are thus left, last but not least, with a final question. Indeed, in how far has ICT influenced or determined the methodology of text editing? I have the impression that we see little evolution in that field, if none. Historians continue to follow the path Dom Jean Mabillon showed more than three hundred years ago, followed by Karl Lachmann, productive

---

[15] http://www.enc.sorbonne.fr/
[16] http://www.fordham.edu/halsall/sbook.html
[17] https://www.dans.knaw.nl:10091/
[18] http://www.cessda.org/about/
[19] http://ahds.ac.uk/

in the second quarter of the nineteenth century, and Jacques Bédier a century ago. For me Dom Jacques Froger's book 'La critique des textes et son automatisation' has been a tremendous intellectual leap forward.[20] Published in 1968 it showed how parallel texts could be compared using computer logic and more specifically Boolean mathematics. Notwithstanding the major methodological innovation it suggested, its reception and certainly its application remained limited.

In 1996 on occasion of our Narrative Sources colloquium in Ghent I expressed the hope that within ten years from then the joint efforts of historians and ICT specialists would lead to a fundamental methodological shift, a complete electronic edition.[21] Simply said: entering manuscripts on one side of a maxi-computer like device while moments later at the other side of the machine a critical and moreover reliable edited text would come out. The production would have gone through all the successive actions editors undertake traditionally, and mostly implicitly. Scanning or digitally photographing manuscripts, verifying the quality of these texts, comparing them, establishing variants, choosing between them, drawing a *stemma*, establishing the critical text and the notes, and so on.[22] I hoped all this would have been possible in my reference year 2006. We are not that far advanced by now, although in my opinion all the technology and the software needed are available. Especially in the field of text collation good results are obtained.[23] So I repeat my hope, my wish, and my expectation. When historians and philologists can in the very beginning achieve this goal, with a lot of mistakes, I know, but progressively in a more correct and performing way, then we use ICT where it has been invented for: to free humans from repetitive acts, to provide them with an unlimited artificial memory, and to give them more spare time for creative thinking.[24]

---

[20] Coll. «Initiation aux nouveautés de la science» n° 7, Paris, Dunod, 1968, xii + 280 p.

[21] L. MILIS, 'Vooruitblikken naar het verleden. Over eruditie en de informatiesnelweg', in: L. Milis, V. Lambert, A. Kelders (ed.), *Verhalende bronnen. Repertoriëring, editie en commercialisering*. Gent, 1996 (Studia Historica Gandensia, nr 283), pp. 3-15.

[22] A good example of attempts in this field is: C. MONROY, R. KOCHUMMAN, R. FURUTA, E. URBINA, E. MELGOZA, A. GOENKA, '*Visualization of variants in textual collations to analyze the evolution of literary works in the Cervantes project*', 6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), 2002 (http://www.csdl.tamu.edu/cervantes/pubs/ecdl2002.pdf), downloadable at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.15.2470.

[23] So far the quality of scanning or photographing remains a problem (especially the recognition of the characters in a manuscript text), although the manipulation of space photographs offered from the beginning a lot of help, as the late professor John F. Benton (†1988) (California Institute of Technology, Pasadena, CA) demonstrated already in the early 1970's. A vast number of publications on automatic collation appeared and in recent years philologists obtained good yet burdensome results. An example is: http://www.digital-humanities.org/dh2007/abstracts/xhtml.xq?id=215 (B. VAN ELSACKER, *The Complete Works of W.F. Hermans. Using Automatic Text Comparison and XML for a Voluminous Edition*.)

Important for further development is: http://www.allc.org/content/pubs/pisa_textual.html, the site of the *Association for Literary and Linguistic Computing.*

[24] All the URLs have been checked on April 18th 2009.