A Century of Science Publishing
E.H. Fredriksson (Ed.)
IOS Press, 2001

Chapter 17
Impact of Computers and Communications on Publishing

Nico Poppelier[a] and Einar H. Fredriksson[b]
[a]Penta Scope, Amersfoort, The Netherlands
[b]IOS Press, Amsterdam, The Netherlands

1. Introduction

The application of computers to publishing, and to science publishing in particular, can be traced back to the period immediately after the end of the Second World War. In a paper published in 1945, former vice-president of MIT Vannevar Bush described the use of new technology for the recording, organisation and consultation of information [1]. This paper proved to be a seminal one, since it can be regarded as the start of various streams of development. These developments eventually resulted in applications and technologies we find in our offices and in our homes today.

In the western world most people are nowadays familiar with Internet, in particular with the World Wide Web. However, the World Wide Web is very recent technology. It was developed in the early 1990's at CERN in Geneva by Tim Berners-Lee, Robert Cailliau and co-workers. Internet, and the Web in particular, continues to influence the field of science publishing. Another significant step forward for the dissemination of scientific knowledge, especially in the fields of mathematics, physics and computer science, was the development, by Stanford professor Donald Knuth, of the TeX system (see Ch. 18).

In the second half of the twentieth century the following three trends can be observed. Firstly, computers became smaller, cheaper, faster and more versatile. The result of this is that many people nowadays use computers for their daily work, and in some cases people cannot even do their work without computers. We also find computers in schools, in our homes, and even in places called "Internet cafes". Secondly, various means of telecommunication became more widespread and cheaper. As a result we now have regular telephone service, mobile telephone service, and global computer networking both with and without wires. Thirdly, various people developed innovative applications of computers, other than the pro-

cessing of numeric data (number crunching). Most people in the western world are familiar with word-processing software and calculation tools such as spread-sheets, but not so long ago manuscripts were still written by hand or typed on a typewriter.

When you consider how these trends apply to the area of (science) publishing, many different areas of development come to mind. In this chapter we will discuss the following three areas.

1. Without the Internet as a (nearly) world-wide facility for computer networking, there would not be a World Wide Web. The development of the Internet will be summarised in section 2.
2. As computers became more powerful and more easily available, people began to apply them to the production and processing of documents. An important concept in this field is the 'markup' concept, which will be explained in section 3. A well-known application of markup techniques to the production of scientific documents, perhaps even the best known throughout the world, is Donald Knuth's TeX system, the history of which is described in Ch. 18. Parallel to the fast deployment of TeX within the scientific community in the 1980's and 1990's there was another important activity, namely the development of the international standard SGML, and its application to document models for science publishing. Web technologies such as HTML and XML were derived from SGML. This will be also described in section 3.
3. Most text documents have a linear character, i.e. they are read from start to end. Using computer technology it is possible, however, to create and use non-linear documents, i.e. documents that can be read in various ways. Various people have played a role in this area, but the three most important ones are: Vannevar Bush, Douglas Engelbart and Ted Nelson. Their work, which will be described in section 4, underlies the World Wide Web as a global hypertext system, but also other forms of hypertext or hyperdocuments. The work of Engelbart also underlies the graphical user interfaces of contemporary computers. We do not describe here the work on 'search engines', based on information-retrieval research, which play an important role in the disclosure of the World Wide Web, which contains a tremendous number of pages and keeps growing daily.

2. Development of the Internet

The first public demonstration of what is now known as the 'Internet' was during the International Council for Computer Communications (ICCC) conference in Washington DC in 1972. It was developed under the auspices of the US

Defence Advanced Research Projects Agency (DARPA), which itself had its roots in the reactions in the United States to the Soviet launch of the Sputnik satellite in 1957. New directions in DARPA under J.C.R. Licklider from 1962 lead to the establishment of ARPANET in 1969 under R. Kahn.

In Europe experiments in the United Kingdom and France had been done on similar lines since the mid 1960's. In the United Kingdom, Donald Davies of the National Physical Laboratory (NPL) had in 1965 coined the term 'package switching' for the technology underlying the new computer communications. During the first decade the new technology had only involved a handful experts on both sides of the Atlantic, and it was only after 1972 that applications work did commence which could be said to have any effect on scientific publishing or communications.

It is interesting to note that the engineering association IEEE, with a leading communications society branch, in 1971 found the idea of a conference on computer communications premature: "it would not attract even 100 persons…" Hence ICCC was born and the conference of 1972 was organised independently of IEEE. This conference attracted more than 1100 participants, and the Internet ideas got its first large-scale exposure there. In connection with the next conference, ICCC'74 in Stockholm, the idea of an international journal was born: *Computer Networks*.

The days of conception of the Internet followed the introduction of the concept of time-sharing in the late 1950's, and Licklider reported that he first heard of it during the world computer congress organised by UNESCO in 1959 in Paris. Time-sharing was introduced as a solution to the need for sharing computer resources in a time when computers were few and very expensive. Connecting them via networks was in many situations a more cost-effective way of usage. In the concept of Donald Davies, traditional message switching as seen in telegraphy, albeit slow, was an alternative to using the telephone network. Communication between computers was characterised by short periods of intensive information exchange , for which the concept of 'circuit switching' known from traditional telephony proved less suited. Package switching combined message switching with time sharing. The next step was to define end-to-end protocols whereby each package contains information about sender, destination and how many packages make up the complete message.

The concept of 'open-architecture networking', introduced by Kahn in 1972, showed the way to connect entire networks based on different architectures. This was done with what was called an 'internetworking architecture.' ARPANET provided the testbed for the development of interconnecting networks, and this led to the widely used Transmission Control Protocol/Internet Protocol (TCP/IP) in 1978.

Since the mid 1970's the systematic introduction of Internet-related protocols into many operating systems in the United States provided a main ingredient in the spreading of Internet, partly because TCP/IP had been given away free of charge. The experimental Internet started in 1977 with 100 host computers, mainly within the scientific world. File transfer, electronic mail and discussion groups were among the main applications at the time. The now familiar Domain Name System was developed a few years later, in 1983, in response to the continuous growth of the Internet and the associated administration.

The ARPANET had its demise after 20 years, but around 1990 the management structure of the Internet was still very similar to the structure set up during the DARPA time. Already by the mid 1980's, DARPA was no longer the major funding agency supporting the Internet. Various governmental bodies, both inside and outside the United States, shared in the funding, and there was also an increased interest from the commercial sector. The management structure of the 1980's consisted of an Internet Activities Board (IAB), and a structure of task forces, with each task force focussing on a particular area of technology. Later, the Internet Engineering Task Force (IETF) was introduced to co-ordinate the work of the individual task forces. A major activity was always the standards process, which differed significantly from that seen in the international standards bodies, the International Organization for Standardization (ISO) being a well-known example. The increased variety of requirements from the user community as well as the increasing pressure to make the standards process open and fair contributed to the formation of the Internet Society (ISOC) in 1991. In 1992 the IAB became the Internet Architecture Board, and further organisational change were implemented within the framework of the ISOC.

While the initial usage of Internet was mainly scientific or military, the introduction of personal computers and local-area networks, starting around twenty years ago, have broadened the user communities to include large segments of society. All areas of science, medicine and education have been profoundly affected by networking technologies. The first uses of this in scientific publishing was, outside the physics laboratories, in the scientific abstracting field.

In the late 1970's the use of telephone lines for international calls was very expensive, and the leasing of dedicated lines even more so. The slow growth of messaging between scientists in the 1980's is partly due to the high costs of communications and the relatively low number of personal computers fitted with modems at the time. Experiments with computer conferencing started around the mid 1970's, but didn't have a broad impact on science communications until 1990.

Around 1980, 'intermediate' technologies were introduced. One of these was

videotex, which combined telephone technology with television receivers, using an auxiliary device to connect the both. Storage on initially central computers, outside the US managed by the national postal and telecommunications authorities, and payments through one or other form of subscription arrangements, proved generally difficult to sustain. For general public use, only the French system Minitel proved successful in its first generation. Minitel used a simple, cheap terminal for alphanumerical data. None of these technologies had an impact on the field of scientific communication, however.

The way the DARPA research community collaborated with the computer and communication industry in the mid 1980's included access to the vast experience and use of TCP/IP protocols. This lead to the Interop exhibition and conferences, starting in 1988, which in turn became a promotion vehicle for the acceptance of internetworking by leading industries. This also started to affect the STM publishing environment outside the earlier uses by organisations like *Excerpta Medica* and the National Library of Medicine (NLM). In the commercial sector only Maxwell's Data Group had attempted to make serious use of Internet before 1989. This was the year of birth of the World Wide Web [2]. It would take another five years before Internet, mainly through the rapid deployment of the Web, today one of its principal applications, could commence to make its fundamental impact on science communication and publishing in general.

The Internet is constantly changing and has been developing largely in parallel with traditional telecommunications and telephony on the one hand, and television on the other. The demarcation lines between telecommunications and computer industries are becoming more and more vague as the Internet develops. New services such as digital telephony and TV provided through Internet are currently being introduced. Traditional content companies, e.g. science publishers and media conglomerates, are becoming part of a broader media industry. However, in this industry information costs money, so new concepts of publishing must be developed. Recent development of new concepts and new technology, such as mobile devices and wireless Internet access, is driven by scientific and technological developments, but not with the scientific community as the prime user group in mind.

3. The markup concept

The application of markup techniques to the production of scientific and technical documents has a long history. An interesting account of the early period, until the early 1980's, can be found in Nievergelt [6]. No doubt, the single most well-known application in this area is TeX, the history of which is described in Ch.

18. Parallel to the fast deployment of TeX within the scientific community in the 1980's and 1990's there was another important activity, namely the development of the international standard SGML [7], and its application to document models for science publishing. Web technologies such as HTML and XML are derived from SGML (although in different ways).

Markup is the name given to the technique of adding readable instructions (codes) to a text document. These instructions can describe the meaning of a word or piece of text, or the formatting (presentation), or both. This concept was developed in the late 1960's and early 1970's. Early applications of the markup concept were PUB, developed at the Stanford Artificial Intelligence Laboratory starting in 1971, nroff, developed at the Bell Laboratories for the — then brand-new — Unix operating system in the first half of the 1970's, and Scribe, developed at Carnegie-Mellon University in the late 1970's. These systems influenced each other and also developments such as LaTeX, the widely used macro package for the TeX system, and of course SGML.

4. SGML: a condensed history

In the second half of the 1980's a new international standard called SGML began to draw attention. But the roots of SGML go back more than thirty years ago. In 1969, three employees of IBM, Charles Goldfarb, Edward Mosher and Raymond Lorie developed a new markup language, which they named GML after their initials. Later, they started to call it Generalized Markup Language. It was a means of allowing editing and formatting of text documents, and information retrieval on collections of text documents. GML was used in an IBM system called the Document Composition Facility.

In the late 1960's several people proposed to use the markup concept in a new way: instead of inserting codes that defined formatting, one should use codes that described the *structure* of documents. In 1967 William Tunnicliffe, chairman of the Composition Committee of the Graphics Communications Association (GCA) gave a talk on separating information content from presentation at the Canadian Government Printing Office. Around the same time, book designer Stanley Rice proposed the idea of a universal catalogue of codes for 'editorial structure'. Norman Scharpf, director of GCA, recognised the significance of these trends, and established a generic coding project in the Composition Committee, which developed the GenCode® concept.

Nearly ten years later, a committee of the American National Standards Institute (ANSI) began the development of a text description language based on GML and the GenCode® concept. The first working draft of this language appeared in

1983. Development later continued under the auspices of the International Standardisation Organisation (ISO). Finally, the standard describing this new language, which was baptised Standard Generalized Markup Language (SGML), was published in 1986 as ISO International Standard 8879. The SGML industry gained impetus because of the inclusion of SGML-based document models in the CALS series of standards, which was developed by the US Government. The deployment of CALS in various industry sectors also triggered the development of specialised SGML software, e.g. editors and document management systems.

5. SGML applied

SGML became a useful tool for science publishing with the development of a series of standards for electronic manuscripts by the Association of American Publishers (AAP). In fact, development of these standards had begun before 1983, before SGML became an international standard.

The series of standards created by AAP consisted of definitions of markup conventions, document type definition (DTD) in SGML parlance, for books, journals and articles, especially for science publishing. For that reason, the publications could not only contain text, but also tables and mathematical formulae.

Although this early development was far from perfect, the influence of it is still felt. Many of the document type definitions used by science publishers today, both in the US and in Europe, still show the AAP influence quite clearly. Among the creators and early adopters of the AAP standards were IEEE, the American Chemical Society, the American Institute of Physics, the American Mathematical Society and Elsevier.

The AAP series of standards was later extensively modified, and became known under the new label of ISO International Standard 12083.

The definitions of markup for mathematical formulae in both the AAP series and in ISO 12083 were very much oriented towards the notation of formulae, i.e. their appearance on a sheet of paper or a blackboard. This approach was criticised by several people. Eventually, the discussion of the definition of mathematical formulae in ISO 12083 contributed to the development of MathML (see below). One of the core standards of the World Wide Web, HTML, is an application of SGML. Although earlier versions of HTML were not rigorously defined as an SGML application, for later versions a DTD was developed.

6. XML: son of SGML

More than ten years of experience with SGML led to the development of XML in 1997, by a working group of the World Wide Web Consortium (W3C). This

working group grew out of the consortium's Web-SGML activity, which was intended to bring the power of SGML, in other words complex document structures, to the Web. XML is a simplification of SGML especially intended for wide deployment on the World Wide Web, and is described in a text of merely thirty pages.

Within a few years after publication of the first working draft on XML, in 1997, it has become the core of a growing family of related standards that supplement and extend XML. For various fields of applications, suitable document type definitions are defined or are under development. HTML itself has been re-defined as an XML application called XHTML.

The publication of the XML Recommendation [8] by the W3C is undoubtedly an important milestone in the history of SGML. XML has the same strengths as SGML, but because of its simplicity it is also much easier to implement in software. Therefore, XML will have a significant impact on almost every branch of economic activity, including publishing, more than SGML ever did.

7. XML applied

One of the first applications of XML created within the World Wide Web Consortium was MathML, the Mathematical Markup Language [9].

Even though the World Wide Web was invented as a means of supporting scientific communication, by 1994 the best way of producing scientific documents was to insert images (pictures) of mathematical expressions into HTML documents. The W3C recognised the lack of support for mathematics. W3C staff member Dave Raggett published a working draft for HTML Math in 1994. Discussions on mathematics on the Web were held at the Web conferences of April 1995 (Darmstadt) and December 1995 (Boston). In the summer of 1996 a group of people interested in this subject area got together, and this group later became the W3C Math working group.

MathML was developed with the following goals in mind.

- It should encode mathematical material suitable for teaching and scientific communication.
- It should encode both the notation (appearance) and the semantics (meaning).
- It should facilitate conversions to and from other formats, with output formats including e.g. graphical displays, speech synthesisers, computer-algebra systems, TeX and braille.
- It should support efficient browsing of complicated and lengthy expressions.
- It should be extensible.

- It should simple for software to generate and process, and be suited for template-based and other editing techniques.

For MathML, several well documented and tested versions exist, the most recent one being version 2.0 [9]. MathML will continue to evolve. Various implementations of it exist already, and many more will appear. Eventually, most Web browsers, including the very popular ones, will support MathML. This will greatly improve and enhance the use of the Web as a vehicle for scientific communication, which was its primary goal. MathML will play an essential part in this, and is therefore without any doubt of great importance to science publishing.

8. Non-linear text

In 1945 Vannevar Bush wrote an article [1] that displayed great vision and would influence the work of many people. Bush was director of the US Office of Scientific Research and Development under President Roosevelt. Before that, he had been Vice-President of MIT and Dean of MIT's School of Engineering. In 1950 Bush became the first director of the National Science Foundation (NSF), an institute which he himself had proposed.

In his 1945 article Bush discussed possible ways in which scientific and technical developments that resulted from the war effort could benefit mankind. The first benefit of science and technology to mankind is, according to Bush, man's increased control of his material environment, e.g. food, clothes, shelter, and health-care.

Bush recognised that, as part of their work, researchers make written records of their findings, and that these written records steadily increase in volume. Already in 1945, Bush found the rate of increase of such records alarming. Therefore he asked himself how new technologies could help researchers make more and better use of the results of prior research. He suggested that new recording techniques such as photography would help. One of the techniques he had in mind was the microfilm, and he used the hypothetical example of the Encyclopaedia Britannica reduced to the volume of a matchbox.

Bush also believed that the future would bring advanced computing machines that would handle advanced mathematics. Nowadays we indeed use computers for complicated numerical calculations, as well as for algebraic manipulation of mathematical expressions, thanks to software packages such as Mathematica and Maple.

For Bush, the problem of consulting the growing mountain of information was perhaps the most staggering one. As a solution to this problem he proposed a new device, 'a sort of mechanised private file and library'. This is probably the most important part of his 1945 paper, and it is certainly the one most often cited. Bush

gave this device the name 'Memex', short for Memory Extension. The memex was supposed to be a device for making and following links between documents. Writing about the memex, Bush in fact described hypertext, although that name would be used for the first time by Ted Nelson in the 1960's (see below). Although Bush was clearly thinking of microfiche in his article, he described the idea of the memex in more general terms, which makes his paper all the more impressive. Near the end of his article he writes "Presumably man's spirit should be elevated if he can better review his shady past and analyse more completely and objectively. He has built a civilisation so complex that he needs to mechanise his record more fully if he is to push his experiment to its logical conclusion and not merely become bogged down part way there by overtaxing his limited memory".

Bush believed scientific and technological advances would improve life on our planet. In the first section of his paper he mentions increased control over our environment. Nowadays many people believe man has exerted his control in the wrong way. In the last paragraph of his article Bush writes that the applications of science have made it possible for people to fight each other with cruel weapons, but that it may also allow them to grow, if they don't perish before learning how to use science for their own good.

Doug Engelbart was familiar with the 1945 article of Vannevar Bush: he had read it during his period of service in the US Navy. Engelbart began his career at Ames Research Laboratory, and later moved to the Stanford Research Institute (SRI). Partly due to Engelbart's influence, SRI became the second node on the ARPANET. After many years of thinking about ways of using computers that were regarded as unconventional at that time, he finally got funding to begin his own project. The first report he produced in this project was called 'Augmenting Human Intellect: A Conceptual Framework'. It is clear from his work that Engelbart was influenced by Bush, a fact he acknowledged in a letter to Bush shortly before publication of the abovementioned report. Roughly fifteen years after Vannevar Bush's seminal paper, Doug Engelbart and co-workers at SRI built a prototype system for collaborative work. Thanks to the improved technology Engelbart's team could build their system, which can be regarded as a form of 'memex'.

The system Engelbart's team developed was called NLS (oNLine System). It was a system for editing and browsing hypertext, as well as for sending and receiving electronic mail. It was also made available as a commercial system under the name Augment. Engelbart's work [4,5] had significant influence on the development of software tools for collaborative work, nowadays called 'groupware'. At the Fall Joint Computer Conference of 1968, Engelbart and his co-workers gave a 90-

minute demonstration of NLS. Using NLS, he and colleagues thirty miles away collaborated in a presentation that contained video teleconferencing, and hypertext links in text and graphics. In the demonstration they also used a pointing device that is regarded as the world's first mouse.

Engelbart had a vision of using computer to support the work of groups of people, by giving them means to communicate their ideas and share their knowledge. The phrase he used for this sort of work was 'CODIAK': Concurrent Development, Integration and Application of Knowledge.

Engelbart observed that most human knowledge, from electronic mail and notes to heavy reports and books, are inherently hyperdocument objects: in other words that our knowledge consists of fragments of information that have various types of links or connections.

He also recognised the importance of explicitly structured documents. This would make it possible to select parts of documents, to view documents in different ways, depending on the application, and to link to parts of documents instead of to a document as a whole. Documents according to Engelbart were not restricted to being purely textual. In fact, it was essential for him that documents had mixed content, e.g. text, diagrams, mathematical equations, still or moving images, and sound, "all bundled within a common 'envelope' to be stored, transmitted, read (played) and printed as a coherent entity called a 'document'" [5].

Ted Nelson is known as the inventor of the word 'hypertext' and for his many papers about the hypertext system Xanadu. Nelson was interested in new ways of representing and connecting information, ways that went beyond the possibilities of traditional publishing on paper [3]. One of the metaphors he used was that of cutting and pasting — not in the way it is implemented in modern word-processors, but the way it was done before computers arrived on the publishing scene. When Nelson was working at the *New York Times* (his job was to fill up the pots of glue) he noticed how journalists would cut up articles into pieces, arrange them on a large surface in front of them, and would then paste it together the way they wanted it. Nelson became interested in using computers for arranging fragments of information and linking them together. In his second year of graduation at Harvard, he took a computer course in order to get a better understanding of the capabilities of computers. That was the start of the project that he later called Xanadu, a project that is unfinished, even today. For a presentation at the 1965 conference of the Association of Computing Machinery (ACM) he coined the word 'hypertext', which is a common word in the vocabulary of computer experts nowadays.

Later developments in hypertext include:

- HES (Hypertext Editing System), developed by Ted Nelson and Andy van Dam, at Brown University in Providence (RI) in the late 1960's, and used on the Apollo missions;
- FRESS (File Retrieval and Editing System), developed by Andy van Dam and students;
- Gopher, developed at the University of Minnesota;
- Hyper-G, developed at the University of Graze in Austria by Hermann Maurer and students.

In 1989, Tim Berners-Lee, while working at CERN in Geneva, wrote a proposal for the development of a new system that would allow easy dissemination of information, e.g. about experimental results or accelerator equipment, within CERN and the community of high-energy physics research.

Berners-Lee was familiar with the work of Bush, Engelbart and Nelson. In 1990 Berners-Lee, together with Robert Cailliau and other CERN staff, developed a prototype of his system on a NeXT computer, and called it World Wide Web. The new system found its way outside CERN, slowly at the start, but gradually picking up speed, especially with the launch of the graphical Web browser Mosaic in early 1993.

Systems such as NLS, HES, FRESS, Gopher and Hyper-G were in a way forerunners of the World Wide Web, but some offered features that the present Web does not have — not yet anyway. The World Wide Web can be regarded as the inevitable outcome of a simple addition: Internet technology meets the markup and hypertext concepts. In reality it was not as simple as this of course, and it took the vision and perseverance of people like Berners-Lee and Cailliau to make it happen [2].

9. Conclusion

In the early part of the Internet history, scientists formed the majority of the user community. Due to developments in computer science, e.g. hypertext and groupware, the widespread distribution of personal computers and affordable networking, Internet applications had a tremendous impact on all sectors of society, the World Wide Web being the most recent example. The developments seen in the past decade have changed the focus of Internet and of computers in general from science to commerce and administration. There seem to be no governmental bodies in position to control further developments of the Internet and computer or networking technology in general. For example, the World Wide Web is controlled by a consortium of 500 organisations and companies. By the year 2000,

hundreds of millions of users have become stakeholders — beyond the control of national governments or international organisations. All participants in the science publishing circle have become stakeholders as well, from authors, editors, publishers, agents, booksellers and librarians, to the scientist-user.

References

[1] Vannevar Bush (1945, July) As we may think. The Atlantic Monthly.

[2] James Gillies & Robert Cailliau (2000) *How the Web was born.* Oxford University Press, Oxford.

[3] Theodore H. Nelson (1980) Replacing the printed word: a complete literary system. In: *Information Processing 80.* S.H. Lavington, (Ed.) North-Holland Publishing Company, Amsterdam.

[4] Douglas C. Engelbart (1990) Knowledge-Domain Interoperability and an Open Hyperdocument System. In: *Proceedings of the Conference on Computer-Supported Cooperative Work, Los Angeles CA, 7–10 October 1990.* 143–156.

[5] Douglas C. Engelbart (1992) Toward High-Performance Organizations: A Strategic Role for Groupware. In: *Proceedings of the GroupWare'92 Conference, San Jose CA, 3–5 August 1992.* 77–100.

[6] Jurg Nievergelt, Giovanni Coray, Jean-Daniel Nicoud & Alan C. Shaw (Eds.) (1982) *Document Preparation Systems.* North-Holland Publishing Company, Amsterdam.

[7] Charles Goldfarb, et al. (Eds.) (1986) *Standard Generalized Markup Language (SGML), ISO International Standard* 8879:1986. International Organization for Standardization ISO, Geneva.

[8] Tim Bray, Jean Paoli & C.M. Sperberg-McQueen (Eds.) (1998, February) *Extensible Markup Language (XML) 1.0.* World Wide Web Consortium.

[9] David Carlisle, Patrick Ion, Robert Miner & Nico Poppelier (Eds.) (2001, February) *Mathematical Markup Language (MathML) 2.0.* World Wide Web Consortium.