

# Document: Form, Sign and Medium, As Reformulated for Electronic Documents

Roger T. Pédaque

► **To cite this version:**

Roger T. Pédaque. Document: Form, Sign and Medium, As Reformulated for Electronic Documents. 2003. sic\_00000594

**HAL Id: sic\_00000594**

**[https://archivesic.ccsd.cnrs.fr/sic\\_00000594](https://archivesic.ccsd.cnrs.fr/sic_00000594)**

Preprint submitted on 12 Sep 2003

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Document: Form, Sign and Medium, As Reformulated for Electronic Documents<sup>1</sup>

Roger T. Pédaque, STIC-CNRS  
Contact and comment: [pedauque@enssib.fr](mailto:pedauque@enssib.fr)

Version 3, July 8, 2003

Original french version available :

[http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/05/11/index\\_fr.html](http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/05/11/index_fr.html)

## Abstract

This paper presents group discussions taking place within multidisciplinary topical network 33 of the CNRS Information and Communication Science and Technology (STIC) Department. It attempts to clarify the concept of document in its transition to electronic form, based on research which tends to privilege form (as a material or immaterial object), sign (as meaningful object) or medium (as communication vector).

Each of these terms reflects the radical transformations that are taking place. Their superposition stresses the importance of multidisciplinary for a lucid and complete analysis of the concept and how it is changing.

## Context

Very few scientific papers give a definition of document and even fewer discuss the definition. The document appears as direct subject of analysis in only a few rare scientific communities: *Information Science* researchers, based on work concerning documentary techniques which have considerably changed due to electronic data processing, researchers investigating the digitizing of documents and indexing-cataloguing problems, who have often extended their reflections to electronic document management, and those developing *electronic publishing* tools. Furthermore, documents are discussed when they are essential tools for constructing and progressing in certain fields such as history and particularly archeology, as well as geography, especially maps, and for the texts of laws, regulations and circulars, but as instruments, and only rarely directly.

Many dictionaries, lists of standards and encyclopedias have definitions that are more designations or descriptions than a real reflection on the concept.

From the Latin *documentum* giving the word roots in teaching (*docere* = to teach), to its marginalization by the more recent, more frequent but hardly more accurate term of “information”, the concept appears to be commonly based on two functions: evidence (the “evidence” presented in courts or the elements of a case file) and information (a representation of the world or a testimony). For instance, contemporary archive science recognizes these two functions in stating that documents have “value as evidence” (of activity), which has a somewhat broader meaning than judicial “evidence”, and “value as information”, in the sense given above.

---

<sup>1</sup> This paper is the third version of a working paper. It was a collective work. Around fifty French researchers contributed to the text. A mailing-list has been dedicated to the debate which is still going on. You can propose your own argument by sending an e-mail to [pedauque@enssib.fr](mailto:pedauque@enssib.fr). Only scientific contribution will be taken into account.

A very large number of research papers use different vocabularies, rigorously defined in some cases but often subject to different interpretations, to designate comparable objects. For instance, computer science researchers, from network analysis to data bases, data mining, information search, automatic language processing, or knowledge engineering; corpus linguists, semiologists, psychologists of learning, sociologists of culture or organization, economists of the media or information, jurists of intellectual property rights and generally the humanities use a variety of terms such as information, data, resource, file, written material, text, image, paper, article, work, book, journal, sheet, page, etc. which of course are not synonymous, each of which has a justification in the particular context of the research concerned, but all of which have a relation (generally not assumed) with the concept of document.

Finally, documents are ubiquitous in our daily life (especially administratively and even in scientific activity). The concept is therefore intuitive for all of us, and we do not feel the need to clarify it.

This lack of clarity is today a problem: the electronic form is revolutionizing the concept of document, but there is no way of accurately measuring the impact and consequences due to a lack of clear contours. The transformations occurring in the shift from the widespread medium of paper to the electronic medium, are obvious when it comes to material aspect, cognitive treatment, perception and use. This upheaval, although announced by a few pioneers and prepared by the increasingly obvious convergence between writing and audiovisual, is very recent, still chaotic and undoubtedly irreversible. It is probable that the many researchers investigating these issues from many different angles would have much to gain from an overall view allowing them to see things more lucidly.

The contrast between the relative stability that existed heretofore and the speed and depth of the changes now occurring undoubtedly explains the delay in analysis. There was no need to investigate, except as a historian, an object so commonplace as to be self-evident, and today, not enough time has gone by to be able to see the situation from a distance.

The document was constructed as an object, whose most common material form is a sheet of paper, over a process that lasted for centuries mingling tools, knowledge and status. Over the last few decades with electronic documents, we have entered a new phase, certain of whose features are in direct filiation with the previous period, whereas others on the contrary mark a radical change and perhaps the emergence of a different concept embodying all or part of the social utility that we were accustomed to calling "document". The most obvious manifestation of this change is therefore the loss of stability of the document as a material object and its transformation into a process constructed on request, which can undermine the trust placed in it.

The querying between rupture and continuity does not arise only for the object. The analysis methods and epistemologies are also rapidly changing.

## **A Multidisciplinary Approach**

We feel that these difficulties can only be resolved through a determinedly multidisciplinary approach. Our thinking was confirmed by the CNRS STIC Department, which initiated a multidisciplinary topical network called "Document and content: creating, indexing, browsing" (<http://rtp-doc.enssib.fr>) including around a hundred researchers. The concept of document is not central to some of the disciplines covered by the network and the researchers only have a partial understanding of what this concept covers. The purpose of the network is therefore to shift the focus in order to make the document an essential subject of research, at least for a time, by pooling the partial contributions of the different researchers.

There is not necessarily a consensus between disciplines or even within each discipline on the issues under discussion. Our aim is to clarify and detail the concepts to dispel misunderstandings, open new prospects and identify disagreements which may exist, not to harmonize or define a line, a current or a school of thought. It is our conviction that dialog between disciplines cannot be fruitful unless we have succeeded in identifying the essential concepts so that we can discuss them or use them as basis.

This attempt is not without peril. First of all, it is possible to distort the meaning or remain superficial. In addition, the different bases of the disciplines or currents may be contradictory. In addition to conceptual difficulties, the subject may run into more mundane obstacles. Each specialty naturally develops its own culture and vocabulary, for both good (rigor) and bad (protection) reasons. The same words sometimes have different meanings in different communities and are often even alien to outsiders. In a horizontal text, we need to use a common vocabulary, in all senses of the term, at the risk of caricaturing.

Concretely, this paper is the result of group work within the network. Because of the method used for writing it and the many contributions it embodies, we decided not to give any quotes or direct bibliographical references. Doing otherwise would bias the group dynamics by introducing competition between authors or schools of thought. However, a bibliography can be viewed on the site of rtp-doc.

## Propositions

We will use an analogy with the linguistic distinction between syntax, semantics and pragmatics to organize our propositions. Without going into a discussion on the validity of this analogy or even the legitimacy of this tripartite division used in language science, we can see that it allows a fairly simple classification of current research and its underlying currents. We make a distinction between:

- The document as form; under this heading, we will classify approaches that analyze the document as a material or immaterial object and which study its structure to better analyze, use or manipulate it
- The document as sign; for these researchers, the document is primarily perceived as meaningful and intentional. The document is thus indissociable from the subject in its context which constructs or reconstructs it and gives it meaning; at the same time, it is considered in a documentary system or knowledge system
- The document as medium; this dimension finally raises the question of the document's status in social relations. The document is a trace, constructed or found, of a communication that exists outside space and time; at the same time, it is an element of identity systems and a vector of power.

The analogy with linguistics remains questionable. It could be argued that the first category is more specifically related to morphosyntax and that the second includes both semantics and pragmatics. But this does not prove anything. All we need is for the analogy to be efficient in our work.

Each category should be viewed as a dominant but not exclusive dimension. For instance, researchers who use the document as form approach do not necessarily neglect the other two approaches, but their analysis and reasoning privilege the first approach, and the other two remain complementary or external constraints. The term "entry" would perhaps be more appropriate. Each of these entries is indeed a way of approaching the document, subject of the research, from which the other dimensions will be found through developments, constraints, obstacles or limits

which appear in the primary reasoning. However, each approach probably also tends to relativize the others excessively.

We will discuss each category using the same scheme:

- First, we will identify the main disciplines, know-hows or specialties that privilege this point of view. The aim is not to discuss their validity or scientificity but to review the diversity of the research representing this orientation, without making a value judgment as to its importance
- Then we will suggest an interpretation of the evolution of the points of view in the transition from traditional document to electronic document
- We will gradually construct a definition of the document based on each entry
- We will identify a few outstanding questions in each category, beyond the scope of the current research.

In each instance, we will attempt to identify the essential, without dwelling excessively on nuances, exceptions and special cases. The aim is to understand what is fundamental, not to be comprehensive. As concerns the definition, one method would be to make a systematic search for cases that do not satisfy it and then construct a universal definition. This method does not seem very efficient to us. Our aim is not to answer everything, but to construct a generic definition, even at the cost of identifying exceptions that represent either very special cases or intermediate or transitory situations, when not simply related to an incomplete or incorrect analysis.

In the conclusion, we propose a synopsis of the three entries highlighting the elements of continuity and rupture with respect to the preceding period.

## **Document as Form**

It may be argued that the term “form” is ambiguous, but we use it for lack of a better one. It should be understood here both as “contour” or “figure”; in other words, the document is seen as an object or an inscription on an object, whose boundaries are identified; and as a reference to “formalism”, because this object or inscription obeys rules that constitute it.

Here, the document is viewed as an object of communication governed by more or less explicit formatting rules that materialize a reading contract between a producer and a reader. The document is mainly analyzed from the angle of this implicit communication protocol, irregardless of its specific textual or nontextual contents.

## **Specialties concerned**

From the outset, we must emphasize the particular place occupied by writing, a widely learned technique that places the document from the outset in a fundamental social situation.

The know-hows, professional and other, that privilege this approach, are numerous, and in some cases very ancient, such as calligraphy, typography, and, for other forms of representation, the techniques of music, video and cinema, as well as library science, focused on document cataloguing, classification and management, and also archive science.

It is logical for information specialists who digitize material objects, i.e. image specialists, to have strong ties to these first specialties. They are rapidly interested in the internal structure of documents, with automatic pattern recognition systems, mainly automatic character recognition, then handwriting and page and image layout recognition. In their area, they are confronted with problems of formats, exchange, storage, description, addressing, preservation and processing of

large quantities. This concerns automatic document reading or analysis. The researchers attempt to decode the object by explaining/making use of the underlying communication protocol (the reading contract).

Similarly, all those interested in typographic characters, page layouts, editorial formats, international standards in these areas, text processing, those who construct digital video systems exploit and renew ancient know-hows.

Other information specialties have also chosen this approach. The design of electronic document management systems, as the name implies, is based on the idea that the document preexists as identifiable object, even if it is virtual. Although the starting point is an electronic file, not a concrete object, many of the problems posed in this case are related to the same fundamental interrogations.

Finally, a sudden change of scale occurred with the invention and explosive success of the World Wide Web. An intensive research, design, negotiation, standardization and development activity is deployed around the Web, within and without the W3C consortium. Although these researchers use the word “document” very little, preferring the term “resource” which covers many other objects, many of the questions raised by Web designers in its current version (i.e. before the “Semantic Web”) are also largely related to this first approach. The investigation focuses on how to interconnect the resources on a planetary scale and therefore requires defining standards and systems that can be used on all machines and assigning a unique address to each of these resources. Many of the resources have the features of a document as understood in this first dimension: HTML or XML file, images, audio or video recordings, streaming multimedia, etc.

## Evolution

A first definition of document could be given by the equation: *Traditional document = medium + inscription*. Initially, the emphasis is placed on a medium that can be manipulated (literally), bearing signs that can be interpreted, depending on their form, by sight, hearing or touch, in the case of Braille, and why not other senses tomorrow, with or without prostheses. These signs represent the contents, materialized by an inscription.

The predominant (but not exclusive) traditional medium is paper, and the signs are writing, handwritten or printed. The basic element, i.e. the written page, can be enriched by formatting and text enhancements and extended by binding, footnotes, etc. conferring on the document a very great plasticity and complexity. The “codex” (book with bound pages) is undoubtedly the most sophisticated form of traditional document. Its quality can be measured by the robustness of its “specificities”, practically unchanged for more than a millennium!

At the cost of a major social effort (school, allowing acquisition of the reading protocol), this type of document is directly perceptible, i.e. without any intermediate high tech instruments (except eyeglasses for some), by a variable share of the population of a given society: those who know how to read.

When, in the course of history, this notion (*medium + inscription*) was extended to other forms of representation, such as recorded music, cinema then audiovisual broadcasting, the medium lost its faculty of direct appropriation. The representation was more accessible to direct human perception (it could be decrypted without a complicated learning process), but the reading process became more sophisticated. It is necessary to have a machine to listen to music (recorded on a disk, tape or CD), project a film (recorded on celluloid) or view a video (recorded on tape or DVD). The object (medium) is still necessary for reading, but is no longer sufficient.

Furthermore, first radio, then television broadcasting separated decryption of the signal from transmission. Broadcast audiences hear or watch “programs” whose transmission is outside their control. They have no control over the time of reading, unless they tape the program. In a way, when broadcasting entered the home, it dispossessed them of some of the space-time autonomy they had when manipulating mastered or recorded objects.

Audiovisual broadcasting thus opened the way to evolution in the use of media, but for us the essential mutation is the change of the inscription from an analog signal to a digital signal, with all the data processing facilities this involves. This has radical consequences for all written, pictorial and audiovisual documents. These mutations are reflected in the read-write systems and in the documents themselves.

As concerns systems, first of all, an extraordinary give-and-take between writing and audiovisual is observed. The first integrates a system familiar to the second. It is no longer possible to read without a machine. Although the production of printed documents requires a great deal of machinery, reading of the printed document is, as we mentioned, straightforward or almost so. Optical or magnetic disks, recorded tapes, signal processing and restoring machines as well as network connections must be purchased individually to read electronic documents, even if a printer is then used to return to the previous state. Listeners and viewers can, for their part, go onto the Web and control the start and stop of streaming media, a capability which was lost in broadcasting, where the only way to modulate the uninterrupted flow of programs was to tape them.

The second consequence on systems is the commingling of medium and signal. The concept of medium is becoming more complex and ambiguous. Is the medium the file, the hardware on which it is stored or the surface of the screen on which it is displayed? As it goes through the network, a document is broken into pieces that are copied in routers for a very short time, and may also be stored complete in caches for a variable time. Moreover, the same “media” can contain any type of representation, provided it is digital, and the representations themselves can be combined provided their formats are compatible: tightly interleaved text, still images, audio and animated images can be “read”.

Whereas printing privileged the physical medium because of the technological complexity inherent in the document production activity, electronic publishing has made possible the on-demand production of documents (on the screen or on paper). The medium has thereby lost its privileged status to the benefit of electronic publishing. On this subject, it can be recalled that *wysiwyg* (*What you see is what you get*) was one of the major advances of electronic publishing.

Last but not least, with the parallel growth of computers and telecommunications, the machines themselves, such as laptops, PDAs, telephones and various types of integrated tools, are multiplying and becoming autonomous, in the search for the best way of conciliating the behavior of the readers with their generic and/or specific needs. It is not indifferent for the future of documents that cell phones have spread faster and much more pervasively than computers.

The concept of medium has thus lost its initial clarity. But, in our equation (*medium + inscription*), the transition to electronic form has just as radical consequences on the second term, inscription. Inscription can be considered a type of coding, an operation familiar to computer specialists. They therefore attempted to isolate the logical elements forming this dimension of the document to model them, automate operations and rearrange the perfected elements.

In this area, a comparison can be made with the concept of program as it is often presented in computer science: *program = software + data*. A document could be considered as a special case of computer program whose *software* part is the “structure” and whose *data* part is the “contents”. The equation then becomes: *electronic document = structure + data*. Consistently

with this first entry concerning form, researchers neglect content and closely investigate structure which, by definition, can be modeled and which, in a way, independently of the medium, represents the “reading contract” concluded between the document producer and its potential readers.

The structure varies enormously according to the type of document. Some documents are practically unstructured, such as certain spontaneous works of art or texts where form and content are indissociable. Others, on the contrary, follow rigid formal rules. The structure also differs according to the type of medium. For instance, audiovisual broadcasting introduces a time dimension which is practically absent from written documents. However, the analysis also allowed several levels of structuring to be identified and isolated in the most general case. These levels were constructed from two research currents, one from analog to digital and the other from digital to analog. Before coming back to the concept of structure, it is preferable to understand the logic of their reasoning.

The first current assigned itself the task of converting traditional documents to electronic form to be able to benefit from the performance of computers. In other words, a traditional document makes the transition from one equation to the other: *medium + inscription* to *structure + data*. This operation can consist of digitizing the original document, with the aim of dematerializing it using an image processing and pattern recognition approach. The reasoning can also be based simply on the representation of a document, directly reconstructing the visual equivalent of all or part of its representation without using the old medium. Note that the operation is not socially trivial. It must be possible in both directions, especially for legal reasons. We will come back to this under the third entry. Image processors are in this first current as their research, as their name implies, attempts to reconstruct the image, i.e. the formal representation, of a document.

This is the principle of pattern recognition. To be recognized, a form must first be known. The more the original document is based on generic structures, the easier it is to transpose. The complexity therefore increases when going from typographic characters to graphics, diagrams, then images and finally three-dimensional objects. Although the aim is to reproduce a perception that is similar or analogous to that of the original object, the process is a new translation which may mask significant elements or on the contrary lead to the discovery or rediscovery of new ones, depending on the technological choices made and the future use of the files. (..)

Other researchers start with the final equation (*electronic document = structure + data*); in other words, they go in the opposite direction. They develop algorithms, the essential element at the heart of computer science, to reconstruct the documents, backtracking through their logic or internal structure step by step to obtain a representation that can be viewed on the screen. This second current derives from the routine use of text in programming languages with gradual integration of the concern for form (*wysiwyg*). It led first to office automation tools, encountered researchers developing tools for electronic publishing and was finally confronted with the necessity of being able to exchange documents on a large scale. It then totally exploded with the Web revolution.

These computer specialists reasoned by layers to isolate and separately process the elements of the document structure. They thus discovered or rediscovered the different logical levels of this structure, the lowest of which is that of the text or analog signal which it was attempted to unify as Unicode, MPEG, etc.

The concept of text markup is attached to that of document structure, ever since the transition was made from electromechanical phototypesetters to digital phototypesetters. It gradually established two principles: the markers describe the structure rather than the physical



characteristics of the document and they are understandable both by a program and by a human interpreter.

Without detailing the history of this process, it can be said from the standpoint concerning us that the Web can be described as an infinity of linked documents. Its architecture is based on three pillars: resources identified by a universal addressing scheme (identification), which are represented as a nonexclusive set of schemes (representation) and exchanged using standard protocols (interaction). This architecture assumes that documents can be accessed from anywhere, using any type of hardware and according to the specificities of the user groups.

The two currents, traditional document recognition and direct construction of electronic documents, are not independent. Starting from different points, they converge to reach the same target. In particular, they allowed two basic levels of document structure to be emphasized: the logical structure (the construction of a document as interrelated parts and subparts) and the formal representation of the layout, the “styles” in data processing (for instance, the typographical choices for text). As far as we are concerned, the fundamental revolution is perhaps the gradual uniformization of document format (in the sense of data processing), because it is what allows simple processing of these two levels.

A document should be readable on any type of computer and decodable by a variety of applications. The trend is toward fragmentation: “proprietary” formats are invading the market and condemning “universal” and “free” formats to remain a luxury of specialists. In addition, “nonuniversal” formats lead to situations of illegibility: a program cannot read a file, an application cannot open a document, a Web page cannot be displayed correctly on the screen. Furthermore, the format must be able to transcribe the alphabet: the “format” should be suitable for transcribing several languages. Standardization is therefore essential.

It is probable that the increasingly widespread success of the XML standard and its many particular derivatives marks a new step or even a conclusion of these movements. The XML standard, resulting from computerization of publishing techniques (SGML) and the sophistication of the first Web markers (HTML), integrates structure and content in the same file using a standard text markup language. This equals and even largely exceeds the plasticity and complexity of bound pages we mentioned at the beginning of this section and a few features of which had been lost along the way. But by renewing the terms of the old reading contract whereby the link between perceived representation and logical structure was fixed by the medium, it also introduces new questions. An XML type approach captures the structure and content. The form can then be derived in different ways. It is not represented intrinsically. It can be said that the form is no longer the essential dimension of the document. But much work is being conducted on the different ways of representing and producing the form of an electronic document, especially for XML documents.

Our equation could thus once again be transformed: *electronic document = structure + data* would become *XML document = structured data + formatting* wherein the second part (the “style”) is largely modulatable. In the XML world, the form is defined separately from the data structure, using a stylesheet (XSL).

A possible, but not certain, evolution would be for documents “written” this way to be added to centralized or distributed databases, with all the files increasingly resembling enormous “Lego” sets where building blocks of different sizes, shapes and uses would be arranged in a great variety of configurations. This would be crossing a last step. A document would only have a specific form at two moments: when produced to allow the author to view or hear it to make sure it corresponds to what the author wanted (which is not even necessary if the document is produced automatically) and when reconstructed by a reader. It is very unlikely that the document will be

the same in both cases. Another way of conceiving this evolution would be to consider that the document is now the database itself, whose outputs are only a partial interpretation of its richness. A community of researchers is studying this issue in the context of the semantic web, in terms of “personalizable virtual documents”.

This evolution raises the problem of management of the different points in time or instances of one or more documents and their writing, enriching and rewriting by various participants. It is already complicated to manage the consecutive versions of a document, for individuals, organizations and on the scale of the Web. Procedures need to be invented to relate a text to an author (or group of authors) while allowing each author to appropriate, or reappropriate, all or part of the documents produced by other authors or themselves in order to limit “noisy” proliferation of the different versions of the same information on the network and identify the nature and origins of these modifications with the aim of coherently managing all the currently available electronic documents, independently of their format and status and outside any centralized institution.

We very clearly perceive the premises of this final step and equally well the problems it raises. However, it is much more hazardous to predict the evolution and therefore the consequences except to say that they will definitely be important and durable.

## Definition 1

The observation of this first dimension leads us to formulate a definition of electronic document that is incomplete at this stage but is representative of a major movement taking place. This definition must take into account marginalization of the medium and the basic role now played on the contrary by the articulation between logical structure and styles to redefine the reading contact, understood here as the legibility contract.

*An electronic document is a data set organized in a stable structure associated with formatting rules to allow it to be read both by its designer and its readers.*

This definition is probably too long to memorize easily. We therefore go back to the transformation of the equation:

*Traditional document = medium + inscription to Electronic document = structures + data.*

And we suggest the current evolution whose outcome is still uncertain:

*Electronic document = structures + data transformed to XML document = structured data + formatting, remembering that, *stricto sensu*, the XML standard does not define the formatting, which is defined by XSL.*

## Questions

This first approach, by the form, leaves several questions unanswered. Here, we focus on those concerning the relation between the perceptible world and the digital organization of the new documentary environment.

A first series of questions is related to the display of documents. Whereas “material bibliography” very closely studied all the aspects of the “book” as object, the transition to electronic form appears mainly to have focused on the structure based on the logical entry and for processing purposes. For instance, in this first dimension, researchers like to say that since the structure is integrated in the file, it can be viewed in any way, so that problems of perception are related to another problematic. This conception, taken to the limit, would assume that structure and contents are independent, which is debatable to say the least. The form has meaning and researchers have long been studying, for instance, the cognitive importance of the possible

navigation through hypertext links. However, much work still remains to be done on electronic reading to have a better understanding of the interdependency mechanisms between the two terms of the equation.

It should be noted that this separation destroys the bases of archival diplomatics one of whose purposes is to authenticate the contents of a document by analyzing its form. The result is that authentication (validation) must (will have to be) ensured by other methods, such as technical (electronic watermark) or organizational (trusted third parties). Would it be conceivable for the requirements concerning the form - in particular the authentic form - to be imposed and/or validated by the existence of a "signed" or "watermarked" stylesheet?

These issues are complicated by the fact that a given document can be read on different reading devices. Should we reason as if terminals had no effect on perception? We just need to compare the screens of a computer, tablet, personal digital assistant and cell phone to be convinced of the contrary. This brings us back to the reading medium which we thought had been done away with. These matters are the subject of extensive discussions, in particular within the W3C consortium: (*device independence*).

The progress made in screen display, especially based on the work of Xerox's PARC laboratory and the increasing sophistication of office automation tools, is limited to visual organization, restricted for documents to layout and organization in folders. A few graphics specialists propose interesting compositions. But these efforts appear to be relatively unconnected to the above research. Similarly, e-books and the hopes placed in electronic ink have not yet led to any very promising applications in everyday life, even if the possibility of reconstructing an electronic codex is exalting. Certain of those for whom spatial representation is essential, such as geographers and architects, have already investigated this issue in depth, but they are the exception, not the rule. Once again, it is perhaps audiovisual that is opening promising prospects with augmented reality integrating analog aspects and digital reconstructions.

A second series of questions is related to the long term life of electronic documents. These questions are often discussed. On the one hand, they do not differ from very old problems of archiving and preservation, simply transposed to other techniques. On the other hand, radically new problems have arisen: XML files are theoretically inalterable, provided they are regularly refreshed and preserved under good conditions, since all the information they contain is in digital form. Therefore, some consider that problems of document long term preservation will soon be solved. Conversely, these files are far from representing the form(s) in which the documents are read. A complete memory of these documents would require preserving all the consecutive reading equipment and systems used to access them. Here again, much theoretical and practical work remains to be done.

Finally, without claiming to be complete, there is a third series of questions. A traditional document is a physical object that can be handled. This object no longer exists for electronic documents, to the extent where the document ultimately becomes a sort of puzzle whose pieces are assembled at the request of the reader. However, a reader always accesses a document from a machine, i.e. the terminal on which it is viewed. Will we witness an extreme version of this idea, whereby a document is nothing more than a modern form of magic slate on which significant multimedia items are displayed on request, restricted only by a logic of meaning and specified needs? Or will there be a restructuring of typical documents meeting special needs or situations, whose dynamic will be restricted to strictly defined ranges? And might we not assume that the visual stability of paper, the possibility of grasping it and the fact that all the pages are present together have an important role to play in cognition? If so, shouldn't we encourage efforts toward an "electronic codex"? What impact will the new reading processes have on our learning

processes? What about the (individual or collective) author's legal or merely moral responsibility? Or, more directly related to our entry on form: can the production of a document be separated from its perceptible form and therefore is it simply conceivable to envision a formal rupture between the document produced by the author (who is also the first reader) and the document proposed to the readers? The success of facsimile (pdf) formats is often analyzed as momentary resistance to change. Is it not rather an indispensable perceptive stability? These questions could be summarized in a single question: By doing away with the medium did we not in fact overneglect the form?

## **Document as Sign**

As for the previous entry, the terms of the title of this part should not be interpreted too academically. The sign has long been the subject of much scientific research. Even if we make use of some of this research here, our purpose is not to discuss the concept. Our aim is simply to group and present research that considers the document primarily as a meaningful object. The entry that interests us here is processing of the content. Although the form is sometimes considered, it is only as a meaningful element.

## **Specialties concerned**

This category concerns disciplines substantially different from the previous entry, some claiming to represent historical progress with respect to it, as if going from form to sign meant getting closer to the heart of the problem.

Thus, as regards professional know-how, we go from library science to documentation then to information professionals who, rather than managing objects, provide answers to readers' questions. Or again, electronic document management (EDM) becomes knowledge management (KM) which, over and above a file store management system, directly identifies knowledge useful to an organization. And especially the Web gains an adjective qualifying it as "semantic Web", meaning that better use of the capabilities of interconnected machines could allow online file content processing in view of organizing services that more closely meet the cognitive demands of Internet users.

From an academic standpoint, this category includes first those who work on written material, speech and images, i.e. linguists or semioticians of all schools, those analyzing discourse, corpus linguistics, semantics as well as those constructing automatic language processing tools for translation or automatic data searches. They are coming together with a second category of computerspecialists mostly from the field of artificial intelligence, who, starting from the attempt to model the reasoning process, are attempting to build tools that are also capable of answering questions by searching files. At the same time, the concept of information is being replaced by the concept of knowledge, which has the advantage over the first of integrating reasoning. A new discipline called "knowledge engineering" is now emerging.

Very rapidly, it appeared that research on information about information (metadata) was useful, and even essential in some cases. From cataloguing to indexing, then from thesauruses to ontologies, metadata have become an essential tool and subject of research.

In this case, as for the previous entry, the explosion of the Web modified the situation by changing the scale of available resources. The attempt to construct a semantic Web, initiated by traditional Web architects, was greeted enthusiastically by the researchers of this dimension.

## Evolution

According to this dimension, the definition of a traditional document could be symbolized by the equation: *Document = inscription + meaning*.

In this case, the medium is ancillary, even for the traditional document, providing it preserves the inscription. What is important is the content, materialized by the inscription, which conveys the meaning. The meaning constructs itself according to the document production and distribution context, which determines the interpretation of the content.

Three leading ideas appear to us to form the basis for this dimension, in a conventional semantic triangle. The first concerns creation of the documents, the second their interpretation and the third the signs of which they are composed.

“To think is to classify”; when we produce documents, we isolate and order discourses to help us make sense of the world. Producing a document is a way of constructing or translating our social understanding. The concepts of textual genre and collection are fundamental. Documents are grouped in major categories whose different items are homologous and interrelated. This operation is carried out both upstream (production of the document) and downstream (organization of the collection). The classification varies according to the situation and era. It can be highly formalized or simply implicit. It can refer to very specific organized actions (ID, forms, contracts, etc.) or simple expectations, impressions, feelings (media, fiction, etc.). It marks our social representation and our readings of the world. It necessarily requires a system allowing the document to be placed in a set and retrieved from it, a literal or figurative indexing, and therefore concrete or abstract classification systems.

The second leading idea is interpretation. What links does the document suggest or establish and how? A document is meaningless unless read or interpreted by a reader. To a large extent, the interpretation depends on the context in which it is made. The same document can have different, even opposing, meanings depending on the period and social or individual status of the person interpreting it. In a way, each reader recreates the document when isolating and reading it. A reader must be understood in a general sense as a physical person, a group of people in different spaces and times and perhaps even a machine.

For the dimension now being examined, the document is considered in a dual relation: relation with the documentary world (classification) and relation with the natural world (interpretation). These relations are established through an “expectation horizon”, a set of familiar signs that constructs the reading contract between reader and document by allowing the reader to decrypt the meaning without difficulty as the reader is automatically placed in the interpretation context. The publishers, by their intervention on the text, its layout and their commercial action, are the first artisans of this construction for published documents. The concept of “reading contract”, whose importance we stressed in the previous entry on form, takes on additional substance, since it is also necessary for understanding the document.

The third leading idea concerns the signs themselves. Any object is potentially a sign and could be a “document”. A discussion, which has become a classic, demonstrated for instance that an antelope in a zoo (therefore in a social system of classification) was a document. But a very great majority of documents are constructed from language, mostly written and also spoken. The zoo itself is built around a discourse and the antelope can be said to be “documented”. The same remark can be made about audiovisual documents that are always accompanied by “reading captions” in the form of a very large number of texts from their production up to their exploitation. The structure of the written language, from letters of the alphabet in Indo-European languages to discourse, therefore organizes most documents. They are actually discrete pieces,

more or less separable and recombinable, analyzable, subjected to syntax, discourse structuring and style rules. This use of natural language gives documents very great plasticity.

The information explosion, i.e. the sudden increase in the number of documents which appeared at the end of the 19th century and has continued unabated to the present, led to the invention of what have been called “documentary languages” (bibliographical records, indexes, thesauruses, abstracts, etc.), organized associatively or hierarchically, which are directly derived from the above triad: it was possible to construct an artificial or formal language from the document texts (or images, or the objects themselves) in order to classify them so as to be able to retrieve them on request. Archivists have long collected metadata on documents and their producers, in the framework of archival description, which presupposes the concept of document context as the essential prerequisite for its future exploitation.

The construction of such “languages” raises many problems. First of all, it requires standardization, a number of common rules agreed upon by the protagonists. But agreement is not sufficient; incentive must be added. Each person participating in the common effort must clearly have something to gain from it or it is highly likely that the collective construction will be ineffective. Finally, the languages continually oscillate between the universal and the contingent. This oscillation is often poorly understood. It is not a conceptual weakness or the inability to choose. On the contrary, it is the dynamic underlying the documentary movement, based on the triad presented in the introduction to this entry: signs considered in a dialectic between the general that classifies and the particular that refers (interpretation).

The insistence, justified or not, of documentalists, that they differ from librarians by the service they provide, information retrieval, also reveals a certain conception of information and its independence from the medium. Documentalists say they analyze document content to give the users the answers they want instead of just the document(s) containing these answers. Documentalists thus participate in the interpretation of the available documents, reconstructing for the reader a document or a set of documents that is adapted to the reader’s need.

“Information science” arose from this movement. The term “information” is poorly defined. It is situated somewhere between “data” and “knowledge”. A more correct term would probably be “documentary units”. Information science investigates how the units fit together (a scientific idea is described in a paper, published in a number of journals, disseminated in a book, gathered in a collection, etc.) and are distributed according to highly regular statistical distributions. It attempts to perfect documentary languages and analyzes in detail the search for information as it takes place between a user or reader and an access system.

The electronic form was initially used by documentalists simply as a performing tool for classifying the items of the documentary languages in bibliographical databases. But this situation changed rapidly with the computerized processing of natural language, electronic document publishing and management, the success of the Internet and finally, modeling of the reasoning process.

Automatic language processing is far from limited to documentary applications. However, conversely, document processing is necessarily concerned by the progress made and difficulties encountered with language processing tools where full text analysis is concerned. Computer scientists and linguists have pooled their expertise, using statistical and morphosyntactic tools to create automatic indexing, abstracts and question-and-answer systems. In their area, they followed a path similar to that of documentalists, using filters and computation to reconstruct, if not a language, at least text supposed to represent the document content in a structured format, thereby enabling automatic processing by machines. The results were initially less promising

than had been expected by the promoters. Even the best tools required human intervention, and were more aids than automatic tools.

However, for the net surfer, if not for the initiate, their efficiency is spectacular in their Web application as search engines. It is striking to see that search engines, perhaps because they now concern very large numbers (of both documents and net surfers), now address the old questions of library science, already reformulated by information science: library science laws (Zipf's law), collections (hidden copies), indexing and keywords (metadata), quotes (links), loans (hits), etc., obviously renewed by computation power, using the contributions of automatic language processing, but often due more to empirical fiddling with methods than to a highly rigorous scientific analysis.

As in the previous approach, computer scientists have attempted to isolate and model logical elements. But in this case, they worked directly on content. As above, we could represent the transformation by the first equation: *Document = inscription + meaning* becomes, for electronic documents, *Electronic document = informed text + knowledge*. The replacement of *inscription* by *informed text* addresses the fact that the text (understood in the broad sense, including audiovisual) has been or could be subjected to processing allowing the units of information it contains to be extracted. The replacement of *meaning* by *knowledge* introduces the concept of personalization for a given reader or user.

The announced arrival of the Semantic Web can be understood both as a continuation of these results and as at least a methodological breakthrough if not a rupture. For the first interpretation, it can be noted that the structure of the documents is increasingly formalized (XML) and indexing is stressed (RDF). From this point of view, what is being constructed is a distributed multimedia library on the scale of the network of networks, integrating more performing search tools. The ambition is also broader. The aim is to progress from a Web which is merely a set of files linked together to a network that makes full use of the computation power of the linked machines, in particular for semantic text processing. The use of "metadata" that can be modeled and combined is essential for this purpose. Therefore, in their way, the promoters of the Semantic Web are constructing sorts of documentary languages that they call "ontologies".

The encounter between Semantic Web promoters and knowledge engineering researchers whose objective is to model the reasoning process was then inevitable. The latter have been working since the 1980s on how to extract the reasoning contained in documents. In particular, they integrate the issue of document status, modeling of the reasoning and especially of the ontologies. Ontologies have been defined as representations of a domain, placing the emphasis on the dissociation (temporary in some cases) between heuristic reasoning and a description of the concepts manipulated by these heuristics. This assumed dissociation was also a way of making it easier to model two types of knowledge, initially considered independent. Ontologies are focused on the essence of a domain (such as medicine or a medical specialty, for instance), on its vocabulary and beyond that, on meaning the meaning it conveys. This meaning has two aspects, i.e. that understood by human beings, which is interpretative semantics, and that "understood" by machines, which is the formal semantics of ontology. Ontologies can be seen as richer structures than the thesauruses or lexicons used until now, because they introduce a semantic dimension (the conceptual network) and in some cases a lexical dimension that improves access to documents. But one of the main assets of ontologies is their formal structure, allowing them to be used by computer programs, where thesauruses fail. This formal structure is obtained by decontextualizing the concepts included in the ontology. This makes it necessary, for reasons of

understandability and maintenance, to link the ontology to the lexical dimension from which it arises, to the texts.

Therefore, as for the previous entry, but perhaps in a less advanced way, we are perhaps on the threshold of a new step for electronic documents through the contribution of the Semantic Web. We could represent this step by the transformation of the above equation: *Electronic document = informed text + knowledge* would become *SW document = informed text + ontologies*.

However, documents accessible in a form not including metadata are becoming much more numerous than “indexed” documents. Even worse, competition on the Web is leading to opportunistic indexing strategies aimed at purposely misleading the search engines. It is therefore likely that two parallel dynamics will coexist, at least initially. On the one hand, for self-regulated communities that have an interest in developing performing document searches (experts, business, media, etc.), “specialized languages” will be applied to documents as far upstream as possible in their production, probably by computer-aided manual tools. On the other hand, simpler automatic metalanguages, possibly adapted to searches in broad categories, will continue to be perfected for tools widely used by net surfers.

The trend and progress of research in this dimension exhibit a cyclical aspect: old questions have to be answered anew for changes of medium, scale or tool. The controversial construction of a parallel language emerges each step. The advocates of the previous step thus have the impression that the new arrivals are rehashing old problems, whereas the latter feel that the breakthrough requires them to view all the problems in a new light. It is not really surprising for the construction of such a language to be cyclical. Each change in medium or scale requires reconstructing its structure. The mass of data to be represented and their multicultural and multilingual aspect must now be taken into account. At the same time, the foundations are not called into question, they simply are (or should be) better known and stronger.

## Definition 2

Based on this second dimension, we could present a new definition of document, still not claiming that it completely covers the concept. This definition must take into account the ability to process the content for information searches or simply for identifying the document. It is related to the second part of the reading contract we identified, that of intelligibility:

*An electronic document is a text whose elements can potentially be analyzed by a knowledge system in view of its exploitation by a competent reader.*

Once again, the definition is somewhat laborious. We therefore recall the transformations of our equations, more schematic but easier to remember:

*Document = inscription + meaning* becomes, for electronic documents, *Electronic document = informed text + knowledge*, which could lead to the following on the Semantic Web: *SW document = informed text + ontologies*.

## Questions

In coming closer to human communication, the researchers of this entry considerably increased the complexity of the problems to be solved. There are still many outstanding issues. As regards languages, for instance, there is the problem of the use of the tools on languages with a different structure and written expression than Indo-European languages. Furthermore, the boundary between automation and human intellectual work is far from clearly delimited.

But for what concerns us, it should mainly be noted that for the researchers privileging this entry, the document often appears as a secondary concept, and only the text, the content, really matters.



However, as we saw in the introduction to this approach, the content is only valid in its context. Couldn't it be said that the document is one of the constructions of this context, as it positions the information it contains with respect to that contained in other documents and allows the reader to have an idea of the value of the contents by the status of the document? In other words, can't it be considered that focusing too exclusively on processing of the text underestimates the semantic value of its inclusion in a given document? New research projects are aimed at giving more importance to the material form. The most immediate advance is to benefit from structural (and semantic in the future) markers to modulate text analyses, knowledge identification or annotation. A more detailed analysis would concern integration of material formatting elements such as font, case, indenting, lists, etc. Collaborations between document specialists and knowledge modeling specialists are then necessary.

A series of questions related to the triad given in the introduction to this entry then arises: Is it possible to analyze the meaning of a document without relating it closely or loosely to the set to which it refers (collection, category, footnotes, bibliography, etc.)? In other words, beyond laboratory work on closed corpuses, is it possible to integrate document analysis as "network head", generating structures that address the meaning? This once again raises the question, brought about by electronic documents, of enrichment of documents by links between them, in new situations of hypertextualization or of construction of collections for specific purposes.

Can information be validated other than by the authenticity of the document containing it? The problem of confidence is currently being investigated by computer scientists (in particular researchers in knowledge representation) interested in the semantic Web and by the W3C consortium. They are looking for a technical solution (addition of a formal layer that can be interpreted by software agents) allowing the users/readers of a site to increase or decrease the credibility or confidence that can be attributed to the information the document contains. This technical approach to the problem is related to the idea that on the Web, it is the users who validate information and make a site popular or not. The project could be implemented within a specialized community which has its own conventions, but rapidly reaches its limits on the scale of the Web. The problem is more complex than a "vote" for a site which makes it credible. More sociological approaches are required.

When analyzing the content of a document to generate knowledge models for particular uses, the validity and relevance of the document are undermined by other knowledge sources, i.e. the experts in a field or the users. Depending on the stated goal, the very methods used to extract knowledge from texts assign the same or even higher weight to orally expressed knowledge. The value added brought on by the authentication, certification or recognition of the text may therefore be disregarded in certain cases.

To what extent can a meaningful element be isolated from a set which has a unity of meaning, i.e. the document as a whole? Does not this unity often have a decisive weight in the meaning of the elements it comprises? How is it possible to take into account the global meaning, the semantic unity, of a document if only its parts are analyzed? The questions raised largely exceed those posed about texts in semantics. The passage from text to document probably deserves a more thorough analysis.

The answers to these questions undoubtedly differ according to the types of documents to which they are applied. But for the time being, lacking any real progress on typology, we feel they remain largely open.

## Document as Medium

Let us repeat our precautionary note on vocabulary for the last time: the term “medium” must be understood in the broad sense. It includes all the approaches that analyze documents as a social phenomenon, a tangible element of communication between human beings.

This entry is therefore related to the analysis of communication, a special instance of communication whereby the document is understood as the vector of a message between people. We can thus state that it is another aspect of the reading contract, that of sociability.

## Specialties concerned

It should first be noted that the social domain concerned includes two parts: organizations that use documents for their internal regulation and to achieve the objectives they set for themselves, and open societies or communities in which documents are circulated.

All the above researchers could probably be included in this category, since they are all interested in a social activity, but we include only those whose entry is primarily social before being instrumental.

As concerns traditional know-how as well, the disciplines already mentioned also fall into this category, but we place the emphasis on archivists, whose main mission is to keep a trace of human activity by saving documents as they are produced, and publishers, whose business is to promote the construction and publicity of documents that interest a social group.

The disciplines of the humanities and social sciences that focus on exchanges are potentially concerned by this dimension. Therefore, sociologists, economists, jurists, historians, a few psychologists, a fair number of philosophers and, of course, researchers in communication science, political science and management science are directly or indirectly interested in documents from the approach corresponding most closely to their discipline.

Electronic documents have renewed the interest of many researchers in these disciplines concerning both the phenomenon as a whole and particular situations. For instance, although it is not always assumed, there is a relation between considerations on documents and the new interest in communities of interest, working groups, networking, memory and patrimony, intellectual property, etc.

But the distance between computer scientists and other researchers is larger for this entry than for the others. Very few specialists in social and human sciences are well versed in computer science. Conversely, computer scientists often have a very limited understanding of social issues. This distance sometimes leads to fascinated enthusiasms or, on the contrary, radical rejections.

## Evolution

A document gives status, a materialized sign, to information. It is upheld by a social group that elicits, disseminates, safeguards and uses it. As we suggested in the introduction to this paper, it is compelling evidence of an existing situation and harbinger of an event. It is also a signed discourse which has an author. It is a testimony, even if that was not its purpose when designed. It is material proof.

To be consistent with the previous entries, we propose the third and last definition as an equation: *Document = inscription + legitimacy*. This equation seems to us to represent the social process of document creation. Document status can be acquired under two conditions: To be legitimate, the inscription must have a scope that is beyond private communication (between a few people) and the legitimacy must be more than ephemeral (go beyond the moment of its enunciation) and

therefore be recorded, inscribed. These two conditions mean that although any sign can be a document, a particular sign, even if it satisfies the two dimensions discussed above, is not necessarily a document. For instance, a diary is not a document unless someone takes the initiative of making it public or at least communicating it beyond the circle of relations of its author. And a radio or television program is not a document unless someone records it for future social use.

This statement does not meet with the agreement of all the contributors to this text. For some, the value of a document could preexist its communication or recording.

Document status is not gained once and for all. It is acquired and may be lost and totally forgotten forever. It may also be regained if someone rediscovers and relegitimizes a document which has disappeared from the collective consciousness but has not been destroyed.

However, the equation does not account for the social function of documents. Documents are used to regulate human societies by ensuring communication and the durability of the norm and knowledge necessary for their survival or their continuity. In a way, it could be said that the reading contract, two of whose dimensions, corresponding to the two above entries, we have identified: legibility and understanding, takes on its third dimension with this entry: sociability, i.e. appropriation whereby the reader marks participation in a human society by reading the documents or, conversely, the inscription on an artifact of a representation of the natural world and its inclusion in a collective heritage.

A document is not necessarily published. Many documents, for instance ones on private matters (medical file, private transaction between individuals), or ones containing undisclosed confidential information can only be viewed by a very limited number of people. However, they have a social character in that they are written according to established rules, making them legitimate, are used in official relations and can be submitted as reference in case of a dysfunction (dispute). Conversely, publication, general or limited, is a simple means of legitimization since once a text has been made public i.e. available for consultation by a large number of people, it becomes part of the common heritage. It can no longer easily be amended, and its value is appreciated collectively.

The multiplication of documents is therefore connected with the evolution of societies by two dynamics, one external and the other internal, which mutually reinforce one another: first the social use of the documents and secondly their specific economics.

Political and social organization is based on the production and exchange of documents. Religions and priests, governments and administrations, productive organizations and trade, civil society, in their different components, their historical evolution, their specific geographies and cultures, their changing functions, have used and continue to use documents extensively for their internal regulation and for the competitive assertion of their identity and position. For instance, the main sources of documentary activity in western countries include, without being restricted to:

- In France, the transition from the Monarchy to the Republic, then from a police state to a welfare state and finally, today, integration of the state in more vast groupings such as the European Union, and globalization, have all had consequences on the production, role and number of documents. For comparison, it is sufficient to mention the importance of documents in the parallel history of the administration of China to understand how basic they are, while at the same time specific to each civilization.
- Industrialization, with all the technical, organizational, transactional and accounting know-hows and standardizations which accompanied it, “produced” a considerable

number of documents. It is perhaps the main factor in the documentary explosion mentioned above.

- The progress in science and education considerably increased the number of document producers and consumers, for the internal functioning of science and even more for the popularization of the countless concurrent partial know-hows.
- Exchanges, commercial and other, which exploded with the development of transportation and telecommunications and the opening of borders, use a considerable number of documents to “flow smoothly” (materialization of transactions, technical manuals accompanying products and services, sales information, etc.).
- The development of leisure time, a longer life expectancy and the increase in “public space” are also essential factors in the development of culture and one of its main vectors: documents.

This dynamic has been examined from various angles, but the few attempts at a general understanding appear to us to be more speculative than demonstrative, perhaps because they are so isolated.

The second dynamic for substantiating documents as medium is their internal economics, based on changes in the technologies from which they are constructed (changes developed in the two above entries) and on the document creation processes. These processes require work and the way of doing this work must be found. The creation of a document can be analyzed as an ordinary act of communication with one (or more) senders and one (or more) receivers. Special fields correspond to specific parts of the process and specific areas of application. Systems have been constructed and officialized to satisfy the need for regulation of production. Small and large businesses have arisen to meet the challenge and organizations have taken over the job. There is a cost for setting up and maintaining these systems, which have their own inertia.

Two main research currents are focused on investigating the economics of document creation. The first is interested in organizational communication and studies documents primarily as a business process; the second analyzes communication in the media and investigates the publishing process.

Research on organizational communication studies documents immersed in characteristic business practices and situations and therefore constrained by rule systems. It examines documents on several levels: first of all as written material identified in a context, officialized by rules governing writing, dissemination, use, recording of an intention related to an action, and preserving a trace of the social and technical negotiations conducted around it. This leads in particular to examining document production and management processes, activities which are not restricted to specialized players but which are distributed throughout the organization. It also considers documents as a structuring element of the organization, as coordination support. Finally, it considers documents as instruments used by individual or collective agents in their different strategies. From a methodological standpoint, documents are considered “observable”, allowing the study of relations between players (they are mediators), regulation modes (they are a management tool) and organizational recompositions (they are one of the elements that reveal this).

Initial progress in the analysis concerns only part of documentary activity and is still far from determining the formal contours of a general document economics. Many uncertainties remain to be dispelled, such as those concerning the relations between documentary systems and organizational systems, the evolution of the different sectors of mediation and the economics of libraries or archives. Among archivists in particular, there are many discussions around *records management* and *business reengineering* practices, which are becoming standardized. The

current doctrine (which is however far from being assimilated by institutional document producers) states that the missions (institutional objectives) generate processes (organizational functions), which generate procedures (officialized methods of action), which generate documents (or transactions).

As concerns the media, the economics of several sectors is well known because it has been the subject of particular analyses. Mention can be made of scientific communications with the role of publication of papers, peer review, citations, and prepublications. This is also the case of media aimed at the general public, with the gradation from publishing, an artisanal activity based on the individual sale of objects, the dialectics between backlist and best sellers and distribution networks; and broadcasting, a more industrial activity, organized to capture the attention of millions of people in their homes and sell it to advertisers.

Other topics have been investigated because of the economic stakes they represent, such as intellectual property rights. In this area, certain document properties can be identified by examining the difference between the Latin “droit d’auteur” and the Anglo-Saxon “copyright”. The first privileges the author’s attachment to the work, whereas the second privileges the notion of publication, giving intellectual property rights to the person that takes the initiative. In a way, it could be said that the “droit d’auteur” is a right to the work, whereas the “copyright” is a right to the document.

There is one last point to be mentioned concerning document economics, very important for our subject: the more the existence of a document is known, the more it is read, and the more it is read, the more its existence becomes known. A resonance phenomenon can develop from the relations between readers and those between documents. It can take on different forms and has different names depending on the sector and specialty. Marketing professionals and media strategists make use of this phenomenon regularly by constructing public awareness which then then sell on other media. In the area of scientific communication, the impact factor based on the number of times a paper is cited in other papers, is related to the same process (and has led to the same excesses). This feature may explain many of the characteristics of document distribution: best sellers in publishing, prime time in broadcasting, various fashions, concentration and expansion or the almost perfect regularity of bibliometric laws when large numbers of documents are equally accessible to large numbers of users.

The digital world leads to contradictory movements that are not easy to interpret. The first observation could be the disappearance of a large number of documents, which in their traditional form reported on procedures. This disappearance is difficult to evaluate, because it took place anarchically. The dissemination of computerized tools often led to the dissociation of functions performed earlier by a single type of document. That is the case of civil registers, which continue to be kept on paper for legal reasons but are in electronic form for consultation. Increasingly, the replacement, coinciding with the disappearance of middle management, is total: forms, schedule boards, machine operating schematics, instructions for use which were commonplace in public and private bureaucracies have been replaced by databases and data networks. This movement, dubbed the “computerization of society” a few years ago, risks speeding up even more with the developments mentioned under the above dimensions.

But a concurrent increase in written material and documents is observed in organizations, amplified exponentially by the quality approach. Documents codify social and organizational standards, turning them into supports for action as well as a memory of relations. Storing data and procedures in databases does not detract from their prescriptive value, quite the contrary. For instance, intranets give documents a status (as reference or tool in particular) by associating

identification and circulation rules with them while modeling and anticipating on the possible uses. They amplify the visibility of decisions and activities by making them largely accessible. In this perspective, as never before, a document alone is devoid of meaning, as it is the electronic storage of transactions defined in advance. The display of information, ephemeral and necessarily dependent upon evolving technologies, is not in itself a document, but needs to be validated by certified procedures. Many documents are similar to the transcription of procedures or one of the stages thereof: such documents could then only be understood in relation to the mode of information translation to which the procedures they describe have been subjected. Considering the progress made on electronic signatures, many transactions could be completed in the future without the formalisms adopted for printed documents.

This mutation considerably increases the possibilities of invasion of privacy through information cross-referencing capabilities. In the social sphere, France has established legal protection, fragile considering the development of electronic transactions, by the “Informatique et liberté” (Freedom of information) law. In the economic sphere, certain analysts have seen the emergence of a new economy the hazards of which far exceed the scope of this paper. For our subject, let us retain the idea of a radical change in social and economic structures. As the industrial age was marked by the interchangeability of parts, the information age can be characterized as the possibility of reusing information.

From the standpoint of organizational communication, we may have identified a first change of our equation from *Document = inscription + legitimacy* to *Electronic document = text + procedure*. But this equation does not account for another very important movement now taking place in the media since the advent of the Web.

The Web suddenly projected the electronic age to the scale of society as a whole. To understand the success it has had, measured by its explosive growth with the public and depending on the types of activity, it is necessary to reexamine the spirit underlying its architecture. The organization of the Web is consistent with the guidelines of the designers of the Internet, imagined as a peer-to-peer communication network where each node, large or small, has the same tools and is both producer and consumer. The Web assumes a social design, or rather a social communication, similar to a “Republic of Sciences” or the freeware movement. In such a society, each person is a player and is responsible to the community for his or her acts. Translating this into our area, we can say that everyone is capable of reading or writing documents concerning community life and everyone must be careful to publish only documents that enrich the community. The pioneering geniuses of the Web, a combination of the Internet and hypermedia, built the system they, or rather the information community to which they belonged, wanted.

This idea is very present in many statements and initiatives in this area, beginning with those of the W3C consortium. The information packaging industry (software and telecommunication) is very attentive to these developments, not without discussion and compromises, because they strengthen its positions by promoting the increase in traffic and processing to the detriment of an information content .

But everyone can't speak to everyone, it would be cacophony, so representatives are necessary. Up to now, this difficulty has been solved by using one or more filter systems to select relevant authors and configure representative and useful documents. Such systems have a cost which cannot be diluted in community operation, since the equality between players has disappeared. Only a few write on behalf of others and professional mediators organize the publication and access system as a whole. The editorial system is an avatar of this organization, a compromise between private and public interests.

There is therefore an initial misunderstanding, willed or not, between systems designed by Internet pioneers and supported by the information packaging businesses and the ordinary reality of social communication. This misunderstanding is however very fertile, since it opens a space of exchange for communities in which communication is restricted by the traditional system. It also gives a number of institutions, starting with those of public interest, a simple tool for communicating with the population at large. The Web is an enormous store containing a multitude of linked documents that can be viewed free of charge by the reader.

There are those who believe this organization is only temporary, illustrating the youth of the medium. This analysis may not take the full measure of the rupture that has occurred with the Web. It is also possible that filtering and selection will no longer take place *a priori* on the Web as in traditional media, but will instead be performed *a posteriori* using a “percolation” system whereby the most relevant documents are gradually identified and placed at the forefront by the number of hyperlinks and the operation of the search engines. The main thing would then be the Web itself which, by its continual movement (links are created and destroyed, engines run, pages appear and disappear), would allow identification of documents. The involvement of a substantial number of net surfers, heretofore excluded from the small closed world of the broadcasting media, and the manifest success of this new media in the practices support this hypothesis by providing a dimension and speed unheard of in the ordinary dynamic of legitimacy through renown.

Following a parallel reasoning, but in a more literary perspective, a number of researchers and essayists have predicted that the Web, and especially hypertext and hypermedia techniques, would lead to the disappearance of documents. The conventional triad, author-work-reader, at the origin of the construction of a literary document, could then give way to an interactive process in which the links between accessible pages played a more important role than the text as it was first constructed by the author. However, although interesting experiments in hypertext writing have been and are still being conducted with nonnegligible semantic and cognitive consequences, the explosive development of the Web appears to have led on the contrary to an exponential increase in the number of documents on line. The links between pages appear to be becoming gradually structured to create new paratext standards, reinforcing on the contrary the documentary aspect of the Web.

We could summarize our development on the Web by transforming our initial equation to *Web document = publication + access cueing*. Publishing alone would not guarantee legitimacy. Notoriety would also be necessary by access cueing.

Consistently with our reasoning, the traditional media have not been able to construct economically viable business models for the Web. Only a few sectors which already had affinities with networks have been able to make money: financial information and scientific information. It is also possible that music, because of peer-to-peer exchanges, is in the process of redefining its distribution and pricing mode.

Conversely, the electronic age has strengthened the “old” media, publishing and broadcasting, by allowing them to make substantial gains in productivity by promoting synergies and diversification. For instance, using the same database, a newspaper can publish news in a newspaper and on the Web, broadcast it on the radio and by SMS and audiotel, etc. And each medium can enhance its own areas of excellence (notoriety through television or radio, interactivity through the Web and telephone, appropriation through publishing) and the resonance mentioned above can lead to exceptional profitabilities. These recent changes still need to be evaluated. They also lead to high investments whose returns are not immediate whereas the future is uncertain. After wild success followed by great disappointment, the Web appears as just

another medium whose intrinsic features need to be well understood to articulate it with existing media.

The announced developments of the Semantic Web will probably lead to other developments, in particular in the relations between document and service. But that is still futuristic and difficult to cover in this entry with a social dimension.

### **Definition 3**

In this perspective, we identified strong movements, in some cases divergent, often chaotic. At this stage, it is not easy to propose a definition that clearly reflects this third entry. That is why we will give only a very general definition:

*An electronic document is a trace of social relations reconstructed by computer systems*

We recall below the equations we constructed and transformed: *document = inscription + legitimacy* becomes *electronic document = text + procedure* and *Web document = publication + cued access*.

Despite the difficulty in constructing a definition, we must stress the importance of this third dimension identified in the reading contract, i.e. sociability.

### **Questions**

The first series of questions concerns the notion of archive, for which the basis is the recording and preservation of documents. The role of an archive is to preserve the memory of a human activity. A new, more active, role is emerging for archives with the electronic age, with open archives, retrieval of audiovisual programs at the source or of television broadcasts, archiving of the Web, etc. Many new activities are developing and archive science practices are changing. The *records management* being set up in organizations appears as a requisite for satisfactory electronic archiving. There are also as yet not fully answered questions on a different role to be played: hesitation between a record of past action and a record of an ongoing action; confusion between archiving and publication; simple recording or preparation of future use. Even more so, how to preserve a trace of the continual renewal movement of linked pages?

The second series of questions concerns the concept of attention (types of perception and intention) without which a document cannot have a reader. Human attention is limited by the time available, reader fatigue and by the reader's technical or intellectual faculties. This problem is well known to broadcasters.

Net surfers are necessarily active. There is no way of "hooking" them like broadcast audiences. In other words, the Web combines the freedom of choice of publishing with the accessibility of broadcasting, or extends library services to the entire planet for collection and to the home for consultation. Bibliometric laws and resonance effects might occur on a scale unheard of in some sectors: attention is concentrated very closely on a small number of documents and dispersed on a very large number. These phenomena and their consequences have not yet been studied in detail. Furthermore, the intention of Web promoters is to make sites and documents available to the entire planet on an equal basis. But the penetration of innovations is very unequal. The Web and electronic documents are no exception to this rule. Even worse, media access appears to be the most inequitably shared goods between countries and between the different populations within each country.

The third series of questions concerns the omission of funding of content. The path of least resistance principle applied to the accessibility of the Web means that net surfers prefer to avoid obstacles and barriers to browsing rather than confront them head on. They will therefore



circumnavigate all direct requests for money. In the same dynamic, a militant movement asserts that the Web should be free and the access to knowledge and culture liberated from commercial imperatives.

Opportunism and politics are combining to gradually configure the economics of content on the Web as an institutional B2B (business-to-business) market. Is it certain that this financial structure guarantees the diversity and plurality of online documents in the medium term? Is it even certain that it contains sufficient resources in each sector to keep producing and managing documents?

The idea of translating all existing documents from traditional medium to electronic form is unconceivable. However, the explosive development of computer science makes it necessary to envision an enormous processing effort without which there is a risk of radical amnesia of our documentary culture. In the near future, reasoned choices will be required (what documents should be digitized first?) and tools capable of processing huge volumes of documents will have to be constructed.

## Conclusion

The three entries discussed highlighted several of the basic themes regarding documents, reinforced or undermined by the electronic version. We now need to consider a summary giving a general view covering the three entries, somewhat like obtaining all the colors from the three primary colors. More academically, is it possible to envision a document theory making us better able to measure the present and future consequences of electronic documents?

First, it is obvious that under each entry, we identified stages in the history of the electronic conversion of documents that we can now compare. Traditional documents consist of a medium, a text and a legitimacy. The first stage of electronic conversion, where we probably still are, highlighted the internal structures of the document, the importance of metadata for processing and the difficulty of validation. The second stage, which has undoubtedly already begun but whose conclusion remains uncertain, and which stresses the XML format integrating structure but not form, is based on ontologies for retrieving and reconstructing texts and places the emphasis on personalized access. There is a meaning to this general evolution that must be better understood as to its orientation, consequences and limits.

It should be stressed that the opposition between paper and electronic versions is futile. Almost all current documents have existed in electronic form at one stage of their life and those that haven't risk being totally forgotten. Conversely, numerous electronic documents are printed at some point on a personal printer or a professional printer. What is important is therefore to have a better idea of the concept of document in general, whose electronic form is both revealing and a factor of evolution.

Finally, it should be noted that under each entry, we emphasized the idea of reading contract, expressed as legibility in the first case, understanding in the second and sociability in the third. It is probable that this contract with three aspects, in all the nuances given, represents the reality of the concept of document. A document may finally be nothing more than a contract between people whose anthropological (legibility-perception), intellectual (understanding-assimilation) and social (sociability-integration) properties may form the basis for part of their humanity, their capability to live together. In this perspective, the electronic form is only one way of multiplying and changing such contracts. But the importance it has gained, its performance and its speed of dissemination make a thorough, careful analysis even more necessary. We clearly showed, in

particular in the series of questions, that none of the entries was independent. It would be futile to attempt to separate them. On the contrary, the concept is only clarified by superimposing them. But we also noted that each entry was taken up in a multidisciplinary research movement which has its autonomy and whose specialization involves expertise that is too particular to be fully sharable.

This paper is a call for more indepth studies to compare each of these approaches and investigate how they intersect..

Roger T. Pédaque  
CNRS - STIC  
08-07-2003