Seeking Emergences from Digital Documents in Large Repositories

Davide Musella +
Marco Padula °

1. The evolution of document conservation and use

Museums and libraries have always been cultural repositories for the preservation of our social memory.

In the civilization of ancient Egypt, knowledge was not stored inside buildings, such as libraries or museums made for that purpose : the Egyptians documented their history with sculptures, monuments, and columns covered with writings and bas-reliefs. Anyone who could read had access to it, navigating through their country which represented their endless encyclopedia : every information could be reached and linked to everything else. This gives us the flavor of what *global communication* means. Cultural communication in the Rome of the first century A. C. was analogous : the city was considered a repository of documents, continuously updated and accessible to everybody.

The *mouseion* of Alexandria (third century B. C.) was part of the famous library, and all kinds of documents were housed inside the same building. In thirteenth century Italy the library became an autonomous institution, independent from the museum, which was considered a collection of artifacts after the fifteenth century. This idea spread later from Florence to all of Europe.

The evolution of museums and libraries has continued toward a thematic specialization of the institutions (the natural science museum, technological museum, public library, mathematics library, humanistic library) and a technological specialization of the communication of the collections of documents (exhibition of real objects, consultation of books, scientific films, posters with photos and the explanation of an animal, digital archives with paintings, audio explanations). This has given rise to a very disjointed cultural repository composed of varied collections, each of a specialistic subject, presented in different forms through different media, and located in a different geographical site. It is a long and hard work to retrieve materials for a global documentation according to a cultural design, even on a single topic. In the Global Information Society, the problem of cultural memorization and communication assumes even dramatic importance.

There was a time when digitization gave each different document form (pictures, films, sound, writings) a homogeneous basis so that each one became a numeric sequence. Data models were defined as the formalization of the different real document structures and the ways of manipulating them ; DataBase Management Systems, Information Retrieval Systems, Image Data Banks for automatic information management were implemented on the basis of the models defined.

The growth of computer networks and the birth of information highways have gone hand in hand with a tremendous increase in message exchange : multimedia and hypermedia technology supply more compact supports for data storing, and the tools for their organization according to their conceptual usage ; with cooperative methodologies for both work organization [Ehn, 1988] and system design [Clement, Besselaar, 1993 ; Carmel *et al.,* 1993 ; Bianchi *et al.,* 1996], the information technologies are shortening the historical gap between end users and social needs.

After experiencing the automatic information systems for managing their undertakings, cultural institutions are now entering the world of circulating information, with which environment they partially overlap. Traditionally these institutions are distinguished by the kind of objects they conserve and offer, the digital versions of which were conserved in homogeneous repositories fitting specific data models.

The Global Internet modifies the configuration of these institutions and the scenario in which they exist, combining them in what William Wulf has called the *collaboratory* [Cerf *et al.,* 1993]. People cooperate in their specialized work places, using tools and protocols, as for example in an Intranet context, which accompany them in the Internet world, hiding the differences among the elements, which become components of the collaboratory. The Global Internet is therefore an infrastructure for *cyberspace,* a term introduced by William Gibson in his novel *Neuromancer* [1984], and frequently cited since the information revolution exploded. Cyberspace is the space of pure information, a channel for cultural communication, the collective memory, and, analyzed from an anthropological perspective, it becomes a cultural space, as argued by Pierre Lévy [1994].

In specific cases we speak of *global library* and *global museum,* referring to the *virtual* versions of the workplaces where the information activity begins **(Figure 1).**

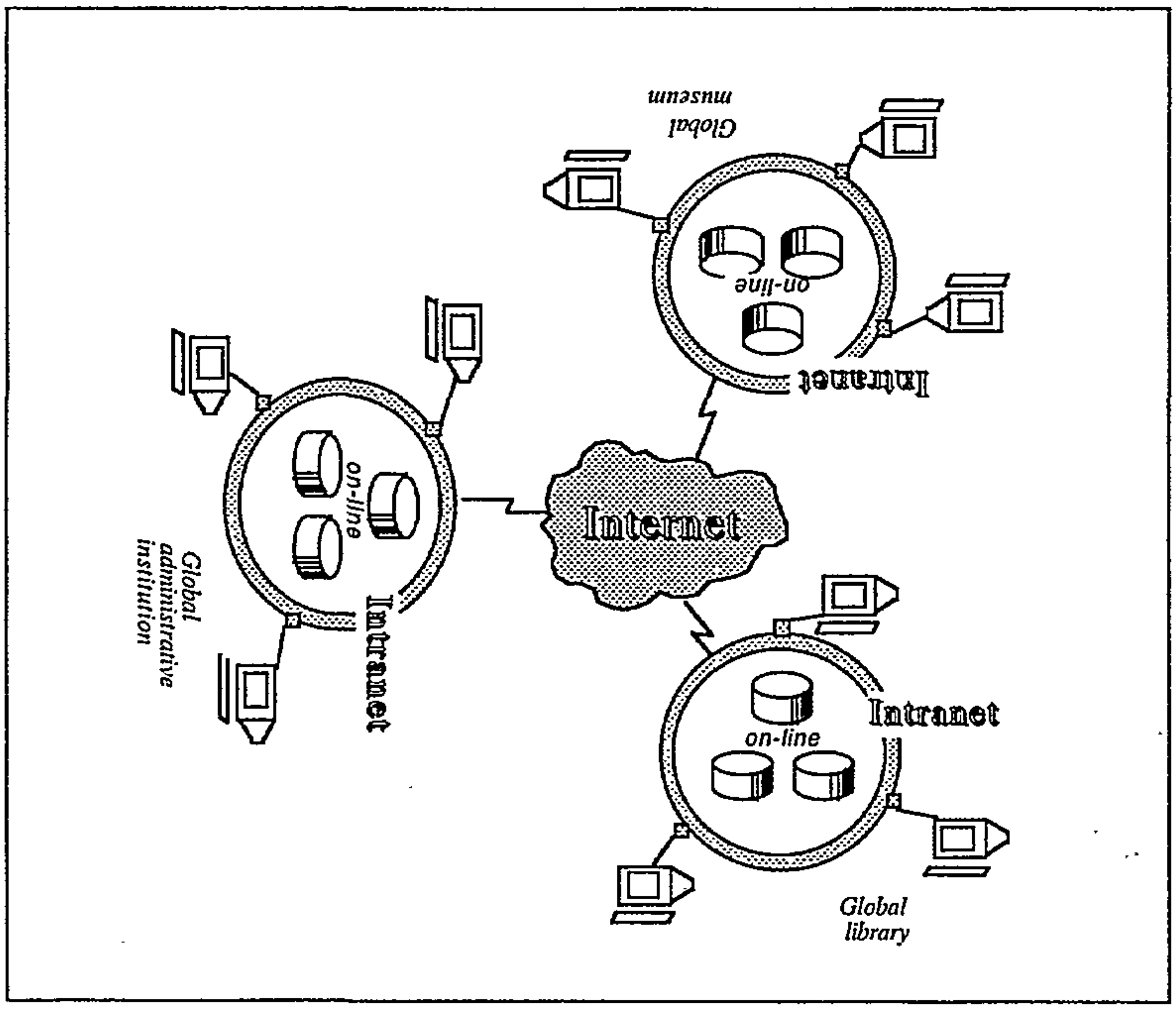

Figure 1

A global museum, or global library, or global institution is a view on the collaboratory from a specific perspective

The mechanisms for Web space indexing have strongly influenced the configuration of libraries. Traditional information centers concentrate in one building all the relevant activities, from document storage to cataloguing by the librarian, indexing on cards in filing cabinets, and search and consultation by users as required. In the global Internet library documents from everywhere are immediately available, and do not need to be moved from where they are authored, while the general index Data Bases, which are sometimes dreamt as *universal* index DBs, are centralized.

The organization of documents is being rethought to allow for the effective disclosure of the emergences which arise from the interconnections of document contents and from the activities of the mass of users. They are being reformulated following the traces of the broader user population and considering the problems from a perspective globally open to the different inputs from society and methodology, from technology and the market [Musella, Padula, 1997].

The role of the technologist is to create a continuously stressed link between speculation and pragmatism, and cun the more abstract ideas for his operating laboratory : while analyzing social needs, he designs solutions based on available or foreseeable infrastructures, methodologies and communication languages.

From an operational perspective, today's challenge lies in the power of the client — server paradigm used to design digital networks architecture. This moves netizens from the role of remote consultants to that of coworkers [Kouzes *et al.,* 1996] : information consumers no longer wait for delivery through the network. They don the netizen uniform and navigate the Internet in search of repositories, and they share all the responsibility for the collaboratory's maintainance.

In the netizen community, personal identities and characteristics lose importance, as do professional profiles and affiliations : only the disseminated contents, the actions performed, and their effects are perceived as effective and relevant. Authors and content suppliers are attracted and compelled to enter the Internet community, and nobody knows who or what they are. Dissemination of their results, namely their publications, is more spontaneous as it requires no authorization, and it is also transient. That means a dramatic extension of popular participation with the opinions of the many and the contribution of their products. But the other side of the coin is an enormous mixture of quality resources and junk in the information available.

The new tools available make electronic publishing and diffusion easy. Many authors are only too happy to bypass editorial procedures and propose their work directly to the World Wide Web community. This does not replace paper editions, but simply move and develop a market sector where the criteria of offer and demand satisfaction change : product quality decreases as the only referee is the author himself, language becomes more original, but also less consistent, typos escape more easily in galley proofs. The counterpart is that works are proposed without delay, providing rapid access to new ideas at practically no cost, as the services of the editors, publishers, book sellers, librarians and indexers are no longer needed.

Information customization [Berleant, Berghel, 1994a, 1994b] addresses a *semantic* manipulation for extracting and synthesizing interesting parts of document content, and a *representative* manipulation to transform them into a form more appropriate to the user's information consumption needs. User needs change quite rapidly, depending on the partial results of information analysis. Therefore, information customization calls for the interoperability of functionalities, which requires a high degree of user interaction and supports the performance in real time of navigation, of document retrieval and filtering, browsing and editing, reporting and annotatation, re-contextualization by settling hypertextual links and reorganizing the collected elements, data externalizing, document content analysis and extraction, followed by its abstraction.

Among the new services that are becoming most relevant (task force groups have been constituted to design the tools and define the protocols and standards) we shall necessarily see the collecting, indexing, assembling and storing of material already on-line. Many robots have been implemented which search for information and organize the collected material, or more precisely (and efficiently) their references, into general archives. However, they do not embody domain expertise, and are quite computationally and memory expensive, and consequently the definition of languages for describing documents for their further indexing, and of rules for robot behavior has attracted many intellectual resources.

Digital technology has enhanced capabilities for transferring, modifying and replicating information for its customization, with a consequent redistribution of local tasks and responsabilities, service providers which are usually contracted outside libraries for cataloguing and indexing, or authors for copy preparation for press. We can expect that the availability of easy to use and powerful tools will emphasize the decentralization of activities and force us to update our skills to meet new needs and focus on what new information could be added by our efforts.

Internet has provided the fertile ground for information activities. The semantic difficulty lies in tracking contents from different scattered collections. Emergent elements are often only implicit in the documental material, as they were in the author's mind : it takes collaboration among differently skilled people customizing the collected information to better study and interpret it, to disclose the emergences that enable the design of new knowledge.

The spiralling wealth of resources hosted by servers and of providers which offer them entangles even more the already complex cyberlinks ; content consumers, attracted by contents which emerge from never ending information combinations, feed the demand for new resources and the clients to exploit them. H. Berghel and D. Berleant have argued that hyperlinks do not scale well [Berleant, Berghel, 1994], but we feel that the problem is not in the complexity of communication channels, nor particularly in the set of cyberlinks (World Wide Web), rather than in information presentation and organization. The complexity of cyberspace, namely the global Internet, suggests and stimulates research for new tools and methods to support information seeking. Traditional technologies for information cataloguing, indexing, retrieval, filtering, and for content extraction can not provide the results of the past, due to the mass of information now involved, its latency time, and the speed demanded by consumers in their searches.

Large scale document customization becomes cyberspace customization, moving the focus from single document manipulation to the re-design of document context, namely its virtual container, according to a new cultural need, or project.

When netizens don the information consumer's uniform, their navigation is drawn to documents customized by means of *cybermaps,* specialized homogeneous collections, or more structured folders dense in content descriptions and references ; when they don the information seeker's uniform to disclose content emergences, they carry on wide research influenced by documents customized by effective cataloguing.

Information seeking used to be characterized by a sequence of query and result analyses. Today this is intermingled with steps for navigation. The vastness of Web space and the availability of search tools have simplified search procedures for particular information, but also worsened the precision of the results. Seeking documents by argument or concept is a difficult and often itèractive task and gives rise to a large number of items which must be analyzed individually [Janes, Rosenfeld, 1996]. This makes it very important that the presentation of retrieval results by means of syntheses of document contents be appealing to the user's interest and lead his navigation to discover the emergences. Appealing, meaningful, and robustly organized descriptions give him greater visibility of where to move, and could therefore resolve the *lost in cyberspace* problem. Cataloguing is fundamental to a good description of Internet resources. This is why it has assumed dramatic importance : today it is a responsibility of the automatic tools (robots), but it is perhaps an unobtrusive task which has been improperly left to automation. In fact, the results obtained are useful but not completely satisfactory, due to the poor characterization of the documents examined. Authors must realize that the externalization of document contents is entirely up to them : documents must be carefully edited (should each component be described ? at what level of granularity ? how should the description be authored and in what detail ? what will the user want to know about it and for what goal ?), taking into account the way robots index and present them through the search engines [Northern Webs, 1996 ; Tomaiuolo, Packer, 1996].

When the exponential growth of the Internet became evident, WWW analysts began to consider document cataloguing and indexing a complex problem of uncertain solution. Many efforts have been directed to realizing new and efficient automatic tools, but without a common standard basis, so that today it is impossible to improve on robot power. More tools have not implied better results. What is needed are the design of a firm infrastructure with a limited number of standards for tools, communication protocols and description languages that scale well while offering at least a minimal effectivenes and are direct and easy to use. They should emerge, if possible, from the *cybercommunity,* both as the result of the activity of task force groups and as the consolidation of practices or informal conventions.

We shall argue here about these languages that can, above all, satisfy the multitude of WWW navigators. A large variety of descriptive languages for cataloguing have been proposed to satisfy the widely varied needs and problems : we have, for example, the Dublin Core [Dublin Core, 1996] and the Warwick Framework [Lagoze *et al.,* 1996], the IDML (IDentify Markup Language) [IDML, 1996], and those adopted by robots such as Altavista [Altavista] or WebCrawler [AOL robot]. None of these languages has had a large circulation.

2. The meaning of the document resource

Traditional libraries, or more generally information centers, have a highly specialized concept that defines documents as single permanent and precisely defined objects, such as as books, papers, journals [Levy, Marshall, 1995 ; Wiederhold, 1995]. These institutions provide for archives organised on uniform criteria which allow to retrieve sets of documents which are associated by common properties. These connected sets present the idea of a *global complex document* intertwined to assist the inclusion of content emergences which have successfully been disclosed, and that fade today, due to the interface of the concurrent and immediate availability of different, independent, and separate archives and catalogues created and organized with completely different criteria and purposes. The possibility of assembling all the heterogeneous data useful for the same project depends on the ease with which we can navigate from one to any other datum, following related ideas. In this way the researcher authors an ephemeral, malleable document, continuously modifying it in its components and in the paths which interconnect these, as he conceives new aspects, or moves to different viewpoints.

Information centers house documents, but what these documents are in the collaboratory is not clear. The document itself, which was conserved and exhibited, or offered for consultation has lost concreteness due to the irrelevance of its support : with the new technologies that allow reliable reproductions of the original document, the value of the document depends only on its ability to communicate. So we must focus first of all on what the documents are [Levy, Marshall, 1995 ; Wiederhold, 1995 ; Shamber, 1996 ; Janes, Rosenfeld, 1996], taking into consideration the new supports which record their contents, the communication space, for example the WWW, where they are distributed, and how netizens use them.

The new collaborative paradigm for information seeking and providing is detailing and accelerating the development and dissemination of knowledge by means of the action of people which are immersed in a networked environment where they navigate, *e-dialogue* with information suppliers and with colleagues, devise new methods for finding and interpreting information and for discussing results through Internet discussion groups, exploit a large variety of material which makes it difficult to follow the structural borders of a document, distributed among the networked and hyperlinked document components. The user of a digital document receives only a copy, from among the unlimited number available, of a requested work. There is no reason for it to be otherwise, for it is a literal image of the master : it could be a selection or a composition with other works which will surely be further processed by the reader. *The document could be completely handled and modified.* The document can be manipulated by reorganizing the representation or the structure of the contents, but also by modifying any part of a digital document circulated on the Internet. The authorship of the document loses importance because the author cannot be identified with any certainty, and its manipulability becomes extreme. The document is virtual, neither its original historical context, nor its originality can be verified. The importance and the value of the document are concentrated in its contextualized conceptual content and in the dynamics of its evolution.

Persistence and transiency have therefore become important factors in the multimedia world [Gudivada, 1995] as fifty million Internet netizens interact with each other and with all the available resources. What is persistent ? What even is of interest or useful for its age, such as historical information ; or what is limited, so limited that very few people interact with it and, therefore, update and develop it. Persistent data are organized and stored in well structured DBMS or IRS for later retrieval. Transient data are accessed simultaneously by many collaborators which communicate interactively and continuously customize them for individual purposes ; they have an immediate use in the more dynamic situations on which evolution feeds, and are thrown into unordered repositories. In the end, which piece of information, what document is worth considering ? How old should it be ?

In the above discussion two standpoints for document modeling intermingle implicitly. The processing and management of a multimedia document requires the definition of an *a priori* model that, on the one hand, outlines the complexity and value of its content (representative value) through a formal description of its components and their relationships, on the other defines the cost for its production and the value of its use through the specification of the modalities and procedures for operating on it and for interacting through it. There can be no doubt that interactivity characterises and valorizes multimedia with respect to a monomedium such as a picture, photograph, cinematic and televised products, and contributes in making it a hypertrophic version of its antecedents, that are the homogeneous media which are fused to build an artificial representation of objects or concepts in their context, which is closely constrained to the digital automatic world, but becomes active and dynamic in the new cultural space.

Some document characteristics are now worth being summarized :

— *stability, i. e.* the rate of change of the document over time ;

— *lifetime* of its meaning or usefulness ; *e. g.* correspondence could be conserved for legal reasons and, after that time, simply for its historical interest ; personal communications and notes may only be useful to the writer, until the conclusion of his book or paper ;

— *version* which assumes importance when instable documents are managed ;

— *composition,* which refers to all subparts and to their inter-links, including those in another archive. What type of data, or material object could be part of a document in a digital library ? Can a digital library be considered a document archive or does it become necessarily a hopelessly disordered environment ?

— *formats* for digital documents are changing. There are some advantages in the traditional paper-based press (at least for the consolidated conventions to which we are accustomed) but organizational structures and formats presenting digital data are very different and will make them obsolete. We actually move much more material than needed, both because documents are rarely structured for efficient use, and because they are usually redundant for our goal.

Like images in a film, multimedia documents are animated entities in the Internet, changing in time during :

— production and updating to show a resource identified by a same URL (Uniform Resource Locator), that evolves with the workgroup activity, determining document stability, lifetime and versioning ;

— composition to include films, sounds, or simulations of events which vary over time. This determines content externalization ;

— dissemination to control the information seekers' perception of its content externalization when they do not access the document by direct knowledge of its URL, but through a search engine associated with a robot which filters the document's history (see for example the discussion about document research and indexing in [Northern Webs, 1996 ; Musella, Padula, 1997]). This transient document is a sequence of snapshots, each of which is partially overlapped to its predecessor : therefore, the time determining the trend of its evolution structures its reading [Zeitoun, 1993], and the robot sampling the snapshots modifies the document reading.

The document's evolution is determined by the process for its production and the value of its use, that is the interactions that take place.

It could be the *means* for navigating, or the *object* of a robot collection, or of a folder, or a guide producer. The interaction simply exploits the document's functionalities in the former case, while it modifies the document's evolution in the latter one. We have already spoken of the robot's interaction ; the collaborators in the workgroup interact to produce and update it, and a folder or guide producer inserts the document into a new context that is different from the original one, and in this way gives the document a new life.

An observer will perceive the effects of the evolution of a document in its different versions, but he will guess the procedures followed for its production only if he knows the axioms, the conventional rules, and the goals adopted to define its model.

Information intensive activities have highlighted the aspect of group cooperation in consuming information, providing contents and seeking emergences. Documents become more and more the result, or the logbooks of e-discussions, communication and report exchanges, annotation, collection browsing. The cooperating group is responsible for the evolution of a document, which is more closely tied to the communication medium than to the object of a conversation, fits more the idea of communication than that of artifact [Shamber, 1996]. The concretization of an emergence and no longer the description of a predefined content, its uniqueness fades, its boundaries are not clearly limited, it is identified by the URL of its components. These are heterogeneously granulated, and very seldom a *main entrance* is clearly indicated (but a set of pages could, or, better, should [Musella, Padula, 1997], be accurately catalogued for a good visualisation of the resource) ; due to its transience, versions of the document (and those of its components) must be tracked in time. Consequently, what is important is the management of the locators of its dynamic components, the access to the relevant information, and not necessarily the production of an artifact which gives concreteness and meaning to the traditional concept of document. We do not believe the formal definition of a document model is essential. What we feel is important is an appealing description, a glimpse of which, in the document's entry pages, will invite the passage or navigation through it, that is invite the user to intervene in the path of the document life. Not everybody agrees and a great effort is being made to redefine precisely the concept of document for the new information society. L. Shamber, for example, proposes a document unit :

> "Consisting of dynamic, flexible, nonlinear content, represented as a set of linked information items, stored in one or more physical media or networked sites : created and used by one or more individuals in the facilitation of some process or project" [Shamber, 1996].

3. Is there any content emergence ?

Internet is more than the sum of its resources. This is a really vague proposition which must be examined in detail to motivate our emergentist standpoint.

Often, what is required to find a particular document is the availability of a quantity of information together with the possibility of reaching successful results in complex documental searches, which were traditionally performed in information centers by following a geographical path through those centers with a greater probability of satisfying the seeker's need of building a global document from specific sets of documents associated by common properties.

Today, due to the intense interaction of so many people, information is assuming a twofold organization, derived on the one hand from networked retrieval requests and designed therefore as distributed archives, folders, cybermaps, etc., and coming in the other from spontaneous human interaction and contributions which makes it uncontrolled, unexplored, unpredictable, disordered but suggestive of the more creative expectations and therefore calls for human cybernavigation.

The importance these issues and the complexity of the cyberlinks have assumed motivate the interest in emergentism, a discipline which has already been applied in technological fields, see for example [Edmonds *et al.,* 1994], and emphasizes the great complexity of the interaction among parts that try, but are insufficient to secure the property of the whole [Sober, 1991 ; Nagel, 1968 (1961)], holding that the Internet may be better understood as an information repository whose global and contextual properties cannot be deduced from the knowledge of its constituent resources and mechanisms but whose phenomenon of a successful contents appearance, finding and assembling according to a researcher's design could possibly be clarified. A content emergence is molded during an information intensive activity and its satisfactory final configuration is frozen into a document, a folder, or a cybermap, new entities which record and organize the knowledge. Then the emergence dissolves, it becomes an axiom, a new part of the repository, the object of new research and interactions, following an evolution independent of the emergence explanation.

We speak of content emergence because not all the documents present somewhere in cyberspace will surely emerge.

Two pressing questions will determine our future studies and clarify ideas : *what document or, more generally, what resource property should we consider ?* This concerns the cataloguing that we have discussed in the previous sections and shall examine more fully in its technical aspects in the following paragraphs. The second question is : *what level of complexity of documents, what inter-relationships and interactions in* *their evolution should we consider in order to reach a good understanding of the phenomenon ?* This is quite a simple matter when seeking documents that contain a specified keyword, or satisfy a Boolean expression : no relationship among documents, or interaction is involved. Matter changes when the documents concern a given topic, have to be collected and conceptually assembled into the global document that is the goal of the seeker, and have to be focused in a dynamic information ocean. An explanation considers the sequence of events and interactions which causes an emergence to be disclosed in that way : *what could we do to find useful documents ? Where could we find them ?* The human factor is here so relevant and manifest in the working method of the researcher. Access to large general repositories and message exchange are only the starting point. After that, he leaves tracks of his modifications, selects the documents which will possibly contribute to the emergence disclosure and assumes information that are used for choosing the actions for following his search path. Bookmarking the Web space and documenting each step and choice is useful for growing an explained emergence configuration.

Going into details : *which are the basic properties of each phenomenon, namely the reasons of the successful access to a document content ? Why that one rather than many others ?*

Appealing [Kahn, 1995 ; Borchers *et al.*, 1996] content externalisation, and the amount of hits, that is the success that the document met with other people, are among the factors that utterly influence the successful reaching of a document content. This is a more detailed level than the previous one and requires to take into account document editing and cataloguing, monitoring of the robots behaviour [Musella, Padula, 1997] and of the user accesses to the servers. Careful performance of these activities, is helpful in understanding what to offer and in which way, that is how the Internet infrastructure could possibly help in emphasizing and disclosing emergences and, therefore, in explaining them.

The two explanation levels suggest an understanding [Darley, 1994] of the complex system Internet which enables a planning, or in some sense a simulation, of the development of an emergence disclosure.

I understand the rules which govern the actions of every single agent and interaction in the system precisely, namely, the communication protocols, the data structures, the servers which constitute the interconnection level of the system, that is the more physical one.

I understand the rules and the arena in which they operate sufficiently well that I can make predictions of the outcome very rapidly from the initial state alone, without having to calculate every interaction. At this level, the interoperability level, methods, information/knowledge repositories, agents, description languages are involved to perform a planning of the research/navigation process which enables rapid and approximate calculation of the expectations with reference to a user model which specifies his needs and goals and his satisfaction scores.

There is a *continuum* between the two understanding levels which allow a more or less detailed planning which in any case remains very far from allowing a prediction of the concluding event : we can only have a feeling about the end of the research process. We could try to identify the minimum amount of understanding which could have a valuable influence in the research/navigation process and conceive a robot behaviour which mimes the process itself, but even an utterly deep and detailed knowledge of the system could not lead to prediction. We have no effective predictive methods but planning ones, despite a very analytic "perfect" understanding, due to the anthropocentrism of Internet : the human interaction and feedback are essential for the Internet system development and, contemporary, they are out of possible control.

4. Languages, methods and standards for document design

The languages for document editing allow, with great ease, to modify both the content and the layout of the document, to highlight parts of it, to link information regarding layout and references (notes, bibliography, index...) to subject matter.

These languages are many and varied : they range from Postscript© to PDF©, from the WORD© format to Rich Text Format, from TeX© to HTML.

A common characteristic of many of them is that they are Markup languages, and descriptive, not procedural Markups.

The difference between a procedural and descriptive Markup language is that the former defines the processing that must be applied in a particular point of the document : for example, it calls out the "Create the index" procedure to format the next paragraph with the parameters of the "Style index : classic", "Position : leading" and "Numbering : Roman numerals".

Instead, a descriptive Markup (tag) system uses tags to define the category of parts of the document : the tag "bibliography", for example, defines the part of the document identified as "this information is a bibliography".

It is this type of language which makes it possible to create textual structures that include both layout information and structural information (title, notes, author, references...), providing a tool that allows even the most unspecialized reader to single out, add to, or modify the layout, personal notes, highlights, or any other element in the structure of the document (**Figure 2**).

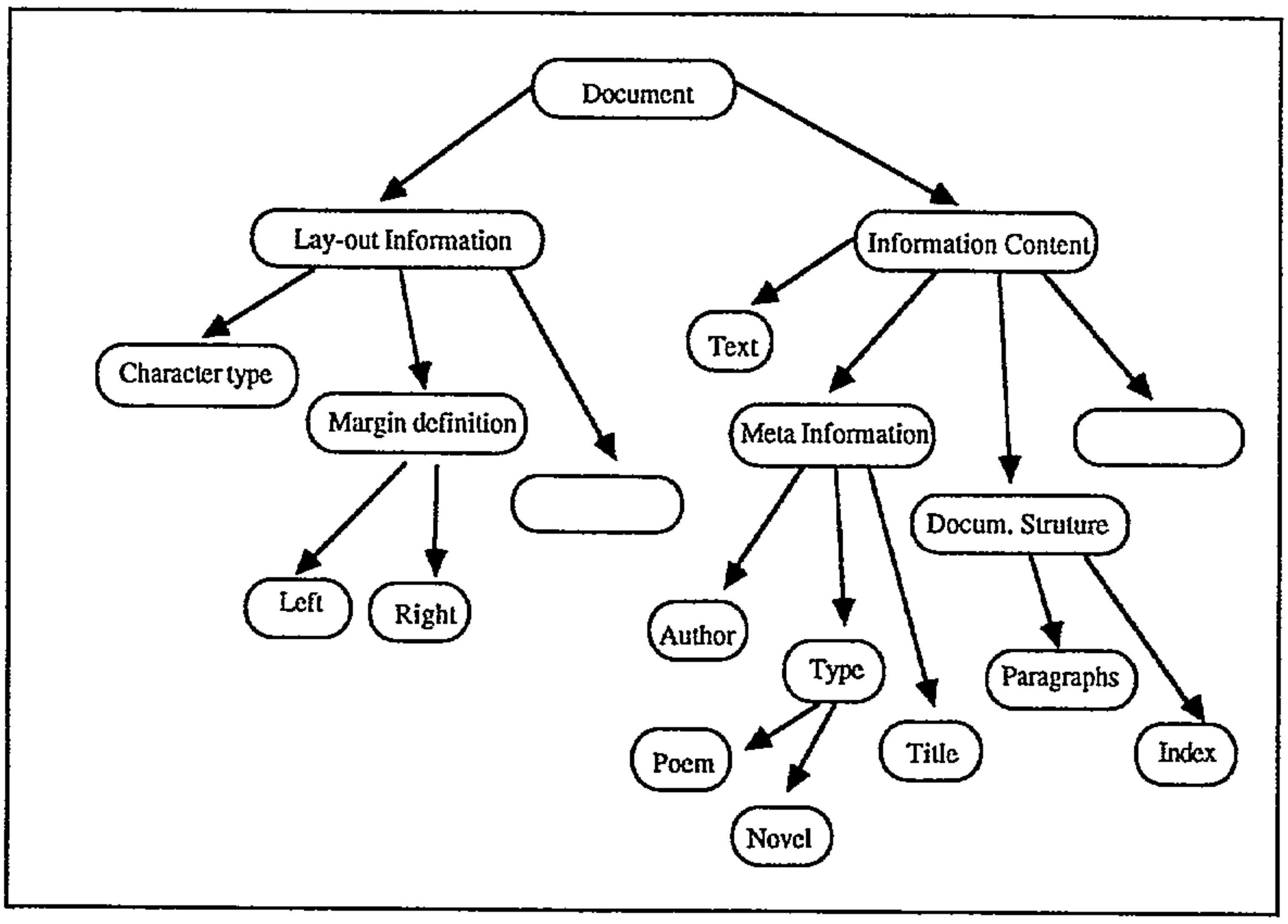

Figure 2

An example of document structure

If we examine the context of the Internet, we see that the standard format for the production of documents in the World Wide Web is HTML [Conolly, Ragget, 1996], which is a descriptive Markup language.

Like most tag languages, its grammar is defined by SGML (Standard Generalized Markup Language) [ISO, 1996 ; Sperberg-McQueen, Burnard, 1993], a language (also Markup) which is recognized internationally as the standard for the description of "marked-up electronic texts". The characteristics that distinguish SGML from other metalanguages are : its emphasis on descriptive rather than procedural markup, its *document type* concept, and its independence of any one system for representing the script in which a text is written.

Let us look at these three characteristics in detail.

SGML contemplates descriptive and procedural structures, but the procedural parts are clearly distinguished from the rest of the code, almost always placing them in different files. The emphasis in any case is on the descriptive characteristics, rather than on its procedural ones, underlining the fact that a text written with this philosophy can easily be processed by different software modules, which can thus apply different procedures to the different modules that compose the text.

Secondly, SGML introduces the notion of a *document type,* and, it follows, a *document type definition* (DTD). Documents are regarded as having types, just as other objects processed by computers do. The type of a document is formally defined by its constituent parts and their structure. A basic design goal of SGML is to ensure that documents encoded should be transportable from one hardware/software environment to another without loss of information.

A document, such as a book, or a manual, or a paper, has a well defined structure : as we have already seen in **Figure 2**, there is a part that defines the layout (at times mixed with the rest, at times separate), and one that specifies the information content. Putting aside considerations regarding the layout, let us see how an information substructure can be added to an existing document in HTML format, in particular, introducing meta information concerning the cataloguing of the document in question.

Since HTML is an instance of SGML, it is better to remain in the SGML context to be able to introduce specifications without modifying the content in an HTML file, but either by adding new specifications to the HTML, or by creating a new DTD to be used in the HTML (SGML contemplates the use of two or more DTD at the same time in the same document).

The first procedure is rather risky, as it entails modifying an internationally recognized standard (HTML) which is hardly ever varied to meet specific needs, unless the reprocessed documents are held to circulate only in a clearly defined circuit where all users will employ the same modified instruments to interpret the new specifications.

The second way is safer, but is based on the assumption that other persons/institutions are not likely to penetrate the same context, of classification in this case.

5. Languages for document cataloguing

Although library techniques have evolved greatly, especially in recent years, they have not been able to convert to the new technologies, such as the Internet, in a complete and efficient manner. We still do not have a single technology that makes it possible to describe information published in electronic form in a univocal and efficient way. This is due mainly to the fact that the community of developers of specifications for the Internet have underestimated the problem.

A number of researchers and organizations have tackled the matter, and have for some time been working on the development of possible solutions of various kinds.

The PICT, the Dublin Core, and the Warwick [Lagoze *et al.*, 1996] are only a few of the innumerable solutions recently proposed by organizations such as the W3C (http://www.w3c.com), or the OCLC (http://www.oclc.org).

But all these techniques come from the scientific environment of library technology, while their application involves general users who have no notion whatever of library science, although they are the authors of the most of the documentation published on the network. This creates a need for a cataloguing method that can be applied by the general public, and still satisfies a minimum of criteria for the correct indexing of documents. Hence the idea of defining a simple and clearly stated grammar of the HTML META tag [Musella, 1996] to use for the description of a HTML document. The realization of this grammar, which is very complex and involves several technical levels up to the protocol specifications, defines a limited group of words (eight in all) reserved only and exclusively to the cataloguing context. Let us see how the methodology is inserted in HTML and what the DTD of HTML looks like, in the specific case of the META tag :

```
<!ELEMENT META     - o    EMPTY        — Generic Metainformation —>
<!ATTLIST  META
           http-equiv NAME  #IMPLIED    — HTTP response header name —
           name       NAME  #IMPLIED    — metainformation name —
           content    CDATA #REQUIRED   — associated information —>
```

These specification remains valid without any technological updating being required to exploit them : to the definition of the DTD must simply be added this comment clarifying the specifications :

Properties	**Description**
Keywords	specifies the keywords describing the document content
Author	specifies the document's author/authors
Timestamp	specifies when the document has been authored in HTTP-date format
Expire	specifies the limit of (or unlimited) validity of the document content in HTTP-date format (or none)
Language	specifies the language in which the document is written : it is composed in the ISO 639 two-letter language code form, followed optionally by a period and a ISO3166 two-letter country code
Description	is associated with a short summary of the document content
Publisher	is the organization responsible for publishing the document.
Revision	is an ordinal number with two or three digits (00, 01, 02, or 000, 001...) specifying the document version

As can be seen, the extension simply involves a more formal re-definition of the attributes of a tag already present in the current version of HTML. This tag can easily be extended to embody more specific concepts by following the above schema, or by introducing a new definition of new tags and new attributes for HTML, or of a new DTD. We have opted for the former to define the methodology's tools, since our aim is to define a general purpose cataloguing system, but this leaves ample possibilities for extensions to satisfy more specific needs.

Other methods, such as the Dublin with its Warwick extension, or the IDML [IDML, 1996], are intended for trained users familiar with more specialized concepts, such as the heritability of attributes, or the concept of object.

Let us see what these methods offer more in detail :

> "The Dublin core is an attempt to formulate a simple yet usable set of metadata elements to describe the essential features of networked documents. The thirteen elements of the Dublin core include familiar descriptive data such as author, title and subject. In the design of the Dublin core consideration was given to mappings between the elements of the core and the existing, more specialized descriptive systems such as library cataloging. So some fields such as coverage and relationship are less typical of descriptive cataloging, and their utilization is reserved to trained cataloguers" [Dublin Core, 1996].

The Dublin core was conceived as a compromise between simple and technical cataloging. The result is a mixed code that satisfies neither librarians nor generic users. It is continuously evolving toward a more technical approach, rather than a simpler one. At present it lacks a complete description of the 13 attributes that would allow its correct use in the META context of a normal HTML page.

These attributes are : *Title, Subject, Author, Publisher, Other Agent, Date, Object type, Form, Identifier, Source, Language, Relation, Coverage.*

A container architecture called the Warwick Framework has been studied to provide a higher-level context for the Dublin Core :

> "This technology defines how the Core can be combined with other sets of metadata in a manner that addresses the individual integrity, distinct audiences, and separate realms of responsibility and management that characterize these distinct metadata sets".

The purpose of this architecture is :

— to allow the designers of individual metadata sets to focus on their specific requirements and work within their specific areas of expertise ;

— to allow the syntax of metadata sets to vary in conformity with semantic requirements, community practices and functional requirements ;

— to promote interoperability and extensibility by allowing tools and agents to selectively access, and manipulate individual packages and ignore others ;

— to accommodate future metadata sets not requiring changes to existing sets.

One of the characteristics of this architecture is that it can be inserted in most contexts because of the formats with which the documents are recorded. The Warwick Framework can be included in a specific document, or be connected to it by special links. Both these possibilities exist in the case of HTML documents. The Framework requires that its packages (part of the Warwick architecture) be strongly typed and defined. This is to permit the browser or agents to determine the type of datum contained in the description package ; a rigorous statement of these types would more closely define package recognition operations. The technique is very much like the procedure with which browsers use MIME types [Borenstein, Freed, 1993] to handle different types of data. It does cause some problems, still unsolved, regarding the codification of Warwick packages. But the true strength of the Warwick architecture lies in its characteristics of recursiveness and distributivity. However, this great strength poses serious problems of implementation, which may prove inefficient.

The IDML technology, instead, was designed to answer the need to classify not only texts, but the objects described in them as well. IDML is based on tables which are not easy to read and whose construction is not well accepted.

IDML introduces three main classes of attributed tags in the HTML document to improve its definition : ID-PUBLISHER defines the author ; ID-INFO, the contents of the site which is considered a single subtree ; and ID-PRODUCT, the products which are sold. A peculiarity of IDML with respect to other techniques is its reference to commercial aspects and the attribute inheritance through all the linked documents. This latter feature is conceptually very interesting, but it is not, unfortunately, supported by robots for indexing : the references to parent documents are not managed, and information inheritance is not controlled. The robots search for documents which are considered autonomous entities with no reference to the links they may have with other documents. Consequently, it is still impossible to formulate an operational definition of document that collocates it within an inheritance hierarchy.

We could describe many other methods, each with its own peculiarities, but all of them with the same singularity of being neither widely used nor accepted as standard, because they are not considered suitable, and, especially, because both the large firms that stand as beacons in the world of the Internet and the great centers of research that propose standards display a general lack of interest in this type of problem.

Moreover, the efforts to improve these methods and ongoing studies to develop technologies that can satisfy the greatest number of people are all to no effect if the authors and editors of networked material do not employ these methods of description in their documents.

One way of getting around this problem could be the further simplification of the methodologies to make them easier to use and manage. Another way, which would probably bring even better results, would be to build servers that check the existence of this information locally, blocking the distribution of documents presented without it. However, this method leads to other problems concerning the concept of provider censorship, which we shall not discuss here.

In the end, the need for a descriptive system remains, and its standard introduction is not foreseen for the immediate future.

(+*Istituto per le Tecnologie Informatiche Multimediali — CNR, Milan, Italy*
davide@jargo.itim.mi.cnr.it
tel. +39 (0)2 70643262)

(°*Istituto per le Tecnologie Informatiche Multimediali — CNR, Milan, Italy*
padulam@acm.org
tel. +39 (0)2 70643271)

References

[Altavista]
"Altavista", *Digital computer*, http://www.altavista.digital.com/

[AOL robot]
"American On Line Robot", *WebCrawler*, http://www.webcrawler.com/

BERLEANT (D.), BERGHEL (H.)
1994a, "The Challenge of Customizing Cyberspace", *The Journal of Knowledge Engineering & Technology*, vol. 7, n°2, p. 33-43.

1994b, "Customizing Information : Part 1, Getting What We Need, When We Need It", *IEEE Computer*, vol. 27, n°9, p. 96-98.

1994c, "Customizing Information : Part 2, How Successful Are We so Far ?", *IEEE Computer*, vol. 27, n°10, p. 76-78.

BIANCHI (N.), MUSSIO (P.), PADULA (M.) *et al.*
1996, "Multimedia Document Management : An Anthropocentric Approach", *Information Processing & Management*, vol. 32, n°3, p. 287-304.

BORCHERS (J.), DEUSSEN (O.), KLINGERT (A.) *et al.*
1996, "Layout Rules for Graphical Web Documents", *Computers & Graphics*, vol. 20, n°3, p. 415-426.

BORENSTEIN (N.), FREED (N.)
September 1993, "MIME (Multipurpose Internet Mail Extension), Part 1 : Mechanisms for Specifying and Describing the Format of Internet Message Bodies", *rfc1521*, http://ds.internic.net/rfc/rfc1521.txt

CARMEL (E.), WITHAKER (R. D.), GEORGE (J. F.)
1993, "PD and Joint Application Design : A Transatlantic Comparison", *CACM*, vol. 36, n°4, p. 40-48.

CERF (W. G.) *et al.*
1993, *National Collaboratories : Applying Information Technologies for Scientific Research*, National Academy Press, Washington D. C.

CLEMENT (A. P.), BESSELAAR (Van den)
1993, "A Retrospective Look at PD Projects", *CACM*, vol. 36, n°4, p. 29-37.

CONOLLY (D.), RAGGET (D.)
May 1996, "Introducing HTML 3.2",
http://www.w3.org/pub/WWW/MarkUp/Wilbur/

DARLEY (V.)
1994, "Emergent Phenomena and Complexity",*http://www.das.harvard.edu/users/students/Vincent_Darley/emergence_alife/emergence_alife.html*

[Dublin Core, 1996]
June 1996, *http://purl.org/metadata/dublin_core_elements/*

EDMONDS (E. A.), CANDY (L.), JONES (R.) *et al.*
1994, "Support for Collaborative Design", *CACM*, vol. 37, n°4, p. 41-47.

EHN (P.)
1988, *Work-Oriented Design of Computer Artifacts,* Pelle Ehn & Arbetslivscentrum.

GIBSON (W.)
1984, *Neuromancer,* New York, Ace Books.

GUDIVADA (V. N.)
1995, "Multimedia Systems : An Interdisciplinary Perspective", *ACM Computing Surveys,* vol. 27, n°4,december, p. 545-548

[IDML, 1996]
1996, "IDML", *http://www.identify.com/welcome/intro.html.*

[ISO, 1996]
1986, *Information Processing. Text and Office Systems. Standard Generalized Markup Language (SGML),* Geneva, International Organisation for Standardization (ISO).

JANES (J. W.), ROSENFELD (L. B.)
1996, "Networked Information Retrieval and Organization : Issues and Questions", *Journal of the ASIS,* vol. 47, n°9, p. 711-715.

KAHN (P.),
1995, "Visual Cues for Local and Global Coherence in the WWW", *CACM,* vol. 38, n°8, p. 67-69.

KOUZES (R. T.), MYERS (J. D.), WULF (W. A.)
1996 ,"Collaboratories : Doing Science on the Internet", *IEEE Computer,* August, p. 41-46.

LAGOZE (C.), LYNCH (C. A.), DANILE (R. jr.)
July 1996, "The Warwick Framework : A Container Architecture for Aggregating Sets of Metadata", *http://www.dlib.org/dlib/july96/lagoze/07lagoze.html*

LÉVY (P.)
1994, *L'Intelligence collective,* Paris, La Découverte.

LEVY (D. M.), MARSHALL (C. C.)
1995, "Going Digital : A Look at Assumptions Underlying Digital Libraries", *CACM,* vol. 38, n°4, p. 77-84.

MUSELLA (D.)
December 1996, "The META Tag of HTML", *draft-musella-html-metatag-03,*
http://jargo.itim.mi.cnr.it/documentazione/draft-musella-html-metatag-03.txt

MUSELLA (D.), PADULA (M.)
1997, "Step by Step Toward the Global Internet Library", *IEEE Communication Magazine,* vol. 35, n°5, p. 64-70.

NAGEL (E.)
1968 (1961), *The Structure of Science,* Indianapolis, Ind., Hackett Publishing Co., Italian translation, Milano, Feltrinelli.

[Northern Webs, 1996]
September 1996, *http://www.digital-cafe.com/~webmaster/set01.html*

SHAMBER (L.)
1996 , "What is a Document ? Rethinking the Concept in Uneasy Times", *Journal of the ASIS,* vol. 47, n°9, p. 669-671.

SOBER (E.)
1991, "Emergence", *in Handbook of Metaphysics and Ontology,* H.Burkhardt, B. Smith, eds., Vol.1, Philosophia Verlag.

SPERBERG-McQUEEN (C. M.), BURNARD (L.)
May 28, 1993, "A Gentle Introduction to SGML", *Draft version2,* http://www.uic.edu/orgs/tei/sgml/teip3sg/

TOMAIUOLO (N. G.), PACKER (J. G.)
1996, "An Analysis of Internet Search Engines : Assessment of Over 200 Search Queries", *Computers in libraries,* vol. 16, n°6, June.

WIEDERHOLD (G.)
1995, "Value and Productivity", *Digital Libraries, CACM,* vol. 38, n°4, p. 85-96.

ZEITOUN (J.)
1993, "Sur certains aspects du temps en synthèse numérique d'images", *Sémiotiques,* n°5, décembre, p.159-168.