

# Porphyry 2001: Semantics for scholarly publications retrieval

Aurélien Bénel <sup>†‡</sup>, Sylvie Calabretto <sup>†</sup>, Andréa Iacovella <sup>‡</sup>, Jean-Marie Pinon <sup>†</sup>

(<sup>†</sup>) LISI – INSA Lyon  
Bâtiment Blaise Pascal, 69621 Villeurbanne CEDEX, France  
Firstname.Surname@lisi.insa-lyon.fr

(<sup>‡</sup>) French School of Archaeology (EFA)  
6 Didotou street, 10680 Athens, Greece  
Firstname.Surname@efa.gr

**Abstract.** We describe the design and algorithms of *Porphyry 2001*, a scholarly publication retrieval system. This system is intended to meet library user studies which advocate human interpretation and social interactions. The metaphors we used are annotations and publication (making public). We first discuss about different philosophical approaches to semantics and choose the more suited to scholarly work: the one considering a transitory, hypothetical and polemical knowledge construction. Then we propose an overview of *Porphyry 2001*: an hypertext system based on a dynamic structure of user annotations. The visualization and evolution of the structure (a dynamic directed acyclic graph) is made more efficient by the use of an *ad hoc* browsing algorithm.

**Keywords.** Patron-augmented digital libraries, user interfaces and visualization systems, semantic nets, browsing/reading/annotating.

## Introduction

Our study is related to a digitalization project by the French school of archaeology in Athens<sup>1</sup>. This project aims at giving online access to one of its main periodical publications: “*La Chronique des fouilles*”, an archeological excavations and findings yearly survey. This corpus has been chosen since although its size is reasonable (about 12,000 pages), it is nearly exhaustive in regards to the past 80 years of archaeological activity in Greece and Cyprus. Besides, the “*Chronique*” is daily read in libraries throughout the world.

We must stress that the information retrieval problem is not new concerning the “*Chronique*”. Publishers have tried for 80 years to make it easier to consult. It is made

---

<sup>1</sup> Ecole française d’Athènes (<http://www.efa.gr>)

of small independent parts (about 50,000) which are hierarchically structured and indexed according to artifact location, dating and typology. Archaeologists who have tried retrieval systems based on automatic indexing or manual indexing using thesauri are satisfied by none of them. The former is said to be inadequate because it deals with out-of-context terms. The latter is considered to be hard to manage over time by the indexing team since it needs periodic updates of both thesauri and indexes in order to reflect science progress. As a first example, there has been several years ago a polemic between two archaeologists about determining in ambiguous cases whether the border of a mosaic was black or white. An automatic indexing system would have extracted only the point of view in the text. Moreover, without knowing whom point of view it is, the index could not have been interpreted. As a second example, when the Copper Age has been inserted in the chronology between the Neolithic Period and the Bronze Age, a thesaurus-based system would have needed plenty of documents to be reinterpreted and re-indexed. Therefore, we had to study the most theoretical aspects of information retrieval (even philosophical aspects), to find an alternative for our system.

## **From semantics theories to workstations**

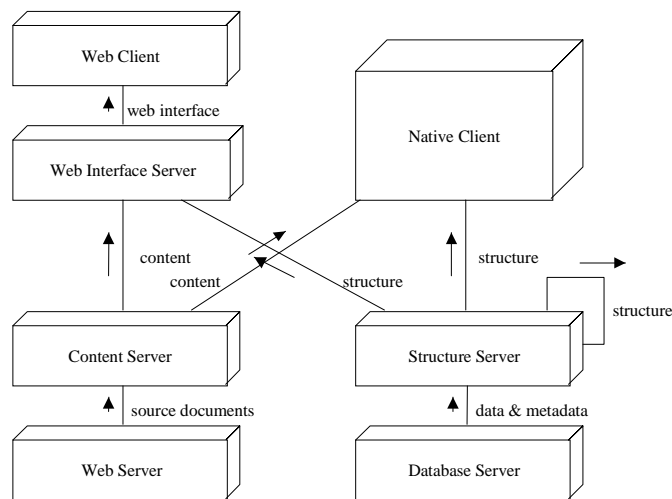
Information retrieval (IR) as defined by Cornelis J. van Rijsbergen [19] aims at matching relevant documents with user information needs. In order to be “computed”, this matching has to be transmuted from the *content* space into the *form* space. Since computers cannot match the *meaning* of the information needs with the meaning of the documents, IR techniques tend to translate information needs into *formal* queries and documents into *formal* descriptions (also called “logical views” [1]). So one of the challenges in IR should be to minimize the gap due to this translation: the gap between *signifiers* and *signified*. This question is a central one in linguistic semantics.

### **Linguistic theories of semantics**

As stated by the French linguist François Rastier [13], there have been, among the various theories of semantics, mainly two streams: the first one (widely spread) from the logic community, the second one (nearly unknown) from the hermeneutic community. Whereas the former focuses on *representation*, the latter focuses on *communication*. Whereas, in the former, properties of a sign occurrence are inferred from relations between types (see “type hierarchies” in John F. Sowa [15]), in the latter, there are only “source occurrences” and “revisited occurrences”. Whereas the former makes the Aristotelian assumption of an *ontology* (from the Greek word “ontos” for “being”), the latter considers a transitory and hypothetical knowledge construction.

Since the system we want to design is for scientists, we will adopt the hermeneutical view of semantics (see our prior work [3]). Indeed, this approach is more adapted to modern science by highlighting the constructivist nature of scientific knowledge (see Karl R. Popper [12] and Thomas S. Kuhn [10]).

It is worth noting that the need for both interaction and communication has been highlighted in experimental studies about traditional library users.



**Fig. 1.** Corpus consulting through Porphyry multi-tiers architecture.

Kenton O'Hara *et al.* [8] studied the document-related research activities of PhD students during a working day. The induced model characterized the work carried out by university library users as going beyond the traditional searching and retrieving of information. In that way, note making, photocopying, bibliographic searching, reading, annotating, information reviewing, and documents writing should be considered as a whole.

In a different setting, Andreas Paepcke [11] interviewed engineers involved for example in customer support or in design in order to learn about their information needs and habits. He concluded that even if retrieving information is central, it is interwoven with discovering, managing, interpreting and sharing information and that all of these activities need the communication between humans.

But oddly enough, very few digital library systems at this time support social interactions [17] and patron-augmentation [7] (see [14] also).

Because science carries more than interactions among individuals, systems should go one step further by supporting groups. Scientific groups (from working groups to colloquiums) are important for knowledge construction in the process of becoming more objective.

Knowing the knowledge of each individual, the question becomes: "What is the knowledge of the group?" Expressed in a different way: "How do we get a syntactic and semantic coherence from these different (and even contradictory) models?" If these questions are opened for Knowledge Managing in general, they have got an answer for years in the scientific praxis: publication.

In the traditional publication process, the "publishers" (real publishers, reviewers, committees...) check submitted papers regarding form and contents in order to ensure its validity for the group. Then the authority given to the publishers is transferred to the papers themselves.

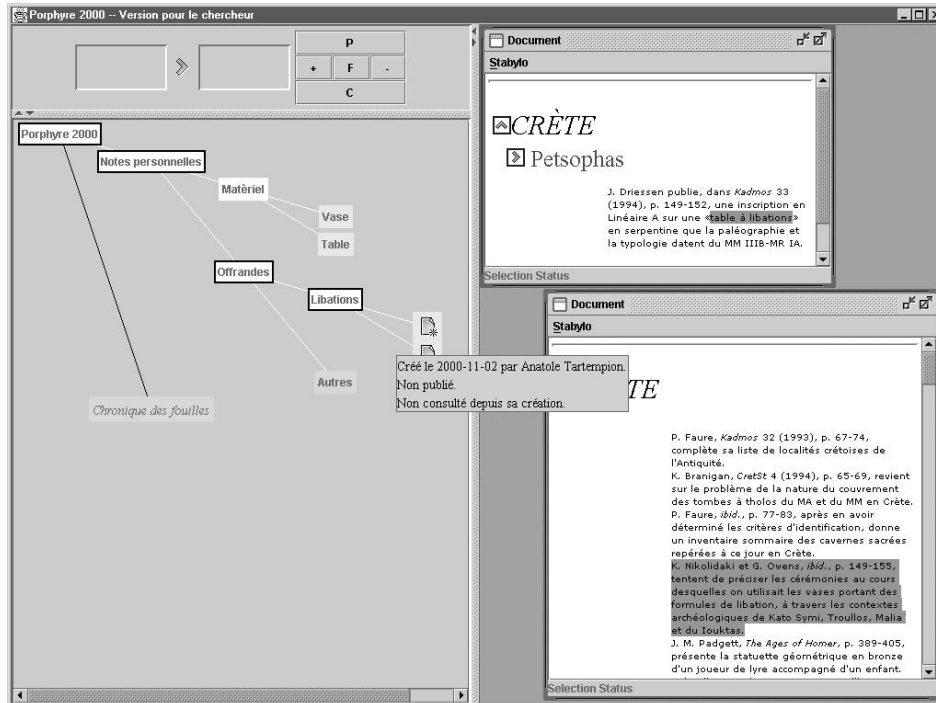


Fig. 2. Porphyry client screenshot.

As a result we propose that in our system scientists can choose to join groups headed by “publishers” they consider as authorities and that their knowledge representations can be “published” through a reviewing process just like in the physical world.

### Porphyry 2001 overview<sup>2</sup>

*Porphyry 2001* is a client-server system aiming at creating, retrieving and sharing documents and annotations. Its architecture (see Figure 1) is grounded on the distinction between content and structure.

The *content server* is a classic web server with an *ad hoc* “servlet”. Given extraction parameters, it is able to deliver fragments from web accessible documents (only plain-text and JPEG images for now).

The *structure server* use a database server to store and retrieve data and meta-data. Both are handled in the same way: they are filtered by the *structure server* and formalized through the same directed acyclic graph model. In this model, if, for two descriptors  $D_1$  and  $D_2$ ,  $D_1 \rightarrow D_2$ , then any document described by  $D_2$  is described by  $D_1$  too. It is worth mentioning that only edges and nodes have significance for the

<sup>2</sup> The Porphyry Project Page: <http://lisi.insa-lyon.fr/projets/descrippr27.htm>

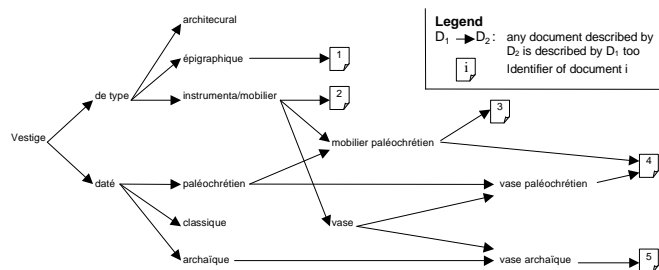


Figure 3 – Sample index structure.

system. But, so that users can interpret the graph, we store labels too. Node labels contain short descriptions and edge labels contain information (see the edge popup label in Figure 2 at the center) about their creation (user, date) and publication (group, date). As long as the formal signification of this framework is kept, users are free to use it in order to represent (see Figure 3): specialization, composition, attributes, attribute values, relations.... The graph being accessed by a user can be split into different sub-graphs depending on their ownership and performance considerations.

Context and structure are combined either by the *native client* (see Figure 2) or by the *web interface server* (so that a classic web client can access it). Although the web interface is dedicated on browsing, the *native client* allow the researcher to upload new documents, to define new fragments, and to modify also his/her own corpus structure.

## Scenario of user interactions

In this section, we will trace step by step an example of computer-human interactions involved in document retrieval. Our schema (Figure 4) will show both the annotation graph as displayed by *Porphyry 2001* and the user actions. As shown in the Figure 4, let us navigate in the Figure 3 indexing graph...

- **Step #1.** The global corpus deals with “*vestige typé*”. More specialized corpora exist dealing with “*daté*” or “*épigraphique*” or “*instrumenta/mobilier*” but not with “*architectural*” (since this descriptor corresponds to no document). When the user selects “*instrumenta/mobilier*”, the system jumps to step #2.
- **Step #2.** The selected corpus deals with “*vestige de type instrumenta/mobilier*”. This describes exactly the document which has the identifier “2”. More specialized corpora deal with “*mobilier paléochrétien*” or “*vase*” but neither with “*architectural*” nor with “*épigraphique*”. When the user selects “*vase*”, the system jumps to step #3.
- **Step #3.** The selected corpus deals with both “*vestige de type vase (instrumenta/mobilier)*” and “*vestige daté*”. Please note that “*daté*” is automatically inferred (since all documents dealing with “*vase*” deals also with “*daté*”). More specialized corpora deal with “*mobilier paléochrétien*”, “*vase paléochrétien*”, “*vase archaïque*”, “*paléochrétien*” or “*archaïque*” but not with “*architectural*”, “*épigraphique*”, “*classique*” or “2”. When the user selects “*archaïque*”, the system jumps to step #4.

- **Step #4.** The selected corpus deals with “*vestige de type vase (instrumenta/mobilier) daté de l’époque archaïque*”. This corpus contains only one document, the one with “5” as its identifier.

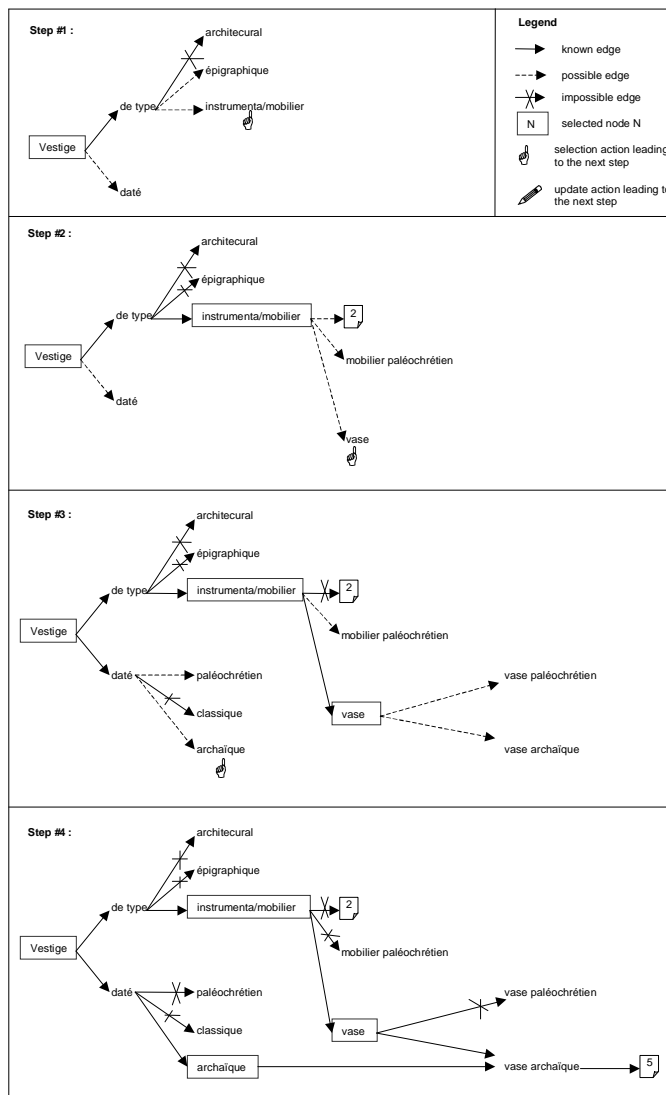


Figure 4 – Retrieval scenario (see index structure in Figure 3).

## Algorithms

In the two preceding scenari, the annotation graph was filtered. As a matter of fact, most of interactive information retrieval systems (see Marti Hearst [9]) focus on reducing cognitive load by filtering information. In our system, the filtering algorithm is a valuable help in navigating through corpora. Since searching and indexing are both corpora discriminations, our filter can be seen as an assistant for both refining a query and reusing descriptors for a new indexing. We will now explain our algorithm in detail.

Gerard Salton in the late 60's [15] defined a set-theoretical model of information retrieval. It deals with a set of "descriptors" and a set of documents. In that way, we can draw the corpus inclusion (see Figure 5) and the request conjunction graphs (see Figure 6). Then, from the mapping of documents with descriptors (see Table 1), we can deduce the mapping of requests with documents corpora (see Table 2). From that point, we can figure out that several corpora can't be obtained by any request (e.g.  $\{B,C\}$ ) and that the same corpus can be obtained by different requests (e.g. request  $a$  AND  $b$  with request  $b$ ). Although these results are widely known, they have been, as far as we know, rarely used as interaction media.

**Table 1.** Sample mapping of documents A, B, C, D with descriptors a, b, c.

		Descriptors		
		a	b	c
Documents	A	X		X
	B	X	X	
	C			X
	D	X		

**Table 2.** Mapping of requests with documents corpora (computed from Table 1).

TRUE	{A,B,C,D}
a	{A,B,D}
b	{B}
c	{A,C}
a AND b	{B}
b AND c	$\emptyset$
c AND a	{A}
a AND b AND c	$\emptyset$

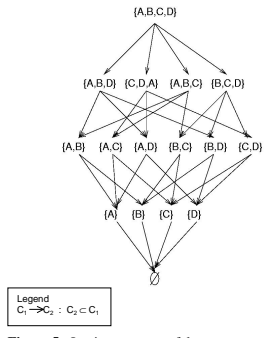


Figure 5 - Lattice structure of documents space.

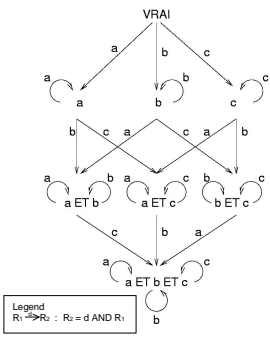


Figure 6 - Lattice structure of descriptors space.

Claudio Carpineto *et al.* [6] used Boolean logic results by removing from the corpus inclusion graph every inaccessible corpora in order to get a static generalization/specialization diagram of documents classes (see Figure 7).

In our approach (see our prior works [2] for more details) we preferred to join together, in the requests graph, requests which describe the same corpus. By doing so, we get a state-chart diagram (see Figure 8) in which states correspond to corpora and transitions correspond to elementary requests. These one-descriptor-requests on transient corpora can be seen as the addition of a descriptor to the global request: a kind of query refinement.

Owing to the preceding state-chart, in a given state (result of the selection of a set of descriptors) any refinement can be said:

- *Impossible*: if it leads from the current corpus to the empty corpus (e.g. descriptor *c* in state *{B}* see Figure 8),
- *Known*: if it leads from and to the current corpus (e.g. descriptor *a* in state *{B}*, descriptor *b* in state *{B}* see Figure 8),

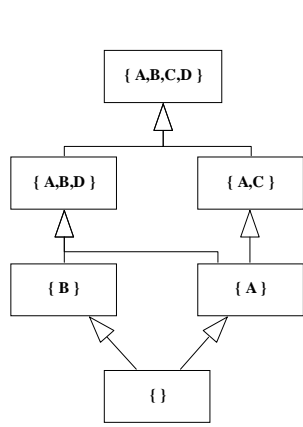


Fig. 7. Class diagram [18] derived from documents space.

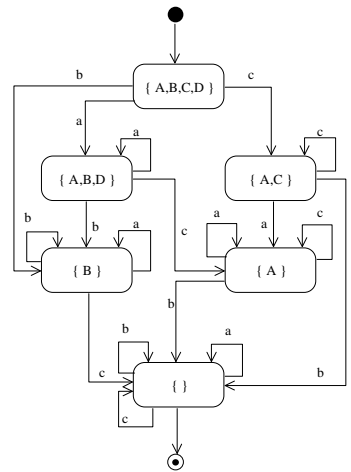


Fig. 8. State-chart diagram [18] derived from descriptors space.

- *Possible*: otherwise.

The filter consists in showing only known descriptors and their “children” and in assigning its corresponding state to any showed descriptor. The scenari shown in Figure 4 illustrates the use of this filter on a simple example. We can also have a look at real size examples in Figure 1.

## Discussion

At this point, a few aspects should be discussed about *Porphyry 2001* system.

### Starting from scratch

The digital library system we have presented is based on patron-augmentation. But, for such an evolutionary approach, the question is: “Evolution? From what?”. Can we give an “empty box” to users? If document retrieving is based on annotations, how could the first annotator of a document retrieve it?

As a matter of fact, some information can be automatically loaded in the descriptor graph:

- The title hierarchy of the semi-structured document (since this structure *describes* each section),
- Manual (or intellectual?) controlled indexes stored in document appendices,
- Automatically extracted key-words or key-phrases (structured for example regarding alphabetical order or formal clusters).

One should note that there are neither thesauri, nor “ontologies”, nor even concepts which are approved by the whole archaeological community. Therefore we cannot reuse them as a bootstrap for our sytem. But, on the contrary, the collaborative use of *Porphyry 2001* system could lead archaeologists to normalize such definitions. From an archeological point of view, this is one of the main challenge of using the system.

### Combinatory explosion

Another important question deals with the algorithmic complexity of our graph filter. On the one hand the system must give results in less than a few seconds (in order to remain usable in an interactive process). On the other hand it is difficult to evaluate the theoretic complexity of the algorithm since there are a very few constraints on the partial order structure.

To find a practical answer, we must consider the use of the system. As we saw, the graph which is browsed by a given user, is made of its personal graph and the graphs from the groups he/she has registered.

Firstly, these graphs are connected only by the “root” descriptor and by identifiers. They are independent indexing dimensions. That is what information sciences call “*facets*”. Because of their independence, the complexity of  $n$  groups is only  $n$  times the complexity of  $1$  group and, moreover, the algorithm can be run in parallel on  $n$  servers (one per group) so that the computing time for  $n$  graphs is nearly the same as for  $1$  graph.

Secondly, the nature of personal graphs and group graphs are quite different. The latter are rather bigger than the former and is updated much less often (only during the *publication* process). So it is interesting to make pre-computation of the group graphs. In fact the involved algorithms consist in recursively deducing relations and in doing basic set operations. We chose to do in advance recursive computations only and to do “on the fly” set operations (database management systems are good for it). It seems to be a good compromise between mass memory use and response time.

The two proposed optimizations (multiple servers and pre-computation) have been implemented and used. We plan for the next months to test them with huge real data.

### **Evaluating interactive information retrieval**

Let us study a few epistemological aspects. Since science is based on *falsification* (see Karl R. Popper [12]), a theory must be *testable* to be said “scientific”. A test is an experiment that can make the theory getting false (by deduction *modus tollens*). A test result is *particular* but not *singular*: it must be obtained and obtained again at anytime by anybody at anyplace. A succeeding test result is a result that “breaks” the theory. Moreover, since methods are based on the domain paradigms (see Thomas S. Kuhn [10]), testing protocols must be validated by the scientific community.

We would like to stress a few points. First, the scientist must *propose* a test but shouldn’t *lead* the test. The *community* should test it. A testable theory is “objective” and doesn’t need anymore the subject who have invented it. By the way, it is psychologically difficult for a human to break his/her own work... On the contrary, it is so great to break somebody else’s work! Second, because we work on *interactive* information retrieval and so with human individuals, it seems to be difficult to get *particular* results. Does it make sense to compare the activity of two users, especially when each of them is the world expert in his/her domain? Does it make sense to compare activities of a same user with two interactive systems (knowing that he/she may have learned “something” during the first activity)?

For us, it would be important if this kind of methodological aspects were discussed by the IR community. If these points were clearer, then we could ask the community to validate or invalidate the following protocol.

Our protocol (its setup is in progress) relies on the “reality” of the test: real users doing their own activities with real information and for a long time. Firstly, we propose to compare two interfaces to access the same data (250,000 records of archeological photographs descriptions): the classical QBE interface (query by example) and ours. Secondly, we propose to log over time the growing of the descriptors graph (for “*La Chronique des fouilles*”) in order to know if the evolution we hope for is real or not.

### **Conclusion**

The system we have presented proposes a framework for free descriptors created either by machines (words or phrases occurrences...) or by human (categories, annotations...). Most of all, owing to collaboration and dynamics, it can be used as a

debate media. Every annotation is dated and authored, so that it can be interpreted, contradicted by another annotation, or considered as obsolete.

We aim at giving the user a system just like the “memex” Vannevar Bush [5] dreamt of. The reader could retrieve his/her former mind “trails” and others’ ones (colleagues, tutors, librarians...). The automatic system would be there to assist the reader in his task by ridding him of repetitive aspects of his/her activity, so that he/she could focus on creative and intuitive aspects of his/her work.

## References

- [1] Baeza-Yates, R., and Ribeiro-Neto, B. *Modern Information Retrieval*, ACM Press and Addison Wesley, 1999.
- [2] Bénéel, A., Calabretto, S., Pinon, J.-M., and Iacovella, A. Vers un outil documentaire unifié pour les chercheurs en archéologie. In: Actes du XVIIIe congrès INFORSID, 2000. pp.133-145. In French.
- [3] Bénéel, A., Egyed-Zsigmond, E., Prié Y., Calabretto, S., Mille, A., Iacovella, A., Pinon, J.-M. Truth in the digital library: from ontological to hermeneutical systems. In: Proceedings of the fifth European Conference on Research and Advanced Technology for Digital Libraries, LNCS #2163, Springer-Verlag, 2001. pp.366-377.
- [4] Berleant, D. Models for reader interaction systems. In: Proceedings of the Ninth ACM Conference on Information and Knowledge Management, 2000.
- [5] Bush, V. As we may think. In: *The Atlantic Monthly*. July 1945.
- [6] Carpineto, C., and Romano, G. Dynamically bounding browsable retrieval spaces: an application to Galois lattices. In: RIAO'94 conference proceedings, “Intelligent Multimedia Information Retrieval Systems and Management”, 1994.
- [7] Goh, D., Leggett, J. Patron-augmented digital libraries. In: Proceedings of the Fifth ACM Conference on Digital Libraries, 2000.
- [8] O'Hara, K., Smith, F., Newman, W., and Sellen, A. Student readers' use of library documents: implications for library technologies. In: ACM Conference Proceedings on Human Factors in Computing Systems, 1998.
- [9] Hearst, M. User interfaces and visualization. In: [1].
- [10] Kühn, T.S. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.
- [11] Paepcke, A. Digital libraries: Searching is not Enough. What we learned on-site. In: *D-Lib Magazine*. May 1996.
- [12] Popper, K.R. *Objective Knowledge: an Evolutionary Approach*. Clarendon Press, 1972.
- [13] Rastier, F. Sens et signification. In: Jacob, A. *Encyclopédie philosophique universelle*, Presses Universitaires de France, 1999. In French.
- [14] Röscheisen, M., Mogensen, C., and Winograd, T. Beyond browsing: shared comments, soaps, trails and on-line communities. In the Third International World Wide Web Conference, “Technology, Tools and Applications”, 1995.
- [15] Salton, G. *Automatic Information Organization and Retrieval*. Chapter: “Retrieval models”. Computer Sciences series, McGraw-Hill Inc., 1968.
- [16] Sowa J.F. Semantic networks. In: Shapiro, S.C. *Encyclopedia of Artificial Intelligence*, Wiley, New York, 1992.
- [17] Tochtermann, K. A first step toward communication in virtual libraries. In the Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries, 1994.
- [18] *UML Notation Guide*. OMG, 1997.
- [19] van Rijsbergen, C.J. A new theoretical framework for information retrieval. In: Proceedings of 1986 ACM Conference on Research and Development in Information Retrieval, 1986.