

# Computer-Processable Information for the Scientists

Serge Abiteboul  
INRIA-Futurs – LRI, University Paris 11

June 2005

## Abstract

A main challenge is to make information available and processable. At a personal or company level, this means leveraging a number of information resources: emails, letters, databases, ldap, reports, contracts, web pages (private or public), programs, images, etc. At a universal level, this is something like the Berlin declaration of open access to knowledge in science and humanities [2]. In this paper, the emphasis is on computer-processable information for the scientists.

## 1 The context

We are concerned here with computing resources that provide scientific information. One form of information is documents, e.g., scientific articles. Other forms are data and knowledge. Data is, for instance, a table describing some measures in an experiment and knowledge the description of the meaning (the semantics) of the table typically based on some terminology (taxonomy). These terms are imprecise and we will not try in this paper to make them more precise. On the other hand, let us try to make more precise what we mean by *information resources*. Today they consist primarily in documents in various formats, typically PDF; on the Web, HTML. Of course such documents will remain an important part of global resources (e.g., scientific publications). But beyond, we view data and knowledge bases available on the network as possible resources. On the Web, these are generally accessed via forms. More generally, any computing program that generates data may be viewed as an information resource. On the Web, these are called Web services.

**More data** With respect to information management, the world is changing. Storage keeps being cheaper and larger (both for disk and RAM). Scientific applications seem to always be eager to attack problems requiring more and more data. The advent of cheap micro-sensor technology will probably also accelerate this explosion of data volume.

**Internet** Storage that used to be a local resource tends to become more and more provided by the network. Indeed, with the Internet, the data used by applications is switching to a world scale. Also, intensive computing power that used to be performed on mainframes tend to move to the grid. All this is placing a lot of stress on communications when network bandwidths are not progressing so rapidly.

**Warning: 2020** In Computer Science, situation changes so rapidly that forecasting is a dangerous science. In spite of grandiose programs, challenges have remained open for decades, and still are to some extent, e.g., automatic translation. On the other hand, the World Wide Web developed at a speed that surprised everyone and lead to totally reshaping road maps.

**Warning: Science** Our topic of research is distributed management of data (and to some extent knowledge) in particular for the Web. Our competence on scientific applications is limited. So, we would like to apologize in advance for possibly naive or erroneous statements.

## 2 The problem

Computing is an enormous accelerator to science. In particular, computer-supported information management has become an essential part of most scientists life. Unfortunately, one sees them more and more confronted to absurd data management tasks and loosing their time doing so. One can admit that if the tools they are offered are more and more sophisticate, they are not adapted to the complexity of their needs. For instance, scientists are sometimes reluctant to use database technology because of the lack of flexibility of such systems in front of their ever-changing needs. Also, they are confronted to the chasm between data and computing. Sometimes the problems are simply economics: better tools exist that are too expensive for their budgets. Sometimes they are cultural: they don't know the tools and prefer to re-implement the wheel; or educational: they don't have the computer-science background to use them properly.

To take fully advantage of the potential of progress in science brought by computing, the information management tools have to be better adapted to the needs of the scientists. For economical reasons, we have to do so in a generic manner: we cannot afford to develop distinct tools for each particular science.

Today, we already have a huge (and rapidly growing) collection of information globally available, namely the Web. Typically, one types in a list of words and receives as answer, pointers to a collection of documents. Besides full-text searching, one can obtain via the Web access to more and more resources, typically databases accessed via forms. One would like to automate the use of such information and view the network, the entire Web, as a computer-supported extension of the scientist mind with information (documents, data and knowledge) and computing resources. This means that the scientist may express queries with a much richer structure than a list of words or a form and a richer semantics as well. The other facet is that the result also has richer structure and should in particular be usable by machines.

To see an example, suppose a scientist is interested in the bacteria that are developing in milk under certain conditions. She should be able to express her needs without being a computer expert. The evaluation of the query may involve a number of tasks. First, retrieving knowledge from some bases, e.g., the bacteria that may develop in milk. Next, understanding which resources may help answer how they develop in milk. Then, obtaining data collected from experiments, from a number of databases. Finally, some computing, e.g., running some simulation, performing interpolation or ranking answers.

For a computer to be able to fully support such functionalities, the system has to understand the content of all the available resources. A great way to do so, is to have knowledge attached to resources. So, we have to deal with knowledge and not simply documents and data.

One of the big success stories of computer science is data management, that has invaded our life with tools such as relational databases. The field benefits from beautiful formal background, mathematical logic, and strong computer science technologies such as indexing, algebraic query optimization or concurrency control. When moving to knowledge, the situation is much less clear: we have to deal with imprecision and complex relationships (in taxonomies), uncertainty, imprecision and incompleteness (notably arising from measures), contradiction and beliefs, quality and trust, etc. Each single of these facets requires non trivially reshaping the mathematical foundations and the computer processing.

When moving to a global management of knowledge, other aspects also complicate the task:

**Volume** In some cases, the volume of data is huge (e.g., biological science) and very distributed (e.g., environmental science). Even if a particular query involves a relatively tiny bit of relevant data, it lives in an ocean of world wide knowledge (lots of junk and lots of irrelevant knowledge). So the technology has to scale to huge volumes and supports sophisticate data filtering.

**Heterogeneity** Even if this may be improved by normalization efforts, it is not conceivable to impose a unique way of describing scientific knowledge. This will be too limiting: scientists do have particular viewpoints, use different taxonomies, different procedures, different softwares. They keep pushing the limits of science, thereby introducing new measures, new concepts, etc.

**Interoperability** In the vision of world wide processable knowledge, knowledge is not isolated in complex systems but distributed over the entire world and *interoperable*. (Some knowledge derived in a system should be reusable by another.) Scientists should be offered systems that allow them to specify, for instance, complex sequencing of information management tasks without having to deal with mundane issues such as translation from one format to another, one protocol to another, etc.

**Extraction** It is much easier to manage knowledge than to extract it. (As an analogy, it is easier to validate a proof than to discover it.) However, the specification of semantics is a tedious task. Some (typically metadata) may be added automatically. In general, semantics should be as much as possible automatically-extracted vs., human specified.

Although solutions to this problem goes through developing better data/knowledge management systems, they also involve social issues that are mentioned next.

Scientists have been trained to publish articles and not data or knowledge. Indeed, such publication is often not well rewarded whereas it is a time consuming task. We have to consider changing the rules to encourage scientists to share their data as well as publish knowledge to simplify the use of that data.

An issue is that people want to keep intellectual property over their data/knowledge. It is easy to cite an article but much less so to give credit to a very large number of systems

that provided data for some experiments. Again, this requires changing the approach to scientific exchanges and evaluation.

Software tools have to take such considerations into account. In particular, information should be tagged with its provenance and include the specification of policies about its use. In any case, these are important information to assess, for instance, the quality of the data.

Finally, scientists in general want to keep control over their information. In particular, they may want to control the access to their own data and to trace its usage (as achieved for instance by water-marking for images). This seems to indicate that the computer architecture of such a system is less that of a centralized system, than of a peer-to-peer system<sup>1</sup>, where the system itself is shared by all the contributors. This fits the original independence spirit of science. Technically, this is not a limitation. Indeed, P2P systems typically scale much better than centralized ones.

### 3 Towards a solution

To be able to take advantage automatically of network resources, some syntactic and semantic description have to be attached to the resources. The syntactic description describes the format of the resources. For instance, on the Web [4], DTD and XML schema are used to describe the types of data that are exchanged, and WSDL types are used to describe the syntax of Web services. But such syntactic descriptions do not suffice. The computer system that is in charge of accessing, reasoning and computing with the information, needs to know the semantics of the information that is exchanged.

Without entering into details, one can classify the functionalities of such a system in some large categories:

1. acquisition: discovering the resources, in particular by crawling the Web, monitoring of (watching) some resources of interest, searching in some service directory (see UDDI).
2. enrichment: ahead of time, we want to perform some complex tasks to facilitate using the resources such as classification, semantic tagging, indexing, etc.
3. integration: it should be possible to integrate heterogeneous information coming from many sources to offer a unique entrypoint for queries, so that the scientist may be given the impression of the existence of one unique very-knowledgeable collection.
4. exploitation: this involve query processing (involving reasoning), report generation or the mining of these resources.
5. interfaces to the scientists: graphical interfaces for publishing and managing data/knowledge, searching for information, specifying complex tasks, etc.
6. publication: last but not least, this works much better if scientists publish resources with associated semantics.

---

<sup>1</sup>In a P2P system, each participant may contribute data or computing resources. Examples of P2P systems are in network content sharing and in distributed grid computing.

These functionalities are discussed in a bit more details in [1].

The system has to seriously address issues such as quality, result ranking, redundancy, cleaning (handling outliers), change control. Finally, it is worth stressing the importance of two aspects:

1. Metadata: these are typically essential forms of knowledge that may often be derived automatically such as the provenance of the data, the date it was created, some classification, etc.
2. annotations and cross references: David Maier makes a nice comparison between the bible (a relatively short text) and the annotations/commentaries (a huge resource). Science is also about *sharing knowledge*, so for scientific data, these are essential.

## 4 Conclusion

Can we facilitate the life of the scientists with respect to their management of information? The previous discussion highlights the need to develop better generic tools for the scientists. Knowledge should be an essential component of such systems. This may seem an unrealistic challenge given the modest progress of knowledge engineering (in my opinion) over the last decades.

There are reasons to believe the situation can be improved.

First, the room for improvement is just so immense, that there are reasons to believe we can do much better. For data management, scientists often use tools that do not take advantage of recent state of the art of research in computer science, not even of industrial products. For this part, the solutions are industrial (transfer of the results of research to the tools), economical (the tools should be affordable by the scientists), and educational.

Next, the situation is improving because of the Internet and the Web:

**Web101** The situation of access to Web knowledge has changed dramatically with the first wave of Web technology: HTTP, HTML, search engines. This turned the Web into a library with virtually unlimited size. However, it is meant for humans and not adapted to machine processing. Furthermore, it does not match the needs of scientist applications in terms of quality.

**Beyond** A second revolution is now occurring with standards that are meant to be used by computers. We now have standards for exchanging data (XML and its family of standards), standards for distributed computing (Web services and their family), emerging standards for knowledge (RDF and OWL). If the situation is still unstable and scientifically not perfect, these standards have the merit to exist and provide a basis for the dream of open access to knowledge.

Until today, all the efforts in knowledge engineering were invested in isolated systems. This resulted in a huge waste of energy. Most real world problems involve an array of functionalities that none of the systems supports, yet they cannot interoperate easily. With the arrival of standards, this issue is being resolved.

Some lessons may also be drawn from one most visible success of the Web, Google search engine. One can argue that a main cause of its success is the use of a ranking based on some mathematical analysis of the link structure of the Web. This works primarily because of the size of the Web: volume turns here into a lever and not a curse. Also, the *semantics* they use for documents (sequence of keywords) was originally much simpler than that used by unfortunate competitors. So, simplicity seems, at least as a first stage, an important aspect to scale to large volumes. This comes in handy when we know that, in general, the scientists publishing knowledge are rarely willing to specify complex semantics to the resource they publish. The lesson may be that, in any case, there is no point in using too complex semantics in a very heterogeneous world.

Also, the situation should improve because of the awareness of the computer science community. The management of information is viewed as one of the main challenges facing the database community [3], and more and more groups are working in the area. If industrial companies don't seem to be so eager to work on the problem, they are actively interested in a related topic, namely, technological watch, that involves similar technologies (acquisition, enrichment, exploitation of large corpus of information) and is viewed as a potentially very large source of profit.

**What are the main risks?** First, there is the risk of developing tools that do not interoperate. As already mentioned, solutions to specific problems may involve an array of techniques. So interoperable tools should be preferred to large computer systems behaving as black boxes. In this context, the absence of a well-accepted standard for describing some simple semantics is worrying.

Another issue is economical. The scientists should be able to afford access to the technology. This means first a preference for open-source software. This may also mean accepting to pay more for software, a change of mentality in some fields.

Finally, as already mentioned, a main issue is social. There needs to be changes in the rules governing the evaluation of science so that scientists are encouraged to publish data and knowledge.

## References

- [1] Serge Abiteboul, Managing an XML Warehouse in a P2P Context. CAiSE 2003: 4-13
- [2] Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, <http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html>
- [3] The Lowell Database Research Self Assessment <http://research.microsoft.com/gray/lowell/>
- [4] The W3C Web site [www.w3.org/](http://www.w3.org/).