Building Web Resources for Natural Scientists

Paul van der Vet*

September 2000

Abstract

Natural scientists increasingly rely on web-based information resources. The speed with which these data will be brought to the scientist's desktop in the near future, however, makes only too clear that automated support for efficient and effective use is in its infancy. In fact, most data are still processed by a combination of manual work and *ad hoc* programming. We propose to alleviate some of the problems by means of a research environment: a web site with highly graphical user interface that allows transparent access to resources and performs a certain degree of fusion of the information found.

1 Introduction

Natural scientists (biologists, chemists, geologists, ...) increasingly rely on web-based information resources. In highly competitive, front-line research areas such as supramolecular chemistry and molecular biology, knowing your way around in often thousands of web sites can make the difference between a fair and a famous group. In a compelling scenario, de Jong and Rip [4] sketch the activities of a research group in molecular biology who just have identified an unexpected experimental outcome. The group is able to interpret the finding and submit a paper to an established journal in a few days by making heavy use of Internet. They have searched data bases and knowledge bases for similar findings and they have remotely run qualitative simulation programs to predict experimental outcomes given particular theoretical assumptions. As a matter of important detail, the group includes a software librarian who knows where to find resources, what they do, and how they can be operated. Sadly, the scenario still is pure fiction.

In the present paper, I will take the information chain (or, rather, a simplified version of it) as a guide for an exploration into the possibilites of web interfaces that serve researchers' needs and may bring the scenario of De Jong and Rip closer to reality. We plan to build one for molecular biology, in cooperation with the Center for Molecular and Biomolecular Information (Nijmegen, the Netherlands) and the European Molecular Biology Research Laboratory (Heidelberg, Germany).

From the point of view of the end-user, the information chain can be roughly characterised as a four-step process.

Step (a) Identify needs. Searching information presupposes at least a rough idea of what is needed. *Step (b) Identify location*, traditionally the province of Information Retrieval

*Faculty of Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands; voice +31 53 489 3694; email `vet@cs.utwente.nl`. This paper appeared in: H. Scholten and M.J. van Sinderen, *Interactive distributed multimedia systems and telecommunication services*, Berlin: Springer, 2000, pp. 205–210.

(hence: IR). On WWW, search engines help sometimes. More often, one knows the location from other sources. *Step (c) Obtain information*, a laborious process in the old days that has changed dramatically through the introduction of WWW. *Step (d) Process information.* Having obtained the information, there remains the frustrating step of putting the information to further use. Data are scattered over many resources and rarely conform to the format demanded by the application at hand. Knowledge is often expressed in a representation language the user's programs cannot handle. Therefore, most processing is still done by hand.

2 Automating the Information Chain

Workers in natural sciences have to cope with a massive information supply. Technical developments will see the merging of steps of the information chain for reasons of both effectiveness and efficiency. Merging steps assumes automated tools that possess awareness of what the information conveys, in other words, of *content*.

ICT research so far has devoted its main efforts on steps (b) and (c) of the information chain, concentrating for step (b) on full-text search. Content providers like Chemical Abstracts Services tend to favour 'classical' approaches to the IR problem. At the same time, they widen their product range to include all kinds of secondary information. For content providers, acquisition of vendors of secondary information is a strategic investment because the company can offer its clients both the means to locate information and the information itself.

Indeed, there is no need to keep steps (b) and (c) separate, as, for instance, the SRS system [1] nicely demonstrates. Once the needed information is located, obtaining it is just a click away. For now, SRS and its likes only provide free information this way. For instance, SRS does not provide links to primary journal articles because accessing the articles themselves takes a subscription. Merging steps (b) and (c) in practice, therefore, presupposes an interface that incorporates an advanced payment system. The end-user has to be shielded from administrative details while at the same time total costs incurred so far have to be known. Also, the end-user may already have a subscription to some of the resources and does not want to pay twice.

Merging steps (b) and (c) leaves step (d). Reuse of information obtained elsewhere requires a dedicated combination of middleware and application software.

3 Systematisation of Domain Knowledge

Automated exchange of scientific information requires a certain degree of systematisation of domain knowledge. Within the field of artificial intelligence, systematisation of knowledge has long been a key issue. One of the important ingredients of any systematisation is a shared commitment to employ particular concepts for well-defined purposes. A concept system or ontology lays down such a commitment (see, for instance, [7], chapter 8). An ontology is a limitative, structured system of concepts. Limitative: the commitment is not to use other concepts. Structured: concepts are related; if possible, concepts are defined in terms of other concepts. Concepts: the contrast intended is with natural-language terms, taking the disambiguation provided by a good thesaurus or keyword system one step further.

A concept has to be identified by a name. Very often, this is a natural-language term, but since ontologies are designed for computer manipulation we can also decide to use another naming system, for instance, a system in which concept names reflect the definitions of the concepts [8]. The translation into something humans understand is then performed by an interface. In fact, developments in imaging and virtual reality allow even further abstraction from natural language. Chemists are already quite used to systems that visualise complex molecules and allow them to grab the molecule and turn it. In such a system, the molecule is depicted as a graph where the nodes stand for atoms and the edges for chemical bonds. The nodes and edges in the picture stand for concepts, and the set of such pictures can be regarded as an ontology of molecules. A further step, using techniques from virtual reality, would be to allow chemists to step into the molecule, inspect parts of it, and (for instance) push or pull to get an idea of the forces governing the molecule's shape. Even more exciting is the possibility to do this while the molecule is involved in a chemical reaction. These possibilities rely in part on the power of pictorial languages that abstract from reality.

4 Research Environments and the Role of Middleware

Processing information obtained from scattered and heterogeneous sources means that the researcher can plug the information she has found into the desktop application she happens to be running. For instance, a paper about the melting point at standard pressure of a particular substance was sought only because that value was needed for a calculation performed by some package. Ideally, the researcher should be able to tell the system so. In response, the system would extract the value from the text in a form suitable for the package and hand it over to the package without further manual intervention. More complicated jobs may require wiring together several databases, knowledge bases, and programs, including the researcher's desktop application. The particular configuration constitutes a research environment. It will only be useful for a short time, say between half an hour and a couple of days, and fine-tuning may occur frequently. We call such systems *coalitions*. A graphical user interface that allows the user to wire the coalition in a manner analogous to making a Lego object will significantly improve the accessibility and manœuverability of these complex information spaces.

Long-lived research environments are equally attractive. To return to the molecule example, we can make a virtual molecule. (The present proposal is inspired by the Virtual Music Center project [5].) A virtual molecule constitutes an environment in which a chemist can 'live'. Obviously, the environment will have to provide transparent access to a wealth of chemical knowledge. Some knowledge can be taken from existing sources, other knowledge can be derived from existing sources. Still other knowledge is simply not available: these are the places where the researcher can make a contribution, if she wishes.

What such an environment effectively does is to provide an overview of what is known about a particular topic. Steps (b), (c), and (d) of the information chain have been integrated into a single environment in which the boundaries between the steps have disappeared. As an interesting bonus, step (a), Identify needs, is also facilitated. Roaming through the molecule is a natural way of browsing. One cannot get lost because the environment is well-defined: any competent chemist knows her way through a molecule.

Research environments function because they integrate information from heterogeneous and scattered resources, both remote and in-house. Resources and programs tend to employ their own formats so that, behind the scenes, conversion operations are going on all the time.

Wiring resources together presupposes that at least directly connected items understand each other's formats. The problem can be solved by standardisation, but standardisation has a bad track record in many domains. For example, in spite of considerable standardisation efforts, there are about twenty different formats in use for files with chemical structural information. Ironically, most of these formats have started their career as proposals for a standard format. There are, I believe, two reasons why standardisation will not work. First, different jobs take different approaches, and no format will ever effectively cater for them all. Second, reaching agreement on a standard is a social process in which the stakes are often high. This also explains why so many standards are unmanageable: they are a compromise between consistency and social acceptability.

A format is a combination of form and content. On closer inspection, disagreement on standards very often involves form rather than content. If parties do not agree on the semantics of information carriers, communication is impossible. Agreement on meta-standards can be formalised by way of a concept system. To make this work, content providers have to provide a syntactic specification of their format accompanied by a semantic part in terms of an explicit ontology. Our own research shows that converters able to convert those data into another format can be generated automatically from unambiguous descriptions of the formats involved. Information is transferred internally in a knowledge representation language that heavily relies on the ontology.

A system of converters programmed by format specifications of the content providers will significantly improve information exchange because it can operate at lower costs, both financial and social, than a system that relies on standardisation. The advent of XML does not change this, if only because current discussions about standardisation of XML tags begin to display some of the symptoms of standardisation trouble discussed earlier: multiplicity of proposals and the tendency to associate particular programs with tag systems, thus destroying the separation of form and content.

5 Research Environments and the Scientific Communication System

The ideal research environment links resources and programs in a way that is shielded from the user. For the user, it appears as a virtual world in which she can move and interact with what is there. Many scientists explore processes that occur at or within three-dimensional structures: molecules, cells, tissues in bodies. For them, a virtual world provides a natural access to information. Realising such environments takes close collaboration between information scientists, computer scientists, and domain experts.

Research environments will function in highly dynamic social contexts which they in turn help shape. We can trace some of these influences by looking at the four functions of the scientific communication system identified by Roosendaal and Geurts [6]: *registration* (registering the research results of an author), *archiving* (making information reliably available), *certification* (assessing the quality of information), and *awareness*. The *awareness* function is the core function of the system. It deals with internalisation of information in an ongoing process of systematisation, comparison, and discovery.

Research environments are built with the awareness function as prime motivation. Researchers feel overwhelmed by the information flood. Indeed, some believe that even without performing experiments (other than, perhaps, confirmatory experiments), many new and

important discoveries can be made by exploring what is known already. This, then, is an important issue on the research agenda that will drive both technical developments and changes in the scientific communication system. ICT research should rise to this challenge. Researchers in developed countries find a heavy computer with continuous Internet access a minimal workplace provision. For them, computers and programs serve exclusively as tools to obtain, process, and disseminate content. I am not entirely sure that this is appreciated fully by ICT workers. For one thing, it entails that tools cannot be developed other than in the larger context of the scientific communication system and its organisational aspects.

While developed primarily to serve awareness, research environments affect the other functions as well. There will have to be some form of certification of the resources accessed by the environment. Certification is a matter of trust between author and reader. When access to the sources is effectively shielded from the user, the maintainers of the environment have to take measures to guarantee a certain minimal quality level or at least to tag sources with an indication of their quality. This shifts the trust relationship to one between maintainers and users of the environment. Trust is furthered when users are able to do assessments themselves, so that they can compare their own assessments with those of the maintainers. Technical support of quality assessment is possible to a certain extent [2, 3].

There are other consequences when the environment becomes widely known and used. There will be demand for use of the site to 'publish' new results. The enviroment then also serves a registration function. Maintainers will have to time-stamp and authenticate new additions. Incorporation of new results inevitably introduces variations in the quality level of the information offered, which puts extra demands on the certification function. Finally, a stable research environment has to address the archiving function. For instance, when the accessability of a particular remote source is believed to be unreliable, a mirror has to be set up.

A research environment obviously is a useful tool to researchers only if technical and organisational issues are addressed with equal emphasis and in their mutual relations. This makes the construction of research environments an essentially multidisciplinary effort. Building research environments promises the identification of new and challenging problems; using research environments promises new perspectives and discoveries. Let's start building.

Acknowledgement

The author is indebted to Hans Roosendaal and Peter Geurts for comments.

References

[1] T. Etzold, A. Ulyanov, and P. Argos, "SRS — an indexing and retrieval tool for flat file data libraries", *Methods in Enzymology* 266 (1996), 114–128.

[2] Jérôme Euzenat, "Building consensual knowledge bases: context and architecture", in: *Towards very large knowledge bases. Knowledge Building and Knowledge Sharing 1995*, N.J.I. Mars (ed.), IOS Press, Amsterdam, 1995, 143–155.

[3] Hidde de Jong, *Computer-supported analysis of scientific measurements*, Ph.D. thesis, University of Twente, Enschede, the Netherlands, 1998.

[4] Hidde de Jong and Arie Rip, "The computer revolution in science: steps towards the realization of computer-supported discovery environments", *Artificial Intelligence* 91 (1997), 225–256.

[5] Anton Nijholt, Joris Hulstijn, and Arjan van Hessen, "Speech and language interactions in a web theatre environment", in: *Proceedings of the ESCA Workshop on Interaction Dialogue in Multi-Modal Systems*, P. Dalsgaard, C.-H. Lee, P. Heisterkamp, and R. Cole (eds.), ESCA/Center for PersonKommunikation, Aalborg, Denmark, 1999, 129–132.

[6] Hans E.Roosendaal and Peter A.Th.M. Geurts, "Scientific communication and its relevance to research policy", *Scientometrics* 44 (1999), 507–519.

[7] Stuart J. Russell and Peter Norvig, *Artificial intelligence. A modern approach*, Prentice Hall, Upper Saddle River NJ, 1995.

[8] Paul E. van der Vet and Nicolaas J.I. Mars, "Bottom-up construction of ontologies", *IEEE Transactions on Knowledge and Data Engineering* 10 (1998), 513–526.