The Impact of Electronic Publishing on the Academic Community

Session 6: Access to scientific data repositories

Political decisions and economic issues

Bernard Levrat

Services Informatiques, University of Geneva, 24, rue du Général-Dufour, CH-1211 Geneva 4, Switzerland

---

Introduction

Just a few words about my background, I started as a physicist and worked at CERN (the European Organization for Nuclear Research) for five years, with the responsibility of data processing for a large online experiment. In 1968, I became professor of computer science, with the mission of introducing computing in the university. My research concentrated on information storage and retrieval and man--machine communication. I am proud of teaching software engineering with modern, real-world oriented methods and it is probably what got me into hot water these last two years.

During the four years I spent as vice-rector of the university from 1991 to 1995, I was responsible for finances, computing and the libraries. Our libraries are part of a consortium called RERO (Reseau des Bibliothèques Romandes), to which all the research libraries of the French and Italian-speaking Swiss cantons belong [1]. They include university libraries as well as reference libraries.

RERO was using a proprietary system called SIBIL (Système Informatique des Bibliothèques de Lausanne) started 20 years ago and written in PL/I [2]. Migrating it was urgent since the running costs were much too high. Because of the large budget savings, it was easy to convince all the partners that it was a good thing to do and I was given the task to organize the migration. I thought that applying the methodology I was teaching would make for a nice, painless transition. I was wrong, but the team I put together managed to do it anyway, acquiring a rare insight in the way a consortium of libraries operates.

My second surprise came from the astronomers. The Rectorate I belonged to encouraged them to apply for the job of analysing the data coming from a gamma-ray satellite. They formed the Integral Scientific Data Center and are nicely set up with a high-bandwidth connection to the observatory [3]. As director of computing services of the University of Geneva, I am a bit worried about the requirements from the future users of what is going to be a data repository of several terabytes.

Accessing the global information resource

Let's start with the remark that the Internet already supports a conservative 30 million computers or (´@200|megabytes@=) 6|petabytes of data. Most of the current human knowledge can be found there in digital form, the question is how to access it. The problem is part of a general concern on the role of digital resources in the future, a set of data from the Hubble telescope being on a par with a digitized version of one of the first editions of the Gutenberg Bible (or of Beowulf).

We want universal access to scientific data repositories and to other information sources such as library catalogues, online journals, better than the current URLs. They should serve at the same time individual users' communities and be part of a global accessing scheme. MARC (machine-readable catalogue) records already include International Standard Book Numbers (ISBNs, books), International Standard Serial Numbers (ISSNs, periodicals) and URLs (Websites) [4]. Insuring the long-term survival of digital information resources will be best handled by specialists. (Should we call them digitarians?)

What about the appropriate software to exploit the data? It is transparent for librarians doing cataloguing and indexing. How will it be done for scientific data? Will the necessary software be freely available to would-be users of the data the way the CERN program library was distributed?

In some cases, local processing may be needed to extract significant subsets of the data. It would be possible to charge for it, the difficulty being to determine beforehand the capacity of the machines without hints of how much they would be used. Standard packages could also be required (Z39.50 servers, AVS [Application Visualization System], Mathematica, Maple, Uniras, Molecular modelling software, etc.). A consortium of users of a given discipline would have to recommend versions and minimum configurations, and provide some sort of help to adjust the tools to the problems.

What is a consortium? It is an association of institutions with common interests that receives some means to achieve a common set of goals. Users are the initiators rather than the government. I hope that the constitution of scientific data repositories will follow the model of libraries rather than having different solutions for individual disciplines.

In international discussions to adopt standards or standard practices, a consortium has a weight proportional to the sum of the weight of its members. Americans have been very good at it, achieving complete dominance in many fields including scientific journals and library automation.

The consortium framework does not exclude a sound financial basis and I would like to offer some relevant examples. (The information concerning these examples is in italics, taken straight from the Web.)

OCLC (Online Computer Library Center, Inc.) [4] is a non-profit computer service and research organization whose network and services link more than 24\000 libraries in the U.S.A. and 63 countries and territories. OCLC services help libraries locate, acquire, catalogue and lend library materials. OCLC has the organizational power and the resources to move towards a global information database. From the outside, it looks very business-like. Should European efforts rediscover what OCLC has developed over the years?

Alexandria Project: the goal of the Alexandria Project Digital Library (ADL) is to build a distributed digital library for geographically referenced materials. A central function of ADL is to provide users with access to a large range of digital materials, ranging from maps and images to text to multimedia, in terms of geographical reference. An important type of query is, "what information is there in the library about some phenomenon at a particular set of places?"

This architecture involves: (i) user interface components that support graphic and text-based access to the other ADL components and services; (ii) a distributed catalogue component that includes metadata and search engines permitting users to identify holdings of interest; (iii) a distributed storage component containing the digital holdings; and (iv) an ingest component allowing librarians to store new holdings, extract metadata from the holdings and add metadata to the catalogue.

Could we consider Alexandria as a scientific data repository in the making?

Limits of models

The political decision to fund a consortium is usually based on a number of economic *a priori* considerations which are not necessarily verified when in operation.

The argument for RERO was that full cataloguing was a time-consuming activity which could be shared among members while local data was easily added when required. True or not, RERO librarians developed a high-quality cataloguing culture which lead to an exceptionally good database, albeit costly to maintain.

Faced with the crisis in scholarly publishing, RERO thought and politicians agreed that a sharing of the less popular titles would preserve the completeness of collections over the network without increasing budgets. That decision boosted a previously quiet activity known as inter-library loans. The requests for photocopies of articles from other expanded manifolds.

Real costs have been replaced by hidden costs, which seem to be very high, at least ten times the amount of money saved by cancelling the duplicate titles. Students used to go to the library, read an article and take a few notes. They now request a photocopy through inter-library loan. Hopefully, electronic publications, cheaper to consult over the Internet, will bring that problem under control, but the effects should be kept in mind when considering savings by having single

copies of important data and documents stored somewhere instead of being distributed when needed.

Repositories viewed as a continuum

As much as I would like to see it happen, the difficulty I see in organizing repositories as a continuum of information is convincing specialized people that it is more useful than having separate solutions. Yet, scientists like to look into disciplines bordering their own. Reviewing bodies need to have common evaluation instruments. Navigation should be uniformly signalled. Partners must be convinced there will be economies of scale and new opportunities.

Proposition for a political agenda

Organizing and accessing scientific data repositories should be entrusted to a consortium, built on existing libraries, broadening their current scope. Models of use with a wider audience including students and an educated public have to be worked out, both for storage volume, processing power, network bandwidth and financial contributions. Help should be provided to move from URL to named resources.

Adopting standards already worked out by U.S.A. organisms does not necessarily mean losing a European identity as far as contents are concerned, provided the storage media permits it. Europeans should make a strong case for Unicode [6] or ISO 10646 to be used everywhere it is possible, especially in indexing material. I may add that standardization rules for bidirectional language display management should be provided as the present rules are very vague. Normalization rules should be specified on searchable items to make searching consistent.

References

1. http://www.rero.ch/reroweb/presentation.html

2. http://www.unige.ch/biblio/opac/sibil/index.html

3. http://obswww.unige.ch/isdc

4. http://www.oclc.org

5. http://alexandria.sdc.ucsb.edu

6. See for example http://www.w3.org/international/0-unicode.html

---