Evolving the Digital Library: Development of 3rd Generation Tools at Los Alamos National Laboratory

Richard E. Luce
Research Library Director & Library Without Walls Project Leader,
Los Alamos National Laboratory
January, 2000

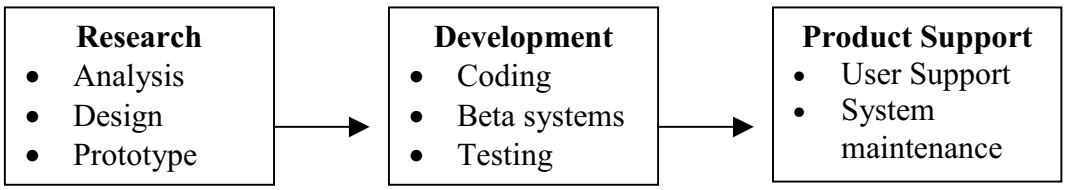
Abstract

The Library Without Walls Project at Los Alamos National Laboratory has developed leading-edge digital library products. This paper outlines new activities aimed at creating a third generation of digital library tools. Current research and development activities are described, with an emphasis on efforts intended to create an adaptive environment for the next generation of digital library capabilities. This third generation capability, the Active Recommendation System, is outlined.

"Our vision is to create a network of knowledge systems that facilitate scientific communication and collaboration" R. Luce

Background

The *Library Without Walls* (LWW) at Los Alamos National Laboratory is widely known as a pioneering, state-of-the-art digital library[1]. The intent of our initial funding in 1994-95 was to support the development of a digital library collection. Having achieved to a significant degree the initial goal, what comes next? The need for strategic innovations to sustain both funding and our development capabilities makes it essential to invest in a research component (front-end feeder) to anticipate and meet future needs. A simplistic view of this structural breakout is illustrated below.



Typically, customer feedback for new products and library literature describe activities in the new development and product support boxes above. However, without sufficient strategic attention to the research and analysis box it is unlikely new products can be developed and brought to market in time to meet expressed user needs. Hence, to avoid lagging the market and competitors, what is the roadmap for the next generation of digital library development? This paper focuses on one component of the research effort to support 3rd generation capabilities underway at Los Alamos.

Current Scene -- Current Development Projects

The LWW team is currently working on several second-generation development projects, which are driven by analysis of customer needs. Representative efforts of current projects include the following:

- **Database Integration and Searching Capabilities** – Two distinct, yet complimentary, components are under development to support this objective. First, we are shifting our current web applications from our library MARC-based system to an integrated full-text retrieval capability that supports relevance feedback. This path was pioneered through the development of our *SciSearch at LANL database*, the first web-based implementation of the Science Citation Index from ISI which contains over 20 million bibliographic records with 205 million citation records, and 115 million hyperlinks. Later databases such as Biosis, INSPEC, Compendex, Energy Database, etc., were incorporated. By combining our locally managed web databases into one common platform, we will enable users to search across multiple local databases spanning 54 million records with a single command.

  The second related capability is dynamic linking to external collections. This entails participation and support of the special effects linking prototype developed by Herbert Van de Sompel from the University of Gent[2]. This work is significant for both publishers and libraries because it allows the user to determine connections of interest. Data regarding the associations utilized by end-users will provide a wealth of statistical associations, which we can use in our Active Recommendation System as described below.

- **Hybrid Information Content** – Our second-generation LWW capabilities require integration and synthesizing the structure and relationships between databases and various types of digital library content. An example is the conversion of 12 years the IEEE content (standards, conference papers, and electronic journals) into a format that can be searched and retrieved in combination with our large array of electronic journals. This resource will be a unique a hybrid between database, sets of journals, and other types of content such as standards and conference papers, which are integrated together supporting dynamic citation linking.

- **Preprints** – We envision the integration of the *xxx.lanl.gov* physics preprint system developed by Paul Ginsparg together with the Research Library's formally published literature, thereby allowing researchers to combine the databases together to retrieve relevant publications and pre-publications. The first step in this evolution was the development of a new user interface for xxx, now renamed the *arXiv* or *E-print ArXiv*. This project has the goal of institutionalizing the xxx e-print archives, while bridging the gap between new publishing models and traditional publication sources.

While the above projects are challenging in themselves, an added challenge is the strategic question remaining: what is next in the research cycle that will feed the next generation of development? Analysis of this question is crucial, given the research and development lead time requirements, if the next generation of products are to be delivered when (rather than after) demanded by our customers.

Future Directions

The research library of the future will be an organization that does much more than collect ever-larger amounts of digital information. Examining current and future functionality of the web, two themes emerge: (1) the demand to organize, provide access and *add value* to the aggregated information to make it useful, especially for scientific collaborations; and (2) the desire to foster and cultivate the *communities* built around common information interests.

We have a charted a path for the following new research and technical development areas for the LWW team to support our third generation of digital library tools and capabilities:

- **Evolving E-prints** – As work with Ginsparg's xxx preprint system or *e-print arXiv* becomes institutionalized under the Research Library/LWW, we are exploring methods to integrate the *arXiv* e-prints with other e-print resources. In order to search and access e-print literature among various archives and disciplines, a common or "standardized" structure for the new metadata descriptors will be required. Such a format obviously must support query, exchange and linking between preprint and e-print servers regardless of discipline.

- **Visualization** – Our users desire more intuitive GUI's when interacting with a variety of information resources. Investigation of visual user interfaces to navigate large metadata sets and interaction and manipulation of smaller sets of data is required. Currently the *LWW* is collaborating with Sandia National Laboratories' *VX Insight* team on the development a visual interface for very large sets bibliographic and citation databases, as well as testing tools for manipulating smaller sets of data.

- **Active Recommendation System** -- A collaborative project between LANL's Computer Research Group and the LWW, we are investigating new methods to identify and connect communities. We are currently developing a computational testbed to deploy an adaptive recommendation environment, called the *Active Recommendation System*.

This paper, one in a series, will describe the *Active Recommendation System* in greater detail, as well as some of the research and design questions we are trying to solve.

New Research Direction: The Active Recommendation System

Problem Analysis

Modern library systems at universities and research institutes are perfect examples of today's complex Distributed Information Systems: they are responsible for serving large and diverse technical communities by providing access to an extensive set of equally large and heterogeneous electronic information resources. As the complexity and size of both user communities and information resources grows, the fundamental limitations of traditional information retrieval systems have become evident. Recent user surveys and interviews of Los Alamos researchers revealed desired functionality, currently unavailable in today's library. (The survey was conducted as one component of the LWW's *Voice Function Deployment* (VFD) process to determine future customer needs. VFD originated in the Quality Function Deployment world). The following limitations with today's systems and capabilities were evident:

- There is no **crossover** of information. It is be very difficult for users to search across databases from different disciplines.
- There is no "**push**" of information. The system does not issue recommendations to its users about related topics that they may be unaware of.
- There is no **user profiling**. The system does not remember user preferences or user-specific keyword categories.
- There is a failure to work at the **concept** level. The system relies on fixed keywords, but does not infer categories of keywords used by communities of users.

The sources of these limitations can be traced directly to a number of technical deficiencies of current distributed information systems, in particular, that they are:

Passive: Information retrieval in distributed information systems is generally unidirectional and query-based , and thus only able to respond to specific user requests. They can generally neither proactively generate information for users, nor even respond to queries in a user-specific fashion. Instead, users must know in advance what information they need, and then try to pull it from the environment.

Semantically Fixed: Semantic tags (keywords) must be provided explicitly by authors (or publishers, librarians, and indexers). The keywords in each document or database are bound to the "concept space"

of these authors, which may be incoherent with the concept spaces of the users, or of the authors of other documents or databases.

Static: Once deployed to users, the knowledge in distributed information systems remains fixed. Any indirect knowledge available through analysis of these structures, or implicit knowledge inherent in the patterns of information retrieval, cannot be exploited to enable push of user-specific content or to enhance semantic representations of content.

Isolated: Knowledge is represented in distinct formats on separate systems. Thus knowledge about the common properties of related domains or databases (available, for example, from an analysis of common structure or directly from users) cannot be exploited.

Existing Technologies for Recommendation Systems

New approaches for information retrieval have been proposed to address these limitations. These active recommendation systems, also known *as Active Collaborative Filtering, Knowledge Mining*, or *Knowledge Self-Organization environments*, rely on active computational environments that interact with and adapt to their users. They effectively push relevant information to users according to previous patterns of information retrieval or individual user profiling.

Typical recommendation systems come in two varieties:

1. **Content-based** systems, where user profiles are created based on the system's keywords. These systems establish a means of recommending documents to users according to their profiles and some kind of semantic metric that describes the relationships between keywords inferred from their association with common documents.

2. **Collaborative systems** do not involve any description of the semantics or content of documents, but rather issue recommendations according to a comparison of the profiles of several users that tend to access the same documents. These user profiles are not based on keywords, but on the actual documents retrieved.

Content-based systems depend on single user profiles, and thus cannot effectively recommend documents about previously unrequested content. Conversely, pure collaborative systems, with no content analysis, match only the profiles of users that (to a great extent) have requested the same exact documents; for instance, different book

editions are considered distinct documents. It is clear that effective recommendation systems require aspects of both approaches.

Systems Development and Research

We are developing and researching recommendation systems for the Los Alamos *Library Without Walls*. These systems will be both collaborative and content-based, and will exploit currently untapped sources of information in distributed information systems. In particular, they will integrate information from the patterns of usage of groups of users, and also categorize database content or semantics in a manner relevant to those groups. Moreover, we intend that the semantic tags and conceptual categories need not be just designed into these systems, but may also be induced and evolved from document content, user-supplied information, and group interaction.

Our overall aims are to deploy software applications within the LWW, and to use the LWW and its user community itself as an object of scientific study. These efforts will provide substantial benefits to the expanding needs of the library by responding to the specific issues revealed in the recent survey, in particular by:

- Establishing **crossover of different subject matter** by enabling search across multiple, interdisciplinary databases.
- Also establishing **crossover among heterogeneous types** of information resources (for example linking abstract indexes such as Inspec with deep-content sources such as e-journals).
- **Pushing** recommendations of related topics that users may not have thought of.
- Expanding the "**semantic space**" of the databases to a more conceptual level, including qualified keywords, keywords derived from the analysis of document content, and higher-level "conceptual" groupings reflecting collaborative usage patterns.
- Deploying a more **personalized human-machine interaction** from a consolidated point of access.
- **Detecting patterns and relationships** of information retrieval leading to the adaptation of the environment to the users.

These overall goals have been pursued during a modest first-year effort to demonstrate fundamental engineering capabilities and scientific results. Initially, an existing prototype of Luis Rocha's *TalkMine* recommendation system will be developed and deployed, and the inherent semantic structure of LWW databases will be analyzed. Later work will see the analysis of customer satisfaction and experimental results, and lay the basis for expansion of these methods in following years.

References:

1. Pack and Pemberton. "*A Harbinger of Change: The Cutting Edge Library at the Los Alamos National Laboratory*" Online Magazine. March 1999, pp. 34-42. http://www.onlineinc.com/articles/onlinemag/pack993.html

2. Van de Sompel, Herbert and Patrick Hochstenbach. *Reference Linking in a Hybrid Library Environment.* Part 1: Frameworks for Linking and Part 2: SFX, a Generic Linking Solution. D-Lib Magazine, April 1999. Volume 5 Issue 4.

3. Rocha, *Adaptive Recommendation and Open-Ended Semiosis.* International Journal of Human-Computer Studies. In Press. [Complex Systems Modeling Team, Computer Research and Applications Group, Los Alamos National Laboratory.]

Further readings:

Henry, C. and Luis M. Rocha. "*Language Theory: Consensual Selection of Dynamics*." In: Cybernetics and Systems: An International Journal. Vol. 27, pp. 541-553. [1996].

Johnson, , Rasmussen, Joslyn, Rocha, Smith, and Kantor [1998]. "*Symbiotic intelligence: self-organizing knowledge on distributed networks, driven by human interaction*". Sixth International Conference on Artificial Life, UCLA, 1998. In Press.

Library Without Walls Anticipates Its Next Generations.
http://lib-www.lanl.gov/lww/lww_bits/Ann_library.html

Luce, R. *"Integrating the Digital Library Puzzle: The Library Without Walls at Los Alamos*". University of Tilburg lectures. August 1998. LA-UR-98-2575.
http://lib-www.lanl.gov/lww/tilberg.htm

Luce, Richard E. "*Library without Walls: Digital Library Developments at LANL's Research Library*". **Bits: Computing and Communications News**, April 1995. pp. 4-5.
http://lib-www.lanl.gov/lww/april95.pdf

Rocha, Luis M. [1997]." *Evidence Sets: Contextual Categories* " In: Proceedings of the meeting on Control Mechanisms for Complex Systems, Physical Science Laboratory, New Mexico State University, Las Cruces, New Mexico, January 1997. M. Coombs (ed.). NMSU Press, pp. 339-357.

Rocha, Luis M. " *Cognitive Categorization revisited: extending interval valued fuzzy sets as simulation tools for concept combination.*" In: Proceedings of the 1994 International Conference of NAFIPS/IFIS/NASA. IEEE Press. pp 400-404. [1994].