

## **Navigating the Distributed World of Community Knowledge**

Bruce R. Schatz, University of Illinois

[schatz@uiuc.edu](mailto:schatz@uiuc.edu), [www.canis.uiuc.edu](http://www.canis.uiuc.edu)

It is common today to state that “The Net is the Structure of the World”. Living on the Net is everyday life for the university faculty and students who are the primary audience for the NSF. Since Living on the Net is also becoming everyday life for ordinary people, pushing new technologies to support the Net must become a major topic for NSF Programs. These trends are well recognized in the NSF Reports on Revolutionizing Science and Engineering through Cyberinfrastructure <http://www.nsf.gov/search97cgi/vtopic> and Science and Engineering Infrastructure for the 21<sup>st</sup> Century <http://www.nsf.gov/nsb/documents/2003/start.htm> .

It is not so common to state that the nature of the world itself has changed. Traditionally, online information has been dominated by data centers with large collections indexed by trained professionals. The rise of the Web has followed this model with many small clients searching few large servers. Even as the Internet has made the historical transition from access to organization, from web browsing to web searching, the paradigm has remained central searching of large archives [Schatz,1997]. The Digital Libraries Initiative, Phase 1, helped push this transition by demonstrating new infrastructure for information searching, such as Google.

The rise of distributed personal computing has radically changed the nature of online information. The traditional components of the publishing cycle, with separate authors and indexers and publishers, are breaking down. Individual persons and individual computers take on all the traditional roles at different times. One exemplar for this trend is peer-peer protocols, such as popularized by music sharing systems such as Napster or cycle sharing systems such as SETI @ HOME.

In the future, online information will be dominated by small collections maintained and indexed by small groups. These digital libraries will store community knowledge, and the great mass of objects on the Net will be stored in these community repositories. New indexing techniques are necessary for special libraries and new searching techniques are necessary for these new indexes. The Digital Libraries Initiative, Phase 2, helped push this transition by demonstrating new collections for many applications.

The Net has already made the transition from data transmission to information retrieval. It is in the process of making the transition from information retrieval to knowledge management. The recording of special knowledge in distributed collections requires different technologies than are customary in knowledge management. In particular, federation across collections is necessary for navigation in the Net. This requires knowledge representations that are comparable across collections, leading to a new paradigm of analysis across repositories [Schatz,2002].

This new paradigm of cross-analysis is a modern restatement of the classical problem of information retrieval called vocabulary switching, where the difficulty was mapping across subject thesauri, or across document contents. Since the first Phases of Digital Libraries Initiatives concentrated on information, the cross-analysis concentrated on basic units, such as

words for documents or textures for images, and terminology switching across subject domains. The Grand Challenge in the 1990s was posed as “semantic interoperability across digital collections”.

The coming phases of research must concentrate on knowledge, so the cross-analysis must concentrate on deeper units, such as cross-languages or cross-cultures. Even with special libraries all in the same language and culture, mapping similar ideas from different ontologies is a hard problem. The fundamental infrastructure question for the future of the Net is “how to make all seem one”. Stated more technically, how to navigate effectively across different collections of different objects represented by different communities at different levels. The Grand Challenge in the 2000s will be “conceptual navigation across community repositories”.

The future Digital Libraries Initiatives must study the problem of analysis in depth, in the context of real examples. Cross-analysis and Special-knowledge pull in different directions, so generic knowledge representations can be shown effective only in practice. What is needed is deep and broad Cyberinfrastructure research, of a particular type. This research deals with supporting the needs of real users on real collections, with emphasis on navigation and representation.

The future Initiatives must thus focus on developing new infrastructure at a scale sufficient to test its utility. Practical construction of information systems, practical development of knowledge management. Full systems with research technologies and complete applications.

The applications where these are tested must serve as effective models for the proposed functions. It is not as important that the applications be specifically the audience for NSF, as that the domains are good models for experimental purposes. Thus, comparative literature in humanities might be a better model for cross-analysis of deep structures than particle physics in science. Or healthcare might be a better model for capturing complete lifestyles than computers.

These observations lead to key structural features of any future Initiatives.

- Projects must be full-spectrum, including systems, users, collections.
- Projects must build new infrastructure and evaluate its utility in real testbeds.
- Applications may be in any subject domain, not necessarily NSF supported.
- Investigators must include representatives for all relevant components.
- Investigators may work at foreign institutions (international collaboration).
- Typical scale is 5 investigators for \$5M over 5 years.
- The numbers of users and the reality of the system must be related to the scale.
- Larger projects are permissible, but must support wide usage over long periods.

### *References*

Schatz, B [1997] Information Retrieval in Digital Libraries: Bringing Search to the Net, *Science* 275: 327-334 (Jan). Cover article for special issue on Bioinformatics.

Schatz, B. [2002] The Interspace: Concept Navigation across Distributed Communities, *Computer* 35(1): 54-62 (Jan). Information Infrastructure article for annual Trends issue.