

Thoughts on the NSF Role: Network-Enabled Frameworks for Knowledge Leveraging

Dave Fulker, 4 June 2003

Introduction

This piece was developed in preparation for an invitational workshop on “post digital library futures” (or “ubiquitous knowledge environments”) June 2003, in Cape Cod. For brevity, the component of the Cyberinfrastructure initiative described herein will be abbreviated KL (for Knowledge Leveraging).

Purpose

The approach described herein is put forward as one that the NSF might adopt in defining a KL component: *for researching and advancing practices in the application of technology to enable knowledge advancement on unparalleled scales and at unprecedented rates.* The problem to be addressed—borne of profound changes in the technologies now available for creating, exchanging, managing, and presenting information—is that longstanding foundations for advancing human knowledge have not yet been fully adapted to accommodate, much less to exploit, these new technologies.

Hence the rate and the breadth of knowledge advancement presently is limited, relative to what could be achieved and sustained with improved infrastructure in four key areas: bibliographic systems, cognition-leveraging tools, social factors, and information architectures.

Impact and Likelihood of Success

Knowledge leveraging frameworks have a special place within the NSF’s Cyberinfrastructure initiative, as they deal with that end of the data-information-knowledge-wisdom spectrum where qualitative human judgment is most critical, where learning is most directly affected, and where matters of meaning and permanence are most germane. Hence it is within a KL component that cyberinfrastructure would have the greatest impact on the Nation’s workforce, on its education systems (including the integration of education with research), and on the legacy that scientific and engineering endeavors create for future generations of researchers and students.

The work to be done has precedents on which one may project a high likelihood of success for KL endeavors. Indeed, important prior results have been achieved in each of the four key areas mentioned earlier: bibliographic systems, cognition-leveraging tools, social factors, and information architectures. Ideas for improvement in each of these areas, relative to prior work and current practices,

are sketched in the sections that follow. Taken together, such improvements (combined with others that cannot be predicted) will accelerate the progress of science and entrain more citizens as players in and beneficiaries of this progress.

Bibliographic Systems

Current practices of knowledge organization, for the general public and for scholars or others with special interests, have been developed and refined over a 150-year period. Four decades of this have incorporated computer technologies, including a 10 years of NSF-supported research on digital libraries, some of which explicitly targeted bibliographic practices.

Despite this rich history, technologies for creating, exchanging, managing, and presenting information have outstripped the capacities of bibliographic systems to 1) paint comprehensive views of the knowledge landscape or 2) fully address the needs of the growing number of people who now expect to navigate this landscape without the expert assistance of librarians. This is not a criticism; the challenges are profoundly difficult and changes occur with astounding speed.

The basic functions of a bibliographic system are *finding, collocating, choosing, acquisition, and navigation* [Svenonius], and each is being affected by rapid technological change. Some key challenges to be addressed by KL-sponsored investigators are discussed below:

Dynamic Content and Atomicity

Resources that are commonplace in the computer age do not necessarily come packaged neatly as indivisible (atomic) units, and some of the most expressive media, such as real-time observational data, are highly dynamic. In *The Intellectual Foundations of Information Organization*, Svenonius points out:

“Documents with uncertain boundaries, which are ongoing, continually growing, or replacing parts of themselves, have identity problems. It is not possible to maintain identity through flux ('On cannot step twice into the same river' [referencing Heraclitus]). ... A snapshot cannot accurately describe information that is dynamic. This is not simply a philosophical matter, since what is difficult to identify is difficult to describe and therefore difficult to organize.”

Unfortunately for bibliographers, such documents are an increasingly important aspect of the knowledge landscape, especially in science and engineering.

Some progress has been achieved in respect to this problem, but an important and achievable goal of KL should be to gain broad acceptance and use of an *operational* definition for uniquely identified digital entities across most or all NSF-sponsored cyberinfrastructure. The associated identifiers would foster interoperability and form crucial foundations for all types of bibliographic systems.

Mixing Multiple Approaches to Metadata Creation

The likelihood of a single approach to creating metadata is decreasing, because multiple methodologies, aligned to specific contexts, are proving successful. Though nearly all of these seek to reduce human effort, whether exerted by expert catalogers or others, the approaches run the gamut from complete automation to a variety of leveraging strategies. Further increasing this diversity, some systems make very effective use of strongly typed metadata fields (such as for geospatially indexed resources) while other successful efforts (such as systems built with OAI-PMH or based entirely on content analysis) employ few data types beyond text.

Hence an important but open question is how to fulfill the bibliographic functions of finding, collocating, choosing, acquisition, and navigation in contexts where the joining of multiple libraries and collections yields a mixture of metadata approaches. Research and development on this challenge must be cognizant of and coordinated with efforts to create the “Semantic Web” and to advance the use of numerous schemas, vocabularies, and ontologies.

Wide deployment of an overarching model that embraces multiple forms of metadata seems essential in a Cyberinfrastructure program. This would serve as a key component in a common technical fabric for linking independent digital libraries and creating coherence on national and international scales.

Universality versus Specialized Expressiveness

Closely related to the foregoing discussion is a need to better understand—from the perspectives of *users* as well as of content and metadata providers—the tradeoffs between metadata universality and specialization. Every bibliographic system designer faces difficult choices between maximizing compatibility with other systems and maximizing expressiveness, relative to the needs of a target audience, often comprising specialists in particular fields.

Solid research data in these matters would greatly improve both planning and effectiveness for a vast array of information and knowledge handling systems.

Scalability of Bibliographic Systems

Finally, there may be no greater problem facing current bibliographic systems than scalability. The problem derives from new user expectations (driven in part by their experiences with Google and its kin) and from increased creation rates of materials to be included in science-related bibliographic systems.

A tempting solution might be to rely entirely on Google or the like, but studies [Sumner] indicate that users expect levels of bibliographic functionality that are beyond what Google or other current systems can deliver, at least for now. Hence an important KL outcome should be to charting a course for bibliographic systems that address the foregoing challenges and simultaneously exploit the scalability of automated approaches to information organization and discovery.

Cognition-Leveraging Tools

Fundamentally, all technologies leverage human capabilities, but computing and networking are especially rich in the forms of augmentation they offer. Of these forms, the leveraging of cognition is one of the most important, enhancing the abilities of humans to understand one another, to gain new perspectives on the universe they occupy and, most fundamentally, to learn.

Among the best outcomes of the NSF's Digital-Library initiatives have been demonstrations of such impacts [ADEPT], but the affected audiences remain relatively small, and the full potential appears far greater than what has been achieved to date. Further, learning and cognition represent areas where research and education meet, so Cyberinfrastructure emphasis on these matters will extend the impact to encompass NSF educational goals and to affect very large audiences, including a significant fraction of the American workforce.

Hence the KL outcomes should include new and improved tools for collaboration and for working (interactively) with all artifacts of scientific progress, including: *observed and simulated data; taxonomies; mathematical expressions; molecular, chemical, and genomic expressions; structural, physical and computational models; tables, graphs, charts, maps and images; field and laboratory notebooks; monographs and other scholarly documents; critical reviews and discourse; ontologies; and bibliographic references to scholarly literature.* Implicit in the need for tools is the need for widely agreed, non-proprietary, digital representations for such artifacts.

Functions addressed by cognition-leveraging tools might include: *spontaneous on-line meetings, collaborative editing (of rich scientific documents), bibliography sharing, curriculum architecting, semantic tagging, knowledge mapping, visualization sharing, (large) data-set structuring, and creating (shared or personal) logs or diaries of experiments and studies.*

Such tools, embedded within digital libraries and elsewhere, will help ensure that the Nation's cyberinfrastructure is an active, not a passive, place for learning by students as well as researchers. This in turn would promote educational enhancements based on inquiry, constructivism, and group learning. [Reeves]

Social Factors

Learning and scientific advancement are fundamentally social activities, despite widely held perceptions about the isolated nature of scientific and scholarly endeavors. Furthermore the Internet is decreasing whatever degree of isolation previously existed, and it is increasing the pace of discourse and interaction underlying scientific progress. The Internet also is causing reconsideration of, and changes in, ethical traditions long championed by academic and public libraries: *fair use, equal access, free speech, and strong practices to ensure privacy.* The Cyberinfrastructure initiative, especially its KL component, must

extend these library traditions, as well as those of customer service, dependability, and longevity [Besser], into the digital era. Doing so will nourish realization of the Internet as a *shared commons for creative work*, critical for scientific progress [Boyle]. As described in the preceding section, cognitive tools and collaboration environments also will encourage this realization.

To gain alignment with key social and institutional needs, the Cyberinfrastructure initiative should explicitly support the formation of communities of practice, built initially by change agents and early adopters of KL systems. Properly chosen and supported, such communities will increase the Cyberinfrastructure impact on both specialized and general-purpose domains, fostering improved science accessibility for all citizens.

These communities are critical to another aspect of the KL component and its digital libraries: *selective, intelligent, targeted collection development*. [Keller] Though no single individual or group can hope to perform this function for the entire community of Cyberinfrastructure users, the problem is tractable within smaller communities of practice, where domains of interest are more limited. Stated another way, smaller communities will create opportunities for excellence where global approaches are likely to yield mediocrity. Such communities can help determine which entities are most deserving of resources to assure their long-term preservation, which is a well established responsibility of libraries.

In a related matter, the KL component should seek a broad geographic presence, with a commitment to embracing a very large number of educational institutions, including those that traditionally have not received significant NSF support. Wide geographic distribution would enable the virtual communities emerging from the KL to be seeded and strengthened by face-to-face interactions. This idea is informed by the success of the regional-networks used to deploy the NSFnet, and it may be advantageously linked to the nascent Institutional Repositories movement [Crow].

Finally, the KL component should support anthropological studies, documenting the emergence of social norms and communities of practice around digital libraries and other aspects of cyberinfrastructure [Khoo].

Information Architectures

Though progress has been achieved in important areas (such as OAI-PMH and DC metadata), a common architecture for digital libraries remains elusive. Yet to be realized fully on a large scale is the goal articulated by Besser for digital libraries: *to deliver information to multiple clienteles, using the same collection to serve many different groups of users, each with its own level of knowledge and modality of learning and interacting*. [Besser] The obstacles have technical and political dimensions, and Cyberinfrastructure has the potential to address both.

Technically, the KL should include one or more frameworks explicitly designed to support data mining within selected segments of the Internet, utilizing content, metadata, and an increasingly rich array of contextual information, such as links, citations, and usage data by audience-type; this might be realized by developing and operating one or more large-scale data warehouses that place particular emphasis—beyond characterizing individual entities and proxies—on relationships among these entities, relationships that are themselves determined by numerous diffuse and diverse processes, with and without human mediation.

Underpinnings for such warehousing—which need both advancement and widespread deployment—include unique-identifiers, nationwide authentication of users (with anonymity protections), high-level semantic markup and associated registries, and representations to deal effectively with programmatic services (and with entities that may be accessed only via such services).

To ensure success, the KL component must deal with the political reality that some degree of standards enforcement may be required. In other words, it may be wise to make Cyberinfrastructure or KL grants contingent upon adherence to some (small) set of protocol and interface standards, such as the warehouse underpinnings described above. Explicit linking of KL to the emerging NSDL core integration system may be part of an appropriate level of standardization. To ensure scalability, dependability, and persistence of resources, the KL architecture also must enable distributed replication and synchronization of entities, including services. This is of course a natural aspect of the Grid piece of the Cyberinfrastructure, but carrots and sticks may be needed to achieve compatibilities on the scales that are required.

Engineering, Operations and Evaluation

Good software engineering, software support, and reliable computer operations are difficult, and often they are expensive. They also are important, as they determine the end-user experience, but such characteristics often are given little support in NSF grants (because achieving them is not considered research). The KL artifacts will have to serve very large audiences dependably, on multiple platforms, and they must be well supported in response to user problems and needs for incremental enhancement. This suggests ample funding for organizations that are committed to both engineering excellence and software support, focused on needed tools and representations. Explicit mention of contemporary approaches to software engineering and support, especially user-centered design and development, ought to be part of the solicitation.

An important means for gaining an appropriate emphasis on customer service, dependability and end-user effectiveness is foster a *culture of evaluation* throughout the KL component. Solicitations should encourage both summative and formative evaluation efforts. A difficult but worthy objective is to embed, within digital libraries and other KL outcomes, instruments for *observing the learning process in relation to resources employed* [Ramaley]. Doing so would

help the Cyberinfrastructure initiative embrace large-scale education enhancement as a first class goal.

Conclusion

All of the above is worth doing, as it will yield a common philosophical and technical fabric for coordinating and using independent digital libraries. [Frew] This would be an important achievement and represent a particularly valuable component of the Cyberinfrastructure initiative.