Roadblocks

Herbert Van de Sompel
Los Alamos National Laboratory, Research Library
Many thanks to Rick Luce for sharing his insights

Introduction

I have no problems with admitting that thinking about a 10-year timeframe poses significant challenges for me. Going through the simple thought experiment of moving myself back in time to 1993, and using that perspective to try and figure which components of the current information infrastructure I would have been able to predict, or which level of global penetration each of those would be able to achieve, leave me rather speechless. Yes, after having seen an early glimpse of the Web through the lens of a Mosaic browser made me understand intuitively that this was going to be important. Real important, as opposed to how, for instance, Gopher was "interesting". But predicting that the Web along with then unknown related technologies would transform our world and world-view in general, as well as everything we did in libraries both traditional, digital and hybrid specifically was way beyond my capabilities as a futurist. Therefore, I have chosen not to directly address the future DL research agenda, but rather, I have concentrated on trying to identify roadblocks in the DL arena that I feel need to be abolished in order to pave the way for the vision expressed in the Cyber infrastructure Report. The items I identify are situated at the level of basic *plumbing* - things that need to be in place so that other great things can happen. Therefore, they may not come across as all that exciting or innovative, however, I feel that failing to address them may seriously impede efforts to move towards an integrated scholarly knowledge environment, as envisioned in the Cyber infrastructure Report. Most issues I identify pose enormous challenges, but they are not merely technical research challenges. Rather, they cover a broad range of research and organizational challenges at both the technical, legal, economical and social level. I present them in no particular order and hope that, when combined with ideas from other participants, the thoughts communicated hereafter will constitute a contribution to the goals of the meeting.

Some perceived roadblocks

(1) DL Theory - The field exists for over a decade, and has produced significant advances. Yet, it seems to me that further advancement may actually be hindered by the lack of an accepted base-line *Theory* for Digital Libraries. I am hereby thinking of a specification such as the one describing the OAIS model, which has significantly enhanced our ability to communicate about, identify research areas, and do actual research in the realm of digital preservation. Establishing such a *Theory* for Digital Libraries could have similar effects in our domain, and might help to escape from the inertia in research that is perceived by some observers. A concrete example might help to emphasize the issue at stake: could it be that the reason that the DL field has so far delivered only one widely agreed upon protocol (the OAI-PMH) is related to the fact that we have not yet agreed upon and named the core components of a distributed DL environment and hence are unable to start defining protocols for their interaction?

(2) DL Context - The DL field largely builds upon Web technology and infrastructure [*which obviously is a sensible thing to do if only in light of sustainability*], yet it does not seem to have a significant impact on their evolution. This seems to be true at the level of applications/services where very few DL research results have been able to transition to a product phase, and even less have evolved into sustainable offerings, commercial or otherwise [*Google being a very prominent counter-example*]. The lack of impact of the DL field also seems to exist at the level of defining essential building blocks for the evolution of the Web infrastructure. Again a concrete example may help: while the DL field agrees that unique persistent identifiers/names for objects/concepts/formats/etc. in DLs are crucial, the manner in which such identifiers can be integrated in the Web infrastructure does not seem to be informed by concerns of the DL community. I am unsure as to what the actual reasons for the perceived situation are, but I do believe the issue needs to be addressed if we want to make DLs an integral part of the Web, and the DL community a respected partner in efforts aimed at the further evolution of the Web.

A similar consideration can be made about the interaction of DL research with other domains that deal with conducting or communicating scientific endeavors using the digital network infrastructure. For instance, to which extent do relevant concepts and results from DL research penetrate into the realm of grid computing development, efforts dealing with mass storage of content/datasets, e-learning R&D, and vice versa? And how can we ensure that such efforts move in a common direction, which will eventually be essential if a fully integrated networked environment for scholarship is to emerge? As is the case with the aforementioned relationship to Web technologies, it seems essential that the DL community be an equal partner in efforts targeted at such integration.

Hence, the fundamental question is how the DL community can be moved into a position that provides adequate guarantees with respect to such interaction. While the suggestion might be totally off base, it occurs to me that in order to give the DL community a strong enough voice in such interactions, the creation of one or more centers of excellence [*like the ones the NSF has created for e.g. nanotechnology*] might be part of an answer.

(3) DL Ecology - The sustainability question is on everyone's lip and it is becoming obvious that it needs answering. Attempts to answer it should not only be about the identification of new business-models, but should include a fundamental questioning of the emerging DLs ecology that organizationally, to a large extent, is a direct transposition of the geographically constrained pre-digital era: a library is an island [*peninsula*] that provides each and every service. Shouldn't we question whether an ecology based on the provision of redundant services, as was necessarily deployed by traditional libraries, remains optimal in a networked environment? Could we not, for example, think of an ecology based on distributed service provision: nodes specialize in specific tasks and exchange their services for those of nodes with other specializations? The OAI-PMH model is a very humble illustration of the idea: some nodes [*data-providers*] focus on storing an exposing metadata, others [*service-providers*] focus on doing something relevant with it. Others models that are conceptually related are the LoCKSS project, the QuestionPoint Collaborative Reference Service, the Seti project, etc. These projects urge us to think about the feasibility of a different ecology, in which a global network of

interacting nodes becomes a robust and persistent global DL that provides a wide range of services.

(4) DL Rights Framework - The thoughts introduced in the above section can be extended to content residing in the DL Ecology: it is high time for serious efforts aimed at deploying an environment in which scholarly assets (technical reports, preprints, research papers, datasets, simulations, metadata records, etc.) behave in a manner that more closely matches the "gift exchange" spirit of scholarship. [*Keep reading, as this is not yet another "ditch Elsevier" serenade.*] As James Boyle brilliantly advocated in his JCDL 2003 keynote, it is time to think about this problem from the perspective of what we actually loose by sticking with the current paradigm? That loss is far from being purely financial [*although one may hardly think about the joint budgets for journal acquisitions of research libraries worldwide, and what could be done with the budget of just 1 year*]. The loss is also about the enormous constraints on the allowed usage of such assets. [*Again, keep reading, because this is not about the legendary professor not being allowed to make 20 copies of her own paper.*] Rather this is about the inability to freely process scientific assets in order to extract knowledge from them, attach knowledge to them, mine them, evolve them, build on them, etc. This is also about the complexities that we are forced to build into our systems in order to accommodate the existing rights framework. We need to think beyond research papers [*although the situation is severe enough there*], and anticipate that datasets, simulations etc. are endangered to go down the same path, and actually in some lucrative disciplines already are. Considering that we cannot change the past, we must understand that it is never too late for a new start, especially in light of the enormous annual growth of scholarly assets. .

(5) *Binding* Scholarly Assets - I perceive a serious shortcoming in the existing scholarly communication mechanism, which I need to explain by a very simple example:

At a certain point, a scholarly paper makes its public appearance in the system as an electronic preprint. Next, it gets peer-reviewed and published in a journal. Then some A&I database providers publish a metadata record describing the paper. Some scholars read the paper, build on it and hence cite it.

Unfortunately, the scholarly system does not record an unambiguous trace of these actions nor of their nature. This is actually true of most value chains that scholarly assets go through: there is no unambiguous, recorded and visible trace of the evolution of a scholarly asset through the system, nor of the nature of the evolution.

The results of this are quite disturbing, and I will illustrate a specific instance of the problem using the simple example:

Services need to go through enormous pains to computationally derive relationships between the preprint, the journal publication, the metadata records, and the citations after the facts. For example, in order to understand that the publishers' metadata, the A&I metadata, and the metadata contained in the citations refer to the same work, complex, and computation-expensive processes need to be run, which actually rarely lead

to perfect results, and which only a few services can actually perform because of the access rights required to do so.

Through the above example, the problem can be misread to be one of computing power, algorithms and access rights. It actually is not. The problem is that relationships, which are known at the moment a scholarly asset goes through a step in a value chain, are lost the immediate moment thereafter, and in many cases are forever lost. The actual dynamics of scholarship, the interaction/connection between assets, authors, readers, quality assessments about assets, scholarly research areas, etc. are lost and are extremely hard to recover after the facts. A whole body of information that could be crucial for the development of knowledge is lost. Therefore, the establishment of a layer underlying scholarly communication – a *Grid* for scholarly communication - that records and exposes such dynamics, relationships, and interactions is desired. In that *Grid*, many types of arcs expressing many different value-chain-based relationships interconnect assets and actors. Such a *Grid* would be instrumental in the derivation of knowledge and understanding from the scholarly assets it binds. It would provide the essential infrastructure required for the extraction of new metrics to assess the quality of scholarly assets, their appropriateness in a given research context, and for the evaluation of the performance of actors in the scholarly system. Such metrics are crucial to avoid information-overload and to pave the way for a different scholarly communication system.

The citing author drags an icon representing the cited paper to his own draft; the author's software records metadata about the cited paper, and its "cited by" relationship with the draft and its author. The citing author posts his paper to the institutional repository, which extracts and exposes all contained metadata, both descriptive and relational. A robot operated by a publisher checks the institutional repository for additions, and selects the preprint to be considered for peer-review and publication. Around the same time, another robot operated by one of the organizations that deliver citation-based services harvests the exposed citation information, and adds it to its collection, keeping track of its origins. Another robot comes by to collect descriptive metadata to add to its scientific altering service. As alerted scholars start reading the preprint, the institutional repository records and exposes the usage – a (probably anonymous) "read by" relationship. A service that provides several quality metrics about scholarly papers collects the usage information. An archiving robot comes by to collect the preprint and all data related to it, archives it, and exposes the fact that it did so to the Grid …