

Toward a Science Extending Digital Libraries, with a Testbed of ETDs

by Edward A. Fox

**Position paper for
15-17 June 2003 NSF-sponsored
Workshop on Post-DL Research Directions**

Introduction:

JCR Licklider, who had been a highly productive researcher working on psycho-acoustics, and whose activities at APRA and MIT (where he directed Project MAC, sometimes viewed as supporting interactions involving “man and computers”) greatly aided the early development of the Internet, wrote “Libraries of the Future” in 1965 [8]. He noted that for the foreseeable future advances in computerization of libraries would be based on a loose coupling of a number of specialized subfields taken from the union of what now we might refer to as CS and LIS. However, he hoped that eventually we could move toward the ideal of an integrative theory to support such efforts.

Licklider might be proud of a successful and broad research program aimed toward a science extending digital libraries. I argue that we can and should develop such an integrative theory, over the next decade. Further, I argue that we should empirically validate such a theory with one or more testbeds which can be devised in a cost effective fashion, wherein almost all of “the defining characteristics of the problems to which a science of information management applies” can be addressed (e.g., [2]).

Addressing George Heilmeyer’s Questions:

We may describe such an initiative by answering George Heilmeyer’s questions, as follows:

- We are developing a science to extend digital libraries.
- We aim for an integrative theory to extend DL efforts, drawn from the union of CS and LIS. The scope includes “network and systems design, human computer interaction, artificial intelligence, information retrieval, information organization, machine translation, database systems, and complexity theory”. To accomplish this we must go beyond first steps in this direction [1, 6].
- Today there is no such integration; rather we have ad hoc combinations of approaches cobbled together through software engineering and human effort. (We have some nice systems, like Greenstone [10], but they are not theory-based.)
- The 5S (society, scenario, space, structure, stream) framework is a new approach that promises to help us build the desired integrative theory [4]. Based on developments at Virginia Tech for the last 5 years, and building on digital library activities funded by NSF since 1991, it has led to a number of doctoral and masters studies, many publications, a large number of presentations, and a good deal of supporting software.

None of our claims have been falsified, and many have been validated, so we believe this approach will be successful to help move us toward an integrative theory.

- We believe that 5S, and other efforts aimed to develop a science to extend digital libraries [1, 6], will have even greater impact than resulted from the development of a theory to underlie work on databases.
- The initial phase of our work on 5S should be complete in 2 years, providing a firm foundation for a much larger and broader effort that should unfold through at least 2013. We aim for the emergence of a theory-based field of DL and information systems, with many milestones and partial results, including, in the near term:
 - A partial / initial set of definitions, lemmas, and theorems for the core theory by the end of 2003 (see Fig. 1)
 - A complete minimalist set of definitions, lemmas, and theorems for the core theory by 2005
 - Version 2 of 5SL [3], an XML encoding that addresses key parts of 5S, and that allows full description of small but complete DLs, by early 2004
 - Version 2 of a metamodel for DLs, prepared using 5SL by DL experts, by the middle of 2004
 - Version 2 of 5SGraph [11], a graphical tool for DL designers (e.g., “digital librarians”), allowing them to specify a small but complete DL through interaction with a representation of the metamodel (see Fig. 2), resulting in 5SL encodings of the desired tailored DL, by the fall of 2004
 - Definitions, lemmas, and theorems that address key aspects of the question of quality, and which relate to practical metrics appropriate for characterizing DLs, by the end of 2004
 - Version 2 of 5SGen [7], a generic generator of tailored digital libraries, that takes the 5SL output produced by 5SGraph [11] (see Fig. 3), building upon open toolkits for building digital libraries from components [9], by the end of 2004
 - Version 3 of an XML format for DL logging, based on 5S, with supporting software and tools [5], by the end of 2004
- The cost for the 5S family of results, in a basic form, is that of funding a medium-size team that is well along in its efforts. Fully developing this effort would require support that has been requested in a medium ITR proposal, as well as support from other groups who would be willing to take up this challenge. It also would require validation in several testbeds, like the one discussed in the next session, where early testing of the theory has led to good initial results.

Testbed of ETDs

In 1987 we began to consider the question of how dissertations might be changed as a result of using new electronic publishing approaches like coding for SGML. By 1997, this led to international efforts coordinated by the Networked Digital Library of Theses and Dissertations [2]. Today there are over 180 members of NDLTD (some of which represent entire states or countries), and our recently held 6th international conference in Berlin had attendees from 46 countries. The resulting emerging DL of electronic theses and dissertations (ETDs) has the following desired characteristics:

- Heterogeneity of information systems and sources (Dienst, Eprints, MARIAN, ODL, OCLC systems, Virtua, etc. – from individual universities, national libraries, companies, etc.)
- Heterogeneity of users and information providers (researchers, students, faculty, etc. – and all of the sources listed above)
- Seemingly unbounded scale of data and users (ultimately, the works of the more than 200,000 students who complete a thesis or dissertation each year, each a small book, which gradually will shift to all include color graphics, multimedia content, and datasets)
- Need for preservation and records management for an indefinite future and into indeterminate future environments (since states like Virginia require preservation for at least 50 years, and many institutions and nations want these primary research records maintained and accessible indefinitely, in diverse locations)
- Evolving legal and social frameworks (since intellectual property rights, laws, and practices regarding education continually change, and already are widely varying around the globe)
- Varying human and organizational capabilities and behavior (since students will always, according to disciplinary background, skills, and training, have varying awareness of constantly changing methods of electronic publishing and practices related to digital libraries)
- Curated information (since graduate programs, registrars, and librarians – as well as scholars supervising graduate research – have varying approaches to curation)
- Collaborative, synchronous processes (since theses and dissertations are developed in a collaborative fashion, where mentoring especially must be supported in a timely fashion)
- Both proactive and reactive uses (so that the results of graduate studies can be disseminated quickly to those interested in related topics, as well as discovered in a multilingual setting by those engaged in new research)
- Enormous range of granularity of data (since entries might be just bibliographic data – now numbering over 4 million, or full theses – varying in length from 10 – 1000 pages, or have large datasets or multimedia content – up to many gigabytes)
- Need for comprehensive metadata (since there is a long tradition of cataloging these works like books, as well as making them available through services like *Dissertation Abstracts*)
- Increasing emphasis on the semantics of and correlative relationships among data (since these works have long bibliographies, and thus fit into a large and complex citation network, and also relate to all other types of scholarship, including the various artifacts studied by architects, art historians, astronomers, medical practitioners, musicians, etc.)

What is particularly important to note about the emerging collection of what will be many millions of ETDs is that this is easily sustainable and will come almost for free! All graduate programs yield theses and dissertations, and so as long as such research continues, it is feasible for the resulting documents to be collected in a distributed fashion. It has been shown in scores of institutions that collecting ETDs saves money relative to paper-based processes, so this is self-funding. Further, whereas paper theses or

dissertations were almost never read, ETDs are often accessed hundreds or thousands of times per year. Now that many institutions are benefiting from this sharing of knowledge, including diverse institutions in developing nations, there appears to be no way to stop the current movement in this direction. Leveraging it to yield an exciting testbed for DL work seems to require a very small but highly worthwhile investment. Further, the whole process is open to flexible development, with little commercial pressure involved. It fits well with current trends toward institutional repositories, which is one of the goals that NDLTD began to advocate in 1995. Finally, it is distributed at its core.

Conclusion

We argue that work like that related to 5S, and testing of such efforts with testbeds like that being developed by NDLTD, are highly cost-effective, and very likely to move us directly in the direction suggested by the call for this workshop.

References

1. D. Castelli and P. Pagano, "OpenDLib: A Digital Library Service System," ECDL 2002, Rome, Italy, 2002
2. E. Fox, "NDLTD: Networked Digital Library of Theses and Dissertations", 2000. <http://www.ndltd.org>
3. M. A. Goncalves and E. A. Fox, "5SL -- A Language for Declarative Specification and Generation of Digital Libraries," in *Proceedings JCDL'2002*, G. Marchionini, Ed. Portland, OR: ACM, 2002.
4. M. A. Goncalves, E. A. Fox, L. T. Watson, and N. A. Kipp, "Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries," Virginia Tech, Department of Computer Science TR-03-04, 2003. <http://eprints.cs.vt.edu:8000/archive/00000646/> (submitted for publication)
5. M. Goncalves, G. Panchanathan, U. Ravindranathan, A. Krowne, E. A. Fox, F. Jagodzinski, and L. Cassel, "The XML Log Standard for Digital Libraries: Analysis, Evolution, and Deployment," presented at Third Joint Conference in Digital Libraries, Houston, Texas, 2003.
6. L. A. Kalinichenko, N. A. Skvortsov, D. O. Briukhov, D. V. Kravchenko, and I. A. Chaban, "Designing Personalized Digital Libraries," *Programming and Computer Software*, vol. 26, pp. 123-133, 2000.
7. Rohit Kelapure, "Scenario-based Generation of Digital Libraries", Virginia Tech CS Dept., Masters Thesis, June 2003 (see from <http://scholar.lib.vt.edu/theses>)
8. J. C. R. Licklider, *Libraries of the Future*. Cambridge, MA: MIT Press, 1965.
9. H. Suleman, "Open Digital Libraries," Virginia Tech Department of Computer Science, Blacksburg, Ph. D. Dissertation, 2002. <http://scholar.lib.vt.edu/theses/available/etd-11222002-55624/>
10. I. H. Witten, R. J. McNab, S. J. Boddie, and D. Bainbridge, "Greenstone: A Comprehensive Open-Source Digital Library Software System," in *Proceedings of the Fifth ACM Conference on Digital Libraries: DL '00, June 2-7, 2000, San Antonio, TX*. New York: ACM Press, 2000, pp. 113-121.

11. Q. Zhu, "5SGraph: A Modeling Tool for Digital Libraries," Virginia Tech CS Dept., Masters Thesis, 2002. <http://scholar.lib.vt.edu/theses/available/etd-11272002-210531/>

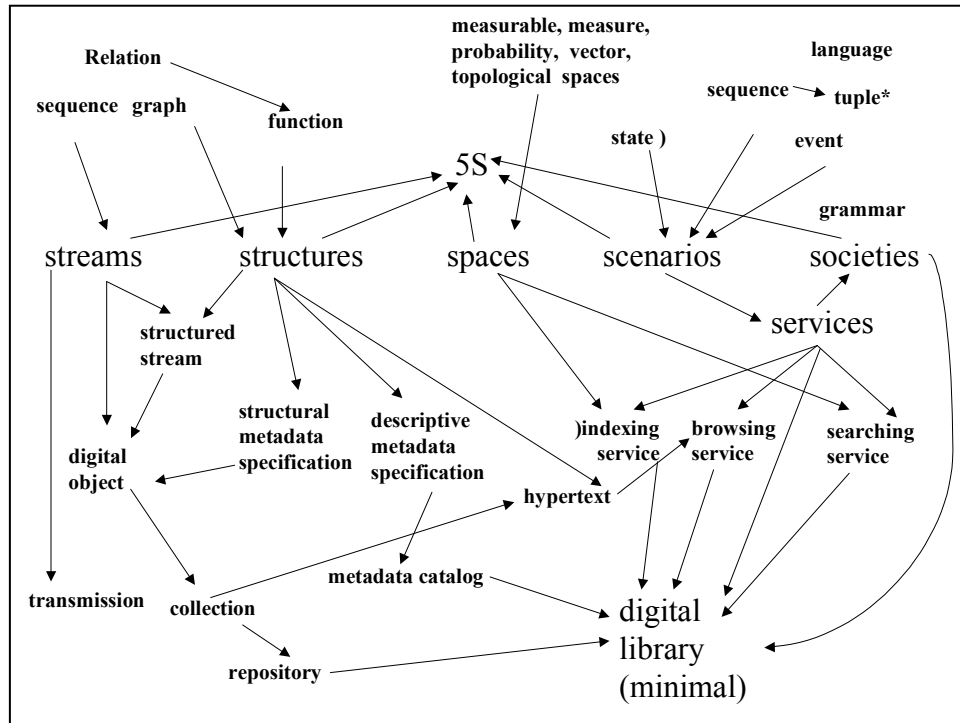


Figure 1. Overview of 5S and DL formal definitions and compositions, taken from [4]

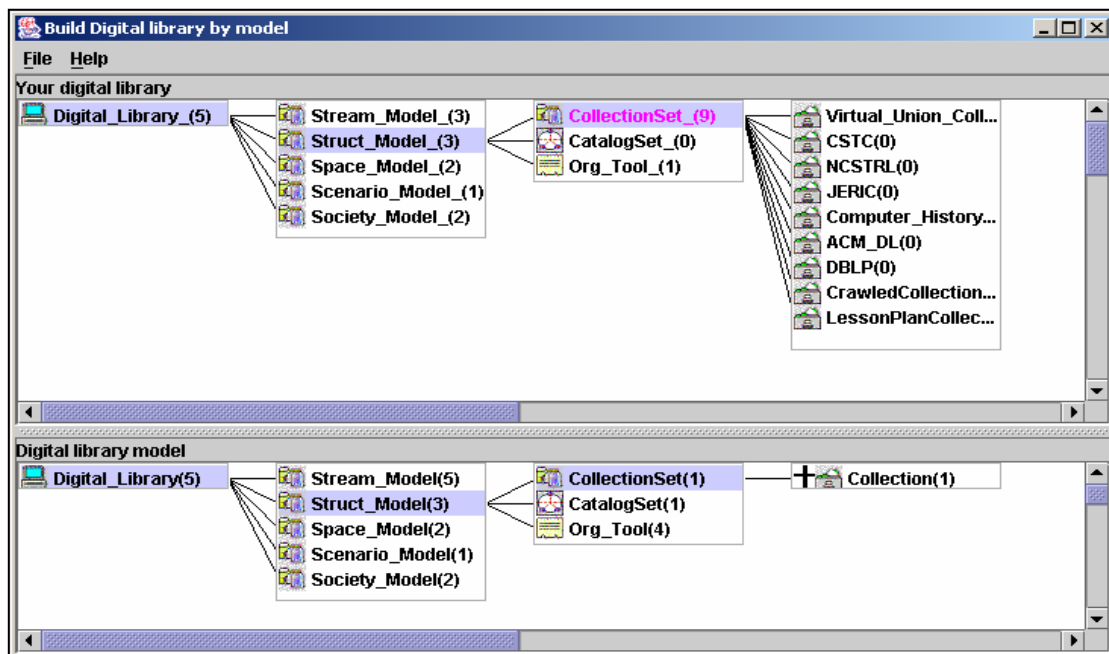


Figure 2. User Interface of 5SGraph, taken from [11]

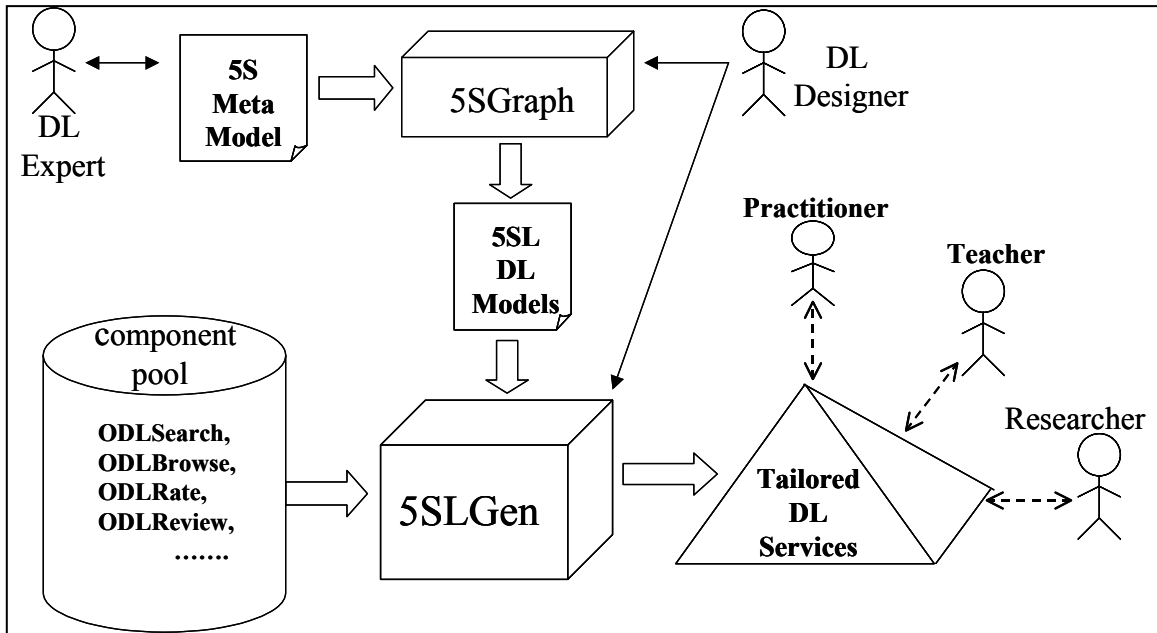


Figure 3. Overview of the architecture for DL modeling and generation, taken from [7]