Periodical Report for SHEFC Funded Research Equipment Initiative Project

From April to December 2002

Developing Terabyte Electronic Infrastructure

Part One:

Cases for Support of Research Strategic Adaptation

Ying Liu

Computing Science Department
The University of Aberdeen

Summary

The study to be reported here is funded by SHEFC under the research equipment initiative. The main objective is to develop *a globally shared digital resources infrastructure* that supports other research projects proposed at the time. This report concludes that

1) Lessons can be learned from ADL (Alexandra Digital Library) and innovations can be inspired by ADEPT (Alexandria Digital Earth Prototype). ADL has now been evolved from "Web driven HTTP client−server/middleware geo−referenced" architecture to ADEPT − "peer to peer network−centric distributing node" infrastructure. This has broadened wider collaborations that Santa Barbara already has. Some partners are newly established in 2002.

2) Aberdeen digital infrastructural development has suffered from two drawbacks: a) other collaborative projects originally proposed are left with no funding, and b) Santa Barbara collaboration with Aberdeen is virtually non−existent.

3) ADL itself has gone through substantial changes and has severe limitations in many aspects. The project must not only be based on the intention for establishing "the first European mirror site of ADL".

4) ADL middleware server download pages have been found, thanks to Google, and the codes are in our machine. What we have been informed about the unavailability of ADL copy is not entirely true.

How to adapt to these changes, to cope with the drawbacks, to take our hardware competitive advantages, and to lead our research in Aberdeen to be productive at present and beneficial for the very near future will be suggested in the second part of this report separately.

1 Introduction

The aim to develop a digital library is not different from the purpose of having a traditional library:

- both make a provision of information to users;
- both are the means for human communication by organizing publications;
- both depend for their very existence on the needs for the communication of knowledge in a social community [Dretske].

What makes them different, however, is that traditional libraries collect printed matters [Kemp]; While digital libraries collect objects only existed in digital forms which need greater computing storage capacity, are less structured and therefore more complex in all the operations being required [Chen] [Greenstein]. Digital libraries are flooded on information highways, but good geo−referenced ones can be hardly seen. A lack of digital content based technology is one of many to blame, so is the infrastructural technology that is unable to collect and index the contents online in the first place. "The (data) network is (still) the least structured organization that can be said to have any structure at all" [Kelly].

This project of research equipment initiative is funded for developing a globally shared digital resources infrastructure to enable geo−referenced digital library services. Specifically it is proposed to creating a generic infrastructure rather than on the applications supported by it. The project has brought contribution and innovation into Aberdeen Scotland. This will be stated in section 2.

Instead of focusing on several infrastructural issues [Liu], the project has been led to establish the first European mirror site of ADL system (Alexandria Digital) in developing services for ADL. ADL was a well known system − a product out of the six digital library projects funded by NSF, DARPA, and NASA in 1995; Its collection and services focus on geographical information: maps, images, geo−referenced data sets with text, and other information sources with links to geographic locations. Now ADL has been evolved to be ADEPT (Alexandria Digital Earth Prototype) by the second phase of the development (from 1999 to 2004) expanding ADL−use−abilities into new fields, e.g., classroom based geo−referencing e−learning applications. Because of the technical uncertainties in ADL evolution and absence of US collaboration, an effort must be made in searching for an approach to adapt to the changes in the coming year. These issues will be addressed in section 3.

Recommendations for developing a terabyte infrastructure will be given in the Part Two separately.

2 Contribution

2.1 Exceptional good valued terabyte machine

The research project has brought the generous support of SUN Microsystems, who, through a combination of a 50% contribution on top of a 30% academic discount, have effectively offered almost £1M worth of hardware − the terabyte machine at not much more than a third of the standard price. If our research strategy centers on the development of an electronic infrastructure, a kind of support environment is firstly much needed for new scientific research collaborations as well as for many far reaching educational implications.

2.2 A provision for overall hardware operational policy

The research project has funded a database/system administrator/researcher. The senior system manager has been working together with the researcher in initiating system administrative policies and regulations necessary to manage and operate the terabyte machine systematically. This is the first step forward to clarifying that the senior system manager is responsible for overall system security, users' authorization, service approvals, and hardware configuration; that the researcher is responsible for overall solution architectural operations, recommendations, request changes, and service maintenance.

2.3 ADL information discovery

A great deal of time and effort have been put into surfing the net. It has been found that the unavailability of a ADL copy as we have been told is not entirely true.

In the October, the downloading page has been found in [Janee1], thanks to Google. A version of ADL middleware server has now been downloaded into our terabyte machine. Although this unknown version of ADL has its severe functional limitations which is to be addressed next, it may still be worthwhile further exploring.

3 ADL and ADEPT

Santa Barbara digital library project (Santa Barbara) has approved to be very intensive in terms of finance, human resources, and technology investment.

Santa Barbara has been evolved through the following two stages: a) Web technology driven client−server architecture for collections as well as services for accessing the collections, and b) Network centric peer−to−peer infrastructure for managing, distributing and integrating the services/collections.

3.1 1995 − 1999: Web technology driven architecture

ADL has three successive versions:

a) a rapid prototype system comprising a relational database of map and imagery metadata, accessed through a desktop GIS (geographic information system) [Frew1].

b) a Web prototype system which replaced the stand−alone GIS with an HTTP server, generating an HTML forms−based user interface accessible via Web [Few2].

c) ADL integrated system which extended the HTTP server into full−fledged middleware, supporting HTTP interfaces to multiple clients, and collections to multiple HTTP catalog databases [Frew3].

In June 1999, the architect of ADL [Janee2] pointed out that ADL requires substantial modifications. Here the problems are put into the following three categories.

S−Category: a service layer is completely missing (S)	
S−Tags−Problem	ADL services are too simplistic. Indeed, a search is one "shot" followed by a stream of result back; a query is limited to the capacity of what one XML protocol tags, because of the next problem.
S−Tags−Semantics−Problem	The semantics of all the metadata are both unknown and unknowable to ADL system.

Furthermore, Janee continues, the architecture does not address holding content at all. Here it is a belief that S−Tags−Semantics−Problem is not entirely responsible for not holding the content, because more other problems described below are significantly inter−related to the holding function too.

O−Category: no inter−operations between content holding resources (O)	
O−Metadata−Content−Problem	Metadata are formulated after the contents are identified by URLs, thus, (see below)
S−Metadata−Content−**O**−Finding−Problem	**Servicing** the content behind those URLs is out of the question.

Note, again, these two problems are not only inter−related, but also external−related to the service layer. Since RDF has something to do with URLs, can these two problems be solved by exploring RDF? Yes and No.

I−Category: no integration with other tools and services (I)	
I−External−**S**−Plug−in−**O**−Problem	ADL' s single, monolithic components support no operations with other tools and services. There is no metadata standard adaptation, no service layer, and therefore no component to be plugged either in or out. Is RDF good enough to represent this and take care of it?

Note, all the above problems have been viewed from an infrastructural perspective. There is no single methodological mechanism that can provide the solutions; Just like no single system can provide all the answers. Because of this, ADL has now embarked on an adaptation of controlling and navigating components.

3.2 1999 − 2004: Moving to network−centric peer−to−peer distributed−autonomous nodes

The focus of ADEPT has now broadened from a geo−spatial library to an integrated environment for managing, querying, and presenting geo−spatial content. While querying and presenting services may not appear to be new, but the controllers and navigators are.

Designing these system managerial components is not based on any theoretical principles, which they never seem to do. Rather, the design is innovated by novel technologies for network−data−management [Janee3].

Solutions (S: service layer, O: operation, I: external integration, SYS: internal integration)	
S−**O**−Repository−Component	collection registry content repository
O−Metadata−To−Metadata−Component	metadata mapper harvest loader item tracker
SYS−**S**−**O**−Component	collection aggregator

.

Although we still need to foresee how these components are developed without any algorithms, they have narrowed their original problematic scope and been on the cutting edge of knowledge technology integration

These ADEPT infrastructural components are likely to provide many opportunities to adapt knowledge organization systems into digital libraries, see, e.g., [Hill], if the components are in the "right places".

Also, since ADEPT may well enrich the services, collaboration with new research partners has also been established to develop new applications. For example, The Kent State University is developing structured models of scientific concepts for organizing, accessing, and using learning materials that ADEPT will provide [Smith]. ADEPT will be able to "plug−in" lscapes (information landscapes) that The University of Georgia is developing, where ontological environments are based on the Earth metaphor lscapes [lsdis].

3.3 Drawbacks in Aberdeen

Aberdeen digital library research was initiated in May 2000 to extend the DNER (distributed national electronic resource) in the area of geo−referenced data, and to offer the potential additional benefits by adapting substantial relevant technical developments from American digital library research. To achieve this, the complexity of establishing a geo−referenced digital library has been recognized. The needs for transferring ADL technologies, developments in areas of geo−referenced databases, dynamic visualization, archiving methods, and landscape image processing were well proposed. The overall research strategy was, and still is, the focus for establishing significant international collaboration between Aberdeen/(Scotland) and Santa Barbra (California) in the creation and standardization of geo−referenced information sources and services on the Web.

However the collaborative projects proposed have been left with no funds. The principal investigator decided in May 2002 to be permanently absent in this project because of his illness. Since then, the collaboration with Santa Barbra has virtually been non−existent.

References

[Chen] Chen, S−S., "Digital Libraries − The Life Cycle of Information", BE (Better Earth) Publisher, 1998.

[Dretske] Dretske, F. I., "Knowledge and the Flow of Information", Basil Blackwell, Oxford, UK, 1981.

[Frew1] Frew, J., Carver, L., Fisher, C., Goodchild, M., Larsgaard, M., Smith, T. & Zheng, Q., "The Alexandria Rapid Prototype: Building a Digital Library for Spatial Information", In: ESRI International User Conference, Environmental Systems Research Institute, Inc, Palm Springs, 1995.

[Frew2] Frew, J., Freeston, M., Kemp, R., Simpson, J., Smith, T., Wells, A. & Zheng, Q., "The Alexandria Digital Library Testbed", D−Lib Magazine, 2 (7/8), 1997.

[Frew3] Frew, J., Freeston, M., Freitas, N., Hill, L., Janee, G., Lovette, K., Nideffer, R., Smith, T. & Zheng, Q., "The Alexandria Digital Library Architecture", International Journal on Digital Libraries, 2 (4), p 259−268, 1999.

[Greenstein] Greenstein, D. & Thorin, S. E., "The Digital Library: A Biography", Report, Digital Library Federation and Council on Library and Information Resources, USA, September 2002, http://www.clir.org/pubs/reports/pub109/contents.html

[Hill] Hill, L, Buchel, O., Janee, G. & Zeng, M. L., "Integration of Knowledge Organization Systems into Digital Library Architectures", In: "Re−conceptualizing Classification research", 13th ASISS SIG/CR Workshop, http://alexandria.sdc.ucsb.edu/~gjanee/archive/2002/kos−dl−paper.doc

[Janee1] −, http://alexandria.sdc.ucsb.edu/~gjanee

[Janee2] −, "Current Architectures & Known Limitations (sign)", http://alexandria.sdc.ucsb.edu/~gjanee/presentation−junkyard.html

[Janee3] −, "The ADEPT Digital Library Architecture",
http://alexandria.sdc.ucsb.edu/~gjanee/archive/2002/jcdl−adapt.pdf

[Kelly] Kelly, K., "Out of Control: The New Biology of Machines, Social Systems and the Economic World", Reading, MA: Perseus, 1995.

[Kemp] Kemp, D. A., "The Nature of Knowledge − An Introduction for Librarians", Clive Bingley Ltd. London, UK, 1976.

[lsdis] − , "ADEPT−UGA Prototype Process", http://lsdis.cs.uga.edu/~adept/process.html

[Liu] − , http://www.csd.abdn.ac.uk/~yliu/presentation/research_issues.ppt

[Smith] Smith, T. R., Zeng, M. L. & ADEPT Knowledge Team, "Structured Models of Scientific Concepts for Organizing, Accessing, and Using Learning Materials",
http://esm.ucsb.edu/fac_staff/fac/frew/cv/pubs/2002_concepts.pdf