It Is the Time for the Digital Library to Meet the Enterprise Architecture

José Borbinha

IST – Instituto Superior Técnico / INESC-ID, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
jlb@ist.utl.pt

Abstract. The purpose of this paper is to raise arguments to support the proposal that we should promote the discussion of the Digital Library in a structured way, aligned with the emerging perspective of the Enterprise Architecture. In this sense, the Digital Library practitioners should be motivated to give more emphasis to the need to better integrate its efforts and body of knowledge with the more generic area of Information Systems, where important concepts, regulations and good practices have been emerging, defined by authorities, the industry and the multiple stockholders of each specific scenario. Concluding, it is time for the Digital Library to mature by recognizing that it is, simply, a case of an Information System, which is specific only in what concerns the requirements derived of its specific business goals.

1 Introduction

The title and motivation for this paper was inspired by [4]. The content was also inspired by [1]. In his paper Michael Lesk was himself inspired by the seven ages of man, described by Shakespeare, giving us that way a very interesting description of the evolution of the area of Information Retrieval. However, after a careful reading we can recognize that the scope of this description covers much more than the traditional area of Information Retrieval, also comprising the area of the Digital Library (DL).

Lesk's paper was written in 1995, on the same time the D-Lib magazine was debuting[1], and was precisely in the first issue of D-Lib that William Arms expressed his eight key general principles for a generic DL architecture.

I propose now to revisit these two works, twelve years after their first publication, with two main purposes in mind: to review their contents at the light of our actual knowledge; to use that effort as a process to try to characterize the actual thinking of the DL as a problem and the main emerging related challenges. The ultimate goal is to raise arguments to prove that, from now, we should not continue promoting the DL by mainly raising generic goals and addressing the technological related issues. Alternatively, the DL community should be motivated to better structure its goals and give more attention to the need to integrate its efforts and body of knowledge with the more generic area of Information Systems, where important concepts have been emerging recently that must not be ignored. Specifically, those are the cases of the concepts of Enterprise Architecture and Enterprise Architecture Framework.

[1] http://www.dlib.org

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 176–185, 2007.

But why is this really important? First, let us develop a simple analysis…

One can conceive "DL deployments" in mainly two scenarios: as a purpose in itself (the DL as the main business goal); or as a contribution to other purposes (technology and processes created from a "DL perspective" in order to be used to support more generic goals). The first scenario will continue sustaining the DL has a relevant concept, where it might be not too difficult to acknowledge the right credits to the right communities contributing for that. It also might be possible to assure that relevance and credits in the second scenario (making the acronym DL[2] equivalent to others such as ERP, CRM, SCM, etc.), but in any of the cases the DL community has to make it happen.

The need to rationalize resources, to apply standard governance's models and business processes, as also the need to accomplish with strict legal and auditing requirements, have been pushing governments and private organizations to promote and impose Enterprise Architecture Frameworks to central administration services, public services and enterprises in general[3,4]. Assuming that DL's technology has reached a maturity for formal deployments at these levels, than those specific requirements concerning management, legal and business issues, and especially concerning accountability, can not be ignored.

2 "Key Concepts in the Architecture of the Digital Library"

Arms' presents eight general principles representing concepts and requirements for the DL architecture. Quoting them in short:

1. **The technical framework exists within a legal and social framework:** "Early networked information systems were developed by technical and professional communities, concentrating on their own needs. The emphasis was on making information available (…) without charge. The digital library of the future will exist within a much larger economic, social and legal framework. (…)"
2. **Understanding of digital library concepts is hampered by terminology:** "(…) Certain words cause such misunderstandings that they are best expunged from any precise discussion of the digital library. The list includes "copy", "publish", "document", and "work". Other words have to be used very carefully and their exact meaning made clear whenever they are used. An example is "content". (…)"
3. **The underlying architecture should be separate from the content stored in the library:** "Separating general functions from those specific to the type of content has other benefits. It encourages different markets to emerge, and allows a legal framework in which storage, transmission and delivery of digital objects is separate from activities to create and manage the intellectual content."

---

[2] DL – Digital Library; ERP – Enterprise Resource Planning; CRM – Costumer Relationship Management; SCM – Supply Chain Management.

[3] "Congress is enforcing its mandate that the Defense Department develop systems compatible with the DOD Business Enterprise Architecture - with the threat of jail time and hefty fines for the department's comptroller." - http://www.gcn.com/print/23_33/27950-1.html?topic=enterprise-architecture

[4] http://www.dmreview.com/article_sub.cfm?articleId=1038091 (Zachman, Basel II and Sarbanes-Oxley).

4. **Names and identifiers are the basic building block for the digital library:** "Names are a vital building block for the digital library. Names are needed to identify digital objects, to register intellectual property in digital objects, and to record changes of ownership. They are required for citations, for information retrieval, and are used for links between objects."
5. **Digital library objects are more than collections of bits:** "A primitive idea of a digital object is that it is just a set of bits, but this idea is too simple. The content of even the most basic digital object has some structure, and information, such as intellectual property rights (…)."
6. **The digital library object that is used is different from the stored object:** "The architecture must distinguish carefully between digital objects as they are created by an originator, digital objects stored in a repository, and digital objects as disseminated to a user."
7. **Repositories must look after the information they hold:** "Since digital objects contain valuable intellectual property, the stored form of a digital object within the repository includes information that allows for it to be managed within economic and social frameworks."
8. **Users want intellectual works, not digital objects:** "Which digital objects should be grouped together can not be specified in a few dogmatic rules. (…) The underlying architecture (…) must provide methods for grouping digital library objects and must provide means for retrieval."

3 "The Seven Ages of Information Retrieval"

Lesk provides an historical description and a vision of the future of the area of Information Retrieval that makes it clearly coincident with the DL.

According to Lesk, **Childhood** (1945-1955) is described as the time when Vannevar Bush proposed is vision for the Memex [2]. The **Schoolboy** (1960s) "…were a time of great experimentation in information retrieval systems". **Adulthood** (1970s) was when "…retrieval began to mature into real systems". **Maturity** (1980s) was reached with "…the steady increase in word processing and the steady decrease in the price of disk space... The use of online information retrieval expanded". Lesk wrote his paper during the **Mid-Life Crisis** (1990s), when "Things seemed to be progressing well: more and more text was available online, it was retrieved by full-text search algorithms, and end-users were using OPACs. (…) Nevertheless it was still an area primarily of interest to specialists in libraries".

After this, it was supposed to come the time for **Fulfillment** (2000s): "Which will it be? I believe that in this decade we will see not just Bush's goal of a 1M book library, so that most ordinary questions can be answered by reference to online materials rather than paper materials, but also the routine offering of new books online, and the routine retrospective conversion of library collections. We will also have enough guidance companies on the Web to satisfy anyone, so that the lack of any fundamental advances in knowledge organization will not matter". Accordingly, **Retirement** (2010) is the age when "…central library buildings on campus have been reclaimed for other uses, as students access all the works they need from dormitory room computer. (…) Most students, faced with a choice between reading a book and watching a

TV program on a subject, will watch the TV program. (…) Educators will probably bemoan this process. (…). As for the researchers, there will be engineering work in improving the systems, and there will be applications research as we learn new ways to use our new systems."

4 The Age of the Digital Library

At a first glance one might be tempted to consider the DL not as a continuum or a specialization of the area of Information Retrieval, but a child of it. This might be an argument for those willing to "reset" Lesk's scale of time, probably in order to give a "second live", or a "second chance" for the DL. I must stress that I disagree of that!

In my opinion, Lesk uses a description of the area of Information Retrieval that really makes it overlap the DL, and his vision is correct. Also, this includes not only the direct references to goals and processes easily identified with that, but also the multiple references to border areas, such as Artificial Intelligence. Lesk is rally talking about the same body of motivations and goals than we have been using as a reference for the DL! In this sense, the DL should be now in its fulfillment age! We have the "Million Books Project"[5]; reference works are common to find as e-books; Yahoo, Google, del.icio.us are fairly well guiding us in the labyrinth of the Web, etc. Z39.50[6], once a specific answer to specific requirements for technical interoperability from specific DL business goals has become irrelevant after the emerging of the web-OPAC, which in itself is disappearing, integrated in the "enterprise portal" or replaced by new processes based on the OAI-PMH[7]. Concerning semantic interoperability, one other common issue in Digital Libraries, it also is a common issue in most of the attempts to integrate businesses and processes among any different organizations. The concept of metadata registries, also usually raised by the DL, started in fact in the industry, due to very practical and generic needs. In fact, since the emerging of HTTP, XML, web-services (whenever they are based on SOAP or simply on REST), etc., that we can not claim anymore any key challenges for technical or semantic interoperability to be specific to the DL. They are simply generic issues in any kind of Information System!

Also automatic indexing, metadata extraction and "knowledge organization" in general are meeting the "traditional" corporate information systems area, trough the vital role played nowadays in any organization by document management systems, enterprise content management (the digital content as asset), and the dematerialization of the processes in general. In those scenarios, the "digital object" is not the exception anymore, but the rule, so even once DL (and archives) very specific issues such as digital preservation have been emerging as a normal concern in any organization, as historical information is making less sense, since all the information available is now critical for any good business governance ("archives" are now "repositories").

Aligned with this tendency, even the roles are changing. And in fact Lesk closes his paper with this very interesting paragraph:

[5] http://www.archive.org/details/millionbooks
[6] http://www.loc.gov/z3950/agency/
[7] http://www.openarchives.org/OAI/openarchivesprotocol.html

> *"Will, in a future world of online information, the job of organizing information have higher status, whatever it is called? I am optimistic about this, by analogy with accountancy. Once upon a time accountants were thought of as people who were good at arithmetic. Nowadays calculators and computers have made arithmetical skill irrelevant; does this mean that accountants are unimportant? As we all know, the answer is the reverse and financial types are more likely to run corporations than before. So if computers make alphabetizing an irrelevant skill, this may well make librarians or their successors more important than before. If we think of information as a sea, the job of the librarian in the future will no longer be to provide the water, but to navigate the ship."*

Accordingly, we can finish this point by concluding that even if there are areas of competence that we can claim as specific to a vision of the DL, we should differentiate its relevance as discipline, with a specific body of knowledge, from the possible applications of that body of knowledge to solve problems in specific scenarios. I mean, I believe that from now the DL community will be not requested anymore to provide technology, but expertise and services. In fact, reviewing Arms' key concepts, we can claim that none of them are really specific to the DL, but instead generic goals, constraints, requirements or good practices that we can find in multiple other areas:

About business goals and business environment:
1. The technical framework exists within a legal and social framework…
7. Repositories must look after the information they hold…
8. Users want intellectual works, not digital objects…

About business concepts and business domain
2. Understanding of digital library concepts is hampered by terminology…
5. Digital library objects are more than collections of bits…

About information systems design and good practices
3. The underlying architecture should be separate from the content stored in the library…
4. Names and identifiers are the basic building block for the digital library…
6. The digital library object that is used is different from the stored object…

5 Enterprise Architecture

The ANSI/IEEE 1471-2000 standard [3] defines architecture as "the fundamental organization of a system, embodied in its components, their relationships to each other and the environment, and the principles governing its design and evolution." According to this, the Enterprise Architecture emerges to help organizations to understand and express their business, structure and processes. The term Enterprise Architecture has, on the same time, two meanings: on one side it is the term given to the map of and organization and the plan for its business and technology continuous change; on the other side it is also the term given to the process to govern all of that.

The ability to have detailed views, planning and analytical knowledge of a system are vital tools to address new unavoidable requirements associated with the Web,

Table 1. The Zachman Framework

View	What (Data)	How (Function)	Where (Network)	Who (People)	When (Time)	Why (Motivation)
Scope	Things important to the business	Processes the business performs	Locations the business operates	Organizations important to the business	Events significant to the business	Business goals/strategies
Business Model	e.g., Semantic Model	e.g., Business Process Model	e.g., Business Logistics System	e.g., Work Flow Model	e.g., Master Schedule	e.g., Business Plan
System Model	e.g., Logical Data Model	e.g., Application Architecture	e.g., Distributed System Architecture	e.g., Human Interface Architecture	e.g., Processing Structure	e.g., Business Rule Model
Technology Model	e.g., Physical Data Model	e.g., System Design	e.g., Technology Architecture	e.g., Presentation Architecture	e.g., Control Structure	e.g., Rule Design
Components	e.g., Data Definition	e.g., Program	e.g., Network Architecture	e.g., Security Architecture	e.g., Timing Definition	e.g., Rule Specification
Instances	e.g., Data	e.g., Function	e.g., Network	e.g., Organization	e.g., Schedule	e.g., Strategy

XML and the concept of Service Oriented Architecture (SOA) [8]. The most important keyword associated with this new scenario is "flexibility"! Under this, the design and development of information systems builds on a global view of the world in which services are assembled and reused to quickly adapt to new goals, business needs and tasks. This means that the configuration of a system might have to change at any moment, removing, adding or replacing services on the fly, in alignment with the new business requirements. This is what Enterprise Architecture provides.

5.1 Enterprise Architecture Framework

Considering that the ultimate goal of the DL is to be able to offer solutions to address problems properly, than we must recognize that such solutions must be always a mix of an organizational structure with the related set of activities and services. Therefore, we'll have an enterprise, in the sense of a business activity. Accepting that, than we should ask now how organizations (enterprises) in other business areas address their issues related to information, processes and technology. That is the scope of the area of Information Systems[8]. The purpose of an information system in an organization is to support processes, and not surprisingly, professionals dealing with that use methodologies, models and frameworks to address their activities.

An Enterprise Architecture framework is a communication tool to support the Enterprise Architecture process. It consists in a set of concepts that must be used to guide during that process. The first Enterprise Architecture framework, also the most

[8] We should remember that the ACM – Association for Computer Machinery, identifies the area of "Digital Libraries" in its classification system with the coding H.3.7, under "Information Storage and Retrieval" (class H.3) and "Information Systems" (class H), as it can be seen at http://www.acm.org/class/

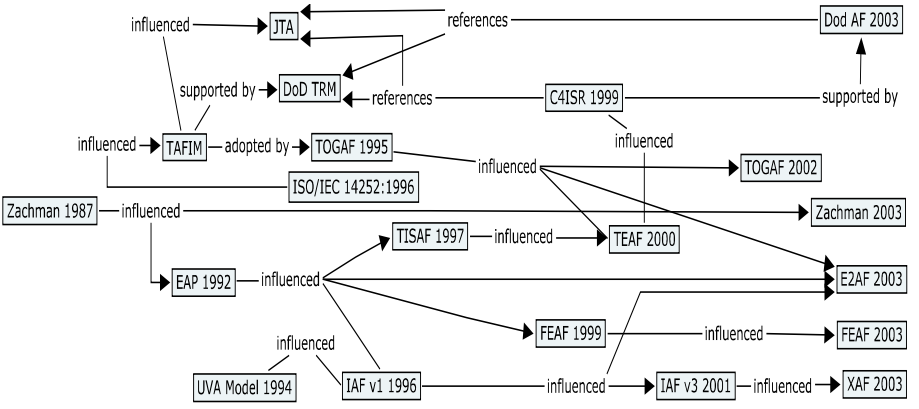

Fig. 1. The Enterprise Architecture Frameworks History Overview[9]

comprehensive and famous of them, is the Zachman framework[10], defined as "…a formal, highly structured, way of defining an enterprise's systems architecture. (…) to give a holistic view of the enterprise which is being modelled." the Zachman framework is resumed in simple terms in Table 1, where each cell can be related with a set of models, principles, services, standards, etc., whatever is needed to register and communicate its purpose. The meanings of the lines in this table are:

Scope (Contextual view; Planner) defined the business purpose and strategy; Business Model (Conceptual view; Owner) describes the organization, revealing which parts can be automated; System Model (Logical view; Designer) outline of how the system will satisfy the organization's information needs, independently of any specific technology or production constraints; Technology Model (Physical view; Builder) tells the system will be implemented, with the specific technology and ways to address production constraints; Components (Detailed view; Implementer) details each of the system elements that need clarification before production; Instances (Operational view; Worker) gives a view of the functioning system in its operational environment.

Concerning the meanings of the columns, What refers to the system's content, or data; How to the usage and functioning of the system, including processes and flows of control; Where to the spatial elements and their relationships; Who to the actors interacting with the system; When represents the timings of the processes; Why represents the overall motivation, with the option to express rules for constraints where important for the final purpose.

From this Framework many other Enterprise Architecture frameworks for specific areas have been developed. Those have been developed by research entities (such as E2A[11]), governmental bodies[12] (such as FEAF, TEAF, TOGAF, etc.) and private

[9] Redrawn from [6] (more details can be found in this reference).

[10] Originally conceived by John Zachman at IBM [9], this framework is now in the public domain, through the The Zachman Institute for Framework Advancement. For more details see http://www.zifa.com

[11] http://www.enterprise-architecture.info/ (Institute for Enterprise Architecture developments).

[12] http://www.eagov.com; http://www.eaframeworks.com/frameworks.htm; http://www.whitehouse.gov/omb/egov/a-1-fea.html

companies (such as the IAF[13], from Cap Gemini). The process has been also influenced by other related activities, as illustrated in the conceptual map in the Figure 1.

5.2 Enterprise Architecture and Governance

Enterprise Architecture is an instrument to manage the operations and future development in an organization. In this sense, in order to practice a correct Enterprise Architecture, planning and development must take in consideration the overall context of corporate and IT governance. This list of references expresses very well the complexity of the Enterprise Architecture process: Strategic Management: Balanced Scorecard[14]; Strategy Execution with EFQM[15]; Quality Management with ISO 9001[16]; IT Governance with COBIT[17]; IT Service Delivery and Support with ITIL[18]; IT Implementation with CMM[19] and CMMI[20].

6 The Goal of the Digital Library

How could we now define the goal of the DL? In my view, this simple statement might be enough to express that: **The goal of the DL is to provide access to selected intellectual works.** This goal comprises this way the three more generic (first level) business processes of the DL: Collection building; Discovery; Access.

We could express this goal with more words, but quite for sure that those would be redundant. We could also express this goal with more details, but quite for sure that such would be only a matter of specialization.

In fact, for a specific case second and other lower level processes must be identified, but these will depend of the specific context (the details of the "Scope" line in the Zachman Framework). For example, storage will be a requirement derived from access. Also the goal to provide access at any moment produces the requirement of preservation. In the same sense, registration is a requirement derived from discovery (to make it to be possible to find or be aware of a resource we produce requirements for cataloguing, indexing, descriptive metadata, etc.). Selectivity can be seen as a goal in itself, from which we can express relevant functional requirements (policies of collection building can be important in educational and professional libraries, in order to promote efficiency for the users), or it can be simply a consequence of a non-functional requirement associated to the fact that it might still be impossible, for a specific system, to provide discovery and access to everything produced by the focused organization (at least for now…).

---

[13] http://www.capgemini.com/services/soa/ent_architecture/iaf/

[14] http://www.balancedscorecard.org/ (The balanced scorecard management system).

[15] http://www.efqm.org/ (European Foundation for Quality Management excellence model).

[16] http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=21823 (ISO 9001: Quality management systems – Requirements).

[17] http://www.isaca.org/cobit/ (Control Objectives for Information and related Technology standard).

[18] http://www.itsmf.org/ (IT Infrastructure Library best practices).

[19] http://www.sei.cmu.edu/cmm/ (Capability Maturity Model for Software).

[20] http://www.sei.cmu.edu/cmmi/ (Capability Maturity Model Integration).

7 Conclusions

Concluding, the DL community must prepare itself for a dignified retirement age by moving its established knowledge from research to engineering, in order to take part in more generic goals[21].

A framework can be described as "a set of assumptions, concepts, values, and practices that constitutes a way of viewing the current environment" [5]. Frameworks can be used as basic conceptual structures to solve complex issues. Concluding, and in alignment with the vision already expressed by the DLF Service Framework Working Group[22], I think that the DL community should "get out of the box" and give more attention to the development of conceptual frameworks giving preference to scopes, goals requirements and processes, in the sense as those concepts are already common in Enterprise Architecture processes ([7] is a classic and stills one of the most cited reference for that purpose) and Enterprise Architecture Frameworks ([6] can be a very simple comprehensive reference for this).

What should it be the process for that and what kind or level of frameworks should we envisage for this work? As also described in [5], "a reference model is an abstract framework for understanding significant relationships among the entities of some environment that enables the development of specific architectures using consistent standards or specifications supporting that environment (…) and is independent of specific standards, technologies, implementations, or other concrete details". Still in [5], "a reference architecture is an architectural design pattern that indicates how an abstract set of mechanisms and relationships realizes a predetermined set of requirements".

Should we have reference models and reference architectures for the DL?

Maybe yes. Maybe it makes sense to develop such references for specific goals and processes, such as Digital Preservation, Institutional Repositories, etc.!

But also maybe not, or at least as some of us have been trying to do it, especially if we give credit to this external observer that wrote one[23]:

> *"A framework should be developed at a particularly high level, encompassing only the common and agreed upon elements of library processes. Whilst you may need to dig deep to collect and confirm processes, the framework itself, I suggest, should remain fairly high -providing individual enterprises the ability to compare, contrast and build upon that framework in their own context. That said, libraries have been around for a very long time, I'm certain that libraries have many business processes that they commonly share.*

---

[21] Off course that the retirement age for the Digital Library will occur naturally, when its children and grandchildren will emerge with new issues and challenges, on the top of its shoulders. Our "intellectual youngest cousin", the Semantic Web, could be one of those descendents, but in spite of the "good schools" where it has been breed and educated, it remains uncertain if it will be able to provide practical value. The Web 2.0, like the "new kid on the block", is bringing new and fresh fascinating ideas, but its informality makes us nervous; it is not clear yet if its actual effectiveness is not only a transient property resulting from the enthusiasm of the schoolboys.

[22] http://www.diglib.org/architectures/serviceframe/

[23] http://ea.typepad.com/enterprise_abstraction/2006/11/dlf_services_wo.html (this entire blog, from Stephen Anthony, deserves a close reading by any Digital Library practitioner).

What am I really saying? I'm saying there are at least 2 levels of architecture here. The high level meta-architecture (framework) that's generally agreed upon amongst libraries, and then there's a true enterprise-level architecture that's needed within an institution to meet specific needs. The enterprise-level architecture should, ideally, use the framework to guide their architecture development and implementations... but a framework can never fully accommodate the specific business needs, planning and implementation required within an organization."

Concluding, maybe it is time to recognise that the focus of the DL should move from the perspective of the engineer to the perspective of the architect[24].

References

[1] Arms, W.: Key Concepts in the Architecture of the Digital Library. D-Lib Magazine, (July 1995), http://www.dlib.org/dlib/July95/07arms.html

[2] Bush, V.: As We May Think. Atlantic Monthly 176(1), 101–108 (1945), http://www.theatlantic.com/doc/194507/bush

[3] IEEE: IEEE Std 1471-2000 IEEE Recommended Practice for Architectural Description of Software-Intensive Systems –Description (October 9, 2000)

[4] Lesk, M.: The Seven Ages of Information Retrieval. In: Proceedings of the Conference for the 50th anniversary of As We May Think, pp. 12–14 (1995), http://www.lesk.com/mlesk/ages/ages.html

[5] OASIS - Organization for the Advancement of Structured Information Standards. Reference Model for Service Oriented Architecture. Committee Specification 1. 2 (August 2006), http://www.oasis-open.org/committees/download.php/19679/soa-rm-cs.pdf

[6] Schekkerman, J.: How to survive in the jungle of Enterprise Architecture Frameworks. Trafford Publishing, ISBN 1-4120-1607-X (2004)

[7] Spewak, S.: Enterprise Architecture Planning – Developing a Blueprint for Data, Applications and Technology. John Wiley & Sons Inc, Chichester (1993)

[8] Thomas, E.: Service-Oriented Architecture: Concepts, Technology, and Design. Prentice Hall PTR, Englewood Cliffs (2005)

[9] Zachman, J.: A Framework for Information Systems Architecture. IBM Systems Journal 26(3) (1987) IBM Publication G321-5298

[24] http://answers.google.com/answers/main?cmd=threadview&id=233551 (Google Answer entry).