DELOS
Network of Excellence on Digital Libraries

Foundations of Digital Libraries

Pre-proceedings of the First International Workshop on "Foundations of Digital Libraries"

Vancouver, British Columbia, Canada, June 23, 2007
In conjunction with ACM IEEE Joint Conference on Digital Libraries (JCDL2007)

Editors:
Donatella Castelli (Italian National Research Council, ISTI-CNR, Pisa, Italy)
Edward A. Fox (Virginia Tech, Blacksburg, Virginia, USA)

June 2007

DELOS: a Network of Excellence on Digital Libraries

www.delos.info

Information Society Technologies

Preface

We are happy to present the papers to be discussed at the 2007 JCDL Workshop on "Foundations of Digital Libraries". This workshop is the first one dedicated to this very fascinating theme, that has gained increasing interest in recent years, although attention to digital libraries (DLs) dates from many years ago.

It is not surprising that DL researchers feel the need of foundations for the Digital Libraries field. The DL universe, in fact, has continued to grow in the last fifteen years and has become a very complex one, producing very heterogeneous models and systems. Now DL researchers are aware that foundations are urgent to avoid that the results of their work are difficult to confront and even harder to combine and reuse to produce enhanced outcomes.

The papers that will be presented for discussion in this workshop aim to contribute to laying the foundations for digital libraries as a whole, as well as continuing the work on the definition of a Reference Model for Digital Libraries launched by the EU DELOS Network of Excellence on Digital Libraries. The common goal is to produce a reference framework wherein new results can be integrated, compared, and discussed. With this view, the workshop gives particular attention to the modeling of three of the main aspects characterizing the Digital Library universe: *Content*, *Architecture*, and *Quality*, in order to contribute to the consolidation of these key concepts. In each of the sections dedicated to these arguments, a number of papers are offered, beginning with those dedicated to the motivations for and approaches to DL foundational work.

Besides researchers, other important communities are interested in the essence of the DL fields, i.e., library users and library providers. We have no doubt that the results of this workshop also will give these communities a better understanding of the DL universe and the opportunity of reasoning about this universe and communicating with a common and well founded concept vocabulary.

June 2007 Donatella Castelli and Edward A. Fox

Organization

First International Workshop on "Digital Libraries Foundations" was organised by Donatella Castelli and Edward A. Fox with the collaboration of the ACM IEEE Joint Conference on Digital Libraries (JCDL 2007).

Organising Committee

Donatella Castelli (Italian National Research Council, ISTI-CNR, Pisa, Italy)
Edward A. Fox (Virginia Tech, Blacksburg, Virginia, USA)

Program Committee

Maristella Agosti (University of Padua, Italy)
Stavros Christodoulakis (Technical University of Crete, Hellas (Greece))
Geneva Henry (Rice University, USA)
Carlo Meghini (Italian National Research Council, ISTI-CNR, Pisa, Italy)
Heiko Schuldt (University of Basel, Switzerland)
Dagobert Soergel (University of Maryland – College Park, USA)

Additional Reviewers

Leonardo Candela (Italian National Research Council, ISTI-CNR, Pisa, Italy)

Table of Contents

Foundations of Digital Libraries

Some Preliminary Ideas Towards a Theory of Digital Preservation

Giorgos Flouris
ISTI-CNR
Via Giuseppe Moruzzi, 1
56124, Pisa PI Toscana, Italy
flouris@isti.cnr.it

Carlo Meghini
ISTI-CNR
Via Giuseppe Moruzzi, 1
56124, Pisa PI Toscana, Italy
meghini@isti.cnr.it

ABSTRACT

The problem of digital preservation is one of the most challenging research problems faced by the community of digital libraries today, receiving growing interest by researchers and practitioners alike. One of the major gaps in the related research is the lack of a general agreement on a formal model to describe the problem or on a formal description of the required properties of a good solution to the problem. This work aims to fill this gap by presenting a number of ideas towards a formal, mathematical, logic-based description of preservation as a scientific discipline, to the end of deriving a methodology resting on solid theoretical grounds. We will present and justify a number of desired properties of such a formalism and introduce a model that handles the static aspects of the problem; some ideas related to the dynamics of preservation will be presented as well.

1. INTRODUCTION

The rapid obsolescence of large volumes of digital (especially "born-digital") data is one of the most challenging problems faced by modern archivists. This problem is commonly referred to as the problem of *digital preservation* [10, 15] and deals with the problem of retaining the meaning of a digital object (file, image, database, document, etc) unaltered for an evolving community of readers. Such readers are usually referred to as the *Designated Community* (DC) of the digital object [3, 12].

The problem of digital preservation is not fully understood to date; even though there is a number of ongoing efforts on the practical and methodological aspects of preservation (e.g., [5, 13, 14]), there are very few efforts in the direction of a formal description of the problem [4]. The introduction of such a formal description would in many ways contribute to the research field of digital preservation. For example, a formal theory could allow the development (and proof) of impossibility and existential results: given the inherent difficulties associated with the problem, we intuitively expect some limitations on what types of digital objects can be preserved; we also expect certain types of DC evolution to be such that no preservation is possible. In addition, a formal theory could allow the grounding of existing (and future) preservation methods upon a common formalism for comparison, and could result to a set of formal desirable properties for evaluating such methodologies [8].

Motivated by the above considerations, we propose certain definitions which are part of a larger ongoing effort towards the development of a formal, mathematical, logic-based description of preservation as a scientific discipline, to the end of deriving a methodology resting on solid grounds.

We begin with a general discussion on digital preservation, addressing some general properties of the problem (section 2). This discussion includes some thoughts on the relationship of our ideas with existing standards, such as OAIS [3]; establishing such a relationship is necessary, as it would eventually allow the connection of this work with existing efforts (such as the CASPAR project [2]). Following that, we introduce a formalism that handles the static aspects of the problem (section 3) and present some thoughts related to the dynamic aspects of the problem (section 4).

2. DISCUSSION ON PRESERVATION

2.1 Types of Preservation

As already mentioned, digital preservation refers to the problem of retaining the meaning of a digital object unaltered for an evolving DC. Let us suppose that the digital object under question is an image, say I, created by a particular individual (called the *producer* [12]), say P; moreover, consider a particular member of the DC (called the *consumer* [12]), say C.

The ultimate goal of preservation is to ensure that C understands I despite the many changes that can intervene as time passes by. Understanding in this context implies accessing, of course, but access alone is (usually) not enough. Below, we sketch the general steps required for C to understand I; notice that most of these steps require the use of some artificial agent (software program, hardware device etc) to apply the relevant transformation:

1. The original input is the physical storage (on some form of long-term storage media) of the sequence of bits which encodes the image I in some format.
2. By *reading* these bits from the storage media, C obtains a sequence of bit values representing the image.
3. By *rendering* these bits, C obtains an image that is

First International Workshop on "Digital Libraries Foundations" Vancouver, British Columbia, Canada, June 23, 2007

some form of light that C's eyes can take in. Rendering presupposes some knowledge on the image *format*.

4. By *interpreting* the image, C figures out its meaning, i.e., the worlds in which the portrayed scene can occur.

The fundamental divide in the above discussion is the separation between rendering the object, the image in our case, and understanding the object. In other words, we regard the above process as *the interpretation of the rendering of the bit stream*. Preservation implies the ability to perform this process *at any time*. This leads to our informal definition of preservation as: *the ability to perform the interpretation of the rendering of a bit stream at any time*. Notice that this involves three steps: producing the bit stream, rendering the produced bit stream and understanding the rendered object. This results to a *decomposition* of the preservation task into sub-tasks, each corresponding to one *preservation type*.

The first type, called *bit preservation*, refers to the ability to produce a particular sequence of bits from a storage media at any time; this can be achieved using error correction techniques, backups, RAID or mirrored disks, media refreshment and other technologies.

The second type, called *data preservation* or *object preservation*, refers to the ability to render the produced bit stream and produce a meaningful output from it at any time. This is the focus of most current approaches to the problem.

The third type, called *information preservation*, refers to the ability to understand the rendered object at any time, i.e., to be able to understand its content by understanding the terms, concepts or other information that appears in it, by placing it in its correct context etc. This is the toughest type of preservation, and is often ignored by existing preservation approaches.

We argue that a complete preservation system should handle all three preservation types. Notice that information preservation applies also for physical objects, whereas the other preservation types only make sense for the realm of digital objects. In what follows, we will not consider bit preservation; for some relevant discussion, refer, for example, to [17]. Our work focuses on information preservation, even though most of the approaches presented here can be easily amended to apply for data preservation as well.

2.2 Preservation in Time and Space

Normally, the process of digital preservation applies when the passage of time renders some digital object incomprehensible by a particular DC. However, we can view preservation as the more general process of allowing an object to be understood by some target DC. The ability of the DC to understand an object may be hindered by several factors, including, but not limited to, the passage of time; the intelligibility of a certain digital object may also depend, for example, on a number of software or hardware modules, or on some background knowledge regarding some particular domain, which may or may not be available to the target DC. This gives rise to two "preservation dimensions": the space dimension and the time dimension.

In the space dimension, the producer needs to formulate the created digital object is such a manner so that the various DCs that he is addressing his data for (which, in general, may have different background knowledge, rendering abilities, hardware, software etc) can understand it.

The time dimension represents the evolution of the knowledge of the DC in time. Such evolution may be, e.g., due to some new discovery, in which case the changes are easy to capture, well-documented and noticeable. However, this is not always the case, as it is possible that the evolution could be due to slight changes in knowledge, jargon, terminology etc, which usually go by unnoticed, but accumulate through time. Thus, the knowledge of the DC should be checked at regular intervals, and, if changes are found, an explicit knowledge shift should be performed to guarantee preservation. This shift consists in the specification of the new knowledge of the DC (i.e., the currently used knowledge) and the change that resulted in this shift.

In fact, both "preservation dimensions" can be essentially reduced to the following problem: given a digital object, carrying a particular meaning, format, etc, as well as a target DC, with some given rendering abilities, software and hardware modules, background knowledge etc, determine the changes required upon the original digital object so that the DC can understand the meaning intended by the object's original producer.

Notice that this formulation makes no reference to the time element, so it avoids the problem of not knowing what a future DC will be like. This way, the preservation problem becomes in many respects similar to a communication problem between two agents and its recursive character is eliminated: we only need to devise a way through which an agent can adequately amend a digital object so as to be understandable by another agent. Once we achieve this, by repeating this process once per agent (i.e., DC), the problem is solved in the space dimension. Moreover, by repeating this process once per agent (i.e., DC) evolution, the DC at time t can play the role of the producer, so as the next-generation DC (at time $t + 1$) will be able to correctly understand the meaning of the digital object, as it was understood by the DC at time t (which is hopefully identical to the meaning intended by the producer at time 0).

2.3 Questions and Answers

In order for preservation to be possible, it is generally necessary for the producer to include in the digital object a certain amount of information on how the object should be interpreted, as well as, possibly, a certain amount of redundancy that will help consumers decipher its meaning. One of the major problems that need to be resolved for preservation is to determine what this information is and how it should be formally represented.

A related issue is which part of the digital object is worth preserving. For example, if the digital object is a text document, then it is composed of various information, including its content, format, fonts, pagination information, attached images or other objects, etc; depending on the context, we may be interested in only a part of this information. Thus, we argue that it is not usually necessary (or possible) to preserve the entire information carried by a digital object; instead, we could isolate and preserve the object's most "useful" or "important" information.

To formalize the above requirements, we will consider that a digital object is a set of *questions* (or *properties*) whose *answers* (or *property values*) will help the consumer understand the (interesting part of the) meaning of the object. Notice that this viewpoint is sufficiently general, as it allows us to include in the preserved digital object some, or all,

of the information in the original object, as well as to include additional, external to the original object, associated information that may be useful for preservation purposes.

2.4 Relevant Questions and the OAIS Model

Determining the information (i.e., questions) worth preserving for the object at hand is not an easy task; it depends on the object type, its content, legal issues as well as on the producer's and consumer's needs, among other things. A great aid in this task is provided by preservation models, such as the OAIS standard [3]. The role of such a model in this respect is to provide a methodological framework and a "best practices" approach towards the aim of determining the most important information related to a digital object.

As an illustration, the categories of information that OAIS prescribes are the Content Information (which is in turn divided into Content Data and Representation Information) and the Preservation Description Information (which is in turn divided into Provenance, Reference, Context and Fixity); the Representation Information is further divided into Structural Information and Semantic Information (see [3] for details). Each of those types of information could be modeled as questions about the object.

3. PRESERVATION STATICS

3.1 Required Model Properties

Before performing any preservation activity, we need to formalize a way to represent a digital object as a set of questions and answers. These should be expressed in some language, let's call it $\mathcal{L}$, which will formally determine the syntactical and semantical rules that can be used for formulating such questions and answers.

We will define $\mathcal{L}$ to be a formal language of a logical nature. There are various arguments in favor of this choice. First, $\mathcal{L}$ has to be formal, like logics are, otherwise no scientific theory of preservation can be developed; second, it must be able to express knowledge, and formal logic has been developed for exactly this purpose; third, it must be suitable to capture question-answering, and the inference relation of mathematical logic allows precisely that; and, finally, logic is a very well studied field of science, offering a very rich set of results from which to draw.

There is an overwhelming array of mathematical logics we could use; at this stage, we do not embrace any of them, because this is not necessary for developing a theory of preservation. The only assumptions made about $\mathcal{L}$ is that it allows us to state queries by talking about otherwise unspecified individuals and that it comes with a formal semantics and an associated inference relation $\models$.

Informally, $\mathcal{L}$ can be viewed as the language which must be "spoken" (understood) by someone in order to be able to understand the (questions and answers related to the) digital object under question. In the process of "reading" a digital object (say a text document), we are often able to draw conclusions that are not direct consequences of the document's content, but are partly based on some background or commonsense knowledge. Such background knowledge is necessary for the correct understanding of a digital object, so $\mathcal{L}$ should be coupled with some domain knowledge, represented by a logical theory $\mathcal{T}$, which is expressed in terms of the language $\mathcal{L}$. Following intuition, $\mathcal{T}$ will be assumed finite and consistent.

Notice that a digital object is nothing more than a bunch of symbols unless coupled with some formal structure that provides the semantics to these symbols. This formal structure is the pair $\langle \mathcal{L}, \mathcal{T} \rangle$ which allows us to understand the "meaning" of a digital object; this pair will be called the *Underlying Community Knowledge* (UCK) of the digital object and each digital object will be considered to be associated to a single UCK, which provides the framework for understanding it.

Notice that the content of the UCK depends on the context. For example, if we are interested in data preservation, the UCK would be a formal description of the underlying format of the digital object; if we are interested in information preservation, the UCK would be a formal description of how the rendered object should be interpreted. Moreover, both the producer and the consumer have a UCK of their own; if this UCK is the same, they can both understand the digital object, and no preservation is necessary. Problems emerge when the UCKs of the producer and the consumer are different, in which case a digital object that carries a particular meaning for the producer may carry a totally different meaning for the consumer, or, more likely, be totally unreadable; this is where preservation comes into play.

As mentioned above, $\mathcal{L}$ allows the statement of queries; such queries will be used to formalize questions. Similarly, the individuals being the answers to such queries will be used to formalize the answers to such questions. Answers to questions should normally encode genuine information about the digital object, in the sense that this information is not implicit in the underlying theory $\mathcal{T}$; however, we can imagine situations where this is not necessarily the case. On the other hand, answers cannot contradict our knowledge (i.e., $\mathcal{T}$). Finally, all answers are assumed to be given by a knowledgeable person, which could be either the producer himself or some other person who can understand the digital object well enough to provide information on it.

3.2 Formal Embodiment of our Requirements

We now have all the ingredients we need to fulfill our goal of determining a formal model for the statics of digital preservation. As mentioned above, such a model should contain a UCK (consisting of a formal language, $\mathcal{L}$ and a logical theory $\mathcal{T}$ from $\mathcal{L}$), as well as a digital object (consisting of a set of queries from $\mathcal{L}$, say $\mathcal{Q}$, and a set of answers to each such query, formalized using a function, say *ans*).

More formally, we define the *Underlying Community Knowledge* (or UCK) as a pair $\mathcal{U} = \langle \mathcal{L}, \mathcal{T} \rangle$, where:

- $\mathcal{L}$ is a logical language, or, more formally, a tuple $\mathcal{L} = \langle \mathcal{L}^L, \mathcal{V}, \mathcal{V}^I, \mathcal{P}, \mathcal{P}^C, \models \rangle$, consisting of the following elements:
  - The set of logical symbols of the language, denoted by $\mathcal{L}^L$.
  - The vocabulary $\mathcal{V}$, which is a set of symbols.
  - A set $\mathcal{V}^I$, which is the subset of $\mathcal{V}$ that contains the individuals of the language, defined as all the elements of the vocabulary $\mathcal{V}$ that can be produced as answers to queries ($\mathcal{V}^I \subseteq \mathcal{V}$).
  - A set of well-formed formulas $\mathcal{P}$, which is a non-empty set containing all the formulas that are allowed in $\mathcal{L}$.

- The set $\mathcal{P}^C$ which is the set of closed formulas of the language $\mathcal{L}$. Obviously $\mathcal{P}^C \subseteq \mathcal{P}$. $\mathcal{P}^C$ in effect splits $\mathcal{P}$ into two disjoints sets, namely the set of closed formulas (i.e., $\mathcal{P}^C$ itself) and the other formulas called open formulas and denoted by $\mathcal{P}^O$; obviously $\mathcal{P}^O = \mathcal{P} \setminus \mathcal{P}^C$. Closed formulas will be used to represent facts (e.g., in the theory $\mathcal{T}$), while open formulas represent queries (for $\mathcal{Q}$).
- A binary relation $\models$ between elements of $\mathcal{P}$ (the inference relation of the logic).

- $\mathcal{T}$ is a finite and consistent theory in $\mathcal{L}$: $\mathcal{T} \subseteq \mathcal{P}^C$.

Each *digital object* is associated to a certain UCK $\mathcal{U} = \langle \mathcal{L}, \mathcal{T} \rangle$ and is defined as a pair $\mathcal{D} = \langle \mathcal{Q}, ans \rangle$ where:

- $\mathcal{Q}$ is a finite, non-empty set of queries in $\mathcal{L}$: $\mathcal{Q} \subseteq \mathcal{P}^O$.
- *ans* is a function associating each query $q \in \mathcal{Q}$ with an answer, that is a set of tuples $\vec{a}$ of individuals in $\mathcal{L}$.

We impose a further requirement on *ans*, by asking that the answers, taken all together, do not break consistency. This means to ask the consistency of the theory: $\mathcal{T} \cup \{q(\vec{a}) \mid q \in \mathcal{Q} \text{ and } \vec{a} \in ans(q)\}$.

Notice that the structure $\mathcal{D} = \langle \mathcal{Q}, ans \rangle$ contains all the questions and answers that were chosen for preservation (see subsection 2.3). Thus, the set of sentences: $\{q(\vec{a}) \mid q \in \mathcal{Q},\ \vec{a} \in ans(q)\}$ is all the information required to enable the interpretation of the part of the digital object that was considered useful for preservation purposes.

Since each preserved digital object is associated to a UCK, we can define the pair $\langle \mathcal{U}, \mathcal{D} \rangle$, or equivalently the 4-tuple $\mathcal{S} = \langle \mathcal{L}, \mathcal{T}, \mathcal{Q}, ans \rangle$, as the *Information Preservation Structure* (IPS) of the digital object. The IPS contains all the information related to the preservation of the digital object, because it contains both the digital object itself (i.e., the questions and answers in $\mathcal{D}$), as well as the description of the meaning of the symbols in $\mathcal{D}$ (i.e., the UCK $\mathcal{U}$).

4. PRESERVATION DYNAMICS

4.1 Preliminary Discussion on the Dynamics

As already mentioned, preservation comes into play when producer's background knowledge is different from the respective consumer's knowledge. Thus, using the terminology introduced so far, the problem of preservation can be defined as follows: given a digital object $\mathcal{D}_O$ whose content (meaning) is understandable using some UCK $\mathcal{U}_O$, a different UCK $\mathcal{U}_N$, and a description of the differences (evolution) between $\mathcal{U}_O$ and $\mathcal{U}_N$, find a digital object $\mathcal{D}_N$, whose content (meaning), understood using $\mathcal{U}_N$, is identical to the content (meaning) of $\mathcal{D}_O$, understood using $\mathcal{U}_O$.

The first problem we have to face in the above process is the identification of the exact changes that led to the new UCK from the old. We argue that the complexity of the UCK structure implies that the changes might be so subtle (or so great) that no automated system (or human being) can determine them by just looking at $\mathcal{U}_O$ and $\mathcal{U}_N$; for example, it is possible that complex changes may overlap and "hide" the effects of each other from an external observer. Therefore, we will make the (reasonable) assumption that preservation takes place while there are still people (human experts) who are knowledgeable of both the new and the old UCK and have kept track and can pinpoint the exact changes that occurred during the UCK evolution.

Given the detailed description of those changes, the purpose of preservation is to determine the changes to apply to the digital object $\mathcal{D}_O$, in order to get the new object, $\mathcal{D}_N$. Such changes should be calculated as a function of the old digital object ($\mathcal{D}_O$), the two UCKs ($\mathcal{U}_O, \mathcal{U}_N$) and the UCK change specification. Notice that this viewpoint allows us to generalize any solutions found, because, once we have found how to preserve an object of some type (i.e., an object associated with some particular UCK) against some particular UCK evolution, we can apply the same solution (function) to all objects associated with the same UCK. For example, if we want to preserve a large number of images of the same format against format obsolescence, all we have to do is determine the correct transformation for one image; then, the same transformation can be applied to the other images.

Our definition makes it clear that, in preservation, the exact syntactical formulation of a digital object is irrelevant; what we are interested in preserving is the *meaning* of the digital object, as derived from the associated UCK.

A final note on the above definition is that it is not always desirable (or possible) to achieve perfect preservation; in some cases, the new DC language ($\mathcal{L}_N$) may be less expressive than the old one ($\mathcal{L}_O$) so the exact meaning of the digital object may not be expressible using $\mathcal{L}_N$; in other cases, part of the meaning of the original digital object may be inconsistent with our current background knowledge ($\mathcal{T}_N$), so, by our definitions and constraints (subsection 3.2), this part should not be preserved.

Combining the above ideas, we conclude that a solution to the problem of preservation should, first, determine a powerful enough formal structure that can describe UCK evolution and, second, define a formal process that will determine the new digital object $\mathcal{D}_N$, as a function of the old ($\mathcal{D}_O$), the two UCKs ($\mathcal{U}_O, \mathcal{U}_N$) and the UCK evolution specification. This function should be such that the meaning of the old digital object is preserved as much as possible, so we should formally define what constitutes "preservation of the meaning" as well.

In the next subsection, we will present some examples that will lead us to some preliminary ideas towards resolving the above issues; a more concrete answer to the above concerns is part of our future work.

4.2 Desired Properties and Examples

Let us consider the example of the evolution of our symbolism from the Roman numerals (I, II, ...) to the Arabic ones (1, 2, ...). An informal description of this evolution could be something like: "the old symbol 'I' evolved to the new symbol '1', the old symbol 'II' evolved to 2, ... etc".

An immediate observation that can be made from this example is that the "language" used to describe the UCK evolution contains terms from both the old (e.g., 'I') and the new (e.g., '1') UCK. Thus, any attempt for a formal description of the UCK evolution should be expressed in a language (i.e., UCK, say $\mathcal{U}_E$) that is at least as expressive as either of $\mathcal{U}_O$ and $\mathcal{U}_N$.

As a second example, let us consider a recent terminology change in the field of astronomy. In August, 2006, during a meeting in Prague, astronomers decided to change the definition of the term "Planet"; in addition, they introduced a new term, "Dwarf Planet" [1]. As a consequence of these

changes, Pluto is no longer classified as a planet, but as a dwarf planet.

The main difference of this example with the previous one is that there is no direct 1-1 correspondence between the meaning of the terms of the two UCKs, because there are new terms that don't correspond to any term in the old UCK (e.g., "Dwarf Planet"), there are terms which don't change name but change meaning (e.g., "Planet") and there are terms that change neither name nor meaning, but, due to other terminological changes, their status with respect to other terms does change (e.g., "Pluto").

The above change types are only a small list of the various changes that could occur to terms; thus, the UCK evolution structure should allow fine-grained information to be captured. If there is a term in the new terminology corresponding to a term in the old (like in the first example), we should be able to denote so; if not, we should be able to express as much as we know about the relationships between the old term and the new terminology.

In addition, even though the above discussion is largely limited to vocabulary changes, this is not the only type of change that a UCK may undergo. More difficult are the cases where the logic itself changes where similar problems may occur. According to [7], the changes that our knowledge may undergo can be classified in three broad categories (levels). The first level (level 1, or logic changes) corresponds to changes in the logical formalism used to describe our knowledge (e.g., removal of a logical operator); the second level (level 2, or language changes) corresponds to changes that affect the vocabulary that is relevant to the domain (e.g., the addition of a concept name or predicate name); the third level (level 3, or KB changes) corresponds to changes that affect our knowledge on the relations between the vocabulary elements (e.g., the addition of logical propositions). To the authors' knowledge, preservation is the only real-world problem in which all three change levels are relevant.

4.3 Ideas Towards a Possible Solution

A possible way to resolve the above problems is to use a mapping from each UCK ($\mathcal{U}_O$, $\mathcal{U}_N$) to the expanded one ($\mathcal{U}_E$). The semantics of this mapping, say f, is that an element x from $\mathcal{U}_O$ (or $\mathcal{U}_N$) "corresponds" to (i.e., has the same meaning as) the element $f(x)$ from $\mathcal{U}_E$. Abusing notation, we will use the same f regardless of whether x is a term, a language symbol, an open formula etc. In effect, f corresponds to a mapping from each of the structures comprising $\mathcal{L}_O$ and $\mathcal{L}_N$ to the respective structure in $\mathcal{L}_E$.

Thus, a structure describing the evolution of the UCKs should consist of an expanded UCK ($\mathcal{U}_E$) and a mapping (f) that provides the correspondences between the various elements of $\mathcal{U}_O$, $\mathcal{U}_N$ with $\mathcal{U}_E$. Using f, we can define what it means to retain the meaning of an element: an element y of $\mathcal{U}_N$ *retains the meaning of* x of $\mathcal{U}_O$ iff $f(x) = f(y)$.

To capture more complex interrelationships between elements of $\mathcal{U}_O$ and $\mathcal{U}_N$, we will use the theory of $\mathcal{U}_E$ (namely $\mathcal{T}_E$) and the $\models_E$ relation of $\mathcal{U}_E$. In particular, to capture a complex terminological relationship between the terms x (of the old UCK) and y (of the new UCK), we represent this relationship using a formula relating $f(x)$, $f(y)$ in terms of $\mathcal{U}_E$ and include it in $\mathcal{T}_E$. Similarly, to capture complex logical relationships between formulas x (of the old UCK) and y (of the new UCK), we include the respective relationship (between $f(x)$, $f(y)$) into the $\models_E$ relation.

The next step is to define what it means for a digital object to *preserve* another. A straightforward definition that uses the notion of "retaining the meaning" is too restrictive, as it is based on both the syntax and the semantics of the involved objects (rather than just the semantics).

Thus, it would make more sense to use some notion of "equivalence" that will allow us greater flexibility on how to preserve a digital object. This idea leads to a number of different definitions, depending on how we formally interpret the term "equivalence". Probably the most interesting way to define this notion is as follows: a digital object $\mathcal{D}_N = \langle \mathcal{Q}_N, ans_N \rangle$ associated to $\mathcal{U}_N = \langle \mathcal{L}_N, \mathcal{T}_N \rangle$ *preserves* $\mathcal{D}_O = \langle \mathcal{Q}_O, ans_O \rangle$, associated to $\mathcal{U}_O = \langle \mathcal{L}_O, \mathcal{T}_O \rangle$ iff $\mathcal{T}_E \cup D_{OE} \equiv_E \mathcal{T}_E \cup D_{NE}$, where: $D_{OE} = \{f(q(\vec{a})) \mid q \in \mathcal{Q}_O, \vec{a} \in ans_O(q)\}$, $D_{NE} = \{f(q(\vec{a})) \mid q \in \mathcal{Q}_N, \vec{a} \in ans_N(q)\}$.

According to this definition, to determine whether $\mathcal{D}_N$ preserves $\mathcal{D}_O$, we take each question-answer pair of the old digital object and map it into its "corresponding" formula in $\mathcal{U}_E$ (using f); the results, taken together, constitute D_{OE}, which is combined with the information on the relationships between the terminology of the old and the new UCK (i.e., the background knowledge of the expanded UCK, $\mathcal{T}_E$). The same process is followed for the new digital object. The definition states that preservation is achieved iff the respective results (for $\mathcal{D}_O$, $\mathcal{D}_N$) are equivalent (under $\models_E$).

As already mentioned, preservation cannot always be perfect; to capture such cases, we would also need to define some notion of *partial* or *approximate* preservation; this is part of our future work.

The final step in the definition of a preservation model is the development of a formal process that will determine the new digital object (i.e., the one that preserves the old) as a function of the old digital object, the two UCKs and the description of the evolution between the two UCKs. To resolve this problem, we need to identify those formulas from $\mathcal{U}_E$ which (a) have an equivalent in $\mathcal{U}_N$ (through f), and, (b) taken together, they satisfy the condition for preservation given above. The exact determination of a step-by-step process for this task is also part of our future work.

4.4 Representing Evolutions

The above structures are useful for theoretical manipulations, but are rather cumbersome in practice without some adequate compact representation. In this respect, the two well-established fields of ontology evolution [11] and belief revision [9] may be of use; these fields are dealing with the representation and determination of changes upon a corpus of knowledge, which could be an ontology (in ontology evolution) or some formal logical theory (in belief revision).

Even though this is a valid option, it should be emphasized that it would only partly cover our preservation needs. The first reason for this is that neither of these fields deals with level 1 changes [7]. In particular, belief revision only deals with level 3 changes, while ontology evolution deals with changes in levels 2 and 3. This restricts the types of UCK evolutions that these fields can describe and handle.

In addition, most of the developments in these fields are based on certain assumptions on the underlying logic; should the UCK logic be different, most of the relevant literature would be inapplicable. For a recent attempt to (partially) overcome this problem, in a different context, see [6].

Another problem that invalidates this option in certain contexts is the "infiniteness" issue. Both belief revision and

ontology evolution use a simple, explicit and straightforward way to represent changes as a list of operations; unfortunately, this would not work in all cases. The example with the Roman and Arabic numerals (subsection 4.2) is an excellent manifestation of this fact: as is obvious from the informal description of that evolution, there is an infinite number of evolutions that took place, one per Roman numeral. Thus, it is not possible to explicitly describe such an evolution in a finite way using the standard methodology; a more compact implicit specification is required.

Unfortunately, this "infiniteness" problem appears more often than not in real-world applications. An everyday example is conversions from one currency type to another, or from one unit of measurement to another (e.g., Celsius degrees to Fahrenheit degrees); in such cases, every symbol (e.g., 18^oC) should be transformed to its equivalent (90^oF) and there is a potentially infinite number of different temperatures (symbols) that could be measured.

One way to address this problem is to describe evolution as the output of a certain algorithm which can be finitely expressed using one of the formalisms developed in computer science (e.g., Turing Machines) [16]. Of course, this option invalidates the use of all representations and methodologies employed in belief revision and ontology evolution.

Despite these deficiencies, we argue that the fields of belief revision and ontology evolution could (and should) be applied for certain types of UCK evolution. Such an option would relieve us from dealing with problems already addressed in these fields, so we believe it's worthwhile to consider it. For example, ontology evolution could handle the astronomy example presented in subsection 4.2.

5. EPILOGUE

This paper reports on an ongoing effort with the ultimate goal of formally modeling the process of digital preservation. We started with a general discussion on the problem, which allowed us to determine the basic properties that such a model should have. This discussion also led to the definition of the vital steps that need to be performed towards this aim, as well as to a number of preliminary proposals that satisfy most of the required properties of such a formalism.

We argued that the process of digital preservation should be described using a model that describes both the digital object under preservation itself (using the questions-answers mechanism) and the general context (semantical, syntactical etc) in which this object is placed (i.e., background knowledge, captured by the UCK structure).

Using these notions, we described the problem of preservation in terms of UCK evolution and argued that, in order to formally model it, we need to define the process that would determine the new digital object as a function of the old digital object, the old and the new UCK, as well as the information on the UCK evolution; the new digital object should be such that the meaning of the old digital object is preserved, so a formal definition of this notion was provided.

We believe that the refinement of those initial ideas will lead to a formal model of digital preservation; such a model would be a significant contribution to the research efforts in the field, as it would allow the development (and proof) of formal results, the grounding of preservation methods upon a common formalism for comparison and the development of a set of formal desirable properties for evaluating preservation methodologies.

6. ACKNOWLEDGMENTS

This work was carried out during the first author's tenure of an ERCIM "Alain Bensoussan" Fellowship Programme; it was partially supported by CASPAR (FP6-2005-IST-033572).

7. REFERENCES

[1] 2006 definition of planet. http://en.wikipedia.org/wiki/2006_redefinition_of_planet.

[2] Caspar: Cultural, artistic and scientific knowledge for preservation, access and retrieval. eu funded project (fp6-2005-ist-033572). http://www.casparpreserves.eu.

[3] *ISO 14721:2003: CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS). Blue Book, Issue 1*, 2002. available at: http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html.

[4] J. Cheney, C. Lagoze, and P. Botticelli. Towards a theory of information preservation. In *Proceedings of the 5^{th} European Conference on Research and Advanced Technology for Digital Libraries*, 2001.

[5] M. Factor, D. Naor, S. Rabinovici-Cohen, L. Ramati, P. Reshef, and J. Satran. The need for preservation aware storage: A position paper. *ACM SIGOPS Operating Systems Review*, 41(1):19–23, 2007.

[6] G. Flouris. *On Belief Change and Ontology Evolution*. PhD Thesis, University of Crete, Greece, 2006.

[7] G. Flouris. On the evolution of ontological signatures. In *Proceedings of the Workshop on Ontology Evolution*, 2007.

[8] G. Flouris and C. Meghini. Steps towards a theory of information preservation. In *Proceedings of the International Workshop on Database Preservation*, 2007. Invited Talk.

[9] P. Gärdenfors. Belief revision: An introduction. In P. Gärdenfors, editor, *Belief Revision*, pages 1–20. Cambridge University Press, 1992.

[10] H. Gladney. *Preserving Digital Information*. Springer-Verlag, 2007.

[11] P. Haase and Y. Sure. D3.1.1.b state of the art on ontology evolution, 2004. available at: http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/SEKT-D3.1.1.b.pdf.

[12] B. Lavoie. The open archival information system reference model: Introductory guide. In *DPC Technology Watch Report 04-01*, 2001.

[13] C. Lynch. Canonicalization: A fundamental tool to facilitate preservation and management of digital information. *D-Lib Magazine*, 5(9), 1999.

[14] P. Mellor, P. Wheatley, and D. Sergeant. Migration on request, a practical technique for preservation. In *Proceedings of the 6^{th} European Conference on Research and Advanced Technology for Digital Libraries*, pages 516–526, 2002.

[15] A. Pace. Coming full circle, digital preservation: Everything new is old again. *Computers in Libraries*, 20(2), 2000.

[16] C. Papadimitriou. *Computational Complexity*. Addison Wesley, 1994.

[17] D. Rosenthal. Engineering issues in the preservation of databases. In *Proceedings of the International Workshop on Database Preservation*, 2007.

Foundations of 3D Digital Libraries: Current Approaches and Urgent Research Challenges

Benjamin Bustos*
University of Chile

Dieter W. Fellner†
Technische Universitaet Darmstadt and Fraunhofer Computer Graphics Institute, Germany

Sven Havemann‡
Graz University of Technology, Austria

Daniel A. Keim§
University of Konstanz, Germany

Dietmar Saupe¶
University of Konstanz, Germany

Tobias Schreck‖
Technische Universitaet Darmstadt, Germany

ABSTRACT

3D documents are an indispensable data type in many important application domains such as Computer Aided Design, Simulation and Visualization, and Cultural Heritage, to name a few. The 3D document type can represent arbitrarily complex information by composing geometrical, topological, structural, or material properties, among others. It often is integrated with meta data and annotation by the various application systems that produce, process, or consume 3D documents.

We argue that due to the inherent complexity of the 3D data type in conjunction with and imminent pervasive usage and explosion of available content, there is pressing need to address key problems of the 3D data type. These problems need to be tackled before the 3D data type can be fully supported by Digital Library technology in the sense of a *generalized document*, unlocking its full potential. If the problems are addressed appropriately, the expected benefits are manifold and may lead to radically improved production, processing, and consumption of 3D content.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.3.5 [**Computer Graphics**]: Computational Geometry and Object Modeling—*Curve, surface, solid, and object representations*

*bebustos@dcc.uchile.cl
†d.fellner@igd.fraunhofer.de
‡s.haveman@cgv.tugraz.at
§keim@dbvis.inf.uni-konstanz.de
¶dietmar.saupe@uni-konstanz.de
‖tschreck@gris.informatik.tu-darmstadt.de

First International Workshop on "Digital Libraries Foundations" Vancouver, British Columbia, Canada, June 23, 2007
.

Keywords

Digital libraries, generalized 3D documents, 3D representation, content-based retrieval.

1. INTRODUCTION

The rapid evolution of information and communication technology has always been a source for challenging new research questions in computer science. What happens regularly is that a new generation of technology makes it suddenly possible to process, store, and/or transmit much larger amounts of information. Thus, a gradual *quantitative* increase can turn into a sudden *qualitative* leap, simply because things become possible that were not possible before. The nightmare from a computer science point of view is the *data grave*: Information that is physically present is "lost" for usage because it is simply not accessible with reasonable user efforts.

Digital Library technology aims to revert the data grave problem into a situation where the stored content is brought to its full productive potential by solving the storage, organization, and content-based access problems. For textual documents, retrieval services attacking the data grave problem are widely available, e.g., in form of desktop search engines. But what is the analogue of full text search in a repository of 3D content?

Content-based Digital Library support for 3D data is highly desirable, as the sources for producing digital 3D content are gaining momentum. We argue that the next major technological revolution will be triggered by massive 3D data sets that we will be generated in the near future. The *modeling bottleneck*, the fact that the creation of digital 3D objects was long much too expensive, is overcome by new technologies. Sources generating massive amounts of 3D data will be *3D scanning* (using scanner devices, see Figure 1 for an example), *photogrammetry* (reconstructing 3D data from 2D images), and *procedural/parametric shape design* (creating new shapes from existing similar, parameterized shapes).

Not only is it easier to *produce* digital shape, also the possibilities to *utilize* and take benefit of the created 3D data sets are increasing. A large shift towards 3D is obvious. On the *PC desktop*, computer games have helped 3D graphics hardware becoming the standard, and after Apple's pioneer-

Figure 1: Creation and markup of a 3D model from range-map input. Left: Original input data, 3 out of 20 range maps taken from a statue are shown un-textured. Top right: Simplified versions of the range maps, textured and un-textured. The gravestone and the statue's cheek were manually segmented for semantic markup. Bottom right: Several range maps were integrated and smoothed.

ing work in MacOS X, with Microsoft Vista, 3D will also be integral part of the Windows desktop. Instead of being optional, 3D on the desktop will actually become a standard. In industry, the prospect of *mass customization* is a driving force behind the digitalization of the whole production chain, relying heavily on 3D models to represent process information.

We argue that in the near future, we will be confronted with massive amounts of 3D content, and that novel 3D Digital Library support will be crucial in making the best possible use of these data amounts. In Sections 2 and 3, we outline the current state of the art in Digital Library support for 3D documents, and identify critical research problems. We argue that if these problems are addressed appropriately, a significant leap ahead in the effective use of massive 3D content will be possible. The efforts required to this end are expected to pay off, as illustrated by potential future 3D applications envisioned in Section 4. Section 5 concludes the paper.

2. 3D DATA AND ITS REPRESENTATION

The 3D data type is a very powerful means of capturing and communicating information. Due to the nature of the data type and complexities involved in acquisition, production and processing of 3D data, a number of serious problems in 3D data representation, encoding, content markup, and data history management exist. To date, these problems have not been sufficiently solved, and they are a major obstacle to a full integration of the 3D data type into Digital Libraries. In this Section, we discuss some of the most important research questions of 3D data management in Digital Libraries, according to our view.

2.1 Understanding 3D shape representations

A fundamental difference between 3D and other media types is that there is no canonical 3D representation. While e.g., the image data type can be seen as a set of color samples organized on a regular grid, representing 3D data is more complicated. Existing approaches can be roughly divided into *surface-* and *volume-based* representations, which in turn can be given in discrete, parametric, or implicit form. E.g., 3D objects can be specified by a set of parameterized surface patches based on splines, or a grid of voxels (a discrete volumetric representation). On top of these two broad categories, *structural* information can represent the relationship between models parts in form of scene graphs or boolean set operations which are highly useful for certain shape modeling or manipulation tasks.

These shape representations are *not* all equivalent, because they differ in their expressiveness (the types of forms they can encode) and consequently, in their semantics (content). E.g., a closed surface bounds a volume, but a volumetric data set contains many surfaces at the same time (e.g., iso-surfaces). They also differ regarding the way we can process and analyze them. Discrete representations relate to sampling theory and may exhibit aliasing effects, while continuous representations are noiseless and usually better suited for analytic processing.

An encompassing 3D shape representation taxonomy covering *all* known 3D representations is needed, to better understand the relationships between the existing representations. This should allow a better tackling of the difficult problem to analyze and relate the content of 3D models, irrespective of the given representation, to support common Digital Library tasks such as organizing objects by similarity, deducing hierarchical catalogue orderings, etc.

2.2 Generic 3D file format

Unfortunately, to date there is no single commonly accepted, comprehensive 3D file format, but application-dependent, proprietary file formats are prevailing. In practice, it is usually impossible to convert losslessly between the different established file formats, which is a fundamental problem for importing content from heterogeneous sources into 3D Digital Libraries. In the CAD domain, where almost exclusively NURBS-based model representations are employed, several long-standing, mature exchange standards such as STEP and IGES exist. The problem with these formats is that over time they have become extremely comprehensive and elaborate, so that implementing converter

programs for these formats constitutes a challenge on its own.

From the research viewpoint, the file format problem has been completely ignored up to now, although it is apparently a significant problem which requires fundamental efforts. The focus in 3D modeling research up to now has been to further extend the set of shape representations, rather than to work on a powerful yet transparent canonical 3D file format. The existence of such a format would not only allow the easy integration of 3D content from heterogeneous sources, but could also support the adoption of advanced 3D representations by real-world 3D applications.

2.3 Stable 3D markups

Another important concern from the Digital Library perspective refers to stable markup methods for 3D content. Given a 3D model or scene, reliable methods are needed which allow the stable identification (markup) of parts of the 3D content for attaching annotations, hyperlinks, cross-references, etc. 3D content is often preprocessed or edited along the 3D application pipeline, which usually significantly affects the 3D content representation. E.g., consider a lossy compression performed on a 3D mesh model prior to its transmission over a network. Mesh simplification (decimation) methods affect the number, position, and connectivity of mesh vertices. Any 3D markup method based directly on the mesh index, and which is not explicitly known to the mesh compressor, must then be considered unstable.

So, the research problem to be addressed is to define *generic, stable 3D markup* methods, by designing methods to robustly reference portions of a 3D model. The markup methods should be independent of the 3D content representation, and robust with respect to certain shape editing and processing operations which might be needed by the applications.

A solution to the problem of updating shape markups during shape processing operations is that the processing algorithm is (a) aware of the markup and (b) keeps track of appropriately defined geometric primitives that are affected by the processing operation. Then, after the processing has taken place, these primitives can be converted back to a markup of the initial type. The crucial point here is that the shape representation must be able to enumerate shape components in the reference. Such shape component enumeration can be regarded as *spatial queries* and take the form of closeness to a point, containment in a frustum, or ray intersection. Identification of an efficient set of shape queries which allow implementation of robust 3D markup remains an important research challenge. Figure 1 (right) illustrates a 3D scene with markups.

2.4 Data origin and processing history

During 3D acquisition, production, and processing, the 3D content is often composed from different, heterogeneous sources, and manually or automatically processed by different users and applications. If in a given 3D model, some local shape detail becomes of specific interest to an analyst, it is a vital feature that it is possible to trace back the origin of the specific detail, its degree of authenticity, and the kind of processing applied on it. To this end (a) suitable standards for describing the *provenance* of 3D content, and (b) a general scheme for capturing the *data processing history* applied on the content needs to be developed. Regarding (b), ideally, the captured information should allow a complete replay of the processing the data has undergone.

The enormous complexity of this problem may not be apparent immediately. First, we have to cope with two levels of heterogeneity, namely, the various shape representations, and the various processing operations possible on these representations. Both are not canonical along the different 3D creation and processing tools available. Second, capturing model editing operations must take place at the right granularity. Practically, it is neither possible not useful to capture each manual editing step individually, but an appropriate level of aggregation has to be chosen. Third, to actually replay the processing operations is extremely difficult: It requires that all tools used in the 3D production pipeline support the processing history and add to it. Practical experience regarding software versions, operating systems, undocumented ad-hoc scripting by users etc. suggests this is a tremendous task. More subtle problems in this context are reported in [10].

3. ORGANIZING AND SEARCHING

The previous Section discussed urgent research problems relating to the representation, storage, and processing of 3D content. Assume those problems were already solved. Technically, it would then be easy to build large repositories of 3D content from heterogeneous sources using crawlers, converters, and storage systems. The second major challenge is then to provide effective *content-based organizing* and *searching* functionality for making use of the resulting large 3D repositories.

One way for organizing and searching 3D repositories is to make use of mark-up, authoring, and editing information, or other meta data associated with the models. Unfortunately, the availability and comparability of such information cannot be assumed for content integrated from heterogeneous sources. Instead, *analysis algorithms* are needed to automatically generate suitable meta data information from the repository. The output of the content-based automatic analysis can then be used for organization and retrieval of the content, as a replacement for or addition to object meta data, as far as such is available. We next discuss key challenges in content-based 3D organization and retrieval.

3.1 Need for a dictionary of 3D features

A fundamental library service is *content organization* in the sense of giving structure which helps the user navigate the repository, and to formulate queries which allow to retrieve content of interest. This structuring needs to be based on attributes or features of the data itself. E.g., in case of text documents, features such as title, author, or the main topics addressed by the text are candidates for structuring of text collections. For the 3D data type, attributes such as author or producer might be consistently specified in a generic format. But what constitutes the actual *content* in a 3D data set, and how can appropriate descriptors be automatically generated from the models?

Conceptually, a suitable content description is expected to be determined by the application domain the content is used in. E.g., for a 3D CAD model, the features relevant to a given 3D model can be expected to depend on the engineering context associated with that model. It can be assumed that a set of model features which are relevant in a given engineering context are not necessarily useful to organize,

say, a repository of models representing historic buildings, as both object types are made use of context of their own conceptual background.

What is missing is an encompassing definition of features (aspects, properties) which are relevant to organize and distinguish collections of arbitrary 3D content. A general *taxonomy of 3D features* in the sense of a 3D dictionary needs to be defined, where the dictionary entries

(1) allow to meaningfully describe any type of 3D content,

(2) are descriptive and discriminating in nature, and

(3) can be robustly extracted (detected) by appropriate automatic analysis algorithms.

Unfortunately, to date no such taxonomy exists, so it is problematic to speculate whether and which 3D analysis algorithms would be capable to robustly detect such features, or how such algorithms should be designed. The problem of defining a 3D feature dictionary is complicated by the fact that it is not clear (a) on which conceptual level the features should be defined, i.e., on the statistical, syntactical, or semantical level, and (b) how the features will relate to the 3D shape representation problem, e.g., if they should be defined based on surfaces, on volumes, or on structural properties. The next Section relies on 3D features to introduce a model of the 3D similarity space useful for designing 3D retrieval systems.

3.2 A model for the 3D similarity space

A most fundamental task in Digital Libraries refers to searching for similar content: The user issues a query to the system, and receives a sorted list of answers. A popular searching paradigm is *query-by-example*, where an exemplary object is provided, and the system returns the most similar elements from the repository. However, the notion of similarity per se is under specified. Like for other data types, for 3D objects many different similarity notions are possible, and the Digital 3D Library should offer support for searching along each of those notions. We propose to organize the space of 3D similarity notions along the three dimensions *similarity type*, *addressed feature*, and *invariance properties*. In the following, we discuss each of these dimensions.

Similarity type

Global similarity considers the similarity between complete 3D object instances, and is used to retrieve whole objects. **Partial (local) similarity** on the other hand bases similarity relationships on correspondences of object parts, not necessarily the objects as a whole. This notion is useful e.g., for retrieving scene models, where similarity may be given by correspondence of individual objects in the scene, not necessarily at the same positions. To his end, the partial similarity makes use of the global similarity notion, applied on individual scene elements. A third type of similarity relates to **functional correspondences**, and can be globally or locally defined. Here, similarity relationships are established between objects or object parts based on application-dependent, functional correspondences. E.g., in a CAD context, complementarity between machining parts could establish a functional correspondence.

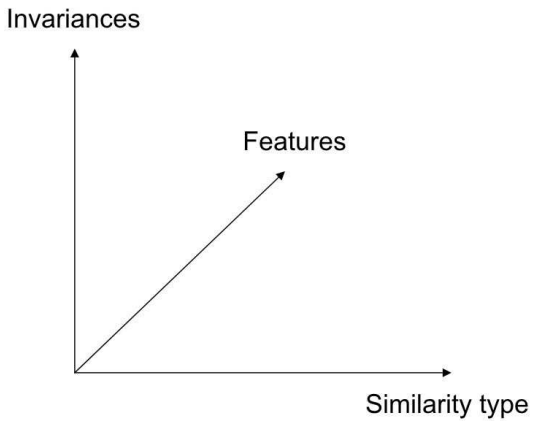


Figure 2: The 3D similarity space model: A combination of similarity type, 3D feature, and invariance setting constitutes a similarity notion.

Addressed features

The similarity types given above can rely on different types of features defined for 3D content description. Important classes of features are based on **geometrical**, **topological**, or **structural** properties of the models. Also, **volumetric** features, or features based on **surface properties** are candidates. It is also possible to consider **annotation** and **markup** information, e.g., processing history or cross-reference information, which might be associated with a given 3D object. Local markups are suited to implement the functional similarity type defined above. However, considering annotation information for 3D similarity evaluation requires a standardized annotation scheme, which allows to compare the individual annotation entries. A major problem in this context is that the types of possible features depend on the 3D representations available, as not all representations allow analysis of all of these features, cf. the discussion in Section 2.

Invariances

The similarity and feature type dimensions are complemented by addition of certain invariance modifiers to the similarity notion. Typical invariance modifiers specify e.g., whether or not **position**, **scale**, or **orientation** of the 3D content in their respective coordinate systems is to be considered when evaluating similarity. Also, invariance regarding the **level of detail** of the 3D content can be desired. Many more invariance modifiers are possible, and in part depend on the type of similarity and feature specified. Integrating such invariance requirements into analysis algorithms is not trivial, but a problem in its own for many existing 3D analysis algorithms.

Figure 2 illustrates the space of 3D similarity notions spanned by these dimensions. The space of possible 3D similarity notions is huge. This model is useful for identifying important similarity notions as well as "blind spots" in this space, which have not been appropriately addressed by research yet (cf. also the next Section). An implication of this large similarity space is that any 3D Digital Library, in which at least some 3D similarity notions are to be sup-

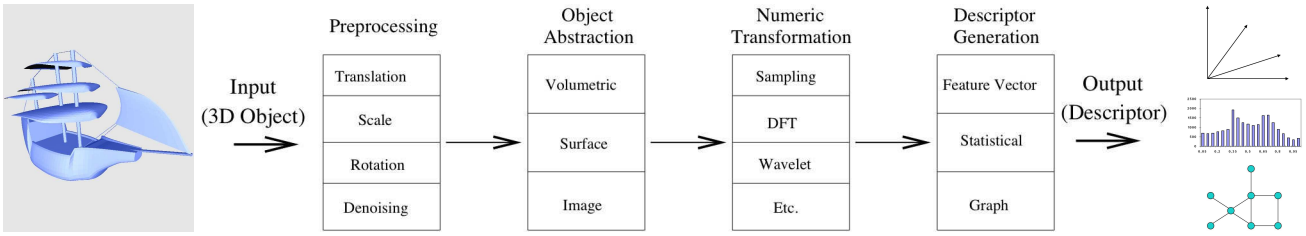


Figure 3: 3D FV extraction process model.

ported, requires significant efforts regarding implementation of 3D analysis algorithms, the output of which is used to quantify the degree of similarity. Also required are then efforts towards the *user interface* side, where the user is to be supported in specifying similarity queries.

3.3 Implementing the similarity notions

Although the 3D similarity notion space is conceptually rich, current methods for retrieval of 3D content mostly focus on the *global geometric similarity* notion. The *transformation* approach determines the similarity between two 3D objects under concern by the cost associated with efficiently transforming (morphing) the global geometry of on object into the other. A simpler, yet efficient approach relies on shape *descriptors*, which are calculated offline for the 3D content. At query time, not the objects themselves, but their descriptors are used for similarity evaluation.

Due to its simplicity and generality, *feature vectors* [4] are often employed as efficient model descriptors. The basic idea is to encode the output of certain shape analysis algorithms in form of vectors of real-valued numbers, effectively representing the 3D content by points in a high-dimensional feature vector space. Distances between the point representations can be calculated, and used as a measure for the (dis)similarity of the underlying objects.

In [3], a process model for the generation of global shape descriptors was presented. Figure 3 illustrates the model which was introduced to capture the essential processing pipeline of most of the current retrieval-oriented shape descriptor algorithms. Briefly, a 3D model is first preprocessed to achieve desired invariance properties. Then, the basis for feature extraction is selected by considering the model as a volume, or by abstracting to its surface or a projection of the model. From this abstraction, low-level features such as the distribution of surface curvature, or shape features calculated from rendered 2D object images, can be captured. From the outcome of this analysis descriptors are formed, with the basic forms being vectors, histograms, or graphs.

To date, a magnitude of low-level 3D analysis algorithms have been proposed, as surveys indicate [2, 12, 9]. Most of them were *heuristically* introduced and motivated by techniques from geometry and image processing. Their suitability for solving the retrieval problem cannot be analytically decided, but needs to be experimentally evaluated by benchmarks [11]. Figure 4 illustrates the evaluation of a number of different low-level descriptors on an exemplary query for a 3D model. As can be seen, each descriptor (one row per query) yields another set of answer objects.

Low-level features are usually efficient to extract and store, and can be quickly evaluated at query time. Besides global shape description, low-level features have recently been used in approaches attacking the partial-similarity problem. These first identify a set of "salient" or "interesting" local features, which are then matched against each other in a second step [7, 6]. Feature vector descriptors can also readily be used together with relevance feedback and machine learning techniques to improve retrieval effectiveness.

The most important drawback, however, is that low-level features are not aware of higher-level *semantic concepts* underlying the objects or object parts, and that the correspondence between low-level features and high-level semantic concepts is not clear for most of the low-level features. A prerequisite before retrieval in 3D Digital Libraries can take place on the semantic level could be the definition of a catalog of semantic shape features, followed by the development of low-level analysis algorithms which can detect, describe, and compare the identified semantic features. Obviously, this is a most challenging problem in 3D content-based retrieval for the following years and beyond.

4. THE 3D DIGITAL LIBRARY VISION

In Sections 2 and 3, we discussed the current situation in representing, describing, and retrieving 3D content. The results achieved so far are remarkable, yet they raise further research challenges which have to be solved to unlock the full functional potential of 3D Digital Libraries. If the problems in representation and content-based organization of 3D content are solved, new and highly productive 3D applications will emerge. Semantically enriched markup, indexing, and retrieval will allow the deep integration of 3D content into Digital Libraries, and fascinating new applications can be envisioned. We sketch some of them in the following.

Intelligent 3D data acquisition

Intelligent 3D scene acquisition will consist of fully automatic segmentation and interpretation of any scanned scene in such a way that each contained object is recognized, its degrees of freedom are identified, and it becomes readily editable in a way that respects its inherent structure and semantics. Appropriate *shape templates* will be associated with the elements in the acquired scene, semantically enriching the data.

Semantic editing and modeling

Based on the recognition not only of low-level shape features, but also of structure and semantics, new possibilities to work with 3D data will emerge. The separation of function and shape will allow for highly efficient editing operations, where the shape of a model can be instantaneously edited by manipulating a few core high-level parameters. Also, the composition of new 3D objects and scenes based

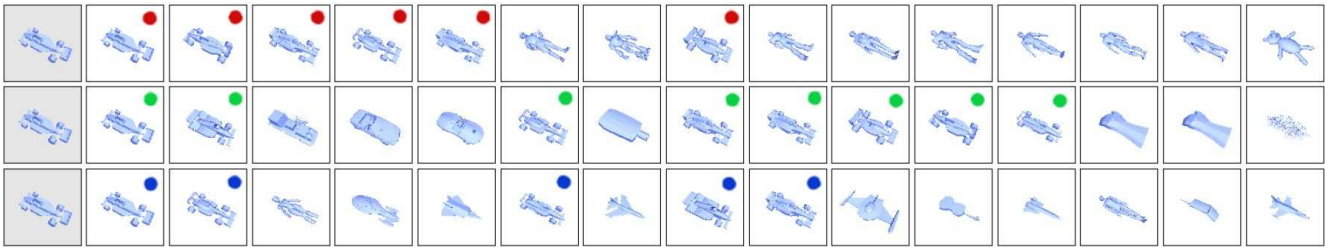


Figure 4: Query-by-example for a Formula-1 racing car model in a 3D repository. Different automatically-extracted low-level object descriptions were used in executing the query, producing different result sets.

on existing content will be greatly simplified, once 3D editors are made aware of semantic properties of the models. A first approach of *modeling by example* [5] illustrates the potential of this paradigm.

Intelligent content-based access

3D search engines will become highly intelligent tools once semantic shape analysis methods are available. If confronted with a user query, the search system will evaluate many different similarity notions on all conceptual levels. From that evaluation, the system will determine the most appropriate similarity notion, and then present the user with the most promising search results. The user will be offloaded from the difficulties in 3D search as currently given, e.g., manual feature selection or supplying much explicit relevance feedback.

Automatic analysis of large 3D collections

Once the representation problems are solved and semantic shape analysis algorithms are available, the fully automatic population of huge 3D Digital Libraries can take place. Content from many heterogeneous sources will be integrated into a decentralized, unified repository. The repository structure will be analyzed, and a suitable organization will be automatically learned from the data.

5. CONCLUSIONS

In this paper, we discussed fundamental aspects and urgent research challenges in 3D Digital Library technology. We argued that based on the technological effects on the production and consumption side, in the near future massive amounts of 3D content will become available. For Digital Library support of these massive data amounts to become effective, a couple of key problems regarding the 3D data level have to be addressed, e.g., in data representation, file format, and stable markup. Furthermore, shape analysis algorithms need to become aware of 3D semantics, to be able to implement advanced automatic organization and retrieval capabilities, and to create large libraries of 3D content that can be effectively searched and accessed. Specifically, low-level features alone are not enough to this end, but defining a catalog of semantic 3D features, and designing algorithms for their robust detection in 3D content are a promising starting point to this end. Once these research challenges are appropriately addressed, Digital 3D Libraries will offer new, highly productive applications in intelligent content acquisition, editing, organization, and accessing.

Acknowledgments

This paper extends previous work by the authors that appeared in IEEE Computer Graphics & Applications in 2007 [8, 1]. The work presented here was supported by the project Probado (www.probado.de) funded by the German Research Foundation (DFG), and by the EPOCH (www.epoch-net.org) and DELOS (www.delos.info) Networks of Excellence funded by the European Commission. It was also partially supported by the Millennium Nucleus Center for Web Research, Grant P04-067-F, Mideplan, Chile (first author).

6. REFERENCES

[1] B. Bustos, D. Keim, D. Saupe, and T. Schreck. Content-based 3d object retrieval. *IEEE ComputerGraphics & Applications*, 2007. To appear.

[2] B. Bustos, D. Keim, D. Saupe, T. Schreck, and D. Vranić. Feature-based similarity search in 3D object databases. *ACM Computing Surveys (CSUR)*, 37:345–387, 2005.

[3] B. Bustos, D. Keim, D. Saupe, T. Schreck, and D. Vranic. An experimental effectiveness comparison of methods for 3d similarity search. *International Journal on Digital Libraries, Special Issue on Multimedia Content and Management*, 6(1):39–54, 2006.

[4] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley-Interscience, New York, 2nd edition, 2001.

[5] T. Funkhouser, M. Kazhdan, P. Shilane, P. Min, W. Kiefer, A. Tal, S. Rusinkiewicz, and D. Dobkin. Modeling by example. *ACM Trans. Graph.*, 23(3):652–663, 2004.

[6] T. Funkhouser and P. Shilane. Partial matching of 3D shapes with priority-driven search. In *Symposium on Geometry Processing*, June 2006.

[7] R. Gal and D. Cohen-Or. Salient geometric features for partial shape matching and similarity. *ACM Trans. Graph.*, 25(1):130–150, 2006.

[8] S. Havemann and D. Fellner. Seven research challenges of generalized 3d documents. *IEEE ComputerGraphics & Applications, Special Issue on 3D Documents*, 27(3):70–76, May/Jun 2007.

[9] N. Iyer, S. Jayanti, K. Lou, Y. Kalyanaraman, and K. Ramani. Three-dimensional shape searching: state-of-the-art review and future trends. *Computer-Aided Design*, 37:509–530, 2005.

[10] M. Pratt. Extension of iso 10303, the step standard, for the exchange of procedural shape models. In *Proc. International Conference on Shape Modeling and Applications (SMI04)*, pages 317–326, June 2004.

[11] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The princeton shape benchmark. In *SMI '04: Proceedings of the Shape Modeling International 2004*, 2004.

[12] J. Tangelder and R. Veltkamp. A survey of content based 3D shape retrieval methods. In *Proc. International Conference on Shape Modeling and Applications (SMI'04)*, pages 145–156. IEEE CS Press, 2004.

EPISTEMIC NETWORKS IN GRID + WEB 2.0 DIGITAL LIBRARIES

MARTIN DOERR
ICS FORTH, HERAKLION
martin@ics.forth.gr

DOLORES IORIZZO
IMPERIAL COLLEGE LONDON
d.iorizzo@imperial.ac.uk

Data-driven research has emerged as a forth paradigm of e-Science, opening up possibilities for the discovery of new knowledge from existing digital resources. Digital libraries need to leverage advances in GRID and Web 2.0 technologies that will drive the infrastructure for building epistemic networks of the future. The development of decentralised core technologies for data integration and co-reference services are essential for creating a sustainable global knowledge network.

Categories and Subject Descriptors:
D.2.12 [SOFTWARE ENGINEERING]:Interoperability
H.1.1[MODELS AND PRINCIPLES]: Systems and Information Theory --- value of information;
H.1.2[MODELS AND PRINCIPLES]:User/Machine Systems ---Human information processing
H.2.2[DATABASE MANAGEMENT]: Logical Design --- Schema and subschema;
H.3.4[INFORMATION STORAGE AND RETRIEVAL]: Systems and Software --- Information networks
H.3.7[INFORMATION STORAGE AND RETRIEVAL]: Digital Libraries
J.5[ARTS AND HUMANITIES]:
General Terms: Theory, Design, Standardization, Management;
Additional Key Words and Phrases: research requirements, cultural heritage, interdisciplinary research, ontology engineering, information integration, epistemic networks, global information access, GRID, Web 2.0.

Data-driven science has emerged as a new model which enables researchers to move from experimental, theoretical and computational distributed networks to a new paradigm for scientific discovery based on large scale distributed GRID networks (OFG, NSF/JISC 2007). The new model is not restricted to the sciences. Hundreds of thousands of new digital objects in multimedia formats are placed on the Web and in digital repositories everyday, supporting and enabling research processes not only in science, but in medicine, education, culture and government. It is therefore important to build infrastructure and web-services that will allow for exploration, data-mining, semantic integration and experimentation across all of these rich resources in large-scale digital libraries where data is properly curated, archived and preserved.

Yet, there is also a growing consensus that traditional libraries and GRID solutions alone are too heavy and administratively burdensome, and that Web 2.0 allows for the development of a more light-weight service oriented architecture that can adapt readily to user needs by using on-demand utility computing, such as mash-up's, surf clouds, annotation and tagging, knowledge sharing, social networks, and automated workflows for composing multiple services. The goal is not just to have fast access to information over distributed networks, but to have the capacity to create new digital resources, interrogate data and form hypotheses about its meaning and wider contexts. (Lagoze 2005) As librarians working in e-Science have increasingly perceived, digital libraries need to dramatically extend the role of traditional libraries by encouraging *collaboration* (allowing users to be both producers and consumers by contributing knowledge actively through annotations, reviews, comments) and *contextualisation* (users expanding the web of inter-relationships and layers of knowledge that extend beyond primary sources). (Borgman 2003)

Clearly what needs to emerge is a mixed-model of GRID + Web 2.0 solutions for digital libraries which creates an *epistemic network* that supports a four step iterative process: (i) retrieval, (ii) contextualisation, (iii) narrative and hypothesis building, (iv) creating contextualised digital resources in semantically integrated knowledge networks to enable new discoveries and social networks. What is key here is not just managing the amount of new data in a digital library, but the capacity to interrogate, contextualise, share and order *existing* resources in a semantically accessible form that creates new knowledge.

Peter Murray-Rust (Murray-Rust 2007) points to a prime example of a scientific discovery that emerged from the re-use of existing resources: Mendeleev's Law of Periodicity: "The law of periodicity was thus a direct outcome of the stock of generalisations and established facts which had accumulated by the end of the decade 1860-1870; it is an embodiment of those data in a more or less systematic expression."

Mendeleev's law emerged from a *concatenation of facts* extracted from the current published chemical literature which appeared in many languages and symbolic formulations; the analysis of *relations* in the data and metadata – the experimental conditions – were critical for establishing his conclusion. Murray-Rust's thesis that *'the current scientific literature, were it to be presented in semantically accessible form, contains huge amounts of undiscovered science'* demonstrates the urgency of developing core digital

library technologies that will allow us to make similar discoveries with existing digital resources.

The core technologies we see as most critical for the development of digital libraries as epistemic networks are:
(i) Data Modelling, Core Ontologies and Document Retrieval by Complex Associations
(ii) Data Integration and Concatenation of Facts for Knowledge Discovery
(iii) Knowledge Management based on Co-reference Services

(i) DATA MODELLING AND CORE ONTOLOGIES FOR COMPLEX RETREIVAL

The single most important obstacle to achieving semantic integration and the contextualisation of information in digital libraries is the fact that traditional digital library metadata repositories do not model contextual relationships. The representation of c*ontent* as well as *context* in digital resources must rely on a generic, or nearly generic, information model. The prevailing assumption has been that a generic 'top-down approach' is required for the semantic integration of digital libraries, but the 'top-down' strategy has proved limited. Generic solutions are generally quick and cheap, but they have a short life span, as can be seen with statistical methods of information retrieval, and *the hypertext model*. The 'top-down' approach is intrinsically short-sighted since its initial conceptualisation can never anticipate future problems and therefore will never be a long-term solution.

An example of the shortcomings of the 'top-down' approach for digital libraries is the Dublin Core metadata element set (DCMI 2006). It is an excellent simplification of bibliographic information that provides a unified data structure for all kinds of materials. However, when more and more cases are squeezed under the same umbrella, so that quite a lot of domain specific interpretation of seemingly common metadata elements become mutually incompatible, then the usefulness of DC brakes. Attempts to fix the problem with 'qualified' Dublin Core Elements only increased heterogeneity so in the end 'qualified DC' was abandoned by the DC Consortium.

However, generic solutions need not be 'top down'; they can also be 'bottom up'. The CIDOC CRM is a 'bottom up' information model that starts from the analysis of real research scenarios and practices of information management in different domains. Our model is based on deep knowledge engineering across disciplines that generalises domain specific cases in order to find the most generic ontological structures and generic processes across multiple domains. The CIDOC CRM is a *'core ontology'* that abstracts hundreds of schemata used for documentation in various museum disciplines into 80 classes and 130 relationships, yet we have found that less than 5% of its concepts are museum specific. It is not huge and messy, but small, compact and focused on *contextual relationships* not objects in isolation. The CRM represents generic kinds of discourse, such as location, participation, part-whole composition, and reveals generic structures that integrate both factual and categorical knowledge in a way that is useful for very specific applications.

Three ideas are central to the CRM: a) The relationship between entities and the identifiers that are used to refer to the entities (including ambiguity of reference) are part of a historical reality that is to be documented, therefore, the CRM distinguishes nodes representing real-world items from nodes representing names *per se*; b) Types and classification systems are not only a means of structuring information about reality, but also represent the historical past as a human construct; c) the CRM analyses the past by dividing it into discrete events. The documented past can be formulated as events involving "Persistent Items" (continuants or endurants) (Crofts et al. 2005) both material (Ceasar, Lucy) and immaterial (The Empire, Hominid). Material and immaterial items can be present in events either through physical information carriers or as concepts.

From this point of view, a picture of history emerges as a network of lifelines of persistent items meeting in space-time events (fig.1). This abstraction turns out to be extraordinarily powerful. Many intuitive relationships are analyzed in terms of events, such as "has creator" or "has origin". With a minimal schema, there arise a surprising wealth of inferences and any event can be described by the CRM. For instance: the life of Caesar.

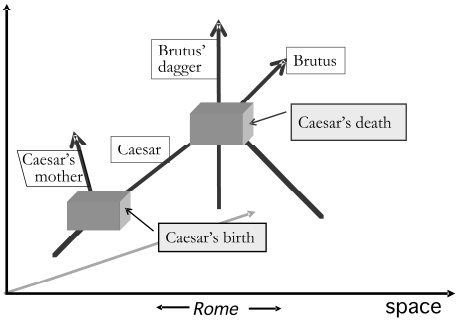


Figure 1: Historical events as meetings of things and people

Complex genetic family relations can be represented by birth events including a father and a mother. The "friend of a friend" application (FOAF) can be based on co-authoring and other common events between people. Influences on lives and achievements can be traced to people meeting or communicating with other people, and the development of ideas, theories and discoveries that lead back to them. Chronologies can be justified by the causal ordering of events. (Doerr , Plexousakis, et al 2004) Experimental knowledge in the sciences is gained by actual human experiments that are carried out by individuals and teams of researchers in space/time; they can be documented as events, independent of subject matter. Calculating statics of bridges or climate models are not covered by ontologies but they can be documented as events. Descriptive sciences, like geosciences and biodiversity

studies, gain knowledge by collecting an immense number of observations carried out by individual scientists and research teams, which can be described as events on a human scale connected to people and ideas. Embedded in all metadata that is stored in libraries, including digital libraries, there is an historical perspective which can be represented as events from which new knowledge can be gained.

(ii) DATA INTEGRATION AND THE CONCATENATION OF FACTS

The CIDOC CRM has developed a model which semantically connects documents in a way that is diametrically opposed to the *hypertext paradigm*. Using a minimal but central part of the CIDOC CRM as an example, we elaborate the problem of extracting knowledge from the contents of documents, metadata and links between documents, into a coherent semantic network. The semantic power of the CRM can be shown with minimal ease by demonstrating how with employing only 3 Classes and 2 Properties from the CRM a network of deep relations can emerge: E5 Event, *P12 occurred in the presence of*; E77 Persistent Item (*Persistent Item* comprises material and immaterial things, including persons); E5 Event. *P7 took place at*; E53 Place. Consider the following data and metadata records:

The State Department of the United States holds a copy of the Yalta Agreement. One paragraph begins, "The following declaration has been approved: The Premier of the Union of Soviet Socialist Republics, the Prime Minister of the United Kingdom and the President of the United States of America … jointly declare their mutual agreement to concert …" (Halsall 1997).

A Dublin Core record about this may read:
*Type:*Text
Title: Protocol of Proceedings of Crimea Conference
Title.Subtitle: II. Declaration of Liberated Europe
Date: February 11, 1945.
Creator:
The Premier of the Union of Soviet Socialist Republics
The Prime Minister of the United Kingdom
The President of the United States of America
Publisher: State Department
Subject: Post-war division of Europe and Japan

Figure 2: Allied Leaders at Yalta

The Bettmann Archive in New York holds a world-famous photo of this event (fig 2).
A Dublin Core record of this image might be:
Type:Image
Title: Allied Leaders at Yalta
Date: 1945
Publisher:United Press International (UPI)
Source: Wikipedia
References: Churchill, Roosevelt, Stalin

Another piece of information comes from the Thesaurus of Geographic Names [TGN], which may be captured by the following data:

TGN Id: 7012124
Names: Yalta (C,V), Jalta (C,V)
Types: inhabited place(C), city (C)
Position: Lat: 44 30 N,Long: 034 10 E
Hierarchy: Europe (continent) <– Ukrayina (nation) <– Krym (autonomous republic)
Note: Located on S shore of Crimean Peninsula; site of conference between Allied powers in WW II in 1945; is a vacation resort noted for pleasant climate, & coastal & mountain scenery; produces wine, canned fruit & tobacco products.
Source: TGN, Thesaurus of Geographic Names

It has long been recognized that the *only element common to all of these records is the date '1945'*; that is why a DC-based or Google search for 'The Yalta Agreement' will never be adequate, since contextual relationships are not represented in their data models.

The information from these three sources can be represented as instances of 3 Classes and 2 Properties of the CIDOC CRM:

(1) *Crimea Conference (E5)*
P12 occurred in the presence of
The Premier of the Union of Soviet Socialist Republics (E77)
The Prime Minister of the United Kingdom (E77)
The President of the United States of America (E77)
Protocol of Proceedings of Crimea Conference (E77)
(2) *Allied Leaders at Yalta (E5)*
P12 occurred in the presence of
Stalin (E77)
Churchill (E77)
Roosevelt (E77)
Photo of Allied Leaders at Yalta (E77)
P7 took place at
Yalta (E53)

(3) *Yalta Conference (E5)*
P12 occurred in the presence of
Allied Powers (E77)
P7 took place at
Yalta(E53)

Resolving in sequence the different ways of referring to the same items, the uncorrelated parts will collapse into a single epistemic network, which connects the text, the image, the place and the people through the historic event:

(4) *Yalta Conference (E5)*
P12 occurred in the presence of
Stalin, Premier of the Union of Soviet Socialist Republics (E77)
Churchill, Prime Minister of the United Kingdom (E77)
Roosevelt, President of the United States of America (E77)
Protocol of Proceedings of Crimea Conference (E77)
Photo of Allied Leaders at Yalta (E77)
P7 took place at
Yalta(E53)

If we collect enough related events, even this rudimentary schema already creates a powerful network for recovering biographical and contextual data about *people*, *documents*, *objects*, and *places*. What we learn from this example is: a) A knowledge network must be built on suitable ontological abstractions that support relevant *contextual relationships* which can be surprisingly simple yet powerful; b) *Advanced reasoning* cannot take place if the elements of the network are not connected. They connect through the domain and range values of the relations that identify items in a domain of discourse. Since identifiers are not usually unique and therefore do not match up, then importance of "duplicate removal" or *co-reference* detection as a process is critical, even though its importance in largely unrecognized in the research on information integration; c) Knowledge about *relationships* comes from the document – either from its proper contents or its "metadata". What actually relates the documents is not a "*hyperlink*", but the fact that they refer to the very same things. These may be events, dates, places, persons, material or immaterial things such as texts, images, names etc. Since the connecting facts are not revealed in the hyperlink, the hypertext model is *fundamentally limited* to manual navigation.

Equally misleading seems to be the paradigm of a document as a "digital surrogate" of a real world item, which is one of the motivations for the RDF syntax. There is a problem however about which of the documents, out of all the documents, about a real world item, should become the surrogate; how should the competition between the properties of the surrogate and the thing itself be resolved?

We suggest that appropriate "digital surrogates" of real world items should be modeled as *surrogate nodes* external to the documents, with no *necessary property* except an identity. The *relations* between surrogate nodes should be seen as *extractions* or summarizations from the documents (see fig. 3); let us call these *facts*. By *facts* we mean the instances of relationships (or 'properties' in the terminology of OWL and RDFS). Constructs like "reification" in RDF and other argumentation models (Roux 2004) make explicit the link between the source provided and the relation.

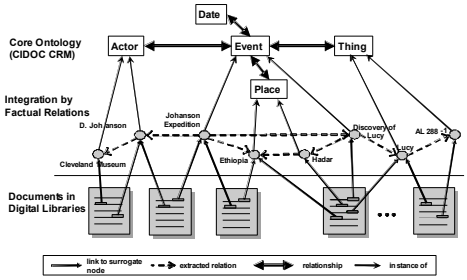


Figure 3: Relations as document summarization

(iii) KNOWLEDGE MANAGEMENT BASED ON CO-REFERENCE

As we have shown, even if we have a global schema, and the means to provide factual relations, one important feature is missing to build a network: *co-reference*. (Levesque 1984) How do we know if two relations relate to the same real world item? This important problem has received little attention, although it is a core technology for epistemic networks. We propose a novel solution to achieve the long-term, scalable integration of facts that will provide knowledge management based on co-reference.

Traditionally, librarians have invested heavily in so-called authority files or knowledge organisation systems (KOS) which register names and characteristics of authors and other items and associate them with a preferred representation in a central resource, and then advise colleagues to use the central resource as a reference to obtain unique identifiers. (Patel et al 2005) In one respect, this does *not solve anything* since we still cannot determine if a local source refers to an item also listed in the KOS. In another respect, the approach has been partially successful. The descriptions increase the chance that an expert of the local source can recognize the item, use the identifier, and then pass this on to colleagues that will also use the identifier for the same item. But using a central resource causes serious scalability problems. Even worse, different communities in different countries tend to create their own authority files with overlapping content, so there is no international process for data integration.

In order to create a truly global knowledge network, one could take advantage of the power of Web 2.0 by creating a social epistemic network, engaging the general public as well as experts, that would publish and preserve each and every detected co-reference together with its sources. We suggest setting up a *Web 2.0 Co-reference Service* supported by a grid service oriented architecture for digital libraries, so that anyone anywhere can publish a co-reference along with its source data to preserve referential integrity; this would achieve more than any single authority file and have an international scope. The epistemic network will grow simply through the efforts of the users. Nothing like this exists on the Web at the moment but it is potentially a way of engaging the public, as in wikipedia, to play a large role in building a global epistemic network.

Intuitively, co-reference should be transitive and form equivalence classes (Levesque 1984) that could scale up to any size. In order to relate the elements of an equivalence class of cardinality *v*, a minimal number of *(v-1)* primary equivalences is needed to derive all *v(v-1)/2* equivalences. This demonstrates the economic power of preserving co-reference knowledge once the networks grow tighter. Each equivalence class can be identified with a *surrogate node* as described above. Co-reference links can then be implemented indirectly as links to a common surrogate node.

We suggest that co-reference detection must be a semi-automatic process within a Web 2.0 service. Massive participation of scholars *qua* experts in this process will be essential since it often requires specialised knowledge and should not be left simply to automated guesswork. As a matter of good practice, it should become a personal product of scholarly research that is properly documented.

There are also economic benefits since data integration is expensive. Mathematical models could be developed that would estimate the time it takes to carry out integration activity and offer a cost-benefit analysis. Further, formal foundations of "data cleaning" could be investigated, such as: to what extent does the propagation of co-reference knowledge allow for inferences or assumptions about other co-references via related facts etc.? Finally, mathematical models could be used to develop effective strategies in peer-to-peer networks of co-reference detection and monitoring of global consistency. A "knowledge economy" would emerge that ensures the long-term integration of digital repositories by preserving knowledge about co-reference.

This idea is radically new, in four respects: (i) The ultimate authority for identifier equivalence are people – the witness or the expert – with knowledge of the two contexts that are to be connected. Co-reference is a valuable element of knowledge that comes at high cost, therefore it should be *curated* and *preserved* for future information systems; (ii) The model suggests that several current approaches of ad-hoc data cleaning and central authorities are ineffective and miss an important part of the problem: *the preservation and control of real-time detected co-references*; (iii) The co-reference model can be implemented in a completely distributed "democratic" manner. Therefore, in contrast to other approaches, it is completely scalable and imposes minimal constraints on the kind of organisation in which it will be implemented; (iv) Problems surrounding *co-reference* act as a perfect proof-of-concept for how Grid + Web 2.0 technologies can be combined to form epistemic networks and provide solutions to the global knowledge management crisis.

Implementation

How can these facts be created in an efficient way? The problem is that a generic model does not suggest what to document in any specific case, it only sufficiently explains what has been documented. It requires constant abstract thinking to match generalizations to specific problems, even though the generalizations are quite obvious after one sees them. For instance, in the CRM finding an object (as in archaeology) would be represented as activity in which an object is present. This abstraction is sufficient for most inferences about an archaeological find. The activity type "finding" would be a term entered as data, but not as part of the core model. An archaeologist entering data however would like to see a field reminding him to enter where and when an object was found. Similarly, other disciplines will have other special things to include. Hence, data entry forms should normally be more application specific than the generic model, even if they are designed to capture data for instantiating the generic model.

It is also good practice for a researcher or documentation specialist to preserve the enriched information unit as a whole, both in order to maintain authorship of the information unit and for future revisions. If data is directly entered into a global semantic network and all knowledge is merged, then the original units are lost. Many traditional relational database schemata are not immune to this criticism. Preserving information units allows an association to be made between them and the people who understand their interpretation and other relevant knowledge, thereby verifying the quality of the contents.

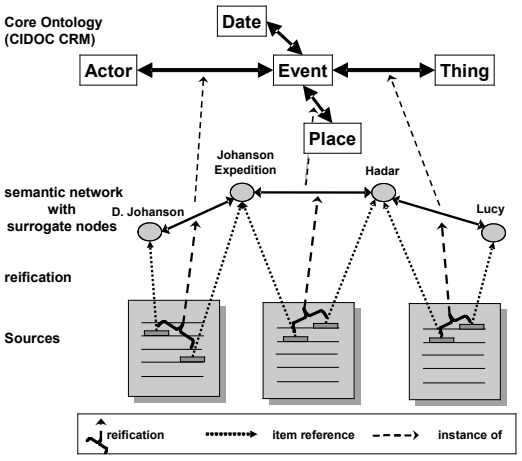


Figure 4: Semantic network linked back to sources via reification links

Finally, large monolithic resources are more sensitive to complete corruption and therefore cause more problems for digital preservation than distributed units. Therefore we propose to logically separate documentation units and primary sources from the network level, and instead to derive data for the network level from the documentation units and primary sources. Duplication of information establishes good practice for digital preservation.

We distinguish three possible architectures to achieve this separation. They have different performance characteristics, but can easily be combined for optimisation purposes: (1) warehousing, (2) mediation services, (3) mixed –model.

1. In a data warehouse-style, facts can be extracted from sources and physically aggregated in a semantic network. The extracted facts directly connect the surrogate nodes (fig. 4). In order to update the network when sources change, it may be necessary to introduce reification statements or similar mechanisms linking facts to their sources. This strategy makes querying, especially joins and deductions, across resources very fast. Updating is more difficult, since individual facts may have multiple sources. Maintaining reification links is relatively expensive. It becomes even more complex when co-reference statements are added and linked to the surrogate nodes. On the other hand, physically (on a dedicated system) creating the network provides more flexibility to actually detect co-reference relations (Doerr, Schaller et al. 2004), because extraction and aggregation can be done in complex processes. Finally, semantic networks are not scalable, or at least no scalable architecture has yet been successfully proposed.

2. Sources are interpreted by a mediation service (Wiederhold 1992). For instance, queries are formulated in terms of the global model and transformed according to the different source models to bring back results conforming to the global model (Calvanese et al. 1998) (fig.5). Assuming mainly a local-as-view (LAV) approach (Cali 2003), this is only possible if the sources have a data structure which can be mapped to the global model. The performance may depend on the degree of heterogeneity of the local source to the global model. For mediation services, it is more difficult to resolve co-reference relations, because queries are expected to be answered in real-time. This would change completely if explicit co-reference relations were available. Joins and deductions are more costly, and require larger temporary computer memory, but with mediation services there is no update problem at all.

3. Whereas the above solutions have been widely discussed in the past decade, we propose here yet another variant. Extracted local facts are represented in terms of the global model as summarization metadata units, which are preserved and remain connected to their sources. Then, co-reference relations could be described by linking to the surrogate nodes the corresponding local nodes, which in turn are linked by local facts (figure 6). The surrogate nodes could, for instance, be implemented by one-to-many XLinks. This strategy doubles the path lengths in the network and makes querying slower, but it has the advantages of avoiding both heterogeneity and reification, and of offering a scalable solution without central update problems.

We suggest that the architectures described above deserve more research about the precise conditions under which they would be most effective, both singularly and in combination. Obviously, querying data paths in solution 3) is more effective the lower the density of co-reference relations relative to the number of local nodes, and the lower the multiplicity of identical facts between the metadata units, because it requires less joins across metadata units.

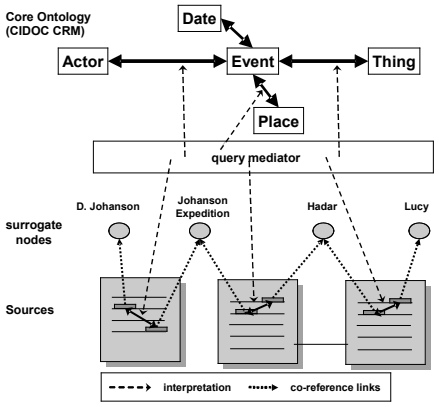


Figure 5: Query mediator interprets source relations

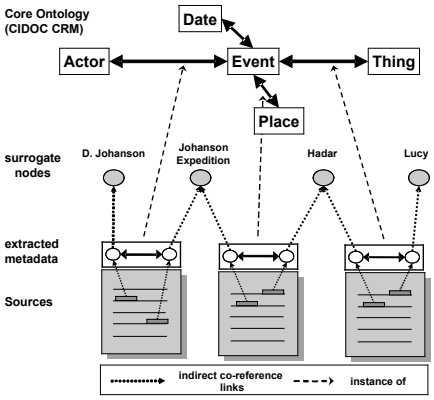


Figure 6: Metadata connected to sources and indirect co-reference links

We have nothing against introducing some limited heterogeneity in solution 3, so that solution 3 and 2 become more similar. In new systems, one could design the data structures of local sources with minimal heterogeneity and published mappings to the global model. In local environments with a low update rate, solution 1 may be most effective, if reification can be simplified. Then, a complete semantic network could take the logical place of a metadata unit in solution 3. Under this aspect, solution 3 could indicate a way to make semantic networks distributed. These are only examples of how these architectures could be combined to produce far more flexible and generic solutions for information integration. For any distributed solution, especially grid-enabled, research about effective indexing would also be a major issue that needs attention. (Podnar et al. 2006).

Solution 3 is particularly suited to natural language processing techniques for knowledge extraction from free text. The CIDOC CRM has a nearly "linguistic" structure and makes this task relatively easy (Genereux & Niccolucci 2006). In particular the event model maps easily to phrases containing action verbs. We suggest that more research should be invested in extracting event-based metadata by semi-automatic methods from free-text. Far too little attention has been focused on this important problem. (Vincent 2005).

In conclusion, we suggest that peer-to-peer networks and GRID technology can provide an effective infrastructure for next generation digital libraries. The use of DataGRIDs will be essential (i.e. nodes with uniform access protocols which can be accessed automatically to follow associations in the way a human would browse the web, thereby collecting concatenated facts and other relations), since they will enable advanced semantics within the emerging global network to perform automated reasoning for executing precise inferences, both categorical and factual, currently impossible on a large scale. This networked infrastructure will support various online services to create a dynamic GRID + Web 2.0 epistemic network that will publish and preserve co-references, create distributed indices, control and monitor consistency, and manage convergence to higher states of integration. This new model of a digital library makes possible advanced reasoning over distributed resources on a global scale, and hence opens up new opportunities for uncovering new discoveries, like those of Mendeleev, from existing resources.

[1] ABERER, K. (et al) 2004, Emergent Semantic Systems, ICSNW 2004, *LNCS* 3226, M. BOUZEGHOUB et al. (Eds.) 14-43.

[2] BILENKO, M., AND MOONEY, R.J. 2003. Adaptive Duplicate Detection Using Learnable String Similarity Measures. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003), Washington DC, August 2003, 39-48.

[3] BORGMAN, C.L., 2003. 'The invisible library: Paradox of the global information infrastructure', Library Trends, 51 (4), 2003.

[4] CALI, A. 2003 Reasoning in Data Integration Systems: Why LAV and GAV are Sibilings. In *Proceeding ISMIS 2003, Lecture Notes in Computer Science* 2871, SPRINGER 2003, 562-571.

[5] CALVANESE, D. (et al) 1998. Description Logic Framework for Information Integration. In *Proceedings of the 6th International Conference on the Principles of Knowledge Representation and Reasoning (KR'98),* 2-13.

[6] CIDOC CRM 2006, The CIDOC Conceptual Reference Model. http://cidoc.ics.forth.gr/official_release_cidoc.html

[7] CROFTS, N., DOERR, M., GILL, T., STEAD, S., AND STIFF M. 2005. Definition of CIDOC CRM http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.doc.

[8] DCMI 2006. Dublin Core Metadata Initiative, "Making it easier to find information" http://dublincore.org/.

[9] DEGEN, W., HELLER, B., HERRE, H., AND SMITH, B. 2001. GOL -Towards an Axiomatized Upper-Level Ontology. Electronics and Computer Science.

[9] DODDS, L 2004. Intro. to FOAF http://www.xml.com/pub/a/2004/02/04/foaf.html

[10] DOERR, M., PLEXOUSAKIS, D., KOPAKA, K., AND BEKIARI, C. 2004. Supporting Chronological Reasoning in Archaeology, Proc. of Comp. Applications and Quantitative Methods in Archaeology, CAA2004, Prato, Italy, 2004. http://www.ics.forth.gr/isl/publications/paperlink/caa2004_supporting_chronological_reasoning.pdf.

[11] DOERR, M., SCHALLER, K. AND THEODORIDOU, M. 2004. Integration of complementary archaeological sources. Proceedings of Computer Applications and Quantitative Methods in Archaeology, CAA 2004, Prato, Italy, 2004. http://.ics.forth.gr/isl/publications/paperlink/doerr3_caa2004.pdf.

[12] GENEREUX, M. AND NICCOLUCCI, F. 2006. Extraction and Mapping of CIDOC-CRM Encodings from Texts and Other Digital Formats. In The e-volution of Information Communication Technology in Cultural Heritage, D. Arnold, et al, Eds. Short papers from the joint event CIPA/VAST/EG/EuroMed 2006, pp. 56-61.

[13] HALSALL, P. 1997. Modern History Sourcebook: The Yalta Conference, http://www.fordham.edu/halsall/mod/1945YALTA.html.

[14] LAGOZE, C., KRAFFT, D. B., PAYETTE, S., AND JESUROGAI, S. 2005. What Is a Digital Library Anymore, Anyway? D-Lib Magazine, Volume. 11, Number 11.

[15]LAKOFF, G. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Univ. Chicago Press.

[16] LEVESQUE, H.J. 1984. Foundations of a functional approach to knowledge representation. Artificial Intelligence, Volume 23, Issue 2, 155–212.

[17] MURRAY-RUST, P., Data Driven Science – A Scientist's View, NSF/JISC 2007 Digital Repositories Workshop, http://www.sis.pitt.edu/~repwkshop/papers/murray.html

[18] OGF 2007 and NSF/JISC Digital Repositories,, http://www.ogf.org/OGF20/events_ogf20.php (Hey) and http://www.sis.pitt.edu/~repwkshop/index.html for discussions of data-driven science as 4th paradigm.

[19] PATEL, M. (et al) 2005. Semantic Interoperability in Digital Library Systems, DELOS–deliverable 5.3.1, June 2005.

[20] PODNAR, I, LUU, T., RAJMAN, M., KLEMM, F. ABERER, K. 2006. A Peer-to-Peer Architecture for Information Retrieval Across Digital Library Collections. In *Proceedings of the 10th European Conference, ECDL 2006*. SPRINGER 2006.

[21] ROUX, V. AND BLASCO, P. 2004. Logicisme et format SCD: d'une epistemologie pratique a de nouvelles partiques editoriales Hermes. CNRS-editions.

[22] VINCENT, K.P. 2005. 'Text mining methods for event recognition in stories', Knowledge Media Institute, The Open University, Milton Keynes, UK, Tech Report KMI-05-2 April 2005. http://kmi.open.ac.uk/publications/pdf/kmi-05-2.pdf.

[23] WIEDERHOLD, G. 1992. Mediators in the Architecture of Future Information Systems. *IEEE Computer*, March 1992.

Leveraging on Associations – a New Challenge for Digital Libraries

Martin Doerr
ICS-FORTH
Heraklion-Crete
Greece
+30 2810 391625
martin@ics.forth.gr

Carlo Meghini
CNR-ISTI
Via G. moruzzi, 1
Pisa, Italy
+39 050315 2893
carlo.meghini@isti.cnr.it

Nicolas Spyratos
University Paris-South
LRI – Bât 490
91405 Orsay Cedex, France
+33169156586
spyratos@lri.fr

ABSTRACT

Decades of research have been devoted to the goal of creating systems which integrate information into a global knowledge network. On the other side, Digital libraries have not overcome the traditional paradigm of delivering a document as ultimate objective. This paper argues that next-generation DL services must be built on accessing associations implicit or explicit in document collections and their metadata. It suggests a new approach to leverage associations based on (i) generic core ontologies of relationship and co-reference links (ii) semi-automatic maintenance of co-reference links by a new kind of service, and (iii) public engagement in the creation and development of the emerging association network.

1. INTRODUCTION

The Web has become an indispensable tool of modern culture. Powerful, but relatively crude search engines organise the enormous amount of information on the internet into simple answers to clear cut, search term-based, questions. What is deceptive about this everyday process is that it flattens rather than deepens and improves knowledge. Research questions which require more than immediate information are thwarted. For instance, we can easily find documents on the Web about Lucy, the hominid, but we have no direct way to discover the locations of finds similar in kind. Even though information on the web is densely linked - the average distance between documents is only 7 successive links [2] - the information itself is not related in a meaningful way. Hypertext links are made for human readers, rather than for machine interpretation. Digital libraries have not overcome the traditional paradigm of delivering a document as ultimate objective. Carl Lagoze states that "..the underlying public key infrastructure that was seen as 'essential to the emergence of digital libraries' remains undeveloped. Despite efforts of the W3C's Semantic Web initiative, the holy grail of semantic interoperability remains elusive" [8]

First International Workshop on ``Digital Libraries Foundations'' Vancouver, British Columbia, Canada, June 23, 2007.

This paper argues that next-generation DL services must be built on accessing associations implicit or explicit in documents collection and their metadata. It suggests a new approach to leverage associations based on (i) generic core ontologies of relationship and co-reference links, (ii) semi-automatic maintenance of co-reference links by a new kind of service, and (iv) public engagement in the creation and development of the emerging network

2. ASSOCIATIONS AND IDENTITY

The ultimate goal of users is not to get an object but to *understand* a topic. Understanding is built on associations. Associations are found in digital objects or metadata. Metadata provide explicit associations in the form of relationships and data paths. Tools may extract associations from digital objects, either by interpretation of data structures or by statistical means such as evaluation of co-occurrence patterns, and save them again as metadata. Indices provide associations, and may also be seen as metadata.

The topic of associations has been faced both in the area of information retrieval and hypertext for many years and the following kinds of associations are widely used in Digital Libraries: Subject relations between documents and classes; subsumption of classes; hypertext links between documents; occurrence and co-occurrences of words. The latter two have weak semantics. There is a vast literature about statistical detection of associations in order to cluster documents by some co-occurrence patterns in the contents. They are mainly used to find similar documents, but *not to exploit the meaning* of the detected associations for understanding a topic. Ontology learning or automated thesaurus construction is a notable exception, but the semantics of the retrieved associations are generic (on a categorical level) and still very weak for subsequent reasoning. Even refined semantics of hypertext links have not brought any break-through in terms of topic-related automated reasoning so-far. It is hard to create powerful expressions from a combination of hypertext links for other purposes than getting documents and automatically following hypertext links readily retrieves the whole Web.

If the semantics of represented relationships are explicit, such as part-whole, membership, creation and participation, then patterns in the network of factual relations (or *material facts* [4]*)*, can reveal new, indirect associations, or can be used for inductive reasoning. There are many relevant applications, in which retrieval and discovery of digital objects themselves is based on simultaneous discovery of indirect associations, such as searching

for related literature based on co-citation [13], based on co-authorship networks ("friend of a friend", [5]), or search for business relations of dependent enterprises. Recently, Amit Sheth has stressed the extraordinary importance of access by factual relationships for the Semantic Web, in particular with respect to business applications [3]. The challenge is *not just to deliver* documents, but to *leverage on* the latent knowledge in the *combined* content of many digital sources.

Factual relations however can form meaningful semantic networks. In order to support any advanced services, relationships (i.e. classes of relations) should conform with a schema or ontology. Even though it is widely believed that there is no global ontology, the acceptance of Dublin Core demonstrates the opposite. If there is one or a few core ontologies, does not make any difference in their ability to give rise to global networks of knowledge. Empirical studies show [10] that the number of relationships in ontologies is orders of magnitudes smaller than that of classes and hence quite manageable. [6], [7], [14] have shown that a core ontology of ten to a hundred relationships can capture semantics of data structures across many domains.

Now, little advanced reasoning can take place if the elements of the network are not connected. They connect through the domain and range values of the relations that identify items in a domain of discourse. The identifiers are normally not unique and therefore don't match. This "duplicate removal" or *co-reference* detection [9] as co-reference is a process widely underestimated in importance for information integration. What actually relates propositions and other contents found in the documents *is not a "hyperlink"*, but the fact that they refer to the *very same items*. These may be events, dates, places, persons, material or immaterial things such as texts, images, names etc. Even terms can often be seen as (conceptual) items of discourse, rather than as expressions of classification. We argue that the actual semantics linking items are *in* the document, and not between them.

So the key to more advanced services seems to be the unique identification of things. The "bad news" is the immense number of things referred, orders of magnitude larger than the number of terms. We suggest a completely different approach: In order to connect facts, an automated system needs not know any detail about the referred items besides that they are identical. *Wherever* the knowledge comes from, it does the job. So, equivalence clusters of explicit co-reference links between respective document parts or elements of database records can replace maintenance of identification data as traditionally done in authority files. This approach is more general, since the former can be generated from latter, but not vice-versa. Therefore we propose a new kind of DL service: *Co-reference Services (CRS)*.

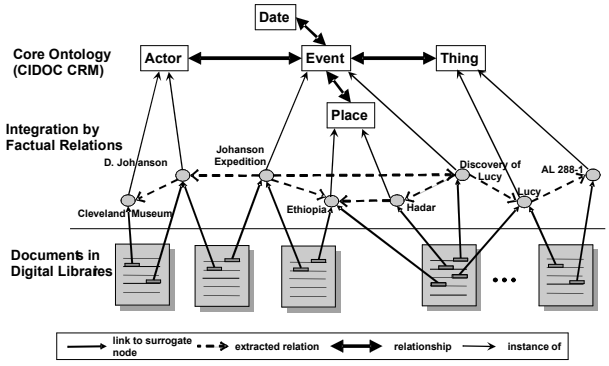
3. ABOUT CO-REFERENCE INFORMATION

Librarians and others have invested heavily in so-called authority files or knowledge organisation systems (KOS) [12], which register names and characteristics of authors and other items and associate them with a preferred representation in a central resource, and then advise colleagues to use the central resource as a reference to obtain unique identifiers. But using a central resource causes serious scalability problems. The standardization process always lags behind reality. Computer scientists tend to regard the recognition of co-reference (duplicate detection) as a question of probability that two items are referred to by similar names or similar properties (e.g., [1]). What is common to both approaches is the fact, that they do not preserve actual knowledge that an identifier a1 in source s1, and an identifier a2 in source s2, refer to the same real-world item. Only very recently, the project VIAF [11] has engaged in correlating two authority files with some nine million person descriptions into what they call a "virtual authority file" by a kind of co-reference links.

If we make the assumption that the maintainer or creator of s1 knows what a1 means, and the maintainer or creator of s2 knows what a2 means, both could *convene* and record the fact of co-reference without any common attribute or authority file. Philosophically, there is only one primary source for the identity of something: a citation in a document or data record field, i.e. "what the author meant by this expression". An record in an authority file poses the same question. All other questions of identity can be seen as elements of the subsequent co-reference problem. If the authors cannot be queried, one may base assumptions about co-reference on known common features of the citations under investigation. Those features may be based on values, such as a common name for the birthplace of a person, which are in turn subject to a co-reference question. Automated data cleaning methods work on the latter base.

Figure 1. Insert caption to place caption below figure

Obviously, co-reference is a question of belief based on explicit or



implicit knowledge and evidence. Therefore we regard a co-reference statement as an elementary piece of scientific or scholarly knowledge, regardless of any heuristic-based software assisting in the identification process. Each co-reference statement allows for the connection of all factual relations to the two identifiers involved.

Intuitively, co-reference should be transitive and form equivalence classes that could scale up to any size. In order to relate the elements of an equivalence class of cardinality *v*, a minimal number of *(v-1)* primary equivalences is needed to derive all *v(v-1)/2* equivalences. This demonstrates the economic power of preserving co-reference knowledge once the networks grow tighter. Each equivalence class can be regarded as a *digital surrogate node* for the referred item. The global number of *surrogate nodes* per real item may be used as an inverse measure for the degree of integration of knowledge sources.

So, if we *publish* a co-reference statement and preserve the referential integrity, we have achieved more than any authority file: we have connected facts from two information assets *to our best knowledge*. (See figure). In contrast to hypertext links, this

information can have a tremendous impact on computer-supported reasoning. A major short-coming of query mediator approaches [15] to information integration is the difficulty to match identifiers on-the-fly. Data warehouse approaches or metadata harvesters are more flexible in this respect, but not as scalable. Explicit co-reference information could close the gap and allow for highly performant hybrid information integration system, i.e. configurations seamlessly including physical and virtual integration systems of metadata

4. CO-REFERENCE SERVICES

We have started to elaborate theoretical foundations for co-reference services, which will be published soon. It has also been subject of several recent applications for European research grants. We present here the general requirements for the envisaged services:

1. A Co-reference Service should be based on common protocols and standards for information access and integration. Webservices in a data GRID environment could provide a beneficial environment.
2. Co-reference links should be persistent and public so that investment pays off. They may be bidirectional or unidirectional. In the latter case harvesting should be foreseen to create the appropriate inverted indices (see 7.). The use of preferred identifiers from an authority file or gazetteer can be seen as a special case of unidirectional linkage, as long as their persistency is guaranteed.
3. Primary Co-reference links should be provided and maintained (curated) by teams having the expertise to assess their correctness, such as librarians, archivists, scholars scientists. Therefore they should be preserved in local, distributed databases ("indices").
4. Social tagging should mobilize the potential of general users and domain experts to enhance and verify co-reference information. Scholars use to spend a large part of their research efforts to collecting and verifying co-reference information. Not all co-reference information is relevant. Social tagging can also create an emergent notion of relevance.
5. Co-reference links must be associated with belief values. Experts distinguish belief values, and trust in sources may differ. Belief values should be used to control precision and recall of retrieval following co-reference links.
6. Duplicate-detection algorithms can be used to populate co-reference indices. Appropriate belief values should distinguish automated from manual sources. Generic Webservice protocols and formats could be beneficial to run intelligent duplicate detection in GRIDs. Duplicate detection algorithms can benefit from co-reference indices.
7. The envisaged open environment requires global coordination: providers may publish bad information, they may not agree, information may be abandoned or relevant areas not covered. Global supervision can be done by open consortia setting the rules and doing central services for appropriate communities. They constitute the co-reference service in the narrower sense. The consortia should in turn collaborate on common standards. Central services are in particular:
   a. Controlling referential integrity and negotiating solutions with primary information providers. Maintaining inverted indices.
   b. Determination of the transitive closures of equivalence clusters. Detection of contradictory information and identification of possible sources of inconsistency. Duplicate detection algorithms can be modified to validate manual co-reference information.
   c. Guiding and monitoring work of primary information providers to conflict resolution, handling of abandoned sources, suggestions for new areas to cover. The employment of authority files can simplify complex co-reference clusters. The service can feed into authority file maintenance.

5. CONCLUSIONS AND FUTURE WORK

We have argued that a next generation of DL systems should leverage on associations across document contents, metadata, indices and collections. We regard explicit co-reference information as enabling factor of great genericity and propose a new kind of DL service integrating data cleaning methods and reference information management in KOS. It has the potential to open up radically new applications on top of DLs. Reasoning services long dreamed of may become feasible in the envisaged connected knowledge networks. To our opinion, the whole area deserves a major research effort. DL research focus should shift from classification to association. We continue research on foundational issues and algorithms for consistency verification and maintenance of co-reference information: How can global consistency be improved in a distributed system? What are the integrating and disintegrating factors?

6. REFERENCES

[1] Bilenko, M., and Mooney, R.J. Adaptive Duplicate Detection Using Learnable String Similarity Measures. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, (Washington DC, August 2003), 39-48.

[2] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. Graph structure in the web. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, *33, 1-6*, (2000) 309–320, ISSN: 1389-1286.

[3] Cardoso, J., and Sheth, A. Eds. *Semantic Web Services, Processes and Applications*. Springer, 2006, 405 pages, ISBN 0-38730239-5.

[4] Degen, W., Heller, B., Herre, H., and Smith, B. GOL - Towards an Axiomatized Upper-Level Ontology. *Electronics and Computer Science* (2001).

[5] Dodds, L. *An Introduction to FOAF*. 2004. At http://www.xml.com/pub/a/2004/02/04/foaf.html, accessed Nov.16, 2006.

[6] Doerr, M. The CIDOC CRM - An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine, 24(3*) (2003).

[7] Doerr M., Hunter, J., and Lagoze C. Towards a Core Ontology for Information Integration. *Journal of Digital Information, 4, 1 (*2003) Article No. 169.

[8] Lagoze, C., Krafft, D. B., Payette, S., and Jesurogai, S. What Is a Digital Library Anymore, Anyway? *D-Lib Magazine*, *11, 11*, (2005).

[9] Levesque, H.J. Foundations of a functional approach to knowledge representation. *Artificial Intelligence, 23, 2,* (1984), 155–212.

[10] Magkanaraki, A., Alexaki, S., Christophides, V., and Plexousakis, D. 2002. Benchmarking RDF schemata for the Semantic Web. The Semantic Web - ISWC 2002: In *Proceedings of the First International Semantic Web Conference* (Sardinia, Italy, June 9-12, 2002). Springer Berlin / Heidelberg 2342/2002, ISSN:0302-9743.

[11] O'Neill, E.T., Bennett, R., Hengel-Dittrich, C., and. Tillett, B., B. *Viaf (Virtual International Authority File)*: Linking Die Deutsche Bibliothek And Li-Brary Of Congress Name Authority Files. In WLIC2006, 2006.

[12] Patel, M., Koch, T., Doerr, M., Tsinaraki, C., Gioldasis, N., Golub, K., and Tudhope, D. *Semantic Interoperability in Digital Library Systems*, DELOS Network of Excellence on Digital Libraries – deliverable 5.3.1, June 2005.

[13] Salton, G. Associative Document Retrieval Techniques Using Bibliographic Information, In *Journal of the ACM*, 10 (1963), pp 440-457

[14] Sinclair, P., Addis, M., Choi, F., Doerr, M., Lewis, P., and Martinez, K. The use of CRM Core in Multimedia Annotation. In *Proceedings of the 1st First International Workshop on Semantic Web Annotations for Multimedia part of the 15th World Wide Web Conference (SWAMM 2006)* (Edinburgh, Scotland, May 2006.)

[15] Wiederhold, G. Mediators in the Architecture of Future Information Systems. *IEEE Computer*, (March 1992).

Extending the 5S Digital Library (DL) Framework: From a Minimal DL towards a DL Reference Model

Uma Murthy
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061, USA
umurthy@vt.edu

Douglas Gorton
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061, USA
dogorton@vt.edu

Ricardo Torres
Institute of Computing
State University of Campinas
13084-851 Campinas, SP, Brazil
rtorres@ic.unicamp.br

Marcos André Gonçalves
Dept. of Computer Science
Federal University of Minas Gerais
Belo Horizonte, M.G., Brazil
mgoncalv@dcc.ufmg.br

Edward Fox
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061, USA
fox@vt.edu

Lois Delcambre
Dept. of Computer Science
Portland State University
Portland, OR 97207, USA
lmd@cs.pdx.edu

ABSTRACT

In this paper, we describe ongoing research in three DL projects that build upon a common foundation – the 5S DL framework. In each project, we extend the 5S framework to provide specifications for a particular type of DL service and/or system – finally, moving towards a DL reference model. In the first project, we are working on formalizing content-based image retrieval services in a DL. In the second project, we are developing specifications for a superimposed information-supported DL (combining annotation, hypertext, and knowledge management technologies). In the third effort, we have used the 5S framework to generate a practical DL system based on the DSpace software.

1. INTRODUCTION

DLs are immensely complex systems which allow information to be stored in an intelligent, usable, and easily retrievable fashion. In order to address the complexity of DLs, Gonçalves, et. al. proposed the 5S framework [8], where they defined a "core" or a "minimal" DL, i.e., the minimal set of components (a metamodel[1]) that make a DL, without which a system/application cannot be considered a DL. According to the framework, the nature of DLs can be described using the 5S's – Streams, Structures, Spaces, Scenarios, and Societies. Together these abstractions provide a formal foundation to define, relate, and unify concepts – among others, of digital objects, metadata, collections, and services – required to formalize and elucidate DLs. A reference model may be considered to be a structure or conceptual framework, which allows the modules of a system to be described and used in a consistent manner. Early versions of the DL reference model, as defined by the DELOS group, seemed to be aiming towards a comprehensive (maximal) representation of a DL, a DL system, and a DL management system [4]. The aim of this model is to facilitate the integration of research and to propose better ways of developing appropriate DL systems/applications.

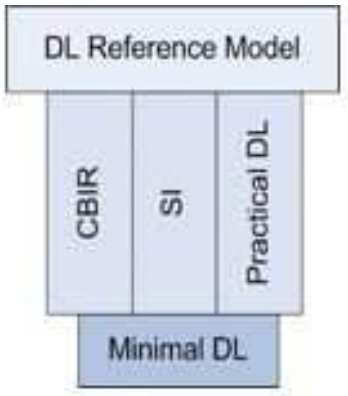


Figure 1: From a minimal DL to a DL reference model.

In this paper, we address three extensions of the 5S framework, going from a minimal DL (as described by the 5S framework) towards a (maximal or comprehensive) DL reference model. Figure 1 depicts this idea, where we consider a minimal DL as the foundation of various extensions, which serve as a base for a DL reference model. In the first extension, we are working on formalizing content-based image retrieval (CBIR) services in a DL (shown as CBIR). Clearly,

[1] *Metamodeling* is the construction of a collection of "concepts" (things, terms, etc.) within a certain domain. A model is an abstraction of phenomena in the real world, and a metamodel is yet another abstraction, highlighting properties of the model itself (from `http://www.wikipedia.org/`).

First International Workshop on "Digital Libraries Foundations" Vancouver, British Columbia, Canada, June 23, 2007

adding images to a DL is important, and since searching is a key service, CBIR services need to be supported. From the earliest days with many DL systems (such as electronic theses and dissertations), annotation was on top of the list of features to add. Also, given the importance of hypertext, having more specificity in hypertext, thus enabling working with information at sub-document granularities, seems to be of value. These ideas relate to our second extension, where we are developing a metamodel for a superimposed information-supported DL (combining specific features of annotation, hypertext, and knowledge management technologies, shown as SI). Finally, our third extension deals with DL generation based on DL software, such as DSpace in this case (shown as Practical DL). This is important because it helps to examine practical DL software functionality and architecture in the context of a formal DL specification, such as the 5S framework.

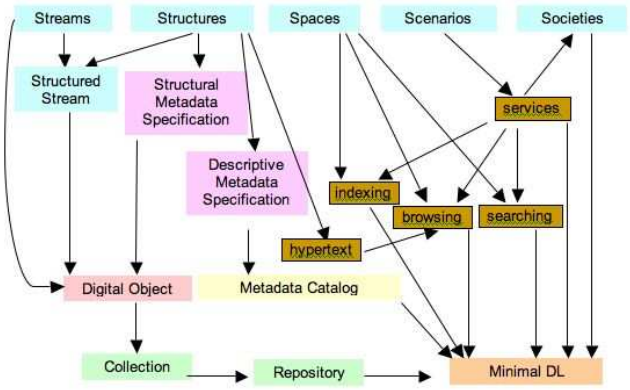


Figure 2: A minimal DL in the 5S framework.

2. 5S FRAMEWORK

Recognizing the difficulties in understanding, defining, describing, and modeling digital libraries (DLs), Gonçalves, et al. have proposed and formalized the 5S (Streams, Structures, Spaces, Scenarios, and Societies) framework of DLs [8]. 5S provides a formal framework to capture the complexities of DLs. The definitions in [8] unambiguously specify many key characteristics and behaviors of DLs. This also enables automatic mapping from 5S constructs to actual implementations as well as the study of qualitative properties of these constructs (e.g., completeness, consistency) [6]. In this section, we summarize the 5S theory from [8]. Here we take a minimalist approach, i.e., we describe briefly, according to our analysis, the minimum set of concepts required for a system to be considered a digital library. ***Streams*** are sequences of arbitrary types (e.g., bits, characters, pixels, frames) and may be static or dynamic (such as audio and video). Streams describe properties of DL content such as encoding and language for textual material or particular forms of multimedia data. A ***structure*** specifies the way in which parts of a whole are arranged or organized. In DLs, structures can represent hypertexts, taxonomies, system connections, user relationships, and containment– to cite a few. A ***space*** is a set of objects together with operations on those objects that obey certain constraints. Spaces define logical and presentational views of several DL components, and can be of type measurable, measure, probability, topological, metric, or vector space. A ***scenario*** is a sequence of events that also can have a number of parameters. Events represent changes in computational states; parameters represent specific variables defining a state and their respective values. Scenarios detail the behavior of DL services. A ***society*** is "a set of entities and the relationships between them" and can include both human users of a system as well as automatic software entities which have a certain role in system operation. These 5Ss, along with fundamental set theoretic definitions, are used to define other DL constructs such as digital objects, metadata specification, collection, repository, and services.

Figure 2 shows concepts in the metamodel for a minimal DL using the 5S framework. For detailed formal definitions of the 5Ss and other DL constructs leading to the definition of a minimal DL, the reader is pointed to [6, 8]. The arrows in the figure indicate that some concepts are used in the definition of other concepts. For example, digital objects are composed of streams and structures. This representation is used in the metamodel figures that follow henceforth in sections 3, 4 and 5. Also, the extension figures in these sections have been drawn with the perspective of showing what needs to be added to the minimal DL. So, all DL concepts defined in the minimal DL (as mentioned in [8]) should be assumed to be in a DL that incorporates the extension.

3. CBIR SERVICES IN A DL

Technological improvements in image acquisition and the decreasing cost of storage devices have supported the dissemination of large image collections, supported by efficient retrieval services. One of the most common approaches involves *Content-Based Image Retrieval (CBIR) systems* [15, 17]. Basically, these systems try to retrieve images similar to a user-defined specification or pattern (e.g., shape sketch, image example). Their goal is to support image retrieval based on *content* properties (e.g., shape, color, or texture), usually encoded into *feature vectors*. One of the main advantages of CBIR is the possibility of an automatic retrieval process, avoiding the work of assigning keywords, which usually requires very laborious and time-consuming prior annotation of images.

Various Digital Libraries (DLs) support services based on image content [3, 5, 10, 14, 18, 19, 20, 21]. However, these systems are often designed and implemented without taking advantage of formal methods and frameworks. In this context, a research initiative is being conducted aiming to extend the 5S DL formal framework [8] for describing services based on image content description. The main contribution of this research is the proposal of several constructs that extend the 5S framework to handle image content descriptions and related services. These constructs can aid understanding of content-based image retrieval concepts as they apply to DLs. They also can guide the design and implementation of new DL services based on image content.

Figure 3 presents the proposed concepts based on the 5S framework to handle image content descriptions and related digital library services. A typical DL service based on image content information requires the construction of *image descriptors*, which are characterized by: (i) an *extraction algorithm* to encode image features into feature vectors; and (ii) a *similarity measure* to compare two images based on the distance between the corresponding feature vectors. The similarity measure is a *matching function*, which gives the

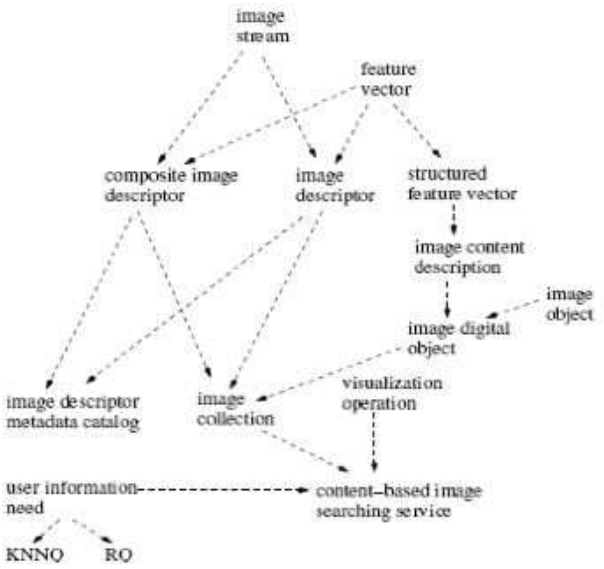


Figure 3: Formalizing CBIR services in a DL using the 5S framework.

degree of similarity for a given pair of images represented by their feature vectors, often defined as an inverse function of the distance (e.g., Euclidean), that is, the larger the distance value, the less similar the images. Structures can be applied to feature vectors for storage purposes (*structured feature vector*) and *image digital object* is defined by extending the original 5S digital object concept by considering image content descriptions. Two typical searching services based on image content can be usually performed: K-nearest neighbor query (KNNQ) and range query (RQ). In a KNNQ, the user specifies the number k of images to be retrieved closest to the query pattern. In a RQ, the user defines a search radius r and wants to retrieve all database images whose distance to the query pattern is less than r.

4. AN SI-SUPPORTED DL

For digital libraries (DLs) to fully support domains such as education there is a need for capabilities that go beyond information seeking-related services. DL users need, but get very little help with:

- Selecting and annotating multimedia information at varying document granularities – parts of a document, to a complete document, to multiple documents
- Linking new content with existing content, at varying document granularities
- Organizing/arranging annotated information.
- Sharing and reusing of new information (annotations, structures, etc) and associated existing information
- Finding and re-finding new information (annotations, structures, etc) and associated existing information through searching/browsing/visualization

An example of such use could be by a Biology professor, who is preparing for a class on the brain. Most of her class

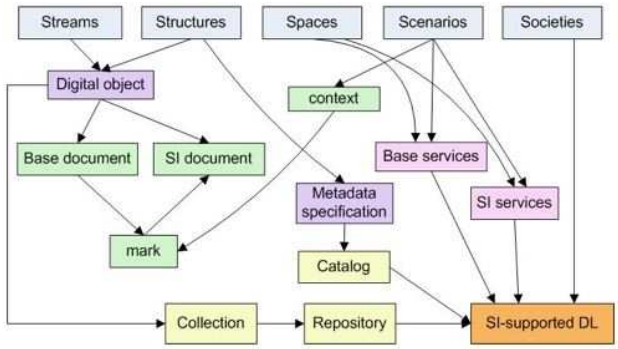


Figure 4: An SI-supported DL using the 5S framework.

material comes from existing (multimedia) resources. For a particular topic, she wants to be able to work with pieces of information in various documents and prepare lecture notes, course materials, presentations, etc. Then, she wants to be able to share all this information with her students and with other faculty, who may have their own representation of the same information.

Existing DLs such as [2, 16] facilitate some of these tasks; however, they provide limited support for working with heterogeneous multimedia formats and/or for working with information at varying document granularities while retaining original information context. We are working towards the development of a Superimposed Information-Supported Digital Library (henceforth referred to as SI-DL), which will bring together *superimposed information* along with traditional DL services that operate in context (of a domain such as education). We believe this will help in building a system with functionality to support annotation, linking, knowledge management and, sharing and reuse of information in tasks such as those mentioned above. Superimposed information (SI) refers to new information laid over existing information [11]. It is supplemental information created to reference, highlight, and extend information present elsewhere. Examples cover a variety of new interpretations, including annotations, tags, citations, indexes, concept maps, multimedia presentations, etc. The focus of SI research is to enable working with sub-document information, such that a user may (a) deal with information at varying document granularity, and (b) select or work with information elements at sub-document level while retaining the original context (by referencing, not replicating, information).

Beginning with development of scenarios and applications (such as [12, 13, 14]), literature review, and brainstorming, we have come up with a preliminary set of specifications for an SI-DL. To ground our work on a firm theoretical foundation, we are extending the 5S framework for DLs to formally define essential units in an SI-DL, resulting in an SI-DL metamodel. These constructs will not only aid in a deeper understanding of SI and related concepts, but also will serve as building blocks for defining various possibilities of an SI-DL.

Figure 4 shows our preliminary work in identifying important SI concepts and their relation to 5S constructs. At the core of an SI system, is a *mark* – an abstraction that specifies an addressable/reference-able region or sub-document, in existing multimedia information of heterogeneous formats.

Marks connect base documents and SI documents. A *base document* is information already existing in the digital library and marks are created in a base document. Marks are used in *SI documents*, which may be constructed by organizing marks in a specific schema/structure. *Context* refers to information and conditions surrounding creation and use of SI including mark creation context, usage context, and context associated with software dealing with base documents and SI documents. Apart from existing *base services* (such as search, browsing and indexing), an SI-DL has *SI services*, which support creation, use and management of marks, context and SI. Finally, creators, viewers, and users of SI form societies that will interact with SI.

5. A PRACTICAL DL

In today's ever-changing world of technology and information, a growing number of organizations and universities seek to store digital documents in an online, easily accessible manner. DLs provide the medium for the online storage and dissemination of such documents and many open source and commercial products are available that help users accomplish that task. While DL software packages enable a broader adoption of DLs, there is still a certain amount of configuration, customization, and data ingestion that must occur in such systems before they are truly optimally usable and set up to serve as many of the institution's needs as allowable. The generation of DLs attempts to abstract some of these processes into a simpler, clearer task where the nature of the desired digital library is described and the generator handles those details with regard to configuration, customization, generation of pertinent code, etc. The intent is to automate these tasks in a way that the DL designer has an appreciation and understanding of the repository to be created but does not need to worry about the underlying technological layer as would be needed if the DL were created manually.

In order to ease the process of creating DLs, we have created a XML-based specification model that describes the nature of possible DLs in MIT and HP Labs' DSpace DL software [1]. We base our work on DL specifications on Fox and Gonçalves' work with the 5S Framework for Digital Libraries [8] and its domain specific digital library declaration language, 5SL [7]. While the original work with DL specification with 5SL was complete, it was more suited to theoretically describe DL systems. In this work we move to a more practical DL metamodel and apply the aspects of 5S and 5SL to describe the nature, structure, and functionality of a modern DL system such as DSpace.

Figure 5 represents the essential components of this metamodel. In order to continue to use a 5S driven organization and separation of the concerns of a digital library, it is necessary to examine the DSpace functionality and architecture in the context of the 5 S's. Thus, we decompose the functionality, structure, and services of DSpace into the aspects that the 5S framework suggests. Because DSpace is a mature, open source software project that has much built-in capabilities as well as customizations via source code and other avenues, our work focuses only on the most commonly used aspects of DSpace. For example, the main DSpace organizational components are Collections and Communities, where Communities are sets of Collections with documents of similar content and subject matter, which we apply to the original hierarchical 5SL constructs of Collections and Col-

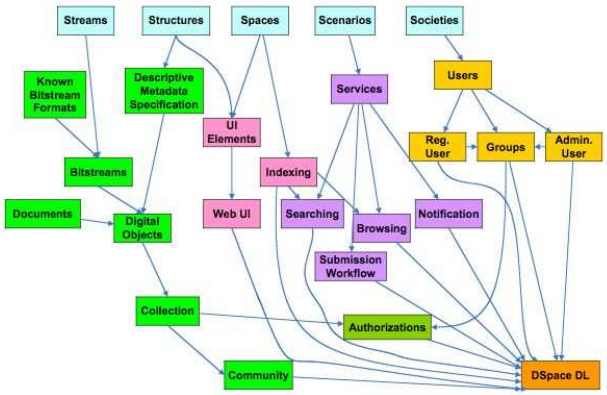


Figure 5: A practical (DSpace) DL using the 5S framework.

lectionSets. Each XML element representing either of these aspects of a DSpace DL also has sub-elements which describe metadata characteristics of each, such as a name, description, and textual components to be used in interfaces. Users that are desired in a DSpace DL are described in a Society sub-model, split into collections of ÔManagers' for administrative users and ÔActors' for regular users. Each type of user requires a few defined metadata elements needed in DSpace such as a password, name, and phone number. Groups of users are also similarly defined.

This work with DSpace generation provides a good proof of concept for applying past work with DL specification and generation to a widely used repository system but there is still much work to be done with DL specification and generation in general. Choices needed to be made to decide which DSpace functionalities were supported for specification and generation, and due to that some functions were unable to be created programmatically by the generator. Much additional work can be done to provide a more comprehensive and all encompassing specification and generation ability for DSpace. Similarly, there are many DL packages out there that have different strengths and are well suited for different applications–the eventual move toward more generalized ways of specifying and generation DL systems would lead to a more streamlined consistent installation and generation path for all these systems. For details on this work, the reader is pointed to [9].

6. TOWARDS A DL REFERENCE MODEL

We have described three different efforts in progress, which build upon a common foundation – the 5S minimal DL framework. Of course there are other extensions also needed. However, one needs to start somewhere and certainly these extensions serve as distinct and valuable starting points. We consider the development of the aforementioned extensions as a step towards understanding, comparing and combining results achieved in different areas of DL development – thus, serving as a base for the development of a DL reference model. Just as there needs to be eventual movement towards a broadly applicable model for DL specification, a more framework oriented approach for the generation of DLs based on specifications is also a direction that would allow for easier, more consistent DL generation. Figure 6 shows such a generation process. We begin by defining a meta-

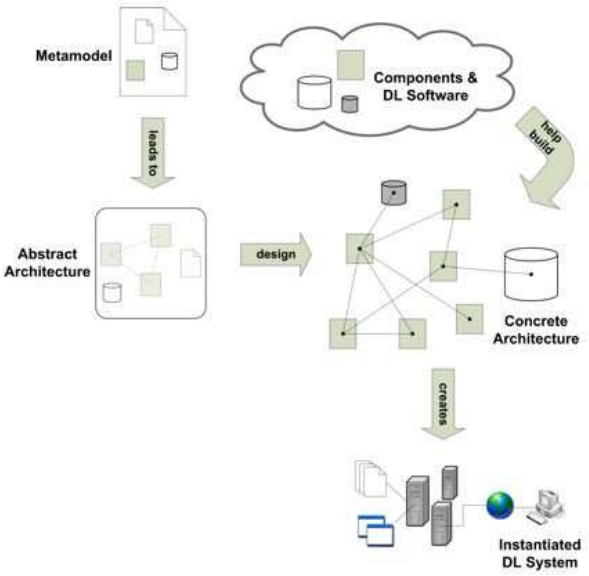


Figure 6: A DL generation process.

model of constructs, or building blocks for the specific DL we want to generate. Specific instances of this metamodel may be derived that represent a user's desired DL system and make up an abstract DL architecture. Based on the declared DL and available software components (and systems, such as DSpace), a concrete architecture may be created for that DL, which finally may be built into a DL system.

Acknowledgments

Thanks go to sponsors of our work, which include: AFOSR (grant F49620-02-1-0090), AOL, IMLS, NSF (grants DUE-0121679, DUE-0435059, IIS-0325579, IIS-0535057) and CNPq project 5S-VQ (grant MCT/CNPq/CT-INFO 551013/2005-2), FAPESP, FAEPEX, Microsoft Escience project, and the Microsoft tablet PC grant. Thanks also go to those in Virginia Tech's Digital Library Research Laboratory, and to all those involved in our various projects.

7. REFERENCES

[1] DSpace Federation, `http://dspace.org/`.

[2] Agosti, M., Albrechtsen, H., Ferro, N., Frommholz, I., Hansen, P., Orio, N., Panizzi, E., Pejtersen, A.M. and Thiel, U. DiLAS: a Digital Library Annotation Service. Presented at the International Workshop on Annotation for Collaboration, Paris, Framce, 2005.

[3] Bergman, L.D., Castelli, V. and Li, C.-S. Progressive Content-Based Retrieval from Satellite Image Archives. D-Lib Magazine, 3 (10).

[4] Candela, L., Castelli, D., Ioannidis, Y., Koutrika, G., Pagano, P., Ross, S., Schek, H.-J. and Schuldt, H. A Reference Model for Digital Library Management Systems. `http://www.delos.info/index.php?option=com_content&task=view&id=345&Itemid=#docs`.

[5] French, J.C., Chapin, A.C. and Martin, W.N. An application of multiple viewpoints to content-based image retrieval. In proceedings of the 3rd ACM/IEEE-CS joint conference on digital libraries, (Houston, Texas, 2003), IEEE Computer Society, 128-130.

[6] Gonçalves, M. Streams, Structures, Spaces,Scenarios, and Societies (5S): A Formal Digital Library Framework and Its Applications. PhD Dissertation, Computer Science, Virginia Tech, Blacksburg, 2004. `http://scholar.lib.vt.edu/theses/available/etd-12052004-135923/`.

[7] Gonçalves, M. and Fox, E.A. 5SL: a language for declarative specification and generation of digital libraries. In proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, (Portland, Oregon, USA, 2002), ACM Press, 263-272.

[8] Gonçalves, M.A., Fox, E.A., Watson, L.T. and Kipp, N.A. Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. ACM TOIS, 22 (2). 270-312.

[9] Gorton, D. Practical Digital Library Generation into DSpace with the 5S Framework. Master's thesis, Computer Science, Virginia Tech, Blacksburg, 2007. `http://scholar.lib.vt.edu/theses/available/etd-04252007-161736/`.

[10] Hong, J.S., Chen, H.Y. and Hsiang, J. A Digital Museum of Taiwanese Butterflies. In proceedings of the Fifth ACM Conference on digital Libraries, (San Antonio, Texas, United States, 2000), 260-261.

[11] Maier, D. and Delcambre, L. Superimposed Information for the Internet. in WebDB Workshop, (1999), 1-9.

[12] Murthy, U., Ahuja, K., Murthy, S. and Fox, E.A. SIMPEL: a superimposed multimedia presentation editor and player. In proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries. 377.

[13] Murthy, U., Richardson, R., Fox, E.A. and Delcambre, L. Enahcing Concept Mapping Tools Below and Above to Facilitate the Use of Superimposed Information. In proceedings of the Second International Conference on Concept Mapping, (San Jose, Costa Rica, 2006).

[14] Murthy, U., Torres, R.d.S. and Fox, E.A. SIERRA: A Superimposed Application for Enhanced Image Description and Retrieval. Lecture Notes in Computer Science: Research and Advanced Technology for Digital Libraries, 2006, 540-543, `http://dx.doi.org/10.1007/11863878_63`

[15] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A. and Jain, R. Content-Based Image Retrieval at the End of the Years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (12). 1349-1380.

[16] Sumner, T., Ahmad, F., Bhushan, S., Gu, Q., Molina, F., Willard, S., Wright, M., Davis, L. and Janèe, G. Linking learning goals and educational resources through interactive concept map visualizations. International Journal on Digital Libraries, 5 (1). 18-24.

[17] Torres, R.d.S. and Falcão, A.X. Content-Based Image Retrieval: Theory and Applications. Revista de Informática Teórica e Aplicada, 13 (2). 161-185.

[18] Vemuri, N.S., Torres, R.d.S., Gonçalves, M.A.,

Fan, W. and Fox, E.A. A Content-Based Image Retrieval Service for Archaeology Collections. In proceedings of the European Conference on Digital Libraries, (Alicante, Espanha, 2006), 438-440.

[19] Wang, J.Z. and Du, Y. Scalable integrated region-based image retrieval using IRM and statistical clustering. In proceedings of the 1st ACM/IEEE-CS joint conference on digital libraries, (Roanoke, Virginia, USA, 2001), 268-277.

[20] Wang, Y., Makedon, F., Ford, J., Shen, L. and Goldin, D. Generating fuzzy semantic metadata describing spatial relations from images using the R-histogram. In proceedings of the 4th ACM/IEEE-CS joint conference on digital libraries, (Tuscon, AZ, USA, 2004), 202-211.

[21] Zhu, B., Ramsey, M. and Chen, H. Creating a Large-Scale Content-Based Airphoto Image Digital Library. IEEE Transactions on Image Processing, 9 (1). 163-167.

The Age of the Digital Library

José Borbinha
Instituto Superior Técnico / INESC-ID
Rua Alves Redol, 9, Apartado 13069, 1000-029 Lisboa, Portugal
jlb@ist.utl.pt

ABSTRACT

Digital Libraries can be characterized mainly as a converging point where disparate communities have been meeting to address common issues related with the creation, management and usage of digital information. The range of issues that have been scrutinized is impressive, but the approaches have been mainly chaotic and unstructured. All of this makes it very difficult, if not unrealistic, trying to related the area with a specific body of knowledge, as if it were the case of a normal discipline.

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.

Categories and Subject Descriptors

H. Information Systems, H.3.7 Digital Libraries, K.6 Management of Computing and Information Systems.

General Terms

Standards, Systems Issues, System Design

Keywords

Enterprise Architectures, Enterprise Architectures Frameworks, Digital Libraries

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, comprising also the area of the Digital Library.

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 Digital Library 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 actually the Digital Library as a problem and the main emerging related challenges.

The ultimate purpose of this exercise is to raise arguments to prove that, from now, we should not continue promoting the Digital Library by mainly raising generic goals and addressing the technological related issues. Alternatively, the Digital Library 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.

But why is this really important? First, let us follow a simple analysis:

> One can conceive "Digital Library deployments" in mainly two scenarios: as a purpose in itself (the Digital Library as the main business goal); or as a contribution to other purposes (technology and processes created from a "Digital Library perspective" in order to be used to support more generic goals). The first scenario will maintain the Digital Library has a relevant concept, where it might be not too difficult to acknowledge the right credits to the right communities contributing for that. It might be also 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 Digital Library community has to make necessary efforts to it that happen.

Now, let us try to conclude why it is important to align the Digital Library with the concepts of Enterprise Architecture and Enterprise Architecture Framework:

> 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

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

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

administration services, public services and enterprises in general3,4. Assuming that Digital Library's technology has reached a maturity for deployments at these levels (the real life…), than those requirements can not be ignored, especially by those in the management, legal and business front edges.

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

Arms' presents eight general principles representing concepts and requirements for the Digital Library architecture:

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."
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"

On the other side Lesk provides an historical description and a vision of the future of the area of Information Retrieval that makes it clearly coincident with the Digital Library.

According to Lesk, **Childhood** (1945-1955) is described as the time when Vannevar Bush had is vision of 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

In a first glance one might be tended to consider the Digital Library 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 Digital Library. 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 Digital Library, and his

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

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 Digital Library!

In this sense, the Digital Library should be now in its fulfillment age. And that is the fact! The "Million Books Project", just to cite one example, is pursuiting the 1M books milestone[5]; reference works are common to find as e-books; and on-line directories are fairly well guiding us in the labyrinth of the World Wide Web (Yahoo, Google, del.icio.us, etc.).

Therefore, the Digital Library should be going to the age of retirement. And in fact it looks like that! This is an empiric statement[6], but I think that I am not going against the actual generic perception of the community if I say that very few specific Digital Libraries challenges can be identified nowadays (if not none at all…).

For example, interoperability was a very specific issue in the Digital Library. Z39.50[7], once a specific answer to specific requirements for technical interoperability from specific Digital Library business goals, has become irrelevant after the emerging of the web based OPAC, which in itself has a tendency to disappear, integrated in the "enterprise portal" and of web-services solutions such as SRU[8] and OAI-PMH[9]. Concerning semantic interoperability, one other common issue in Digital Libraries, is also 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 Digital Library, started in fact 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 specific of the Digital Library. They are simply generic issues in ay class 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 Digital Library very specific issues such as the digital preservation have been emerging as a regular concern in any organization. "Archives" are becoming "repositories"; historical information does not make sense anymore, as all the information available is now critical for any good business governance.

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

> "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 of a concrete vision of the Digital Library, 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, from now the Digital Library community will be not requested anymore to provide technology, but expertise and services. In fact, reviewing now Arms' key concepts, we can claim that any of them are really specific of the Digital Library, but instead generic goals, constraints, requirements or good practices that we can find in multiple other cases. I think that such will result more evident if we reorganize Arms' arguments this way:

– 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.

---

[5] http://www.archive.org/details/millionbooks

[6] The author has analytical work in progress that tending to demonstrate this statement…

[7] http://www.loc.gov/z3950/agency/

[8] http://www.loc.gov/standards/sru/

[9] http://www.openarchives.org/OAI/openarchivesprotocol.html

View	What	How	Where	Who	When	Why
Scope	Things important to the business	Processes the business performs	Locations in which 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

Table 1: The Zachman Framework

The purpose of having detailed views, planning and analytical knowledge of a system can be tracked to a long time ago. But in several scenarios its addressing is now not only a possible purpose, but also key vital tools to address new unavoidable requirements

A new world for Information Systems in general arrived recently with the technology associated with the Web, 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 Digital Library is to be able to offer solutions that, for specific situations, the problems are properly addressed, than we must recognize that such solutions are always a combination 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 with information, processes and technology. That is the scope of the area of Information Systems[10]. 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 comprehensive and famous of them, is the Zachman framework[11], 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. In this sense, the meanings of the lines in this table are:

- Scope (Contextual view; Planner): The business purpose and strategy. It defines the context for the other views.
- Business Model (Conceptual view; Owner): A description of the organization, revealing which parts can be automated.
- System Model (Logical view; Designer): The 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): How the system will be implemented, with the specific technology and ways to address production constraints.
- Components (Detailed view; Implementer): Detail of each of the system elements that need clarification before production.
- Instances (Operational view; Worker): A view of the functioning system in its operational environment.

On the same time, the meanings of the columns are:

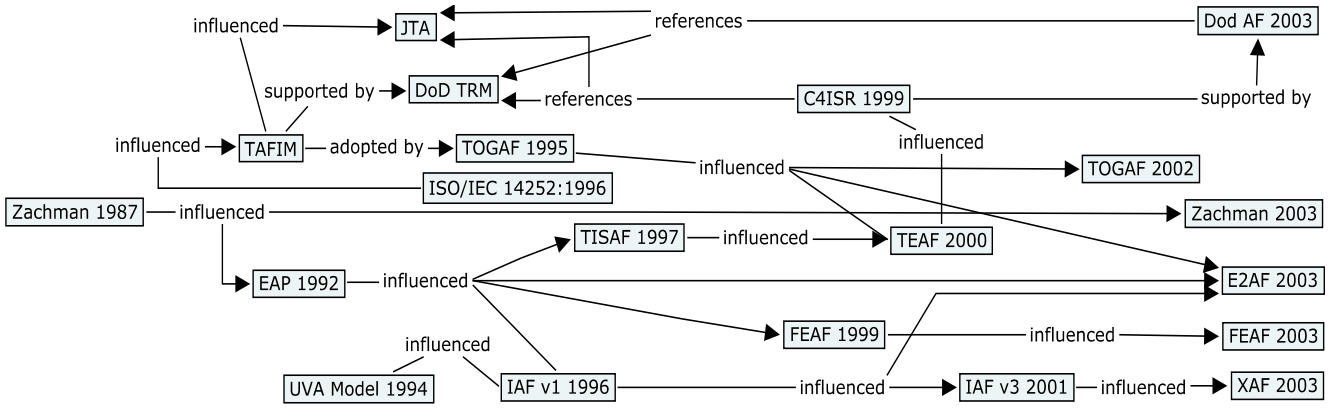
- What (Data): The system contents, or data.
- How (Function): The usage and functioning of the system, including processes and flows of control.
- Where (Network): Spatial elements and their relationships.
- Who (People): The actors interacting with the system.
- When (Time): The timings of the processes.
- Why (Motivation): Motivation for the system and rules for constraining it (applied mainly to the Why and How views).

From the Zachman Framework many other Enterprise Architecture frameworks for specific areas have been developed. Those have been developed by research entities (such as E2A[12]),

[10] 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/

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

[12] From the Institute for Enterprise Architecture developments - http://www.enterprise-architecture.info/



by governmental bodies[13] (such as FEAF, TEAF, TOGAF, etc.), and by private companies (such as IAF[14], from Cap Gemini). This 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 for that expresses very well the complexity of the Enterprise Architecture process:

- Strategic Management: Balanced Scorecard[16]
- Strategy Execution: EFQM[17]
- Quality Management: ISO 9001[18]
- IT Governance: COBIT[19]
- IT Service Delivery and Support: ITIL[20]
- IT Implementation: CMM[21] and CMMI[22]

[13] http://www.eagov.com; http://www.eaframeworks.com/frameworks.htm;

http://www.whitehouse.gov/omb/egov/a-1-fea.html

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

[15] Redraw from [6] (more details can be found in this work).

[16] The balanced scorecard management system - http://www.balancedscorecard.org/

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

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

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

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

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

6. The Goal of the Digital Library

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

- 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 impossible, for a specific system, to provide discovery and access to everything produced by the focused organization (at least for now…).

All of this means that we must make a special effort to rethink the structure for the thinking of the Digital Library.

7. Conclusions

Concluding, the Digital Library community must prepare itself for a dignified retirement age by moving its established knowledge

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

from research to engineering, in order to take part in more generic goals[23].

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[24], I think that the Digital Library 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 Digital Library?

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

But maybe not, or at least as some of us have been trying to do it, especially if we give credit to someone else that wrote once[25]:

> "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.
>
> 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 Digital Library should move from the perspective of the engineer to the perspective of the architect[26].

8. 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) pp. 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. 9 October 2000

[4] Lesk, M. The Seven Ages of Information Retrieval. Proceedings of the Conference for the 50th anniversary of As We May Think, 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, 2004. ISBN 1-4120-1607-X

[7] Spewak, S.: Enterprise Architecture Planning – Developing a Blueprint for Data, Applications and Technology. John Wiley & Sons Inc (29 October 1993). ISBN 978-0471599852

[8] Thomas, E.: Service-Oriented Architecture: Concepts, Technology, and Design. Prentice Hall PTR (1 September 2005). ISBN 978-0131858589

[9] Zachman, J. A Framework for Information Systems Architecture. IBM Systems Journal, vol. 26, no. 3, 1987. IBM Publication G321-5298.

[23] 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.

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

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

[26] http://answers.google.com/answers/main?cmd=threadview&id=233551

Towards a Reference Quality Model for Digital Libraries

Maristella Agosti
Dept. of Information Engineering
University of Padua
Via Gradenigo, 6/b – 35131
Padova, Italy
agosti@dei.unipd.it

Nicola Ferro
Dept. of Information Engineering
University of Padua
Via Gradenigo, 6/b – 35131
Padova, Italy
ferro@dei.unipd.it

Edward A. Fox
Dept. of Computer Science
Virginia Tech
Blacksburg, VA 24061
fox@vt.edu

Marcos André Gonçalves
Dept. of Computer Science
Federal University of Minas Gerais
Belo Horizonte, M.G., Brazil
mgoncalv@dcc.ufmg.br

Barbara Lagoeiro
Dept. of Computer Science
Federal University of Minas Gerais
Belo Horizonte, M.G., Brazil
barbara@dcc.ufmg.br

ABSTRACT

This paper discusses the importance of defining a Reference Quality Model for Digital Libraries. Current approaches for Digital Library (DL) quality evaluation are presented. Our view of the steps necessary to achieve this goal is given and discussed.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

General Terms

Design, Measurement, Theory

Keywords

digital libraries, quality, evaluation, reference model

1. INTRODUCTION

In this paper, we discuss issues related to defining a quality model for digital libraries in the light of the recent efforts for building a Reference Model for Digital Library Management Systems [1] [2].

The idea of a Reference Model is to lay the foundations for the digital library field as a whole. The lack of agreement on these foundations has led to a number of uncoordinated efforts that are hard to combine and reuse to produce enhanced outcomes.

[1] `http://www.delos.info/ReferenceModel/`

First International Workshop on "Digital Libraries Foundations" Vancouver, British Columbia, Canada, June 23, 2007

One very important aspect of building such a Reference Model is to capture the notion of quality in DL, i.e., how can one define what is a "good" or "successful" digital library? [9, 18]. Evaluation models for digital libraries have been proposed [9, 18, 6, 7], but some of them were not built having a common DL model as their foundation.

Defining quality and quality measures for digital libraries will allow:

- to detect problems in the system and obtain information to fix them;
- to follow the evolution of systems and their several components (e.g., collections, catalogs, services);
- to evaluate contents to be inserted in the system and check if their quality is compatible with contents already in the system; and
- obeying certain constraints, to compare two or more systems, with regard to some of their components.

This is a very complex task. Quality measures need to be formally defined so they can be captured and quantified. Even if successful in proposing a theoretical quality model, we need to take the appropriate steps to support its operativeness. The model needs to be evaluated and tested in many settings. Standards to capture the necessary information (e.g., log formats) to allow the computation of quality measures need to be established. Once all of this is done, tools can be developed to help with such evaluations. Standard collections can be built to test services, among many other possible tasks.

This paper is organized as follows. Section 2 presents current approaches to quality in digital libraries and DL quality evaluation. Section 3 presents what we think are the necessary steps to produce and give support to such a Reference Quality Model. Section 4 concludes the paper.

2. APPROACHES TO QUALITY IN DIGITAL LIBRARIES

2.1 Broad Studies and Conceptual Frameworks

In [17] and [15] DL evaluation challenges and requirements are enumerated and an evaluation conceptual framework is suggested. The evaluation requirements should answer questions such as: "Why to evaluate?", "What to evaluate?", and "How to evaluate?". The framework considers that the evaluation should deal with performance aspects of parts of the DL system. Thus, the performance can be evaluated according to the effectiveness (how well the system performs its tasks), efficiency (what are the costs for the system to perform the tasks), or a combination of these two factors.

In [16] Saracevic provides an overview of the work on DL evaluation. He analyzed about 80 evaluation studies along the lines of:

1. constructs that were evaluated: the evaluated construct can be a specific digital library or a DL related process;

2. context in which the evaluations were conducted: the evaluation can deal with human, system, usability, anthropological, ethnographic, sociological, or economic aspects;

3. criteria that were chosen as a basis for evaluation, i.e., the judgment standard that was defined for the evaluation. The criteria depend on the context, for instance, for an usability evaluation, the criteria "effort to understand" and "error rate" can be used as a basis for the evaluation;

4. methods that were used during the evaluation; some of those methods are: surveys, structured interviews, observations, case studies, focus groups, transaction log analyses, experimentation, and usage analysis.

In [6] is defined a DL conceptual model to develop test suits that would satisfy the needs of researchers in the DL evaluation area. This model is based on four main dimensions: data & collection, system & technology, users, and usage. The idea consists of using the relationships between these dimensions to create a set of evaluation criteria that, when answered, would generate DL detailed descriptions. These descriptions can be applied to define test-beds or to compare digital libraries.

Nicholson [14] presents a conceptual framework to guide holistic evaluations of library services, considering different points of view: from the user, library staff, and decision makers. Using a matrix of topics and perspectives for measurement, the evaluator can choose what to evaluate and how to evaluate it. This matrix presents the following views:

- Internal View of the System (what are the components of the system): compares components of the system against some type of standard. To evaluate it, staff interviews and surveys, and audits of collections, system, or staff can be used.

- External View of the System (how effective is the system): the user presents a query to the library and evaluates the usability of the system and the returned results. To evaluate it, interviews and focus groups can be used.

- External View of Use (how useful is the system): the user presents the overall usefulness of information obtained through the system. Surveys, interviews, focus groups, and user citation tracking can be used to evaluate it.

- Internal View of Use (how is the system manipulated): interactions between users and a system are analyzed to understand how a system is manipulated. This can be evaluated through the analysis of logs and user behavior.

Tsakonas et al. [19] developed a framework to evaluate the interaction between the user and the DL. An interaction is composed of three components: the user, the content and the system. The work considers three categories of evaluation criteria which define relationships among components: usability (the quality of the direct interaction between the user and the system), usefulness (whether the user needs are being fulfilled by the content), and performance (considering the system response). These categories can be applied to highlight requirements, parameters and metrics for the interaction evaluation.

In [7], the goal is to provide a set of flexible and adaptable guidelines for DL evaluation, outlining the main directions, methods, and techniques for assessing the components of a DL. Besides that, a study about existent DL evaluation approaches is performed, describing the main models to be applied during an evaluation. After this discussion, a framework based on [16] is described, trying to cover most of the aspects that can be found through the several levels of an evaluation process.

In [18], Shen proposes a model of DL success from the end user perspective, based on the integration of various research studies of different areas (e.g., digital libraries and information systems). This model helps to define when and how to measure the different quality aspects. In addition, numeric indicators for the quality of union catalogs and union services are specified.

Aimed at an evaluation from the user point of view, in [12] DigiQual is developed, a protocol based on a similar project for traditional libraries, which helps the DL administrators to understand the quality notion of the users of their system. The protocol defines that the users can answer about 12 quality themes throughout the time, systematically, to identify the best practices for a DL system.

Proposals that develop and present standards for log formats aimed at registering data for evaluation, such as [10] and [11], contribute towards a Reference Quality Model since they provide ways for storing information for assessment. In [11], a multi-level logging schema is proposed that accounts for a large amount of data about users, systems and user-system interactions. Some of this information is difficult to capture (e.g., information about the user behavior may require observing or interviewing the user, which may be very time consuming). Because of that, Klas et al. [11] focus their work on the concept level, which comprises general DL events such as *search*, *browse*, and *navigate*. They store information about services, like the timestamps for the start and end of an event, and the errors that may have occurred. Their proposal is built on top of the work of Gonçalves et al. [10], which describes an XML-based log format that captures detailed information about users and system behavior.

Gonçalves et al. [9] define an explicit formal/quantitative

quality model for digital libraries based on the 5S formal framework for digital libraries [8]. The model is validated through its application to several DLs in different scenarios. A tool implementing a portion of the model has been developed [13].

2.2 The DELOS Approach

2.2.1 The DELOS Reference Model

Digital Library is a complex concept which can be expressed using different perspectives and viewpoints. The DELOS[2] approach for the representation of this many-sided concept has been to start an effort for developing a Reference Model [2] where a framework of three tiers to represent three levels of abstraction is used to represent: the DL, the DLS and the DLMS. The DL is the level where the digital contents are kept, and the DLS is the level of all the organizational and software application components that are able to manage the contents, providing useful services to the interested users with the support of a DLMS.

The DELOS Reference Model [2] aims at providing a representation which characterizes existing and future DLMS from at least the four perspectives: DL end-users, designers, system administrators, and application developers. It introduces the main concepts, the relationships between these concepts, and the constraints that hold among them. It also prescribes aspects that are mandatory for this type of information system. Figure 1 (extracted from [4]) represents the highest level concepts of the DELOS reference model:

content is the entry point for all the concepts related to the content that is managed and disseminated by the DL, e.g., collections, information space model, metadata, ontologies;

user is the root for concepts like roles, communities, and profiles, that represent aspects of the DL users;

functionality is the entrance to that part of the model which concerns DL functions;

architecture regards software components, hosting nodes and how these are linked and constrained;

quality groups qualitative parameters characterizing the digital library behavior within a given operational domain;

policy covers all the concepts that are related to established procedures or plans of actions governing the DL, such as collection management, preservation, and access rights.

From a final user point of view, a DLS is the collection of tools he can use to access and browse the collection of digital information objects – the Digital Library – that is of interest, where the management and the keeping over time of the objects is done by a DLMS and the maintenance of the collection of objects is secured by an organization in charge of it. As outlined in the DELOS Digital Library Manifesto [4], at least three types of conceptually different "systems" can characterize the digital library universe: the DL, the DLS and the DLMS, which are hierarchically related; so are their models, i.e., the DL model is included in the DLS one, and the latter is included in the DLMS model.

[2] `http://www.delos.info/`

2.2.2 Notion of Quality in the DELOS Reference Model

The notion of "quality", which is one of the highest level concepts of the DELOS Reference Model [2], as seen in Section 2.2.1, can be considered at each of the three levels of abstraction: DL, DLS, and DLMS. This means that we can define quality parameters for the information objects, for the services given to the users, and for the system that supports the management of the services. Once the quality parameters are defined, the control of them can be pursued making use of specific control tools and mechanisms. Quality encompasses the characteristics of a DL and the resources that it contains that can benefit from being measured and monitored. The quality is expressed by a set of quality parameters; each parameter can be measured; those measurements are mostly related to the contents and the functionalities.

As far as the content is concerned, the quality of each information object needs to be verified over acquisition and lifetime, because of that it becomes necessary to define a set of quality parameters that can be expressed via a value assigned as result of measurement, where the act of measuring includes a quality parameter in accordance with a selected process and a unit of measurement. The value of a quality parameter is obtained via the selected process, that does not depend on individual perception.

One parameter related to the content assesses the information object quality of being complete. This parameter encompasses the extent to which an information object is of sufficient breadth, depth, and scope for the task at hand, as pointed out in [3].

Authenticity is a content quality parameter which measures whether an information object retains the property of being what it purports to be; this definition takes into account the results and experience of the InterPARES project [5]. The provenance content quality parameter concerns the origin or earliest known history of an information object. This parameter is particularly important when dealing with scientific data. The provenance of data must be tracked since a scientist needs to know where the data came from and what cleaning, rescaling, or modelling was done to arrive at the data to be interpreted [1].

The DLS is the system in charge of implementing the DL. It is composed of components and hosting nodes. As a consequence, the DLS inherits quality related concepts from the DL while it needs new relationships that make it possible to assign such parameters to the entities it deals with, e.g., a hosting node. In particular, the quality assigned to a component supports the DL system administrator during the component selection and configuration phases.

The DLMS is a software system with diverse components. Like other well-constructed software systems, the DLMS has been conceived and developed applying principles and methods of software engineering. Taking into account that the fundamental principles of software engineering are applicable throughout the software life cycle, the DL designer, the DL system administrator, and the DL application developer need to make reference to those general principles. In particular they need to refer to software engineering best practices regarding software quality measurement.

From a final user point of view, taking note that the final user mostly uses an access function to search and browse the DL with the final aim of having delivered a copy of information objects of interest and of certified quality, the final user is interested in quality and quality control over

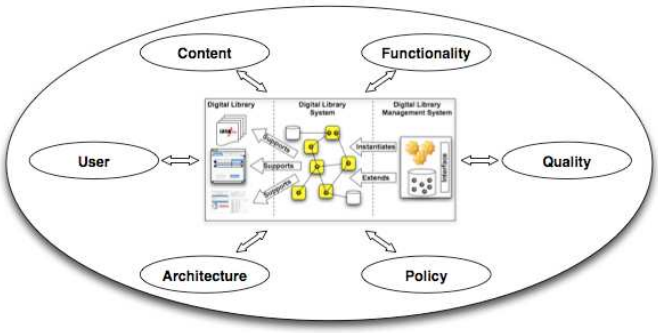


Figure 1: Highest Level Concepts of the DELOS Reference Model.

the contents of the DLS, and the notion of quality of the search for contents, where the searching is implemented by a searching service that has to be evaluated before, to be given for use to the final user.

3. STEPS TOWARDS A REFERENCE QUALITY MODEL FOR DLS

We envision the following steps in order to achieve the goal of building a Reference Quality Model for DL:

1. Contribute to the definition of a Reference Model for DL

   A set of concepts (at least a minimal set), defining what aspects, that have to be taken into account in a digital library, are going to be defined.

2. Formalization of the Model

   To support precision and accuracy in the definition of the concepts in the Reference Model, there is the need to formalize pertinent aspects.

3. Definition and Formalization of Quality Indicators

   Quality dimensions for several of the concepts defined in the Reference Model need to be defined. Numeric indicators for each quality dimension will then be proposed based on the formalization of the concepts in the Reference Model provided in Step 2.

4. Defining the context for each quality dimension in light of the Information Life Cycle

   Each quality dimension needs to be associated with one phase of the Information Life Cycle (i.e., Creation, Distribution, Seeking, and Utilization). This will set the context for specifying when we can apply and compute the respective numeric indicators for each quality dimension and how to use the results of the quality analysis.

5. Discussion with the community and reformulation

   The model needs to be discussed with the community and to be validated by it. Several reformulations to accomodate several different perspectives may be necessary.

6. Providing Support for the Model

   Once we have a solid version of the Reference Quality Model, tools implementing the numeric indicators for each dimension and supporting the envisioned evaluation process need to be build. We will also need standards such as a standard log format to help to capture the necessary information for evaluation.

4. CONCLUSIONS

A co-operative work to make some steps towards the definition of a complete Reference Quality Model for DL has been initiated having in mind the objective of defining, and developing, a model where all previous relevant experiences come together in a synergistic way.

Acknowledgments

This work was partially supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618). Support also was provided through the US National Science Foundation through grants IIS-0535057, DUE-0532825, DUE-0435059, and IIS-0325579.

5. REFERENCES

[1] S. Abiteboul, R. Agrawal, P. Bernstein, M. Carey, S. Ceri, B. Croft, D. DeWitt, M. Franklin, H. Garcia-Molina, D. Gawlick, J. Gray, L. Haas, A. Halevy, J. Hellerstein, Y. Ioannidis, M. Kersten, M. Pazzani, M. Lesk, D. Maier, J. Naughton, H.-J. Schek, T. Sellis, A. Silberschatz, M. Stonebraker, R. Snodgrass, J. D. Ullman, G. Weikum, J. Widom, and S. Zdonik. The Lowell Database Research Self-Assessment. *Communications of the ACM (CACM)*, 48(5):111–118, 2005.

[2] M. Agosti, L. Candela, D. Castelli, N. Ferro, Y. Ioannidis, G. Koutrika, C. Meghini, P. Pagano, S. Ross, H.-J. Schek, and H. Schuldt. A Reference Model for DLMSs – Interim Report. In L. Candela and D. Castelli, editors, *Deliverable D1.4.2 - Reference Model for Digital Library Management Systems [Draft 1]*. DELOS, A Network of Excellence on Digital Libraries – IST-2002-2.3.1.12, Technology-enhanced Learning and Access to Cultural Heritage – `http://146.48.87.122:8003/OLP/Repository/1.0/Disseminate/delos/2006_WP1_D142/content/pdf?version=1` [last visited 2007, March 23], September 2006.

[3] C. Batini and M. Scannapieco. *Data Quality*. Springer-Verlag, Berlin, Germany, 2006.

[4] L. Candela, D. Castelli, Y. Ioannidis, G. Koutrika, P. Pagano, S. Ross, H.-J. Schek, H. Schuldt,and C. Thanos. Setting the Foundations of Digital Libraries: The DELOS Manifesto. *D-Lib Magazine*, Vol. 13 No. 3/4, March/April 2007.

[5] L. Duranti. The long-term preservation of accurate and authentic digital data: the INTERPARES project. *Data Science Journal*, 4:106–118, 2005.

[6] N. Fuhr, P. Hansen, M. Mabe, A. Micsik, and I. Solvberg. Digital libraries: A generic classification and evaluation scheme. In *Proc. of the European Conf. on Digital Libraries*, pages 187–199, Heidelberg, 2001. Springer.

[7] N. Fuhr, G. Tsakonas, T. Aalberg, M. Agosti, P. Hansen, S. Kapidakis, C.-P. Klas, L. Kovas, M. Landoni, A. Micsik, C. Papatheodorou, C. Peters, and I. Solvberg. Evaluation of digital libraries. *Int. Jour. of Digital Libraries*, 2007.

[8] M. A. Gonçalves. *Streams, Structures, Spaces, Scenarios, and Societies: A Formal Framework for Digital Libraries and Its Applications: Defining a Quality Model for Digital Libraries*. PhD thesis, Virginia Tech CS Department, Blacksburg, Virginia, 2004. URL - http://scholar.lib.vt.edu/theses/available/etd-12052004-135923/.

[9] M. A. Gonçalves, B. L. Moreira, E. A. Fox, and L. T. Watson. What is a good digital library? - defining a quality model for digital libraries. To appear in Information Processing and Management, 2007.

[10] M. A. Gonçalves, 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. In *Proc. of the 3rd ACM/IEEE-CS Joint Conf. on Digital Libraries*, pages 312–314, Washington, DC, USA, 2003. IEEE Computer Society.

[11] C.-P. Klas, N. Fuhr, S. Kriewel, H. Albrechtsen, G. Tsakonas, S. Kapidakis, C. Papatheodorou, P. Hansen, L. Kovacs, A. Micsik, and E. Jacob. An experimental framework for comparative digital library evaluation: the logging scheme. In *Proc. of the 6th ACM/IEEE-CS Joint Conf. on Digital Libraries*, pages 308–309, New York, NY, USA, 2006. ACM Press.

[12] M. Kyrillidou and S. Giersch. Developing the DigiQual protocol for digital library evaluation. In *JCDL '05: Proc. of the 5th ACM/IEEE-CS Joint Conf. on Digital Libraries*, pages 172–173, New York, NY, USA, 2005. ACM Press.

[13] B. Lagoeiro, M. A. Gonçalves, and E. A. Fox. 5squal: A quality tool for digital libraries. In *Proc. of the 7th ACM/IEEE Joint Conf. on Digital Libraries*, page (demonstration accepted), New York, NY, USA, 2007. ACM Press.

[14] S. Nicholson. A conceptual framework for the holistic measurement and cumulative evaluation of library services. *Jour. of Documentation*, 60(2):164–182, 2004.

[15] T. Saracevic. Digital library evaluation: Toward evolution of concepts. *Library Trends - Special issue on Evaluation of Digital Libraries*, 49(3):350–369, 2000.

[16] T. Saracevic. Evaluation of digital libraries: an overview. Presentation at the DELOS WP7 workshop on the evaluation of digital libraries, Department of Information Engineering, University of Padua, Italy, October 2004. URL - http://www.scils.rutgers.edu/∼tefko/DL_evaluation_Delos.pdf.

[17] T. Saracevic and L. Covi. Challenges for digital library evaluation. In *Proc. of the 63rd Annual Meeting of the American Society for Information Science*, volume 37, pages 341–350, 2000.

[18] R. Shen. *Applying the 5S Framework to Integrating Digital Libraries*. PhD thesis, Virginia Tech CS Department, Blacksburg, Virginia, 2006. URL - http://scholar.lib.vt.edu/theses/available/etd-04212006-135018/.

[19] G. Tsakonas, S. Kapidakis, and C. Papatheodorou. Evaluation of user interaction in digital libraries. In M. Agosti, N.Fuhr, ed.: Notes of the DELOS WP7 Workshop on the Evaluation of Digital Libraries, 2004. URL - http://dlib.ionio.gr/wp7/workshop2004/_program.html.

Author Index