Proceedings of the 4th International Workshop on **Semantic Digital Archives** (SDA 2014)

held as part of the
International Digital Libraries Conference (DL2014)
September 12, 2014 in London, UK

http://sda2014.dke-research.de

Edited by:

Thomas Risse, L3S Research Center, Hannover, Germany
Livia Predoiu, University of Oxford, United Kingdom
Andreas Nürnberger, Otto-von-Guericke University of Magdeburg, Germany
Seamus Ross, University of Toronto, Canada

CEUR Workshop Proceedings
http://ceur-ws.org
ISSN 1613-0073

Vol-1306
urn:nbn:de:0074-1306-1

Preface

The 4th Workshop on Semantic Digital Archives (SDA 2014) has built upon the success of the previous editions in 2011 to 2013 and has been held as part of the International Digital Libraries Conference (DL2014) on September 12, 2014 in London, UK. Organized as full-day workshop, SDA 2014 has aimed to promote and discuss sophisticated knowledge representation and knowledge management solutions specifically designed for improving Archival Information Systems.

Archival Information Systems (AIS) are becoming increasingly important. For decades, the amount of content created digitally is growing and its complete life cycle nowadays tends to remain digital. A selection of this content is expected to be of value for the future and can thus be considered being part of our cultural heritage. However, digital content poses many challenges for long-term or indefinite preservation, e.g. digital publications become increasingly complex by the embedding of different kinds of multimedia, data in arbitrary formats and software. As soon as these digital publications become obsolete, but are still deemed to be of value in the future, they have to be transferred smoothly into appropriate AIS where they need to be kept accessible even through changing technologies.

The successful previous SDA workshops showed: Both, the library and the archiving community have made valuable contributions to the management of huge amounts of knowledge and data. However, both are approaching this topic from different views which shall be brought together to cross-fertilize each other. There are promising combinations of pertinence and provenance models since those are traditionally the prevailing knowledge organization principles of the library and archiving community, respectively. Another scientific discipline providing promising technical solutions for knowledge representation and knowledge management is semantic technologies, which is supported by appropriate W3C recommendations and a large user community. At the forefront of making the semantic web a mature and applicable reality is the linked data initiative, which already has started to be adopted by the library community. It can be expected that using semantic (web) technologies in general and linked data in particular can mature the area of digital archiving as well as technologically tighten the natural bond between digital libraries and digital archives. Semantic representations of contextual knowledge about cultural heritage objects will enhance organization and access of data and knowledge. In order to achieve a comprehensive investigation, the information seeking and document triage behaviors of users (an area also classified under the field of Human Computer Interaction) needs also to be included in the research.

One of the major challenges of digital archiving is how to deal with changing technologies and changing user communities. On the one hand software, hardware and (multimedia) data formats that become obsolete and are not supported anymore still need to be kept accessible. On the other hand changing user communities necessitate technical means to formalize, detect and measure knowledge evolution. Furthermore, digital archival records are usually not deleted from the AIS and therefore, the amount of digitally archived (multimedia) content can be expected to grow rapidly. Therefore, efficient storage management solutions geared to the fact that cultural heritage is not as frequently accessed like up-to-date content residing in a digital library are required. Software and hardware needs to be tightly connected based on sophisticated knowledge representation and management models in

order to face that challenge.

In line with the above, we invited contributions to the workshop that focus on:

- Architectures and Frameworks for semantic AIS and Archival Information Infrastructures (AII)
- Semantic (Web) services implementing AIS & AII
- Contextualization of digital archives, museums and digital libraries
- Linked data for AIS, AII, museums and digital libraries
- Ontologies for AIS, AII, museums and digital libraries
- Semantics of complex content (e.g. Social Media, Multimedia)
- Information integration/semantic ingest (e.g. from digital libraries)
- Semantic search & information retrieval in digital archives, digital museums and digital libraries
- User interfaces for (semantic) AIS, AII, digital museums & semantic digital libraries
- Semantics for Preservation Processes and Protocols
- Preservation of work flow processes
- (Semantic) provenance models
- Semantics for the appraisal and selection of content
- Evolving semantics in long-term archives
- Trust for ingest & data security/integrity check for long-term storage of archival records
- User studies focusing on end-user needs and information seeking behavior of end-users
- Implementations & evaluations of (semantic) AIS, AII, semantic digital museums & semantic digital libraries
- Semantic long-term storage & hardware organization for AIS & AII & digital libraries

We received submissions covering a broad range of relevant topics in the area of semantic digital archives. With the help of our program committee all articles were peer-reviewed. These proceedings comprise all accepted submissions which have been carefully revised and enhanced by the authors according to the reviewers' comments.

We sincerely thank all members of the program committee for supporting us in the reviewing process. Altogether, the diversity of the papers in these proceedings represent a multitude of interesting facets about the exciting and promising research field of semantic digital archives and semantic digital archiving infrastructures. During the workshop itself we had many fruitful and inspiring discussions which would not have been possible without the well done presentations and the interested audience. Many thanks to all workshop attendants for a great workshop!

We would also like to thank Sun SITE Central Europe for hosting these proceedings on http://ceur-ws.org.

December 2014

T. Risse, L. Predoiu, A. Nürnberger, and S. Ross

Organizing Committee

- Thomas Risse, L3S Research Center, Hannover, Germany
- Livia Predoiu, University of Oxford, Oxford, UK
- Andreas Nürnberger, Otto-von-Guericke University, Magdeburg, Germany
- Seamus Ross, University of Toronto, Toronto, Canada

Program Committee

- Vassilis Christophides, Foundation of Research & Technology, Hellas, Greece
- Kai Eckert, University Library of Mannheim, Germany
- Alejandra Gonzalez-Beltran, University of Oxford, UK
- Marco Grassi, Università Politecnica delle Marche, Italy
- Tudor Groza, University of Queensland, Australia
- Armin Haller, CSIRO, Australia
- Andreas Harth, KIT, Karlsruhe, Germany
- Steffen Hennicke, Humboldt-Universität zu Berlin, Germany
- Stijn Heymans, SRI International, USA
- Pascal Hitzler, Wright State University, USA
- Christian Keitel, State Archive of Baden-Württemberg, Germany
- Claus-Peter Klas, FernUniversität in Hagen, Germany
- Birger Larsen, Royal School of Library and Information Science, Denmark
- Thomas Lukasiewicz, University of Oxford, UK
- Mathias Lux, Klagenfurt University, Austria
- Maria Vanina Martinez, University of Oxford, UK
- Annett Mitschick, TU Dresden, Germany
- Knud Möller, Talis, Birmingham, UK
- Kai Naumann, State Archive of Baden-Württemberg, Germany
- Gillian Oliver, Victoria University of Wellington, New Zealand
- Jacco van Ossenbruggen, VU University Amsterdam, Netherlands
- Andreas Rauber, Vienna University of Technology, Austria
- Sebastian Rudolph, Karlsruher Institut für Technologie, Germany
- Heiko Schuldt, Universität Basel, Switzerland
- Mike Salampasis, Alexander Techn. Educational Inst. (ATEI) of Thessaloniki, Greece
- Kunal Sengupta, Wright State University, USA
- Gerardo Simari, University of Oxford, UK

- Herbert van de Sompel, Los Alamos National Laboratory Research Library, USA
- Marc Spaniol, Max-Planck-Institut Saarbrücken, Germany
- Manfred Thaller, University of Cologne, Germany

Table of contents

Invited Talk

Understanding Web Archives 1
Helen Hockx-Yu

Semantics for Interoperability and Identification

Self-contained Information Retention Format for Future Semantic Interoperability 4
Simona Rabinovici-Cohen, Roger Cummings and Sam Fineberg

Identification Semantics for an Organization Establishing a Digital Library System.....16
Angela Di Iorio and Marco Schaerf

Persistent E-mail Identification is Viable! 28
Stefan Haun and Andreas Nürnberger

Culture meets Semantics

The Digital Online Museum: A new Approach to Experience Virtual Heritage.....38
Tilman Deuschel, Timm Heuss and Bernhard Humm

Relations for Reusing (R4R) in a Shared Context: An Exploration on Research Publications and Cultural Objects 49
Andrea Wei-Ching Huang and Tyng-Ruey Chuang

Understanding Web Archives

Helen Hockx-Yu

Head of Web Archiving
British Library, London, UK,
Helen.Hockx-Yu@bl.uk

Abstract. This talk provides an insight into web archives by examining the "unknown" aspects beyond the archived web pages, or the "text". It argues that web archives have a rich set of semantics which when explored offers a new way of understanding their characteristics. It showcases examples of British Library's work beyond the "document-centric" approach of providing access.

Keywords: Web archives, exploration, semantics

1 Introduction

The effort to archive the web started in the mid-1990s, a few years after the web was born. This was initiated by the Internet Archive in the US. Many national libraries and archives, which traditionally have the duty to preserve a nation's cultural and scientific heritage, followed the suite and started actively collecting web content. Internet Archive's Wayback Machine[1] is the earliest and most comprehensive web archive to date, containing over 435 billion web pages archived from 1996. Many national heritage organisations have established collections covering their respective national web domain or subsets of it.

There are however issues related to the access and use of web archives: it is often restricted by legal requirements on one hand, in exchange for reproducing copyrighted material for the purpose of cultural heritage, and by the (single) envisaged use case on the other [HY14]. The latter is based on the assumption of web archives consisting of historical documents (web pages) used for reference. Researchers access previous states of individual web pages and websites in a web archive, which are selected, described and grouped together by curators, in the same way as printed books and journals. The over-focus on "documents" or "text" means contexts of archived material tend to be ignored or regarded as irrelevant.

2 Understanding Web Archives

A common assumption is that web archives contain copies of older versions of websites which are no longer current and have been replaced by the "live"

[1] https://archive.org/web/web.php

version. Brügger and Finnemann argue that archived web resources are "reborn", different from digitised and born digital collections and from the live web in many ways [BF13]. Using the British Library's web archive as an example, this talk examines in detail the many boundaries and imitations related to web archive, determined by purpose, strategy, legal requirements and technological choices. It also points out a fundamental oversight which impacts users' interpretation or understanding of web archives: very little is explained or made clear to the users beyond the actual HTML pages (or the "text"). A typical example of this is the common error message "Resource Not in Archive", which is presented to the end-users when a requested URL cannot be found in the archive. This could be caused by many reasons: some are intended, introduced by things like data limitation at crawl time or content beyond the scope of the crawl; others relate to technical limitations, e.g. dynamic content which the web crawlers are not capable of collecting.

3 More to "text"

Effort started to emerge in recently years which moves away from the level of single webpages or websites to the entire web archive collection. Using visualisation and data analytic techniques, new ways have been developed to view web archives, offering opportunity to unlock embedded patterns and trends, relationships and contexts, which are not possible by consulting websites individually. This is in alignment with the changes in scholarly practices as researchers increasingly take advantage of new possibilities offered by technology. New methods of scholarship are emerging, which challenges the primacy of "text" as object of study. This talk references the concepts of "paratexts"[Nie10] and "distant reading"[Mor00], as theoretical basis for using web archives as scholarly sources. The role of web archives is to provide services supporting scholars who read texts differently.

This talk focuses on a range of non-text attributes of web archives (including an example visualisation or demo for each), explored by the British Library or others, as additional ways of understanding web archives. Scholars are encouraged to explore other contextual or "para-textual" content in the web archives, such as viral content and crawl logs.

- Statistical overview, scale and distribution of a web national domain
- Size: bytes
- Space: geo location, postcodes
- Type of content, e.g. file format, language
- Structure, linked entities and networks
- Evolution, pattern of change over time, e.g. domain names
- Correlation, e.g. between certain term and historical event

This talk also discusses the general issues related to analytical access, such as researchers' scepticism or suspicion about hidden algorithms behind analysis, and how biases in data and how data collection decisions lead to variances in outputs.

References

[BF13] Niels Brügger and Niels Ole Finnemann. The web and digital humanities: Theoretical and methodological concerns. *Journal of Broadcasting & Electronic Media*, 57(1):66–80, 2013.

[HY14] Helen Hockx-Yu. Access and scholarly use of web archives. *Alexandria*, 25(1):113–127, August 2014.

[Mor00] Franco Moretti. Conjectures on world literature. `http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature` (accessed on 17 November 2014), 2000.

[Nie10] Niels Brügger. Website analysis: Elements of a conceptual architecture. `http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/cfis_skriftserie/012_brugger.pdf` (accessed on 17 November 2014), 2010.

Self-contained Information Retention Format For Future Semantic Interoperability

Simona Rabinovici-Cohen[1], Roger Cummings[2], and Sam Fineberg[3]

[1] IBM Research – Haifa
simona@il.ibm.com
[2] Antesignanus
roger@antesignanus.com
[3] HP Storage
fineberg@hp.com

Abstract. Long term preservation of digital information, including machine generated large data sets, is a growing necessity in many domains. A key challenge to this need is the creation of vendor-neutral storage containers that can be interpreted over time. We describe SIRF, the Self-contained Information Retention Format, which is being developed by the Storage Networking Industry Association (SNIA) to support this challenge. We define the SIRF components, its metadata, categories and elements, along with some security guidelines. SIRF metadata includes the semantic information as well as schema and ontological information needed to preserve the physical integrity and logical meaning of preservation objects. We also describe how the SIRF logical format is serialized for storage containers in the cloud and for tape based containers. Aspects of SIRF serialization for the cloud are being experimented with OpenStack Swift object storage in the ForgetIT EU project.

1 Introduction

Generating and collecting very large data sets is becoming a necessity in many domains that also need to keep that data for long periods. Examples include genomics, medical records, astronomy, atmospheric science, photographic archives, video archives, and large-scale e-commerce. While this presents significant opportunities, a key challenge is providing economically scalable storage systems to efficiently store and preserve the data. This includes both the data itself as well as semantic metadata necessary to enable search, access, and analytics on that data in the far future.

The Storage Networking Industry Association (SNIA) conducted a "100 year archive" survey. It found that 83% of the organizations surveyed have digital assets they need to retain for over 50 years, and 53% have information they need to retain "permanently". Recognizing these challenges, SNIA formed the Long Term Retention (LTR) group [1] to address storage aspects of digital retention. LTR is working on the Self-contained Information Retention Format (SIRF), to create a standardized vendor neutral storage format that will help its users

interpret preservation objects in the future even by systems and applications that do not exist today. SIRF provides strong encapsulation of large quantities of metadata with the data at the storage level, and enables easy migration of the preserved data across storage devices.

Both cloud storage and tape technologies are viable alternatives for storage of data for the long term. Cloud technology is emerging as an infrastructure suitable for building large and complex systems, presenting a scalable and cost-effective alternative to the traditional storage systems. Thus, the cloud is clearly an attractive platform for long term preservation solutions, and in particular, cloud storage can be leveraged for preservation-aware storage [2].

Tapes are attractive for long term data retention as their expected lifetime is higher than that of other types of media and their cost is considerably lower. Moreover, The SNIA Linear Tape File System (LTFS) takes advantage of a new generation of tape hardware to provide efficient access to tape using standard, familiar system tools and interfaces. This paper combines SIRF with cloud technology, as well as separately combines it with tape technology.

A core standard for digital preservation systems is the Open Archival Information System (OAIS)[4], an ISO standard since 2003 (ISO 14721:2003 OAIS). OAIS metadata can also include semantic metadata [3] to facilitate the preservation of schemas and ontological information. However, OAIS is a high-level reference model, which means it is flexible enough to be used in a wide variety of environments. More detailed steps and workflow stages need to be developed for the implementation of an OAIS based system. SIRF adds more detail to the metadata needed in the storage container.

SIRF uses cases and functional requirements were described in [4] along with the substantial differences from other formats. Our main contribution in this paper includes the definition of the SIRF format for long term storage containers. We define the SIRF catalog metadata, its categories and elements along with the rationale behind them. To show that SIRF can be combined with different types of underlying storage containers, we describe SIRF serialization for the cloud and SIRF serialization for tapes. We also provide some implementation overview of SIRF aspects in OpenStack cloud object storage[5] that is being examined in the context of the ForgetIT[6] European Union integrated research project.

The rest of this paper is organized as follows. In section 2, we discuss the business need of storage containers for long term retention. In section 3, we introduce the SIRF container format, its components and metadata. Section 4 defines the serialization for cloud and for tapes. Section 5 describes some aspects of experimental usage of SIRF in ForgetIT project for concise managed preservation of personal data and organizational web sites. In section 6, we review related work and conclude with a summary and some future work.

[4] http://public.ccsds.org/publications/archive/650x0m2.pdf
[5] http://www.openstack.org/software/openstack-storage
[6] http://www.forgetit-project.eu

2 Business Need for Long Term Retention

While no one wants to lose their digital content, the cost of maintaining integrity and access is significant, in both money and effort. And unlike paper based content, the lifespan of digital content can be very short unless if proactive steps are being taken to protect it. The use of a storage container format like SIRF adds little expense and greatly increases the sustainability of data. However, this is not adequate unless if the cost of preserving content is less than the (potential) cost of losing it.

In a business context, there are three major reasons why content is preserved. These are: to preserve history, to mitigate risk or meet a legal mandate, and for future value of information. One or more of these may apply, and the amount an entity is willing to spend will differ depending on how well these reasons are aligned with the business goals of an organization.

One of the main reasons why people and organizations preserve content is to preserve history. In the case of an individual, it may be photos, videos, and other content preserving one's life history. In a business context, libraries, national archives, historians, and others have a primary mission to preserve history.

Another often cited reasons for preserving data is for "risk mitigation", or in some cases for "legal mandate". These are closely related reasons because legal mandate is often looked at through the lens of legal risk. For example, an often cited legal mandate is in healthcare, where medical organizations are required to retain information for the lifetime of a patient. This seems like a difficult requirement, especially since records are often maintained in private doctors' offices and other places that may not exist 50 or 75 years into the future. Anecdotal evidence shows that medical records are not maintained that long. So, why is this happening? It is because records retention is expensive, and there are no penalties for losing information. That is not to say that doctors and hospitals don't try, rather they won't spend the necessary money.

Regarding future value of information, one obvious example is in the entertainment industry. Movies, TV shows, music, and other content can be re-sold and repurposed decades after its creation. This can result in many dollars in revenue. So not surprisingly, organizations like the Motion Picture Expert's Group are at the leading edge of digital preservation. Entertainment companies spend significant amounts of money retaining their content so that they will have it available to repurpose. However, this does not mean they can retain everything. With the advent of digital movie production, the amount of data that can be generated during the creation of a single film is immense. Therefore, even here where future value is tangible, some hard choices need to be made.

So, how does SIRF help? SIRF brings down the expense of preservation, because data can remain accessible even if the software that created the data no longer exists. SIRF reduces the complexity of logical and physical migration, making it easier for businesses to justify. By using SIRF today, it becomes possible to retain more information, and to retain information with a lower perceived future value. This is unlike proprietary and undocumented formats, which become useless soon after a business stops paying for support.

3 The SIRF Format

3.1 SIRF Components

Archivists and records managers of physical items such as documents, objects, records, etc., avoid processing each item individually. Instead, they gather together a group of items that are related in some manner - by usage, by association with a specific event, by timing, and so on - and then perform all of the processing on that group as a unit. Once assembled, an archivist will place the collection in a physical container (e.g. a file folder or a filing box of standard dimensions), and that container is attached with a label that gives an overview of the container content e.g. name and reference number, date, contents description, destroy date.

We propose an approach to digital content preservation that leverages the knowledge of the archival profession and helps archivists remain comfortable with the digital domain. We define a digital equivalent to the physical container - the archival box or file folder - that defines a collection, and which can be labeled with standard information in a defined format to allow retrieval when needed. SIRF is intended to be that equivalent - a storage container format for a set of (digital) preservation objects that includes a catalog with metadata related to the entire contents of the container as well as to the individual objects and their interrelationship. This logical container makes it easier and more efficient to provide many of the processes that will be needed to address threats to the digital content.

SIRF is a logical container format for the storage subsystem, appropriate for the long-term storage of digital information. It is a logical data format of a mountable unit e.g. a filesystem, a cloud container, an object store, a tape, etc. It assumes the mountable unit includes an object interface layer that constructs objects out of the sectors and blocks.

Figure 1 illustrates the SIRF container, which includes the following components:

- A magic object that identifies whether this is a SIRF container and gives its version. The magic object is independent of the media and has an agreed defined name and a fixed size. It also includes the means to access the SIRF catalog (for example, the catalog's location).
- Preservation objects that contain the actual data to be preserved. An example preservation object can be the OAIS Archival Information Package (AIP). The container may include multiple versions of a preservation object and multiple copies of each version, but each specific preservation object is generally immutable.
- A catalog that is updateable and contains semantically enriched metadata needed to make the container and its preservation objects portable, accessible, and understandable into the future without relying on metadata external to the storage subsystem.

While traditional storage systems include only limited standardized metadata about each object, SIRF provides the semantically rich metadata needed for long

term preservation and interpretation of information, and ensures its grouping with the data. This rich metadata is defined in the catalog in a logical format to allow its serialization for different storage technologies. We show its mapping to some of today's storage containers (cloud storage and tapes), but as new storage technologies become prevalent in the future, additional mappings will need to be defined.

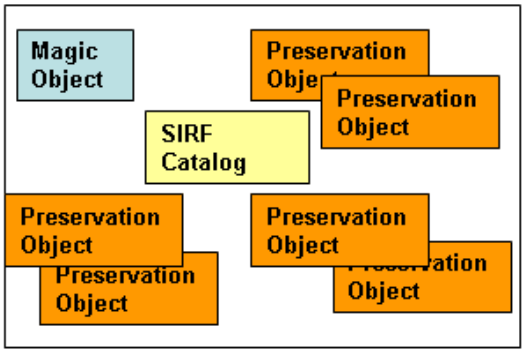


Fig. 1: SIRF Components

3.2 SIRF Catalog Metadata Schema

The SIRF catalog is an object that includes metadata about the preservation objects (POs) in the container and their schema and interelationships. It has a well-defined standardized format so it can be understandable in the future. The SIRF catalog is separated from the metadata contained in the POs themselves because a strict standardized format is difficult to impose on the POs that are generated by different applications and domains. Additionally, the SIRF catalog includes some metadata that is not included in the PO e.g. fixity value of the whole PO. Including this metadata within the PO changes the fixity value of the PO making this metadata inherently incorrect.

The SIRF catalog includes metadata related to the whole container as well as metadata related to each preservation object within the container. Both types of metadata are divided into categories, elements and attributes organized in a hierarchical representation. The full metadata definitions and the rationale behind them are defined in SIRF draft specification[7]. Here we provide some example categories for the whole container metadata in subsection 3.2.1 and for each preservation object within the container in subsection 3.2.2 below.

3.2.1 Container Information Metadata Schema. The metadata for the whole container includes the categories Specification, Container ID, State, and Container Provenance.

[7] http://www.snia.org/tech_activities/publicreview, to appear

The Specification category includes information about the specification used. As the specification may evolve over time and distinct storage containers may use different SIRF specifications, it's important to include the exact version of the specification in the SIRF catalog including specification ID and specification version.

The Container ID category includes the container unique identifier such as the tape ID for tape based storage containers or cloud container ID in case of cloud storage.

The State category is an indication of the progress of any activities that are to be carried out against a container. For example, if a container holds many preservation objects, state may indicate whether all of the objects intended for a container have been included or not. Or, state may indicate an in-process migration of a container. Multiple state entries are allowed in case if there are multiple pending activities.

The Container Provenance category is metadata describing the history of the information in a SIRF container (e.g., its origins, chain of custody, preservation actions and effects). The Provenance information may vary depending on the type of information being preserved or its intended audience and it may be large. Therefore, it is included in the catalog by reference, and the actual information is stored in another preservation object. The container provenance information stored in SIRF may be in the W3C-PROV format, or any other well known provenance format. Regardless of the perspective from which provenance metadata is derived, it is critical for understanding the container, its history, its context and meaning.

3.2.2 Object Information Metadata Schema. The metadata for each preservation object includes several categories; from which we'll describe here: Object IDs, Fixity, and Audit log.

The Object Identifiers (IDs) category is used to identify a PO and to link to other POs. Managing identifiers over the long term raises issues such as: how to ensure uniqueness of identifiers over long term, how to handle evolution of identifiers over time, how to ensure scalability of identifiers.

SIRF helps to address these issues by enabling redundancy in identifiers and registering the evolution (genealogy) of POs. Hence, a PO in a SIRF container can have multiple identifiers as redundant identifiers. This increases the chances that at least one of the identifiers will survive for the long term. Nevertheless, at any time, at least one of the identifiers should be persistent and unique.

Fixity is used to demonstrate that the content information has not been altered in an undocumented or unauthorized manner. The fixity information can be seen as an integrity check value. Fixity is sometimes computed via simple cheap functions such as a CRC, or it can include a stronger and more expensive (in execution time and space) cryptographic hash function such as MD5 or SHA-512. No matter how strong the fixity computation functions are, they are likely to become obsolete in the far future when larger amounts of storage and stronger computing power are available. Thus, the preservation system should

be allowed to update fixity functions in the future, as existing ones become obsolete. Consequently, the SIRF catalog allows for multiple fixity algorithms and values for a given PO.

The audit log category is provided as a place for preserving any important information about how an object has been accessed or modified. The extent and contents of an audit log depend on the needs of the specific preservation data store and its use case. Distinct domains have different audit logs regulations e.g., SEC is for the US financial market domain, FDA is for the US medical domain. In SIRF, audit logs are stored in the catalog as links to preservation objects.

3.3 SIRF Container Security Guidelines

Some of the legal mandates for information retention also incorporate requirements for privacy and access protection. Where such security-based requirements exist, they add another level of complexity to long-term retention of the SIRF container. Much of this additional complexity results from the fact that the security-based requirements tend to mitigate against other retention requirements. For instance, while retention generally seeks to make information widely available and usable, security tends to restrict access to ensure that information privacy is maintained.

Information security also adds significantly to the amount of metadata that must be maintained within the container to ensure future usability of the information. Most obvious is the need to identify the encryption scheme used, and the need to maintain information about the different types of access that should be granted to the information. All access information needs to be based on the definition of abstract roles rather than specific people because, given the time periods being addressed by long-term retention, people will change job functions, organizations will grow, merge, or disappear, and uses for the information may significantly alter. A long-term retention system must be able to continuously add new users and associate them with existing roles, and change the roles assigned to existing users.

The management of keying information, whether related to the encryption of information or to the authentication of the roles assigned to specific users, presents a specific challenge in terms of long-term retention. Clearly such information cannot directly be located within the container itself, but sufficient metadata must be included in the container to allow the keying information to be located, validated, and verified.

ISO/IEC 27040 draft is being created to address the security of both local and cloud-based security systems. It emphasizes that there are integrity, authentication, and privacy threats that are particular to long-term storage systems. It also notes that the long lifetime of information within such systems enables attacks that require a large amount of access to the information but which can be disguised as many small requests over an extended period of time. It highlights the importance of maintaining a log of attack attempts, compromises, and system and user changes, and notes that such a log must also be maintained for

the long-term. In the current version of SIRF, we support some initial security guidelines via e.g., the Fixity and the Audit Log categories.

4 SIRF Serialization for Cloud and for Tape

The SIRF serialization for cloud/tape specifies how a cloud container or a tape container becomes SIRF-compliant. A SIRF-compliant cloud container or tape container enables future's cloud/tape clients to "understand" containers created by today's cloud/tape clients even though the properties of the future client is unknown today. By "understand", we mean we can identify the preservation objects in the container, the packaging format of each object, its fixity values, etc. (as defined in the SIRF catalog).

For the concrete serialization we chose specific standard based storage containers. For the cloud, we chose CDMI[8] and OpenStack object storage while for tapes we chose LTFS[9] based tapes. No single technology will be usable over the time spans mandated by current digital preservation needs. SNIA CDMI and LTFS technologies are among best current choices, but are good for perhaps 10-20 years. SIRF provides a vehicle for collecting all of the information that will be needed to transition to new technologies in the future, and it can be serialized for future technologies as they emerge.

For the serialization step, we classify the preservation objects as either simple preservation object or composite preservation object. A simple PO contains just one element and is mapped to one object in the CDMI cloud or one file in the LTFS tape. A simple PO can be for example a jpg photo or a tar file. A composite PO contains several elements and a manifest that combines the elements. The composite PO is mapped to several objects in the CDMI cloud or a number of files in the LTFS tape.

4.1 Serialization For Cloud Storage

The Cloud Data Management Interface (CDMI) is an ISO/IEC 17826:2012 standard created by SNIA that defines an interoperable format for moving data and associated metadata between cloud providers. CDMI has several implementations including an open source implementation for OpenStack Swift[10] cloud storage.

A CDMI cloud container can be qualified as a SIRF container when:

- The SIRF magic object is mapped to the CDMI container metadata.
- The SIRF catalog is an object in the CDMI container formatted in JSON (self-describing) that includes one containerInformation section and multiple objectInformation sections - one for each PO within the container (self-contained). This object should be indexed (if possible). There is a CDMI extension to support indexing with object granularity.

[8] Cloud Data Management Interface - http://www.snia.org/cdmi
[9] Linear Tape File System - http://www.snia.org/ltfs
[10] Swift - https://wiki.openstack.org/wiki/Swift

– A SIRF PO that is a simple object (contains one element) is mapped to a CDMI data object.
– A SIRF PO that is a composite object is mapped to a set of data objects (one for each element) and a manifest data object that includes information about the elements.

The interface to the SIRF-compliant CDMI container is the ordinary CDMI Application Program Interface (CDMI API). In addition, the CDMI API can be used to store and access the various preservation objects and the catalog object.

For example, figure 2 depicts a CDMI container named "Patient Container" that is SIRF-compliant and includes medical encounters and images for the patient. Assume each encounter is a simple preservation object; each image is a composite preservation object; and since the container is SIRF-compliant, it also includes a catalog object.

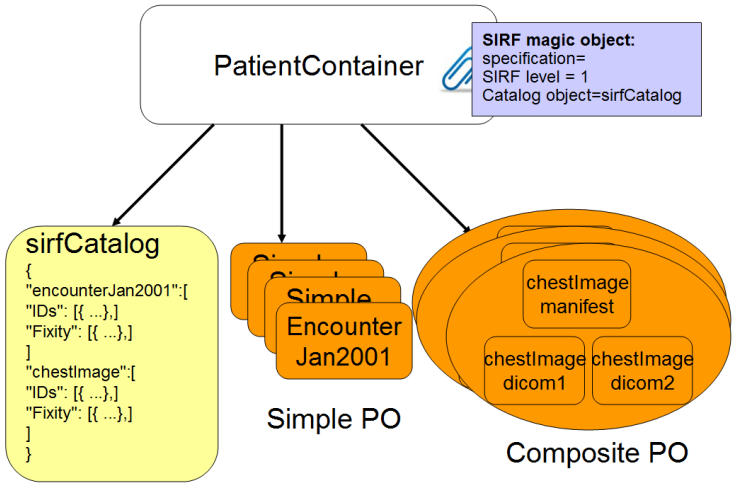


Fig. 2: SIRF Seralization for CDMI Example

4.2 Serialization For Tapes

The Linear Tape File System (LTFS) format specification defines LTFS Volumes. An LTFS Volume holds data files and corresponding metadata to completely describe the directory and file structures stored on the volume. Files can be written to, and read from, an LTFS Volume using standard POSIX file operations. The LTFS Volume includes an index in XML that contains metadata similar to information in disk-based file systems such as file name, dates, extent pointers, extended attributes, etc. LTFS is becoming the standard for linear tape and is being formalized through SNIA.

An LTFS volume is comprised of a pair of LTFS partitions: a data partition (DP) and an index partition (IP). Each partition contains a Label Construct followed by a Content Area. As depicted in figure 3, a LTFS tape container can be qualified also as a SIRF container when the volume format is as follows:

- The SIRF magic object is mapped to extended attributes of the LTFS index root directory.
- The SIRF catalog resides in the index partition and formatted in XML (self-describing) that includes one containerInformation section and multiple objectInformation sections - one for each PO within the container (self-contained). LTFS application has rules to indicate what to store in the index partition. That method can be used to indicate to store the SIRF catalog in the index partition. Alternatively, the index partition can include a reference to the SIRF catalog that will reside in the data partition.
- A preservation object (PO) is mapped to an LTFS file or set of files. In case the PO is a simple object composed of one element, it is mapped to a LTFS file. In case the PO is a composite object composed of several elements, it is mapped to a set of LTFS files (one for each element) and a manifest file that its content includes information about the elements.

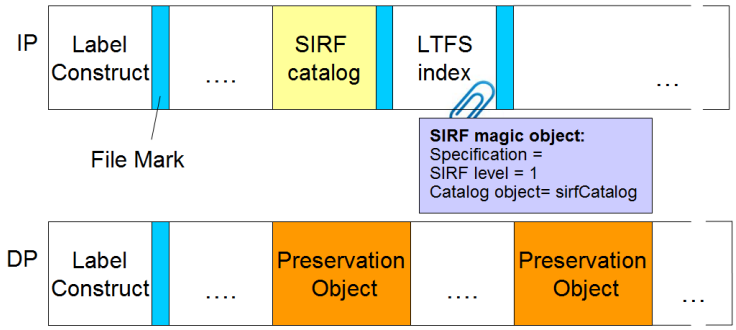


Fig. 3: SIRF Seralization for LTFS Volume

5 SIRF in ForgetIT

The European Union integrated project ForgetIT investigates ways for concise long term digital preservation and its adoption for personal data and organizational web sites. It combines three new concepts: managed digital forgetting inspired from human brain and cognitive psychology; smooth transition between data active use and its preservation; contextualized remembering keeping the archive understandable and useful.

The ForgetIT Preserve-or-Forget framework uses the DSpace open source as its preservation system, where the archival storage is Preservation DataStores (PDS) in the Cloud [2] that provides preservation-aware storage services based on the OAIS model. PDS includes the Preservation Engine and the Storlet Engine. The Preservation Engine transforms the logical OAIS functions and information objects into processes and physical storage objects. The Preservation Engine sometimes requires performing data-intensive computational tasks, such as transformation, migration, fixity checks, and data analysis. When the Preservation Engine requires performing such tasks, it uses storlets - computational

modules running in a sandbox close to the data. Offloading OAIS-based functionality to the storage decreases probability of data loss, simplifies the applications and supports automation of preservation processes.

The Storlet Engine [5] provides the cloud storage with a capability to include storlets that run within the storage in a sandbox that provides isolation. It is plugged into a private cloud or object storage such as OpenStack Swift and provides a powerful extension mechanism that makes the storage flexible, customizable and extensible. By using storlets, the client benefits of reduced bandwidth (reduce the number of bytes transferred over the WAN), enhanced security (reduce exposure of sensitive data), cost saving (reduce infrastructure at the client side), and compliance support (improve provenance tracking).

PDS in ForgetIT implements some aspects of SIRF. It creates the various identifiers used for maintaining the evolution of POs, which can be stored in the Object IDs category in the SIRF catalog.

Regarding the fixity category in the SIRF catalog, PDS developed a fixity storlet that can compute multiple fixity values for each PO, and new hash functions can be uploaded to the storage as older ones become too weak or even obsolete.

While the ForgetIT POs are generated by different applications and domains (personal and organizational use cases), the SIRF catalog presents a standardized format that can be interpreted in the future.

6 Discussion and Conclusions

6.1 Related Work

Storage aspects of archiving and preservation systems have been the focus of a growing number of studies. You et al. [6] present PRESIDIO, a scalable archival storage system that efficiently stores diverse data. Adams et al. [7] studied scientific and historical archives, covering a mixture of purposes, media types, and access models. Based on this study, they identify areas for improving the efficiency and performance of archival storage systems.

Long-term preservation systems differ from traditional storage applications with respect to goals, characteristics, threats, and requirements. Baker et al. [8] examine these differences and suggest bit preservation guidelines and alternative architectural solutions that focus on replication across autonomous sites. Storer et al. [9] discuss security threats that arise when storing data for long periods of time. This includes common threats such as loss of integrity, failure of authentication and compromise of privacy, as well as new specific threats such as slow attacks.

Dappert and Enders [10] discuss the importance of metadata in a long term preservation solution. The authors identify several categories of metadata, including descriptive, preservation related, and structural, arguing that no single existing metadata schema accommodates the representation of all categories. The work surveys metadata specifications contributing to long-term preservation.

6.2 Conclusions and Future Work

Moving forward, digital content preservation will have many technical and cultural challenges. As digital technologies continue to replace physical ones, these challenges must be solved to prevent us from losing a generation of content.

SIRF, the Self-contained Information Retention Format, was developed to address the growing necessity to preserve digital information over long periods of time. SIRF does this by acting as the digital equivalent of an archivist's "box". SIRF preserves data and metadata as a single unit and provides a catalog containing the basic metadata needed to access and preserve content. This aids in the future understanding of data, and in the migration to new storage devices and formats.

We have shown that the SIRF can be serialized for a variety of storage technologies including LTFS based tape and CDMI cloud containers. This should provide a means for preserving information for the next years, and a vehicle for migrating to whatever new storage technologies become prevalent in the future.

In future work, we would like to improve support for the security guidelines developed in ISO/IEC 27040. Also, we would like to experiment SIRF in other projects and serialize it for additional storage containers.

Acknowledgments. This work was partially funded by the European Commission in the context of the FP7 ICT project ForgetIT (under grant no: 600826).

References

1. SNIA Long Term Retention (LTR) group. URL: `http://www.snia.org/ltr`
2. Rabinovici-Cohen, S., Marberg, J., Nagin, K., Pease, D.: PDS Cloud: Long Term Digital Preservation in the Cloud. In: IC2E 2013: Proceedings of the IEEE International Conference on Cloud Engineering, San Francisco, CA (March 2013)
3. Brunsmann, J.: Product Lifecycle Metadata Harmonization with the Future in OAIS Archives. In: DC 2011: Proceedings of the International Conference on Dublin Core and Metadata Applications, Hague, The Netherlands (2011)
4. Rabinovici-Cohen, S., Baker, M., Cummings, R., Fineberg, S., Marberg, J.: Towards SIRF: Self-contained Information Retention Format. In: SYSTOR 2011: Proceedings of the International Systems and Storage Conference, Israel (2011)
5. Rabinovici-Cohen, S., Henis, E., Marberg, J., Nagin, K.: Storlet Engine: Performing Computations in Cloud Storage. IBM Technical Report H-0320 (August 2014)
6. You, L., Pollack, K., Long, D., Gopinath, K.: PRESIDIO: A Framework for Efficient Archival Data Storage. ACM Transactions on Storage **7**(2) (July 2011)
7. Adams, I.F., Storer, M.W., Miller, E.L.: Analysis of Workload Behavior in Scientific and Historical Long-Term Data Repositories. TOS **8**(2) (2012)
8. Baker, M., Shah, M., Rosenthak, D., Roussopoulos, M., Maniatis, P., Giuli, T., Bungale, P.: A Fresh Look at the Reliability of Long-Term Digital Storage. In: Proceedings of the 1st ACM SIGOPS European Systems Conference. (2006)
9. Storer, M.W., Greenan, K.M., Miller, E.L., Voruganti, K.: POTSHARDS - A Secure, Recoverable, Long-Term Archival Storage System. TOS **5**(2) (2009)
10. Dappert, A., Enders, M.: Digital Perservation Metadata Standards. Information Standards Quarterly, Special Issue on Digital Preservation **22**(2) (2010) 4–12

Identification Semantics for an Organization, establishing a Digital Library System

A. Di Iorio, M. Schaerf

DIAG - Department of Computer, Control, and Management Engineering Antonio Ruberti - Sapienza University of Rome, Italy

Abstract. The Sapienza Digital Library collects digital resources from the different University's Organizations, representing the multidisciplinary Sapienza University's community. The underlay of the metadata infrastructure was built on digital library standard metadata semantics and was used for exchanging package, between the archival systems that manages different services for the established digital library. The semantics adopted for the metadata infrastructure can be exploited, not only for the actual digital library services, but also for connecting the resources to the Linked Open Data Cloud through authoritative identifiers.

Keywords: Digital Libraries, Metadata Semantics, Organization metadata

1 Introduction

The paper describes a specific aspect of the development of the Digital Library System of the Sapienza university (Sapienza Digital Library `http://sdl.uniroma1.it`). The approach adopted collects information, regarding the Organizations involved in the management of the digital resources' life-cycle.
In order to manage the complexity of the Sapienza University's organizational framework, a workflow for building digital resources, based on the Organizational semantics, was designed at the first stage of the project's development.
The creation and the maintenance of an identification system, based on semantics used at national level, and mapped onto other identification systems, internationally used, was necessary, in order to make feasible the retrieval of relevant information in the Linked Open Data Cloud[1] through an authoritative identifier.
The system had been resulted essential, in the entire life-cycle of the project's development, in order to refer unambiguously to the digital resources among the project's participants. In addition it was supportive for testing and improving of the overall system's information infrastructure, for refining the metadata structures, and for curating the data.
The semantics of the SDL metadata infrastructure were used for building self-documenting packages containing metadata and objects, and for exchanging packages between different digital repository systems. The digital repositories,

[1] Linked Open Data, `http://linkeddata.org`

sharing the SDL semantics, uses the exchanging package for replicating digital resources and for distributing digital library services.

2 Background

The Open Archival Information System (OAIS) [4] defines the OAIS itself as "An Archive, consisting of an organization, [...]of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community." In addition, "The Archive is responsible for creating and preserving Provenance Information from the point of Ingest; however, earlier Provenance Information should be provided by the Producer. Provenance Information adds to the evidence to support Authenticity."
The DL.org [1] booklet remarks that "The Organization Domain stems from the Organization core concept and it is conceived to represent the main settings for characterizing the DL service..."
In our project the Organization Domain is identified by the establishing organization which indeed is, the Sapienza University. The digital resources' management of the Sapienza Digital Library(SDL)[5] was founded on the cited reference models, in particular considering the relationship between the provenance information and the Organization responsible for the management of digital resources. The production process of the OAIS Information Package (IP), used in the different functional scenario (Submission, Archiving, and Dissemination), was designed following the strategy of capturing relevant information about its custody, and exploiting the identification information associated to the Sapienza's Organizations. The self-documenting digital resource produced, can be used by other application systems sharing the standard metadata semantics, used by the SDL metadata infrastructure.

3 The long term scope of the system architecture

The system architecture was conceived with the scope of the Long Term Digital Preservation(LTDP) of materials for the multidisciplinary community of Sapienza.
The replication of the produced OAIS IPs in different repositories geographically separated, and the heterogeneity of the supporting technologies and methodologies[9][2], were considered influencing requirements, in the design of the overall architectural system.
As consequence, the initial scope of building a digital library was extended, and turned toward the conception of an infrastructure for digital library, and digital preservation services.
Following this conception, the metadata infrastructure had to be agnostic about the technological platform, in order to re-use information and objects in different digital systems, as well as in different semantic contexts. Nevertheless much

of the semantics, used for the values of the metadata elements, are often under the competence of the managing Organization. The semantics used if not well-documented and structured can be an obstacle, for the automatic management of data and documents, and consequently can have have a strong impact on the long term management of the digital resources.
Under this belief, the work-flow for building digital resources was conceived for absorbing information, conveying the custody chain of the management activities performed by different Organizations.
In other words the overall management of a digital resource, during its creation process, is permeated by the Organization's context information, connecting the digital resource to its "real" Organizations involved in the management of its production's .
An abstract representation of the main components of the overall architecture of the system is showed in the Figure 1. The main components can be divided in three categories: the pre-ingestion systems preparing the digital resources, the Digital Library Management System (DLMS)[1], performing the OAIS functional services[4], and the dissemination system. The system's components performing specific function in the architecture are briefly described in the following list.

- The Massive conversion system performs the retrospective conversion of existing digital materials, and related content's description, standardized or not standardized: it was developed for the need of Sapienza, extending a PHP/Mysql application, Bringing Digital Environment (BriDgE)[2].
- The Cataloguing system properly developed for describing collections of heterogeneous materials to be digitized.
- The DLMS as defined by the Delos Reference Model[3]: was developed extending services of Fedora Commons[4].
- The web portal of SDL, which manages the public interface of the system.

The Cataloging system and the web portal had been developed using Drupal[5] that uses services managed by the DLMS. The Italian University consortium Cineca[6], as technological partner of Sapienza for SDL, has developed the DLMS and the Cataloguing and the web portal systems.
Actually the repository, archiving the digital resources managed by the DLMS, is located in Bologna (the location of the Cineca's headquarter).
The exchange of IPs between pre-ingestion systems (Massive conversion and Cataloguing) and the DLMS, is performed between Sapienza repositories in Rome, and the DLMS's repository located in Bologna. This preservation strategy respects the influencing requirements of the LTDP: the digital resources replication in different repositories geographically separated, and the heterogeneity of the supporting technologies and methodologies[9][2].

[2] Bringing Digital Environment (BriDgE), `http://bri-dge.sourceforge.net/`
[3] DELOS Reference Model for Digital Libraries, `www.delos.info/ReferenceModel`
[4] Fedora Commons, `http://fedora-commons.org/`
[5] Drupal, `http://www.drupal.org/`
[6] Cineca website, `http://www.cineca.it`

The OAIS IPs produced by the pre-ingestion systems are the exchanging packages used by the systems supporting the different services. Consequently the IPs produced by the pre-ingestion systems has to be self-documenting, on the base of metadata and identification semantics shared by the SDL systems, geographically separated.

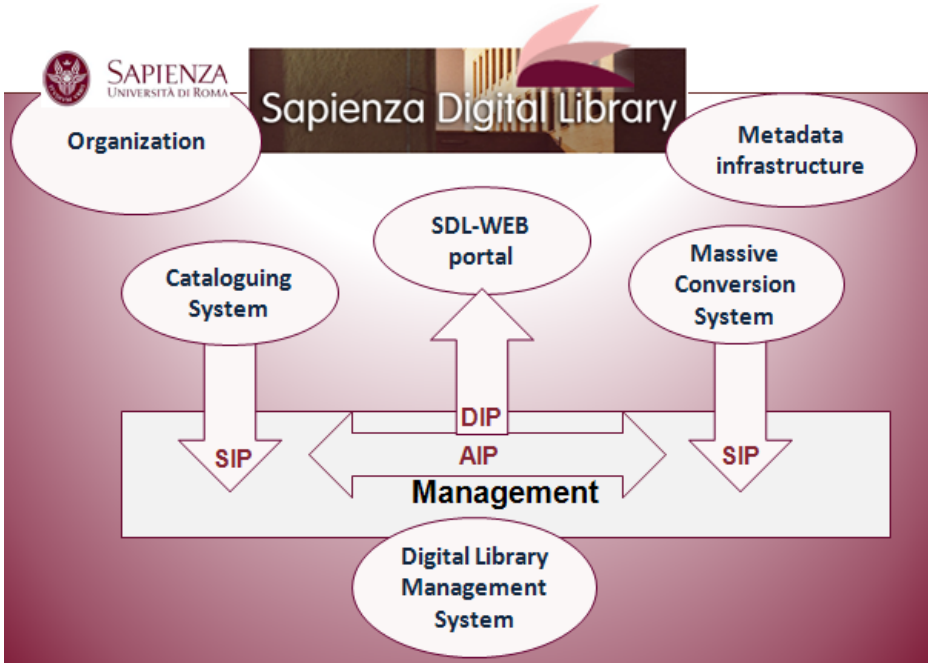


Fig. 1. Abstract overview of the SDL architecture components

4 Approaching the organizational complexity

The Sapienza University's is a complex Organization composed by 63 investigation departments, 56 libraries, 21 museums, 8 administration departments and some research center. We have conceptually considered the Sapienza's Organizations as Organizational units belonging to the Sapienza University.
In order to deal with the Organizational complexity of the Sapienza University, it was deemed essential to devise a metadata infrastructure, not only based on semantics world-wide known, but also with identification semantics aiming to identify unambiguously the Sapienza's Organizational units, involved in the work-flow production of the digital resources.

Furthermore, the long-term focus implies that the metadata infrastructure is able to record information referring to the real evolution of the Organizational units, that are involved in the management of the digital life-cyle of resources. The conception of an holistic approach referring to the Organizations' custody chain, recorded and expressed by the metadata infrastructure was based on the two reference model cited in the section 2 [4][1]. In addition the "Certification (TRAC): Criteria and Checklist"[3] that now is an ISO standard[7], focused on the repository's trustworthiness certification, proves that the first aspect in the checklist, influencing the trustworthiness of the digital repository, is the Organizational infrastructure. Consequently, the information about Organization, establishing an information system, has not to be neglected, but has to be curated and considered as relevant OAIS Preservation Description Information.
Considering the reference models, the long term aspect, and the complex organizational application context of Sapienza, the following requirements for designing the metadata infrastructure and the supporting identification semantics, were deemed essential:

- the unambiguous identification of the Sapienza's Organizations producing digital resources;
- the maintenance of the naming information history, connecting the evolution of the real Organizations with the digital management of the resources;
- the establishment of an identification hierarchy based on the concept of the Organizational Collection.

5 The Digital Library system and the metadata infrastructure

The digital resources managed by the SDL system constitute the digital representation of the Intellectual Entities[10], that are managed under different types of conditions (creation, holding, management etc.), by the Sapienza's Organizational units. The definition of Intellectual Entity, is borrowed from the PREMIS Data Model[10], which defines the intellectual entities as: "a set of content that is considered a single intellectual unit for purposes of management and description: for example, a particular book, map, photograph, or database. [..] An Intellectual Entity may have one or more digital representations." In the SDL system an Intellectual Entity is technically represented by a Digital Resource (DR), that can be considered as the digital embodiment of an intellectual item, and is equivalent to the OAIS IP [4].
By the implementation point of view a DR is physically composed by the set of objects files, that together represent the OAIS Content Information, and the set of metadata represent the OAIS Preservation Description Information.

5.1 The digital library standards adopted

The metadata infrastructure was conceived for supporting different kind of DRs. The DRs can be represented in different formats (still and moving images, texts,

sounds, cartographics, etc) and can be representing different kind of intellectual contents (multidisciplinary knowledge). In order to manage the materials' diversity and to deliver centralized digital library services, based on the metadata, we had considered metadata standards with a sufficient degree of granularity, as well as a sufficient level of semantic interoperability. The analysis of the standards adopted in the digital libraries' scenario had driven to the choice of a very well known standards combination:

- Metadata Objects Description Standard(MODS) which describes the intellectual contents and follows libraries semantics, derived by the MARC 21 semantics[7], the pillar standard of all libraries information systems;
- PREservation Metadata Implementation Strategies(PREMIS) for managing preservation metadata;
- Metadata Encoding and Transmission Standard(METS)[8] for wrapping together metadata belonging to the DR.

Mostly these standards, made up in combination, have covered the need of providing sufficient granularity of information for the intellectual content (MODS), sufficient granularity of information for the digital preservation management (PREMIS), and sufficient granularity and flexibility for supporting the need of managing an Organization infrastructure, using DRs variously structured (METS). Indeed, the encapsulating mechanism provided with METS has allowed not only to include other standard semantics, more relevant to specific aims (like for example Dublin Core (DC)[9] (more interoperable), or NISO Technical Metadata for Digital Still Images Standard MIX[10]), but also supporting the exchange of packages between the architectural components of the SDL infrastructure (see Sect.3).

5.2 Metadata infrastructure and the building blocks

The metadata infrastructure is coded in the adopted standard semantics and is organized on the DRs, that are the essential bricks, building the digital library. Both the massive conversion system, and the cataloguing system produce DRs, encoded in XML[11], and conforming to the metadata standards adopted by the project (see the following Section).
The DLMS ingests DRs produced by both two pre-ingestion systems, in order to start the management of their digital life-cycle[4].
The Figure 2 is a simplified representation of the SDL's DR. On the left is visible how the conceptual OAIS IP is generally divided into two parts: the metadata, and the content objects. On the right is represented how is physically composed

[7] MARC 21 Format for Bibliographic Data, `www.loc.gov/marc/bibliographic/`
[8] Metadata Encoding Transmission Standard, `www.loc.gov/standards/mets/`
[9] The Dublin Core Metadata Initiative, `dublincore.org/`
[10] NISO Technical Metadata for Digital Still Images Standard, `www.loc.gov/standards/mix/`
[11] Extensible Markup Language (XML), `http://www.w3.org/XML/`

inside of the system, as a set of different metadata semantics and a set of object files. Each box is labeled with the name of the related standard XML schema[12] name(see Sect. 5.1).

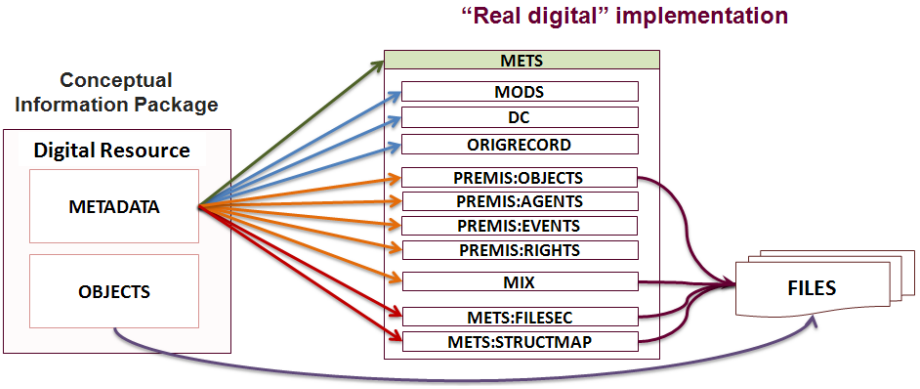


Fig. 2. Digital resource's structure

The descriptive metadata, pointed by the blue arrows, is coded into two descriptive standards. The MODS which reflects the granularity of MARC 21. The DC, which is commonly adopted in other contexts, not strictly related to the libraries world, is consequently considered more interoperable.
The inventory metadata, listing the files' names and locations, and the structural metadata, pointed by the red arrow, are coded in the two relevant METS sections. Both sections of metadata are connected together by METS, which is essentially used for conveying the whole structure of the DR in the XML format. All metadata blocks are unambiguously identified and referred to the Organizational context[8], related to the DRs production.
The system was publicly opened the 20th of December 2013, as Beta version 1.0[13], and is under testing by the communities.
The DLMS is actually providing access, and discovery services to the communities and has ingested more than 11.000 DRs distributed in 22 collections, belonging to 10 different Organizational units. By the end of the year more than 30 new Sapienza's Library will be incrementing the number of digital resources.

6 The identification semantics for Digital Resources managed by an Organization

The SDL DR's production starts from the main constraint of existence about the identifier of one of the Sapienza's Organizational units, which has the man-

[12] XML schema, `www.w3.org/XML/Schema`
[13] Sapienza Digital Library, `sdl.uniroma1.it`

agement responsibility of the DRs. Consequently, the *conditio sine qua non* for the existence of every single DR must be its identification by an identifier based on the Sapienza's Organizational units' identifier. This identifier abstractedly defines the concept of "Organizational collection", that gathers all DRs belonging, owned or managed by a Sapienza's Organizational unit. Consequently, all objects belonging to a DR are identified extending the Organizational collection identifier, which is the root of the identification.
The identifier is necessary for capturing information about the Organization context, which has some responsibility in the SDL DR production: scientific or technical responsibility, objects digitization, metadata editing or management responsibility. The long term focus of the digital library requires to deal with an ever-growing amount of DRs and the re-use in the long term of a DR could result difficult or inconsistent, if it is not possible to have agents of reference about its management.
The semantics adopted for the whole process of SDL's DRs production is based on an identification system that, first of all, aims to identify the Sapienza University ownership of the digital library service. In addition it identifies the Sapienza's Organizational unit, having the initial management responsibility of the resource's digital born under the Sapienza domain (selection or creation of the digital materials). The production method designed for building DRs allows to produce self-documenting IPs, where the documentation is based on the structured semantics, referring to the Organizational context.

6.1 The Organizational collection and the identification family

The Organizational collection in the conception of the SDL is the digital embodiment of the Organization's collecting actions, that consist of the digital production, preservation and fruition. The collected digital item is represented physically by the OAIS IP which is the DR in the SDL context.
The Organizational collection is the set of the whole digital production made, managed or owned by the Sapienza's Organizational unit that has the responsibility of the DRs created for the SDL. The abstract concept of the Organizational collection refers the contextual information about the Organization and set the basement of the identification semantics of the referred DRs.
By means of the Organizational collection identifier, we captured information about the organizational context where the DR was born, and produced for the ingestion in the SDL's DLMS. We have also leveraged on the identification information for relating other information, about context and provenance[4][6] related to the DRs.
This is the reason why the related Organizational collection's identifier is considered the first mandatory information, for submitting the resources to the system.
In order to respect the LTDP requirement, allowing the DRs re-use, we have considered essential to use identification semantics, already used by a national identification system, where the main organization Sapienza and its Organizational units are hierarchically represented.

Respecting the hierarchical structure of the University, the SDL identification system has adopted an identifiers' family derived and extended from the Italian National Bibliographic System[14] where the Sapienza University is identified by the identifier "RMS".
This is the main identifier, which associated with descendant identifiers, unambiguously identify the Organizational collection, and build relationships with other entities involved in the DRs management: objects, agents, events and rights[10].
The well-defined structure of the SDL identification system has allowed to enrich resources and the pertaining objects with contextual information about Sapienza organization.
In addition, the registration of the Sapienza University to the international identification MARC organization code [15], identified by "itrousr", and semantically mapped to the same level of the italian "RMS" identifier, allows to set the DRs context also at international level. Indeed, the replication of such code as mandatory administrative metadata in each SDL's DR, makes possible its connection to the Linked Open Data Cloud[16].
The open world "itrousr" identifier, exposed by the Library of Congress Linked Data Service(LCLOD)[17] in the Cultural Heritage Organization identification system as authoritative identifier, makes each DR, belonging to the local Sapienza domain, worldwide reachable through the exposed identifier "`http://id.loc.gov/vocabulary/organizations/itrousr`", and by virtue of the mapping between the local ("RMS") and global identifier("itrousr").

6.2 The Organization as the source of the identification system

The SDL identification system is structured on four layers, extended from the main layer, represented by the "RMS" identifier of the Sapienza Digital Library, and going down to the following hierarchical layers, that are also sampled in the Fig.3:

- the root identifier corresponding to the Organizational Collection (see subsect. 6.1), in the showed case, "RMSAR" identifies the Sapienza's Library of Architecture;
- the digital collection identifier, corresponding to the SDL aggregation level for managing DRs, which in many cases is directly identified by the Organizational collection itself. In the showed case, the Library of Architecture collects the digitized books from its holdings, directly collected as DRs of the Organizational collection "RMSAR". In addition the same library collects a special collection "RMSAR_SEVERATI" collecting images, donated by an Architecture's Faculty member;

[14] Anagrafe Biblioteche Italiane `http://anagrafe.iccu.sbn.it/opencms/opencms/`
[15] MARC code list for organizations `http://www.loc.gov/marc/organizations/org-search.php`
[16] Linked Open Data, `http://linkeddata.org`
[17] Library of Congress Linked Data Service, `id.loc.gov`

– the DR (Figure 1) identifier, in the figure the "RMSAR_00000025" is a digitized book of architecture, and "RMSAR_SEVERATI_00000001" is a photograph digitized and containing Brasilian buildings relevant for the architecture interest;
– the digital objects identifier, represented by the DR's identifier and the order number of the object, as example the book's page "RMSAR_00000025_0324".

The replication of the higher layer's identifiers over the identifiers of the lower layers, allows to reuse the single objects in other contexts, without ambiguity about the pertaining DR of the objects, and from the root identification layer, back to the responsible Organization. The multiple representing format (in the example jpg and tif) are managed by the system, using the reference of the digital object's identifier.

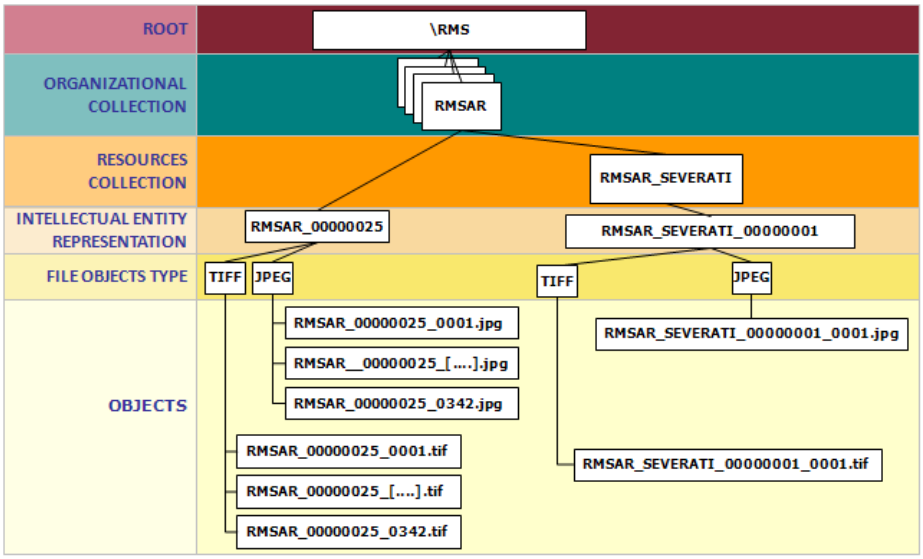


Fig. 3. SDL resource identification layers

7 Conclusions and future developments

The management of the identifiers based on semantics, derived by the Organizational collection conception, matches with two Ontologies, recommended recently by the W3C.
The Organization Ontology(Org-O), originally developed for use by data.gov.uk[18], represents the formal definition resulted from real implementations and uses. The

[18] Opening up government `http://data.gov.uk/`

core class in the ontology is the "Organization" class which represents "a collection of people organized together into a community or other social, commercial or political structure". The main class "Organization" of the ontology Org-O, semantically speaking, matches to the Sapienza University. While the Org-O subclass "OrganizationalUnit", matches with the Sapienza's Organizational units.
The matching conceptualization between the "OrganizationalUnits" class of the Org-O and Sapienza's Organizational units associated to the SDL's Organizational Collections, and unambigously identified, will drive the reasoning systems to retrieve information about DRs belonging to the pertaining "Organization" or "OrganizationalUnit".
The identification system based on semantics locally defined, but world wide processable by means of dereferenceable URI like the LCLOC identifier (see Sect.6.1), allows to make all belonging DRs reachable by URI through the Organization ontology support.
Coherent to this scenario is the ontology aimed to model the information about "entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness"[19], known as Provenance Ontology.
The Prov-O Ontology(Prov-O) [11] is provided with a Data model, that simply defines three core types of classes: Agent, Entity, and Activity and related relationships. Focusing on the topic of this paper we underline the fact that the Agent defined as main class in the PROV-O data model, can be connected to the Org-O's Organization concept by means of the Agent subclass "Organization". The Organization subclass is defined in the Prov-O as "An organization is a social or legal institution such as a company, society, etc.". Also in this case the matching of PROV-O definition with the SDL Organizational units, and its Organizational collection digital conceptualization, allows to connects classes of information and relationships with the information collected in SDL, where the identification semantics drive to the relevant values.
The recommendation by W3C of this two ontologies demonstrates the global interest, around the traceability of digital assets back to the Agents responsible for their management, harmonically with the SDL's Organizational collection conception, where the agents belong to the context information referred to the Organization.

References

1. Candela, L., Athanasopoulos, G., Castelli, D., Al., e.: The Digital Library Reference Model. Tech. rep., DL.org: Coordination Action on Digital Library Interoperability, Best Practices and Modelling Foundations (2011), http://bscw.research-infrastructures.eu/pub/bscw.cgi/d222816/D3.2bDigitalLibraryReferenceModel.pdf
2. Caplan, P., Kehoe, W., Pawletko, J.: Towards interoperable preservation repositories (tipr). In: Proceedings of the 2010 Roadmap for Digital Preservation Interoperability Framework Workshop. p. 16. ACM (2010)

[19] PROV-Overview, http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

3. CLR (Center for Research Libraries and RLG Programs): Trustworthy repositories audit and certification checklist (2007), http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf
4. Consultative Committee for Space Data: Reference Model for an Open Archival Information System (OAIS), Recommended Practice CCSDS 650.0-M-2 Magenta Book (2012), http://public.ccsds.org/publications/archive/652x0m1.pdf
5. Di Iorio, A., Schaerf, M., Bertazzo, M.: Establishing a digital library in wide-ranging university's context: The Sapienza Digital Library experience. In: Digital Libraries and Archives, 8th Italian Research Conference on Digital Libraries, IRCDL 2012, vol. 354 CCIS, pp. 172–183. Springer (2013), http://www.scopus.com/inward/record.url?eid=2-s2.0-84873865280\&partnerID=40\&md5=d8b5b1f12a673c347ec521d4a4e8b391
6. Di Iorio, A., Schaerf, M., Guercio, M., Ortolani, S., Bertazzo, M.: A digital infrastructure for trustworthiness The Sapienza Digital Library experience. In: Bridging Between Cultural Heritage Institutions, pp. 59–69. Springer (2014)
7. ISO (the International Organization for Standardization): Iso 16363:2012 - space data and information transfer systems – audit and certification of trustworthy digital repositories (2012), https://www.iso.org/obp/ui/#iso:std:iso:16363:en
8. Parsons, M.A., Godøy, Ø., LeDrew, E., De Bruin, T.F., Danis, B., Tomlinson, S., Carlson, D.: A conceptual framework for managing very diverse data for complex, interdisciplinary science. Journal of Information Science 37(6), 555–569 (2011)
9. Payette, S.: The state of technology for digital archiving. arXiv.org http://arxiv.org/pdf/1403.7748v1.pdf
10. PREMIS Editorial Committee: PREMIS Data Dictionary for Preservation Metadata version 2.2. (2012), www.loc.gov/standards/premis/v2/premis-2-2.pdf
11. W3C: Organizational Ontology (2013), http://www.w3.org/TR/vocab-org/

Persistent e-mail identification is viable!

Stefan Haun and Andreas Nürnberger

Data and Knowledge Engineering Group,
Faculty of Computer Science,
Otto-von-Guericke-University Magdeburg, Germany
http://www.dke.ovgu.de

Abstract. Persistent identification of entities in Personal Information Management (PIM) is necessary to enable stable, long-term references in archives and semantic applications. In the case of e-mails, the standard offers Message-IDs (MID), which are widely deployed. However, stores do not use the MID but rather rely on a path, which is likely to change, to refer to e-mails and thus do not offer a stable identification. We show that MIDs are viable to identify and retrieve e-mails from an IMAP store in real-world scenarios. The presented concept can be integrated into any store, but we also offer a software solution that serves as an additional layer above the store and allows real-time access over MID. We propose a validation method to prove that the concept is working and some applications that are enabled by e-mail identification are sketched.

1 Introduction

Sending e-mails has long replaced traditional letters, especially in the business and research context. Being digital, e-mails can be easily stored and accessed from different locations, building up large archives as part of the personal information managed by each user. Although a central element of communication, e-mails are still tied to special software, the Mail User Agent, instead of being integrated into the overall workflow. Part of the reason is the lack of means to reliably identify an e-mail message within an archive. While the Internet Message Format–the de-facto e-mail standard–provides a globally unique Message-ID (MID) and this identifier is present in each e-mail, it is not used for identification [11]. Instead, e-mails are referred by their folder and a running number (UID)–both most likely to change if the user decides to put the message elsewhere or other messages appear [8].

Using a stable and persistent identifier for e-mails enables novel applications: Most archives are only accessible in a read-only mode, either due to technical reasons, e.g. WORM and similar media, or because of law regulations or policies. Therefore references cannot be adapted and identifiers need to be stable in the first place. Semantic applications store outside references to e-mails. With today's stores, those references become stale if a message is moved. As a result, references to e-mails are either not available or must be enforced by a very tight integration, locking out other applications. A further benefit of stable identifiers

is the intrinsic cross-referencing given through the fact that identical messages have the same identifier, even across archives, i.e. an e-mail has the same MID at the sender and the receiver. When researching e-mail archives, for example in legal cases or historic research, this effect leads to higher efficiency.

We present a concept that enables referencing of e-mails by their Message-IDs. While this concept can be implemented into any software solution, we developed a prototype that works on top of IMAP-enabled stores, thus allowing additional functionality without changing running systems.

In the following, Related Work towards the topic is discussed. We then present the Message-ID Index, that adds MID-based references to an IMAP store, followed by a discussion of a validation method. Based on the availability of persistent identification, a set of enabled applications is sketched. The paper closes with a conclusion and the list of references.

2 Related Work

In [4] we argued towards persistent entity identification in Personal Information Management: It is necessary for recognition, dissemination and (external) cross-references to digital objects. *Uniform Resource Identifiers* (URIs) provide an established scheme for identification in the context of Internet communication and semantic technology [1].

Tools for alternative access to an IMAP store already exist. For example, [3] presents an IMAP plugin for *SquirrelRDF*[1] that allows to pose *SPARQL*[2] queries to an IMAP store. However, the proposed solution uses anonymous or generated node names for identifiers. While a *MessageID* attribute is provided, the paper clearly states that it is derived from the message number, which is even more volatile than the UID value and therefore should not be used.

The *Internet Message Format*, base for the e-mail format, is defined in RFC 5322 [11]. However, identification has been specified in a much earlier draft as RFC 724 [10]. RFC 2111 [6] specified a URI form of Message-IDs and is the base for the representation we chose to identify e-mails. The relevant parts of these specifications will be further elaborated in the next section. As the implementation resembles an offline IMAP store, many of the operations are described in RFC 4549 [7].

In [9] an analysis of stability and reliability of digital identifiers in *digital forensics*, including a survey on Message-IDs, is presented.

To the authors' knowledge current standard-conforming IMAP server implementations do not support an efficient query by Message-ID. The only known solution that uses an e-mail's Message-ID for identification is the *Gnowsis*[3] semantic desktop [12].

[1] http://notes.3kbo.com/squirrelrdf

[2] SPARQL is an acronym for "SPARQL Protocol And RDF Query Language". See http://www.w3.org/TR/rdf-sparql-query/ for further reading.

[3] http://www.semantic-web.at/de/gnowsis

3 Message-ID index

E-Mail Identification. Even without semantic applications and archiving, message identification is necessary for communication between a Mail User Agent (MUA) and an e-mail store. In the following we concentrate on the *Internet Message Access Protocol (IMAP)* as defined in RFC 3501 [2], which is supported by most server implementations and has become a wide-spread method for accessing e-mails from remote or distributed devices, e.g. from a desktop PC or smart phone. The IMAP standard defines the *Unique Identifier (UID) message attribute* as means of identification of a single message within the store. The UID is defined as an integer value that is unique within a specific IMAP folder. While this values is mean to be stable, the IMAP server may decide to re-organize the folder and thus change the UID for each message. Using folder and UID allows efficient storage and access during IMAP sessions, but is not viable for long-term identification.

A solution, however, is already embedded in each e-mail: The *Internet Message Format* defined in RFC 5322 [11], which is the base description for e-mails, contains a set of *Identification Fields* and, more specific, the `Message-ID` (MID). The MID is intended as a globally unique identifier embedded in each e-mail, which is currently used to generate threaded folder view, i.e. show the tree structure of messages within a single folder. There is a major downside to the Message-ID: The generation is left to the Mail User Agent. A malicious user can try to spoof an existing MID and mask other e-mails, if the MID is known. This attack is similar to other attacks on the message meta-data and providing a solution must be left to research on e-mail security. The form is roughly described in RFC 5322 as *localpart@domain*, were the domain should match the mail server's domain. The local part can either be a sequence number, a pseudo-random number or a hash of e-mail meta-data. Often a mixture is used in combination with the recipient address. However, efforts to create a recommendation for the Message-ID format never made it to the RFC catalog.[4]

Default operations of an IMAP store require fast access to folders and random access to messages within a folder, leading to a default hierarchy of message lists within a folder tree and the above identification scheme. While the MID is accessible from each single message, it does not help IMAP operations to make them available on a higher level, therefore this operation is not supported. There are IMAP implementations that use the Message-ID as file name for disk storage, but even then the folder tree, represented by directories, and the message lists apply.

To make the Message-ID quickly accessible, we propose an index on top of the IMAP structure that maps and updates Message-IDs to locations within the IMAP store. For an efficient solution in terms of runtime-efficiency, the index should be included in an IMAP server implementation. However, this would limit the use to a specific system and requires effort beyond the proof of concept, i.e. requires to adapt a software component that needs to be very reliable and can

[4] A draft can be found at http://tools.ietf.org/html/draft-ietf-usefor-message-id-01

cause severe data loss on misbehavior. It is, however, feasible to integrate the index into future implementations or create a specialized version that leverages features of a specific IMAP store.

Index Structure. The index is structured as follows:

$$
\begin{aligned}
&\langle\text{Message-ID}\rangle \quad \longmapsto \quad (\langle\text{Folder URI}\rangle, \langle\text{UID}\rangle)_{\text{unique}} \\
&\langle\text{Message-ID}\rangle \quad \longmapsto \quad (\langle\text{reference type}\rangle, \langle\text{Message-ID}\rangle) \\
&\qquad\qquad\qquad\quad \text{with} \quad \langle\text{reference type}\rangle \in (\texttt{Reference}, \texttt{In-Reply-To}) \\
&\qquad\qquad\qquad\quad \text{and} \quad \texttt{In-Reply-To} \rightarrow \texttt{Reference}
\end{aligned}
$$

Message-IDs are mapped to locations, which consist of a folder and the message's UID. The location pair is unique, i.e. there can be only one message at a specific location. However, since copying messages is allowed, the same e-mail can be found at several locations. While the resolution from a Message-ID to a message is unambiguous, the mapping to a location is not. The first line is sufficient to resolve MIDs, but we decided to index reference fields as well to allow quick searches for related e-mails. Those references are `In-Reply-To` for answers to a specific e-mail and `Reference` for a more generic reference, e.g. all e-mails from a discussion thread. The `In-Reply-To` field implies the `Reference` field and the reference is stored only once. Note that referenced e-mails are not necessarily available. For example, the user may have deleted the original e-mail before getting an answer, hence the `In-Reply-To` field points to an e-mail that does not exist locally. Still the identifier is valid and might be resolvable by a third-party user, e.g. the sender who kept the e-mail he answered.

Challenges. The main challenges are to create and update the index. Creation requires to crawl all available messages in order to extract their MID and references and add them to the index. While this process takes a while, it is only necessary on setting up the index. Subsequent runs can ensure consistency to avoid missed updates, but are by design not necessary. Keeping the index up to date is the larger problem. As the user moves messages around, a number of location entries change: First the moved message now is at a different location and the index entry for this message must be updated. The UID value should not change often, but if the server decides to re-organize the folder structure[5], all locations for this folder are invalid and must be retrieved again. For a responsive system, this re-indexing must be finished before the next Message-ID resolution.

The index implementation is very defensive against invalid entries. Therefore each location mapping is checked before it is returned to the caller. Checking is done by loading the message at the denoted location and comparing its Message-ID with the stored value. If there is no match, the entry is discarded as invalid. When no valid entries can be found, there is either no e-mail corresponding to the Message-ID in the store or the message is in a folder pending for re-indexing, in which case the resolution is stalled until the index is consistent

[5] The UIDVALIDITY attribute allows to detect when UID values have become invalid.

again. Regarding the latter case, there are two look-up contexts with different performance behavior: For resolution of a Message-ID to an e-mail, only one valid entry is needed, as all entries will point to the same content. However, for a list of all locations, the index must be consistent, otherwise the result may be incomplete. In a responsive user interface, this distinction should be made to avoid unnecessary long processing times.

Optimization. Having thorough checks on all mappings allows an optimization in the update process: When an e-mail is moved only the new location is stored. Since the same e-mail can be stored at different locations, a complete re-check of all locations would be necessary when a message is updated. This way, only the new entry must be added, which is a much faster operation. The other way around, uniqueness of locations allows to overwrite entries if the location mapping shows to a different Message-ID. These optimizations lead to faster indexing, but may result in an inconsistent index. Further performance tests are needed to decide whether these optimizations are necessary or if there is a faster solution.

Operations. Besides updates and consistency checks, which run automatically in the background, the index offers two main operations to a user: First, a Message-ID can be resolved either to its content, i.e. the complete message, or to the set of locations where the respective e-mail can be found. The main difference between those resolutions is that finding an e-mail requires only one valid location while finding all locations needs a completely consistent index and therefore has to wait until all update operations are finished. However, as the main purpose of the index is the resolution of Message-IDs to e-mail content, the faster query method will most likely be used. Second, the index allows to retrieve references to a specific e-mail, allowing to query preceding e-mails or related e-mails from an ongoing conversation. Note that this information come solely from the header fields within the message and are not inferred by further content analysis. If a Mail User Agent fails to set the respective fields, these references will not be available. However, apart from some Web-based mail systems, the data quality seems to be very good.

Architecture. Figure 1 shows the component diagram of the index. The Application has access to the underlying IMAP store for message content access and IMAP-related operations. Thus the default IMAP behavior is not hindered. To look up a Message-ID, the Application sends a Query to the Resolver, which in turn reads the mapping from an SQL DBMS. The result is checked against the IMAP store and, if valid, returned to the Application, otherwise invalidated in the DBMS. When the Resolver invalidates an entry, the Crawler is triggered, which in turn will read e-mails from a provided folder and update the respective mappings in the DBMS. The Observer is notified by the IMAP store if the message count in a folder changes, i.e. if messages are added, moved or removed, and triggers the Crawler to update the mappings. These notification can also

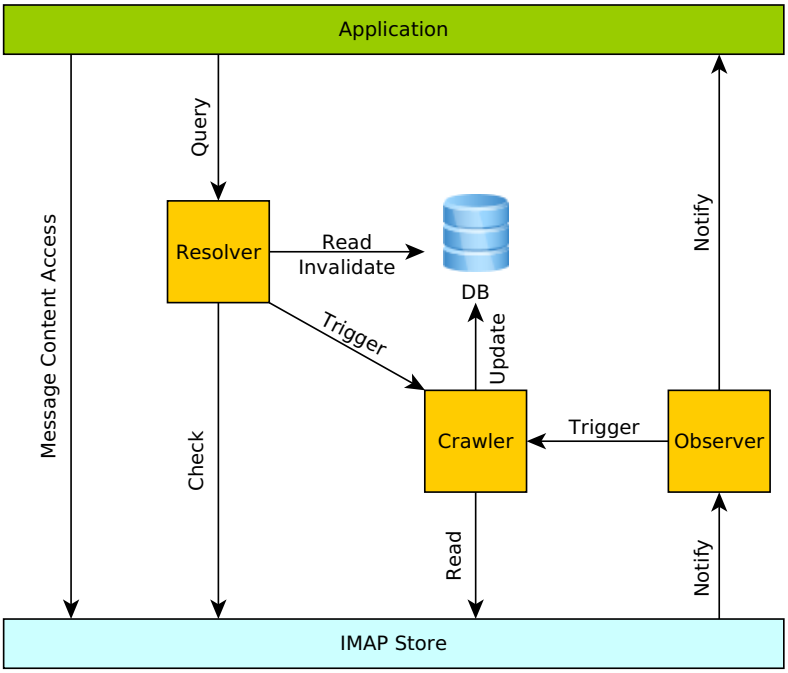


Fig. 1. Component diagram of the Message-ID index.

be received by the Application to update displays or trigger other application-dependent reactions.

Runtime Environment. We implemented the index using *Java*[6] and the *Java Mail API*[7] for generic IMAP access. For the IMAP store we are currently using the *Courier MTA*[8]. The index is stored in a *MySQL*[9] DBMS.

Our deployment environment is quite distributed, i.e. IMAP store, DBMS, index and application run on different machines in different networks. The first implementation could crawl 100.000 e-mails in about 2 hours. None of the machines showed substantial load, so most of the time goes into network communication. The current implementation uses parallelized crawling and bulk access to the database, so that with a rate of 50 e-mails per second the complete store can be crawled in half an hour.

[6] https://www.java.com/, the implementation uses the J2SE6 standard
[7] https://javamail.java.net/, Version 1.5.2
[8] http://www.courier-mta.org/imap/
[9] http://www.mysql.com/

4 Validation Methods

Due to time constraints, validation is not yet finished. However, we want to present our validation concept and first results.

The goal of validation is to show that with the index, resolution of Message-IDs to e-mail locations within an IMAP store is 1) faster than without an index and 2) fast enough for user interaction. Additionally, we measure the "usefulness" of the index at the time of access, i.e. the hit/miss ratio.

We expect that any of the supported operations is faster with the index than without. This follows from the rationale that a completely inconsistent or empty index would crawl the IMAP store, which is about the same operation as Message-ID lookup without an index. It is more interesting to see if the index is fast enough for user interaction. Studies have shown that a user becomes impatient after waiting two seconds for a response and annoyed after four seconds [13]. Since the look-up results from the index will most probably need further processing, even two seconds may be too much. However, we chose them as an upper limit: In an optimal operation, no query will take longer than two seconds. The *percentage of queries that take longer than N seconds* measures the performance, where $N = 2s$ can measure usability-critical performance. A better boundary for access times might be found from upcoming applications. First results show access times around 10 ms for a consistent index, which is coherent with the fact that only single SQL query and the retrieval of one e-mail header in an already open IMAP connection are necessary. For an inconsistent index, the access time is directly related to the time it takes to crawl the folder containing the message. Our current test system[10] achieves 50 e-mails per second. The folder structure in this system consists of smaller "work" folders with less than 100 e-mails and large "archive" folders with up to 15.000 e-mails. Crawling the larger folders takes several minutes, but only occurs when the UIDVALIDITY changes, which has not happened in 3 months of running the system. The smaller folders undergo much more fluctuation, but can be scanned within 2 seconds. The observation component even shortens these values: Changes to a single message are reported within one second with a rate of 20 messages per second for bulk operations.[11] For most of the time, the 2-seconds-limit for usable access was met.

The *readiness* of the index can by evaluated by the ratio between hits, i.e. found entries, and misses in terms of invalid entries or entries that where not available due to re-indexing. It is important not to evaluate if entries could not be found, as they may come from Message-IDs for e-mails that were never available in the store. To count as a *miss* the entry should have been available in a perfect index. We expect misses on two occasions: Either a message has been added, but is not yet available in the index or messages have been removed

[10] A quad-core Intel Atom platform. However, these values are only first estimates as the test setup is not yet fit for clean statistics.

[11] IMAP bulk operations appear when a large number of messages are marked as read, as supported folder-wise by many clients, or a large number of messages is moved to another folder.

from a folder, but re-indexing is still pending. The observer module is directly informed about e-mails that have been added to a folder, so new messages are available very quickly. However, it may take time to re-organize if the message numbers have changed. So in most cases not the moved, but the remaining messages are affected. We expect that in a real scenario this lessens the number of misses, as the focused message is always readily available. However, re-indexing is relevant if an application frequently requests all storage locations. Furthermore we measure the *time the index spends in an inconsistent state* with pending re-indexing. We observed only two scenarios where the index was not ready within one second after a change: 1) When emptying the trash, as each message is reported individually. 2) When marking a large number of messages read, which happens often for mailing list conversations. As observation and crawling are parallelized processes, other folders were not affected and changes in these folders were available within one second. Therefore we could not observe a significant amount of cache misses.

A trivial approach towards a test scenario is a random distribution of operations (add, move, delete e-mails) and accessed e-mails. However, this pattern is far away from typical use cases. After their receipt only a small fraction of e-mails will ever be accessed again. As a consequence, there is a rather small set of "hot" e-mails a user might ask for, while the larger part will never be accessed. We suspect that those relatively new e-mails are in a distinct set of folders, like the INBOX or folders based on current project, while the rest has been moved to archive folders where they will most likely stay, so that for every (re)move operation only a very small part of messages is affected. Therefore, a more realistic test-scenario will weight the access based on the age of an e-mail and on the folder. A third test scenario comes from user observation. Due to privacy reasons it is however very hard to get the respective data. Observing user behavior on e-mails relies on an index on the user's e-mail store and tracking of live usage data in a critical part of everyday communication. Therefore the test set will only be very small. While the last scenario is hard to acquire, we expect the best evaluation results since it matched actual access patterns best.

5 Outlook on Applications

In this section we present three applications that are enabled by fast Message-ID resolution.

Related E-Mails. The first application allows to view related messages in a *Mail User Agent (MUA)*. While a tree view within a folder is often available, it is not possible to display a series of messages across folder boundaries, since related messages cannot be found fast enough. With the index, however, related messages can be easily retrieved and displayed in a tree alongside a selected message. This allows the user to easily navigate through a set of messages, regardless of the folder they are stored in. Since older messages are often moved to archive folders, this application allows the user to see all related messages even if they have already been archived.

Misplaced E-Mails. Related e-mails are often stored in the same folder, e.g. based on a project or the communication partner. When a message is accidentally misplaced it is very hard to recover this message later on.[12] A quick access to message references, provided by the second part of the index, allows to check if there are any stray messages. While it cannot be assumed how a user organizes e-mails, the amount of e-mails references that cross folder boundaries can be used to determine if a message may be in the wrong folder and where it should be. An appropriate notification can inform the user about the potential mishap and offer a quick solution.

External References. A third application directly uses the fact that Message-ID references can be easily resolved: Using an add-on, the browser can be enabled to understand the MID URI scheme. When confronted with a respective URI, the message will be looked up and displayed in the browser or opened directly in the MUA. As a result it is now possible to send links to e-mail messages, e.g. via e-mail, instant messenger or embedded into a document, regardless where they are stored.

6 Conclusion

Although e-mails play an important role in modern communication, there is no applicable way of referencing them. We have shown that a generic IMAP store can be enhanced so that the already existing Message-ID field is viable for persistent, long-term identification and reference of e-mail messages in archives and semantic applications. We have also presented a set of applications that are enabled by fast e-mail identification via Message-ID.

The validation concept is ready and our next step will be a long-term validation of the index on realistic scenarios to proof its usefulness. As mentioned before, evaluation on real data is challenged by privacy issues. During further research we will try to build a test set based on the ENRON e-mail dataset [5] to diminish the issue of private test data. We were also able to find a number of people who are willing to run an analysis tool on their mailbox to acquire accumulated statistics about folder structures and message distribution over time.

For the software itself, further implementation will include more statistics for the validation process and a better recoverability on link failures to the IMAP server.

Afterwards the applications sketched in the outlook will be implemented.

References

1. Berners-Lee, T., Fielding, R., Masinter, L.: Uniform Resource Identifier (URI): Generic Syntax. RFC 3986 (January 2005), http://tools.ietf.org/html/rfc3986
2. Crispin, M.: Internet Message Access Protocol - version 4rev1. RFC 3501 (March 2003), http://tools.ietf.org/html/rfc3501

[12] The best way to hide a book in a library is to put it into another shelf.

3. Eynard, D., Recker, J., Sayers, C.: An IMAP plugin for SquirrelRDF. Tech. rep. (October 2007), http://www.hpl.hp.com/techreports/2007/HPL-2007-161.html
4. Haun, S., Nürnberger, A.: Towards persistent identification of resources in personal information management. In: Predoiu, L., Mitschick, A., Nürnberger, A., Risse, T., Ross, S. (eds.) SDA. CEUR Workshop Proceedings, vol. 1091, pp. 73–80. CEUR-WS.org (2013), http://dblp.uni-trier.de/db/conf/ercimdl/sda2013.html#HaunN13
5. Klimt, B., Yang, Y.: Introducing the enron corpus. In: CEAS (2004), http://dblp.uni-trier.de/db/conf/ceas/ceas2004.html#KlimtY04
6. Levinson, E.: Content-ID and Message-ID Uniform Resource Locators. RFC 2111 (March 1997), http://tools.ietf.org/html/rfc2111
7. Melnikov, A.: Synchronization Operations for Disconnected IMAP4 Clients. RFC 4549 (June 2006), http://tools.ietf.org/html/rfc4549
8. Melnikov, A., Newman, C.: IMAP URL Scheme. RFC 5092 (November 2007), http://tools.ietf.org/html/rfc5092
9. Pasupatheeswaran, S.: Email 'Message-IDs' helpful for forensic analysis? In: Proceedings of the 6th Australian Digital Forensics Conference. School of Computer and Information Science, Edith Cowan University, Perth, Western Australia (2008)
10. Pogran, K., Vittal, J., Crocker, D., Henderson, A.: Proposed Official Standard for the Format of ARPA Network Messages. RFC 724 (May 1977), http://tools.ietf.org/html/rfc724
11. Resnick, P.: Internet Message Format. RFC 5322 (October 2008), http://tools.ietf.org/html/rfc5322
12. Sauermann, L.: The Gnowsis Semantic Desktop for Information Integration. In: Proceedings of the IOA 2005 Workshop at the WM. Springer (2005), http://www.dfki.uni-kl.de/šauermann/papers/Sauermann2005a.pdf
13. Shneiderman, B.: Designing the user interface: strategies for effective human-computer interaction. Addison-Wesley, Reading (1998)

The Digital Online Museum
A new approach to experience virtual heritage

Tilman Deuschel[1,3] and Timm Heuss[1,2]and Bernhard Humm[1]

[1] Darmstadt University of Applied Sciences,
Darmstadt, Germany
{Tilman.Deuschel,Timm.Heuss,Bernhard.Humm}@h-da.de
[2] University of Plymouth
Plymouth, United Kingdom
Timm.Heuss@plymouth.ac.uk
[3] Cork Institute of Technology
Cork, Irland
Tilman.Deuschel@mycit.ie

Abstract. This paper describes a novel approach to satisfy the needs of museum's website visitors with a unique experience that cannot be reproduced in the museum itself. We aim at providing a continuous and lasting experience, without the emphasis of a single, final result - a process we call digital strolling. The view supports this process by displaying results as a path on which the user strolls. To enable the user to find new and unexpected inspiration, recommendations to related exhibits are proposed in different dimensions to vary the user's path. The common approach of image retrieval as the sole method to generate recommendations of related exhibits is not sufficient. Authored tagging is still the better but more costly solution. The proposed approach claims to fill the gap between current digital museums and the needs of the digital museums' visitors.

Keywords: Digital Museum, Digital Strolling, Semantic Search and Indexing, Semantic Tagging, Virtual Heritage

1 Introduction

In the federally funded project Mediaplatform, we are researching new and enhanced ways of searching and displaying the online collections of galleries, libraries, archives and museums (GLAMs) together with the German Städel Museum [23]. Within this context, we have built a standard information retrieval system, that supports the proposed browsing-based usage paradigm by hosting various kinds of media meta data, recommender logics and media stocks. The application, developed within this research has to meet high requirements in feasibility, usability and performance, as it will represent the Städel Museum [23] in their 200 year anniversary in 2015.

Modern museums, like the Städel Museum [23] can only exhibit nearly 2% of their art stock. Therefore the objective is not to attract more visitors to the

museum, but to share the existing art stock with the public in a digital museum. According to Weng, the main features of a digital museum is to archive, exhibit and educate in the same way as the physical museum pursues this objectives [29] and most existing research creates digital musems that are a representation of a physical museum [28, 17, 19, 18, 29, 27, 12]. The findings of Marty [14] indicate that these approaches are not sufficient. This paper proposes a new approach to close the gap of previous work and the described demand of museum website visitors.

2 Proposed Approach

2.1 Digital Strolling

Modern digital museums present their exhibits only in a traditional information filtering view [17, 19, 18], by a virtual representation of the museum [27], or by providing pipelined, curated paths [19, 12]. But the findings of Marty [14] indicate that users call for a different experience: "[...] online museum visitors are interested in having access to unique experiences that cannot be duplicated in museums".

Therefore we address this need by letting the user digitally stroll through the exhibits to discover unexpected results, similar to the sightseeing metaphor, established by Tezuka and Tanaka [26]. Whilst their work employs a split interface, separating a map where the user can digitally stroll from the media content that can be explored, we propose to not separate content from navigation to reinforce the strolling experience. This experience is based on physical strolling applications for museums like [2] and [20], but provides more degrees of freedom. Instead of just chaining content elements, we let the user decide which recommendations and paramedia to follow. The digital art stock of every museum consists of exhibit images and paramedia like paratexts. Paratexts are information around the exhibit that add extra meaning to it [1]. All paramedia and the exhibit's image as the central anchor build an information cluster. This cluster is strictly hierarchically structured, to increase the user's understanding, and thus the user's engagement with the system [3]. The information cluster also contains recommendations to similar exhibits with each other.

We call the experience digital strolling, when one or more information clusters are displayed as search results, related exhibits are recommended and all the clusters the user interacted with build a path, which is entirely available for the user. Similar to a browser history the information cluster string together as a path. In difference the user shall be able to modify the path by following the recommendations within the path or drag the clusters to compare one or more exhibits.

2.2 Ranking

Interactive storytelling systems [20, 2] demonstrate the necessity of providing ranked virtual heritage artefacts in order to present a coherent information

stream that can be discovered during the digital strolling. Automated processes like image retrieval are employed by e.g. Hong et al. [6] or the *Google Art Project* [5] but these methods still have issues in identifying contentual relationships. For example, Figure 1 shows that also unrelated images are presented as allegedly related images, because they are similar in shape and color composition, but not contentual. In order to support digital strolling, image retrieval methods are not enough. Instead, there are strong suggestions to involve a user into the curation and categorisation process of an online catalog, while an initial expert tagging is a suitable and expected starting point [22]. A classification harnessing the users' own words for describing an exhibit can furthermore create a contemporary understanding of the exhibitions, as advised by [1].

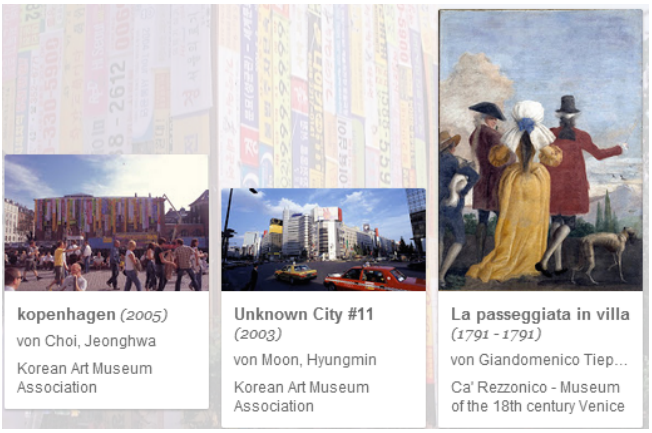


Fig. 1. The *Google Art Project's* Image-Retrieval-Based Recommendation.

Thus we propose a method taking several factors into consideration: domain experts maintain semantic tags for each collection item in multiple dimensions, which also act as starting points if no other data about the user exists. In further expansion stages, these defaults are augmented with additional tags gathered from automatically Named Entity Recognitions (NER), based on text meta data fields. These tags are biased by interpretations of user's implicit and explicit interactions with collection items.

On the other side, [22] concludes with the finding that "simply providing Web 2.0 interactions, such as tagging and commenting, is not enough". A combined method of several factors as mentioned above might therefore be the right approach.

3 Exhibit platform

The exhibit platform represents a novel user interface for rich media databases as found in museums. The user can query and browse through the results and

also get inspired by the system's recommendations of related works for each result in a map of media tiles, as shown in Figure 2. Based on modern web technologies (HTML5, CSS3 and JavaScript) the application runs on computers with large screens, as well as on mobile devices. Its responsive design makes it independent to different screen sizes. For very small screens, like smartphones, a different interaction concept is employed that suits the smartphone's interaction possibilities. Therefore a different view is employed by adaptive web design.

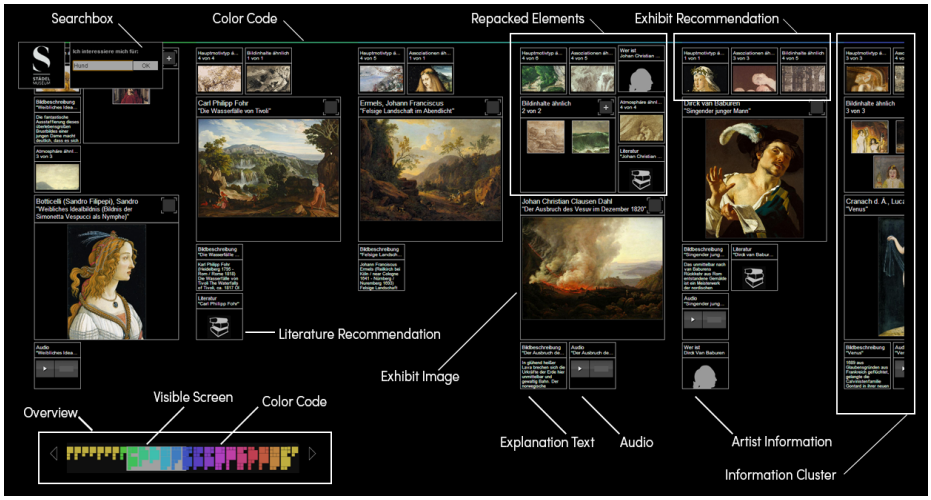


Fig. 2. Annotated tablet view of exhibit platform displaying a digital strolling path.

3.1 Use Case

It is intended to share virtual heritage artefacts from the art stock of the Städel Museum [23] with users, not to attract them for a physical museum visit. The user can stroll through, filter and discover images of exhibits, information about the exhibit in text, audio and video and information about the artists.

3.2 Design Principles

Role of Speed Several studies and best practises clearly indicate the importance of speed for online portals and search interfaces [15, 4, 21, 11]. At *Amazon*, for example, a 100 ms delay results in 1% of revenue drop [11]. A study at Bing showed that delays even under half a second have negative business impacts [21]. We therefore pay special attention to performance, which impacts the way we designed the architecture and the processes, especially in terms of semantic indexing as explained in the following chapter.

Responsivity and Adaptation Half of the German population owns a smartphone (effective date: February 2014) [24]. Hence it is necessary to support mobile devices and stationary devices at the same time to address most users. By employing Responsive Web Design (RWD) it is possible to provide a website that changes its layout to make use of different screen sizes [13]. Adaptive Web Design (AWD) is a different approach and provides two websites with a switch that reacts accordingly to the detected device class [25]. RWD is used when the interaction mechanisms stay the same on the different device classes and only the layout shall be adapted to the change of available screen real estate. This can be achieved with a packery system as described by [7]. AWD is used when the interaction mechanisms change on the different device classes. These techniques are poorly supported by the related works described. Recently the *Google Art Project* [5] introduced both RWD and AWD. We recommend the combination of RWD, for adapting to different screen sizes that are typical for high-performance devices, such as tablets and notebooks or desktops, and AWD, for switching to a website that employs different interaction mechanisms that are suited for the small screen.

Usability aspects Experts prefer common search portals like *Google* over curated, openly-ranked, reviewed, domain-specific search portals [9]. Krug recommends to design a self-evident GUI in order to minimise confusion potential [10]. He also advises to design for scanning web sites rather than reading them to increase the user's engagement. Clear visual hierarchy is a strong factor to increase the user's engagement [3].

3.3 Architectural Overview

The architecture has to deal with many different and sometimes opposing requirements. On one side, the system needs to cope with data from various sources with different formats and licenses. The Mediaplatform should be a place where different data can be used symbiotically, where the uses of single dataset can multiply. On the other side and as discussed previously, there are strong constraints regarding the application performance, feasibility and usability aspects. Reflecting those requirements, the Mediaplatform was designed as a client/server architecture shown in Figure 3.

The first and most important step from the various data sources towards a working application is a process we call Semantic Extraction Transformation Load (Semantic ETL). Its purpose is to read the multiplicity of various input data, transform the data formats into a unified entity model which is suitable for different kinds of media, add or apply authority files, and finally streamline the data bases to the common usage patterns of end users of the Mediaplatform.

As mentioned, performance is one of the most important criteria, thus this process results in a highly optimized *Apache Lucene* search index, capable of answering common information needs with a single query. Once the index as well as other pre-processed structures are created in an overnight batch job, both

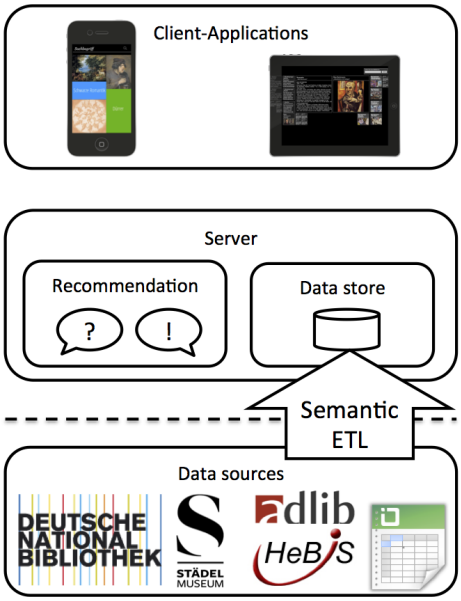


Fig. 3. Client-Server-Architecture of the Mediaplatform.

constitute the backend of all online-operations of a server. Beside queries that go directly to the prepared *Lucene* index, the server contains a recommendation component capable of making semantic suggestions based on authority files or semantic taggings.

3.4 Semantic Image Recommendation

We created a hybrid combination of several approaches introduced previously: On one side, we employ an automatic index and ranking mechanism of semantic tags, on the other side we try to foster the user's curiosity by hiding these criteria behind the actual exposition items and by offering them several possible paths through the collection. This includes automatically indexed, human-defined tags for each collection item in the six dimensions atmosphere, association, main motive, main motive type, emotion, and subject. Currently, in the first version of the application, those tags are defined exclusively by experts as a preliminary step and then remain hidden behind the scenes after being indexed in the Semantic ETL process - the GUI sorely works with images. This way, users will be offered six different graphical paths to continue their digital strolling through the museum in their own likeness, each showing up to four highly related images within the respective dimension as teasers, it is also possible to reveal all related images. Discovered images are no longer offered as recommendations for the further strolling path.

Behind the scenes, we determine how strong two items are related by the number of tags they share in one of the dimensions, e.g. the main motive. So in

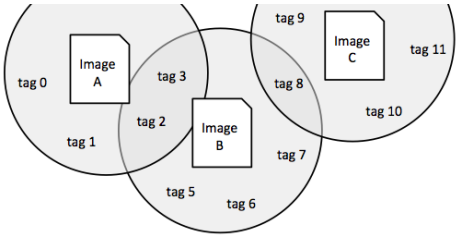


Fig. 4. Finding related images by the tag they have in common.

Figure 4, for example, for the given number of taggings in one given dimension of the images A, B and C, we try find the largest intersection, in this case $B \cap A$, followed by $B \cap C$. The most related image of B would therefore be Image A over Image C.

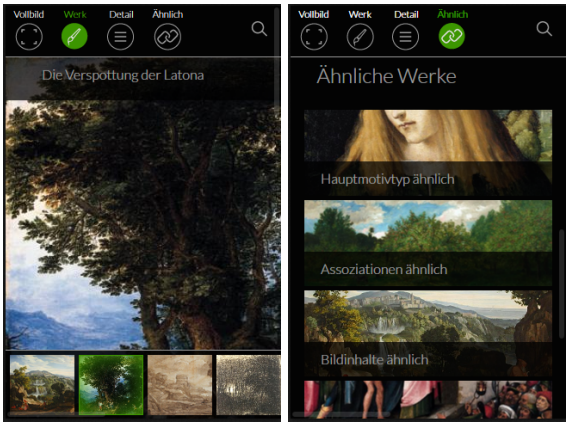
It is easy to translate this Set Theory problem into an Information Retrieval problem: for a given image, we basically formulate a query which OR-chains every tagging of the image in the given dimension. By the example of image B, a resulting query would be "tag 2" OR "tag 3" OR "tag 5" OR "tag 6" OR "tag 7" OR "tag 8". Thereby, we let the Information Retrieval framework - Apache Lucene fed with Boolean-Should-Queries [16] - determine the proper ranking for the relatedness of collection items. This results in a very fast and powerful semantic search feature as foundation for digital strolling, benefiting from automatic indexing as well as human curated guidance.

3.5 Client-Applications

The prototype welcomes the user with featured exhibits and topics where the user can select an element or send a search query. Then the results are displayed horizontally by animations to increase the overview for the user. Gesture support ensures flawless interaction on touch devices. Each result element is an information cluster, consisting of different media types like an image of the exhibit, information about the exhibit as text, video and audio. The elements within an information cluster are layouted according to a rectangle packer, similar to [7]. On different screens the elements are automatically repacked to fit to the available screen size in an optimal way. Each information cluster itself can be dragged to a new position enabling the user to compare the exhibits. Hence the fact, that the amount of information clusters can increase so that the available screen real estate is not sufficient to display them all at once, the user can scroll horizontally. An overview snippet provides the overview of all results like a radar, displaying the different packed information clusters in a minimised way.

For every exhibit the system provides suggestions of related exhibits in the six dimensions atmosphere, association, main motive, main motive type, emotion, and subject. Following a related exhibit inserts a new information cluster about this particular exhibit at the left of the predecessor.More recent results are displayed on the right, according to the European reading direction, thus the user

walks along a path. Color codes help to orientate the user which results belong to the same query and where has a related exhibit been explored. Too many results would decrease the performance on tablets and decrease the overview, therefore an oblivion mechanism has been integrated. User interaction updates the information cluster's date of creation. Once a dynamically computed threshold has been exceeded, the oldest information cluster is removed from the view. The web-based platform can be accessed on desktops, tablets and smartphones. For the latter an easified digital strolling mechanism is presented, that displays only one information cluster at a time, see Figure 5.



(a) Search results. (b) Similar results.

Fig. 5. Smartphone views for digital strolling.

4 Related Work

4.1 The digital museum as an extended physical museum

Existing research focused on extending the museum's physical showroom. Wang and Shen [28] refined 3D digitalisation techniques to present the exhibits as virtual three dimensional representation. Our approach does not focus on recreating the virtual museum as the physical museum. It introduces a possibility that cannot be reproduced in a physical museum. Closer to our approach is the *Google Art Project*. It presents digital collections of exhibits with the main focus on providing gigapixel images. The users are able to filter images based on properties like the exhibit material, year of production, artist name, etc. For each exhibit, the system offers three different related exhibits that share a similar motive or color composition. Unfortunately there is no information provided on which basis the images are related to each other. The results create the impression, that only image retrieval methods are employed. In contrast to the *Google*

Art Project an initial step in our approach is manual expert tagging. We want to extend this by image retrieval and user generated tags.

4.2 Strolling

Tezuka and Tanaka [26] show that the user satisfaction of a library information filtering system can be increased by employing a sightseeing metaphor. Literature is represented as places of interest on a map. In contrast to our approach we do not use a map, but the content itself to build a path. Both approaches have in common, that the user can stroll digitally to encounter unexpected information.

Similar to our digital strolling approach are on-site museum guides, that help the user navigation through collections of a physical musem. Rocchi et al. propose such a guide based on manually tagged images and texts of cultural heritage artefacts [20]. The visitor uses a location aware handheld device, which displays a life-like avatar, presenting paratexts that are related to the approached exhibits. The presented content is a seamless stream of information generated from text files, which are read aloud automatically, and images, which are presented in a cinematic slideshow. The text-image pair is joined with other text-image pairs based on their contentual context. Each text has a topic and rhetorical relations and each image is tagged concerning its content. Damiano et al.[2] follow a similar approach to support museum visitors without a predefined path. They also present a storytelling agent as an in situ museum guide exploiting mobile devices and position aware information retrieval for exhibits in front of the device. The information units are tagged semantically enabling a seamless playback. The users can stroll in the exhibit rooms freely because the information units can be combined in any constellation possible. These on-site museum guides demonstrate the necessity of semantically tagged content in order to be able to provide a digital strolling experience.

5 Conclusion

In this paper we have presented a novel approach of a web-based digital museum that supports a unique experience that cannot be duplicated in museums: the digital strolling. We have described a semantic search feature called digital strolling, that enables the user to freely discover virtual heritage artifacts. Thereby, a hybrid ranking feature takes several factors in multiple dimensions into consideration. Also we have described a prototype that establishes the proposed method. We believe the basic idea behind digital strolling can be ported on more and different use cases as well as on other underlying technologies. We can conclude that it is possible to fill the gap that exists between current digital museums and the needs of the digital museums' visitors.

6 Future Work

First informal usability test with students and museum experts show, that this approach is promissing: The individuals instantly interacted in a strolling man-

ner and reported that the recommendations raised their interest to follow them further. In order to be able to draw reliable conclusions about the proposed approach we will perform a standardised usability evaluation including usability walkthroughs and usability questionnaires according to the international norm for designing interactive dialogues [8].

According to Christensen [1] it is necessary to establish the user's participation in an exhibition platform, because they can create a contemporary understanding of the exhibitions, beyond the scientific understanding. We will research how user input can be considered in an automated tagging mechanism, e.g. images that are often compared obviously share a property that is relevant for the users.

Currently, the process of tagging collection items is of manual kind. However, we see great potential in supporting humans in this task with automatic NER. We want to explore a carefully selected middle course between strictly predefined curated paths on the one side and following a community-driven approach on the other.

References

1. J. R. Christensen. Four steps in the history of museum technologies and visitors' digital participation. *MedieKultur*, 50:7–29, 2011.
2. R. Damiano, C. Gena, V. Lombardo, F. Nunnari, and A. Pizzo. A stroll with carletto: adaptation in drama-based tours with virtual characters. *User Modeling and User-Adapted Interaction*, 18(5):417–453, 2008.
3. S. Djamasbi, M. Siegel, and T. Tullis. Visual hierarchy and viewing behavior: an eye tracking study. In J. A. Jacko, editor, *HCI'11 Proceedings of the 14th international conference on Human-computer interaction: design and development approaches*, pages 331–340. Springer-Verlag, Berlin and Heidelberg, 2011.
4. Forrester Consulting. ecommerce web site performance today, 2009. http://www.damcogroup.com/white-papers/ecommerce_website_perf_wp.pdf.
5. Google. Google Cultural Institute, 2014. http://www.google.com/culturalinstitute.
6. J.-S. Hong, H.-Y. Chen, and J. Hsiang. A digital museum of taiwanese butterflies. In P. J. Nürnberg, D. L. Hicks, and R. Furuta, editors, *the fifth ACM conference*, pages 260–261.
7. E. Huang and Korf, Richard, E. Optimal rectangle packing: an absolute placement approach. *Journal of Artificial Intelligence Research*, 46(1):47–87, 2013.
8. ISO. Din en iso 9241-110 Ergonomie der Mensch-System-Interaktion - Teil 110: Grundsätze der Dialoggestaltung (iso 9241-110:2006); Deutsche Fassung EN ISO 9241-110:2006.
9. M. Kemman, M. Kleppe, and S. Scagliola. Just Google It - Digital Research Practices of Humanities Scholars. *arXiv:1309.2434 [cs]*, Sept. 2013.
10. S. Krug. *Don't make me think! A common sense approach to Web usability.* New Riders, 3rd edition edition, 2013.
11. G. Linden. Make Data Useful, 2006. http://www.gduchamp.com/media/Stanford DataMining.2006-11-28.pdf.
12. Louvre. Louvre visitor trails, 2014. http://www.louvre.fr/en/routes.
13. E. Marcotte. Responsive web design, 2010. http://alistapart.com/article/responsive-web-design.

14. P. F. Marty. Museum websites and museum visitors: digital museum resources and their use. *Museum Management and Curatorship*, 23(1):81–99, 2008.
15. M. Mayer. In search of ... a better, faster stronger web, 2009. http://assets.en.oreilly.com/1/event/29/Keynote%20Presentation%202.pdf.
16. M. McCandless, E. Hatcher, and O. Gospodnetic. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., Greenwich, CT, USA, 2010.
17. Metropolitan Museum of Art. The metropolitan museum of art - home, 2014. http://metmuseum.org/.
18. Museum of Modern Art. Moma — museum of modern art, 2014. http://www.moma.org/.
19. Powerhouse. Powerhouse museum — science + design — sydney australia, 2014. http://www.powerhousemuseum.com/.
20. C. Rocchi, O. Stock, M. Zancanaro, M. Kruppa, and A. Krüger. The museum visit: generating seamless personalized presentations on multiple devices. In J. Vanderdonckt, N. J. Nunes, and C. Rich, editors, *IUI '04 Proceedings of the 9th international conference on Intelligent user interfaces*, volume 9, pages 316–318. ACM, New York and NY and USA, 2004.
21. E. Schurman and J. Brutlag. The User and Business Impact of Server Delays, Additional Bytes, and HTTP Chunking in Web Search Presentation, 2009. http://cdn.oreillystatic.com/en/assets/1/event/29/The%20User%20and%20Business%20Impact%20of%20Server%20Delays,%20Additional%20Bytes,%20and%20HTTP%20Chunking%20in%20Web%20Search%20Presentation.pptx.
22. R. Srinivasan, R. Boast, J. Furner, and K. M. Becvar. Digital Museums and Diverse Cultural Knowledges: Moving Past the Traditional Catalog. *The Information Society*, 25(4):265–278, July 2009.
23. Städel Museum. Homepage, 2014. http://www.staedelmuseum.de/sm/.
24. Statista. Anzahl der Smartphone-Nutzer in Deutschland bis 2014 — Statistik, 2014. http://de.statista.com/statistik/daten/studie/198959/umfrage/anzahl-der-smartphonenutzer-in-deutschland-seit-2010/.
25. SYZYGY. Responsive vs Adaptive, 2014. http://www.syzygy.de/studien/responsive-vs-adaptive.
26. T. Tezuka and K. Tanaka. Traveling in digital archive world: Sightseeing metaphor framework for enhancing user experiences in digital libraries. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, E. A. Fox, E. J. Neuhold, P. Premsmit, and V. Wuwongse, editors, *Digital Libraries: Implementing Strategies and Sharing Experiences*, volume 3815 of *Lecture Notes in Computer Science*, pages 23–32. Springer Berlin Heidelberg, Berlin and Heidelberg, 2005.
27. The State Hermitage Museum. The state hermitage museum, 2014. http://hermitagemuseum.org/html_En/index.html.
28. N. Wang and X. Shen. The research on interactive exhibition technology of digital museum resources. In *2013 IEEE International Conference on Green Computing and Communications (GreenCom) and IEEE Internet of Things(iThings) and IEEE Cyber, Physical and Social Computing(CPSCom)*, pages 2067–2070.
29. T. Weng. The 19th century official paris salon exhibition digital museum. *WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICATIONS*, 12(6):1903–1912, 2009.

Relations for Reusing (R4R) in A Shared Context:
An Exploration on Research Publications and Cultural Objects

Andrea Wei-Ching Huang and Tyng-Ruey Chuang
Institute of Information Science, Academia Sinica, Taipei, Taiwan.
{andreahg, trc}@iis.sinica.edu.tw

Abstract: Will the rich domain knowledge from research publications and the implicit cross-domain metadata of cultural objects be compliant with each other? A contextual framework is proposed as dynamic and relational in supporting three different contexts: *Reusing*, *Publication* and *Curation*, which are individually constructed but overlapped with major conceptual elements. A *Relations for Reusing (R4R)* ontology has been devised for modeling these overlapping conceptual components (*Article*, *Data*, *Code*, *Provence*, *and License*) for interlinking research outputs and cultural heritage data. In particular, packaging and citation relations are key to build up interpretations for dynamic contexts. Examples are provided for illustrating how the linking mechanism can be constructed and represented as a result to reveal the data linked in different contexts.

Keywords: citation, context, cultural heritage, curation, ontology, packaging, publication, R4R, research data, reuse, sharing

1. Introduction

A digital object *Y* curated in a digital museum, is a cultural object *Y* with metadata descriptions. This cultural object *Y* reused by an academic article is not a cultural object but a science object *Z* that can be viewed under different context perspectives. By a definition of Zimmermann et. al., "*when the contexts of two entities overlap and part of the context information become similar and shared*," a shared context emerges [1].

Embedded information has been well preserved and curated in research data repositories and in Libraries, Archives and Museums (LAM) databases, but has not been explored for their potentials in enriching each other's contexts. For instance, cultural objects are mostly preserved with metadata information, but part of the data may come from the outputs of research projects. As for research data, the interpretation of domain knowledge is professionally established from scholarly publications which are comprehend by articles' textual descriptions, or by supportive evidences like associated publications (i.e. data and code), and these supportive evidence may come from cultural objects curated in LAM collections. Thus, is there a shared context between these two domains that can serve for a common understanding? And, how can a shared context

4th International Workshop on Semantic Digital Archives (SDA 2014).

between these two help us enrich contextual information and make our data better? In practice, will linking data from scholarly publications to metadata-rich LAM collections foster contextualizing research outputs? Will linking data from LAM collections to research publications increase the reuse and the remix of cultural heritage for a broad range of disciplines? And, in particular, what kinds of relations exist, or need to be established for a shared context? Finally, how these relations can be represented?

In this study, we hope to contribute to open a new dialogue among researchers from across different communities who share a common interest in understanding the potential of data sharing and reusing accross different domains. In the meantime, three more recent developments provide the potential of relating data in a wide range of contexts: (1) An increasing development on data publication and citation principles which is participated vividly by research communities like CODATA[1] , Research Data Alliance (RDA)[2] and FORCE11[3]. At the same time, the opportunity that open science movement presents for research reproducibility is taken from joint publications of articles, datasets and software codes. (2) The choice of linked data approach for data publication in research domains such as the VIVO project [2] and Linked Science and Education [3]; in cultural heritage data, efforts like LODLAM community[4] and the Europeana project [4], or in specific library catalog cases in LIBRIS [5], Library of Congress [6], WorldCat Work of OCLC[5] are examples in which this trend is well-justified. (3) As an overlapping of data publication and citation, open science as well as linked data developments, cases like publishing semantic enriched articles [7], source code Linked Data repository [8], and the emerging code citation mechanism[6] are such examples just to name a few.

However, citations need context [9], linked data is not enough only for research data [10], and the lack of theory and "object-rich but resource-poor" problems are identified in cultural heritage domains [11]. Therefore, above mentioned developments with these problems have motivated us to the design of a contextual framework to disclose context by a systematic approach in the next section.

2. A Contextual Framework for a Shared Context

For modeling and representing contextual linking, we follow the operational definition of [1] for determining the design space of context models. The five essential contexts are time, location, individuality, activity and relations. And in specific to model the activity, we further adopt Courtright's theoretical concept of actors-in-context which combines a relational view on activities of users, information systems and information

[1] http://www.codata.org/task-groups/data-citation-standards-and-practices
[2] https://rd-alliance.org/
[3] https://www.force11.org/
[4] Linked Open Data in Libraries, Archives, and Museums (LODLAM): http://lodlam.net/
[5] http://www.oclc.org/data.en.html
[6] https://github.com/blog/1840-improving-github-for-science

existence that context not only shapes action but is also shaped by it [12]. Our framework consists three major parts: (1) three contexts relate actors' levels with associated activities as *Reusing, Publication* and *Curation*[7]. (2) a Representation-Preservation-Interpretation setting is established. (3) Nine contextual elements are derived and extended from a contextual study on cultural heritage objects, and are further adjusted to accommodate particular settings. Table 1 provides a summary of this contextual framework, and the following offers theoretical backgrounds in details.

Table 1: A Contextual Framework for relating Reusing, Publication and Curation Contexts

activity \ setting	Representation	Preservation	Interpretation
Reusing *(User Level)*	**Application** (VII) *(ex. Reusing Cases)*	**Authorization** (VIII) *(ex. Policy/Licence)*	**Utilization** (IX) *(ex. Citation or Packaging Relation)*
Publication *(Author Level)*	**Identification** (I) *(ex. DOI/URI/URL)*	**Physicalness** (II) *(ex. Article, Data, Code)*	**Intangibleness** (III) *(ex. Domain Vocabulary & Citation)*
Curation *(Curator Level)*	**Classification** (IV) *(ex. RRObject/RRPolicy)*	**Authentication** (V) *(ex. Metadata/Provenance)*	**Ontological Relations** (VI) *(ex. Domain-independent Ontology /Relations ontologies/ R4R ontology)*

(1) Three dynamic activity contexts: *Reusing, Publication* and *Curation.*

From session one, we realize the importance of modeling publication and reusing contexts. However, Contextualizing only for these two activities is not enough since this framework is also to assist system designers, developers and curators for their practices. Thus a third Curation level is added for two more reasons: (1) Zimmermann et.al [1] defines activity context as a context which decides to its current needs and covers current and future activities. In other words, curation activity not only determines current needs of curators but also future activities like publication or reusing. Similarly, the publication activity serves publication-now and reusing-in-the-future purposes. (2) As [12] indicates that technology has a dual role in context, technology variations depend on other contextual elements while at the same time technologies influence information practices. In other words, a shared context between *Reusing* and *Publication* emerges as a technical dimension for the *Curation*. In short, three activity levels are situated in a multiple, overlapping, and dynamic context because *Publication* involves both publication and curation activities, and *Reusing* involves reusing, publication and curation, while *Curation* cannot exist without considerations of two other activity contexts.

(2) A perspective setting: *Representation-Preservation-Interpretation*.

In considering theoretical issues for a contextual framework, a Representation-Preservation-Interpretation setting is established from Charles Sanders Peirce (1839–1914)'s triadic sign theory: {Representation, Object, Interpretant} that a sign constituents three basic parts with a relation that a something, Representation, brings its Interpretant sign determined or created by it, into the same sort of correspondence with its Object, as that

[7] Three activity contexts are italics with the first word capitalized.

the something (Representation) stands to the Object [13]. Here, we define a contextual setting as a sign with the triadic relation [13]:

- The **Representation** is a representation of the activity context setting itself, and is the form that the setting takes. For instance, in *Reusing*, the Representation is the application cases employed to determine a resource to be used by oneself or others.
- The **Object** is the entity to which the context setting points, refers or applies. In this study, it is the specific preservation object that the authors, users, and curators refer to. The original "Object" has been adjusted to the object preservation for "Preservation" to describe associated activities.
- The **Interpretant** of a contextual setting is the Interpretation that is made of the setting. In this study, the interpretation is taken from the view of [1] on Relations Context that context information captures the relations an entity has recognized to the others.

The triadic sign theory has been empirically applied as an analytical framework for dynamic and complex composition such as for social tagging [14] and semantic web [15]. Furthermore, according to Tim Berners-Lee's own words, the Semantic Web is "a fervent desire to implement some ideas of Charles S. Peirce"[8]. Thus, we use this triadic relation that has also influenced Resources Description Framework (RDF) data model (Subject-Predicate-Object) to some degrees, as a basis to construct the context model as a triadic setting: Representation-Preservation-Interpretation. In addition, [12] argues that contextual elements must be explicitly linked to particular information practices, and the variability must be distinguished among actors and contexts. Thus, contextual elements need to be constructed within the *Representation-Preservation-Interpretation* setting and three dynamic activity contexts: *Reusing*, *Publication* and *Curation*. Next, we will move to disclose what contextual elements are constructed.

(3) Nine contextual elements: eight dimensions about context and its role are suggested by Beaudoin as technical, utilization, physical, intangible, curatorial, authentication, authorization, and intellectual [16]. The eight dimensions were generated for digital preservation of cultural heritage. For more context needs in this study, we adjust and extend technical, curatorial and intellectual dimensions to identification, application, classification and ontological relations. Table 1 is summarizes this framework. Details of these nine contextual elements associated with specific contexts and settings are introduced by using cases to illustrate how they can be applied in session 4[9]. Thus, we brief here four new contextual elements that are different from Beaudoin's work.

(I) Identification is a representation for disclosing the Intangibleness of the physical objects. In this framework, it is a publication-level representation for disclosing the existence of article, data, or code that can be identified for publication. It is restricted

[8] http://www.w3.org/DesignIssues/CG.html

[9] See more possible scenarios for different contents http://guava.iis.sinica.edu.tw/r4r/examples/possible_scenarios_for_different_contexts

by the *Curation*, and can be potentially utilized for the *Reusing*. For instance, when publishing linked data, it requires using URIs as names for things, the URIs are curated in restrict rules of the curation activity, and can be potential utilized for *Reusing***.**

(II) Application is a specific result or application cases like remixing or reusing, a representation for determining the Utilization of the presence of Authorization objects like digital policy or license that concerns the needs of users for *Reusing*.

(III) Classification is a classifying representation brings relational interpretations for Authentication elements (ex. metadata or provenance). It is a curatorial-level representation since it is the main task for curators to curate metadata about datasets. And metadata is interpreted by domain ontologies in the *Publication*, but interpreted by domain-independent ontologies in the *Curation*. For instance, the catalogue metadata of European Union Open Data is available as linked data[10], and uses the Data Catalog Vocabulary (DCAT) [11] to classify seven basic classes for catalogue metadata[12].

(IV) Ontological Relations is an interpretation for Classification that represents authentication elements such as metadata or provenance at the curatorial-level. Since contexts are changeable, we extend Beaudoin's intellectual dimension [16] and focus on the construction of a fundamental relationships for dynamic contexts and a domain-independent ontology formation. For instance, the Fedora relationship ontology[13] is used to model partial and provenance relations that can be shared across in its Fedora Ontology. Similarly, R4R ontology is designed for such functions.

To sum up, in *Publication*, an Identification name (ex. URI) is published and brings the interpretation by the network linkages of Intangibleness (ex. a domain vocabulary or citation), which determined or created by it, into the same sort of relation to the Physicalness (ex. data), as that in which the Identification stands to the Physicalness. Similarly, the rules are applied for *Reusing*: Application-Authorization-Utilization as well as for *Curation*: the Classification-Authentication-Ontological Relations. In practice, this framework is a conceptual tool to help us establish relations if we want to use the shared context for modeling *Reusing* and *Publication*. Since these two contexts share *Curation*, according to [1] we should start to establish relations between these two by examining what major preservation objects can be found in the *Curation* context.

3. Relations for Reusing (R4R) Ontology

For a light-weight design purpose, R4R consists 15 terms only: 7 classes and 7 properties plus one exceptional property *Cites*. Figure 1 illustrates the conceptual model of

[10] http://open-data.europa.eu/en/linked-data
[11] http://www.w3.org/TR/vocab-dcat/
[12] Catalog, Catalog record, Dataset, Distribution, Concept scheme, Concept, and Organization/Person
[13] http://www.fedora.info/definitions/1/0/fedora-relsext-ontology.rdfs

the R4R, and a full specification can be accessed online[14]. In the following, we will brief the major structure, and discuss our modeling decisions. Two crucial components as individual class concepts are identified in this model, namely, Reusing Related Object (RRObject) and Reusing Related Policy (RRPolicy). RRObject distinguishes R4R's basic components of described targets, creating the unique identification of the related objects, from RRPolicy being packaged for more specific combinations of provenance and license. The primary consideration for designing R4R is that it should on the one hand being capable of describing the combination of RRObject and RRPolicy, while on the other hand still allowing to just represent RRObject alone without packaging the RRPolicy. This is a decision made from reasons:

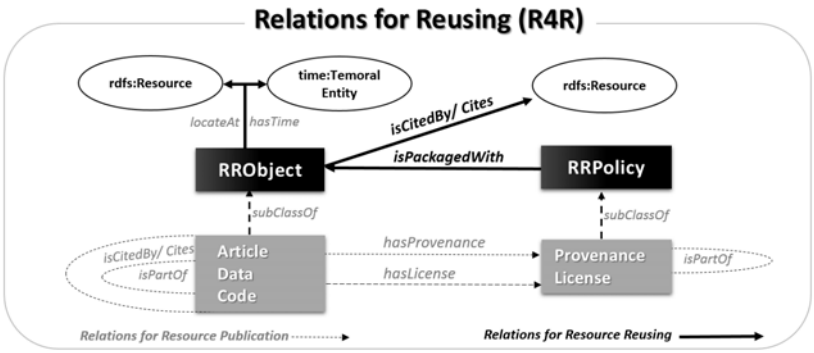


Figure 1: Relations for Reusing (R4R) Conceptual Model.

(1)Provenance and license concerns are not fully taken and implemented in existing practices, or have been curated as metadata in local curation that are not accessible or downloadable. Thus we use *hasProvenance* and *hasLicense* for relating local curation or for sharing publications. For *Reusing*, the context transitions occur, and according to [1], context attributes will change from one context entering another, thus Provenance or License, or both can be packaged with RRObject for reusing purposes. For such using of the relation, *isPackagedWith*, RRObject (article/data/code) and RRPolicy (provenance/license) are reachable and accessible for changing the original *Publication* and *Curation* contexts to a shift of the *Reusing* context.

(2) *isPartOf* and *isCitedBy/Cites* like *hasProvenance* and *hasLicense* that can relate internal relations within subclasses of RRObject (article/data/code). Meanwhile, these two relations can also be used for describing external relations. *isPartOf* describes partial relationships with temporal and spatial constraints. A *isPartOf B* only if A and B share the same time and location. This design helps to clarify relations of collections and items since temporal and spatial attributes of collections constrain item-level attributes. It also helps semantic publishing that one partial paragraph, session, chapter or even a sentence can be represented as an RRObject for article enrichments.

[14] http://guava.iis.sinica.edu.tw/r4r

(3) *isCitedBy* is distinguished *from Cites* for temporal constraints. Normally, when A *isCitedBy* B implies the publication time of A occurs before B. However, it is also possible that A and B are mutual-cited at the same time. For instance, two articles publishing in the same journal and citing each other are common research practices.

(4) Relations between Data and Code in current practice are sometimes *isPartOf*, sometimes *isCitedBy*, since dataset and code are quite often published together as Data. When Data and Code share the same temporal and spatial attributes, and data modelers wish to distinguish the two, it can be described as Code *isPartOf* Data.

(5) Citation is one of the most important traces to link contextual information from the original to many interpretations of the reused. In *Publication*, authors create their works by citing references as evidences/interpretations. In *Reusing*, afore mentioned publications become other's evidences/interpretations. As indicated by [1], when the activity (like citation) predominantly determines the relevance of context elements in specific situations, citation thus becomes one of our major interpretations for relations.

(6)Packaging relation in R4R is a relation between RRObject and RRPolicy. *isPackagedWith* is utilized only when *Reusing* occurs. It is a design specific to differentiate interpretations of metadata/provenance and license in different contexts. In *Publication*, metadata/provenance are curated for local preservation, and may be interpreted by domain vocabularies as a reflection of the author. In *Reusing*, metadata/provenance, and license are necessary components for Authorization and Authentication, therefore RRPolicy needs to be packaged to be able to be reused or remixed.

In sum, the design concept of R4R components are more toward modularity, in which components can be separated and recombined in different contexts, at different time. This is important because R4R wish to describe the future relations which will grow and evolve like future citations, provenance changed, or license policy changed.

So far we have dealt only with the contextual framework and the R4R ontology that reveal how context shared or changed can be modeled through establishing and explor-

Figure 2: A Data-Paper like publication in digitalarchives.tw

ing relations. But how a shared context between different domains like research publications and LAM collections help us enrich contextual information and make our data better? In the fowling, we will use R4R and different contexts to represent an example of interlinked data between research publications and a cultural object curated in LAM.

4. A Use Case from the Digital Archives Taiwan

Digital Archives Taiwan (digitalarchives.tw) consists collections of five million digitized cultural objects contributed by the largest memory institutions in Taiwan, and spanning various domains (history, art, biodiversity, geology, geography, ethnology, anthropology, etc.). The collection of Digital Archives Taiwan curated both in item and collection levels is indexed and catalogued through the Union Catalog (catalog.digitalarchives.tw) for data aggregation, representation, and citation. Figures 2 shows one item[15] that is published as a form which is similar to "data papers" (dataset descriptions for scientific research) or "nanopublications" (small units of publishable information with unique identifiers) [16]. Each item page constitutes: (1) The collection object and its basic information (Scientific Names and Vernacular Name); (2) Link to the original database; (3) Metadata Description; (4) Contact Information for Licensing; (5) Citation Information (bibliography and the unique URL). In addition, this item has an archive record ID, S010384, and it will be discussed in following sessions several times, thus we use daT(S010384) as a substitute name for this collection item[17].

The daT(S010384) has the Union Catalog metadata which uses Dublin Core for curation schema. The item also has a citation spec[18] and the license information is expressed by a contact information. The following shows how we use R4R in Turtle syntax to model this cultural object being curated and published in the Union Catalog. For *Curation*, daT(S010384) is being classified as RRObject (Classification) using R4R

```
@prefix dc:   <http://purl.org/dc/elements/1.1/> .
@prefix time: <http://www.w3.org/2006/time#> .
@prefix r4r:  <http://guava.iis.sinica.edu.tw/r4r#> .
@prefix :     <http://www.example.com/data#> .

:daT_S010384
    a r4r:data, r4r:RRObject ;
    r4r:locateAt :URI_S010384 ;
    r4r:hasTime  :t3 ;
    r4r:isPartOf :daT_Collection ;
    r4r:isCitedBy <http://www.plosone.org/article/#> .

:daT_Collection
    a r4r:data, dc:Collection ;
    dc:publisher "Digital Archives Taiwan";
    dc:provenance:daT_Metadata   .

:t3
    a time:Instant ;
    time:inXSDDateTime "2012-01-01" .
```

[15] http://catalog.digitalarchives.tw/item/00/61/e8/e2.html
[16] http://nanopub.org/wordpress/
[17] All figures presented in this paper are published with high-resolution gif files in the reference [25].
[18] http://digitalarchive-taiwan.blogspot.tw/2012/02/blog-post.html

ontology (Interpretation) to relate its metadata description (Authentication). For *Publication*, daT(S010384) is published using an R4R Identification that brings the Interpretation of Dublin Core and citation relations to it (Intangibleness). For a Shared Context, the relation is established by modeling daT(S010384) as subclass of RRObject (in *Curation* level) to be r4r:Data (Physicalness in Publication level), and using *hasTime* and *locateAt* to relate the Representation of two contexts, and prepare for the possible future *Reusing* emerging context.

A simple *Reusing* is presented by a citation relation. The daT(S010384) has been cited in a science articles' material and method session[19]. For a simple citation modeling, we can add this citation in local metadata using *isCitedBy* relation. The science paper may be benefited from this citation since the daT(S010384) is also curated under a catalog structure of domain knowledge interpretation from the international scientific standard of the biological classification: Domain/Kingdom/Phylum/Class/Order/..., as well as a hierarchy which includes the project information about the source organization and project details[20]. For a complex *Reusing*, these rich domain knowledge can be packaged for more application uses. For instance, we assume there is a digital plant atlas of natural museum in Europe, called PA. In their plant atlas, lacking of digital collection in Asia is one of major problems. PA finds that a plant specimen collection in Digital Archives Taiwan is proper for their uses. The first problem PA will encounter is the authorization of each digital item. The second problem is that they have to validate each item's collection and digital process for data quality. The third problem is that even each item in Digital Archives Taiwan is well documented and accessible through hyperlinks to original data repositories, PA does not want to manually click through all the links. Thus, if a machine readable and executable license and provenance are provided, not only PA but any other users can easily select, reuse or remix this digital collection. Taking daT(S010384) for example, the item can be modeled by provenance information using PROV-O ontology[21]. An example of this is described in [25].

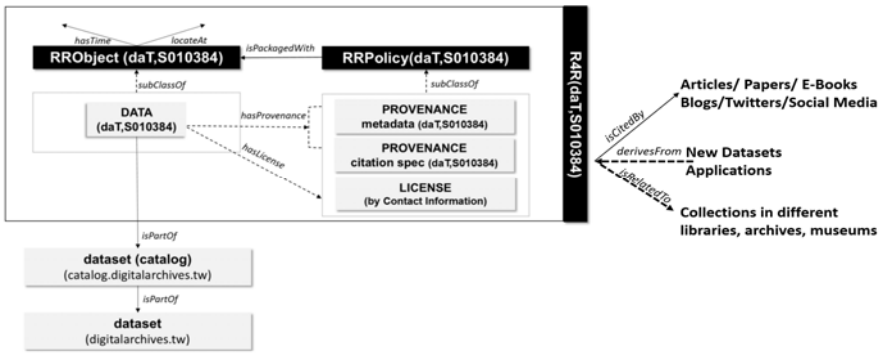


Figure 3. R4R for modeling one collection: daT(S010384)

[19]http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0077626#pone-0077626-g001

[20] http://guava.iis.sinica.edu.tw/r4r/examples/the_story_of_dat_s010384

[21] http://www.w3.org/TR/prov-o/

In short, when provenance or license is not ready to be packaged or not for releasing openly, we can use RRObject individually by publishing their unique identifications embedded with domain knowledge or citation interpretations through *hasProvenance* and *isCitedBy* to relate provenance information at the metadata level, and citation relations between article, data and code internally or externally. Once the RRObject is packaged with RRPolicy as R4R(daT,S010384), it is ready for other resources to connect and reuse by policy-aware tools for license like Semantic Clipboard [17], and by capturing provenance through ontology use like PROV-O at multiple layers [18]. It can also be easily used and relate to many forms of resources and from different domains. It can also be related to similar collections of other libraries, archives and museums; reused and recreated by other works. Or it can be embed in the package format of digital publishing like EPUB for E-books (see Figure 3).

5. Related Works

Although context modelling has been discussed in Artificial Intelligence literatures, the use of mathematical theory and logical formalization is beyond the scope of this paper. Instead, relations modeling that tries to classify linking structures in an attempt to make complicated relationships easier for semantic representation is most related to our work. For instance, the Fedora Relationship Ontology has been developed for representing object-to-object relationships in the Fedora architecture for complex object modelling [19]. And another useful example of relation representations supporting domain concepts interlinked by logical constrains is provided by the case of OBO Relation Ontology[22] in biomedical and life science. This ontology later influences the design of the Artifact Relationship Ontology (ARO) that has been designed specifically for comparing museum objects [20].

In addition, the Literature Object Re-use and Exchange (LORE) relationship ontology, a simplified version of IFLA FRBR is presented in [21] to facilitate reuse and exchange LAM collections for research purpose. Relations like authorship relations (i.e. creators, agents, or organizations), object attribute relations (metadata descriptions), or preservation and derivation relations are major concerns for LORE, and that results in more than one hundred relations are defined. Although many relation concepts of LORE are similar to R4R, it is taken from a bibliographic perspective. LORE uses its own definitions to represent similar and provenance information, while R4R recommend users to reuse SKOS[23], which can reference other concepts using a variety of semantic relationships, as well as PROV-O in the afore mentioned example. Most importantly, modeling compound and complex objects as employed in Fedora, Research Objects [10], and LORE alike is not the aim of R4R that takes the Shared Context for a design space, and aims to meet data publication, citation, and reusing for Open

[22] http://www.obofoundry.org/ro/
[23] http://www.w3.org/TR/2009/REC-skos-reference-20090818/

Science that needs to distinguish reusing, publication and curation for different contextual constructs. Table 2 is a summary of above mentioned relation ontologies, a full view of comparison can be accessed in [25].

Table 2: A Comparison of five relation ontologies

	Fedora	OBO/RO	LORE	ARO	R4R
Time	2005	2005	2009	2013	2014
Domain	independent	Life Science	Research	LAM	independent.
Concept		OBO Foundry /Other Biomed..	9 Classes from IFLA FRBR.	OAO+ Greek Vase Ontology	7 Classes
Relation	21 relations: (10 reverse)	13 relations : with logical definitions	133 relations: 63 reverse rel. +7 individual rel..	16 relations: classified by 5 levels	8 relations: (7 + 1 exceptional)
Location	---	V	V	---	V
Partial	V	V	V	V	V
Similar	V		V	V	SKOS
Provenance	V	V	V	V Open Annotation Ontology (OAO)	V PROV-O
Citation	---	---	V	---	V
Bundle	---	---	V	---	V
License	---	---	---	---	V
Compare	---	V	---	V	---
Definition	---	V	---	---	---

6. Conclusion

As responding to recent developments (Session 1) that have challenged research data, archival and cultural heritage communities for a contextual framework to support a dynamic and shared context environment, we have proposed a framework (Session 2), and to the establishment of an ontology, Relations for Reusing (R4R), that can facilitate the representation of contextual links between resources in diverse contexts (Section 3). In section 4, we use R4R for representing different contexts that can enhance semantic relationships of research publications and cultural objects when both are contextually linked. Section 5, related works are discussed and presented with a comparison on five existing relation ontologies that distinguishes the R4R from previous works.

The advantage of designing a new conceptual model to describe relations in a shared context is to ensure articles, datasets, software codes, provenance and license information can be treated as first-class contextual objects. At the same time, the module-like design of RRObject and RRPolicy can be practiced in isolation, and the unifying representation of their relations is semantically enough but not so structurally heavy-weighted that curators or researchers find it difficult to apply.

In sum, the daT(S010384) is a digital object with rich metadata descriptions being curated in *Curation* context. It is published as a cultural object Y, with unique identification, and being cited as a science object Z, interpreted by the citation relation for more professional interpretations. At the same time, the citing research can be benefited from the implicit information embedded in the institution's cataloging vocabularies for more

domain knowledge. Through the exploration of the Shared Context and R4R representation, the daT(S010384) now is capable to move from its traditional role and to "act as a citation of active knowledge" indicated in [22]. Creating knowledge out of interlinked data [23] is thus one step forward by packaging provenance and license for a policy-aware *Reusing* context. As a result, when data sharing needs not to remove the data's initial context but embedded in a shared context, the difficulty to interpret the reused data [24] may be expected positively through the use of the contextual framework and R4R ontology proposed in this study.

REFERENCES

1. Zimmermann, Andreas, Andreas Lorenz, and Reinhard Oppermann. An operational definition of context. *Modeling and using context* (2007): 558-571.
2. Krafft, Dean B., et al. Vivo: Enabling national networking of scientists. *Proceedings of the Web Science Conference*. Vol. 2010. 2010.
3. Keßler, Carsten, Mathieu d'Aquin, and Stefan Dietze. Linked Data for science and education. *Semantic Web* 4.1 (2013): 1-2.
4. Haslhofer, Bernhard, and Antoine Isaac. data. europeana. eu: The europeana linked open data pilot. *International Conference on Dublin Core and Metadata Applications*. 2011.
5. Malmsten, Martin. Making a library catalogue part of the semantic web. *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications* (2008): 146-152.
6. Ford, Kevin. LC Classification as linked data. *Italian Journal of Library and information science*, 4.1 (2013): 161.
7. Shotton, David. Semantic publishing: the coming revolution in scientific journal publishing. *Learned Publishing* 22.2 (2009): 85-94.
8. Keivanloo, Iman, et al. Towards sharing source code facts using linked data. *Proceedings of the 3rd International Workshop on Search-Driven Development: Users, Infrastructure, Tools, and Evaluation*. ACM, 2011.
9. Wendl, Michael C. H-index: however ranked, citations need context. *Nature* 449.7161 (2007): 403-403.
10. Bechhofer, Sean, et al. Why linked data is not enough for scientists. *Future Generation Computer Systems* 29.2 (2013): 599-611.
11. Skinner, Julia. Metadata in Archival and Cultural Heritage Settings: A Review of the Literature. *Journal of Library Metadata* 14.1 (2014): 52-68.
12. Courtright, Christina. Context in information behavior research. *Annual review of information science and technology* 41.1 (2007): 273-306.
13. Peirce, Charles Sanders. "Elements of Logic", Chapter 2: Division of Signs. In: C. Hartshorne and P. Weiss (eds), *Collected Papers of Charles Sanders Peirce (2)* (Thoemmes Press, Bristol, 1998) :134–272
14. Huang, Andrea Wei-Ching, and Tyng-Ruey Chuang. Social tagging, online communication, and Peircean semiotics: a conceptual framework. *Journal of Information Science* 35.3 (2009): 340-357.
15. Legg, Catherine. Peirce, meaning, and the Semantic Web. *Semiotica* 2013.193 (2013): 119-143.
16. Beaudoin, Joan E. Context and its role in the digital preservation of cultural objects. *D-Lib Magazine* 18.11 (2012): 1.
17. Seneviratne, Oshani, Lalana Kagal, and Tim Berners-Lee. Policy-Aware Content Reuse on the Web. *The Semantic Web - ISWC 2009* (2009): 553-568.
18. Carata, Lucian, et al. A primer on provenance. *Communications of the ACM* 57.5 (2014): 52-60.
19. Lagoze, Carl, et al. Fedora: an architecture for complex objects and their relationships. *International Journal on Digital Libraries* 6.2 (2006): 124-138.
20. Yu, Chih-Hao, and Jane Hunter. Documenting and sharing comparative analyses of 3D digital museum artifacts through semantic web annotations. *Journal on Computing and Cultural Heritage (JOCCH)* 6.4 (2013): 18.
21. Gerber, Anna, and Jane Hunter. Authoring, editing and visualizing compound objects for literary scholarship. *Journal of Digital Information* 11.1 (2010).
22. Srinivasan, Ramesh, et al. Digital museums and diverse cultural knowledges: Moving past the traditional catalog. *The Information Society* 25.4 (2009): 265-278.
23. Auer, Sören, and Jens Lehmann. Creating knowledge out of interlinked data. *Semantic Web* 1.1 (2010): 97-104.
24. Borgman, Christine L. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63.6 (2012): 1059-1078.
25. Associated data publication can be accessed at http://guava.iis.sinica.edu.tw/r4r/examples