



*Empowering users: an active role  
for user communities*

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

**Parallel sessions II**  
***Sustainable policies for digital culture preservation***

## **Maria GUERCIO**

### **Introduction to the session: conceptual framework and chain of custody for sustaining the digital preservation\***

\* This introduction is partially based on the keynote contribution presented at the conference *Perspectives on Metadata*, held in Vienna, 12-13 November 2009,  
<https://fedora.phaidra.univie.ac.at/fedora/get/o:45908/bdef:Asset/view>.

#### **A premise**

This introduction will be dedicated to present a common perspective on digital preservation by assuming that basic requirements for its success have conceptual and organizational nature, as increasingly recognized by the literature and the research outputs in the field. The metadata for preservation, the early adoption of adequate formats, controlled methods and good technical standards for acquiring digital resources play their role for ensuring the sustainability of the function, but they need to be included within a comprehensive and convincing intellectual framework and well state responsibilities. If the specific applications and related tests are not included within a systematic and robust theoretical infrastructure, the fragmentation is not avoidable and the risks for failure increase. This is why we have to put the accent on the relevance of the main goal and principles of the entire system (the defense of its trustworthiness and credibility) and its roots (the conceptual framework) and on the correct identification of responsibilities and procedural rules (the custodial environment as a chain of custody and its certification), both required for developing new products and implementing the existing solutions.

This introduction will start from two assumptions:

1. first of all, the challenges still open, specifically for handling the creation and preservation of digital resources depends on the recognition of their dynamic nature and the related need for handling as part of continuing and ongoing processes: the digital world offers a rich series of tools for the identification and capture of metadata and information on the basis of their position and encoding: they can appear as attributes of the resource itself, i.e. in the face of the digital object, as logical and physical components of its form, they can play as external elements (i.e. in a database system), but they also can act as implicit information within the procedural, technological or juridical contexts and they have to be captured and, even more, understood and maintained;
2. a pragmatic effort is required but it must be strongly rooted on consistent theory and principles specifically if we want to play with advanced technologies): it must be able to combine the best models for interdisciplinary approach, to avoid a useless overloading of detailed but not always useful information and to take into account in the application the promising outputs of the most recent research projects (PLANETS, CASPAR, INTERPARES, PREMIS just to mention those already known for their successful achievements and presented and discussed in this conference.

InterPARES is here considered as a conceptual framework thanks to principles, policies and procedures tested in many case studies and based on a consistent dictionary. The OAIS standard is recognized as a reference model for information architecture but also – specifically in the CASPAR project - as an implementation system. The guiding principle of CASPAR has been the application of the OAIS Reference Model to research, develop and integrate advanced components to be used in a wide range of preservation activities and to create a specific framework as a software platform for preservation that enables the building of services and applications that can be adapted to multiple areas, specifically to cultural, scientific and performing arts domains (that is dynamic sectors which require very complex and really evolving solutions).

CASPAR and PLANETS conceptual models have included multiple relevant results achieved in the field of preservation in the course of the last decade research efforts: the principles of InterPARES itself, the OAIS general framework, the checklist for auditing digital repositories developed in the TRAC report (Trusted Repository Audit Checklist) and in the RAC recommendations (Repository Audit and Certification), the PREMIS schema developed as metadata for digital preservation, the ISO standard CIDOC (Conceptual Reference Model) for developing ontologies and mapping metadata schemas with semantic functionality. The motivation was the creation of digital repositories and the development of framework and services for preservation based on an integrated approach to be applied to differentiated and complex archival and information systems.

The contributions presented in this session have made constant reference to these results in the specific effort for developing concrete domain-centric solutions.

The definition (and the agreement on the role) of a conceptual framework for ensuring both the consistency and the efficiency of the digital repositories requirements and of the preservation action in terms of policy,

procedures and responsibilities is a key basic issue, a condition to transform into an interrelated approach the individual solutions based on metadata identification and extraction or on the development of persistent identifiers criteria as it will be illustrated and discussed further in the course of this session.

The solidity of this analysis and chiefly the consistency of its implementation need some general statements. Specifically we could/would agree at least on the fact that the handling of digital assets as reliable, accurate and authentic heritage implies the clarification of the principle of trustworthiness.

If we look at the applications developed at national level, in most cases we could see continuing and exacting attempt for integration of principles and tools as outcome of research projects and standards development. But the fragmentation is difficult to overpass and it is even more complex to build a organic scenario.

### **The conceptual framework and the principle of trustworthiness for digital preservation**

The information and record preservation is increasingly based on concept of trust, specifically if the environment becomes digital.

First of all, it is suitable to share the definition of this term and clarify the connection between the concept of trust and the nature and quality of the digital heritage to be preserved, because the questions related to the metadata collection but also those concerning the responsibilities and the technological and organizational contexts for preservation are involved in this analysis and cannot be used conveniently and efficiently without this clarification.

In the dictionary (Merriam-Webster, s.v.) trust is identified as “a charge or duty imposed in faith or confidence or as a condition of some relationship”, a sort of “glue which binds that relationship together”<sup>1</sup>, whose ingredients have to be identified and described for effectiveness of the custody.

The custody can play successfully its role if all the elements and activities involved in this function can imply or presume a trustful handling and accomplishment.

According to the recent CCSDS guidelines, still published as draft (Recommended practice: Requirements for bodies providing audit and certification of trusted digital repositories, <http://wiki.digitalrepositoryauditandcertification.org/bin/view/Main/ReqsForAuditors>) the trust is at the basis of the certification process and at the centre of the whole process for providing solidity and efficiency to the curation action in the digital world. It involves a large community:

“to give confidence to all parties that a management system fulfils specified requirements. The value of certification is the degree of public confidence and trust that is established by an impartial and competent assessment by a third-party. Parties that have an interest in certification include, but are not limited to

- the clients of the certification bodies,
- the customers of the organizations whose management systems are certified,
- governmental authorities,
- non-governmental organizations, and
- consumers and other members of the public”.

It requires the identification of reference principles able to inspire confidence. This kind of principles includes (according to the CCSDS report):

- “impartiality,
- competence,
- responsibility,
- openness,
- confidentiality, and
- responsiveness to complaints”.

Each single attribute should be evaluated and transformed into procedures, rules, tools and metadata collection in a way to provide frames and contents for the evaluation of requirements and the recognition of the quality of digital repositories and their management and preservation systems.

Specifically, a more detailed exam of the core definitions could be of help for investigating the efficient use of metadata finalized to

- foster the credibility of the repository as trustworthy custodian on the basis of its capacity of securing integrity and authenticity of their digital contents through a standardized accumulation of descriptive and management information,
- control the cost of descriptive function “by using a simple [and standardized] encoding scheme and by ingesting metadata on transfer from public sector institutions”,
- enlarge the range of interrelations by “exchanging finding aid metadata with metadata harvesters from all kinds of communities”.

<sup>1</sup> See Jennifer Borland, *Trusting Archivists*, in “Archivi & Computer”, 2009, 1, pp. 95-106.

We do not have here time for this analysis, but it is important to recognize, within this perspective, the risk of fragmentation in the collection of all these information elements<sup>2</sup> and the low capacity of the present schemas and standards to document comparatively processes and describe them with an holistic and dynamic approach, the only one capable of dealing with the continuing evolution of the technological complexity. Of course, this last aspect, the most crucial for preserving the digital resources, requires the design of the digital preservation work as a chain of custody based not only on content identification, description and protection but also and with an increasing emphasis on the requirements for certifying institutional dedicated repositories, common policies and well defined and documented responsibilities.

### **The chain of custody: requirements, policy, responsibilities**

“The enduring trustworthiness of our documentary heritage is becoming a central responsibility of its designated custodian”<sup>3</sup>, as neutral third party on the basis that “it has no reason to alter the records and no interest in allowing others to do so, and must have the knowledge necessary to implement procedures that ensure the integrity and accuracy of the records”<sup>4</sup>. This assumption is today at the centre of a common effort made by the professionals involved in digital documents and in digital forensics, all of them persuaded that the core concepts concern the creation of a multilayer approach able to verify the integrity and authenticity of the resources at various levels of analysis:

- on the basis of the elements on the face/in the form of the resource and its attributes and metadata,
- from the circumstances of its maintenance and preservation: “an unbroken chain of responsible and legitimate custody is considered an insurance of integrity until proof to the contrary”<sup>5</sup>,
- from the integrity of essential metadata related to the resources handling and preservation as a further requirement for attestation of integrity and authenticity (individuals/offices involved, indication of annotations, of technical changes, of presence or removal [and the related time] of digital signature and other digital seals, the time of transfer to a trusted custodian, the time of planned deletion, the existence and location of duplicates outside the system,
- as inference on the basis of the trustworthiness of the record/document/information system in which the records/documents/information exist.

As Luciana Duranti has recently clearly expressed, “the authenticity...is a removable responsibility, as it shifts from the creator’s trusted ...keeper, who needs to guarantee it for as long as the record is in its custody, to the trusted custodian, who guarantees it for as long as the record exists”<sup>6</sup>.

If the framework and some basic principles seem today accepted and constitute the basis for the future implementation, some relevant details stay undetermined.

### **What is still missing**

1. consistent and accepted terminology and definitions used across domains and requested to be well understood beyond the professional communities involved in digital curation environment with specific reference to the fact that:

- definitions related to the attributes of preservation are not clearly expressed and present dangerous ambiguities<sup>7</sup>,
- new terms or the revision of traditional expressions (i.e. significant properties<sup>8</sup>) can produce dangerous misunderstanding;
- OAI glossary has still inconsistencies even if the standard is a fruitful framework for implementing digital curation/preservation environment and has the ambition and the capacity to define concepts for a

<sup>2</sup> See Kai Naumann, Christian Keitel, Rolf Lang, “One for Many: A Metadata Concept for Mixed Digital Content at a State Archive”, *The International Journal of Digital Curation*, 2009, 2, <http://www.ijdc.net/index.php/ijdc/article/viewFile/120/123>: “It is the diversity of these objects which represents the key challenge in devising a metadata concept to describe, preserve and distribute them. They all need to be located on the existing finding aid system, regardless of their media format”. See also Pikka Heutonen, “Creating Recordkeeping Metadata”, *Atlanti*, 19 (2009), pp. 67-76.

<sup>3</sup> L. Duranti, *From Digital Diplomats to Digital Records Forensics*, in print.

<sup>4</sup> *Ibidem*, with specific reference to Bernard D. Reams Jr., L. J. Kuttan, and Allen E. Strehler, *Electronic Contracting Law: EDI and Business Transactions*, 1996-97 Edition (New York: Clark, Boardman, Callaghan, 1997), p. 37.

<sup>5</sup> L. Duranti, *From Digital Diplomats to Digital Records Forensics*, cit.

<sup>6</sup> *Ibidem*.

<sup>7</sup> M. Day, *Preservation metadata*, <http://www.slideshare.net/michaelday/preservation-metadata>.

<sup>8</sup> The definition of significant properties is emblematic of the pointlessness of this new term: “the characteristics of digital objects which must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what the purport to record” (see Andrew Wilson, but also InSPECT - Investigating the Significant Properties of Electronic Content over Time). The term seems to concentrate what common sense normally does.

general frame: the new version (under final approval) has not been able to solve all the uncertainties even if a serious improvement is easily recognizable.

2. development of interrelations and concrete and open cooperation among relevant projects and standardization process (like PREMIS, InterPARES, PLANETS, CASPAR, DRAMBORA, RAC, CIDOC) with the aim of building an interoperable framework and diminishing the present fragmentation for a better orientation of the users.

As a consequence:

3. integration of models, schemas and business solutions to be developed in the application scenarios for handling relevant tasks as:

- authenticity and its presumption,
- storage systems in independent environment,
- automated metadata extraction: on this last point, some efforts have been made recently, but the results are slow and not enough convincing. The time is not enough to enter into details. Two recent contributions to the field could be taken into account: Kim-Ross research on automated genre classification and the FinnONTO project developed in Finland<sup>9</sup>.

The complexity and the contradictions of the digital world could have two opposite consequences, as directly experienced by many e-government legal frameworks and preservation projects: frustration and inactivity on one side, free attitude for creating, testing and supporting innovation on the other side without avoiding or hiding difficulties. Of course the last possibility requires capacity, courage and most of all confidence on the professional accumulated knowledge. The session has offered the opportunity to share ideas and increase the quantity and the quality of this knowledge in one of the most complex and relevant task we have to face, rich of promises and contradictions. One more reason to thank the organizers for this event and all of contributors for their efforts.

---

<sup>9</sup> Y. Kim, S. Ross, "The Naming of cats. Automated Genre Classification", International Journal of Digital Curation, 2 (2007), 1, <http://www.ijdc.net>; Pikka Heutonen, "Creating Recordkeeping Metadata", Atlanti, 9 (2009), pp. 67-76. For the FinnONTO project see [www.seco.tkk.fi](http://www.seco.tkk.fi).

## **Miquel TERMENS, Mireia RIBERA, and Alice KEEFER**

### **Does "long-term preservation" equate to "accessibility forever"?**

#### **Abstract**

This paper proposes a reflection on the points of convergence between the fields of digital preservation and digital accessibility, in terms of both research and development. The two areas have little exchange between them. But, if we look more closely, we find numerous elements – such as objectives, procedures and unresolved problems – that coincide.

Starting with objectives, each area strives to serve users who, at first glance, are quite different: digital preservation is aimed at future users that will use digital platforms that are still unknown, whereas accessibility focuses on current users with disabilities or within disabling contexts. But on closer look, there are parallels between the two groups of users. In both cases there is a considerable lack of understanding about the true needs of users and many unknowns about their technical usage requirements.

As to procedures, the standards for preservation (ISO 14721:2002 – OAIS) and accessibility (CWA 15778:2008 – Document Processing for Accessibility) share obvious similarities. Both propose a model in which there are entry formats, internal formats, and output or dissemination formats. The criteria for format selection in both fields are frequently quite similar.

Finally, there are common, unresolved problems. In the field of preservation a debate has long existed about which “significant properties” need to be preserved, whereas with accessibility, in the absence to date of serious consideration about elements such as emotional aspects, this debate is just beginning.

In conclusion, there is an evident need and rationale for establishing bridges between the two fields in order for them to learn from one another. If they join forces, it is quite possible that common solutions can be found.

**Keywords:** Digital preservation; Digital accessibility; Long term access.

#### **Introduction**

The goal of digital preservation is to allow documents produced in the past to be accessed in the future. For this reason “access” – or “accessibility” – of preserved documents is one of the recurring themes in this area. At the same time the term “digital accessibility” has another meaning: it is the combination of techniques that make it possible for digital documents to be used by anyone, regardless of possible disabilities: vision impairment, motor difficulty, deafness, etc. The aim of this paper is to relate the “accessibility” concept of digital preservation with this second meaning and comment on similarities and differences between the two areas.

Digital preservation and accessibility are two distinct areas, but we believe that it is worth noting the important similarities in the problems that each seeks to resolve. As such, we believe that the way in which solutions are sought in one area can, at the very least, shed light on issues under consideration by the other. For example, both areas have addressed how to select and maintain important document features for subsequent access: to future generations, in one case, and to users with sensory disabilities, in the other. Another relation between the two areas is the uncertainty surrounding the needs of real users, either because they are future users and we do not know what technology they will be using; or because the technology for assisting them currently advances at such a fast pace that their needs adjust continually to the ever-increasing capacity of new systems.

In spite of these points in common and the fact that both areas work with the same elements – digital objects – they do so in separate ways, in terms of the persons, institutions and standards that are devoted to them.

One example that reveals how preservation and accessibility are not marching in unison is that of open repositories of scholarly material, promoted by universities and other research institutions: given the importance of the stored content, preservation aspects are being given attention but, paradoxically, little attention is being paid to the current accessibility of these same documents. [1]

#### **Beneficiaries**

On a conceptual level both preservation and accessibility share the broad goals of working to serve all types of users, but the reality does not reflect this ideal. For example, the directives for the accessibility of web content recognize explicitly that they do not include users with cognitive disabilities. [2] Also, even if documents are created following the existing standards, their accessibility is not guaranteed. Some producers in targeting a specific audience may decide that a given property is not essential, even though this will cause the product to be inaccessible for other groups for which the eliminated property may be very important. [3]

Even though generic techniques exist that permit specific documents –or parts of documents—to be accessible to all users, in many cases it is absolutely necessary to know the potential audience in order to apply the most relevant solutions. For example, designing a product for the prelingually deaf may result in a sharp reduction of textual language, even though the resulting product is then ill-suited for persons with visual impairments. Similarly, in the field of commercial publishing it has proven impossible to create via a single production line a digital work that responds adequately to all situations, publishing channels, and needs. Thus, the CWA recommendation gives examples of good practices with diverse scenarios, but it makes it evident that each situation will require solutions adapted to its own users. So in the end technical efficiency and economic viability are the parameters that determine the adoption of particular accessibility solutions.

In preservation the vision is similar, but expressed using a different terminology. As the OAIS standard itself states, the purpose of preserving digital information is to “make it available for a Designated Community” [4]: the intended future users of the preserved digital objects. And why is this so? For a very simple reason: the awareness of the difficulty –if not to say the impossibility– of fully preserving all original properties of the digital objects. Again, technical constraints and the need for economic viability lead to solutions in which only some significant properties, or essential elements, of the objects are preserved. And this poses a question –which elements are essential?– that can only be answered from the perspective of a given designated community. [5][6] Even for the experts this is not an easy matter because the choice of significant properties is subjective, making it difficult to arrive easily at agreements.

There are two main streams of thought regarding accessibility [7]: the user-centred design, more inclined to create specific solutions for different communities (the elderly, those with motor disabilities, etc.); and the universal design that promotes the idea of a single design to serve all publics. Nonetheless, both visions share the belief that documents produced with accessibility in mind will end up being better for everyone. On the other hand, with preservation there is the growing tendency to design systems adapted to a specific community of users, in which the preservation of given significant properties are prioritized over others. As a result in the future we may find documents that are valid for one community, but perhaps totally unintelligible or unusable for others.

These choices – be they related to accessibility’s audience or to preservation’s designated user community - lead to a renunciation of the digital object’s universal applicability and can prove difficult for different sectors’ experts to accept. For example, questions arise such as: Why renounce the subtitles of certain videos? Why not preserve the original typography and colour of a catalogue of artworks?, etc.

## Problems

Making digital documents and computer applications accessible, as well as preserving all types of digital objects, are stimulating missions, but at the same time difficult to accomplish fully. The basic principles can clash with formidable technical, economic, and management difficulties.

The first difficulty is the broad reach of the missions: at present it is impossible to make all content accessible and to preserve all that needs to be preserved. Priorities must be established and the criteria can vary: the easiest, the most economical, the most scalable, the most heavily used, etc. Therefore, prioritization requires policies to be applied. And the other side of the prioritization coin is the renunciation: of what (for the moment, perhaps) will not be accessible or will not be preserved. The policies of prioritization are painful because implicitly they go against the global aims: some disabled persons will not have access to content that they perhaps will need; others, in the future, will not have access to specific data or testimony from our time.

Traditional accessibility solutions, such as screen magnifiers or screen readers, are built upon the applications and thus are not well integrated into operating systems and other programs. In the long run, the solution will lie in incorporating accessibility into all phases of the development of hardware, software and content, as well as having it present in the workflow of document management. In digital preservation the use of proprietary file formats multiplies the challenges of managing their preservation, as does the plethora of existing formats. A similar reflection can be made regarding the limited support that standardised metadata schema receive from many software applications, not to mention file formats that do not admit metadata.

Legal barriers are also common to both accessibility and preservation. In the analogical world, in many countries the law protects the rights of disabled persons by setting limits to the intellectual property rights in order to allow for the publication of books in Braille. In the digital world, there tend to be fewer exceptions, a situation that leads to increased expense for the rights to publish accessible works. This has opened new fronts for the struggle to broaden rights concerning digital documents. [8] In other cases the problem does not stem from the document itself, but rather from the existence of proprietary reader software that impedes the introduction of elements that contribute to accessibility.

The problem is similar with preservation. Laws protect the rights of copyright holders by prohibiting the reengineering or decompiling of software, or the modification of content formats, to cite three of the major

techniques applied in many preservation scenarios. Certainly, in recent years there have been numerous initiatives to permit such activities in the context of preservation. But at present, many preservation-related actions currently take place in an environment of questionable legality, at the least. Similarly licenses and usage restrictions –sometimes in the form of Digital Rights Management (DRM) – are also barriers for producing documents that are accessible to all. [9] They also act as barriers for full preservation.

The timing of implementation is also important. The law in many countries protects the publication of accessible versions of textbooks for disabled students, especially the visually impaired. Nevertheless, the procedure for exercising this right can be slow, and may not conclude until after the publication of the commercial work. Therefore, users dependent on the accessible version receive the work much later than others. [10] Similarly, preservation is still seen in many scenarios as an activity taking place at the end of the “normal” life of a digital object and hence is not considered until it is time to “store” the object. At that stage, immediate treatment –for example, migration or other procedures – may be necessary, which might have been spared had the digital objects been created in accordance with preservation requirements.

### **Technological foundation**

A comparison of the principal technologies of accessibility and preservation leads us to conclude that there are important similarities both in the procedures and recommendations promoted in each area. The major ones are: the transformation of file formats as a basic technique for facilitating present or future access; the standardization and use of structured and open formats; and the requirement to make full use of metadata.

For accessibility, the use of standards in software and open formats facilitates interoperability and, therefore, the integration of technical aids for reading the documents. The use of structured formats from the XML family facilitates the transformation of documents and, therefore, it too aids in the generation of versions adapted to the needs of different user communities. [11] DAISY is perhaps the format that is currently experiencing the greatest development along these lines, within the publishing sector. [12]

In the field of preservation, many experts have prioritized the reduction of formats used and the appropriate choice of formats for subsequent preservation. The choice of formats is frequently made during the creation, or even during use, of the document and thus remains beyond the scope of preservation actions. However, some programs encourage the use of open, interoperable and standard formats. [13] An appropriate characterization of files and the proper structuring of contents in them can also contribute towards subsequent preservation tasks, such as migration.

Whether from the vantage point of accessibility or of preservation, the volume of digital production is so great that it is virtually impossible for all files to be handled appropriately after the fact, e.g., after their creation. This leads to the recommendation that files should be standardised at their point of origin, as a means of reducing variability. It should not surprise us, then, that accessibility experts are promoting the adoption among publishers of standards and common formats, for both webs and textbooks. [14] This would enable the publishing chain to generate specific products with varying presentations and formats geared to the needs of each user. In preservation, there are many more sources of content generation, since digital objects created within the publishing world account for only a small fraction of the total number of items to be preserved. Some current attempts for influencing how digital objects are created are centred within the public administration and some scientific fields and it remains to be seen if and how it will spread to other areas in the future.

### **Conclusion**

Accessibility and preservation serve different objectives even though they act on the same types of materials. In this work we have seen some of their similarities: in strategies, in approaches to challenges, and in technological underpinnings. We have also seen that some proposed solutions are stymied by pre-existing legal conditions. Likewise we have shown how practical concerns – such as technical expediency and economic viability - can lead to actions that are frequently more limited than the respective movements' overriding aims.

These common elements lead us to believe that a greater degree of understanding between the two communities would be beneficial to both sides. Surely each could learn something from the other and, in so doing, shed more light on its own approaches. Also, collaboration would be beneficial in order to reach common objectives, such as the promotion of open and structured formats.

Finally, it is worth remembering that the two communities maintain close relations with certain stakeholders: universities, public administration and libraries. Equally important to both is the expansion of e-government as the main transforming engine for practices related to the creation and management of digital content. This common ground could facilitate points of encounter and contribute to working together towards common solutions.



## References

- [1] Kelly, Brian (2006): "Accessibility and Institutional Repositories". UK Web Focus. <http://ukwebfocus.wordpress.com/2006/12/12/accessibility-and-institutional-repositories/>
- [2] Web Content Accessibility Guidelines (WCAG) 2.0. W3C Recommendation 11 December 2008 (2008). <http://www.w3.org/TR/WCAG20/>
- [3] CWA (2008): CEN/ISSS CWA 15778:2008 – Document processing for accessibility. <ftp://ftp.cenorm.be/PUBLIC/CWAs/DPA/CWA15778-2008-Feb.pdf>
- [4] Consultative Committee for Space Data Systems (2002): Reference model for an open archival information system (OAIS). Blue Book. <http://public.ccsds.org/publications/archive/650x0b1.pdf>
- [5] Cedars Project (2001): The Cedars Project Report. April 1998 – March 2001. <http://www.leeds.ac.uk/cedars/pubconf/papers/projectReports/CedarsProjectReportToMar01.pdf>
- [6] Hedstrom, Margaret; Lee, Christopher A. (2002): "Significant properties of digital objects: definitions, applications, implications". Proceedings of the DLM-Forum 2002. Luxembourg, Office for Official Publications of the European Communities. p. 218-223.
- [7] Seale, Jane K. (2006): *E-Learning and Disability in Higher Education: Accessibility Research and Practice*. Oxford, Routledge.
- [8] Commission of the European Communities (2008): *Green paper: copyright in the knowledge economy*. Brussels, Commission of the European Communities. [http://ec.europa.eu/internal\\_market/copyright/docs/copyright-info/greenpaper\\_en.pdf](http://ec.europa.eu/internal_market/copyright/docs/copyright-info/greenpaper_en.pdf)
- [9] Kramer, Elsa F. (2007): "Digital Rights Management: Pitfalls and Possibilities for People with Disabilities". *Journal of Electronic Publishing*, 10 (1). <http://hdl.handle.net/2027/spo.3336451.0010.106>
- [10] Keil, Sue; Parris, Delith, Cobb, Rory; Edwards, Angela, & McAllister, Richard (2006). *Too little, too late - provision of school textbooks for blind and partially sighted pupils*. London, Royal National Institute of the Blind.
- [11] Paepen, Bert; Engelen, Jan (2002): "Using XML as a Reading Enabler for Visually Impaired Persons". 8th International Conference, ICCHP 2002. *Lecture Notes in Computer Science*, 2398, p. 382-389.
- [12] Kahlisch, Thomas (2008): "DAISY: an opportunity to improve access to information for all". *Information Services and Use*, 28 (2), p. 151-158.
- [13] Arms, Caroline; Fleischhauer, Carl (2005): "Digital formats: Factors for sustainability, functionality and quality". Proceedings Society for Imaging Science and Technology (IS&T) Archiving 2005. Washington DC. p. 222-227.
- [14] Lockyer, Suzanne; Creaser, Claire; Davies, J. Eric (2005). "Availability of accessible publications: designing a methodology to provide reliable estimates for the Right to Read Alliance". *Health Information and Libraries Journal*, 22 (4), p. 243-252.

## Sven SCHLARB, Andrew N. Jackson, Max Kaiser, and Andrew Lindley

### The Planets Testbed: a collaborative environment for experimentation in digital preservation

#### Abstract

This paper presents the Planets Testbed, a web-based application that provides its users with a controlled collaborative environment for scientific experimentation in digital preservation. The paper gives an overview about the core concepts of the Planets Testbed and describes how the application supports the user community in preserving the digital cultural heritage.

**Keywords:** Planets project, Testbed, digital preservation, long term preservation

#### Introduction

The Planets Testbed is one of the core results of the FP6 Planets Project (<http://www.planets-project.eu>) which aims to create a software suite capable of addressing the digital preservation challenges that libraries, archives and the digital preservation community are currently facing.

The Planets Testbed is more than a software package – it is a central environment (consisting of software, hardware and data) for testing the performance and capabilities of tools for digital preservation. The tools are offered as web services which can be combined in complex workflows. Measurement processes are highly automated, allowing large amounts of tool evaluation results to be collected via mass experimentation.

The Planets Testbed is essentially community software dedicated to people dealing with long term preservation issues on a day-to-day basis. In the following, we will provide an overview of the Planets Testbed and discuss its role for the dedicated user community and for the preservation of the digital cultural heritage.

#### The Planets Testbed

##### The Planets Testbed Environment

The Planets Testbed provides a web-based software allowing to explore and test preservation services. This software relies on a Planets-wide, interoperable infrastructure, through which different tools can be invoked in a uniform way: the Planets Interoperability Framework. It defines the generic interfaces enabling the seamless integration of a large number of tools each of which provides a specific functionality required for performing long term preservation tasks.

##### The Experiment Process

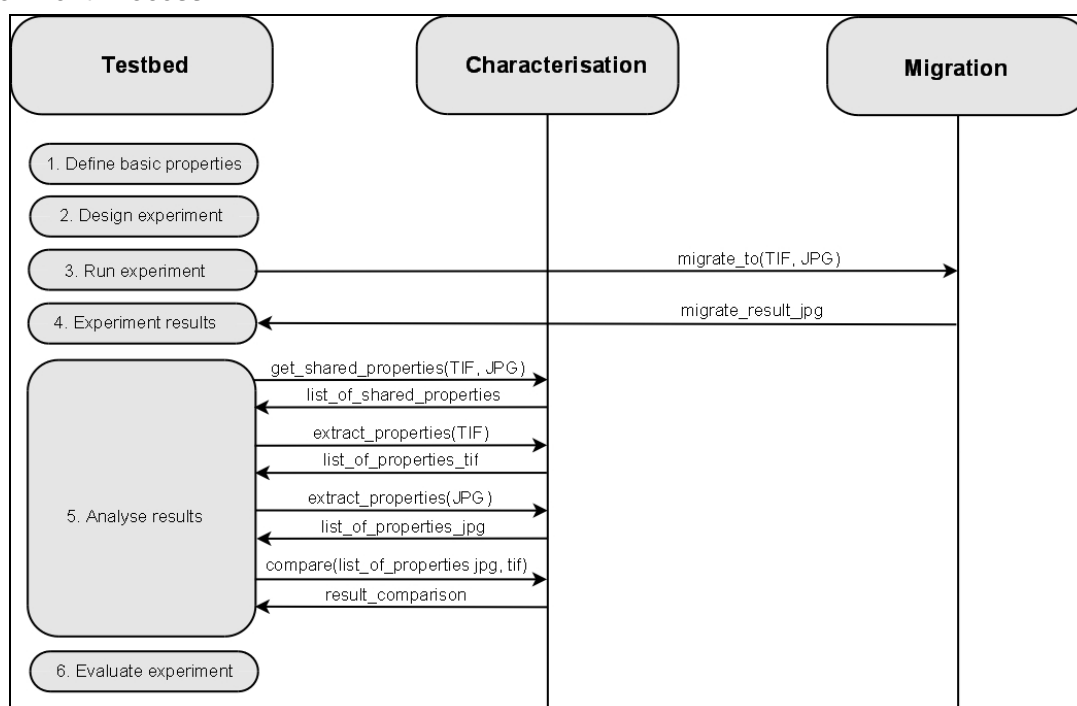


Figure 1: Example of a Planets Testbed experiment process

Different kinds of experiments are divided into different 'Experiment Types' (see section 1.3). Each experiment type of is based on a workflow which itself consists of a sequence of preservation service operations.

Using the Planets Testbed web application, the user is guided through six steps of an experiment process, as shown on the left-hand side of Figure 1. The following walk-through will use an example which might play a role in a real institutional process: The automated characterisation and migration of digital content. To be more concrete, this example refers to the migration of a single TIF file to a single JPG file, and subsequently the comparison of the properties of the input and output files.

#### Define Basic Properties

In the first step of an experiment, basic experiment metadata is recorded. A user is required to enter a name for the experiment along with some basic information about the experimenter. The user can also supply information on the overall purpose and focus of the experiment, and references to relevant experiments, scientific publications or web resources.

#### Design Experiment

The experiment type can be selected here. A simple graphical representation of the experiment workflow is presented to the user. Configuration of this workflow depends on the experiment type, but in most cases, this involves browsing and selecting available services and selecting digital objects to experiment upon. The digital objects can be chosen from the data sets available in the Testbed or from content the user has uploaded.

Taking the example of the migration experiment, the workflow is configured by selecting a migration pathway, composed of the starting format TIF, the target format JPG, and a migration service (e.g. ImageMagik).

#### Run experiment

Once designed and configured, the experiment can be submitted for approval. At this point, the administrator in charge of the Planets Testbed is given an opportunity to prevent the experiment from being executed, for example if it is likely to put an unreasonable load on the server if executed at that time. Experiments that require only modest resources are automatically approved, and can be executed straight away.

Following approval, the user can initiate execution of the workflow. The Planets workflow execution engine then takes each digital object, and passes it through the specified chain of services.

#### Experiment results

In this step, the user can inspect the experiment result objects, overall success rates and basic performance statistics, e.g. whether all migration actions successfully created new digital objects. The user is also given the opportunity to re-run the experiment in order to collect additional data.

#### Analyse results

If characterisation tools are available for the digital objects which are part of the experiment, they can be used to analyse the properties of the digital objects. In our migration example, there are two digital objects, an input TIF file and the resulting JPG file which have different file format specific characteristics. Based on the common set of properties of these file formats which are determined by a Planets characterisation service, the values can then be automatically compared using the metrics that apply to the different properties.

#### Evaluate Experiment

The final step of an experiment allows the user to judge the overall performance of the preservation workflow. The experimenter can also provide a brief written report about the experiment's outcome. The result can then be more widely shared between Planets Testbed users, so that others can learn from the results or even setup an equivalent experiment in order to reproduce and verify the outcomes of other experimenters.

### Planets Testbed experiment types

An experiment type defines the generic structure and data flow of an experiment, and there are many kinds of experiments to be explored other than the migration experiment outlined as an example above. In the following, we shortly describe the experiment types that exist so far.

- *Characterisation Experiments*

A characterisation experiment allows for direct comparison of characterisation tools against each other or against a set of authoritative property values.

- *Validation Experiments*  
A validation experiment is used to test whether a digital object is well-formed and valid with respect to a particular format.
- *Emulation Experiments*  
Emulation generally refers to imitating a (usually obsolete) soft- and hardware environment within another (usually up to date) soft- and hardware environment. In the Testbed, an Emulation experiment creates an emulation session for a digital object which is then visualised the imitated soft- and hardware environment. By that way, the user can record how well the object is being rendered with respect to this specific environment.
- *Execute Plato preservation plan*  
"Plato" (see [1]) is one of the outcomes of the Planets project, and is a web based software for creating a preservation plan for preserving a specific collection or a part of a collection of digital objects. The concrete recommendation of the preservation plan ends up in an "executable preservation plan" which can then be evaluated by a corresponding Planets Testbed experiment.

It is to be expected that the existing experiment types do not cover all the requirements for the different experiment scenarios the long term preservation community might require. If an experiment does not fit with one of the existing experiment types, a new experiment type must be set up by a Testbed administrator contacted through the Testbed helpdesk (see end of section 3).

### **Sharing knowledge with the Planets Testbed community**

The Planets Testbed is community software in the sense that it allows reviewing and even reproducing existing experiments by all community members. New experiments can reference existing ones and refine or give a statement on existing experiment results. In that way the community members contribute to a continuously growing and reliable knowledge base on digital preservation.

The main goal of the Planets Testbed in this aspect is to enable community members to share their research results amongst cultural heritage institutions all over Europe. The Planets Testbed acts as the central experimentation platform gathering knowledge about long term preservation topics in various dimensions: In the first place, an experiment can focus on performance and reliability of long term preservation services and the underlying software components themselves. Then, the annotated experiment datasets contain information about special cases (an extreme value for a file format specific parameter, for example) and important properties of digital objects. And finally, the Planets Testbed establishes a procedure to share meaningfully aggregated results with other Planets software, like Plato (see [1]), for example.

### **Knowledge about long term preservation services**

A wide range of preservation services have been developed by the Planets project, and the Planets Testbed aims to make them available for public use. Each service is supplied with metadata describing the supported formats, migration pathways, the identity of the service creator, the location of the endpoint which makes the Planets service available and so on. The Planets Testbed makes it easy to explore this information which is continuously managed and maintained.

### **Knowledge about experiment datasets**

Some experiment types require information about the data an experiment is based upon. The Planets Testbed integrates annotated datasets (corpora) in order to be able to check the output of a service against recorded metadata. As a simple example, if an identification tool is tested against an object of a known format (e.g. PDF file), the Planets Testbed can compare the embedded properties against the results from the identification service. This allows the scope and accuracy of identification tools to be closely examined. Similarly, validation services can be exercised using carefully constructed valid and invalid documents, testing the edge-cases of format specifications. For example, the Isartor test suite (<http://www.pdfa.org/doku.php?id=pdfa:en:isartor>) can be used to detect whether validation tools can spot PDFs that are invalid with respect to the PDF/A-1 (ISO 19005-1:2005) specification.

### **Contributing to the Planets-wide knowledge base**

By standardising and sharing results, the Planets Testbed acts as a central point for accumulation and aggregation of data from many experiments and across institutional boundaries. From this rich dataset it should be possible to determine the robustness and performance of particular preservation tools and techniques in an objective manner.

The results are stored centrally and can be used as a basis for future development of a knowledge base.

## The Public Planets Testbed

The Planets Testbed software will be made publicly available by the Planets project. A full installation requires all of the different preservation services to be installed, each of which may have different software dependencies and operating system requirements. The publicly available central Planets Testbed addresses this problem by providing as many tools and services as possible – pre-installed, configured and ready for testing. The Planets Testbed can be accessed using a web browser and allows interested parties to evaluate all the preservation services and strategies supported by Planets using their own data or benchmark content. Additionally, it is possible to download and install individual Planets Testbed instances. The software installer makes it easy to deploy the Planets Testbed locally, but can only provide limited functionality out of the box. The public Planets Testbed is available at <http://testbed.planets-project.eu/testbed>, hosted by HATII at the University of Glasgow. It is currently in beta release phase and selected external parties have accounts granted. The service will go completely public in beginning of 2010, but it is already possible to ask for an account at [helpdesktb@planets-project.eu](mailto:helpdesktb@planets-project.eu). Further information about the Planets Testbed, also about upcoming training workshops can be found on the Planets website.

## Conclusions

The innovative aspects of the Planets Testbed are the ways in which experimental data is collected, analysed and shared. The Planets Testbed provides a single interface to a wide range of hardware and software benchmarking environments, so that data can be collected reliably and reproducibly.

The Planets Testbed is also building corpora of digital objects with well-known properties. These properties, in combination with a number of innovative Planets software technologies, allow for the outputs of preservation services to be analysed rigorously and automatically.

Finally, the Planets Testbed defines standard semantic structures to contain these results, permitting community-wide aggregation of experimental results and experiences using the tools and services needed for long-term preservation of the digital cultural heritage.

The Planets Testbed will be made available to the digital preservation community as a free service by beginning of 2010.

## References

- [1] Christoph Becker, Hannes Kulovits, Andreas Rauber, and Hans Hofman, Plato: a service oriented decision support system for preservation planning, JCDL '08:Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (New York, NY, USA), ACM, 2008, See <http://doi.acm.org/10.1145/1378889.1378954>, pp. 367-370.
- [2] Petra Helwig, Judith Rog, Caroline van Wijk, Eleonora Nicchiarelli, and Manfred Thaller, Test methods for testbed, Tech. report, 2007, See [http://www.planets-project.eu/docs/reports/Planets\\_TB3-D2\\_MethodsForTesting.pdf](http://www.planets-project.eu/docs/reports/Planets_TB3-D2_MethodsForTesting.pdf).



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

**Jan HUTAŘ, Andrea FOJTU, Marek MELICHAR, and Bohdana STOKLASOVÁ**

## **Czech National Digital Library and long-term preservation issues**

### **Abstract**

The Czech Republic has earned worldwide recognition for its remarkable results in the area of cultural heritage preservation. However, digitisation and digital preservation are significantly hindered by a lack of resources. This results in a relatively slow pace of digitisation. Furthermore, it leads to serious delays in dealing with, and solving, current digital preservation issues.

This paper explores a “National Digital Library” project, which has been accepted by the Ministry of Culture as a candidate for European funding under the Integrated Operational Programme. The National Library of the Czech Republic, along with the Moravian State Library in Brno, have prepared an ambitious project with two main goals – 1) to accelerate the digitisation; 2) to establish a trusted long-term preservation repository. 1.2 million documents should be digitised within the next 20 years. The most fragile documents should be digitised during the five-year project between 2010 and 2015.

The following paper reflects on the issues we have to face in preparation for this project. The mass digitising encourages the institutions which are involved to make a number of organisational changes. There are also a number of strategic decisions to be made (national/institutional digital preservation policy formulation, national bibliographic identifier scheme implementation). And there are also a number of technical tasks with the possibility of an enormous future impact (like decisions on what file formats and metadata formats to use for this mass digitising, how to choose LTP system software, how to include the data from previous projects etc.)

The paper concludes that planning large-scale digitising needs significant administrative, organizational and political preparation, which may be more overwhelming than the technical part of such a project. Involved institutions must be ready for a business change, well before the scanners produce the first pages.

**Keywords:** digital preservation; national policy; mass digitisation; EU project

### **Historical background**

In the National Library of the Czech Republic (NLCR) digitisation started in the early 1990s and the webarchiving was launched in 2000. The first digitisation projects, financed from public grants of the Ministry of Culture of the Czech Republic (MCCR), were focused on old prints and manuscripts, later also on historical newspapers. Digitisation was then considered to be a way of preservation as it reformatted the documents in danger of deterioration. Later, naturally, the focus had changed, and turned more towards the end users needs. Our digital libraries tried to comply with existing international standards, and the digital content is being integrated into portals like TEL, EUROPEANA. Even though the Czech Republic is a small country, it has earned worldwide recognition for its long tradition and remarkable results in the area of culture heritage preservation: in 2005, the NLCR was awarded the first UNESCO/Jikji Memory of the World Prize for its contribution to the preservation and accessibility of our documentary heritage.

Until today the main projects have produced 80TB of data, and yet they cover only a small fraction of our national cultural heritage. With the current pace of digitising, we would be working for the next 300 years to make accessible the nation’s cultural heritage in digital form. However, lack of sufficient funding slows down the digitisation process and leads to delays in dealing with digital preservation issues.

### **National Digital Library (NDL)**

In 2005 MCCR with NLCR jointly prepared the National preservation policy for the traditional and electronic library documents until the year 2010. [1] Hereby proposed funding was 8 million Euro. The whole policy stayed only on the level of declaration, and the financing was never approved by any of the next Governments.

The image below explains the organization of digital preservation as described in the National preservation policy for the traditional and electronic library documents until the year 2010.

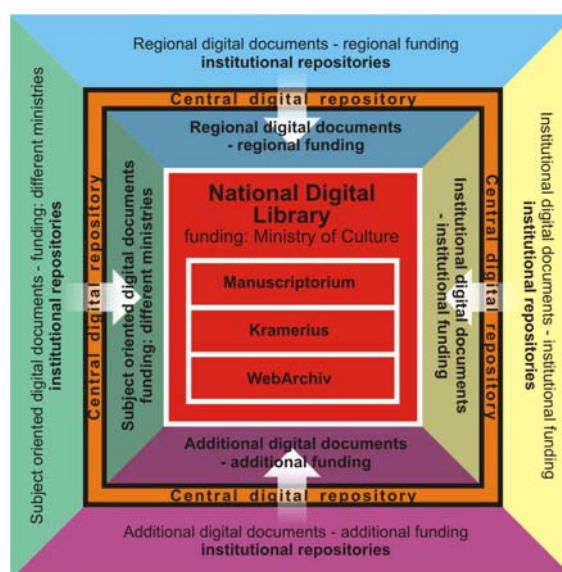


Figure 1: Project of the National Digital Library and its central repository

NDL forms the heart of the whole national system of culture heritage preservation. NDL contains the core of our national cultural heritage. These documents, digitised or born digital, are acquired, preserved and disseminated within three large national projects funded by the Ministry of Culture.

- Manuscriptorium [2] is a system which gathers information on historical book resources, linked to a virtual library of digitised documents.
- The Kramerius [3] project focuses on the preservation of and accessibility to 'modern' periodicals (from year 1801 onward), books and other documents in danger of acid paper degradation.
- WebArchiv [4] is a digital archive of Czech web resources

Digital documents held by any Czech library, museum or archive can be selected to become part of the NDL. Digitising and preservation of these documents is to be funded by the MCCR. Digital data not selected for the NDL can also be deposited into the central repository, but their long-term preservation has to be funded from other resources. Other institutions may not be interested in depositing their data into the central repository. In such cases they have to secure their own financing, but their data can be integrated into the national access portals, provided that they comply with the metadata standards.

NDL operates in the broader context of a wider national digitising strategy of the MCCR, which also covers archival documents, museum collection, architectonic monuments, performing arts and media etc. However this wider strategy is currently only operating on a conceptual level, it's not a policy with proposed financing schemes.

### **NLCR and digital preservation issues: Current state of the art**

NLCR is the key player in the area of long-term preservation in the country. Other libraries in the country rely on NLCR and wait for a solution they can follow. The nationwide standards for the digitisation projects are set by NLCR. Most of the projects use the same metadata schemas, the same access applications, the same or similar file formats.

The current state of digital preservation in NLCR is far from ideal. From the OAIS point of view NLCR only implements the archival storage module. Current installation, the Central Data Storage (CDS), is built on IBM products. Two IBM Systems Storage DS 4800 are installed, one in Klementinum and the second one in Hostivař data centre (18 km distance). The Tivoli Storage Manager (TSM) is used for the back-up and archive services, together with an IBM tpe library. We also have archival back-up and an archiving strategy in place. Of course disaster recovery services are in place and replication between localities helps to protect data against physical destruction and human or software error.

Some parts of OAIS functions are secured by number of applications used in the digitisation workflow or in the access applications. The ingest processes are limited to the hash function checks, consistency and completeness of the package and batch. Access applications use separated data storage and metadata database, and the long-term archival copies are seldom used. There is no metadata management module upon the archival data. Data arrives with several different formats of metadata, and there are often problems with

identifying the archival documents, and providing links to the access copies and library catalogues. Many of the basic OAIS functions which enable digital object management and preservation planning are missing.

Access is secured by servers access applications, which have more or less reliable authorization and authentication mechanisms. Each of them has its own store of the user copies. International standards as DC and MARC21 (used as MARXML) are used in most Czech libraries. Since 2007 we have used a structural, administrative and technical metadata scheme based on METS and PREMIS (the PREMISobject part).

NLCR have participated or NLCR currently participate actively in a number of European projects in the area of digitising (ENRICH, TEL+, TEL-ME-MOR, EDLnet) and digital preservation (DigitalPreservationEurope). Involvement in the DPE project was extremely important. In 2007 the central digital repository of the NLCR went through an audit based on the first generation of the DRAMBORA toolkit, and this helped the staff to realize the risk related with current underdeveloped preservation solutions.

### **Dramatic step forward: Towards the IOP project**

The situation described above is destined to change very soon. The NLCR, along with the Moravian State Library in Brno (MLB), have prepared an ambitious project, which should be financed mainly from the Integrated Operational Program Smart Administration in 2010-2015. The feasibility study for the project was finished in September 2009, and the project calls should be open at the end of the same year. The project, called again "National Digital Library," has the following main goals:

- to accelerate digitisation (building two digitisation centres with robotic scanners in Prague and Brno for mass digitisation)
- to improve long-term preservation and access to digital objects (building a trusted and certified digital repository using two geographically separated localities - Prague and Brno, 180km from each other, purchasing digital preservation system software and tailoring it to the needs of the project).
- to secure wide dissemination of the national cultural heritage in digital form in a user friendly environment (using national aggregators and portals, possible also upgrading the technology of the national meta-search tools)

The digitising centers and the long-term preservation system have to be integrated into the existing infrastructure of the two participating libraries. Some further steps are needed to achieve permanent financial, technological and administrative sustainability of the created systems and of the digitised data and access to them.

The core of the Czech national cultural heritage (documents published in the country since 1801 + historical documents until 1800 stored in Czech libraries) form approximately 1.2 million documents, that is about 350 million pages. Many of the documents are printed on acid paper and/or are highly used and their digitisation is therefore very urgent. The projects infrastructure should allow digitising these 350 million pages within the next 20 years. The most fragile or highly used documents should be digitised during a five-year project itself between 2010 and 2015. The results of the project will be digitisation of 540.000 documents published since 1801, 20.000 documents published before 1800, archiving Czech web, all together producing about 1,5PB of data. The total budget of the project should be 27 million EUR (85% from European funding and 15% from co-funding).

### **The issues and decisions we have to face**

The whole IOP project draws attention to a number of issues which need to be clarified before the mass digitisation will start. During the preparation of the project we have to face a number of organizational issues, and make a number of strategic and technical decisions. It is clear now, that organizational and process changes are on the same level of importance as those of a technical nature.

### **Staffing and organization**

First of all, even though the NLCR has been running a digitising line and some software and hardware infrastructure for many years, the processes of mass digitisation will require much more technology and staff on all levels. This will inevitably lead to a number of organizational and management changes. There was no "digital preservation department" in the NLCR before 2008, and no IT experts who would be able to coordinate and run the necessary environment, or survey the work of service companies. As well as more skilled staff members, changes to corporate culture are also necessary. The need for closer cooperation between different parts of the library became crucial in the process of project preparation. The communication channels still have to be improved between the IT departments, the new digital preservation team, the cataloguing department and the digitising team. Also, an experienced project manager will be needed to manage and control the entire project. All this means shifting the organization to a more business like culture, and to ensure there is a more cooperative environment inside the institution. Existing workflows in the library are currently undergoing





Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

reviews, so that we can locate the staff which could be relocated to digitising document selection and preparation. The preparation of the documents for digitising requires in many cases conversion of traditional catalogue records from scanned catalogues into Aleph database, or creating new catalogue records, de-duplicating the existing records etc.

This kind of project also needs the support of the top management of the library and the library funding body. Large digitisation still needs advocacy even inside the institution, as the library functions are multiple and the project will certainly affect the daily business in most departments.

### **Strategies, policies and politics**

In a project of such extent we must consider the number of stakeholders' needs. The project is of potential interest to Authors' right holders associations, politicians, producers of HW and SW, similar projects in the country, and in other libraries. On the political level, the coordination with other Smart Administration projects would be necessary.

What more, NLCR has no clear policy statement on digital preservation, specifying the responsibilities and extent of preserved materials with clear long-term cost estimation and technical specifications and requirements. Even the Library Statute is missing a reference to this problem. NLCR prepared an internal draft of the general digital preservation policy of the institution, and suggested a review mechanism of this policy document. But this document has to be approved by the NLCR steering board, which has to recognize the budget and staffing. Besides, some more strategic decisions are to be made on the national level very quickly, especially about the URN:NBN identifier system implementation, and the national bibliographic number implementation. Both will be essential for the success of the mass digitising and preservation projects.

### **Technical issues**

Naturally there are many more down to earth issues, which the project team had to face. Even little decisions are of potential large future impact. The first challenges we faced was to measure the amount of pages and documents we will have to digitise and the data amounts, in order to set the final sizes of the data repository and provide an estimation of the necessary scalability, setting the staffing needs of the project, measuring the economic efficiency of the project, etc. After completing this demanding process we realised that in many areas we only have sufficient data to make rough estimations, even though the founding agency would need a solid exact numbers.

We had to make decisions about the use of file formats, compression level, bit levels, metadata schemas, and other standards. We are determined to follow acknowledged standards field. [5] We expect to use mathematically lossless JPEG2000 for preservation master files and lossy JPEG2000 as a user copy. The OCR files in METS ALTO, and metadata files (XML METS with PREMIS, MIX, MODS or MARCXML). The webarchiving will move towards the WARC format from currently used ARCs. We have to design specific ingest workflows for various types of incoming digital documents, and decide how the existing digital data will enter the new digital preservation storage. This might be quite a time consuming and complex process of migration of old data and metadata into new formats, which would then be ingested into the new system. As we have about 7 millions of pages, this could take months.

At the highest production speed we expect to produce and archive more them 70 000 of pages a day on four robotic scanners in Prague and two scanners in Brno. This will also change the requirements on the access applications, and some will have to undergo technological reconstruction.

Since we have no real IT expert team capable of large scale programming and we do not expect to hire such a team in the future, we decided to search for a commercial solution for the long-term preservation system. Advantages seemed obvious – buying a “ready-to-go” system, which possibly already has some implementations in other libraries/archives. Even though the system will need some adjustments, it will fulfill most of the requirements for secure storage of the mass digitising production. Future support, development activities and upgrades will be done based on the requirements of more users of the system.

According to our present knowledge, there are only three commercial solutions available on the market now (SDB by Tessella, Rosetta by ExLibris, DIAS by IBM). After lot of workshops, corresponding with all providers, our team saw all of these systems running in the real implementations. In August 2009 we sent out an RFI to these three companies to design a complex solution for the central digital repository, which should guarantee the long-term digital preservation and dissemination of digital objects (including metadata). The RFI consisted of the description of the required system and list of requirements which we asked the companies to report back on. Their responses included cost estimations for the LTP system software, databases or other dependent software components, installation and setting costs, maintenance costs, required hardware infrastructure, and a prognosis of the running cost for the next ten years. We were also interested in receiving an estimation of time needed for LTP system implementation (from the contract to pilot phase and to productive phase).



*Empowering users: an active role  
for user communities*

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

The collected data were used in the feasibility study [6], with similar data from the producers of scanning workflows and scanners. The project was submitted in October 2009 and tenders covering different parts of the project are expected in spring 2010.

## **Conclusion**

If it were only the technical decisions about file formats and metadata schemas, the preparation of this mass digitising project would be much simpler. However, the organizational and management aspects can influence the project results much more intensively than we could have expected, and may require much attention and workforce.

## **References**

- [1] NLCR, 2005. National preservation policy for the traditional and electronic library documents until the year 2010. Draft.
- [2] <http://www.manuscriptorium.com>
- [3] <http://kramerius.nkp.cz/kramerius/Welcome.do?lang=en>
- [4] <http://en.webarchiv.cz>
- [5] Florida Digital Archive, 2003. Recommended Data Formats for Preservation Purposes in the Florida Digital Archive. <<http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf>>.
- [6] NLCR, 2009. Feasibility study for project National Digital Library.

## **Friederike KLEINFERCHER and Kristina KOLLER**

### **Cultural Heritage: from the library shelves to network residents**

#### **Abstract**

In the context of the eSciDoc project, the Max Planck Digital Library and the FIZ Karlsruhe are building an e-research environment for multi-disciplinary scientific research organizations. Based on the eSciDoc infrastructure, several solutions for the end-user will be developed and provided as open source software. One of them is ViRR (Virtueller Raum Reichsrecht), a solution to support collaborative and interdisciplinary research on text resources like manuscripts or books. A user-centred approach was applied to define necessary functionalities and adequate graphical user interfaces. ViRR provides several smaller flexible tools in one web interface for the creation and enrichment of metadata, for the modelling of the structure of a work and for the enhancement of the collection with related resources such as annotations and transcriptions. One of them is a configurable online editor for defining the structure of the digitized work in accordance with the structure of the original resource.

This paper will give an overview of the ViRR solution which was developed to support researchers from different backgrounds working together on text resources. Additionally, we will outline eSciDoc, the underlying infrastructure of the ViRR solution.

**Keywords:** eSciDoc, digitized text resources, collaborative workbench, online editor

#### **Introduction**

In the context of the eSciDoc project (<http://www.escidoc.org>) the Max Planck Digital Library (MPDL) has developed a web based solution for different user groups (researchers, librarians) to make their textual holdings online available. ViRR [1] enables the enrichment, dissemination and preservation of digitized cultural heritage like manuscripts or books. Its aim is mainly to support scholars in the humanities in the analysis and evaluation of text resources.

The MPDL is a scientific service unit within the Max Planck Society (MPG), which consists of about 80 institutes from various scientific disciplines, and therefore the development of services and solutions has to deal with requirements from diverse research contexts. During the development of the ViRR solution, the general approach was to start with specific requirements from a pilot community, and then identify generic services, which can be re-used by other disciplines. The aim is to develop a solution which can fulfil most of the diverse requirements of working with digitized text resources within the MPG.

The name ViRR derives from the content of the first collection, which consists of about 20.000 scans of legal artifacts from the period of the Holy Roman Empire provided by the Max Planck Institute for European Legal History (<http://www.mpier.uni-frankfurt.de>).

#### **Working with Digitized Text Resources**

Solutions, which support scholars in their work with digitized text resources, differ in their focus and quality as working instruments. The very basic level is the mere digital representation of a single text resource with basic browsing functions and without any sophisticated user management or re-use options.

A more enhanced level offers functionalities to intellectually enrich digitized text resources. Hereby, the scholars and librarians are able to uncover the "hidden" information, which cannot be provided by a mere digital representation. Some of these functionalities imply the capturing and enhancement of structural metadata and semantics, ideally in different standard formats like METS [2], MODS or TEI. Detailed information about the composition of a resource might be gathered, such as the pagination (logical and physical) or the structure of a work (see e.g. [3, 4]). Standardized interfaces support the re-use of this additional information in other contexts, such as library catalogues, aggregated viewing environments or mash-up services, and allow the integration of external knowledge bases, such as dictionaries or viewing tools.

Having the resources and the related information on the web, the logical consequence is the support of collaborative scenarios from various disciplines, which might assist the creation of knowledge related to the artifacts [5]. The possibility to describe different entities of a resource on a semantic, lexical, etymological or pragmatic level, and to describe the relations of these entities to other resources such as annotations, transcriptions, images or dictionaries, enables a real workbench scenario for scholars in the humanities.

To provide a sustainable solution for supporting these different aspects of a workbench, we have chosen a gradual approach in the development: providing an online editor for the enrichment of structural information, at the same time developing robust content models, to enable future interlinking to other artifacts.

## The eSciDoc Solution ViRR

The eSciDoc solution ViRR combines a set of tools (components) for publishing scientific content in one user interface. This includes the two key features, the electronic modeling and editing of the original source material (ViRR Editor, see Fig. 1) and its online representation in a digital library (ViRR Viewer, see Fig. 1). These features are often separated from each other and realized in different tools, so that data transformations between these tools become a necessary drawback. The integrated design of ViRR allows users to perform all working steps within one software solution.

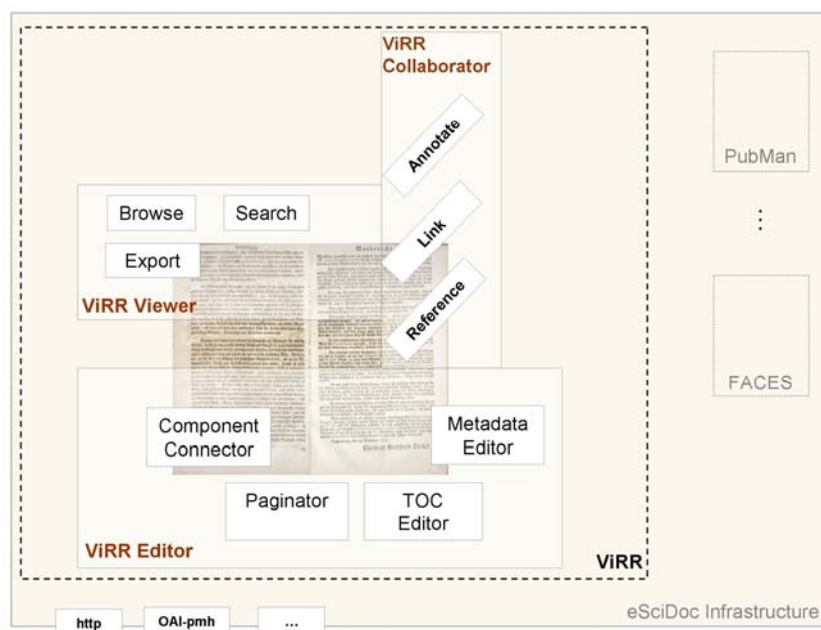


Figure 1: Overview of the different components of ViRR, embedded in the eSciDoc infrastructure

The core of ViRR is the online editor for the creation of electronic representations of cultural artifacts. While browsing through the scans (Fig. 2), several independent working steps are supported: the semi-automatic recording of the logical pagination next to the already available physical one (Fig. 3), the gathering of the structure via building a hierarchical tree based on different structural elements like, for example introduction, chapter or paragraph (ToC editor, Fig. 4) and the assignment of corresponding scans and descriptive metadata to these structural elements (metadata editor, Fig. 4). All of these working steps are presented in one complex, but flexible workspace. This design was chosen due to different user groups (e.g. librarians, scientists) with various working methods. It allows every user to configure the editor workspace based on his focus of work by providing relevant and hiding distractive information for each working step separately. Further on, all working steps can be performed in any order or can be mixed up depending on the individual needs of the user. Created data (structure, pagination, metadata) can be published online at any time during the editing process and therefore immediately be reused by other users.

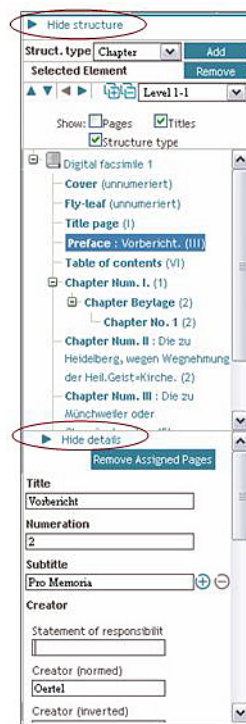
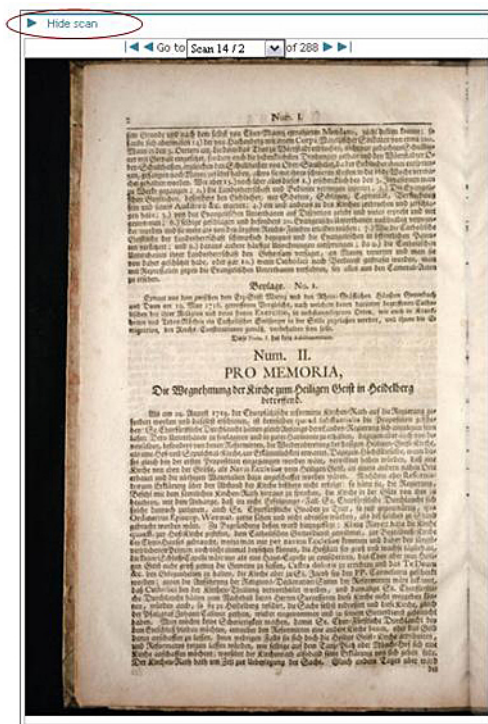


Figure 2: Parallel browsing during the editing process

Figure 3: Paginator

Figure 4: ToC and metadata editor

Within the ViRR Viewer, the content of the collection (multivolumes, volumes and monographs) is navigable via a browsing tree. Each work can be browsed separately in a configurable workspace where the user himself can decide whether he wants to see the bibliographic metadata, the logical structure in form of a table of contents or some parts of it, the scans, or a mixture of all of them. The offering of such customizable viewing sections provides each user an optimized environment to focus on his special interest.

### ViRR Collaborative Aspects

In a next step, the ViRR solution will be enhanced with a new component, the ViRR Collaborator (as presented in Fig. 1), with the aim to improve the scientific value of the digitized collections by revealing hidden semantics and relations between various disciplines.

The provision of adequate collaboration tools is especially of interest when dealing with different research contexts: investigating textual aspects focus on certain details of a collection (e.g. transcriptions or the identification of text fragments) whereas studies on visual aspects focus on e.g. high resolution scans and referencing of image parts. Others might be interested in the collection as such by e.g. browsing through the scans and investigate the metadata. The challenge is to identify the generic functionalities for annotating and sharing, and to provide a working environment adaptable to the requirements of different holdings. Different collaboration tools can be applied like graphical annotations, e.g. by integrating the enhanced viewing environment DigiLib (<http://digilib.berlios.de>), or textual annotations. Further on, transcriptions of the original text corpora will be included to improve the semantically exploitation and retrieval of the digitized works. For easy creation and quality assurance of metadata, we will aim to integrate discipline specific authority data, either stored externally or provided by the eSciDoc service CoNE (Control of Named Entities [6]).

For supporting collaborative work around different collections we would like to enable users to invite others to co-work on a collection by assigning fine granular access rights to private content.

### The eSciDoc Infrastructure

The collaborative refinements of the ViRR solution are mostly enabled by its underlying technical infrastructure. The eSciDoc infrastructure [7, 8] is designed as a service-oriented architecture. It is an open source joint development of the Max Planck Society and the FIZ Karlsruhe, funded by the German Federal Ministry of Education and Research (BMBF).

A service-oriented architecture fosters the reuse of existing services; therefore an eSciDoc service may be reused by other projects and institutions and become a building block within a broader e-Science infrastructure

[9]. The data storage system for the eSciDoc infrastructure is based on the Fedora Commons platform (<http://www.fedora-commons.org>).

The eSciDoc content model primarily consists of two generic objects called item and container. An item object, in case of ViRR, is the digital representation of a cultural artifact (e.g. scanned page) and contains metadata (such as MAB, MODS) and optionally components (such as jpeg, pdf). A container object is an aggregation of objects (items or containers) such as a journal issue which aggregates several articles. Using this content model, ViRR specializes item and container objects into volume, multivolume, monograph, ToC, and scan (see Fig. 5).

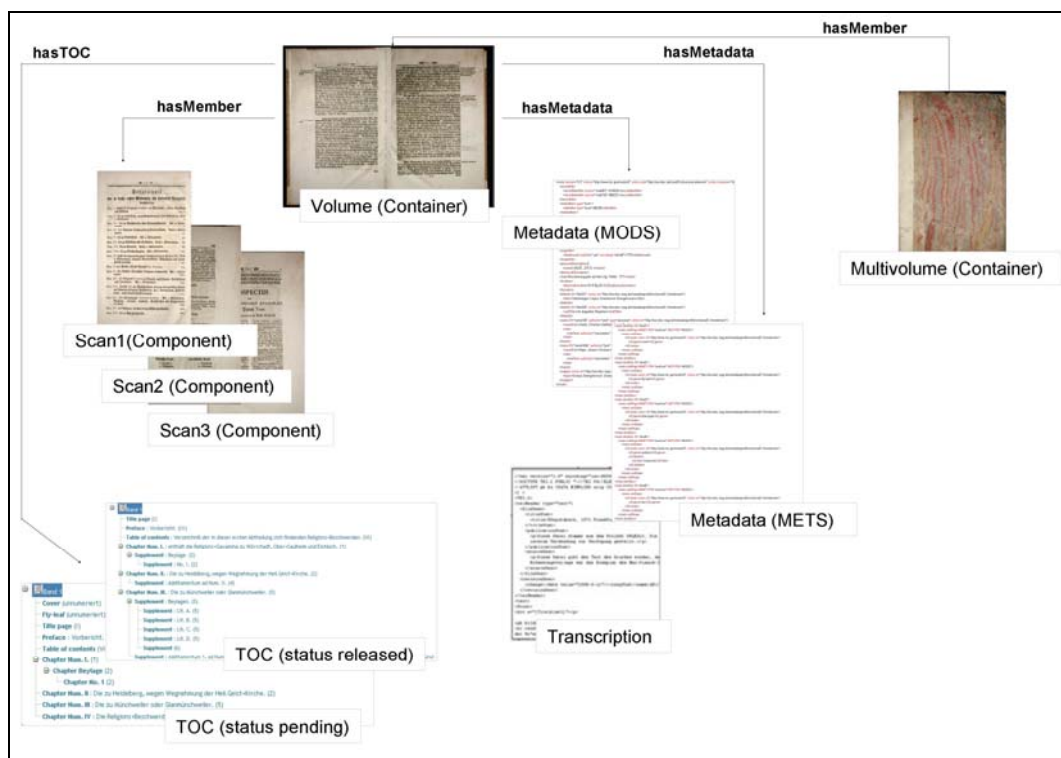


Figure 5: Digitized book content model

For example a digitized book is expressed in eSciDoc as a container, consisting of multiple items such as scans, transcriptions, and structural metadata. This container additionally holds the bibliographic metadata of the book (expressed in MODS). Additionally the generic data model of eSciDoc enables the integration and representation of data from diverse disciplines. The definition of new content models for other research data such as digitized journals or collections of images with discipline specific attributes can easily be integrated into the infrastructure, by defining a new content model with corresponding metadata profile.

As ViRR is fully embedded in the eSciDoc infrastructure it can profit from all existing eSciDoc services. Especially persistent identification (CNRI Handle or other), versioning, preservation (incl. PREMIS metadata) or the support of multiple metadata profiles (Dublin Core, MODS, custom profiles) would require, without the availability of eSciDoc, complex and time consuming development efforts for each new type of data.

eSciDoc is an open source project, setting a high priority in the implementation of standardized interfaces like oai-pmh, sword (<http://www.swordapp.org>) or RSS. Such an orientation fosters the integration of eSciDoc and eSciDoc-based solutions and their exploiting by other projects like the German national standardized viewing platform DFG Viewer (<http://dfg-viewer.de>). eSciDoc solutions are also evaluated in the context of other national or European initiatives like TextGrid (<http://www.textgrid.de>) or DARIAH (<http://www.dariah.eu>).

ViRR itself, besides offering functionality to process and disseminate data, provides as well services such as on-the-fly transformation of data. These can be used by other solutions, forming together an open accessible net of research data.

## Conclusion

A range of requirements from different research disciplines exists for the handling of digitized cultural heritage on the web. Based on our experiences, this range can not be fulfilled by a monolithic software alone. One

possibility to handle this range is to use an extensible infrastructure like eSciDoc, which focuses on standardization to support interoperability and therefore allows data exchange with services from other providers. So the data of eSciDoc solutions can be further re-used by external tools.

With the approach of using an underlying extensible infrastructure for the development of the ViRR solution, we are confident to fulfil most of the requirements arising from diverse disciplines concerning the work with digitized text resources, which is especially important in a heterogeneous research organization like the Max Planck Society.

## References

- [1] [http://colab.mpg.de/mediawiki/ViRR:\\_Virtueller\\_Raum\\_Reichsrecht](http://colab.mpg.de/mediawiki/ViRR:_Virtueller_Raum_Reichsrecht)
- [2] McDonough, J. P.: METS: standardized encoding for digital library objects. In: International Journal on Digital Libraries (2006)
- [3] Gow, J., Buchanan, G., Warwick, C., and Rimmer, J.: Document Structure in Humanities Collections. <http://www.cs.ucl.ac.uk/staff/J.Gow/papers/DocumentStructure.pdf> (accessed 08 Sept. 2009).
- [4] Bainbridge, D., Thompson, J., and Witten, I.H.: Assembling and enriching digital library collections. In: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries. IEEE Computer Society, Washington (2003) 323-334.
- [5] Hyman, M. D., and Renn, J.: Toward an Epistemic Web. Unpublished manuscript prepared for the Dahlem Workshop on Globalization of Knowledge and its Consequences. (2007) <http://archimedes.fas.harvard.edu/mdh/epistemic-web.pdf> (accessed 09 Oct. 2009)
- [6] [http://colab.mpg.de/mediawiki/Service\\_for\\_Control\\_of\\_Named\\_Entities](http://colab.mpg.de/mediawiki/Service_for_Control_of_Named_Entities)
- [7] Bulatovic, N., Tschida, U., and Gros, A.: eSciDoc - a service infrastructure for management of Cultural Heritage content. In: Proceedings of the 14th International Conference on Virtual Systems and Multimedia, (Eds.) Ioannides, M.; Addison, A.; Georgopoulos, A.; Kalisperis, L. ARCHAEOLOGIA, Budapest (2008) 138-143.
- [8] Dreyer, M., Bulatovic, N., Tschida, U., and Razum, M.: eSciDoc – a Scholarly Information and Communication Platform for the Max Planck Society. In: German e-Science Conference, Seq. No.: 315471.0 (2007).
- [9] [http://colab.mpg.de/mediawiki/ESciDoc\\_SOA\\_AtGlance](http://colab.mpg.de/mediawiki/ESciDoc_SOA_AtGlance)

**Felix ENGEL, Claus-Peter KLAS, Holger BROCKS, Alfred KRANSTEDT,  
Gerald JÄSCHKE, and Matthias HEMMJE**

**Towards supporting context-oriented information retrieval in a scientific-archive based information lifecycle**

**Abstract**

Supporting access to archived scientific publications, supplementary data, and multimedia objects as a basis for various types of reuse in scientific work processes and in publication processes is still an open issue in many ways. Reuse comprises, for instance, the subsequent verification of the content or its exploitation with a novel purpose. Retrieval approaches that factor in the versatile context of the archived data and documents can contribute to supporting reuse beyond traditional indexed based retrieval. The capturing of additional metadata during all life phases of digital objects before, during and after archival is a prerequisite to this approach. This paper motivates the usage of captured context data of digital objects for the purpose of enabling efficient reuse of preserved digital objects.

**Keywords:** OAIS, IR, context, scientific publishing

**Introduction**

An important goal of Digital Preservation (DP) is to enable the reuse of digital content. Reuse of digital content covers its subsequent verification and its exploitation with a novel purpose. Understanding the nature of the digital content and its origin supports information seekers in identifying relevant elements in archive collections and in interpreting them correctly. But the preservation of digital content, especially in the long term, covers periods of time, during which the nature of digital resources as well as their usage settings change [10]. As consumers cannot refer back to the creators, reuse of preserved digital objects depends on proper descriptions provided through the archive.

The SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg) project, co-funded by the European Commission under the seventh RTD Framework Programme, aims to develop a next generation digital preservation framework. The context model developed within the project provides an infrastructure-independent representation of the attributes associated with and (implied) relations between digital objects. Provided that the archive manages, preserves and makes available context data about digital objects, the SHAMAN context model is a potentially invaluable source for context-oriented retrieval on the archive holdings.

For SHAMAN context is not only defined by the discrete digital objects themselves, but also by the processes, in which they were created, ingested, accessed and reused. Processes are organized along phases within an Information Life Cycle Model. From an archive-centric perspective, each phase identifies one distinct stage in the life cycle of digital objects.

Context comprises information about the preserved object itself, but also the relations between objects. Hence, capturing of contextual data is of great interest for enabling advanced retrieval in archival access, in addition to supporting preservation actions. Through the preparation of context data the retrieval is not restricted on full text index, but could be opened to retrieval approaches on relations between objects. The retrieval results in this case are not necessarily preserved objects but could also be sets of contextual data.

This paper motivates the approach of context oriented Information Retrieval (IR), in an archive based life cycle with a focus on scientific publishing. This comprises current context oriented approaches in IR as well as the current approach towards a Context Model and an Information Life Cycle Phases Model in the SHAMAN project.

**Contextual Data in the Domain of Scientific Publishing**

Today, scientific publications are expected to be by origin born digital. They are presented and discussed at conferences and preserved over time in archives. Conferences take a prominent part in scientific research, because they are used to present works and ideas, to discuss new products, to determine trends, to socialize, and to initiate co-operation and collaboration. Conferences pay special attention to the assembling and the provision of scientific contributions.

Publications document the scientific contributions of the conference. Publications take various forms with individual strengths and weaknesses in distribution, storage, capacity and access capabilities. The abstract book documents the scientific contributions of the conference. Traditionally printed, abstract books nowadays are distributed as net publications. The conference web site allows for interactive structured access to the



abstracts along date and time, type of presentation, topic and presenter, embedded in the overall scientific program of the conference.

Data collections incurred and managed in the course of one conference and/or consecutive editions of one conference features heterogeneous material with metadata and multitude of relationships. Entities in a collection comprise amongst others, abstracts, papers, posters, presentations, authors, sessions, and topics. Data and document collections comprise two general types: self-contained documents that can be considered complete and well-established (for example, presentation slides, posters, the printed conference abstract book), and the multitude of data, texts, images, and document parts gathered or produced in the course of the conference. This material includes amongst others, organizational data including conference participants, presenters, events, sessions, talks, and topics, as well as structured text information from conference contributions, especially abstracts and their tables and figures.

For a conference scientific contributions are accepted, indexed and re-viewed. Speakers get invited, a scientific program is set-up, categorized and linked thematically.

### Information Life Cycle Model

Context data of digital objects evolve in different phases of existence. Context is guided by the processes in which the digital object is created, preserved, accessed and reused. Today, archives often depend on deriving metadata from the digital object obtained from the producer together with a minimum metadata set called-in by the archive. A good share of the imprint of the digital object gets lost during its transit into the archive. Opening-up the context of digital objects requires the capturing of context during all life phases of the digital object. Those life phases of a digital object are modeled in the archive-centric Information Life Cycle Model, depicted in Figure 1. The model distinguishes five relevant phases:

- Creation: new information comes into existence.
- Assembly: denotes the appraisal of objects relevant for archival and all processing and enrichment for compiling the complete information set to be sent into the future, meeting the presumed needs of the Designated Community. Assembly requires in-depth knowledge about the Designated Community in order to determine objects relevant for long-term preservation together with information about the object required for identification and reuse some time later in the future.
- Archival: addresses the life-time of the object inside the archive.
- Adoption: encompasses all processes by which accessed archival packages are unpacked, examined, adapted, transformed, integrated and displayed to be usable and understandable for the consumer. This includes also emulation activities if needed. The adoption phase might be regarded as a mediation phase, comprising transformations, aggregations, contextualisations, and other processes required for re-purposing data.
- Reuse: means the exploitation of information by the consumer. In particular, reuse may be for purposes other than those for which the Digital Object was originally created. Reuse of Digital Objects can lead to the Creation of other, novel Digital Objects. Reuse also may instigate the addition or updating of metadata about the Digital Object held in the archive. For example, annotation changes informational content and affects the relationships existing between the Object and other Digital Objects.

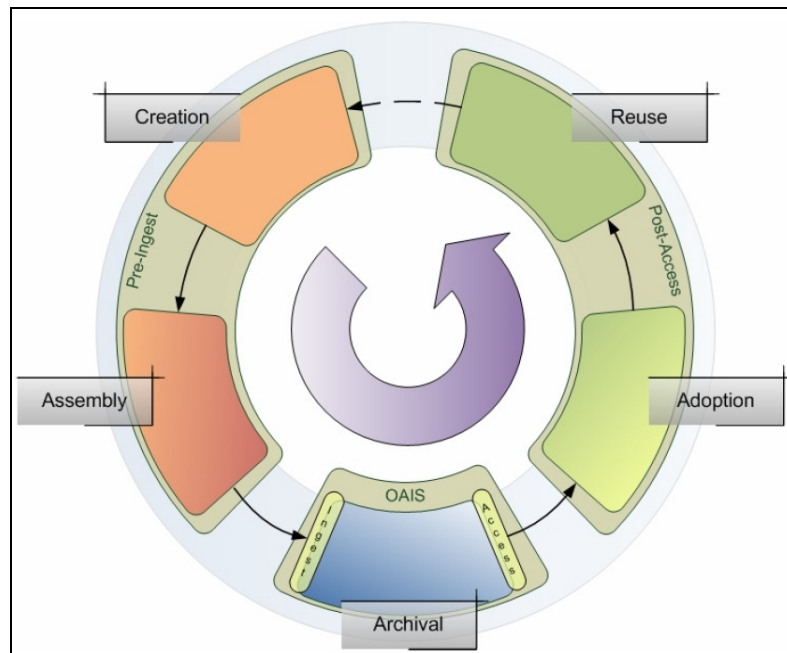


Figure 1: Information Life Cycle Phases

## Context and its Representation

The pursued approach regarding context is Digital Preservation (DP) centric, following the Open Archival Information Systems (OAIS) Reference Model [1]. OAIS is a framework of terms and concepts providing a standardization of archival systems.

Context accords to the interrelated conditions in which something exists or occurs [3]. This expresses generally what all context definitions have in common. This statement implies for digital resource management, that the context of a digital object is complex, possibly containing concepts which are shared with other objects. This might be the process environment in which they are created, the associated actors, resources and information objects and also the preservation environment in which they are stored.

Furthermore, different domains and different scenarios have different requirements towards a context definition. Currently six content components are distinct in the context approach of the SHAMAN project: Document Context, Production and Reuse Context, Preservation System Context, Modeling Change Context, Social- and Enactment Context.

The most important context component for scientific publishing is the Production and Reuse Context (PRC). This context component corresponds to the producer and (anticipated) consumer environment, i.e. the respective designated communities creating and accessing digital objects. The creation environment includes the actors and resources involved, but also a formal representation of the organizational and technical processes carried out in the production of a digital object. To re-trace information paths, the representation of the production context has to be maintained during the transition from the production into the preservation environment. The reuse of preserved digital objects depends on a proper description of the significant properties and the associated domain-specific knowledge. This description allows for the efficient access and usage even from outside the designated community.

Especially in the PRC it is obvious that context is not only defined by the digital objects themselves, but it is also defined by the processes, in which they were created, preserved, accessed and reused. Domain-specific groundings provide interfaces to the relevant concepts and topics of the designated communities addressed, in addition to formalizations of the organizational structures involved, including associated role assignments. Concluding from this three distinct concepts are encountered, which are strongly involved in defining context. Those are:

- Domain: the concepts specific to the domain and their relations. For instance in the domain of scientific publishing: Abstract, Abstract Book, Presentation or Supplement.
- Enterprise: the structural layout of an organizational environment. For instance in the domain of scientific publishing: Affiliation, Persons or Roles.

- Process: the processes and their associated activities, including information about their implementations (service invocations): Submission, Indexing or Reviewing.

If context data should be preserved over time, a model for representation and organization of data is required. As a structured representation form of concepts and their relations, the usage of ontology is appropriate. An ontology represents concepts and their relations to one another. This could be seen as a formal model of a specific domain (see e.g. [5]). Ontologies are used to establish a common understanding about knowledge existing within a domain. One important aspect of ontologies is that they formally express the semantics of each element contained, enabling individuals and machines alike to access and process the knowledge represented. Rules and inference (or reasoning) mechanisms can be employed to derive new insights, i.e. making so far implicitly existing knowledge explicit.

The ontology used in SHAMAN is conceptually structured in the three sections Domain Ontology, Enterprise Ontology and Process Ontology. Those ontologies are consolidated through the ABC ontology [8], which was formally developed to model resources and their spatial, temporal, structural and semantic relationships.

### Context-oriented Information Retrieval

Basing on the context notion and the representation of context as described in the previous sections, retrieval could be extended in two ways: firstly through the creation of an additional full-text index, containing the indexed context data and secondly through retrieval mechanisms on base of the relations between archived objects. Such relations between archived objects evolve through similar context attributes values. Those attribute values in the domain of scientific publishing are for instance the same author, the same conference, the same reviewer or common keywords. These data should be accessible through query, browsing with visualization support. The result of such a context oriented query is then not restricted on the archived objects; rather this could be a set of context data.

Context data in the domain of scientific publishing which can be expected to aid the retrieval of relevant publications for the purpose of scientific reuse are, for example

- Representation types such as abstract, presentation slides, poster or full paper;
- Embedding in the world of scientific discourse along citation nets, roles, interest and competence profiles of persons and organizations, and discussion threads;
- Implicit and explicit relationships to other documents like review reports and conference reports.

Those data could support, for instance, the retrieval of information for a state of the art research. Once a first relevant publication was found it could be used as the starting point to search for similar publications. Similarities according to publication context are, for example, but not limited to: publication origination from conferences with similar subject focus, by origination from the same conference, its conference sessions, its tutorials or its keynotes. Furthermore, it could be valuable to find publications with the same key words, publications which are referred to the source publication or the publications that refer to the source publication. Different approaches for defining the concept context exist in IR. A user centric approach has been done for instance by Järvelin et al. in [6]. They stated that context is given through dependencies in time, place, history of interaction, task at hand and some other factors. Another approach towards a context definition in IR has been outlined by Cool et al. in [2]. They classify IR context in four different levels, namely: information environment, information seeking, IR interaction and the query level.

Some conceptual and implementation work on context based IR is already done. Melucci for instance presents in [9] a context model and the application of the model for ranking. Some context based IR support tools are implemented in Daffodil. This is an experimental system for IR and collaborative services in the field of higher education for the domain of computer science and others [4]. Daffodil comprises, for instance, an Author Net, which depicts relations among authors stored in a database and is used for ranking and the search for central actors in a set of documents or central actors for a specific author. Daffodil furthermore implements a Citation and Co-Author Browser, which are similar the Author Net, as well as an adaptive suggestion tool, which is based on the current situational user context [7].

But even if some particular solutions towards context oriented retrieval are implemented yet, the retrieval in preservation systems access lacks of offering a holistic model of digital object context for different domains and the preparation of context data for usage in retrieval.

For such a context oriented IR process it is essential to:

- define a holistic and adaptive context model, in order to serve the requirements of different domains
- provide mechanisms to capture relevant context data during the ingest phase
- prepare the context data in order to make them usable for retrieval
- offer an appropriate query- or browsing format in order to query the context data
- offer an appropriate way for presentation

The support of all those requirements is a task for future scientific work.

## Conclusion

In this paper the advantage towards context oriented IR in an archive information life cycle is motivated. A context notion on basis of ontology is presented in order to model the context of preserved digital content. The ontology based representation provides valuable additional information for IR through the description of relations. By means of the archive-centric information life cycle model, the important phases for capturing context are presented. The domain of scientific publishing was used to illustrate the usage of this retrieval approach.

## References

- [1] CCSDS. Reference Model for an Open Archival Information System (OAIS). Blue Book 1, Consultative Committee for Space Data Systems, January 2002. Recommendation for Space Data Systems Standards, adopted as ISO 14721:2003.
- [2] Colleen Cool and Amanda Spink. Issues of context in information retrieval (IR): an introduction to the special issue. *Information Processing Management*, 38(5):605–611, 2002.
- [3] Merriam-Webster Online Dictionary. context; cited 30.04.2009. "ONLINE" <http://www.merriam-webster.com/dictionary/context>.
- [4] Norbert Fuhr, Claus-Peter Klas, André Schaefer, and Peter Mutschke. Daffodil: An Integrated Desktop for Supporting High-Level Search Activities in Federated Digital Libraries. In *Research and Advanced Technology for Digital Libraries. 6th European Conference, ECDL 2002*, pages 597–612. Springer, 2002.
- [5] Nicola Guarino. Formal ontology and information systems. In Nicola Guarino, editor, *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98*, pages 3–15, 1998.
- [6] P. Järvelin, K. & Ingwersen. Information seeking research needs extension towards tasks and technology. "ONLINE" <http://InformationR.net/ir/10-1/paper212.html>, 2004.
- [7] Claus-Peter Klas, Sascha Kriewel, and Matthias Hemmje. An Experimental System for Adaptive Services in Information Retrieval. In *Proceedings of the 2nd International Workshop on Adaptive Information Retrieval (AIR 2008)*, October 2008.
- [8] Carl Lagoze and Jane Hunter. The ABC Ontology and Model. In *Dublin Core Conference*, pages 160–176, 2001.
- [9] Massimo Melucci. A basis for information retrieval in context. *ACM Trans. Inf. Syst.*, 26(3):1–41, June 2008.
- [10] Ute Schwens and Hans Liegmann. Langzeitarchivierung digitaler Ressourcen. In Rainer Kuhlen, Thomas Seeger, and Dietmar Strauch, editors, *Handbuch zur Einführung in die Informationswissenschaft und -praxis*, volume 1 of *Grundlagen der praktischen Information und Dokumentation*, chapter D9, pages 567 – 570. München : Saur, 5., völlig neu gefasste Ausgabe. edition, 2004.

**Maurizio LANCIA, Brunella SEBASTIANI, Roberto PUCCINELLI, Marco SPASIANO, Massimiliano SACCONI, Luciana TRUFELLI, Emanuele BELLINI, Chiara CIRINNÀ, Maurizio LUNGI**

**Towards a European global resolver service of persistent identifiers**

**Abstract**

In this paper we present the Italian initiative that involves relevant research institutions and national libraries, aimed at implementing an NBN Persistent Identifiers (PI) infrastructure based on a novel hardware/software architecture. This solution can be the base infrastructure towards the implementation of the European Global Resolver Service of PI.

The proposal is about a distributed and hierarchical approach for the management of an NBN namespace and illustrates assignment policies and identifier resolution strategies based on request forwarding mechanisms. Starting from the core motivations for the assignment of “persistent identifiers” to digital objects, this paper outlines a state of art in PI technologies, standards and initiatives, and illustrates other NBN implementations. The structure and goals of our initiative are described as well as the features already implemented in our system and the results of our testing activities.

The paper ends with a proposal for the extension of this approach to the EU scenario.

**Introduction**

Stable and certified references to Internet resources are crucial for all the digital library applications, not only to identify a resource in a trustable and certified way, but also to guarantee continuous access to it over time. Current initiatives like the European Digital Library (EDL) [1] and Europeana [2], clearly show the need for a certified and stable digital resource reference mechanism in the cultural and scientific domains. The lack of confidence in digital resource reliability hinders the use of the Digital Library as a platform for preservation, research, citation and dissemination of digital contents [15]. A trustworthy solution is to associate to any digital resource of interest a PI that certifies its authenticity and ensures its long term accessibility. Actually some technological proposals are available [24], but the current scenario shows that we can't expect/impose a unique PI technology or only one central registry for the entire world. Moreover, different user communities do not commonly agree about the granularity of what an identifier should point to.

In the Library domain the National Bibliography Number (NBN – RFC3188) has been defined and is currently promoted by the CENL. This standard identifier format assumes that the national libraries are responsible for the national name registers. The first implementations of NBN registers in Europe are available at the German and Swedish National Libraries.

In Italy we are currently developing a novel NBN architecture with a strong participation of the scientific community, led by the National Research Council (CNR) through its Central Library and ITC Service. We have designed a hierarchical distributed system, similar to the DNS, in order to overcome the criticalities of a centralised system and to reduce the high management costs implied by a unique resolution service. Before describing our system in detail, we will provide in the following sections an overview of available PI technologies.

**Persistent Identifier standards**

The association of a PI to a digital resource can be used to certify its content authenticity, provenance, managing rights, and to provide an actual locator. The only guarantee of the actual persistence of identifier systems is the commitment shown by the organizations that assign, manage, and resolve the identifiers [25], [26].

At present some technological solutions are available but no general agreement has been reached among the different user communities. We provide in the following a brief description for the most widely diffused ones. Only the NBN [3] standard will be described in details in the next section.

The Document Object Identifier system (DOI [11]) is a business-oriented solution widely adopted by the publishing industry, which provides administrative tools and a Digital Right Management System (DRM).

Archival Resource Key (ARK [10]) is an URL-based persistent identification standard, which provides peculiar functionalities that are not featured by the other PI schemata, e.g., the capability of separating the univocal identifier assigned to a resource from the potentially multiple addresses that may act as a proxy to the final resource.

The Handle System ([12], [26], [27]) is a technology specification for assigning, managing, and resolving persistent identifiers for digital objects and other resources on the Internet. The protocols specified enable a distributed computer system to store identifiers (names, or handles) of digital resources and resolve those handles into the information necessary to locate, access, and otherwise make use of the resources. That information can be changed as needed to reflect the current state and/or location of the identified resource without changing the handle.

Finally, the Persistent URL (PURL [13]) is simply a redirect-table of URLs and it's up to the system-manager to implement policies for authenticity, rights, trustability, while the Library of Congress Control Number (LCCN [14]) is the a persistent identifier system with an associated permanent URL service (the LCCN permanent service), which is similar to PURL but with a reliable policy regarding identifier trustability and stability.

This overview shows that it is not viable to impose a unique PI technology and that the success of the solution is related to the credibility of the institution that promotes it. Moreover the granularity of the objects that the persistent identifiers need to be assigned to is widely different in each user application sector.

### **NBN overview**

The National Bibliography Number (NBN) [3] is a URN namespace under the responsibility of National Libraries. The NBN namespace, as a Namespace Identifier (NID), has been registered and adopted by the Nordic Metadata Projects upon request of the CDNL and CENL. Unlike URLs, URNs are not directly actionable (browsers generally do not know what to do with a URN), because they have no associated global infrastructure that enables resolution (such as the DNS supporting URL). Although several implementations have been made, each proposing its own means for resolution through the use of plug-ins or proxy servers, an infrastructure that enables large-scale resolution has not been implemented. Moreover each URN name-domain is isolated from other systems and, in particular, the resolution service is specific (and different) for each domain.

Each National Library uses its own NBN string, independently and separately implemented by individual systems, with no coordination with other national libraries and no commonly agreed formats. In fact, several national libraries have developed their own NBN systems for national and international research projects; several implementations are currently in use, each with different metadata descriptions or granularity levels.

Examples are the DIVA project [16], EPICUR [18], and ARK at National Library of France [17].

There are some important initiatives at European level like the TEL project that it is in the process of implementing a unique system based on NBN namespace within the European Digital Library (EDL). The adoption of NBN identifiers is needed for implementing the 'National Libraries Resolver Discovery Service' as described in the CENL Task Force on Persistent Identifiers report [19].

In our opinion NBN is a credible candidate technology for an international and open PI infrastructure, mainly because it is based on an open standard and supports the distribution of the responsibility for the different subnamespaces, thus allowing the single institutions to keep control over the persistent identifiers assigned to their resources.

### **The NBN initiative in Italy**

The project for the development of an Italian NBN register/resolver started in 2007 as a collaboration between "Fondazione Rinascimento Digitale" (FRD), the National Library in Florence (BNCF), the University of Milan (UNIMI) and "Consorzio Interuniversitario Lombardo per l'elaborazione automatica" (CILEA). After one year of work a first prototype demonstrating the viability of the hierarchical approach was released. The prototype leveraged some features of DSpace and Ark and provided a basic PHP web interface for library operators and final users. The hierarchy was limited to a maximum of two levels.

The second and current phase of the Italian NBN initiative is based on a different partnership involving Agenzia Spaziale Italiana (ASI), Consiglio Nazionale delle Ricerche (CNR), Biblioteca Nazionale Centrale di Firenze (BNCF), Biblioteca Nazionale Centrale di Roma (BNCR), Istituto Centrale per il Catalogo Unico (ICCU), Fondazione Rinascimento Digitale (FRD) and Università di Milano (UniMi).

The Italian National Research Council (CNR) developed a second prototype based on Java Enterprise technologies and web 2.0 user interface, which eliminated the need for DSpace and Ark and the two-level limit and introduced new features. CNR and FRD hold property rights of the software and will release it as opens source under the terms of EUPL license. In order to encourage its adoption by other national registers a supporting community will be established.

The results are available as an installable software; future objectives have been defined in order to extend functionality and integrate the system within an international infrastructure. To this end, the Italian group is currently establishing international collaborations.

In the following we provide a description of objectives, governing structure and licensing policy defined for the

Italian initiative.

The initiative aims at:

- 1) creating a national stable, trustable and certified register of digital objects to be adopted by cultural and educational institutions;
- 2) allowing an easier and wider access to the digital resources produced by Italian cultural institutions, including material digitised or not yet published;
- 3) encouraging the adoption of long term preservation policies by making service costs and responsibilities more sustainable, while preserving the institutional workflow of digital publishing procedures;
- 4) implementing a new service based on URN, similar to other national systems but with a more advanced architecture in order to achieve distribution of responsibility for name management;
- 5) extending as much as possible the adoption of the NBN technology and the user network in Italy;
- 6) developing an inter-domain resolution service (e.g., NBN Italy and NBN Germany, or NBN Italy and DOI) with a common meta-data format and a user-friendly interface (pre-condition for global resolver);
- 7) creating some redundant mechanisms both for duplication of name-registers and in some cases also for the digital resources themselves;
- 8) overcoming the limitation imposed by a centralised system and distributing the high management costs implied by a unique resolution service, while preserving the authoritative control.

In order to define organization and policies for the Italian register, a governing board has been established, where BNCF, BNCR, CNR, FRD, ICCU are represented. The governing board defines the top-level structure of the Italian NBN domain hierarchy and the policies for overall infrastructure management, sub-domain creation/removal and PI assignment.

### **The distributed architecture approach**

The proposed architecture, starting from [22], [23] and taking into account the URN standard requirements as [20], [21], introduces some elements of flexibility and additional features as shown in [29]. At the highest level there is a root node, which is responsible for the top-level domain (IT in our case). The root node delegates the responsibility for the different second-level domains (e.g.: IT:UR for University and Research, etc.) to second-level naming authorities. Sub-domain responsibility can be further delegated using a virtually unlimited number of sub-levels (eg.: IT:UR:CNR, IT:UR:UNIMI, etc.). At the bottom of this hierarchy there are the leaf nodes, which are the only ones that harvest publication metadata from the actual repositories and assign unique identifiers to digital objects.

Each agency adheres to the policy defined by the parent node and consistently defines the policies its child nodes must adhere to.

It is easy to see that this hierarchical multi-level distributed approach implies that the responsibility of PI generation and resolution can be recursively delegated to lower level sub-naming authorities, each managing a portion of the domain name space. Given the similarity of the addressed problems, some ideas have been borrowed from the DNS service.

Within our architecture each node harvests PI information from its child nodes and it is able to directly resolve all identifiers belonging to its domain and sub-domains. Besides, it can query other nodes to resolve NBN identifiers not belonging to its domain. This implies that every node can resolve every NBN item generated within the NBN:IT subnamespace, either by looking up its own tables or by querying other nodes. In the latter case the query result is cached locally in order to speed up subsequent interrogations regarding the same identifier.

This redundancy of service access points and information storage locations increases the reliability of the whole infrastructure by eliminating single points of failure. Besides, reliability increases as the number of joining institutions grows up.

In our opinion a distributed architecture also increases scalability and performance, while maintaining unaltered the publishing workflows defined for the different repositories.

### **Policy**

The trustability and reliability of an NBN distributed infrastructure can be guaranteed only by defining and enforcing effective policies. To this end the Italian NBN governing board is going to release a general policy that will have to be signed by all the participating agencies.

We have performed an initial analysis to detect problems and issues that the policy should address. In our opinion each agency should satisfy some requirements, which are both technical and organisational, and should commit in respecting some guidelines.

### *Organisational requirements*

Each participating agency should indicate an administrative reference person, who is responsible for policy compliance as regards the registration and resolving procedures as well as for the relationships with the upper and lower level agencies, and a technical reference person, who is responsible for the hardware, software and network infrastructure.

### *Technical requirements*

The hardware hosting an NBN register/resolver should be housed in a managed hosting infrastructure, with uninterruptable power supply and high-speed network connection. An agency that does not have an internal server farm may outsource hosting services to an external provider, which fulfils the technical requirements.

The hardware architecture should be redundant in order to guarantee no single point of failure.

In our opinion it would be also useful to identify and monitor some simple service level indicators, such as service response time and up time, and define thresholds that each agency should respect. Each domain maintainer could monitor its child sub-domains and notify them service level violations. The policy should also define how violations should be dealt with.

### *Guidelines*

The policy should define rules for:

- 1) generating well-formed PIs;
- 2) identifying the digital resources which “deserve” a PI;
- 3) identifying resource granularity for PI assignment (paper, paper section, book, book chapter, etc.)
- 4) auditing repositories in order to assess their weaknesses and their strengths (the Drambora toolkit may help in this area).

### **Testing activities**

After developing a first working prototype, collaborations have been established with several research institutions in order to create a community where final users and software developers are both represented. Several institutions are already involved in user requirement definition or have declared their availability to join the NBN network. These institutions are: the University & Research Group (ISS, INAF, INFN, INGV, ASI, ENEA, INOA, APAT, University of Pisa, University of Rome ‘Sapienza’, the University of Florence, the Florence University Press, University of Milan, i.e.).

A first testbed has been deployed where users can execute test cases and provide feedback to the developers in terms of bug/defect notifications, change or enhancement requests and new requirements. On the other hand the developers perform technical tests to evaluate performance, scalability and reliability of the infrastructure and implement what needed to satisfy user indications.

The testbed is configured as follows:

- a) central node at BNCF, responsible for the Italian sub-domain (NBN:IT),
- b) a second level inner node at CNR, responsible for the “University and Research” sub-domain (NBN:IT:UR),
- c) a second level leaf node at FRD, responsible for the local NBN:IT:FRD sub-domain,
- d) a third level leaf node at UNIMI, responsible for the local NBN:IT:UR:UNIMI sub-domain,
- e) a third level leaf node at CNR, responsible for the local NBN:IT:UR:CNR sub-domain.

The second level CNR inner node (NBN:IT:UR) aims at implementing the University and Research National Registry. It currently aggregates the records generated by the UNIMI and CNR leaf nodes for the resources stored in their local repositories. The FRD node generates NBNs for resources stored in a local Dspace repository. A first set of tests has been performed to verify functionalities and behaviour in a distributed environment using different metadata sets.

Performance was not the main focus in this phase and this is the reason why the servers used to set up the infrastructure are neither particularly powerful nor up to date.

First feedbacks from users are positive as regards registering and resolving functionalities. The system harvests resources, assigns NBNs and provides access to metadata and documents as expected. As regards duplicate discovery via hash comparisons, it has been pointed out that this mechanism works only if the compared files are identical, but fails even if they differ for a single bit. It has also been remarked that currently it is not possible to represent within the identifier the “part of” relation between two digital objects. This means that if we want to assign identifiers both to an entire document and to parts of it (e.g. a picture) there is currently no commonly agreed way to represent this inclusion relation in the final part of the persistent identifier. Finally, the need for higher-level services has been expressed by several parties, first of all the possibility of producing reports about the number of publications deposited in a sub domain within a certain period. This problem is tightly related to the duplicate detection one. If the latter is not solved, resource accounting statistics may be



affected by errors whose impact cannot be estimated at the moment.

### **Towards the European Resolution Service**

In this paper we have described a new software application for a distributed and hierarchical NBN register/resolver infrastructure. The main technical problems pointed out so far pertain to the identifier uniqueness guarantee. The proposed solution of using MD5 hash codes partly resolves this issue but poses performance problems and does not cover cases where the same content is represented in different formats. A more comprehensive solution will probably involve the comparison of a strictly defined set of metadata. This means that strict rules and clear responsibilities must be defined as regards data entry in the digital libraries.

From a political point of view the short-term objective is to enlarge the group of supporting institutions in order to create a first nucleus of a credible NBN national infrastructure. On a larger scale, CNR and FRD participate to the PersID project, funded by the Knowledge Exchange consortium and the SURF foundation, and aimed at developing a European Global Resolver. The adoption of our software as top-level node manager will be taken into consideration in the following months.

In our opinion it is also important to identify high-level value-added services (such as digital resource accounting) that could be built on top of the infrastructure. This would probably favour the diffusion of NBN persistent identifiers.

From the technical point of view the next steps will include performance testing and tuning, in order to define the hardware requirements for a production infrastructure that would guarantee the necessary service levels.

The testbed will be enlarged in order to include a leaf node installed at the University of Bologna, which will harvest records from the "Magazzini digitali" project repository. The goal of this project is to enable the BNCf digital library to harvest doctoral thesis from the University of Bologna Eprints repository, in order to accomplish their legal deposit. In this case the resources already have an NBN name. A new NBN record will be created in our registry using the existing identifier, which will be associated to the new URL assigned by legal deposit at BNCf.

A research group has also been established to thoroughly examine the duplication problem and its possible solutions. In this field hash codes different from MD5 could provide better performance with respect to comparison operations. The same group will also address the problem of the "part of" relation representation. Finally, we are going to investigate ways to establish permanent and reliable connections between NBNs and other persistent identifiers such as DOI, which would favour the implementation of a multi-standard global resolver.

### **References**

- [1] European Digital Library  
<http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/edproject/>
- [2] Europeana [www.europeana.eu/](http://www.europeana.eu/)
- [3] IETF RFC 3188 Using National Bibliography Numbers as Uniform Resource Names  
<http://tools.ietf.org/html/rfc3188>
- [4] IETF RFC 2141 URN Syntax <http://tools.ietf.org/html/rfc2141>
- [5] C. Lagoze and H. V. de Sompel. The Open Archives Initiative Protocol for Metadata Harvesting, version 2.0. Technical report, Open Archives Initiative, 2002.  
<http://www.openarchives.org/OAI/openarchivesprotocol.html>.
- [6] Dublin Core Metadata Initiative. Dublin Core Metadata Element Set, Version 1.1.  
<http://dublincore.org/documents/dces/>.
- [7] MPEG-21, Information Technology, Multimedia Framework, "Part 2: Digital Item Declaration," ISO/IEC 21000-2:2003, March 2003.
- [8] METS, <<http://www.loc.gov/standards/mets/>
- [9] Herbert Van de Sompel et al. Resource Harvesting within the OAI-PMH Framework D-Lib Magazine December 2004 Volume 10 Number 12 ISSN 1082-9873
- [10] J. Kunze. The ARK Persistent Identifier Scheme. Internet Draft, 2007 <http://tools.ietf.org/html/draft-kunze-ark-14>.
- [11] Norman Paskin. Digital Object Identifiers. Inf. Serv. Use, 22(2-3):97–112, 2002 <http://www.doi.org>
- [12] Sam X. Sun. Internationalization of the Handle System - A persistent Global Name Service. 1998.  
<http://citeseer.ist.psu.edu/sun98internationalization.html>. [www.handle.net](http://www.handle.net)
- [13] Persistent URL <http://purl.oclc.org>
- [14] Library of Congress Control Number <http://www.loc.gov/marc/lccn.html>
- [15] David Giaretta, Issue 1, Volume 2 | 2007 The CASPAR Approach to Digital Preservation The International Journal of Digital Curation



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

- [16] Andersson, Stefan; Hansson, Peter; Klosa, Uwe; Muller, Eva; Siira, Erik Using XML for Long-term Preservation: Experiences from the DiVA Project
- [17] Bermes, Emmanuelle, International Preservation News, Vol 40 December 2006, pp 23-26 Persistent Identifiers for Digital Resources: The experience of the National Library of France  
<http://www.ifla.org/VII/4/news/ipnn40.pdf>
- [18] Kathrin Schroeder. Persistent Identification for the Permanent Referencing of Digital Resources - The Activities of the EPICUR Project Enhanced Uniform Resource Name URN Management at Die Deutsche Bibliothek. The Serials Librarian, 49:75–87(13), 5 January 2006.
- [19] CENL Task Force on Persistent Identifiers, Report 2007  
[http://www.nlib.ee/cenl/docs/CENL\\_Taskforce\\_PI\\_Report\\_2006.pdf](http://www.nlib.ee/cenl/docs/CENL_Taskforce_PI_Report_2006.pdf).
- [20] Sollins, Karen Architectural Principles of Uniform Resource Name Resolution (RFC 2276)  
<http://www.ietf.org/rfc/rfc2276.txt>
- [21] Masinter, Larry; Sollins, Karen Functional Requirements for Uniform Resource Names (RFC 1737)  
<http://www.ietf.org/rfc/rfc1737.txt>
- [22] E. Bellini, M. Lunghi, E. Damiani, C. Fugazza, 2008, Semantics-aware Resolution of Multi-part Persistent Identifiers, WCKS 2008 conference.
- [23] E. Bellini, C. Cirinna, M. Lunghi, E. Damiani, C. Fugazza, 2008 Persistent Identifiers distributed system for cultural heritage digital objects, IPRES2008 conference
- [24] H.-W. Hilse, J. Kothe Implementing Persistent Identifiers: overview of concepts, guidelines and recommendations, 2006, ix+57 pp. 90-6984-508-3 <http://www.knaw.nl/ecpa/publ/pdf/2732.pdf>
- [25] DCC Workshop on Persistent Identifiers, 30 June – 1 July 2005 Wolfson Medical Building, University of Glasgow, <http://www.dcc.ac.uk/events/pi-2005/>
- [26] ERPANET workshop Persistent Identifiers, Thursday 17th - Friday 18th June 2004-University College Cork, Cork, Ireland, <http://www.erpanet.org/events/2004/cork/index.php>
- [27] Handle System website, <http://www.handle.net/>
- [28] Wikipedia Handle page, [http://en.wikipedia.org/wiki/Handle\\_System](http://en.wikipedia.org/wiki/Handle_System)
- [29] E. Bellini, C. Cirinnà, M. Lunghi, R. Puccinelli, M. Lancia, B. Sebastiani, M. Saccone, M. Spasiano - Persistent identifier distributed system for digital libraries - IFLA 2009 Conference – Milan



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

**Jeremy W. HUNSINGER**

## **Where did the user's go? A case study of the problems of event driven memory bank**

### **Abstract**

The April 16 Archive (<http://www.april16archive.org>) at the Center for Digital Discourse and Culture is a memory bank of user contributed digital artifacts relating to the event, the April 16 Tragedy at Virginia Tech. A memory bank collects things that people contribute to it; usually digital originals, related to something worth remembering. These digital memorabilia form an esoteric collection that before its collection would have become ephemeral and likely lost (Goff, 2008). However, as collected, they should, as they are tied directly to the act of contribution and the more significant relations beyond that between the individual and the event, have significant social and emotional ties to the contributors and the community. This paper argues that those ties are fading. The April 16 Archive, as an event driven memory bank, originated from a passionate and committed community of users who shared the emotional and social attachment surrounding the event. The paper describes the tensions in the development and maintenance of the archive as the various communities have, through time, grown farther from the event, placing it further into their communal memories. In doing this, I hope to provide insights into the problems that develops in event driven digital archives as its communities grow apart. I also hope to share some of our experiences in developing and maintaining an event driven archive using web 2.0 oriented software.

**Keywords:** Memory Bank, Audiences, Users, Archives, Web 2.0

### **Introduction**

As I sit in my office watching the tail of the web server logs scroll through my terminal window for a few minutes considering how I should start this paper, I am struck by how rarely the topic of this paper, the April16archive.org website is appearing in those logs. This observation is in part the basis for this paper. Event-driven archives, like the April 16th Archive struggle to maintain users overtime as the memories of the event fades, even if the effects of the event do not.

On April 16th, 2007, Seung-Hui Cho, shot and killed 32 people and wounded many other faculty, staff, and students Virginia Polytechnic Institute and State University (Virginia Tech) . He left wounds both physical and emotional. While I was not on campus that year, I was still affiliated with Virginia Tech, as I am today, and like many I was overwhelmingly concerned for the safety and wellbeing of my colleagues at Virginia Tech. This event was a shock across the global university system and has had broad ranging effects from changing how universities deal with mentally ill students, to how universities manage security, and how we communicate with students, families, faculty, staff, and the greater audience. In short, this event was tragic and transformative, changing the ways universities and higher education operates in many ways.

Within a few days of the event, my colleague Brent Jesiek, who was managing the Center for Digital Discourse and Culture (CDDC) and now is an Assistant Professor at Purdue University, met with colleagues on campus from departments in the social science and humanities and they discussed options for preserving elements of the event that might be overlooked or uncollected elsewhere. Their idea was to move beyond the archival mission and into the memory mission, even toward a memorial mission, which would enable more personal and shared narratives, such as podcasts, blog posts and similar media to be captured as part of a memory bank like occurred for Hurricane Katrina and September 11th with their respective memory banks . Unlike a normal archive, the idea for the April16archive was to be more expansive:

“This project contributes to the ongoing efforts of historians and archivists to preserve the record of this event by collecting first-hand accounts, on-scene images, blog postings, and podcasts. It is our sincere hope that this site can contribute to a collective process of healing, especially as those affected by this tragedy tell their stories in their own words. The April 16 Archive runs on Omeka, a "digital memory bank" platform that uses the Internet to preserve the past and make memories available to a wide audience for generations to come”.  
(<http://april16archive.org/about>)

Contacting colleagues George Mason University's Center for New Media and History (CNMH), Prof. Jesiek discussed the memory banks for hurricane Katrina and Sept. 11 and inquired as to the nature of the software used. The CNMH had software in development that was to become Omeka that they contributed for the CDDC to use on this project. Omeka was still in development at that stage, but one of the CNMH developers had recently graduated from Virginia Tech's history program, between his efforts and the rest of the CNMH staff, within a few days, the software to launch the April16tharchive was in Prof. Jesiek's email inbox. Eight days after the tragic events of April 16, on April 24th, the first objects from the general public began to be donated to

the archive. On April 30th, the College of Liberal Arts and Human Sciences did a formal press release announcing this new memory bank to the [1].

Omeka is a memory bank application that allows the collection and discussion of digital collections in a web 2.0 environment. It is the platform that CDDC uses to host the April16archive today. Omeka is an object-oriented php/mysql web-based software application that can easily be installed and used for a variety of purposes. One such purpose is the memory bank. A memory bank attempts to collect and preserve memories--contributed objects such as images, videos, even word documents. Almost all the material contributed are ephemera, which might not be otherwise preserved were they not found, made, or even constructed to be preserved by their contributors.

### Ephemera, Events, Memories and Audiences


In the April16archive, we focus on one event and this has dramatic effects in the two plus years that we have been in existence. As an event-driven archive--an archive that documents a temporal event that occurred at a certain time--the April16archive faces the challenge of sustaining a group of users or even a strong audience. In the beginning of the archive this issue was not foreseen. In November of 2007, we presented at the 48th annual Rare Books and Manuscripts section of the ACRL's Collecting for Contemporary Events seminar at Johns Hopkins University, where reported that at that time we had almost 1000 items in our digital collection, and today we have just over 1200, though we have secondary collections like the April 16th Archive Frontpages collection, which adds to that number. In addition, in the last year we have had few, actually very few, contributions to the archive that were not created in house. The drop off in contributions was matched by an increase in spam, which we eventually controlled. However, new contributions are occurring in the order of 1 or 2 every 3-5 months. In short, it appears as if, as one might expect that the memory of the event has faded and with that fading of memory, there has been a fading of the number of contributions.

Contributions to the archive are all similar to the image below:

## Memorial After Remembrance

View of the April 16 memorial with Burruss Hall in the background. Photo taken April 17, 2008.

Licensed under [Creative Commons Attribution-NonCommercial-ShareAlike 3.0](#)



Tags: [memorial](#), [remembrance](#), [anniversary](#)

## Citation Information

Brent Jesiek, "Memorial After Remembrance." *The April 16 Archive*, Item #2643 (accessed October 14 2009, 1:14 pm)

Screenshot of archive material donated under Creative Commons License

In this image we can see all of the user-contributed content of a contribution to the archive. The author provides title, description, license information (if appropriate), tags and the content itself, which in this case is a large picture of which the smaller thumbnail is represented here. The contents of the archive are entirely searchable by author, tags, and almost any conceivable search, such as date, etc. The system provides a clear APA citation for the material, which is useful for future users.

Users are a perpetual question with the archive, while there are interesting academic politics surrounding the archive, its contents, and who can and should use them for what purposes, the real issue is less those politics than the lack of use and users in general. From the initial set of content creators and contributors as indicated

in earlier discussion, now we are basically only receiving spam, which we filter. The material growth of the archive is minimal in the last year, and without continued effort and as such funding, we doubt it will have any more substantial growth. The competition amongst various parties about who does what with what parts of which archive, especially the tensions between research, memorial and archival missions highlight the differences of users of the current and future archive [2]

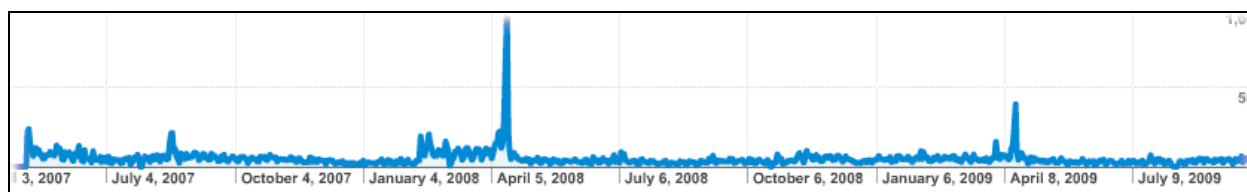
If we think of three sets of users of the archive being contributors/memorialists, researchers including news reporters, and archivists, there seems to be tensions between what they want and what the archive provides. The archive only contains the material contributed and other than two objects, it shows all of that contributed material to the public at any given time. However, the availability of the material does not make it useful to the university researcher who wants to use the material as representing individuals instead of as documentary material, as we did not request rights to use the material to research the creators from the creators. However, over time, the distance between creators and material is fading, much like the memory is fading as described above, and with that fade, the human subjects issues are also fading.

Fading memories and forgetting is quite normal in regard to such tragic events. Forgetting as a social and political process is important for the reconstitution of subjectivities and social, political relations [3][2].

### Fading relations, Fading Memories, Fading Interests

The question that drives this paper and the descriptions so far has been, "what is this archive?", but the question that this paper answers is, "What happened to the users of this archive?". To answer that question, we will inspect and describe the phenomena of our users as represented in the Google Analytics (TM). We use Google Analytics (TM) for user data because they strip out 99% of the bot and otherwise inhuman access, representing the human, and thus representing the marketable to Google, access to the page. By leveraging this tool, I was able to develop reports at several levels of analysis comparing the 2007, 2008, and 2009 data sets available.

The first data object is on the one hand the most revelatory, and on the other hand, the least data intensive to understand. From the date we turned on Google Analytics (TM) for this in May 2007, the data is fairly consistent, with a few peaks, such as the start of school in Fall 2007, the anniversary of the event, the Tragedy of Northern Illinois University on Feb. 14 2008 and the second anniversary of Northern Illinois and third anniversary of the Virginia Tech Tragedy. The peaks indicate high points of use. Correlating with these reference peaks there are contacts from media and other researchers about some topic via email and phone. We can see that other than on such peak events, the number of visits to the archive is relatively low, averaging below 100 on any given day for the past few years, though it has been decreasing over time, which is clear from the data representations.



May 3rd 2007 to August 2009 user visits report from Google Analytics(TM)

Investigating the peaks reveals a bit more below, I have representations of two peaks. The peaks represent the second and third anniversaries of the April 16 archive. These two peaks provide insights particular to the archive's significant event. In the 2008, the first anniversary, we can see fairly interesting behavior for an internet archive, the use of the archive for the month of April was 4546 visits with around 3 minutes and 42 seconds per visits with 33,345 page views and only a 44.74% bounce rate. This means that people are coming to the archive and looking around for a significant period of time, clicking from page to page, from object to object. Surprisingly, the time spent on the site and number of pages per visit goes down for the 932 visits to the site and 6491 page views on April 16, 2008. This is likely related to the increase in new visitors over the average to the month of April. Many people will still remembering and revisiting their content in this period. People unfamiliar with the material in the archive are less likely to spend time on this archive when other archives with different content are also available.



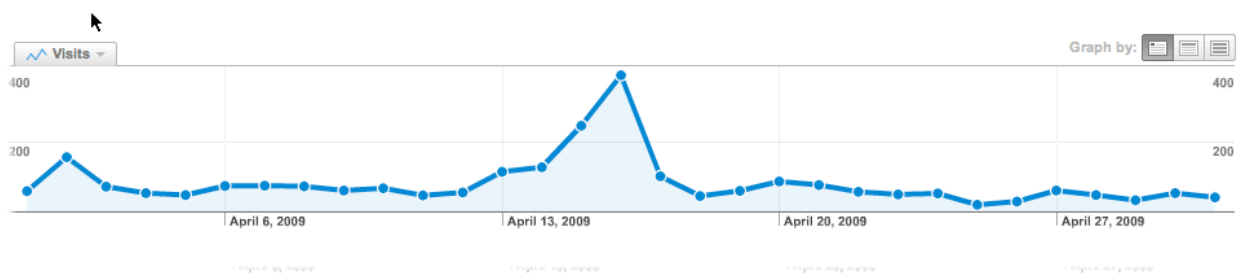
#### Site Usage



#### Site Usage



While the low numbers in 2008 are interesting to some extent, they gain their strength as indicators of a loss of contributors/interested parties in relation to the even lower numbers in 2009.



#### Site Usage



#### Site Usage



The third anniversary of the April 16 Tragedy, shows the archive getting significantly less traffic in the month of April, at slightly more than 1/2 the visits and 1/3 the page views, the bounce rate is 1/3 higher, and we have fewer pages per visit. Before the third anniversary we did launch the frontpages archive, but that has had little effect to the primary site. The day of April 16 is significantly less in all statistics in comparison to the prior archive with our key indicators of pages per visit and average time on site, which we take to be indicators of interest to the site and the material falling off significantly.

## Conclusion

With people spending less time on the site, viewing fewer pages on the site, I feel fairly safe in saying that there is likely less interest in the event across various user groups, and with this loss of interest we have a loss of activity. This loss of activity does not hurt the legitimacy of the archive for researchers, but does likely relate to fading memories and the peripheralization of the event to people's lives. This change seems lessens the interaction with the content creators that contributed materials to the archive, which in terms removes over time



some of the web 2.0 orientation of the archive. I suspect that most event-driven archives face the same issues of community and fading memories.

**References:**

- [1] Jesiek, B. K., & Hunsinger, J. (2008). THE APRIL 16 ARCHIVE: Collecting and Preserving Memories of the Virginia Tech Tragedy. In B. Agger & T. W. Luke (Eds.), *There is a Gunman on Campus* (p. 22). Lanham, Maryland: Rowman & Littlefield.
- [2] Jesiek, B. K., & Hunsinger, J. (2009). Collecting and Preserving Memories From the Virginia Tech Tragedy: Realizing a Web Archive. In N. Brügger (Ed.), *Web Histories*. London: Peter Lang.
- [3] Auge, M. (2004). *Oblivion*. University of Minnesota Press.

**Thomas RISSE, Julien MASANÈS, András A. BENCZÚR, Marc SPANIOL**

## Turning pure Web page storages into living Web archives

### Abstract

Web content plays an increasingly important role in the knowledge-based society, and the preservation and long-term accessibility of Web history has high value (e.g., for scholarly studies, market analyses, intellectual property disputes, etc.). There is strongly growing interest in its preservation by libraries and archival organizations as well as emerging industrial services. Web content characteristics (high dynamics, volatility, contributor and format variety) make adequate Web archiving a challenge.

LiWA will look beyond the pure “freezing” of Web content snapshots for a long time, transforming pure snapshot storage into a “Living” Web Archive. In order to create Living Web Archives, the LiWA project will address R&D challenges in the three areas: Archive Fidelity, Archive coherence and Archive interpretability. The results of the project will be demonstrated within two application scenarios namely “Streaming Archive” and “Social Web Archive”. The Streaming Archive application will showcase the building of an audio-visual Web archive and how audio and video broadcast related web information can be preserved. The Social Web application will demonstrate how web archives can capture the dynamics and the different types of user interaction of the social web.

**Keywords:** Web Archiving, Rich Media, Spam Detection, Crawl Coherence, Terminology Evolution

### Introduction

The Web today plays a crucial role in our information society: it provides information and services for seemingly all domains, it reflects all types of events, opinions, and developments within society, science, politics, environment, business, etc. Due to the central role the World Wide Web plays in today's life, its continuous growth, and its change rate, adequate Web archiving has become a cultural necessity in preserving knowledge. Consequently a strong growing interest in Web archiving library and archival organizations as well as emerging industrial services can be observed.

However, web preservation is a very challenging task. In addition to the “usual” challenges of digital preservation (media decay, technological obsolescence, authenticity and integrity issues, etc.), web preservation has its own unique difficulties:

- distribution and temporal properties of online content, with unpredictable aspects such as transient unavailability,
- rapidly evolving publishing and encoding technologies, which challenge the ability to capture web content in an authentic and meaningful way that guarantees long-term preservation and interpretability,
- the huge number of actors (organizations and individuals) contributing to the web, and the wide variety of needs that web content preservation will have to serve.

A first generation of Web archiving technology has been built by pioneers in the domain like the Royal Library of Sweden and the Internet Archive based on existing search technology. It is now time to develop the next generation of Web archiving technology, which is able to create high-quality Web archives overcoming the limitations of the previous generation. The aim of the European funded project LiWA is to create innovative methods and services for Web content capture, preservation, analysis and enrichment.

In the following section we first give an overview about the current state in Web archiving. Afterwards we will introduce in more detail the Living Web Archives project followed by an overview of the approaches to address the previously mentioned issues. Furthermore we will give an overview of the applications to be developed within the project. Finally the paper concludes and gives an outlook on the remaining project life time.

### The Living Web Archives Project

The LiWA project, started in February 2008, brings together a consortium of highly qualified researchers (L3S Research Center, Max Planck Society, Hungary Academy of Science), archiving organizations (European Archive Foundation, Sound and Vision Foundation (NL), National Library of the Czech Republic, Moravian Library) and a commercial company (Hanzo Archives). It is the intention of the project partners to turn Web archives from pure Web page storages into “living Web archives” within the next three years. Such living archives, will be capable of: handling a variety of content types; dealing with evolution as well as improving long-term content usability. In order to create Living Web Archives, the LiWA project addresses R&D challenges in the three areas: Archive Fidelity, Archive coherence and Archive interpretability:



- **Archive Fidelity:** development of effective approaches and methods for capturing all types of Web content including the Hidden and Social Web content, for detecting capturing traps as well as for filtering out Web spam and other types of noise in the Web capturing process.
- **Archive Coherence:** development of methods for dealing with issues of temporal Web archive construction, for identifying, analysing and repairing temporal gaps as well as methods for enabling consistent Web archive federation;
- **Archive Interpretability:** development of methods for ensuring the accessibility, and long-term usability of Web archives, especially taking into account evolution in terminology and conceptualization of a domain;

The results of the project will be demonstrated within two application scenarios namely “Streaming Archive” and “Social Web Archive”.

### **LiWA Approaches**

In the following sub-section we give an overview about the selected approaches in the four research areas covering the three objectives of the LiWA project. These approaches were developed after getting a detailed understanding of the requirements and the system architecture. The requirements analysis collected the requirements from three different angles. The user angle describes the desirable usage of web archives by libraries and archives. The technical angle collects functional requirements necessary to meet the user requirements of libraries and archives and the intention to extend the current state-of-the-art in/of web archiving. Finally the architecture angle defines functional requirements necessary to integrate LiWA services into one advanced web archiving infrastructure.

#### **Capture of Rich and Complex Web Content**

The aim of this working area is to improve dramatically the fidelity of Web archives by enabling capture of content defeating current Web capture tools. This comprises the ability to find links to resources regardless of the encoding using virtual browsing, the detection and capture of structural hidden Web and the capacity to handle streaming protocols to capture rich media Web sites. In order to develop an interpretation/execution-based link extractor for complex and dynamic objects, potential Javascript rendering engines for tasks were identified and tested. The comparison lead to select "WebKit" for implementation as it offers a huge number of features like JavaScript getters and setters, DOM class prototypes, significant JavaScript speed improvements, support of new CSS3 properties. DOM manipulation issues were analysed in depth to develop better links extraction. Various strategies to manipulate DOM from Webkit were tested. The result is a customized version of WebKit for the special use of link extraction.

For capturing rich media open source modules and helper application to support AV applications were tested. The Mplayer was selected as the basis for the helper tool implementation. In order to develop an improved rich media capture module, the crawlers were de-coupled from the identification and retrieval of streams and then moved to a distributed architecture where crawlers communicated with stream harvesters through messages.

#### **Data Cleansing and Noise Filtering**

The ability to identify and prevent spam is a top priority issue for the search engine industry [1] but less studied by Web archivists. The apparent lack of a widespread dissemination of Web spam filtering methods in the archival community is surprising in view of the fact that, under different measurement and estimates, roughly 10% of the Web sites and 20% of the individual HTML pages constitute spam.

Spam filtering is essential in Web archives even if we acknowledge the difficulty of defining the boundary between Web spam and honest search engine optimization. Archives may have to tolerate more spam compared to search engines in order not to loose some content. Also they might want to have some representative spam either to preserve an accurate image of the Web or to provide a spam corpus for researchers. Therefore the main objective of spam cleansing in Web archives is to reduce the amount of fake content the archive will have to deal with. The envisioned toolkit will help prioritize crawls by automatically detecting content of value and exclude artificially generated manipulative and useless content.

The current LiWA solution is based on the lessons learned from the Web Spam Challenges [2]. As it has turned out, the feature set described in [3] and the bag of words representation of the site content [4] give a very strong baseline. Therefore the LiWA baseline content feature set consists of the following language-independent measures: the number of pages in the host, the number of characters in the host name, in the text, title, anchor text etc; the fraction of code vs. text, the compression rate and entropy; and the rank of a page for popular queries. Within this set we use the measures for in- and outdegree, reciprocity, assortivity, (truncated) PageRank, Trustrank [5] and neighborhood sizes, together with the logarithm and other derivatives for most



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

values. Whenever a feature refers to a page instead of the host, we select the home page as well as the maximum PageRank page of the host in addition to host-level averages and standard deviation.

In addition LiWA services intend to provide collaboration tools to share known spam hosts and features across participating archival institutions. A common interface to a central knowledge base will be built in which archive operators may label sites or pages as spam based on own experience or suggested by the spam classifier applied to the local archives.

As a major step in disseminating the special needs of Internet Archives, we propose tasks for a future Web Spam Challenge [6]. We generate new features by considering the temporal change of several crawl snapshots of the same domain [7]. In addition by the needs of collaboration across different archival institutions we also provide training labels over one top level domain and request prediction over a different domain.

### **Archive Coherence**

A common notion of “coherence” refers to the explanations given in the Oxford English Dictionary (cf. <http://dictionary.oed.com>) describing coherence as “the action or fact of cleaving or sticking together”, which - in terms of a Web site - results in a “harmonious connexion of the several parts, so that the whole 'hangs together'”. From an archiving point of view, the ideal case to ensure highest possible data quality of an archive would be to “freeze” the complete contents of an entire Web site during the time span of capturing the site. Of course, this is illusion and practically infeasible. Consequently, one may never be sure if the contents collected so far are still consistent with those contents to be crawled next. However, temporal coherence in Web archiving is a key issue in order to capture digital contents in a reproducible and, thus, later on interpretable manner. To this end, we are developing strategies that help to overcome (or at least identify) the temporal diffusion of Web crawls that last from only a few hours up to several days. Therefore, we have developed a coherence framework that is capable of dealing with correctly as well as incorrectly dated contents[8]. Depending on the data quality provided by the Web server, we have developed different coherence optimizing crawling strategies, which outperform existing approaches and have been tested under real life conditions. Even more, due to the development of a smart revisit strategy for crawlers we are also capable of discovering and (as a consequence) of ensuring coherence for contents, which are incorrectly dated and thus not interpretable with conventional archiving technologies. Current results make temporal coherence of Web archiving traceable under real life applications and provides strategies to improve the quality of Web Archives, regardless of how unreliable Web servers are.

### **Archive Interpretability**

The correspondence between the terminology used for querying and the one used in content objects to be retrieved is a crucial prerequisite for effective retrieval technology. However, as terminology is evolving over time, a growing gap opens between older documents in (long-term) archives and the active language used for querying such archives. Language changes are triggered by various factors including new insights, political and cultural trends, new legal requirements, high-impact events, etc.

An abstract model has been developed [9] that allows the representation of terminology snapshots at different times (term-concept-graphs). From this we derived that the act of automatically detecting terminology evolution given a corpus can be divided into two subtasks. The first one is to automatically determine, from a large digital corpus, the senses of terms. Such a word sense discrimination module has been implemented and successfully been tested on the Times corpus that covers 200 years of news articles. Current work focuses on the second step – the detection of terminology evolution. In this step the word clusters detected in the first step are tracked over time to detect evolution and to derive mappings.

### **Applications**

The LIWA Technologies can be used either at crawl-time or after completion of the crawl, integrated with existing web archiving workflow. In order to test and apply these new methods and results, an integration platform of the modules is being built both by the European Archive Foundation (using open source tools) and by Hanzo Archives.

Two applications scenarios are developed in LIWA to illustrate the possible use of these technologies in real world scenario whose scope is wider than what LiWA specifically addresses.

### **LiWA technology for content and context in Sound and Vision archive**

The Netherlands Institute for Sound and Vision is one of the largest audio-visual archives in Europe. The cultural heritage preservation policy of the Institute implies that the AV archive should preserve the Dutch audiovisual cultural heritage. As the Internet is increasingly becoming an important source for (user generated) audiovisual cultural heritage content, Sound and Vision has a strong commitment to capture information

available on the Web. More specifically, the institute is eager to capture broadcast related websites, including streaming content. However, as capturing streaming content from the web is difficult, until now only a selection of user generated video content is downloaded manually from the Internet. With the streaming content capturing technology developed in the LiWA project, Sound and Vision is able to address the capturing of Dutch cultural heritage content in a much more efficient way.

Besides being a potential provider of audiovisual content, the Web is regarded as a valuable source for gathering contextual information that relates to the collections. Context information is relevant for both documentalists, and also other users interested in a specific broadcast or a broadcasting related topic, such as journalists, teachers or researchers. Typically, these users have to use different interfaces for different sources to search these sources. Ideally, Sound and Vision provides these users with a single interface that allows searching both the digital asset management system of the AV archive (iMMix) and related web content. The LiWA application Streaming demonstrates how broadcast related potential end users could access web content. The archived content will be used as test data for the development of the Sound and Vision context data platform that specifically addresses the linking of web context to the digital asset management system of Sound and Vision.

### **Social web application**

Social web sites typically contain highly inter-linked content and use dynamic linking, widgets and tools as well as high degree of personalisation. Capturing social web sites is extremely challenging and cannot be fully achieved using current methods and tools. Social web thus represents one of the greatest challenges in web archiving.

With the Social web application, LiWA intends to demonstrate a dramatic improvement in both archive structure and content completeness so that the rapidly evolving and increasingly diverse content of the social Web is captured more accurately and evenly. The aim of the application is to show how the LiWA technology fits in the workflow of an active Web archiving institution, by considering a real-life scenario of the National Library of the Czech Republic. The application is designed as a set of independent modules developed in LiWA as described in section 2. The modules can be readily integrated with existing Web archiving workflow management tools. A Web archiving institution can choose to deploy all of the modules or just some of them, depending on its needs and particular workflow. The application is designed as generic and can be used to enhance archiving of any type of web content, not just social web.

### **Conclusions & Outlook**

In this paper we presented important issues in Web archiving and introduced the Living Web Archives project, which aim is to overcome these limitations. For research areas have been identified namely Capturing of Rich and Complex Web content, Data Cleansing and Noise Filtering, Archive Coherence and Archive Interpretability. Promising solutions have already been developed and continuously being enhanced in the second half of the project. Furthermore the presented application showcases will be implemented.

### **Acknowledgments**

This work is funded by the European Commission under LiWA (IST FP7 216267).

### **References**

- [1] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.
- [2] C. Castillo, K. Chellapilla, and L. Denoyer. Web spam challenge 2008. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.
- [3] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: web spam detection using the web topology. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 423–430, 2007.
- [4] I. Bíró, D. Siklósi, J. Szabó, and A. A. Benczúr. Linked latent Dirichlet allocation in web spam filtering. In *AIRWeb '09: Proc. 5th int. workshop on Adversarial information retrieval on the web*, 2009.
- [5] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proc. of the 30th Int. Conference on Very Large Data Bases (VLDB)*, pp. 576–587, Toronto, Canada, 2004.
- [6] A. A. Benczúr, M. Erdélyi, J. Masanés, and D. Siklósi. Web spam challenge proposal for filtering in archives. In *AIRWeb '09: Proc. of the 5th Int. Workshop on Adversarial Information Retrieval on the Web*. ACM Press, 2009.
- [7] M. Erdélyi, A. A. Benczúr, J. Masanés, and D. Siklósi. Web spam filtering in internet archives. In *Proc. of 5th the Int. Workshop on Adversarial information retrieval on the web (AIRWeb)*, 2009.



- [8] M. Spaniol, D. Denev, A. Mazeika, P. Senellart and G. Weikum. Data Quality in Web Archiving. In Proceedings of the 3rd Workshop on Information Credibility on the Web (WICOW 2009) in conjunction with the 18th World Wide Web Conference (WWW2009), Madrid, Spain, April 20, 2009, pp. 19-26.
- [9] N. Tahmasebi, T. Iofciu, T. Risse, C. Niederee, and W. Siberski; Terminology Evolution in Web Archiving: Open Issues; In Proc. of the 8th Int. Web Archiving Workshop 2008, Aarhus, Denmark

**Sam COPPENS, Erik MANNENS, Tom EVENS, Laurence HAUTTEKEETE,  
and Rik VAN DE WALLE**

## **Digital long-term preservation using a layered semantic metadata schema of PREMIS 2.0**

### **Abstract.**

In Belgium, many institutions have a lot of information stored on analogue carriers. This information is likely to get lost if no digitized copy of the information is stored for the long term. Long-term preservation is subjected to many risks. Overcoming those risks starts with describing the data thoroughly. The metadata needed for long-term preservation are descriptive metadata to search and manage the whole archive, binary metadata to describe the bitstreams, technical metadata describing the files, structural metadata for the representation information, preservation metadata for keeping track of the provenance of the data, and rights metadata. Therefore, we developed a layered semantic metadata schema. The top layer holds the descriptive metadata, the bottom layer holds all the information necessary for long-term preservation. The top layer consist of an OWL representation of Dublin Core, while for the bottom layer we developed an OWL representation of the preservation standard PREMIS 2.0, extended with a vocabulary defining the legal roles of a person, organization, or software. This way, our model offers all the necessary metadata for long-term preservation.

**Keywords:** digital preservation, PREMIS 2.0, ontology, semantic web

### **Introduction**

In Belgium, the broadcasters, cultural organizations, private persons, and government institutions possess thousands of hours of speech and image material which is stored on analogue carriers. This material belongs to the most important cultural heritage in Flanders. At this moment, the analogue carriers are degrading and are continuously losing quality, making the data inaccessible. Disseminating and storing the content digitally overcomes this problem only temporarily. Furthermore, this digital content has to remain intact and accessible over time, e.g., 20, 50 years or longer. Digital long-term preservation forms the solution for this issue. The project BOM-VI (Preservation and Disclosure of Multimedia Data in Flanders, [1]) initiates the digital long-term preservation of the cultural heritage in Flanders and researches the problems encountered with digital long-term preservation.

In this paper, we present our layered semantic metadata model. First, in chapter two, we introduce the different kinds of metadata that are needed to overcome all the risks involved in long-term preservation and show how our proposed, layered, semantic metadata model relates to those risks. The semantic model consists of two layers: the top layer delivers the descriptive metadata, and the bottom layer is responsible for the binary metadata, the technical metadata, the structural metadata, the preservation metadata (provenance metadata, fixity metadata, and context metadata), and the rights metadata. This way, all the metadata for describing the content for the long-term, are covered by the layered semantic metadata model. For the top layer, we use a Web Ontology Language (OWL, [2]) representation of Dublin Core [3], which is described in chapter three. For the bottom layer, depicted in chapter four, we developed an OWL representation of the preservation standard Preservation Metadata, Implementation Strategies 2.0 (PREMIS 2.0, [4]). This PREMIS OWL schema (PREMIS OWL, [5]) not only covers the necessary metadata described in chapter two, but also stores the semantics of the metadata for the long term. This can be very important due to, e.g., terminology changes. This schema is accompanied by a vocabulary describing the legal roles that a person, organization, or software application can have.

### **Metadata levels for long-term preservation**

When preserving digital multimedia data for the long term, the digital archive demands some specific requirements. On the one hand the software and hardware of the digital archive have to guarantee access to the information during a long time. On the other hand human input is necessary in the form of archive descriptions, work processes, and the use of standards to keep the information accessible and interpretable as long as possible to the user community. Based on the Open Archival Information System (OAIS, [6]) reference model, the data has to be described on three levels to guarantee long-term preservation. On every level, there are possible risks for loss of data, which can be minimized by describing the data thoroughly.

On the lowest level, a digital file consists of bits and bytes which can change by external influences, like corruption of carriers, migrations, etc. On this lowest level, binary metadata and fixity metadata are needed to correct these errors and to guarantee authenticity of the data.

On a higher level, file formats and compression formats like AVI, MP3, and JPEG describe the way the bits can be transformed to an interpretable multimedia representation. When a file format becomes obsolete, the archive has two solutions to preserve the stored data: migration or emulation. Metadata is needed to support these actions. At this level, it is also very important to preserve the look and feel of the objects when migrating file formats. Thus, a rich description of the look and feel is also necessary. For this level we need technical metadata, for describing the files, structural metadata, for describing sets of files and their relations, e.g., a book which is represented as a set of scanned TIFF images, and provenance metadata, for describing the history of the content information: the original owners of the data, the processes that determined the current form of the data, and the available versions.

On the highest level, the information should remain interpretable. Institution structures, terminologies, the designated community, and the rights of an object or institution can change over time. To keep the information interpretable, enough information should be included in the archived package. At this level, the archive needs descriptive metadata, for a general description of the object, e.g., MARC, rights metadata, for describing copyright statements, licenses, and possible grants that are given, and context metadata, for describing the relations of the content information to information which is not packed in the information package. Examples of context metadata are related datasets, references to documents in the original environment at the moment of publication, helper files, and the language.

When developing a metadata schema for the long-term preservation of digital multimedia, metadata descriptions on all levels have to be taken into account, going from bit level descriptions to descriptions of the intellectual content. To realize this, we developed a layered semantic metadata schema. The top layer offers the descriptive metadata. The bottom layer takes care of the preservation metadata, rights metadata, binary metadata, technical metadata, and structural metadata necessary for deep archiving. For the top layer, an OWL representation of Dublin Core is developed. For the bottom layer, an OWL representation of the preservation standard PREMIS 2.0 is developed. This standard is based on the OAIS reference model. This schema describes the data on all necessary levels.

### **Top layer: Dublin Core**

Descriptive metadata describes the content of the data: subject, author, date of creation, file format, etc. This metadata makes it possible to manage and search the complete digital archive. When archiving data coming from different sectors like the broadcast sector, the libraries, the cultural sector, and the archival sector, a problem arises concerning descriptive metadata. Many of the institutions already have descriptive metadata. Are these descriptive metadata stored as metadata or as data? Both strategies have their advantages and disadvantages. When archiving these descriptions as metadata, the archive has to provide a metadata schema. The choice of this schema is a non-trivial task. The metadata schemes used for the descriptions are very domain-specific. To store the descriptive metadata lossless the descriptive metadata schema should be some kind of smallest common multiple of all the descriptive metadata schemes offered by the institutions. This would be a huge metadata schema, impossible to maintain. That is why the descriptive metadata is archived along with the data in their original metadata format, e.g., MARC, so there is no information loss. On top of this metadata, the archive offers a broadly accepted descriptive metadata schema. This gives the archive the necessary tools to search the whole archive. When finding the data of interest, the original metadata that is stored as data can still be presented to the users.

Dublin Core was chosen to describe this top layer of descriptive metadata. Dublin Core is a broadly accepted descriptive schema. The power of this schema is its simplicity and generality. It consists of fifteen fields among which creator, subject, coverage, description, date. It can answer to the basic questions: Who, What, Where, and When. All the fields in Dublin Core are optional and repeatable. This makes it possible to map relatively easily almost all the descriptive metadata schemes to Dublin Core whereas many institutions already support Dublin Core.

### **Bottom Layer: PREMIS OWL**

For this layer, we developed an OWL schema of the preservation standard PREMIS 2.0. PREMIS is a preservation standard based on the OAIS reference model. The preservation standard is described by a data model. The data model of PREMIS consists of five semantic units or classes important for digital preservation purposes:

- Intellectual Entities: a part of the content that can be considered as an intellectual unit for the management and the description of the content. This can be for example a book, a photo, or a database.
- Object: a discrete unit of information in digital form.
- Event: An action that has an impact on an object or agent.

- Agent: a person, institution, or software application that is related to an event of an object or is associated to the rights of an object.
- Rights: description of one or more rights, permissions of an object or agent.

Intellectual entities, events, and rights are directly related to an object. An agent can only be related to an object through an event or through rights. This way, not only the changes to an object are stored, the event involved in this change is also described. These relationships offer the necessary tools to store the provenance of an object properly. Fig. 1 clarifies the data model of PREMIS.

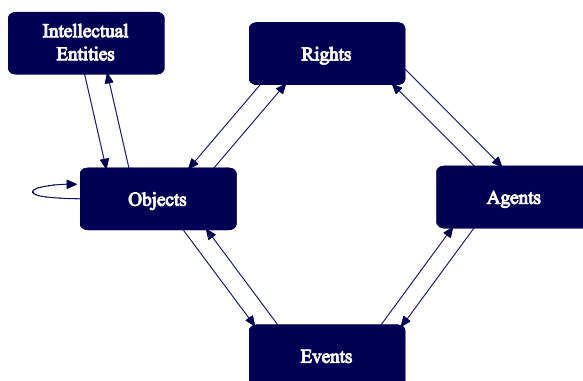


Fig. 1. Data model of PREMIS

## Object

The Object class describes a unit of information in digital form. It is related to the intellectual entity class. This intellectual entity is described by descriptive metadata. This descriptive metadata are very domain-specific. For this, there exist already a lot of descriptive metadata models. Therefore, the description of the intellectual entity is out of scope for PREMIS. In our implementation, the top layer describes the intellectual entity.

An Object class knows three subclasses:

- File: a file is an ordered sequence of bytes that is known by the system.
- Bitstream: a bitstream is the actual data inside a file.
- Representation: a representation is a set of files with structural metadata needed for a complete description of an intellectual entity.

The Object class possesses all the necessary features to describe the object on the different levels. The minimum information for describing an object (File, Bitstream, or Representation) are `objectIdentifier`, which gives the identifier of the object, `objectCharacteristics`, needed for the Bitstream subclass and the File subclass, which gives the necessary technical and binary metadata, and `storage`, necessary for describing a File or Bitstream, which indicates either the location the data is stored, either the medium the data is stored on. An object can be described further into detail using `preservationLevel`, because some repositories offer the opportunity to define a preservation level for an object, `significantProperties`, defining some significant properties of the object, which need to be preserved when, e.g., migrating the data, `originalName`, for indicating the original names of the packages delivered to the repository, `environment`, which describes the environment the user needs to render the content and interact with the content, `signatureInformation`, for storing digital signatures generated during ingest into the repository, and finally, `relationship`, which relates to structural metadata to assemble complex objects.

For linking object information to events, intellectual entities, or rights statements, the object class offers three properties, i.e., `linkingEvent`, `linkingIntellectualEntity`, and `linkingRightsStatement`.

## Event

An event aggregates all the information about an action that involves one or more objects. This metadata is stored separately from the object metadata. Actions that modify objects should always be recorded as events.

The Event class is described at least by an `eventIdentifier`, `eventType`, e.g. capture, creation, and an `eventDateTime`. This information can be extended using the `eventDetail` property, which gives a more detailed description of the event, and the `eventOutcomeInformation`, which describes the outcome of the event, in terms of success, failure, or partial success. These properties are able to describe any event altering an object. The Event class can be related to an Agent class or Object class via the resp. properties `linkingAgent` and `linkingObject`.

## Agent

This class aggregates information about attributes or characteristics of agents. Agents can be persons, organizations or software. This class provides the necessary tools to identify unambiguously an agent. The minimum properties needed to describe the Agent class are `agentIdentifier` and `agentType`. Optionally, an agent can also be described using the `agentName`. This is just enough to identify the agent.

An agent can hold or grant one or more rights. It may carry out, authorize, or compel one or more events. An agent can only create or alter an object through an event or with respect to a rights statement. The relationships between an agent and an object through an event or rights entity make it possible to describe the whole provenance of an object.

## Rights

The minimum core rights information that a preservation repository must know, is what rights or permissions a repository has to carry out related to objects within the repository. These may be granted by copyright law, by statute, or by a license agreement with the rights holder. Rights entities can be related to one or more objects and one or more agents.

Every Rights class can be related to different RightsStatements. A RightsStatement knows three subclasses: the Copyright subclass, the License subclass, and the Statute subclass. These three subclasses offer the necessary metadata for describing, rights information, i.e., copyrights, licenses, and statutes. Every RightsStatement is described at least by a `rightsStatementIdentifier`, and has also the optional property `rightsGranted`, which describes the actions the granting agency has allowed the repository. The RightsStatement class can be related to an Object class or Agent class via the optional, repeatable object properties: `linkingObject` and `linkingAgent`.

This part of the PREMIS OWL schema is extended with a vocabulary that describes the roles agents can have concerning a rights statement. This vocabulary is based on the results of research performed within the project BOM-VI. To fully describe the rights of an object, all the persons, involved in the production of the described object, should be taken into account which is for many organizations impossible. Therefore, a checklist was made with the most important rights and rights holders that should be described. Based on this checklist a vocabulary was made to describe these important legal roles of an agent, e.g., author, composer, conductor.

## Conclusion

When preserving digital information for the long term, different metadata are important. Descriptive metadata are needed to describe the intellectual entities, binary metadata, technical metadata, and structural metadata are essential for the description of the data on all levels (bitstream, file, representation). Preservation metadata is necessary to describe the provenance of the data, to guarantee the authenticity of the digital data, and to provide a context. At last, rights metadata also needs to be stored.

The two-layered, semantic metadata schema described in this paper offers all these metadata. The top layer takes care of the descriptive metadata. An OWL representation of DC was chosen for this layer. The bottom layer carries the binary metadata, technical metadata, structural metadata, preservation metadata, and the rights metadata. For this layer an OWL representation of PREMIS 2.0 was developed. To describe the rights in a more detailed manner, the PREMIS OWL schema was extended with a vocabulary defining the different legal roles of persons, organizations and software. By describing the data with this layered metadata schema, all the risks that come with long-term preservation are minimized. By splitting up the semantic schema in two layers, the top layer with the descriptive metadata can be made public and weaved into the web of data, if the rights permit it. The bottom layer remains closed for the public and is responsible for the long-term preservation of the data.

## References

- [1] Preservation and Disclosure of Multimedia Data in Flanders, <https://projects.ibbt.be/bom-vi/>
- [2] Dean, M., Connolly, D., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., Stein, L. A.: OWL web ontology language reference, W3C Working Draft, <http://www.w3.org/TR/2003/WD-owl-ref-20030331> (2003)
- [3] The Dublin Core Metadata Initiative, DCMI, <http://dublincore.org/> (2009)
- [4] Higgins, S.: PREMIS Data Dictionary, Digital Curation Centre (DCC), <http://www.loc.gov/standards/premis/v2/premis-dd-2-0.pdf>, Glasgow (2007)
- [5] Coppens S., Mannens E., Van de Walle R.: PREMIS OWL, Semantic Model of PREMIS 2.0, <http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/premis.owl>
- [6] Consultative Committee for Space Data Systems (CCSDS): Reference Model for an Open Archival Information System. Blue book. Issue 1, 148 p., CCSDS, Washington (2002)





Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

## **Daniel TERUGGI**

### **PrestoPRIME: keeping digital content alive**

FP7 Integrated Project, third call

Started on January 2009, duration 40 months

Partners : INA (F), BBC (GB), RAI (It), JRS (Au), B&G (NI), ORF (A), ExLibris (Israél), Eurix (It), Doremi (F), Technicolor (NI), IT Innov. (GB), Vrije Universiteit Amsterdam (NI), Universität Innsbruck (Au, European Digital Library Foundation (NI)

Getting into the digital world is a highly challenging issue for any Audiovisual collection due the complexities and costs of digital transfer. However a still more challenging issue is how to remain in the digital world within its continually changing context. Information systems, formats, definitions change continuously and induce regular actions on contents, descriptions and systems so to guarantee access on the long-term. There is no clear vision how all these actions affect contents or how to ensure that these actions will not modify them.

For Digital Libraries this represents a major evolution in their conception: initially build around digitised books, they have started including all kinds of cultural digital contents from Archives, Museums and Audiovisual Libraries. The main challenge in past years was bringing analogue contents to the digital world, then, conceiving specific tools for the management, description and structuring of digital contents. The challenge today for content holders is to make all contents available to the citizen, in secured environments and respecting intellectual property. Major European projects advance in this direction through intelligent access portals with millions of contents.

A large array of problems still remains within digital libraries; they may be particular to kind of content or media or transversal to different domains. They concern the permanence of Digital objects through time (a major concern for any digital content owner); Interoperability of contents and metadata; management and handling of Rights; Content tracking and identification and; a less technical however important issue: how to economically foster major digitization actions and digital preservation.

### **Specific domain**

The Audiovisual domain presents challenges and issues that need a specific approach:

- The mass of accumulated material since the beginning of film and broadcast industry is huge (estimated in 100 million hours)
- Most of it is still in an analogue format
- It is totally dependent of access technology (machines, readers)
- Digital born material presents similar problems of technological access and format dependency
- Professional audiovisual material increases, in Europe, at a rate of circa 5 million hours per year
- Audiovisual contents need efficient and extensive metadata for archival and access purposes
- They represent huge storage volumes (1 million hours of video at a 1mb/s rate, which is a non professional rate, represent 3,6 Petabytes of storage)
- It generally presents complex right ownership situations or poorly identified ownership
- If not managed with a long-term perspective they are easily subject to loss or inaccessibility
- Audiovisual archiving represents a continuous cost for content holders, archiving needs to be carefully evaluated, planned and implemented

Past projects like Presto and PrestoSpace, have developed specific technologies and business plans in order to address, on an industrial basis, the problem of analogue to digital migration. The results of the project have been brought back to the community through preservation machines, tools for storage management, tools for restoration, metadata platforms, business models and strong methodologies for an industrial approach .

PrestoSpace opened the gates to analogue to digital migration through its achievements and the strong price-reducing factor it brought to the activity in the domain. It also structured the actors of the community and set the basis of an efficient interaction among them. However these important steps, as essential as they have been, are not enough to inscribe the preservation of Europe's Audiovisual Heritage on the very long term. Managing and securing huge masses of digital contents with an appropriate description and identification is an indispensable step to assure content owners of the durability of their assets and to promote large access programs to contents.



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

## **PrestoPRIME: an Integrated Project for Digital Audiovisual Contents**

Getting into the Digital world represented the first indispensable step in order to address the problem of Digital Audiovisual Archives as a whole. A large awareness has been built on the necessary actions that need to be done in order to avoid the analogue black hole. Preservation and documentation functionalities have been improved and brought to common working environments and new semantic tools are being largely developed to improve access, recognition and identification of contents.

Still, mainly due to the volumes, the complexity and the diversity of the audiovisual domain; there is a strong need for research and development of tools and methodologies to assure what is the new challenge for digital contents: long-term preservation, identification and access. The considerable growth of the number of audiovisual contents from the past and regularly produced by professionals and non-professionals, associated to the increasing circulation of those images, introduces two new problems that deal directly with preservation:

The origin and identification of contents: Who made it? Who does it belong to? Can I use it? Where is the original? Can I find it in a better quality?

Content archiving and preservation: Who is keeping them? Does he have the mission to do it? With which time perspective? Will the contents be there in a hundred years? Will I be able to access them? How can I secure my own contents?

The PrestoPRIME project addresses the new challenges that need to be tackled in order to guarantee long-term access and usability of contents. The project is structured in four domains that constitute a global approach to an organised structure for audiovisual contents permanence. It brings together two research domains concerning

- Digital permanence, long-term storage and content identifications
- Interoperability and quality assessment

These central aspects are dealt in close relationship with two other major issues which are:

- A Right management environment to model European right legislation and propose effective tools for content exchange at a European level
- A Competence Centre dedicated to Audiovisual preservation, restoration and documentation issues, serving as a reference institution for content holders, service providers, industrials and research projects.

## **The Competence Centre, a central concept for PrestoPRIME**

The Competence Centre is a major tool for the PrestoPRIME project, its function is to foster, accelerate and become a reference for Preservation and Migration actions in the Audiovisual domain. Based on the results of previous projects for all what concerns analogue to digital migration; it will incorporate the results of research and development issued from the project, thus being functional from the beginning of the project. Furthermore, its role as a federator of actors within the audiovisual domain: -audiovisual archives and collections - service providers – industrials - academic and industrial research; its structuring actions: - registers of experts – registers of technologies – registers of works; will make the Competence Centre become a central actor for the domain, in relation the major access projects like Europeana and in close relationship with the European Commission.

The Competence Centre will be launched in October 2010 through its portal.

## **Raffaele CIAVARELLA, Alain BONARDI, Guillaume BOUTARD**

### **Virtualization of real time audio processes: towards a musical notation of contemporary music**

#### **Abstract**

Contemporary musical production makes an heavy usage of digital artefacts, either hardware or software based. Since the middle of the 80es, an important issue has been recognized: the fast obsolescence of hardware and software products endanger seriously the future of this production. The question is not only to preserve the results, by recording them, but to preserve the ability to reperform the works live, as we do today for music of the last centuries.

We will present the methodology we develop in the ASTREE project for building knowledge in relation to musical works, and particularly for digital processes that are considered as specific music instruments. We will discuss the different issues in preservation of contemporary music, and show that one of the most prominent is the lack of formalized knowledge about the digital musical instruments, their notation, and their integration into musical score. We will present our efforts towards building an organology of real-time audio processes, and show that this can be the basis for an adequate musical notation and its integration in musical score.

**Keywords** : Music, Contemporary, Digital, Preservation

#### **Introduction**

##### **A brief history**

The first interactive works combining performers and realtime electronic modulation of their parts have appeared in the middle of the 80es. Electronic devices, either hardware or software, have been interfering with various musical configurations: the instrument-computer duet, for instance in Manoury's works (Jupiter, for flute and computer, 1987-1992 ; En Echo, for voice and computer, 1993-1994); the works for ensemble and live electronics, such as Fragment de lune (1985-1987) by Philippe Hurel; the works for soloists, ensemble and electronics, such as Répons (1981-1988) by Pierre Boulez.

After nearly 25 years of interactive works, institutions have become aware that this type of music is completely dependant on its hardware and software implementation. May the operating system or the processor evolve, the piece cannot be reperformed. This is for instance what nearly happened to Diadèmes, a work by composer Marc-André Dalbavie for alto solo, ensemble and electronics. First created in 1986 and honoured by the Ars Electronica Prize, the work was last performed in 1992. In december 2008, its american creation was planned, more than 22 years after its premiere in France. But the Yamaha TX 816 FM synthetizers previously used are no longer available, and the one still present at Ircam is nearly out of order. Moreover, composer Dalbavie has tried several software emulators, but none of them was suitable according to him to replace the old hardware synthetizer.

In april 2008, Dalbavie and his musical assistant Serge Lemouton decided to choose another technique: they built a sampler. It is a kind of database of sounds produced by an instrument. The sounds have been recorded from the old TX 816 at various pitches and intensities. This solution enabled to reperform the piece having a kind of photography of the previous sounds. When no sound corresponding to a given pitch exists, the sampler is able to interpolate between existing files, to give the illusion the missing note exists.

One quickly understands the maintenance activity to be able to reperform a work is a never ending activity that should moreover respect a minimum of authenticity.

##### **Aims of preservation**

Having in mind that the aim of preservation is to make possible new performances of the works, it becomes clearly not sufficient to preserve the outputs – audio or video recordings – even if these recordings are clearly part of the objects to be preserved.

At the very core of performance is the real-time process, often called “the patch” : it is software that takes data in input – directly from the performer, or from prerecorded data, audio or video, and process them before rendering the output on speakers or a display. The real-time process is the expression of the ideas of the composer regarding the use of digital material, it is then the main object to be preserved, in addition to the score (French composer Philippe Manoury often calls the digital material he is using a “Digital Orchestra”).

## Risks and strategies

### The different strategies

Active preservation of realtime interactive music involves various aspects and is based on various actions. The first step is the physical conservation of all elements necessary for the reperformance of the work: the score, the patches, the various instructions, etc. At Ircam, the Mustica server provides patch files and instructions of implementation for a selection of nearly 60 works.

Another possible strategy is emulation, definitely one of the most difficult. Bernardini and Vidolin [1] quote the example of Stockhausen's Oktophonie, which requires an ATARI-1040 ST computer that no longer exists. There are Atari emulators running on other computers but nobody knows whether the Notator program used by Stockhausen will run on an emulator, though communities of users may give some help...

Migration is the most widespread activity to achieve reperformance. Many composers had their works transformed from one technical environment to another. All institutions in the field of electronic arts face migration necessities. At IRCAM, important pieces using Next computers were moved to Macintosh machines at the end of the 1990s.

Last but not least, virtualization means describing electronic modules using abstractions. At IRCAM, Andrew Gerzso has completed an important work aiming at finding representations as independent as possible from technical implementation for signal processing modules in *Anthèmes 2*, by Pierre Boulez, for violin and live electronics. The effects used in the piece have been added to the score as if they were instrumental parts. This is according to us the ultimate level of virtualization. It has also to be noticed that current musical notation (Common Western Musical Notation) is virtual, in the sense that it is independent from any implementation : we can play music written for any instrument on another instrument, the paradigms of notation being sufficiently abstract in order to achieve this goal.

### The musical notation issue

The need to integrate new technology in the musical score has been recognized a long time ago [2]. But, despite numerous tries, it seems that few systems have emerged. One can for instance examine the problem of notation for spatialization.

Spatialization of sound seems to be now a well-known domain, where numerous realizations have been made, and that offers a wide range of experiments. But there are few certainties, few theoretical studies and few references [3].

Concepts and terms used are vague, not precisely defined, and their acceptions are different according to the point of view of the actor, depending on numerous factors [3] :

- actions have different meanings according to the point of view : producer (an audio engineer...) or receiver (a listener)
- descriptors have different meanings according to the point of view envisioned : reality, image of the reality, or conceptualization of reality.

In this context, it becomes very difficult to envision a musical notation that takes into account all these different point of views, before having realized a unification, or merely a standardization of the domain, putting in relation the different points of view expressed.

Moreover, some tries to achieve notation of spatialization of the score becomes dependent on the physical implementation, like the following one that is dependent on a 5.1 system [4]:



Figure 1 : a notation of spatialization for a 5.1 system

### The need of a rationalized approach

From the remarks exposed above, we recognized the need to build a common base about the digital musical instruments, from where we could extract and constitute the knowledge basis for a rationalized approach of the musical notation issue. This approach is the basis of the ASTREE project that is exposed below.

### The ASTREE strategy

The ASTREE methodology

The ASTREE methodology is twofold:

First, existing processes have to be translated into a common language. Second, from that language, we will rebuild the original processes, or equivalent one, we will analyze them by applying data mining techniques, and we will also generate an automatic documentation.

The ASTREE methodology can be summarized in the following figure:

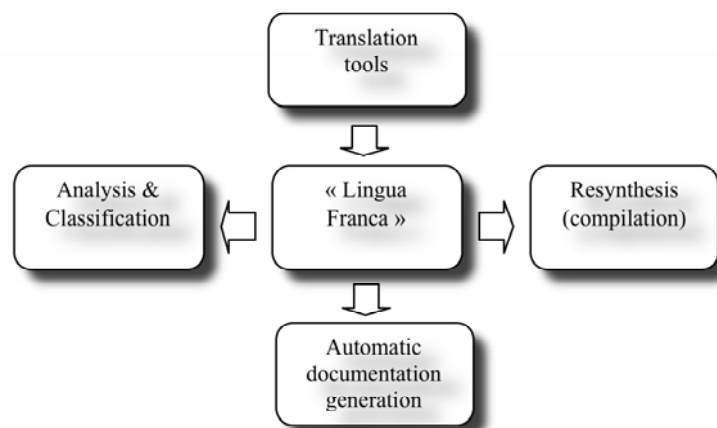


Figure 2 : the ASTREE methodology

#### Lingua Franca

At the very core of the methodology is a common language, a “Lingua Franca”, that should be completely independent from any hardware or software. Furthermore, this language should be sufficiently expressive to convey the meaning of current existing real-time processes, and at the same time very concise in order to be easily analyzed.

The FAUST language, developed in Grame since 2000 [5], is partially consistent with these requirements. It is a signal processing language, expressed as an algebra, that is aimed originally to process audio signals at a fixed rate, but is currently extended in order to become able to process vectors and matrices, with multi-rate capacities.

The FAUST language is sufficiently expressive for the expression of current objects in the domain of audio processing, at least for the synchronous part.

#### Translation

We develop tools for translating currently existing processes, built with current environments in use for contemporary music, for instance Max/MSP, but also for other environments like PureData (or even MatLab).

These translation tools are essential, not only for translation of existing material, but also for future material. Users, like composers or computer music designers that assist the composer in his task, are unlikely to use directly an algebraic language like the FAUST language. They will certainly continue using graphical programming environments like Max/MSP or PureData that let them free to experiment and build tools by successive refinements, rather than expressing tools in a language that imposes more or less a preliminary analysis and modeling.

#### Resynthesis

On top of this language, we can then add the ability to build processes, that in turn will become dependent on the machine, but that can be immediately compared to the original in terms of results: one can compare the output given by the newly implemented process the original output, as soon as the original process and its translation in the Lingua Franca are available.

The FAUST language can currently be translated in C++, and then compiled in a machine executable. We will also adapt the translation tool in order to work in the reverse order, and obtain Max/MSP or PureData implementations from the FAUST expression.

The purpose of this resynthesis is not only to be used in new performances, but, at the time of archiving processes, to prove that the FAUST expression of the process is sufficient. We can prove it through two stages : first, by doing a reverse translation in the original language, we can prove that no information was lost in the process, and second, by doing a new implementation, and comparing outputs with the original, we can make an a priori evaluation of the authenticity of the translated process towards original.

#### Automatic documentation generation

The code obtained as well as the source code can be analyzed in order to extract from there an automatic documentation. We can document input and output data, control parameters, extract comments and structures, and generate figures and mathematical expressions corresponding to the source.

#### Analysis and classification

By applying data mining techniques to the data set obtained by applying translation tools to existing processes, we intend to start a process that will end in an organology of digital instruments. Not only the processes themselves will be analyzed, but also the annotations made by users, as well as the automatic documentation previously obtained, and particularly the control parameters, names of input and outputs data, and dependencies. The relationship to works, where processes are in use, and metadata (authorship...) could also be analysed.

### Conclusion

For preservation of digital material that is produced today, our approach is to use the tools that are available today, and particularly digital tools. To build a database is not sufficient, there is also the need of building knowledge out of it. To this end, we will use the most recent techniques in data mining and data analysis: neural networks, bayesian networks, or fuzzy logic. We will also validate the results obtained with these techniques by using statistical methodologies.

For building executable programs, we have to use the most recent techniques, for instance parallelism in order to get the best of multi-core processors. We have then to automatize the whole process, in order to save time and effort, including automatic generation of documentation, or automatic translation of objects from one expression to another one.

This does not exclude other approaches, and particularly those based on human reasoning and human activities. Our opinion is that our approach will give them some ways to explore, as well as a strong basis for experimentation and validation of new ideas.

### Acknowledgements

This work is possible due to a grant of French National Research Agency, under the number 08-CORD-003. We would thanks the ASTREE partners, and noticeally Yann Orlarey, Laurent Pottier and Pierre Jouvelot.

### References

- [1] Bernardini, N., Vidolin, A. 2005. Sustainable Live Electro-acoustic Music. In Proceedings of the International Sound and Music Computing Conference, Salerno, Italy, 2005.
- [2] Assayag, G., « Problèmes de notation dans la composition assistée par ordinateur, », Rencontres Musicales Pluridisciplinaires. Musique et Notation GRAME, Lyon, 1997
- [3] Merlier, B. "Vocabulaire de l'espace et de la spatialisation des musiques électro acoustiques : présentation, problématique et taxinomire de l'espace", Electroacoustic Music Studies Network – Beijing 2006
- [4] Elleberger, E., "Notation symbolique musicale dans le domaine de la spatialisation", <http://www.cmusge.ch/HEM/archivage/recherche/Spatialisation.htm>
- [5] Orlarey, Y., Fober, D., Letz, S., " An Algebraic approach to Block Diagram Constructions" Actes des Journées d'Informatique Musicale JIM2002, GMEM Marseille, p.151-158

**Dennis MOSER**

## **Second looks at Second Life: considerations on the conservation of digital ecologies**

### **Abstract**

Libraries, archives, and museums (LAMs) have been the curatorial stewards of cultural heritage for some 5,000 years. The emergence of virtual worlds/immersive online realities as containers/conveyors of cultural heritage is presenting new preservation challenges to LAMs. Second Life is one virtual world that has achieved a level of ubiquity as to serve as model for all such digital environments, having a sufficiently large enough number of common problems from which to learn. The sheer complexity of the Second Life environment argues for a more "ecological" approach. The use of Second Life as an environment for government, commerce, education, and art has resulted in the creation of digital cultural heritage materials that are at risk. Successful approaches are examined, alternatives considered, and new directions are recommended.

**Keywords:** digital preservation, digital ecologies, digital cultural heritage, Second Life, machinima

### **Introduction**

The virtual world Second Life is an example of why the library, archives, and museum (LAM) community needs to become more directly engaged in the decision-making processes of the digital preservation concerns surrounding virtual immersive environments. Taking the conservation of whole ecologies as a preservation strategy is appropriate to providing the context of meaning and relationship in Second Life that might otherwise be lost if current "data set" practices are maintained.

This departs from the usual concepts of digital preservation that focus upon the preservation of data as found in aggregate file structures or systems and their content. The latter, while capable of considerable automation, lends itself to the fragmentation of context and a susceptibility of losing the various relationships that may have initially existed between the aggregates of the data. Such losses are unacceptable and can be avoided by adopting more holistic approaches.

### **Definitions**

Second Life is but one of many virtual worlds, yet it possesses qualities that make it a model for working with other similar immersive and interactive digital environments. It has the defining quality (for "virtual worlds") of what has been described as "persistent user-modifiable content"; this quality is best characterized as content which is created by users that persists within the virtual environment regardless of whether the creator/user is present and can be further modified by other users (Duranske, 2007) [1].

This is important, as it is an operative distinction from those games that utilize virtual, immersive environments. A virtual world may contain games (there are many games that are played in the Second Life environment), but games rarely achieve this status of "virtual world."

### **The Legacy**

Libraries, archives, and museums (LAMs) have been curatorial stewards of cultural heritage for some 5,000 years. The nature of that stewardship has been evolving along with the available technologies. As our cultural manifestations have become increasingly digital in nature, LAMs have incorporated digital technologies to document, acquire, present, and preserve that culture. Documentation has become a part of the historian's domain and virtual worlds, while certainly *au courant*, have existed and flourished for a modestly significant amount of time while gaining the attention of historians who seek to record that emergence.

The definition of "significant time" is related to the rapidity with which digital technologies are embraced and then discarded. Complex environments were being created and used as early as 1998, in such projects as the Virtual UC Santa Cruz, which certainly foreshadows much of what we see today in Second Life. Perhaps not so coincidentally, Philip Rosedale founded Linden Lab in 1999, and for the first two years, the lab's work on Second Life was developing it as an immersive, objective-driven game. From 2001 onward, the focus of Second Life development by Linden Lab has been user-generated content and a community-driven experience [2].

### **The Present**

The emergence of virtual worlds as conveyors of cultural heritage is presenting new preservation challenges to LAMs. It is important to understand some of the distinguishing characteristics of what we mean by "virtual worlds." Virtual reality is no longer a new concept and the number of virtual worlds is growing regularly. There are whole "universes" of virtual worlds of varying complexity, depth, scope and breadth.



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

The highly social dimension that now exists in Second Life is of extreme importance. This increased “sociality” demands a critique of Second Life historiography; indeed, it requires more of an archivist’s, an historian’s, or an ethnographer’s approach and sensibility to talk about how and what to preserve in Second Life, not the perspective of a computer scientist or game programmer.

It is important to distinguish between the different approaches needed to preserve game content versus virtual world content. Games, whether of the massively-multiplayer online type (MMOGs) such as *Eve*, *World of Warcraft*, or *Everquest*, the single player, or the various player versus player type, require different means of capturing the essence of the game, its story arcs, structures, and rules than when trying to capture the ecological complexity of a socially-driven immersive world. It is, however, difficult to decide the appropriate method for capturing MMOGs that approach the qualities that define a virtual world, as defined above. In such cases, a more ethnographic or documentalist approach is certainly appropriate in order to address the highly social nature of such games. Additional difficulties lie in Second Life’s proprietary underlying environment, the user-created content with its issues of “ownership” and the currently “closed” nature (i.e., the rules, regulations, and requirements for entry and use of the software) of the environment.

The use of Second Life as an environment for government, commerce, and education raises legal and fiduciary issues. One has to consider if current records management practices are sufficient. A recent conversation with the founder of a group called “Archivists of Second Life” (an organization in Second Life having as a mission, among others, providing “leadership in the identification of records/archives of historical value to the residents of Second Life”) revealed no encounters with anyone with a records management background [3]. Given the legal requirements and mandates usually associated with government, business, and education, the failure of Second Life to be on the proverbial radar of records managers is troubling at best and represents an area for further research, outreach, and education in itself.

### **The Future(s)**

Linden Lab continues to evolve Second Life in new, but not totally unexpected, directions. The sheer scale of user numbers has drawn attention from the usual suspects, intent on making their fortunes. There is constant talk of the Lab selling or going public. Linden Lab has very deliberately announced various initiatives and strategic partnerships that suggest that they have an interest in promoting Second Life as a platform and technology for others to essentially license from the Lab. Distinction is now made between “Second Life Online Virtual World” and the “Second Life GRID Virtual World Platform” on the Second Life website [4]. The “business” of Linden Lab, or as it is increasingly self-identifying, “Linden Research, Inc.,” is very deliberately appealing to enterprises, government, and education entities to use their platform. While this latter may be a shrewd business strategy, it carries considerable ramifications for any kind of preservation of content. Further, as these entities begin using the Second Life platform to conduct their business, there continue to be concerns about the records of transactions that occur.

That Linden Lab is diversifying their offerings should hardly be surprising in light of the competition from other companies and initiatives to provide similar platforms for virtual worlds. The Lab has released the Second Life viewing client (the “Viewer”) to the open source community, but still maintains very tight control over the Second Life server code, as this is clearly seen by them as where money can be made. In a research partnership with IBM, experiments were conducted on 30 June, 2008, to see if alternative grids would be compatible to allow an avatar from the “official” Second Life grid to traverse to other grids not running on Linden Lab servers. The experiments were a qualified success in that the avatars were able to travel, but the “inventory” did not transfer across with the avatar.

The Open Grid Public Beta, as the experiment was known, was the attempt to develop virtual worlds with compatible underpinnings allowing inter-operability and new levels of customization, user control, or security. Along with the Open Grid, there is a competing free open source initiative called the OpenSimulator that is quite forthright in attempting to create a virtual environment “similar to Second Life” [5]

More recently, Blue Mars, which is still in beta, is gaining attention, but fails to offer some of the inclusivity that is one of Second Life’s hallmarks. By comparison, Second Life runs on Windows, Linux, and Mac OS X; Blue Mars is exclusively Windows. Blue Mars is gamer-oriented (reminiscent of the early incarnations of Second Life) and unlike Second Life, is not user-content driven. In direct competition with Second Life, Blue Mars is making an appeal to educators and business.

### **Conclusions**

Digital preservation, at the most fundamental granularity, is rightly focused upon the preservation of “data” and there are best practices and guidelines in place for doing just that. But the preservation of digital culture, and especially virtual worlds, is more than simply “saving” data.



The preservation of digital culture must include and retain the context of that data which comprises the culture. This is an elusive goal and is at the root of the need for understanding that “the back up is not the archive.” It is important to look at some of the “outside strategies” for more inclusive documentation and context-creation (Moser, 2009) [6]. This includes the utilization of open source solutions for institutions desiring the use of virtual worlds without tying them to the less-open environment of Second Life proper. Running one’s own servers (especially Second Life viewer-compatible ones such as the aforementioned open source OpenSimulator), means being able to more confidently fulfill the obligations and legal requirements that may be incumbent.

A change in documentation is also needed. A more “ethnographic” approach, akin to archeology or cultural anthropology, is beneficial and appropriate for the documentation of digital culture. We are talking about whole systems, or preservation of an ecology, not “data set” preservation. The tools of the ethnographer and cultural anthropologist need to be adapted to use in Second Life. Documentary film in those fields has been highly effective; the application of this approach to Second Life is just beginning. Live motion screen capture and the use of machinima (i.e., in-situ created animated videos of the worlds) are providing documentaries that reflect the richness of the environment. This also includes the use of inworld tools that “follow” the avatar to record the important and elusive avatar-avatar interactions. No one tool is sufficient to capture this environment. We must consider the ecological approach, where each element has a direct connection to the whole. Such environments are multi-modal; our tools must likewise be multi-modal.

Bruce Damer is a pioneer in this approach. Damer has been documenting the history of virtual worlds, focusing upon environments with social interactivity as the emphasis, as opposed to the “gamer”-oriented ones. He has produced videos documenting some of the earliest worlds. His videos approach the subject from within their environments, allowing us a window into those worlds. As these predate the use of machinima, they do have some shortcomings not seen in Second Life machinima. They are still useful if only for the avowedly historical perspective utilized. His footage of Bonnie Devarco’s Virtual UC Santa Cruz is an example succinctly documenting early virtual worlds.<sup>6</sup>

Damer’s approach and work contrast sharply with the approach of his affiliate, Henry Lowood. Lowood is a key member of the “How They Got Game” collaboration at Stanford Humanities Lab, itself a part of the Preserving Virtual Worlds project, funded by the National Digital Information Infrastructure Preservation Program (NDIIPP) funded by the U.S. Library of Congress. Sadly, Lowood’s approach conflates games with virtual worlds with less than satisfactory results. Virtual worlds simply are not the same as games. An example of this “game” approach to a virtual world, “Tabula Rasa: The Final Stand” is lacking in depth and context. The approach simply fails to capture the enormity of what is being examined. Tabula Rasa was a MMOG that was, by definition, a virtual world. What Lowood’s approach has given us is snapshots when we need panoramic videos.

## References

- [1] From the Editor: Are MMO Games “Virtual Worlds?” Benjamin Duranske, 25 February, 2007. <http://virtuallyblind.com/2007/02/25/from-the-editor-are-games-virtual-worlds/> (Accessed, 15 October 2009)
- [2] From the wiki, [http://en.wikipedia.org/wiki/Second\\_Life](http://en.wikipedia.org/wiki/Second_Life) ... the entry has an overview of “what” Second Life is. It provides an excellent explanation of the breadth of experience to be found there. (Accessed 15 October, 2009)
- [3] Taken from the “Archivists in Second Life” Group Charter, the mission statement says they exist to:
  - \*To promote the profession of records/archives preservation and records/archives access in and through Second Life.
  - \* To provide education, research and networking opportunities for archivists in and through Second Life.
  - \* To provide leadership in the identification of records/archives of historical value to the residents of Second Life.”
- [4] This from the Linden Lab URL, <http://lindenlab.com/>, with its links to the online virtual world page, <http://secondlife.com/>, and the GRID page at <http://secondlifegrid.net/>. The GRID is described as a “Virtual World Platform for Business, Education, & Government | Second Life Grid”.
- [5] The OpenSimulator project, while declaring itself to be alpha software, is compatible with the official Second Life viewer, runs on a variety of operating systems including various flavors of Linux, Mac OS X, Free BSD UNIX, and several versions of Windows. It is being promoted as a fully open source alternative to the closed source Linden Lab’s Second Life server.



- [6] "The Avatar in The Archives: Issues of Documentation and Preservation of New Media Art and Virtual Worlds", Dennis Moser, May, 2009. LIDA 2009 Conference Proceedings, LIDA 2009, Dubrovnik/Zadar, Croatia.
- [7] The Internet Archives contains some 96 entries by Damer; while this is a keyword-driven count, most of those are linking to the videos from his collection. The Devarco video can be found, directly, at this URL: [http://www.archive.org/details/vw\\_virtual-ucsc-devarco](http://www.archive.org/details/vw_virtual-ucsc-devarco)



## **Bernard SMITH**

### **Conclusions and report from the parallel sections**

Five years ago the Fondazione Rinascimento Digitale (<http://www.rinascimento-digitale.it/>) came into existence with the explicit task to promote the application of information and communication technologies in the field of cultural heritage. The young Foundation decided almost immediately to hold a conference. In this very same wonderful Teatro della Pergola (<http://www.teatrodellapergola.com/>) the Foundation organised a conference on 14-16 Dec. 2006 on the topics of access and preservation (<http://www.rinascimento-digitale.it/conference2006.phtml>). The conference looked at how new technologies were transforming knowledge and imposing new organisational requirements on our cultural institutions. Most of the papers were loosely classified as either on “digital libraries” or “digital preservation”.

After the success of its first conference the Foundation has continued to work on projects in the broad fields of digital preservation, digital repositories, digital libraries and archives, and persistent identifiers.

Some 18 months ago it was decided to organise this second conference. Over the last 2-3 years much has changed. We saw IFLA (<http://www.ifla.org/>) in their 2009 conference in Milan focus on the future evolution of the library - and the way “libraries would drive access to knowledge”. We saw ICA (<http://www.ica.org/>) becoming increasingly preoccupied with the challenges in exploiting new technologies to preserve “born digital” material. And we saw in the 2009 conference “Museums on the Web” (<http://www.archimuse.com/mw2009/>) major presentations on the institutional changes brought on by social media, on the creation of wiki communities, on digital asset management and digital preservation, on museum Web 2.0 sites, and on young audiences and creators.

So it was against this background that our conference title “CULTURAL HERITAGE: an active role for user communities” was conceived. We felt that the twin topics of access and preservation were just as valid as 3 years ago. However today it seems that users are not only able to adapt to technological changes faster than cultural institutions, but they are also driving innovation, becoming content producers and pushing institutions towards a new user-institution relationship. The Foundation was very fortunate to find support from the Italian Ministero per i Beni e la Attività Culturali and the US Library of Congress, and this produced a great cooperative effort in creating the sessions format, in providing speakers, in attracting high-quality papers and posters, and in promoting the event.

In addition to the support of many prestigious authorities and institutions, around 400 people attended the conference (including the pre- and post-conference tutorials). On the first day we were welcomed by representative from the Comune and Provincia of Firenze, the Regione Toscana, the Ente Cassa di Risparmio di Firenze (the parent organisation of the Foundation), the US Library of Congress, the Italian Ministero per i Beni e la Attività Culturali, and the European Commission.

We had 12 substantial invited talks on the state-of-art and state-of-practice in access and preservation. We had 24 papers presented in 2 parallel sessions: Digital library applications & interactive Web, and Sustainable policies for digital cultural preservation. And we had a poster session with another 11 papers. So a total of 47 speakers and presenters came together from 14 different countries. We saw speakers from major institutions such as the US Library of Congress, Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane, the Italian Archivio di Stato, The British Library, the French Institut National de l’Audi-visuel, the Austrian National Library, the Estonian National Archive, the European Digital Library Foundation and the European Commission. We also saw speakers from a multitude of prestigious academic institutions (North Carolina, Bath, Pisa, British Columbia, Helsinki, and Barcelona come to mind), and from major research labs. such as CNR, IRCAM, IBM and Max Planck.

So all the building blocks for success were present - a good cross section of high-quality cultural institutions and academic-research organisations presenting their activities and latest results.

Let us look more closely at the actual content. I want to retain our 2 traditional topics: access and preservation. My comments are largely based on the presentations and posters from the 2nd day. Many of the invited talks are in themselves masterful overviews of the state-of-art and/or state-of-play in specific fields of relevance to cultural institutions - and I would be doing the authors an injustice to try to summarise their contributions in a few lines. As such my comments should be read along with the collection of invited papers (and a longer version of these comments is available on the conference Web pages).

Equally I will not try to summarise all the papers presented. I have decided to pick out some points that I felt most relevant at the time. These are my personal comments and conclusions, and in no way do I intend to reflect negatively on the quality of the papers not mentioned. And naturally I hope I have understood and



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

captured the key messages of the authors - and I here present my apologies if I mis-quote or mis-represent someones work.

### **Firstly Access**

In this conference we saw two distinct trends concerning access. The first trend was towards very practical, large-scale Digital Libraries with an abundance of high-quality content being digitised and put online, and the second trend concerned a multitude of experiments with Web 2.0 technologies and social networks. Sitting between these two trends were a series of papers looking at the risks and benefits in adopting Web 2.0 technologies and social networking - and there were some concrete suggestions as to how to exploit the opportunities and manage to risks.

### **Large-scale digital libraries**

How could I not but start with Jill Cousins of the European Digital Library Foundation who presented Europeana (<http://www.europeana.eu>) and Max Kaiser of the Austrian National Library who presented EuropeanaConnect (<http://www.europeanaconnect.eu>). Already today the prototype portal links to more than 5 million objects from more than a 1,000 European institutions and collections. And they promise 10 million items by 2010 and 25 million items by 2012. Europeana now has to integrate differing vocabularies, resources discovery tools, harvesters, metadata registries, and a multitude of licence agreements - then they want to add semantic processing, GIS- and time-related query options, etc. - and provide mobile access and on-demand ebooks. I admired the courage and optimism of Europeana - two essential qualities when trying to build a sustainable, large-scale pan-European Digital Library infrastructure and services.

Silvia Gstrein & Günter Mühlberger from the library of the University of Innsbruck described a trans-European ebooks on-demand network bringing together 20 libraries from 10 countries, including 6 national libraries (<http://books2ebooks.eu>). The approach taken gets users to co-fund the initial digitisation of rare or out-of-copyright material. Once the initial users demand has been met, the digitised book is made freely available to the public. What interested me was that a user survey indicated that around 70% of users were prepared to pay 50€ to get a copy of an out-of-print book.

And this is not all - over the 2 days we learned about a real abundance of high-quality content being put on line:

We heard Laura Campbell of the US Library of Congress mention the American Memory Website <http://memory.loc.gov/ammem/index.html>, which makes freely available more than 14 million historical primary source materials. In addition we also have the 200,000 documents in the Global Gateway (<http://international.loc.gov/intldl/intldlhome.html>) and the 100,000 newspaper pages in the US National Digital Newspaper Program (<http://chroniclingamerica.loc.gov>).

On the 1st day Daniel Teruggi mentioned that French Institut National d'Audiovisual provides access to more than 100,000 documents and more than 5,000 hours of audio-visual material.

Thomas Kirchhoff and his co-authors from the museum information systems group in Konstanz University presented BAM the German cultural heritage portal for libraries, archives and museums (<http://www.bam-portal.de>). This is another major portal effort to provide online access to 41 million records in the form of catalogues, repertories and inventories.

Lauri Leht from the National Archives of Estonia (<http://www.ra.ee>) talked about digitising around 5 million images, most from church books, and also putting online all 8 million archival heading thus allowing users to avoid searching through paper records.

Aly Conteh from the British Library & Asaf Tzadok from IBM in Haifa used the example of the National Library of Australia's newspaper Website (<http://newspapers.nla.gov.au/ndp/del/home>), which today has put 104,000 articles online.

Andrea Fojtu of the Czech National Digital Library ([http://www.ndk.cz/project/view?set\\_language=en](http://www.ndk.cz/project/view?set_language=en)), discussed their plans to digitise 1.2 million documents or 350 million pages over next 20 years.

Christoph Müller from the Ibero-American Institute in Berlin (<http://www.iai.spk-berlin.de/>) talked about their plans to put online around 1 million items (plus 900,000 press clippings, 200,000 microforms, more than 70,000 maps, etc.).



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

Friederike Kleinfurher & Kristina Koller from Max Planck Digital Library looked at a development within the eSciDoc portal (<https://www.escidoc.org>). They mentioned that their ViRR project (<http://test-virr.mpdl.mpg.de:8080/virr/>) contains about 20,000 scans of legal artefacts from the Holy Roman Empire.

### **Portals, quality content, usage and sustainability**

In looking at the emergence of these large-scale digital libraries we can see a move by cultural institutions towards services based upon portals. They want to present their content in a more user-friendly way, to offer new levels of interactivity, and to introduce user-oriented services sometimes working together with specific communities of interest.

The list of digital library initiatives given above is just the tip of the iceberg - there are hundreds of other large- and small-scale projects being planned and implemented around the world. Yet in talking with authors most felt that that authentic high-quality content was still lacking on the Web. This means that institutions still saw a need to continue to digitise and expand their online digital collections.

Yet I was worried by the absence of real data on the actual usage of the services already available. And I also missed a real discussion on the sustainability of the investments already made.

### **But what about standards?**

Standards were not a specific topic in many of the papers presented at this conference, nevertheless the impression was that we have moved from a situation of uncertainty (as seen in the 1st conference in 2006) to one of relative clarity and even apparent abundance - few authors mentioned the lack of, or complexity of, today's standards as a barrier or risk.

In my opinion this may not be the true situation. Firstly this conference was associated with pre- and post-event workshops on long-term preservation and Dublin Core. Both offering ample opportunities to focus on standards such as RDF and metadata, digital formats, as well as tools, practices, and approaches to risk management, etc. More importantly some authors hinted at the fact that smaller institutions appear still to have a very naive approach to standards for both digitisation and long-term preservation. I still think that there is a place for standards development and promotion through very practical guidelines, trails and experiments. Equally the success of the pre- and post-conference workshops shows, if anything, an increasing need for tutorials and training courses.

On the other hand we heard from several authors that one of the main advantages of Web 2.0 as a technology platform is that it is an existing (at least semi-standardised) infrastructure and is cost-effective for institutions, even if the different social networks are not interoperable (Kelly & Oppenheim). I might add that the novel integration of RFID tags, GPS and semantic Web by Kauppinen (representing the consortium behind the SMARTMUSEUM project <http://smartmuseum.eu>) also has the advantage of adopting what are becoming cheap and well-defined industrial standard components.

### **Web 2.0 technologies: risks & benefits**

Looking beyond these large digital library efforts we can see more experimental projects exploiting Web 2.0 technologies and social networking as a way to involve users in the creation and maintenance of distributed collections of cultural material.

Smiliana Antonijevic of the Royal Netherlands Academy of Arts and Sciences working with Laura Gurak from the University of Minnesota looked at trust in online interaction. They noted that modern-days users want fast, accurate systems that have some embedded intelligence and can be customised. However they also noted (and I think rightly so) that above all users want systems that are trustworthy. Some of today's digital repositories are certainly moving in that direction - becoming reliable and persistent over time and engendering trust with their users. On the positive side, involving users can transform a static and historical content authority into a dynamic, multi-faceted and evolving body of localised knowledge. The down side is that information provided by users can be incorrect, incomplete, misleading, corrupt or highly biased. The authors called for cultural institutions to understand how to protect their status as trusted authorities and to learn how to transfer trust to external sources of information. However the authors also noted that the main socio-cultural features of trust have remained stable over the years and much can be learned by harvesting results published over the past 30 years.

Brian Kelly from UKOLN & Charles Oppenheim from Loughborough University echoed the point of view expressed by Antonijevic & Gurak, but they went one step further. On the one hand they recognised that Web 2.0 concepts were moving into the cultural institutions (they mentioned Library 2.0 as being an accepted expression - it even has its own wiki entry at [http://en.wikipedia.org/wiki/Library\\_2.0](http://en.wikipedia.org/wiki/Library_2.0), and Archive 2.0 and Museum 2.0 being not far behind). On the other hand they also extended the list of concerns and risks:

services may not be secure, reliable or interoperable, they are open to misuse, and there are still open legal issues concerning the relationship between cultural institutions and users as content providers (and there are also outstanding copyright issues, the risk of misleading or inaccurate information, a failure to respect data protection laws and personal privacy, and the ever-present risk of posting illegal content). On the positive-side the authors noted that social networks are popular and easy to use, they can engage with new user communities, and are cost effective because they exploit an existing infrastructure. So the real issue is to help cultural institutions learn how to manage the risks involved in building and exploiting social Webs. The authors went on to propose a risk/benefits framework where institutions should be explicit about intended use, benefits, risks, miss opportunities when not adopting a new technology as well as the costs when adopting it, how to minimise risk, and the need to clearly document the evidence used in the analysis.

There were three further papers that highlighted the difficulties in understanding and meeting user needs. The first paper was by Alida Isolani from Scuola Normale Superiore di Pisa looking at empowering users without weakening the digital resources. The authors provide a collection of Renaissance texts for humanities scholars (<http://bivio.signum.sns.it>). The contents are valuable and regularly accessed by academics, but the advanced retrieval tools available are not well used - and users tend to access the site in a “traditional way”. The authors concluded that the tools need to be simplified - by making them more complex! More services have to be offered (analysis, note, mark, correct, edit, etc.) and more formats have to be supported. It will be interesting to see if this approach really increases user demand. The second paper was from Fred Stielow of the American Public University system looked at community building in an online university context. The author extended the list of very practical risks/problems he faces daily - ranging from the problem of price negotiations in today's chaotic rights marketplace, through the need to keep costs down when creating metadata and catalogues, to ways to improved tailoring for individual students. The third paper in this group was from Jeremy Hunsinger of Virginia Tech., who looked at the problem of usage of an event-driven memory-bank (<http://www.april16archive.org>). The author reminded us that the memory bank is a collection (or memorial) of digital artefacts contributed after the April 16 tragedy at Virginia Tech. (where 32 people were killed). He mentioned that today his real problem is now a lack of visitors or users, and the author asks “what happened to the users?” without really being able to find an answer!

### Looking beyond the risks: practical experiences

So a risk-benefit analysis is an absolute must, but there are still many ways to exploit Web 2.0 and social networking, keep the risks low, and obtain some valuable and practical results. Lets look rapidly at some practical experiences.

Cèsar Carreras & Frederica Mancini of the Universitat Oberta de Catalunya looked at Web production by small sized institutions. The authors discussed the aims and fears of institutions when faced with the Web 2.0 and the development of social networks. Users can express preferences and opinions, and this virtual community can represent a new life for a small institution (encouraging physical visits, promoting daily discussions, creating empathy with “friends of the museum”, stimulating user content production and commentaries). But how to do this properly? There are risks: sterile tools that create nothing new, alienation of the users through abusive advertising, deforming the institutional identity, etc. In concluding the authors discussed the different approaches. The key for a small institution appears to be to create a local community of interest that supports not only content creation but also has reliable elements of content quality checking and validation (through local professionals, teachers, etc.). The authors stressed that quality checking is an expensive process for a small museum, and yet poor quality content can undo all the benefits that a institution creates in its local community.

Lauri Leht of the Estonian National Archives (<http://www.ra.ee>) looked at involving users in enriching digitised archival material. They have digitised around 5 million images, and put online all 8 million archival heading. Archive volunteers have been employed doing quality checking, helping to understand the content and describe the content in a structured way, and in collecting similar data from different archival sources (remembering that documents are in Estonian, German and Russian with differing alphabets and full of errors).

Aly Conteh from the British Library & Asaf Tzadok from IBM in Haifa talked about ways to foster user collaboration during mass digitisation - using as an example the National Library of Australia's newspaper Website (<http://newspapers.nla.gov.au/ndp/del/home>) which supports collaborative correction of Optical Character Recognition (OCR) output. Generally speaking over 20% of the text of an early 19th century newspaper will not be correctly recognised (and this is equally true of many types of historical text). In-house

checking and re-keying is not a valid option when digitising millions of pages. The authors argued for collaborative user correction and validation to improve the accuracy of OCR results. And improved OCR means better text mining, resource discovery, and overall accessibility. Already this newspaper project, in its first 6 months, brought together nearly 3,000 people to correct 104,000 articles.

The authors concluded by suggesting a hybrid approach: improved OCR technologies for automated text recognition linked with collaborative correction (not just correcting errors but also helping to train and enhance the OCR's engine vocabulary and language analysis features).

### **Top problems: cost, expertise, and information management**

Before moving on the topic of preservation, I would like to close this section on access by referring to the paper of Wendy Duff and co-authors from Toronto University. They looked at the impact of new technology on the museum environment in the US - and based their discussion on semi-structured interviews with 16 US-based senior museum professionals. The 3 most common challenges facing those interviewed were: cost of designing, implementing and maintaining technology, the lack of in-house expertise, and information management.

The starting point is a series of bold quotes saying that "a museum without a collections database and a Web presence is hardly considered professional" and that museums are moving from "object-centred to experience-centred design". However on the down side "not all institutions are using online access equally well" and many funding agencies and museums professional don't fully understand the challenges IT poses for museums.

Finding from the interviews included:

There was no consensus about the extent to which the core-mission of museums has been impacted by new technologies (have they changed the core mission of the museum, does it help to attract a broader audience, or connect better with the local community, or change the way the museum works, or has it altered the way a museum sees itself, ...).

However most agreed that museums have been physically transformed by the proliferation of new technologies (multi-media installations have changed the way exhibitions are held, changes made in dissemination and collections management, for some interviewees 3D imaging is become an increasingly important tool, technology also helps professionals remain curious and creative in developing their expertise and plans for the future, and technology is now an essential tool in linking objects with the information about them, even if the management of legacy data remains a challenge).

For the majority of interviewees the major challenges are: cost of designing, implementing and maintaining technology, the lack of in-house expertise, and information management. Databases need to be created, data needs to be migrated and cleaned, metadata created and maintained, vocabularies need to be agreed upon and shared, ..., and all this takes time, expertise and is expensive. Despite some people being technologically savvy, the majority of people were seen as not computer literate, so high-tech services might be simply an over-kill. Some people noted that poor quality or out-of-date information distributed over the Web can reflect negatively the reputation of the museum and its staff. And many museums don't understand fully the cost/benefit of introducing new technologies.

An important point made by the authors was that many museums don't appear to be dealing in the most efficient and cost effective manner with long-term digital preservation, e.g. digital photos are just being dumped on CD-ROM's and stored on shelves.

### **And now Preservation:**

In this conference our aim was also to review progress in digital preservation technologies and applications (and we should not forget that there was a pre-conference event dedicated to the basic concepts and practices of long-term digital preservation).

Sven Schlarb from the Austrian National Library (and his co-authors from The British Library and ARC - the Austrian Research Centers) looked at the Planets Testbed (<http://www.planets-project.eu>) which is a Web-based application that provides a controlled collaborative environment for scientific experimentation in digital preservation. The authors outlined how the testbed was used, how a tool was tested and assessed, and how the results analysed. Tools can be compared, preserved objects can be validated, emulation experiments performed, and a preservation plan created with recommendations. There is already a community of users sharing the experiments and a lot has been done to provide access to results (preservation services are offered, annotated datasets are available, validation services can check for valid and invalid document types, etc.). The authors closed by noting that the testbed will soon be a freely available public service.

Sam Coppens and co-authors from Ghent University looked at digital preservation using a semantic metadata schema of the PREMIS 2.0 preservation standard (<http://www.loc.gov/standards/premis/>). The authors kicked-off with an impressive list of all the different types of metadata that are needed: descriptive for search and general archive management, binary to describe the bitstreams, technical describing the files, structural for the representation information, preservation indicating provenance, context, etc., and finally rights metadata. The authors have extended PREMIS 2.0 to include the legal roles that people, organisations or software application can have. In concluding the authors stated that employing a 2-layer model allows the upper level with descriptive metadata to be made public (rights permitting), whilst the lower level with the legal roles remains in the hands of the institution.

Christoph Müller from the Ibero-American Institute (and his co-authors including from IPK-Fraunhofer) looked at user demands and preservation requirements for digitisation. The Institute in Berlin (<http://www.iai.spk-berlin.de>) is Europe's largest special collection on Latin America, Portugal, Spain and the Caribbean. The paper looks at the differing, often conflicting, requirements of scholars, librarians, and users. For example scholars want digitised copies to be as authentic as possible and tend to focus on making rare and unpublished material available. Librarians want digitisation to integrate well into their workflow and enable automated quality controls and indexing during scanning. Users (students, public, etc.) want content and context, want full-text search, and want fast and easy access (in particular for exam preparation). The authors now have a "wish list" for features of a future digitisation system, starting with flexible automated digitisation, then interactive quality control, excellent picture quality, easily generated metadata, etc.

Andrea Fojtu and co-authors looked at long term preservation in the Czech National Digital Library ([http://www.ndk.cz/project/view?set\\_language=en](http://www.ndk.cz/project/view?set_language=en)). The authors discussed their digitisation and long-term preservation objectives (e.g. digitisation of 1.2 million documents or 350 million pages over next 20 years using robot scanners). They rightly identify the organisational challenges as being as important as the technical issues (nice expression "institutions must be ready for a business change, well before the scanners produce the first pages"). The authors provided a long list of practical suggestions, ranging from the creation of a digital preservation department to the changes needed in existing in-house workflows and the relocation and retraining of staff.

### **More on metadata!**

Thomas Risse from the L3C research centre (and co-authors from the European Archive, the Hungarian Academy of Science, and the Max-Planck-Institut für Informatik) looked at how to turn stored Webpages into a living Web archive. The authors started by noting that Web archival has value (for scholarly studies, market analyses, IPR disputes, etc.), and there are now emerging industrial services in addition to the usual library and archival organisations. However Web content is highly dynamic, volatile, and in many formats. In addition physical media decays, technologies become obsolescent, encoding standards change, authenticity and integrity are difficult to maintain, etc. To go beyond just "freezing" Web pages, the authors looked at archival fidelity (capturing also the hidden and social Web, but not spam), coherence (identifying, analysing and repairing temporal gaps), and interpretability (ensuring accessibility and usability of the archive including the evolution of terminology, etc.). The authors discussed 2 applications: a "social Web archive" (for dynamic and varied user interactions) and a "streaming archive" (for audio-visual content) - all within a EU-funded project called LiWA (<http://www.liwa-project.eu>).

Felix Engel from FernUniversität Hagen (and his co-authors from Deutsche Nationalbibliothek and the company GLOBIT) looked at context-oriented scientific information retrieval with the specific aim to enable reuse of scientific publications, data and multimedia objects. This requires the capture and storage of additional metadata during all life-phases of the digital object, before, during and after archival. As noted by the authors this supports the goal of digital preservation by enabling reuse (and without being able to contact the object creators). Thus born-digital objects are defined not only as themselves, but also by life-cycle processes such a creation, appraisal, archival and adoption (unpacking, ingestion, adaption, transformation, display, emulation, access, aggregation, contextualisation, etc.) and reuse (including updates to the metadata).

Maristella Agosti and her co-authors from the University of Padua looked at cross-language access to archival metadata. The authors argued for an approach that would allow archival metadata to be both easily machine processable and permit cross-language solutions developed in the library community to be easily adopted by archivists.





Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

## Going beyond “conventional” digital preservation

Jerome Barthelemy from IRCAM in France (and his co-author including from McGill University) looked at real-time audio processing and a notation for contemporary music. The authors want not only to preserve music but also preserve the ability to re-perform the works live, e.g. for modern interactive works that are today completely dependent on a specific hardware and software implementation. They claimed that it is necessary but not sufficient to simply record and preserve outputs. The actual hardware and software used (called a patch) to process the input (from the performer or pre-recorded) must also be preserved. An alternative might be to develop an emulator, but this looks to be fraught with difficulties and uncertainties. Migration, moving from one technical environment to another, has its place. However the authors put forward the idea of virtualisation, or describing the electronic modules employed using abstractions. So a representation of signal processing modules can be found to describe say a violin played together with live electronics, and this can be scored alongside the instrumental parts. Now comes the issue of musical notation.

Dennis Moser from the University of Wyoming looked at conserving digital ecologies such as Second Life. The author's premiss was that our libraries, archives and museums will need to preserve complex environments such as Second Life (<http://secondlife.com/> a user-generated and community-driven “experience”). He argued that it is inappropriate to simply store files, losing the relationships that existed between aggregates of data. Massive-multiplayer online games can pose problems when trying to capture the stories, structures, rules, etc. and the complexity is increased by the closed proprietary environment used in Second Life and with the user-generated content that has separate ownership. Moser argued for a more “ethnographic” approach when dealing with worlds such as Second Life, i.e. preserving an ecology rather than a data set. He suggests that producing video documentaries, with for example machinima (<http://www.machinima.com>), can go some way to capturing what actually happens inside Second Life.

## A need to combat fragmentation in long-term digital preservation work

More generally the different papers and presentations on digital preservation highlighted the complex nature of the problem. In particular when dealing with environments that change and evolve in a disordered and quite rapid way.

We saw in some papers tools being developed that look to be based upon self-defined principles and methods, but which are specific to individual sectors.

The risk is fragmentation. Different ideas and approaches can rapidly lead to isolation and dead-ends when set against a rapidly changing technological and organisational landscape - even more so when users look to be driving innovations.

What we need is to share results and experiences in a cross-domain confrontation. We need to promote a common understanding of the different scenarios and frameworks that underpin efficient digital preservation policies. We need a set of networks (regional, national, European) to:

Avoid useless duplication and foster a single-minded concentration on the long-term sustainability of approaches;

To test research results (to breaking) across a set of complex, cross-disciplinary tasks;

To offer high-quality training/educational events with a focus on real-world problems and using real content.

## From “what might be” to “what is”

At the start of these conclusions I mentioned the objectives of the Fondazione Rinascimento Digitale, but today the real question is concerns what we can expect from the Foundation in the coming years. I personally think that the key will be to make its research, training courses, workshops, and above all its results as relevant as possible to cultural heritage professionals and academics. But to do so it will need feedback - positive and negative. Please go to the Foundations Website - look at the results, use them, adopt them, and tell the Foundation what you think. It needs constructive criticism in order to progress. And suggest to the Foundation what you think it should be doing next.

But criticism is not enough, it needs also to be congratulated when it has done something positive. And I think this conference is a positive result. What we have seen over the last 2 days has been less to do with “what might be” and more to do with “what is” - that is real-world considerations on building large digital collections, the practical reality in working with Web 2.0 technologies, the risks and benefits in working with users within large social networks, and the state-of-play in long-term digital preservation.

For making this conference happen our thanks must go to the Fondazione Rinascimento Digitale, and to the Italian Ministero per i Beni e la Attività Culturali and the US Library of Congress for the support they provided. In addition we had an impressive list of sponsors: the Ente Cassa di Risparmio di Firenze (the parent organisation of the Foundation), the Comune and Provincia of Firenze, the Regione Toscana, and UNESCO. and an equally



impressive list of supporters: CNR, W3C, Liber, IFLA-PAC, European University Institute, europeana, Planets, CIVITA, and many more.

Our thanks must also go to the authors, speakers, and session chairs for providing the content of our conference and for making it such an intellectually stimulating event. Equally our thanks go to all the participants who attended all the sessions, asked questions, created debate, and who made this conference so dynamic and - in many ways a real, tangible, albeit "old-fashioned" social network.

As a final comment and with the desire to build on the embryonic community created over the 2-day conference I ask the organisers to:

Post on the conference Webpage a simple link-page listing all the links mentioned in all the different papers, posters, and presentations (pointers to collections, tools, projects, etc.);

Send out a questionnaire to all attendees asking for comments concerning the conference content and organisation;

Consider ways to build on the community spirit established over the 2 days through a short regular newsletter or even a dedicated Facebook page (the approach must be validated with the user community).



*Empowering users: an active role  
for user communities*

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

**Papers accepted**

## **Maristella AGOSTI, Nicola FERRO, and Gianmaria SILVELLO**

### **Enabling cross-language access to archival metadata**

#### **Abstract**

In this paper we analyze the ratio between Digital Library (DL), archives and multilingualism. We focus our attention on the interoperability issues that need to be faced when you attempt to make different cultural institutions cooperate, to allow a selective and pinpoint online access to their resources, and to enable cross-language retrieval of their materials.

#### **Introduction**

Digital Library (DL) systems have been becoming the fundamental tool for managing, exchanging and searching cultural digital resources and as a research field has seen continuous growth over the last ten years. The central role of DL in fostering access to our cultural heritage is also enhanced by the European Commission which financially supports many projects related to DL, such as the TELplus project<sup>1</sup>, which aims to offer a free service to access the resources of the 48 national libraries of Europe in 20 languages, or the Digital Repository Infrastructure Vision for European Research (DRIVER) project<sup>2</sup>, the goal of which is to develop a pan-European Digital Repository Infrastructure by integrating existing individual repositories from European countries and developing a core number of services, including search, data collection, profiling and recommendation. Furthermore, the "European Commission Working Group on Digital Library Interoperability has the objective of providing recommendations for both a short term and a long term strategy towards the setting up of the European Digital Library as a common multilingual access point to Europe's distributed digital cultural heritage including all types of cultural heritage institutions" [4]. In particular, the recipient of these recommendations is Europeana<sup>3</sup>, which aims at addressing the interoperability issues among European museums, archives, audio-visual archives and libraries for the creation of the "European Digital Library". From this picture we can see that DL are not merely the digital counterpart of traditional libraries, but they are the fundamental tool for pursuing interoperability between different cultural organizations such as libraries, archives and museums. Collecting and managing the resources of these organizations is fundamental for providing wide, distributed and open access to our cultural heritage.

Currently, libraries are the foremost components of DL, this is due to the availability of technologies well-suited for them and that have been adopted by DL since their conception such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) that is the standard de-facto for metadata exchange in distributed environments and the Dublin Core<sup>4</sup> (DC) metadata format which is a tiny and lightweight metadata format that is getting the preponderant mean to exchange information. Archives and museums should adopt these technologies to exploit the services offered by the DL systems; two European projects pursue this goal: the APENet<sup>5</sup> (Archives Portal of Europe on the Internet), which aims to build an Internet Gateway for Documents and Archives in Europe, and the Athena (Access to cultural heritage networks across Europe) project<sup>6</sup>, which aims to reinforce, support and encourage the participation of museums and other institutions coming from those sectors of cultural heritage not fully involved yet in Europeana. Unfortunately, the process of adopting these technologies and exploiting the DL system advanced services is not as straightforward as it is for the libraries; this is due to the nature and the organization of the archives and of the museums as cultural institutions. In this paper we shall concentrate on archives because the problematic issues of museums can be related to those of archives; indeed, often museum resources are described and organized as archival resources. The archival structure is deeply hierarchical and the relationships between the documents must be retained to express their full informational power. These characteristics lead to the development of metadata standards such as the Encoded Archival Description (EAD) which are not particularly well-suited to be used within the DL systems. These standards may be a barrier towards the interoperability between the cultural institutions and towards the automatic processing of the data. These difficulties have moved archives away from full participation in DL, in particular they have limited the access to several services offered by DL systems. For both archives and

1 <http://www.theeuropeanlibrary.org/telplus/>

2 <http://www.driver-repository.eu/>

3 <http://www.europeana.eu/>

4 <http://www.dublincore.org/>

5 <http://www.apenet.eu/>

6 No Website yet available.

libraries, multilingual access to the resources is a key point especially in the European context; indeed, multilingualism also promoted the CACAO European project<sup>7</sup> which aims to offer an innovative approach for accessing, understanding and navigating multilingual textual content in digital libraries. Furthermore, the CACAO infrastructure will be adopted by "The European Library" to promote aggregation of different contents at the European level. In this paper we analyze the problematic issues which could prevent the use of the multilingual services within the archival digital resources. Moreover, we shall propose a methodology that permits us to exploit the techniques adopted by the libraries with the archival metadata, enabling a multilingual access to these valuable resources.

The paper is organized as follows: Section 2 introduces the three main techniques to address metadata-related challenges in a multilinguistic environment. In Section 3 we briefly describe the archival organization and we explain why EAD metadata format does not work well in distributed and multilingual environments. In Section 4 we present our methodology which maps the EAD files into a combination of sets and DC metadata enabling the use of the cross-language techniques. Finally, in section 5 we draw some conclusions.

### Cross-Language Access: Metadata-Related Challenges and Solutions

In the European Union (EU) there is a huge need to provide cross-language access to information; this is due to the diversity and multilingual EU environment where there are 23 official languages spoken in 27 member states. Cross-language access to information leads to problems of both semantic and syntactic interoperability [6]. Many solutions such as those adopted by the CACAO Project aim to address these problems mainly through the use of metadata, which provide access to a multilingual corpus of cultural resources.

A system which has to provide cross-language access to information must address two important metadata-related challenges which can be tackled by specifying the language of the metadata fields [6]: false friends and term ambiguity. To address these issues three main solutions are usually considered:

- Translation: A query formulated in the user language is automatically translated in the other supported languages and then submitted to the system. This solution is not free from the false friends issue.
- Enrichment of Metadata: The aim is to make the intended meaning of information resources explicit and machine-processable, to allow machines and humans to better identify and access the resources. The language would thus be provided in the metadata itself.
- Association to a Class: Terms are associated to a fairly broad class in a library classification system such as the Dewey Decimal Classification (DDC). This is a common solution for the term ambiguity problem and is similar to synsets used in WordNet<sup>8</sup>.

The specification of the language of metadata field enables the full exploitation of metadata for cross-language purposes. If metadata do not come with or cannot be enriched with the language of the field, it is useful to rely on the association to a class technique. This useful technique relies on the use of the subject field of metadata; it is not always possible to determine the subject of a metadata or of a term. This is particularly true for archival metadata where determining the subject can be very difficult.

### Archival Metadata and the EAD Format

An archive is a complex cultural organization which is not simply constituted by a series of objects that have been accumulated and filed with the passing of time. Archives have to keep the context in which their documents have been created and the network of relationships among them in order to preserve their informative content and provide understandable and useful information over time. The context and the relationships between the documents are preserved thanks to the strongly hierarchical organization of the documents inside the archive. Indeed, an archive is divided by fonds and then by sub-fonds and then by series and then by sub-series and so on; at every level we can find documents belonging to a particular division of the archive or documents describing the nature of the considered level of the archive (e.g. a fond, a sub-fonds, etc.).

The union of all these documents, the relationships and the context information permits the full informational power of the archival documents to be maintained. In the digital environment an archive and its components are described by the use of metadata; these need to be able to express and maintain such structure and relationships. The standard format of metadata for representing the complex hierarchical structure of the archive is EAD [7], which reflects the archival structure and holds relations between documents in the archive. In addition, EAD encourages archivists to use collective and multilevel description, and because of its flexible structure and broad applicability, it has been embraced by many repositories [7]. The use of EAD is widespread

<sup>7</sup> <http://www.cacaoproject.eu/home/>

<sup>8</sup> <http://wordnet.princeton.edu/>

in the United States of America and also in the EU; for instance the “Nationaal Archief”<sup>9</sup> in the Netherlands preserves a big collection of EAD metadata in Dutch or the “Archives Napoleon”<sup>10</sup> is based on EAD metadata in French. It is important to include archival metadata in DL because they retain unique and valuable information and at the same time it is very useful to enable multilingual services to access and retrieve them.

Unfortunately, the structure of EAD turns out to be a very large eXtensible Markup Language (XML) file with a deep hierarchical internal structure. On the other hand, EAD allows for several degrees of freedom in tagging practice, which may turn out to be problematic in the automatic processing of EAD files, since it is difficult to know in advance how an institution will use the hierarchical elements. The EAD permissive data model may undermine the very interoperability it is intended to foster. Indeed, it has been underlined that only EAD files meeting stringent best practice guidelines are shareable and searchable [10]. Moreover, there is also a second relevant problem related to the level of material that is being described. The EAD schema rarely requires a standardized description of the level of the materials being described and this possibility is often ignored, as pointed out by Pitti in [7]. Therefore, the access to individual items might be difficult without taking into consideration the whole hierarchy. This issue compromises the possibility of automatically enriching the metadata for multilinguality purposes. A single EAD metadata is used to describe an entire archive, thus in a single metadata we can find very different subjects. With this organization it is very difficult to disambiguate the terms or to identify the subject of metadata; with the EAD metadata the “association to a class” solution is essentially unworkable. Moreover, sharing and searching archival description might be made difficult by the typical size of EAD files which could be several megabytes with a very deep hierarchical structure. Indeed, each EAD file is a hierarchical description of a whole collection of items rather than the description of an individual item. On the other hand, users are often interested in the information described at the item level, which is typically buried very deeply in the hierarchy and might be difficult to reach.

### **A Methodology to Enable Both Cross-Language Access and Exchange of EAD Metadata**

In [2] a solution was proposed to enable the sharing of EAD metadata in a distributed environment and enabling the variable granularity access to the data; this solution maintains also the integrity and the structure of the described archive exploiting OAI-PMH inner structure and the DC metadata; indeed, it is based on a methodology which enables an EAD file to be represented as a combination of OAI-sets and several DC metadata. To properly understand this methodology it is worthwhile briefly describing the functionality of OAI-PMH called selective harvesting and how its internal organization based on OAI-sets can be used to express a hierarchical structure as an organization of nested sets [3].

Selective harvesting is based on the concept of OAI-set, which enables logical data partitioning by defining groups of records. Selective harvesting is the procedure which enables the harvesting only of metadata owned by a specified OAI-set. In OAI-PMH a set is defined by three components: setSpec which is mandatory and a unique identifier for the set within the repository, setName which is a mandatory short human-readable string naming the set, and setDesc which may hold community-specific XML-encoded data about the set. OAI-set organization may be hierarchical, where hierarchy is expressed in the setSpec field by the use of a colon [:] separated list indicating the path from the root of the set hierarchy to the respective node. For example, if we define an OAI-set whose setSpec is “A”, its subset “B” would have “A:B” as setSpec. When a repository defines a set organization it must include set membership information in the headers of the records returned to the harvester requests. We exploit this structure to represent a hierarchical structure such as a tree data structure as an organization of nested sets as shown in Figure 1. Here we can see that each node of the tree can be mapped into a set, where child nodes become proper subsets of the set created from the parent node. Every set is subset of at least one set; the set corresponding to the tree root is the only set without any supersets and every set in the hierarchy is subset of the root set. The external nodes are sets with no subsets. The tree structure is maintained thanks to the nested organization and the relationships between the sets are expressed by the set inclusion order [3]. This methodology allows us to decompose the EAD tree structure into an organization of OAI-sets where the elements belonging to a set are metadata records. The structure of the EAD is maintained by the OAI-sets and the data are mapped into many DC records. As far as the mapping of the actual content of EAD items into DC records is concerned, we adopt the mapping proposed by Prom and Habing [9]. Our solution differs from [9] from a syntactic point-of-view: we propose to maintain the hierarchical structure of EAD throughout an organization of OAI sets containing the DC records mapping the content of EAD items. In [9] the hierarchical structure is maintained by means of several pointers connecting the DC records to the original EAD file.

<sup>9</sup> <http://www.nationaalarchief.nl/>

<sup>10</sup> [http://www.archivesnationales.culture.gouv.fr/chan/chan/archives\\_napoleon-averti.htm](http://www.archivesnationales.culture.gouv.fr/chan/chan/archives_napoleon-averti.htm)

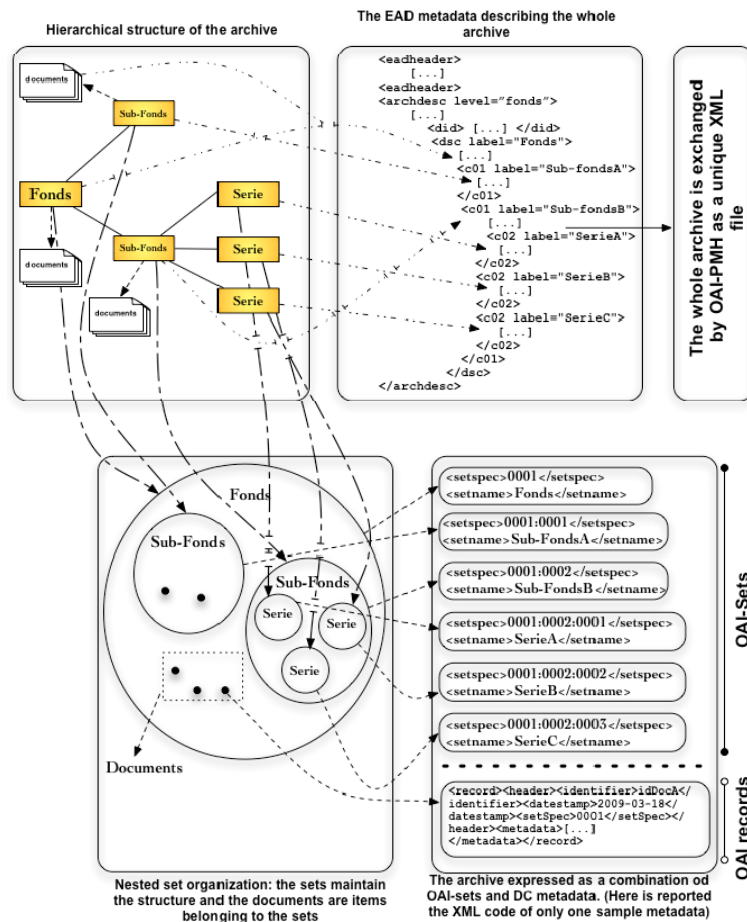


Figure 1 An EAD file mapped into a collection of OAI-sets and DC metadata records.

In Figure 1 we can see two approaches to representing the archival organization and documents. The first approach is the EAD-like one in which the whole archive is mapped inside a single XML file. All information about fonds, sub-fonds or series as well as the documents belonging to a specific archival division are mapped into several XML elements in the same XML file. With this approach we cannot exchange precise metadata through OAI-PMH, rather we have to exchange the whole archive. At the same time it is not possible to determine a specific subject or to access a specific piece of information without considering or accessing the whole hierarchy.

By means of our approach, which graphical representation is shown in the lower part of Figure 1 we can transform archival metadata into a collection of DC metadata and OAI-sets. This solution is particularly well suited for use in the context of the several European projects and in particular for the CACAO project which relies on OAI-PMH to harvest the metadata and on DC records as minimum metadata requirement. In this way the solutions proposed to enable cross-language access to digital contents can be applied also with the archival metadata opening these valuable resources to a significant service offered by the DL technology. Indeed, the decomposition of an archive from a single EAD file into several DC metadata makes it easier to determine the subject of each single metadata and thus to apply the "association to a class" solution; in the same way the metadata enrichment can be adopted because the DC metadata are well-suited to automatic processing. As we can see, thanks to this methodology, the cross-language solutions developed for the library context can be easily adopted in the archival context without any additional efforts.

## Acknowledgments

The work reported in this paper has been partially supported by a grant from the Italian Veneto Region. The study is also partially supported by the TELplus Targeted Project for Digital Libraries, as part of the eContentplus Program of the European Commission (Contract ECP-2006-DILI- 510003).

## References

- [1] A. Bosca and L. Dini. CACAO Project at the TEL@CLEF 2008 Task.
- [2] N. Ferro and G. Silvello. A Methodology for Sharing Archival Descriptive Metadata in a Distributed Environment. In Proc. 12th European Conf. on Research and Advanced Technology for DL (ECDL 2008), pages 268-279. Lecture Notes in Computer Science (LNCS) 5173, Springer, 2008.
- [3] N. Ferro and G. Silvello. The NESTOR Framework: How to Handle Hierarchical Data Structures, In Proc. 13th European Conf. on Research and Advanced Technology for DLs (ECDL 2009), pages 215-226. Lecture Notes in Computer Science 5714, Springer, 2009.
- [4] S. Gradmann. Interoperability of Digital Libraries: Report on the work of the EC working group on DL interoperability. In Seminar on Disclosure and Preservation: Fostering European Culture in The Digital Landscape. National Library of Portugal, September 2007.
- [5] K. Kiesling. Metadata, Metadata, Everywhere - But Where Is the Hook? OCLC Systems & Services, 17(2), pages 84-88, 2001.
- [6] B. Levergood, S. Farrenkopf, and E. Frasnelli. The Specification of the Language of the Field and Interoperability: Cross-Language Access to Catalogues and Online Libraries (CACAO). In Proc. Of the Int'l Conf. on Dublin Core and Metadata Applications 2008, pages 191-196.
- [7] D. V. Pitti. Encoded Archival Description. An Introduction and Overview. D-Lib Magazine, 5(11), 1999.
- [8] C. J. Prom. Does EAD Play Well with Other Metadata Standards? Searching and Retrieving EAD Using the OAI Protocols. Journal of Archival Organization, 1(3), pages 51-72, 2002.
- [9] C. J. Prom and T. G. Habing. Using the Open Archives Initiative Protocols with EAD. In Proc. 2nd ACM/IEEE Joint Conference on Digital Libraries, (JCDL 2002), pages 171-180. ACM, 2002.
- [10] C. J. Prom, C. A. Rishel, S. W. Schwartz, and K. J. Fox. A Unified Platform for Archival Description and Access. In Proc. 7th ACM/IEEE Joint Conf. on DL, (JCDL 2007), pages 157-166. ACM, 2007.



## **Paolo Budroni**

### ***Rethinking how to shape a new matrix for the protection and retention of cultural heritage***

#### **Abstract**

Institutional repositories are currently evolving into a single core of digital collections. Necessary for this evolution is the task of charting the principles used in creating a matrix, which can be derived from the rules for the preservation of cultural heritage in order to endure over the course of time. These principles should be derived in such a way that they develop far beyond what the usual technology-dependent advice and methods would. Technologies come and go, but preservation is a task that should be carried out independently of the currently common line of thinking. Technology with know-how in preservation matters is only one aspect. Preservation is a cultural mission, which must guide our actions. The code that we select in the task of preservation, makes our culture bound behavior reproducible.

**Keywords:** Open Access; open university; information storage; dissemination of information; cultural policy

#### **Introduction**

In 1984, Italo Calvino was officially invited by Harvard University in Cambridge, Massachusetts, to present the "Norton Lectures" which appeared in the preface of his posthumously published book(1). These were a series of six lectures focussing on the topic of "Poetry". This referred to any form of poetic communication, namely, of any: literary, visual, musical form. This brings us to the heart of the matter, since the objects of his studies also refer to objects with a multimedia format that are stored in our current digital repositories for posterity. This is especially true for institutional repositories in the scientific establishment, which is now evolving toward a core of digital collections and whose job is the careful storage of cultural heritage, including all teaching and research output in such a way that it remains accessible, usable and understandable over the long term.

Italo Calvino, as is further stated in the preface, was almost obsessed with clarifying this topic. In this passage, he presented some literary values that needed to be preserved for the next millennium. Calvino named these values as: "Six memos for the next millennium." These are in order: lightness, quickness, exactitude, visibility, multiplicity and, not carried out by him, consistency. With that which Calvino meant in his Norton Lectures on poetry, the author of these lines identifies himself and thus, makes the following issues the focus of his essay:

- What principle is recognized in the field of preservation of cultural heritage and more specifically, how can one use it to create a code that is both derived from universal rules and can be reproduced at any time?
- What are the key ideas and possibilities available to us that will withstand deterioration?
- Technologies come and go. What would be beyond the usual technology-dependent advice and methods of use that would enable us to conduct the act of preservation effectively at all times, regardless of the respective current mindset and the technological platform?
- After these considerations, a momentous question should also be strongly posed: Why are the local research funders so quiet on this topic??

#### **Closing the Generation Gap**

The cultural differences between the younger generation and their parents are expressed today in such a way that we experience in the field of "new technologies in the workplace" the following remarkable fact: the parental generation, in the application of these technologies, obtains the support of their next generation and, consequently, often learns from their next generation while dealing with these technologies. The transfer of know-how and experience therefore occurs in persons from younger to older alike. At the same time, we are seeing for the first time in history, and within the last few years, how a rich, technical apparatus is created – accompanied by the respective know-how – for which the particular content literally has yet to be found. This is now being created mostly by commercial "content providers" with the benefit of hindsight. Formerly it was the reverse: first came the content (presentation of a problem), and then came the solution as a derivative of know-how. In summary: Technology today is the "form", and chronologically speaking, is first in the world, whereas only until the second instance does one wonder, "cui prodest?" Content or various sources of content then result, because they were made for this purpose. Does technology now drive methods and processes or vice-versa? Now it behaves in such a way that today's administrators of their cultural heritage have been trained at a time when there were other methods and decision-making processes in the scientific establishment, and they therefore often applied an apparatus that can be derived from this earlier thinking. It is not timely. These circumstances lead to a strengthening of the generation gap.



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

Corollary to the formation of a matrix: The methods handed down in the scientific establishment from one generation to the next as well as decision-making processes should be called into question to the extent where the following principle may come to bear: What really matters is not whether a particular system is perfect or true ("true" is an arbitrary term), but rather how well it functions for the respective user/user groups. Efficiency is the measure of this "truth".

### **Digital Archives and Cultural Heritage: Rethinking Workflows, Administration, Decision-Making Processes**

The last years were marked by the so-called convergence of technologies. What was left out of the general discourse was the convergence of knowledge (or better stated: the convergence of disciplines). The focus was not, as apparently widely believed, on extreme specialization, but on the convergence of disciplines. The projects that are practiced today arise from this convergence. Digital archives can be taken as an example here. Really successful projects are solely those projects where different disciplines are integrated into one: law, computer science, linguistics, psychology, philosophy of science, communication science, economics, sociology, and so on and so forth.

Admittedly, the Internet was created in academic circles and used for the first time on a broad basis. However, there are a large number of findings, knowledge-related processes and functions that our academic communities are only just now beginning to understand on a broader basis where they in turn then very slowly process such data. Chronologically this takes place after someone else, namely the next generation, has already successfully used it. In this context, the ease with which one can take advantage of the terms for digital repositories offered by Calvino for the Norton Lectures, is truly amazing.

First, a brief note on the term consistency: here, the "shelf life of the data" is meant and the remarks on preservation can be found again throughout the entire text.

*Multiplicity:* This not only stands for the consistency of the data (e.g. metadata) and of work processes including the complexity of the resulting processes in the digital archive (e.g. with preservation and functions of reuse), but also the complexity of the system itself. The next generation has been practicing networking and the building of communities for years. The young students of today are the young scientists of tomorrow. They are the ones who cavort in the social web for years and for example, create lists of "friends" on their Facebook pages and link to multimedia content. That is exactly what they are doing, deliberately and consistently, including the citability of dataset: How many repositories would have to be made available to our traditionally run, educational scientific institutions in Europe in order to benefit – assuming academic methods – similar communities of students and their teachers involved in research?

Corollary to the formation of a matrix: It is time to create similar codified opportunities for our repositories, such as chat rooms for setting up a digital repository. Another measure would be to allow in our universities access to the functions of the systems in order to open them up also to guests (befriended scientists or partners in a project). Privileges and access policies should be the same as for the "own community".

*Lightness:* Lightness requires doing without the exclusive use of centralized logic, central systems, and "central intelligence".

In the context of digital archives, the lightness of operation and the "lightness" of the data are essential characteristics one expects from digital archives. This lightness is synonymous with intuitive controls and should be realized by using common and generally accepted standards. A generally accepted standard does not necessarily mean "a certified standard". The "lightness of the information structured in accordance with generally accepted standards" may be the lowest common denominator of the demands of all concerned, qualified providers of data to a scientific digital archive. Defining "lowest common denominator" could result in the following task: The task of the system is the fostering of communication between those things that are different. In this case, the differences would not be blurred, but, on the contrary, strengthened by highlighting the characteristics of individual digital objects. (Regarding the digital objects: no matter what format and type, they must be provided with contextual information and equipped with technical, descriptive, and long-term digital preservation metadata at their inception).

We have seen how some platforms have prevailed worldwide. What each of these have in common is namely their lightness. I will mention four of them at random: Napster (who can still remember the peer-to-peer exchange of data and the author alert systems used then?), eBay, Amazon, and YouTube. Let's stick with YouTube for a moment and take as an example the lightness of use in uploading content. This is not just about the operation, here the focus is primarily on access to the data delivery process (anyone can upload content) and on access to information (anyone can download content). The lightness is shown in a further example: The

next generation uses Flickr as one of the world's best online photo management and sharing applications. Why doesn't something similar happen in the scientific establishment?

Corollary to the formation of a matrix: The scientific community should have the complete opportunity to store images, pictures and powerpoint presentations in institutional repositories of their institutions quickly, easily and inexpensively (see also multiplicity). Here the emphasis is on the processes of targeted publishing, quoting, commenting, sharing and reuse by and with the interested public, or community.

Another corollary to the formation of a matrix: The previous corollary implies that digital content is reliably archived long-term, provided with appropriate metadata, and more easily and always searchable via a persistent signature (assignable to the digital content and/or the respective author with a digital author identifier).

*Quickness:* The speed of the system is determined by the normalization of data (removal of redundancy) and by dividing it into areas that are associated with a specific task. The concept of the system operator should be determined primarily by efficiency and intuitive recognition.

The efficiency of the user interface – interaction concept – should be distinguished by the fact that the information content on the screen is not too compressed. From the above platforms, I now move to eBay. The operation of the platform requires a low level of literacy, although the services offered can be very complex (e.g. the clarification of payment terms, the resolution of legal issues and questions of logistics.) The user focuses only on his projects, all collateral duties will be met by the system (e.g. allocation of tags for indexing.) The same goes for Amazon. In all of these systems used worldwide, ethnic, political, or linguistic borders are irrelevant, rather, a variant of the game of accessibility becomes effective.

Corollary to the formation of a matrix: The assignability of information should take place quickly and in one effort. The user should always know where he happens to be, what he is doing and how he can cancel a transaction. He may never get lost in the system. Accessibility plays an important role. Accessibility is not just a purely technical issue to be solved, rather, it must be a fixed part of the deliberations at the stage of conceptual planning.

*Exactitude:* The accuracy of the descriptive data is to be achieved through standardization and coding. The submission of data should be conducted according to a set standard, supported by an information code for the individual entries. Thereby, one or more subject catalogs should be used, which are used to classify which have equivalents in other languages and which have cross-references. The systematic accuracy should be determined by syntactic accuracy (the syntax of the user input is determined and controlled through the system in each case), and by semantic accuracy. At this point, an incidental remark made in 1984 by Calvino on "exactitude" can be quoted. For him, exactitude was three things, and for this essay, I will employ namely his second point mentioned: "The evocation of clear, distinctive and memorable visual images, in Italian we have an adjective for this that does not exist in English nor German: *icastico* from the Greek *eikastikós*."(2)

Here it is quite remarkable how, still in subsequent years with the next generation, the term "icon" could prevail in everyday language in its wide-ranging application, and now in both languages precisely this semantic aspect in English and German has become indispensable.

Corollary to the formation of a matrix: Also in this case, the effectiveness is the measure of true accuracy. In general, no "blank spaces" should exist (i.e. "null values" for information derived from queries should not be allowed.)

*Visibility:* In this context, visibility is considered the ease of use. The user should be able to carry out actions based on on-screen information and interact with the system, or alternatively, to retain on-screen information as a consequence of actions the user takes. Furthermore, the degree of traceability is not only significant for the objects (e.g. the origin) but also for the work processes (that is, the traceability in the search history and its results).

Corollary to the formation of a matrix: Not only the visibility of the repositories should be increased. Institutional blogs should be conducted at the interfaces to the repositories of our institutions. The digital content, the content of certified repositories, should be organized in relation to each other or linked (e.g. in order to create new collections of digital objects and to scientifically annotate them). Qualified content should be posted, linked and annotated. Even certainly the linking of the repository and long-preserved material should be linkable to and from platforms like, for example, Twitter. We need a cross-disciplinary approach.

### **Limits to the Digital Preservation of Cultural Heritage**

What prevents us from following these corollaries? Mainly honored traditions and processes (methods), even if some or even all of the above conditions are satisfied. In addition, crucial is the lack of confidence in the know-

how developed to date, as well as in the expertise of active promoters of these processes. The next generation (and a part of it is our young scientists) is often unfortunately not taken seriously because they are considered too young (which is also the reason why this writing does not include something about Twitter-like services nor something about the possibilities that would result from safeguarding cultural heritage for use in mobile applications. At this very instance, developed know-how in general seems methodologically too young.

Corollary to the formation of a matrix: We need to foster the building of confidence in expertise, competence and the available e-infrastructure. It would also be recommended to develop certification mechanisms for digital archives, so that reliability would be resultant and thus the citability of datasets would be enabled. Quality Assurance in all of its facets would then be a part of the certification mechanism.

## Conclusions

In safeguarding cultural heritage, we need a more sophisticated way of thinking for developing solutions. The approach to the design of systems – including systems of thought – should be crossdisciplinary. We need a crossdisciplinary approach.

The instrument itself, the digital repository, should be designed from the beginning as a multimedia marketing tool that enables information transfer and communication between users (data suppliers and consumers) in order to make content consistently available. The possibilities for distribution of information would be far more diverse and knowledge would not only be stored but its implementation would be easier. To accomplish this, a different access system would have to be designed from the beginning, including a sophisticated rights management system. Of course, the data provider should retain all sovereignty over the data. These are not empty words as the solution for these issues today is not of a technical but exclusively of a political nature. Access and restrictions (technically and legally speaking) are mostly an expression of political will. The same goes for accessibility: it is not only challenged technically, but also in all of its associated processes, in data production ranging from the delivery of data to the final data output.

We should therefore redefine the role of users. Users can, in principle, be individual users or institutions. Users can also be subdivided into groups of data providers and consumers. This necessitates a policy of open and free access not only for the consumption of published information, but also in the very process of publishing itself.

With regard to users, generally speaking, they should be empowered, especially that user access should be enhanced, particularly in the following two roles and processes:

The user as a data supplier with free access to the digital archive

The end-user as the beneficiary of the digital content in the digital archive.

It should be possible to guarantee this end-user free access to all published information. He should be given more rights and functionality. This requires a different approach with the wishes of the end users (focus groups) on the system. For example, for the purposes of the reuse of digital content, the end-user should be in a position to be able to implement the new knowledge gained by linking it with other content online (e.g. formation of collections of data sets inside of the repository). In addition, he could be empowered to link individual objects together so that he can therefore “form virtual dossiers” which he can then make available to other users in his community.

Interoperability with other systems and trust in online interaction should be guaranteed to the user (individual user or institution) as a data supplier. For information providers, who are not from the next generation, special training should take place offering these users more expertise, especially in the areas of preservation and reuse of information and techniques of self archiving.

Finally, perhaps the most important corollary on the formation of a matrix comes this time in a personal form (and please forgive the repetition):

What really matters is not whether a particular system conforms to a “true” norm (“true” is an arbitrary term), but rather how well it functions for you the user. Effectiveness is the measure of this “truth”.

## References:

- [1] Italo Calvino, *Lezioni americane. Sei proposte per il prossimo millennio*. Garzanti, Milano, 1988.
- [2] Cited from the German translation of Calvino's work, page 83, in: Italo Calvino, *Sechs Vorschläge für das Nächste Jahrtausend*, Harvard Vorlesungen, Carl Hanser Verlag, Munich Vienna, 1988

## **Cèsar CARRERAS, Federica MANCINI, and Group ÒLIBA**

### ***Improving cultural web production in small-size institutions: strategies to promote heritage in a productive basis***

#### **Abstract**

Social Media is transforming visitors' behavior in Internet, they become active participants in the creation of knowledge instead of passive viewers. Memory institutions are slowly introducing those media to respond to the latest social demands, which involved opening their institutions to public contributions and opinions.

Public create content, develop local or distant virtual communities, through which interests and information are shared. People that belong to those communities express preferences, feel accepted by other peers and develop a more direct relationship with museum staff. Despite the lack of enthusiasm from curators and unsolved problems, the Web 2.0 phenomenon seems to offer new strategies to attract audiences and encourage users to get involved with their cultural institutions.

The present paper attempts to discuss these new strategies by analyzing the aims and expectations as well as their comprehensive fears. Two different case studies in which we are currently involved, allow us to discuss in detail two successful experiences: the Immigration History Museum of Catalonia (<http://www.mhic.net> - Spain) and the Civic Museum of Rovereto (<http://www.museocivico.rovereto.tn.it> - Italy). Although both case studies present differences in the type of collection and objectives, they have employed the new social media to reinforce the existent real communities around both institutions. Keeping strong ties with their local communities also through Internet provide them with original content that can be attractive to distant visitors as well.

Our paper pretends therefore to develop a further reflection on visitors' cultural content creation and strategies that small and medium-size institutions with limited-budget can take on to disseminate their original cultural heritage.

**Keywords:** communities, cultural content creation, Web 2.0

#### **Introduction**

The "mediamorphosis" (Fidler, 1997)<sup>1</sup> currently taking place in cultural institutions has evolved as a tool to support the creation of relevant information. Museums already use Internet as a new mean of communication that allows them to access new publics. However, the development of social networks and the Web 2.0 supposes them a new challenge with potential huge benefits in terms of social response. Providing participatory tools is a way to satisfy public's requirements, who wish to express their own preferences and points of view to Museum curators. Although such freedom of speech may generate uncomfortable situations for those cultural institutions, the phenomenon represents new life for institutions with a virtual community around them.

Indeed virtual communities seem to be useful to attract new audiences because through social groups, people tend to establish emotional relationships (Rheingold, 1994)<sup>2</sup>. It is easy to imagine that through those Web 2.0 applications, lasting relationships can sprout amongst museum's users as well as between them and the institution itself. Museums ply new Web 2.0 tools to encourage virtual audience to repeat a physical visit to the centre, promote daily discussions and often involve visitors with empathy to become "friends of the museum" (Von Appen, Kennedy, Spadaccini, 2006)<sup>3</sup>. Besides, creating virtual communities increases the average time spent in the website, so a feeling of belonging towards the institution (Sommavilla, 2007)<sup>4</sup>.

Likewise, the movement of cultural information, as well as producing very effective phenomena of viral marketing, encourages interested parties to select information according to personal preference and stimulates production of content and users' commentaries. Such content gives fresh-air to traditional institutions such as museums since it is continually renewed by the own public. This democratic approach becomes more reliable

1 Fidler, R. (1997). *Mediamorphosis*. Understanding New Media. Colorado, Pine Forge Press.

2 Rheingold, H. (1994). *The Virtual Community: Homesteading on the Electronic Frontier*. Secker & Warburg. London. Online (updated) edition available from: <http://www.rheingold.com/vc/book/> [Accessed 23 September 2005].

3 Von Appen, K., Kennedy, B. and Spadaccini, J. (2006). *Community Sites & Emerging Sociable Technologies*, in J. Trant and D. Bearman (eds.). *Museums and the Web 2006: Proceedings*, Toronto: Archives & Museum Informatics, published March 1, 2006 at <http://www.archimuse.com/mw2006/papers/vonappen/vonappen.html> editor's note. URL corrected Jan. 21, 2007.

4 Sommavilla D (2007). Nielsen//Netratings comunica la dimensione del Web 2.0. Per la prima volta in italia uno studio quantitativo del fenomeno. [http://www.netratings.com/downloads/n Nielsen\\_netratings\\_Web20.pdf](http://www.netratings.com/downloads/n Nielsen_netratings_Web20.pdf).



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

to the eyes of other users and at no cost for museums, whose only task would be limited to check the correctness of the received material.

Despite these undeniable advantages, it is necessary to combine appropriate strategies with the use of social networks. Otherwise, social media are likely to remain sterile tools, sometimes even harmful to the own aims of cultural institutions. Indeed there are a lot of unfortunate examples of museums that end for abusing the advertisement mailing without drawing a suitable communication strategy. Similarly many of the tools used are often unable to stimulate the creation of communities around the museum, leaving many opportunities offered by technology, unfathomable.

As use of these tools is a fairly recent phenomenon, it is appropriate to do some research on how they can become effective in institutions' strategies. Cultural institutions reflect their own identity in the way they are seen by public. Therefore, communication strategies are key issues in the future museum projection outwards, becoming a kind of institutional DNA.

Therefore, the advent of communications and information technology, gradually introduced by institutions for enhancement of cultural heritage, should be used according to the institutional aims, celebrating its distinctive characteristics and peculiarities. An unconscious use of them increases the risk of deforming the identity of the institution in favor of momentary trends, perhaps destined to disappear in a few years. For this reason, here we aim to analyze two cases with which we were lucky enough to work directly. Both, while adopting two different approaches towards the use of participatory technologies, seek to combine the contribution of users in the production of knowledge with the original mission of the institution, representing two success experiences on which to center our reflection.

### **The museum as a services' lab: the case of the Museo Civico di Rovereto**

(<http://www.museocivico.rovereto.tn.it>)

The Civic Museum, since its opening in 1855, has been defined an institution closely related to its territory. Its main purpose is recording the environment where the community lives, in which citizens can rediscover their own roots, obtain local information and interact with the social and economic tissue. Following such innovative tradition, the Civic Museum of Rovereto decided to open the institution to technological developments and new opportunities' offered by Web 2.0 while remaining faithful to its objectives. The combination of tradition and innovation leads in the case of the Civic Museum to provide services to its citizens and to a community of local actors willing to invest the needed resources for sustaining the museum. The collection of scientific data obtained by its historical equipment, combined with an ongoing dialogue with the productive forces for the promotion of knowledge and citizens' participation, have therefore stimulated a community cooperation in which the museum is rooted.

The website of the Civic MUuseum of Rovereto (<http://www.museocivico.rovereto.tn.it>), for example, offers hundreds of thousands of digitalized cards related to the different fields, providing the public with an archive, continually updated, also available for online access. Some of these cards are geo-referenced, with a GIS supported by the University of Siena, and regularly consulted by experts and professionals of different sectors after subscription.

The virtual space shows both the artifacts owned by the museum as well as an interconnected network of additional information which informs citizens about a phenomenon that occurs in its territory educating them to the use of precautionary measures. External collaborators of the Civic Museum, specially trained for updating the portal content, exchange and transform data gathered in a collaborative way in the fieldwork and made it available to the public in real time.

The Web Directory is reserved to research groups to share scientific tools and information. Access to that space is made possible just by requesting it to the museum. Therefore, the Museum has developed a kind of virtual lab for an active scientific community related to the local museum, which also disseminates such specialized information very fast.

The museum, as a great cultural workshop, can manufacture and sell the knowledge which is constantly added to its container thanks to the community cooperation. Some environmental phenomena taking place around Rovereto, for example, were monitored through citizens' contribution and scientists who scanned and shared the results of their investigation on the web page or on the Web Directory. This approach gives identity to the community and encourages his actors to support the museum by participating in its activities. Especially the latent communities and the ones active just in a physical way are being reinforced by the museum's intervention and by the real and virtual initiatives launched by it.

The dialogue between the productive forces and the great work of economic animation has created enormous repercussions not only in terms of money but also for contacts and new opportunities. The budget of the Civic museum for example is due only in small part to the revenues of public institutions. The rest of the cost of the

structure is covered by selling services to businesses, professionals, governments and to a variety of subjects demonstrating the effectiveness of the model undertaken.

Another important information channel through which the museum proposes scientific news in video format is Sperimentarea.tv (<http://www.sperimentarea.tv/>). The scientific Web TV is an open laboratory, which combines museum's experiences with students' creativity, teachers, researchers and professionals. The television station on air offers a diverse programming broadcast on a fixed schedule besides sections of video on demand where users can choose both large movie productions by international filmmakers and curious, interesting, explanatory movies uploaded by researchers, students and users. Contents can be viewed on any PC or portable audio-visual media, including mobile phones and sent to other users with a function of email alert. Using the audio-visual medium, therefore, the museum draws the attention of the general public, showing how science and technology disciplines are at the service of people.

Thanks to these initiatives provided to its citizens, the Museum has managed to cluster around them a vast network of experts, local businesses and citizens, strongly motivated to cooperate with the institution because of its services strongly useful to all of the residents in the area. These networks of contacts allow the museum to be self-sustainable. Choosing to sell services it opted for the development of a virtual environment accessible only to the professional community, to ensure the quality and reliability of the published information. Analyzing the log data from the Museum website, it is remarkable that users of the Civic Museum download 50 gigabytes per year on average. Although the time's visit is long, it is assumed that users accessing to the platform are really interested in content posted or services provided. The contact established with some of them through the museum in fact confirmed to us that many of them were professionals and experts, who access the virtual resources of the Civic Museum often because it offers information not available through other sources.

### **Life stories': the Immigration History Museum of Catalonia**

(<http://www.mhic.net>)

A completely different approach is offered by the new-born Museum of Immigration History of Catalonia, which opened in 2004 in Sant Adrià de Besòs, a small-town in the neighbourhood of Barcelona (Spain). The Museum attempts to record the life experiences of all the immigrant people that came to Catalonia and the city of Barcelona for a job opportunity and finally settled down here. As happens in other regions and countries all over the world (i.e. USA, Australia, Argentina...), their progress and wealth was due to some extent to the labour force that came from other regions. These anonymous stories are not normally part of the traditional history Museum that is why Immigration Museums have grown as independent institutions.

Sant Adrià de Besòs is an immigrant town, whose people came from different parts of Spain and nowadays from other countries (i.e. Africa, America and Asia) to work in the local industries and construction. The initial aim of the Museum was to create special links with the local community, because the real collection of the Museum was the testimonies of anonymous people who wanted to explain their story.

At the time the Museum was being developed, the UOC (Universitat Oberta de Catalunya) was involved in a European project called COINE to generate a digital archive on-line with testimonies of local communities. It was believed that the Immigration Museum was an excellent field for a pilot experience, so the application was implemented here at the time the portal was being created.

It was such a sophisticated application with metadata tagging and thesauri, as well as multimedia files (i.e. audio, video, images) that most potential content providers, who were old people, were quite afraid of taking part. Therefore, different tests of usability demonstrated that complex applications could not be attractive enough for old people, who did not want to invest much time in explaining their life story.

On the contrary, they could spend sometime in front of a videocamera or taperecorder, so museum curators recorded their stories in those formats. When the social media started to become popular, all those testimonies were posted in channels such as Youtube for video or Odeo for audio. Despite the fact that this was a possible solution, it involves people coming to the Museum and arranging a date for recording themselves. Therefore, the advantages of Internet as a way to break barriers of time and space could not apply here.

Old people were afraid of recording themselves on-line because they require an assistant to help them in the first steps to play with multimedia formats as well as the sophisticated metadata tagging, whereas a presential alternative was too time consuming.

Social media, in this case multiuser-blogs, have become an excellent alternative for the Immigration Museum. The idea behind was a collaboration with the local schools asking students to become assistants of their families relatives. Students for particular schools and courses under the supervision of their teachers have been providing life's stories of their relatives including any kind of media (i.e. images, text, video). They write the story, scan images or provide digital ones and videos if they have and publish in the blog.

Here, the most important issue is who controls the quality of the life's story, in this case the own teacher. Schools take part in the activity as part of their own curricula and an interesting way to show concepts such as

multiculturality and identity. Young students from different origins explain immigration stories that have common traits with their own schoolmates. The possible fears of the Museum of publishing inappropriate contents are vanished since contents are controlled by teachers before publishing. As you can see, it is not opened application of social media, but a regulated one based in an existing local community that provide contents to the institution.

The final aim of the Immigration museum is creating a completely digital archive that combines objects, documents and memories from the own museum collection obtained from researchers linked to the institution together with contributions of the local community. However, the more data is updated in this virtual archive more need will be to create somekind of metatags to favour intelligent search in such database. Probably, this documentation task should rely on the Museum staff.

Combination of specialised documentation with local contributions make the Immigration Museum archive quite an interesting website for experts, which have widely used so far those documents available. One of the problems that should be address is how to allow other users non-related to the local schools to introduce their own stories. Probably, another Web 2.0 application could be the answer to such requirement, but there are still questions about administration and content control that ought to be born in mind.

## Conclusions

Thanks to these two cases analyzed, it is possible to see how the public input in the production of content represents a huge resource for cultural institutions. Through the introduction of participatory applications it is possible to organize a network of contacts that contribute to raise the importance of the web page and as a result of the museum. However, as mentioned previously, it is important that the participative applications used are accompanied by strategies that involve users, motivating them to become concerned. To do so the museum should promote the creation of communities around it and prove useful services to the public.

The web services for example can be structured primarily on the needs of the local community as well as on the objectives of the institution. The resources that the local community can offer are enormous if coordinated with the services and benefits offered by the museum. These if properly designed and accessible to the public can motivate a large number of users to provide their contribution. To offer new services to citizens or to the community around the museum it is necessary to maximize the potential of existing technologies and find the appropriate tools and strategies to achieve the museum's objectives. This is supposed to develop a participative approach in the creation, use and administration of local cultural content that meet practical needs of information and learning.

In fact the activity of a museum (organization of exhibitions and events, renovations, acquisitions, etc.) brings the institution to collaborate with different communities and offline groups: students, scientific associations, schools and voluntary associations. It is also important to encourage this "public" to join the museum's online conversation, providing them specific content and tailored web spaces. Likewise the on-line activity of the museum should create the conditions for groups of people with common interests to join and form an active virtual community on the web.

If there are functional needs (Giacoma, Casali, 2008)<sup>5</sup> in the local tissue to which the museum is able to answer it is very likely that the community can grow. The museum should also be able to design a strategic network to meet practical users' requirements and at the same time to please those relational motivations (Giacoma, Casali, 2008) triggered in social contexts, the satisfaction of which leads to the recurrence of the experience. To enhance users' motivation to participate it is therefore necessary, besides offering specific services, to stimulate their curiosity as well as their desire to share interests and to feel part of a group. 2.0 applications that currently are spreading in the portals of museum's institutions therefore require:

- An active community considering them a means for obtaining some benefits whether they are informative, fun, learning or otherwise.
- Strategies for monitoring of content produced by its community.

Institutions have two options using social networks open to users' contributions. The first one is to have a team of professionals responsible for the control of the material uploaded by users and for the verification of its ethics and honesty. The second one is to leave the community free to regulate and manage inappropriate content as in the Wikipedia. Unfortunately, this second option works only if the active community is composed by a large number of users that can adequately take care of removing or correcting inappropriate or offensive comments. If not there would be the risk of leaving unsuitable material online for a long time or fail to refute or correct the improper content before others use them. This could greatly lower the quality of services offered to the community and undo the benefits that a cultural institution should be able to ensure to its audience.

<sup>5</sup> Giacoma, G., & Casali, D. (2008). Elementi teorici per la progettazione dei Social Network. Tratto da Issuu:

[http://issuu.com/folletto/docs/elementi\\_teorici\\_per\\_la\\_progettazione\\_dei\\_social\\_n](http://issuu.com/folletto/docs/elementi_teorici_per_la_progettazione_dei_social_n)





Small and medium-sized institutions cannot always afford to devote internal resources to the review of the content posted by users because of budget and at the same time when they decide to use participatory applications they do not yet have virtual communities around them able to supervise the content in an independent way. The participation of communities already physically active in the museum, as well as the latent ones, can therefore be fundamental to support small institutions in managing users' contributions.

In both the analyzed cases for example it is guaranteed the quality of published users' content through two different strategies derived from different stories and objectives, but equally effective because based on an active local community, interested in the outcome of its collective collaboration. Virtual communities, born from a real local need can increase and include new distant communities by offering interesting content and a strategic example of how sharing life experiences and knowledge. These two cases thus represent two examples of good practices from which to take the cue when open the door to the plurality of voices of the network.

## **Susanna CHOULIA-KAPELONI, Jenny ALBANI, Polyxeni BOUYIA, Michelle KONDOU, Charikleia LANARA, and Sofia TSILIDOU**

### **E-wandering through the glories and vicissitudes of the Roman Agora and Hadrian's Library at Athens**

#### **Abstract**

This paper briefly deals with a digital application designed for presenting the Roman Agora and the Library of Hadrian, two adjoining civic structures situated in the historic center of Athens. Although they have suffered destruction and alteration in form and function over the centuries, they have played an active role over the years in the life of the city and are at present two major archaeological sites of Athens.

The paper is divided into three parts. The first involves the architecture and history of the buildings. The second part concerns the scope of the project, and the final section gives a short description of the digital application.

The project is conceived as a virtual tour through the two monuments allowing the users to explore them interactively feature by feature and phase by phase. The application includes maps, plans and perspective reconstructions of the monuments, engravings by travellers, photographs, QVR Panoramas, Google Maps and informative texts. A dynamic timeline allows users to follow the most important historical events concerning the city of Athens during the Roman, Byzantine, Ottoman and Modern era. Personalities associated with the history of the monuments are highlighted and specific architectural terms elucidated.

The on-line digital application was designed and financed by the Directorate of Museums, Exhibitions and Educational Programmes of the Hellenic Ministry of Culture and was produced and animated by the company Minimatik, visual + interactive communication and coordinated by Makebelieve, design & consulting. As soon as the project is completed in both Greek and English, free access will be possible through the official website of the Hellenic Ministry of Culture ([www.culture.gr](http://www.culture.gr)). One of the goals of this project is to function as a model in the future for similar applications dealing with other monuments all over Greece.

**Keywords:** The Roman Agora, Hadrian's Library, virtual tour, Athens, archaeological sites

#### **Introduction**

The aim of this paper is to present briefly an on-line digital application for the Roman Agora (=Market) of Julius Caesar and Augustus and the Library erected by the Roman emperor Hadrian at the heart of Athens, the metropolis of Classical civilization. These are monumental building complexes, initially two porticoed enclosures, lie next to each other in the historical center of the capital of modern Greece (Hellas) and, albeit altered in form and function, constitute a major landmark of its topography.

This on-line digital application is a virtual tour through these two monuments. It was designed and financed by the Directorate of Museums, Exhibitions and Educational Programmes of the Hellenic Ministry of Culture and was produced and animated by the Greek company Minimatik, visual + interactive communication and coordinated by Makebelieve, design & consulting. As soon as the project is completed in both Greek and English, free access will be possible through the official website of the Hellenic Ministry of Culture ([www.culture.gr](http://www.culture.gr)).

These monuments are ideal for a digital presentation in view of:

- their elaborate architectural form and function,
- their long history,
- their nodal location within the historical city center and
- their modern function as major and popular archaeological sites.

The state of preservation of both structures varies from fair to poor and the ruins seen today represent a mixture of different building phases. As a result, understanding the Agora and the Library is a challenging task both for tourists and scholars. This bilingual digital application presents them concisely and accessibly to the public worldwide.

The present paper is divided into three parts. The first deals with the architectural type and history of the buildings, the second with the objectives of the digital application and the final section offers a short description of the project.

#### **The Monuments: brief presentation**

The Roman Agora and the Hadrian's Library were built in Athens under Roman authority. The former was financed by Julius Caesar (51-47 BC) and Augustus (19-11/10 BC), whilst the latter was conceived and

donated by the philhellene emperor Hadrian (131-132 AD). Their construction reveals the personal interest of Roman rulers in Athens.

### The Roman Agora



Figure 1 - The Gate of Athena Archegetis

The Roman Agora (1) was built on the site of a crowded open market, which extended up to the principal commercial and civic center of Classical Athens, the famous Athenian Agora. The Roman Agora consists of a rectangular building complex around an open court surrounded by porticoes (stoas). Its plan, which is a quadriporticus in form, recalls that of Roman fora, which catered for religious, political, military and commercial activities. Its two entrances, which face each other, are enhanced by monumental façades (propyla). At the west gates, known as the 'Gate of Athena Archegetis' (=the patron goddess Athena), the road connecting the Ancient Agora with the Roman Market terminated. This road, paved and flanked by shops, was reserved for pedestrians and was called the Wide Road. The Ionic façade on the east side was placed off-center, since it marked the end of an old street. The north side of the structure has not been completely excavated. The Roman Market at Athens had storerooms (horreae) on the west side, shops (tabernae) on the east side and a fountain on the south side. Although some offices connected with the operation of the market (e.g. control of prices and weights) may have been housed in rooms across the south side of the structure, it would seem that the office of the market officials is to be sought west of the Roman Market. The paucity of shops may suggest that they were intended only for wholesale traders, whereas retail trade may have taken place in the court and the stoas.

Before the construction of the Roman Market a marble octagonal tower existed directly east of it (2). It combined a hydraulic clock in its interior and a sun-dial and vane on its exterior. This sophisticated ancient mechanism, an invention of Andronicus of Cyrrus, in Macedonia, is known as the 'Clock of Andronicus from Cyrrus' or 'Tower of Winds', due to the relief frieze around the upper part of the exterior of the building that displays personifications of eight winds. This unique structure is apparently a creation of the 2nd century BC, a period when technology excelled and developed rapidly.

### The Roman Agora from the Byzantine to the Modern period



Figure 2: The Church of Prophet Elijah & the Church of the Taxiarchis, Engraving by Th. du Moncel.

A three-aisled basilica was built within the Roman Agora during the Early Christian period (4th-6th century AD). It was converted into a mosque in the Ottoman period (1456-1830 AD). Likewise, in the Early Christian times the Tower of Winds served as a baptistery of a nearby church, whilst in the 18th century it housed an Ottoman Tekkés (holy place). Two adjacent domed cross-in-square churches, the Church of the Taxiarchis (3) and the Church of Prophet Elijah (4), were erected on the north side of the Market complex in the Middle Byzantine period (11th-12th century AD). The Church of the Taxiarchis was demolished in 1852 and was replaced by a new church dedicated to the Archangels and the Virgin. The Church of Prophet Elijah was refurbished in a rudimentary fashion after the Greek War of Independence (1821-1828 AD), in order to serve as a hospital and in 1848 it was finally demolished.

During Ottoman rule in Greece the little Church of the Soteira tes Pazaroportas, dedicated to the Virgin, was built on the north end of the West Gate of the Market. It was demolished after the liberation of Greece (1829 AD).

### Hadrian's Library



Figure 3: Hadrian's Library

The Library (5), directly north of the Roman Market, was a rectangular two-storied building around a peristyle courtyard with a pool in the middle. It was provided with reading rooms, spaces for the storage of papyri and with lecture halls. The building was approached from the west through a propylon with four Corinthian columns of Phrygian marble which was flanked on both sides by a row of seven columns located on pedestals of green Karystian marble. This substantial intellectual and cultural center of Athens, which also housed the archives of the city, was badly damaged during the invasion of the Heruli, a Germanic tribe, in 267 AD.

### **Hadrian's Library from the Byzantine to the Modern period**

In the early 5th century AD, the Library of Hadrian was repaired. During the first half of this century, the central pool was filled in and on this spot a luxurious tetraconch church (6) was built by the Eparch of Illyricum, Herculus, or, in the view of other scholars, the empress Eudocia, a native of Athens. At the end of the 6th century, the so-called Tetraconch was severely damaged, probably because of some Slavic invasion, and converted into a three-aisled basilica, which was destroyed at the late 11th century AD.

A little, single-aisled domed church, the Megale Panagia (7), dedicated to the Virgin, replaced the basilica. At its south end a second church, dedicated to the Holy Trinity and of the same type and size, was annexed during the 17th century or a little after 1715. After the liberation of Greece (1828), the Megale Panagia housed the state collection of antiquities. It was demolished, after being burnt, in 1885 to allow archaeological excavations. The Hagioi Asomatoi "sta skalia" (8), a church dedicated to the Archangels, abutted the north colonnade of the west monumental façade of the Library during the 11th-12th century AD. From 1576, when the church was renovated and decorated with wall-paintings, if not from the time of its construction, it belonged to the eminent Byzantine family of the Chalkokondylai. This church in its turn was demolished in 1849.

During the Ottoman occupation of Athens, a commercial center with more than 100 shops, known as the "Upper Bazaar", grew up in the area of the Library. This extremely lively area operated until 1884, when it was destroyed in an enormous conflagration. In the southwest corner of the Library the residence of the Turkish governor of the city, Voevodaliki (9), was also erected. In 1835 the governor's mansion was converted into barracks and later into a prison.

### **The Roman Agora and Hadrian's Library at present – the scope of the project**

Today the Roman Agora and the Library of Hadrian form two adjacent archaeological sites, where systematic excavation, restoration and rehabilitation continue. Their location at the heart of the historical center of Athens, in the pivotal area of Monastiraki, opposite the Metro station of the same name, makes them a familiar landmark for Athenians and a popular sight for both Greek and foreign visitors. Very few, however, are aware of the historical events associated with their erection, the politics pursued through it, their architectural prototypes, their various building phases, and the drastic changes in their use over the twenty centuries of their existence.

The digital application informs its visitors of these matters by means of a virtual tour through the monuments, in time and space. References to related events and personalities, a combination of texts and images presented interactively make the e-wandering an engaging experience.

The goals of this project are:

- to use the Internet as a medium to permit free use and easy viewing of the monuments by people all over the world,
- to invite the visitor to the website, whether an ordinary inhabitant of Athens who is in the habit of hurrying past the monuments or a potential tourist, to stroll through them,
- to help users grasp the history and cultural context of the monuments and their individual features and recall this information when they finally visit the sites in person,
- to offer a better understanding of the monuments for those who have already visited them and
- to function as a model in the future for similar applications dealing with other monuments, which can then be produced in co-operation with various provincial Ephorates of Antiquities of the Hellenic Ministry of Culture.

### **On-line digital application for the Roman Agora and Hadrian's Library**

The aim of the on-line digital application for the Roman Agora and Hadrian's Library is to supply students, scholars and the general public with easily accessible, up-to-date and expert material on a digital and visually dynamic platform. The concept behind the navigation design is to invite users to experience the two monuments, rather than simply to access information, by taking advantage of the interaction possibilities the web offers as a medium. As part of this goal, a virtual tour and dynamic timeline were implemented, thus allowing users to explore the two monuments feature by feature and phase by phase and to follow the complex patterns of construction, modification and destruction from antiquity to the present day. Thematic essays written and reviewed by experts are accompanied by a plethora of images, plans and maps and two QVR Panoramas. A simplified version of the virtual tour has also been made available in a customized fully functional Google Maps environment, which offers a completely interactive experience.

### **Virtual Tour & Timeline**

A two-dimensional interactive resizable plan including nearly thirty clickable individual features, including buildings, parts of buildings and major monuments, inside and near the two monuments, functions in

conjunction with an interactive timeline. The visitor can navigate the monuments interchangeably, by place and/or time.



Figure 4: Snapshot detail of the Virtual Tour and Timeline.

The timeline is divided in four main periods, Roman, Byzantine, Ottoman and Modern. Clicking on one of the four periods highlights all the monuments that were either built or reconstructed during this period. Each period is subdivided in a number of dates. Hovering over a date displays monument-specific and general information, thus helping the user grasp the history and cultural context. Clicking on a specific date highlights the respective feature

in the plan and brings up basic information, in the form of text and image, regarding the feature, within the current interface. Clicking on an individual feature in the plan highlights the feature and corresponding date, while realigning the timeline if necessary, and bringing up basic information.

The user may view complete information on the feature by clicking on a 'read more' link which opens a minisite in a new window. The minisite contains extensive information divided by tabs (history, description, special features, references etc.). Each tab contains an image gallery, corresponding to the information provided.

### Main Navigation – Information Indexing

The website offers a different means of accessing the rich information associated with the monuments, dividing it into thematic sections (Monuments, People, and History). This twofold method of accessing information caters for accessibility issues and ease of use. Thus the user may choose how to navigate, either through lists or engaging in the interactive experience of the Virtual Tour & Timeline section.

### Google maps & QVR Panoramas

Another section of the website is focused on the present. The monuments are localized in their contemporary environment through the use of the Google maps application and overlays. Two QVR Panoramas provide a current view by means of video footage, accompanied by texts offering an ideal walkthrough.

### Conclusions

The combination of archaeological data and modern technology is now a reality and one of the objectives of the Hellenic Ministry of Culture. It is desirable that users of Internet be acquainted with Greek monuments by electronic means before their physical visit to Greece. The digital application presented above is the first effort in this direction, but it is hoped that it will serve as a model for similar projects pertaining to other monuments all over Greece.

### References

- [1] J. M. Camp, *The archaeology of Athens*, New Haven-London 2001, passim. M. Hoff, *The Roman Agora at Athens* (diss. Boston University 1988). Idem, "Early history of the Roman Agora at Athens" in: *The Greek Renaissance in the Roman empire*, Papers from the Xth British Museum Classical Colloquium, BICS: Suppl. 1989, 1-8. Idem, "The so-called Agoranomion and the Imperial cult in Julio-Claudian Athens", *Archäologischer Anzeiger* 109 (1994) 93-117. T.L. Shear, Jr., "Athens: From city-state to provincial town", *Hesperia* 50 (1981) 356-377. A. Choremi-Spetsieri, "Πολεοδομική εξέλιξη και μνημειώδη κτήρια στην Αθήνα κατά την εποχή του Αυγούστου και του Αδριανού", Αθήναι. Από την κλασική εποχή έως σήμερα (5ος αι. π.Χ. – 2000 μ.Χ.), Athens 2000, 166-193. D. Sourlas, "Νεότερα στοιχεία για τη Ρωμαϊκή Αγορά της Αθήνας", in: S. Vlizos (ed.), *Athens During the Roman Period. Recent Discoveries, New Evidence* [Μουσείο Benaki, 4th Supplement], Athens 2008, 99-114.
- [2] H. J. Kienast, *Ο Πύργος των Ανέμων. Οι Αέρηδες*, Athens 2007. J. Von Frieden, *OIKIA KYPPHETOY – Studien zum sogenannten Turm der Winde in Athen*, Rome 1983.
- [3] Ch. Bouras, "The Middle-Byzantine Athenian Church of the Taxiarchis near the Roman Agora", in: J. Herrin, M. Mullett and C. Otten-Froux (eds.), *Mosaic. Festschrift for A. H. S. Megaw* [British School at Athens, Studies 8], Great Britain 2001, 69-74.
- [4] S. Sinos, "Die sogenannte Kirche des Hagios Elias zu Athen", *Byzantinische Zeitschrift* 64 (1971), 351-361.
- [5] J. Travlos, *Pictorial Dictionary of Ancient Athens*, London 1971, 244-252. D. Willers, *Hadrians Panhellenisches Programm: Archäologische Beiträge zur Neugestaltung Athens durch Hadrian*, Basel 1990, 14-21. J. Knithakis, E. Symboulidou, "Νέα στοιχεία δια την Βιβλιοθήκη του Αδριανού", *Αρχαιολογικόν Δελτίον* 24 (1969), Μελέται, 107-117. I. Tigginaga, "Η μεγάλη ανατολική αίθουσα της



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

- βιβλιοθήκης του Αδριανού (βιβλιοστάσιο). Αρχιτεκτονική μελέτη – πρόταση συντήρησης και αποκατάστασης", Αρχαιολογικόν Δελτίον 54 (1999), Μελέται, 285-326. A. Choremi-Spetsieri, "Library of Hadrian at Athens. Recent Finds", *Ostraka* 5 (1995) 137-147. A. Choremi-Spetsieri, I. Tigginaga, "Η Βιβλιοθήκη του Αδριανού στην Αθήνα. Τα ανασκαφικά δεδομένα", in: S. Vlivos, *op. cit.* (n. 1), 115-131.
- [6] I. Travlos, "Το τετράκογχο οικοδόμημα της Βιβλιοθήκης του Αδριανού", *Φίλια έτη εις Γ. Ε. Μυλωνάν, I*, Athens 1986, 343-347. *Idem*, *Πολοδομική εξέλιξις των Αθηνών*, Athens 2005 (3rd ed.) 139, 141.
- [7] Ch. Bouras, "Επανεξέταση της Μεγάλης Παναγίας Αθηνών", *Δελτίον της Χριστιανικής Αρχαιολογικής Εταιρείας ΚΖ'* (2006) 25-34.
- [8] E. Touloupa, "Ο Άγιος Ασώματος στα σκαλιά", *Ευφρόσυρον. Αφιέρωμα στον Μανόλη Χατζηδάκη, II*, Athens 1992, 593-600.
- [9] J. Knithakis, F. Mallouchou, G. Tigginaga, "Το Βοεβοδαλίκι της Αθήνας", in: Ch. Bouras (ed.), *Επώνυμα αρχοντικά των χρόνων της Τουρκοκρατίας*, Athens 1986, 107-124.

## **Alida ISOLANI, Dianella LOMBARDINI, Claudia LO RITO, Daniele MAROTTA, and Cinzia TOZZINI**

### **Empowering users without weakening digital resources: is this possible?**

#### **Abstract**

The aim of this paper is to address an issue regarding the digital resources created for the Humanities by both Signum-SNS and INSR. This issue focuses on the fact that these digital resources are not fully exploited by users. As these resources are created for Humanities scholars, CRIBeCu (now Signum) has started a pioneering work by means of a synergy between humanists and computer scientists. This has allowed cutting-edge research to proceed along with applied research and attention to users. Furthermore, humanists have suggested IT research and they have received, in turn, inputs from informatic results.

It can be claimed that the major potential of digital resources lies in their flexibility, although such a flexibility implies an high level of complexity. Despite the facilities put at disposal of users, the latter are discouraged by difficulties involved in the use of them.

Two typologies of resources have been created:

- digital collections for XML documents' search and consultation (e.g. BIVIO)
- collaborative tools for XML documents' management and advanced search (e.g. TauRo)

As statistical analysis demonstrates, users approach digital resources according to a traditional perspective: digital tools are consulted as a digital reproduction of paper documents, while the tools specifically developed for text management and analysis are disregarded. This approach enables users to exploit only the basic functions of the digital resources so that performance and fruition are less effective.

In this paper we will examine the reasons determining this phenomenon in order to develop a strategy which can contribute making digital resources more effective.

**Keywords:** digital library; search engine; XML document; collaborative tool; Renaissance

#### **Introduction**

Signum [1] (formerly CRIBeCu) is a computer science laboratory of Scuola Normale Superiore (Pisa) that provides and designs digital resources for specialists in the Humanities. It includes a team of humanists and computer scientists who cooperate with Istituto Nazionale di Studi sul Rinascimento (Florence).

The major purposes of Signum are digital humanities research, effective application of this research to digital resources and, finally, exchange of ideas with similar research groups.

The results of the activities carried out by this laboratory can be evaluated through an examination of the feedback from users.

Accordingly, this paper will present an inquiry made by Signum about the attitude of users towards two main typologies of resources, exemplified by the following projects: BIVIO – Virtual library online [2]; TauRo – search and advanced management system for XML documents [3], which have been developed by Signum.

The results of this inquiry point to a limited exploitation of digital resources. This indicates that either users are unable to work with these tools or that the latter are inadequate.

A detailed analysis of the actual needs and capabilities of consumers is required to provide more effective tools. In particular, it has been observed that in recent years users tend to prefer new digital tools to the resources provided by computational linguistics.

#### **Case studies analysis**

BIVIO was created as a response to a need to have Renaissance texts available and searchable online. BIVIO can be seen as a model case study. The presentation page of the project may be quoted: "The purpose is to guide philosophical, historical, artistic, philological research to create a virtual library, able to offer rare texts in their more significant editions and translations, made available thanks to adequate IT systems that guarantee multi-level information retrieval: from the easier, as words frequency, to the more sophisticated, apt to analyse the content". This statement clearly demonstrates that the purpose of the project is to stimulate text analysis through specific research tools. It is also stated that "the project offers aids (e.g. quotations lists and iconographical corpora) to a deep comprehension of the period in question".

On the website of BIVIO a resource access has been arranged which is parallel to those of libraries catalogues, and provides textual documents. An IT system has also been developed, which enables to analyse and compare texts in an innovative way. This system provides information which differs from any 'paper-like' approaches, such as visualisation of occurrences in the text retrieval results (snippet lists), which allows

comparison between different occurrences of the same word and text search based on both two distant words and different variants of the same word. Nevertheless, it has been noted that these available tools are not fully exploited by users.

Let us focus on the reasons determining this phenomenon.

Analysis of access statistics indicates that BIVIO users are essentially humanistic operators. Indeed, the outer websites usually reaching BIVIO are mostly academic sites or sites which are concerned with themes treated in BIVIO (fig.1). Another evidence confirming a correspondence between the theoretical and the actual target of BIVIO regards the researches made by main search engines: 70% of them queries citations, works titles or specific authors (fig.2).

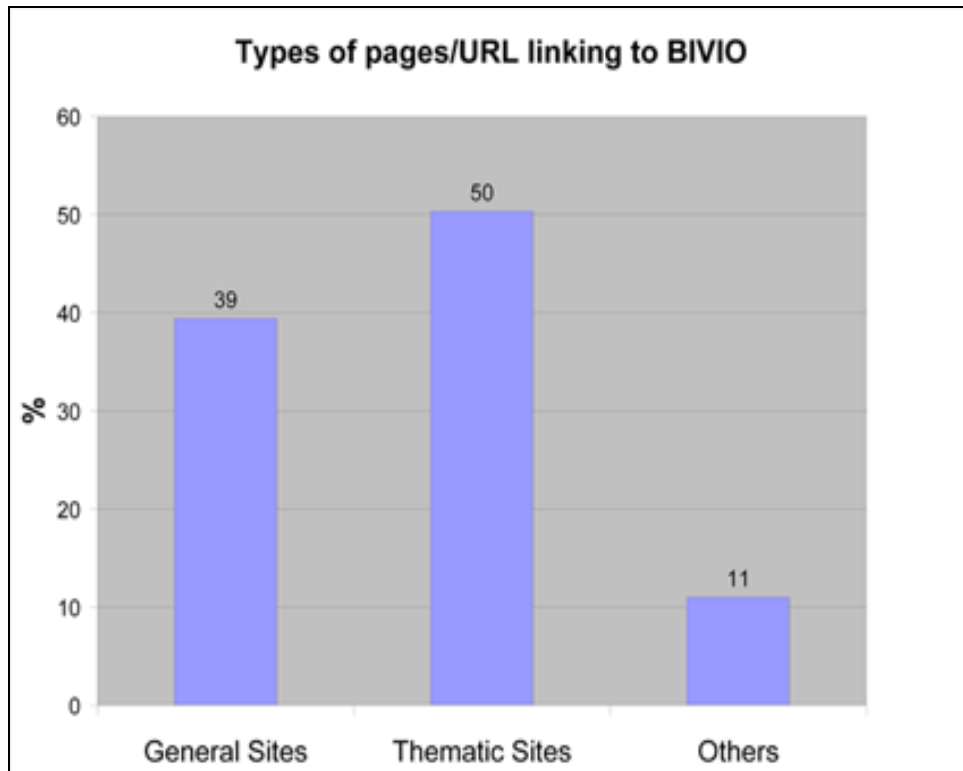


Figure 1 - Types of pages/URL linking to BIVIO

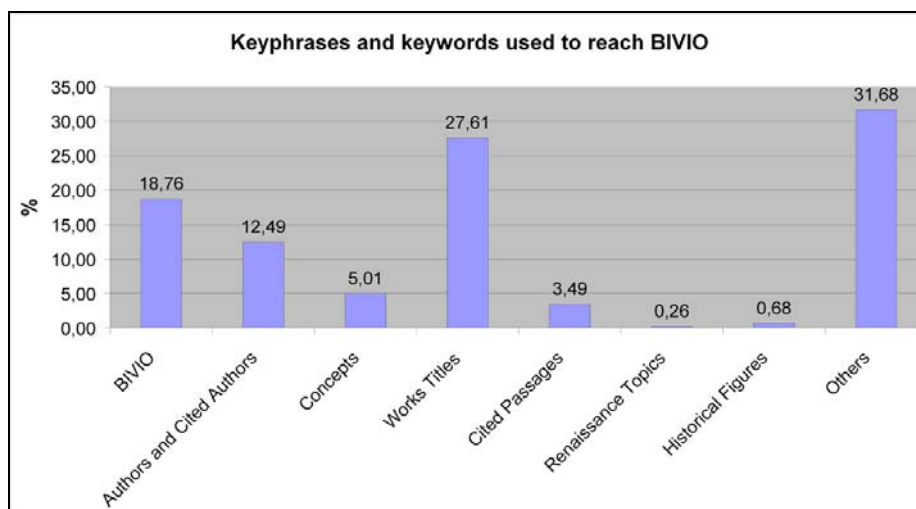


Figure 2 - Keyphrases and keywords used to reach BIVIO



Moreover, about 75% of users add BIVIO to their favourite links, which indicates an appreciation of this resource.

These data seem to demonstrate that BIVIO fully matches the needs which originally led to its creation, and that the users feedback is positive. However, a closer examination of these data rather indicates that users do not exploit the full potential of the tools provided by Signum.

Apart from the evaluation of any visits with a duration of less than 30 seconds (about 50% of the total), statistical analysis reveals that the interaction with the site reflects a traditional approach while the innovative retrieval tools appear to be scarcely used (fig.3).



Figure 3 - Most visited pages

The results of this statistical analysis as well as the continuous interchange with the humanists may be helpful for interpreting the general picture above. Accordingly, some hypotheses can be advanced which have an important bearing on the promotion of original digital humanities research and new applications to cultural heritage:

users have a traditional working approach; generally speaking, they prefer to read texts while rarely asking for analysis tools.

users perceive information provided by BIVIO tools as something similar to paper data (rhyme concordance, word occurrence, lexicon, etc.) which are generally used for particular text analysis and for specific studies.

a specific training to use the IT tools applied to cultural heritage is lacking; as users may be not familiar with digital resources, they persist in using only the basic functions of these resources.

users do not trust research systems based on IT tools and so results are not considered reliable.

BIVIO does not fully satisfy Humanities scholars demands.

It has been suggested that hypothesis 5 can be easily verified through a direct inquiry by users. Statistics, however, lead us to discard this approach because users without any adequate training in managing these resources may not have a critical opinion.

Hypotheses 1 and 3 specifically refer to users and their training while hypotheses 2 and 4 are connected with the nature of the tools under examination and their appreciation by users.

In order to settle these problems, it is needed to promote the dissemination of digital resources such as BIVIO through a specific training for users. Furthermore, generalist systems may suggest to adopt strategies developing an easy access to digital resources, which maintain a high, scientific and reliable standard of the product.

Differently from a digital resource such as BIVIO, Signum has projected TauRo not in view of the actual needs of users, but rather by means of a new approach typical of Web 2.0. Therefore, a collaborative tool for XML documents' management and advanced search has been created, which is able to exploit all the capabilities of a search engine previously realized by Signum: TauRo-core.

The idea of bringing together a community of XML documents users was substantially unsuccessful because a very few users have been involved. As a consequence, collaborative tools are not currently used and users limit their access to public resources consultation (according to TauRo statistical data).

Moreover, TauRo is not able to test the capabilities of the search engine because users do not use its advanced functions and only scroll documents or make simple queries.

Users show a similar attitude to two different digital resources such as BIVIO and TauRo for similar reasons.

Furthermore, the collaborative nature of a resource like TauRo leads us to focus on this specific aspect. Users load a very few documents, create a very few collections, and essentially do not share them. The following hypotheses can be put forward to explain this phenomenon:

XML format is not familiar amongst humanists.

XML format is used by scholars who do not like to share their own work.

TauRo system is too complex for users.

All these hypotheses suggest that a general solution for these problems is to simplify TauRo. This is possible by making TauRo even more complex and by developing a system that, thanks to a wide range of services, will make it more accessible for a larger audience and will attract specialists, who will be encouraged to use it as a working platform. Signum will also try to make TauRo interface more user-friendly. Indeed, in the future, users will be able to load documents in more common formats, which will be converted into XML format by the system, so as to preserve all the capabilities of the search engine. TauRo will also be provided with proper tools to analyse, note, mark, correct and edit loaded documents.

## Conclusions

Analysis of users' attitude is important in order to develop new strategies for future applications, and to open a new path toward both humanistic and computer science research.

In this paper, we have focused on some problems regarding the use of two model digital resources realized by Signum. Furthermore, some solutions have been suggested on the basis of the key idea that digital humanities research does not simply target users demands, but it also helps to acquire new skills and to master new working methods.

Users training needs to be urgently increased for a more effective and aware approach to digital resources, while the latter should be more accessible and useful.

## References

- [1] <http://www.signum.sns.it/>
- [2] <http://bivio.signum.sns.it/>
- [3] <http://tauro.signum.sns.it/>



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

## **Thomas KIRCHHOFF, Werner SCHWEIBENZ, Jörn SIEGLERSCHMIDT**

### **BAM - A German portal for cultural heritage as a single point of access for users**

#### **Abstract**

BAM – the joint portal for libraries, archives, museums in Germany intends to become a single point of access for cultural content and serves users who do not want to search several different databases at different servers using different search interfaces and vocabularies for access. In addition to combining different information services from different institutions in one point of access, BAM can also serve as a portal for a single institution's libraries, archives, museums and media centres. BAM also tries to increase the visibility of the digital objects in the collections of the participants by cooperating with Wikipedia Germany and enriching articles with a link to content in BAM.

**Keywords:** Cultural heritage, portal, museums, libraries, archives, access

#### **Introduction**

When looking for digital cultural heritage information, users do not care whether the information they require is stored in a library, an archive or a museum [1, 2]. In the digital realm it is no longer relevant whether the original materials that are now available in a digital form were stored in a library or a museum or an archive [3]. The current development of libraries, archives or museums goes towards a digital memory institution where the information of all institutions is available online. BAM – the joint portal of Libraries (in German: Bibliotheken), Archives, Museums intends to set up such a digital memory institution for Germany providing a single point of access to users who do not want to search several different databases at different servers using different search interfaces and vocabularies. Such a single point of access is a major improvement because in Germany does exist a lot of digital resources but they are scattered all over the Internet like islands in the sea. In order to find these materials, the users have to know that these islands of digital materials exist, where they are located and what kind of resources they hold. So the users have to do some island hopping in order to find the information they are looking for. In addition, to access such a treasure island, they need to know the magic words Open Sesame as in Ali Baba's tale in One Thousand and One Nights, i.e. they must understand the various interfaces, know the right terminology and the underlying indexing structure for the database for each and every information resource. From the users' perspective it would be more effective and convenient to have one platform where they can stop and search all the available online databases - a single point of access.

#### **BAM – A Joint Portal for Libraries, Archives, Museums**

BAM (Fig. 1)[4] started as a project funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) in 2001. Since 2007 a consortium of library, archive and museum institutions hosts the BAM portal, among them the Bibliotheksservice-Zentrum Baden-Württemberg (BSZ), a library service centre that hosts the portal. At the moment BAM contains more than 40 million digital records contributed by several major German academic libraries, by sixteen museums and museum networks, and several major archives (cf. Table 1).

BAM total number of digital records	41 195 322
Libraries	37 175 528
Northern German Union Catalogue GBV (some 330 scholarly libraries)	~20 M
Southwestern German Union Catalogue SWB (some 1200 scholarly libraries)	~13 M
State Library of the Prussian Cultural Heritage Foundation, Berlin	~3 M
Central Index of Digitized Imprints (ZVDD)	~0,5 M
Archives	2 905 652
State Archives of Baden-Württemberg	1,7 M
State Archives of Hesse	0,8 M
Federal Archive of Germany	88 K
Municipal Archives (Freiburg, Heilbronn, Reutlingen, Mainz)	86 K

Museums	291 563
Architecture Museum of the TU Berlin (collection of technical plans and drawings)	69 K
Historical Museum of the City of Leipzig	141 K
The Prussian Cultural Heritage Foundation, Berlin	11 K
digiCULT Schleswig-Holstein	18 K
Foundation Haus der Geschichte, Bonn / Leipzig	6,5 K
German Historical Museum, Berlin	6,5 K
Other sources (Kalliope portal)	822 708

Table 1: The total number of digital records in BAM

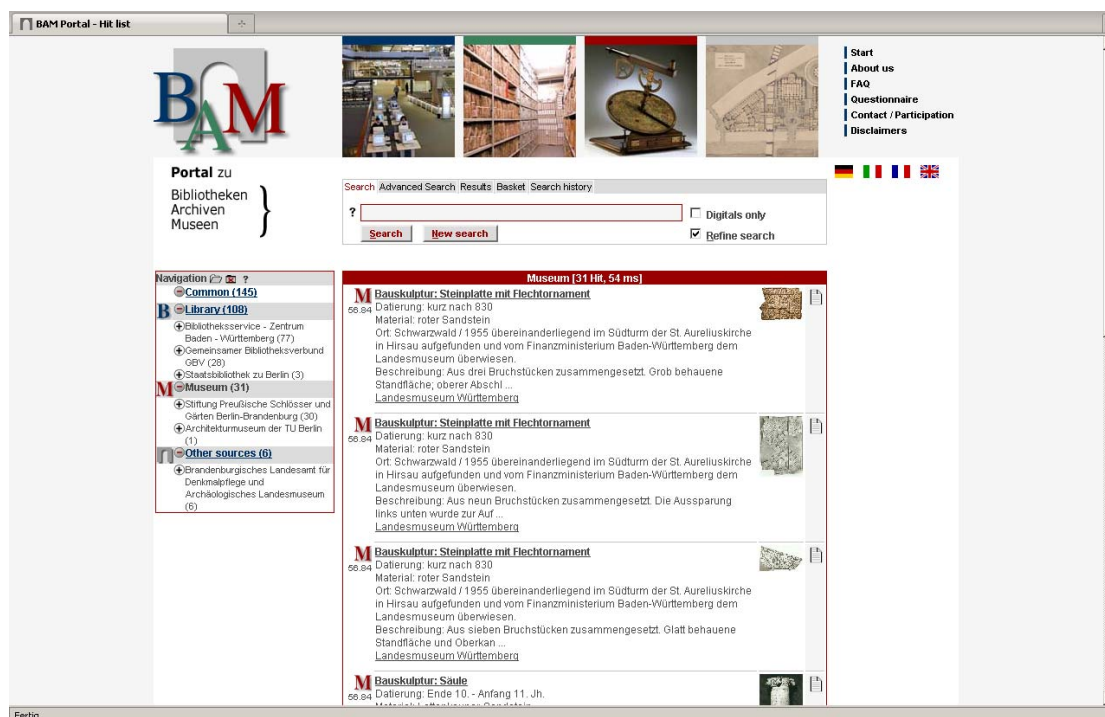


Figure 1: The BAM portal

The BAM portal offers the participating institutions a joint cross-institutional platform for digital catalogues, repertories, and inventories. Therefore, metadata of the participating institutions are collected, stored, indexed and made searchable on the BAM server, while the media content, i.e. the digital materials such as images and – in theory also text, audio and video, is stored in the online databases of the participating institutions who keep full control over and responsibility for their digital materials using BAM only as a gateway and as a means to increase their visibility on the Web by contributing to large digital collection that attracts user traffic. For smaller institutions without an online database of their own, a hosting service is offered by BAM. Such smaller institutions can store both the metadata and the media content of their digital collections in the BAM database which allows them to present their content on the Internet without having to maintain a complex web presence including an online database. As a bonus for sharing their content via BAM, these institutions can include a search form on their websites in order to present their own content on their own homepage. This option is important for institutions with limited resources.

To the present day, BAM is the only German cultural heritage portal on a national level as the German Digital Library (Deutsche Digitale Bibliothek, DDB) is still under construction and is not going online before the end of 2011. Therefore, BAM is currently a single point of access for all users who are searching items of cultural content on the German Web. As a consequence, the potential range of users is very broad, the major target audience being scholars, students, but also a general public of interested laypersons. As it is considered a central educational and scientific resource, access to the portal and the content of the participating institutions is free of charge.

## BAM Local - Uniting Different Branches of an Institution in one Portal

Apart from serving as a portal for different institutions, BAM is also applicable for an individual institution or a city or region who wants to make accessible its digital collections from different branches such as libraries, archives, museums, photo libraries and media centres at a single point of access. The so called "BAM local" presents a single institution's or city's or region's collections from different sources in a single portal and in this way creates a single point of access for potential users.

The advantage of a "BAM local" application is obvious: most institutions or cities or regions maintain different information services which can only be accessed from individual Web-based applications such as Online Public Access Catalogues in one or many libraries, from search engine interfaces of different Web-based database applications in museums, archives and media centres. With "BAM local", all these different content providers can unite their collections in one metadata database with a single index and interface. The "Google slot" of BAM can be integrated into almost any Web design by a simple HTML form and the user will be transferred to the BAM results page which can also be adapted to the institution's or city's or region's corporate design. In this way, "BAM local" is applicable for many purposes.

## Increasing Content Visibility by Collaborating with Wikipedia

In addition to serving as a central point of access, BAM tries to increase the visibility of the digital content of all participating institutions by collaborating with Wikipedia Germany. In August 2007 an alliance was formed that allows Wikipedia users to connect the encyclopaedia's web links section to a predefined query in BAM using a specific BAM Template (Fig. 2). Both information services can take advantage of this alliance: Wikipedia Germany offers its users a wide range of sources to investigate and BAM increases the visibility of its partners' digital content and draws traffic to their Web sites. Until December 2008 more than 900 BAM links have been created in Wikipedia and the process goes on, continually increasing the number of links.

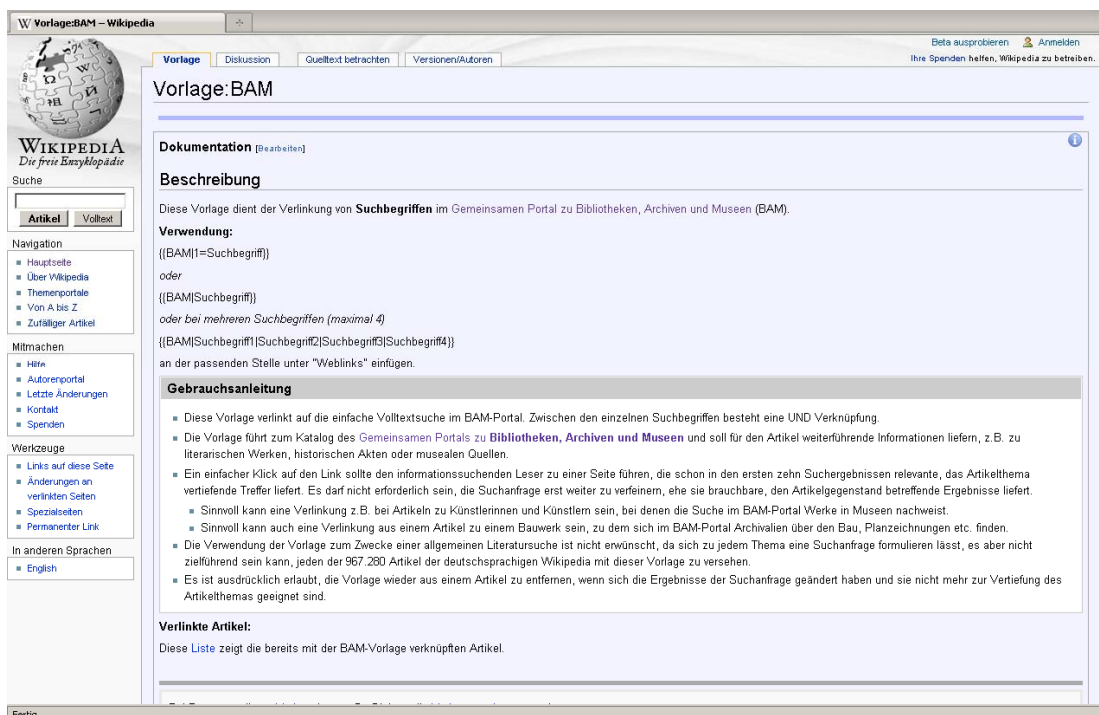


Figure 2: The BAM template in Wikipedia

## BAM and its Users

A detailed analysis of log files has not yet been carried due to lack of time and personnel. Hence the above mentioned target audience of the BAM portal has to be investigated further. The results of a preliminary examination of the BAM log files shows that there are more than 1 000 visits per day or around 30 000 visits



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

per month (from June 2008 to May 2009). These numbers are small compared with those of major search engines, yet it is a reasonable start and a point from which to continue to build a stable and large BAM community. Especially the link to Wikipedia has increased the traffic considerably as the current examination indicates.

## Conclusions

BAM – the joint portal for libraries, archives, museums in Germany intends to become a single point of access for cultural content on the German Web. In this way, BAM serves users who do not want to search several different databases at different servers using different search interfaces and vocabularies for access. To do so, BAM combines the different online information services from different institutions in one point of access. In addition, BAM can also serve as a portal for a single institution's libraries, archives, museums and media centres by combining their digital collections in one index under one search interface that can be integrated into the institutions corporate design. Apart from this, BAM also tries to increase the visibility of the individual digital objects in the collections of the participating institutions by cooperating with Wikipedia Germany. A Wikipedia template containing a predefined query to BAM can be added to any Wikipedia article and enrich it with a link to media content in BAM. Therefore, from our perspective, BAM is a successful tool to empower users who are looking for digital cultural heritage content on the German Web.

## References

- [1] Hedegaard, R. (2003) Benefits of Archives, Libraries and Museums Working Together. In Access Point Library: Media - Information – Culture. Proceedings of the World Library and Information Congress: 69th IFLA General Conference and Council in Berlin, Germany, August 1-9, 2003  
<<http://www.ifla.org/IV/ifla69/papers/051e-Hedegaard.pdf>>, accessed: 09/29/09.
- [2] Martin, R. S. (2003) Cooperation and Change. Archives, Libraries and Museums in the United States. In: Proceedings of the 69th IFLA General Conference and Council, August 1-9, 2003, Berlin. 1-10.  
<<http://archive.ifla.org/IV/ifla69/papers/066e-Martin.pdf>>, accessed: 09/29/09.
- [3] Kraemer, H. (2001) Museumsinformatik und digitale Sammlung. [engl.: Museum Informatics and Digital Collections.] WUV-Universitäts-Verlag, Wien.
- [4] <<http://www.bam-portal.de>>, accessed: 09/29/09.
- [5] Kirchhoff, T.; Schweibenz, W.; Sieglerschmidt, J. (in print) Archives, Libraries, Museums and the Spell of Ubiquitous Knowledge. In: Archival Science – Special Issue on Digital Convergence in Libraries, Archives and Museums. Springer.

**Lauri LEHT**

## **Involving users in the content enrichment process of digitized archival material**

### **Abstract**

The National Archives of Estonia has intensively digitized its most used archival units and developed internet solutions for presenting these materials free for everybody online since 2005. Among other solutions there are opportunities for users to add value to the digitized materials. Users can enter names of people described on pages of archival units to facilitate searchable access to mostly church books. Users can create custom online databases about the content of archival units and make accessible for everybody. Users can interact with each other in the forum and point to an exact spot on a page to ask help for reading hand-written text. Users can create their own link collections of digitized pages and spots on pages. Experience of the National Archives of Estonia says that users should be involved in time consuming processing of digitized materials for adding searchable data to them.

**Keywords:** digitization of archival materials, user involvement, content enrichment, online access

### **Results of digitization of archival records**

The National Archives of Estonia (NAE) has digitized around 50 000 archival units with 5 million images comprising around 5 Terabytes of data during 2005 to 2009. Most of these archival units are church books from the 18th to 20th century that are the most used materials by genealogists in the Estonian archives. By the end of 2009 almost all Estonian church books are available online.

The essence of genealogical research in Estonia has changed radically since 2005 when the first digitized materials were published online at the environment of digitized content of NAE called Saaga ([www.ra.ee/saaga/](http://www.ra.ee/saaga/)). Until then genealogy was a hobby for a few who could allow themselves spending time in the reading rooms of archival buildings mostly during working hours but also Saturday mornings and scrolling through the microfilms of church records.

When Saaga environment was opened genealogy became a popular way of spending free time for many people. All one needs for it is an average computer with an average broadband internet connection. Saaga is available for everybody 24 hours a day. Only free registration and afterwards logging in is needed for access. A drop-down of 25% of the quantity of physical users in the reading rooms was a logical result for these developments.

The other issue that has severely influenced the behavior of archival users and the character of usage of archival materials is the mass input of headings of archival units, series and archives into the archival information system (<http://ais.ra.ee/>). The input process was started in 1999 and the database was published online in 2004. By the end of 2009 all the headings (about 8 million) will be inputted and be available online – Estonia is then amongst the few countries where 100% of the archival descriptions are digital.

Since 2004 users make their searches in the archival information system instead of paper records. This has changed the variety of archival units used – new groups of archival materials are accessed that were practically not used before because of low knowledge amongst users. Users also spend no more time searching for the archival units in the reading rooms as they mainly do it online and order documents to the reading rooms before their visits.

### **Options for user involvement**

As the amount of digitally available archival descriptions and digitized images has risen steadily and there are much more archival users than before, the amount and potential of the user community has grown massively. Genealogists have nonprofit amateur unions where they share knowledge and ideas and discuss different problems in forums (e.g. [www.isik.ee/foorum/](http://www.isik.ee/foorum/)). The number of genealogists and other online users outnumbers the amount of archivists and IT developers in the archives a lot. Most of the users are waiting eagerly for every new piece of digitized documents and each new digital description. Several of them are true fans of archival and genealogical studies and have expressed their wish to help NAE in the process of making archival documents digital.

As the archives can not let volunteers do the basic digitization work which involves physical scanning procedures, there are opportunities to involve users in the areas where the archives will probably never have enough resources to do the work. These areas have been described in NAE as the following:

- quality checking of digital archival descriptions for printing errors and logical faults,

- helping other inexperienced users in understanding the content of archival documents,
- describing the content of archival documents in a structured way,
- collecting similar data from different archival documents and making these thematic databases available for public.

### Realization of user involvement in online tools

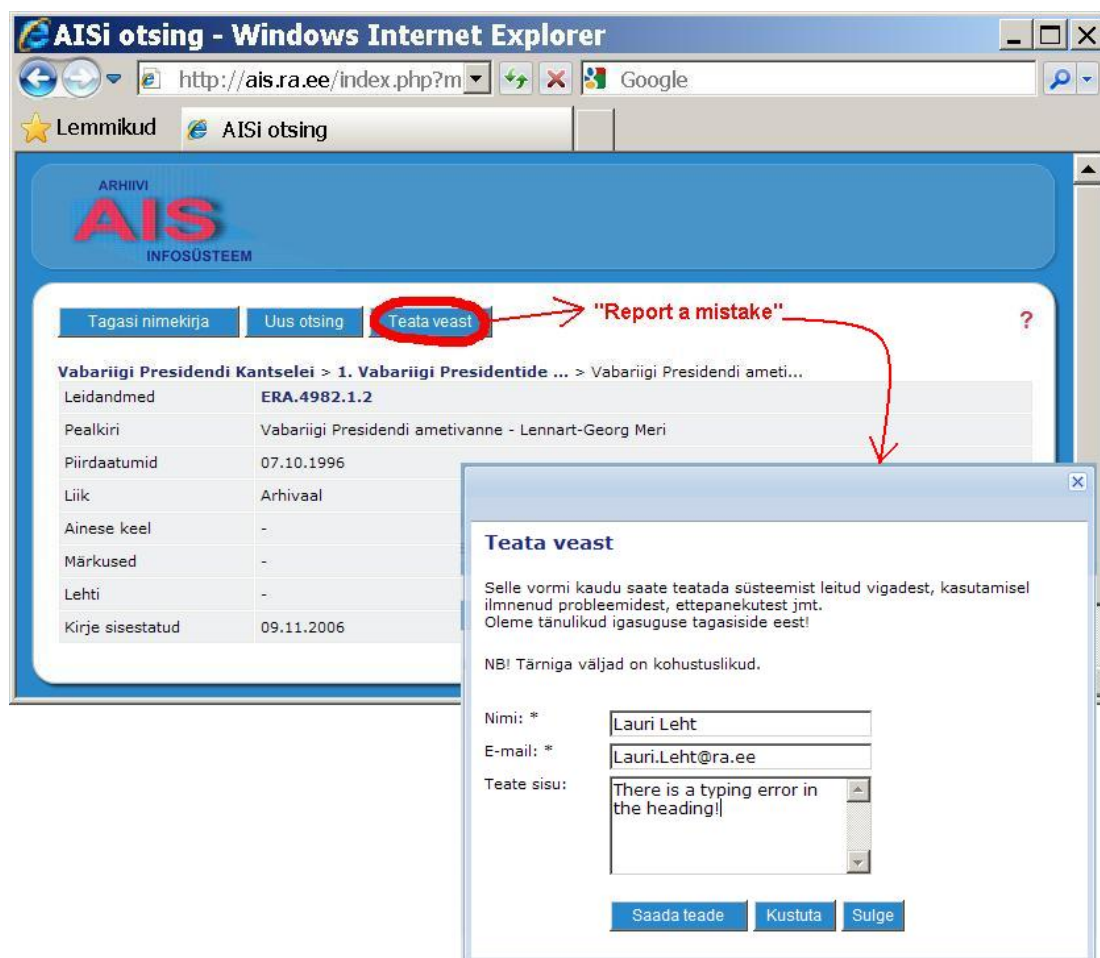
In connection with user groups several tools have been made by the archives' IT developers to satisfy the users' wish to contribute to the digital content and the archives' need to have more searchable data about digitized documents of good quality.

### Quality checking of digital archival descriptions

In the web portal of the Estonian archival information system (<http://ais.ra.ee/>) there are headings and other descriptions of about 8 million archival documents, series and archives.

As this data has been input manually during 10 years, several data has been transferred from legacy databases and the descriptions are in Estonian, German and Russian which use different alphabets then it is a known fact that there are quite many typing errors, data transfer errors, logical descriptive errors etc in the system.

NAE has launched a simple solution to allow users to give feedback about descriptive data that is not correct. The feedback button "Report a mistake" includes the number of the archival unit and other technical data with the user's message about the mistake and it is sent to the database administrator of the system for correction. Every day several mistakes are reported.



The screenshot shows a web browser window titled "AISi otsing - Windows Internet Explorer" with the URL <http://ais.ra.ee/index.php?m>. The page header includes "ARHIVI AIS INFOSÜSTEEM". Below the header, there are navigation buttons: "Tagasi nimekirja", "Uus otsing", and "Teata veast". A red circle highlights the "Teata veast" button, with a red arrow pointing to the text "Report a mistake".

The main content area displays a record for "Vabariigi Presidendi Kantselei > 1. Vabariigi Presidentide ... > Vabariigi Presidendi ameti...". The record details are as follows:

Leidandmed	ERA.4982.1.2
Pealkiri	Vabariigi Presidendi ametivanne - Lennart-Georg Meri
Piirdatumid	07.10.1996
Liik	Arhivaal
Ainene keel	-
Märkused	-
Lehti	-
Kirje sisestatud	09.11.2006

A "Teata veast" dialog box is open in the foreground. It contains the following text:

**Teata veast**

Selle vormi kaudu saate teatada süsteemist leitud vigadest, kasutamisel ilmnenud probleemidest, ettepanekutest jmt. Oleme tänulikud igasuguse tagasiside eest!

NB! Tärniga väljad on kohustuslikud.

Nimi: \*

E-mail: \*

Teate sisu:

Buttons: Saada teade, Kustuta, Sulge

Figure 1. Reporting form for mistakes in the archival descriptions.



## Helping other users understand content of documents

Most of the digitized documents are church books that are up to 300 years old, lots of them are hand-written in German Gothic writing which is quite hard to understand for inexperienced users.

Fortunately there are also several experienced users in the community who have been dealing with the church records for a long time and can give answers for most of the puzzles. As it was possible to free the archivists from the need of giving this kind of explanations online, NAE implemented in its Saaga environment ([www.ra.ee/saaga/](http://www.ra.ee/saaga/)) a solution where users can select one or several areas on a digitized image and post them to a forum where volunteers from the genealogical society are eager to help each other in understanding the meaning of badly-written phrases.

Users can also combine the area selection function and the personal link collection function in Saaga and save their necessary data from the digitized images pointing exactly to the relevant parts on the image.

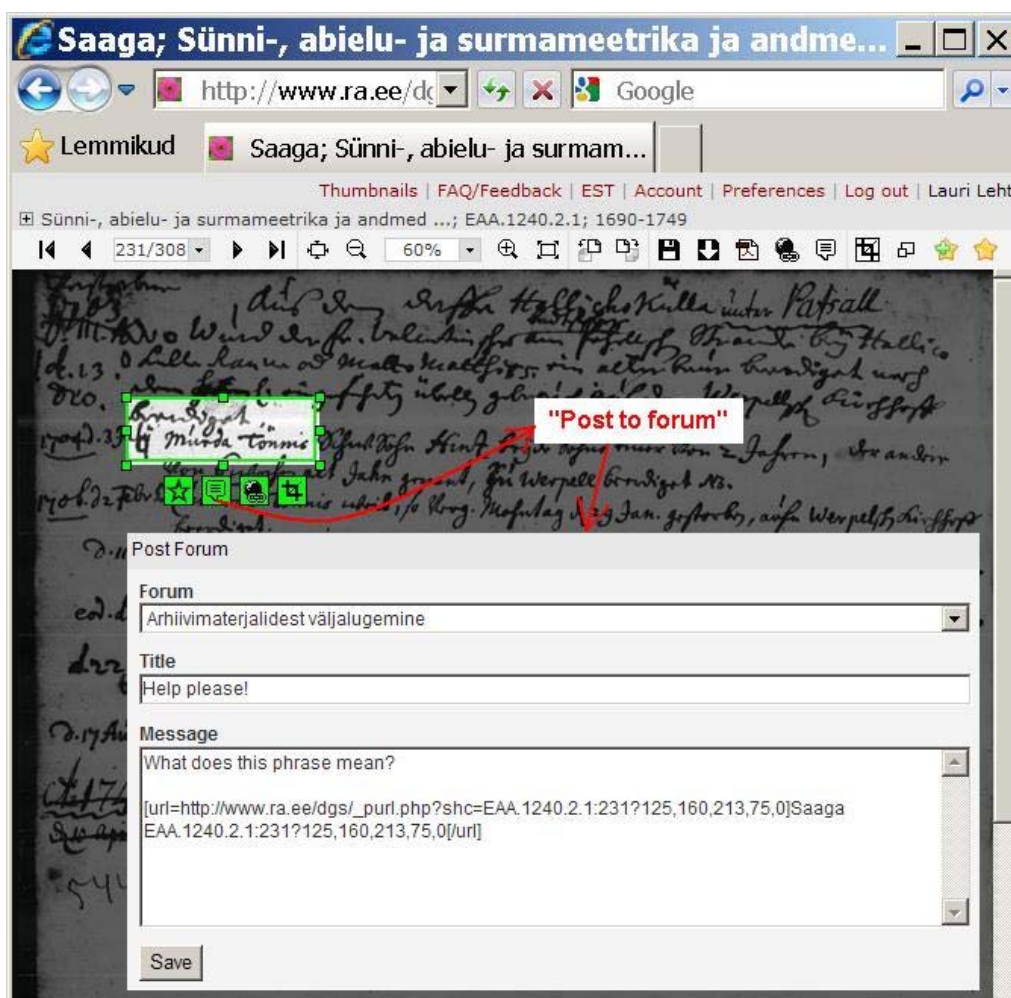


Figure 2. An example from Saaga of cropping and pasting areas of an image to forum.

## Describing content of archival units in a structured way

The images of digitized archival units are raster images where no layers or text is optically recognized. This will probably be so for some more time as the OCR techniques for old hand-written texts are not yet practically available.

Therefore NAE has created and given to the volunteer users a tool for indexing data of names of people from the church books as names are the most used search words and also the real essence of church records ([www.ra.ee/dgs/addon/nimreg/](http://www.ra.ee/dgs/addon/nimreg/)). The genealogical society is doing the work of inputting names with page numbers from the digitized images. All the users of Saaga can help in connecting the church books' page numbers with the actual digital frame numbers as these figures always slightly differ.





Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

VAU - DatabasesPage - Windows Internet Explorer  
 http://www.ra.ee/vau/index.php?page=D  
 Lemmikud VAU - DatabasesPage

VAU virtuaalne uurimissaal  
 RAHVUSARHIIV Lauri Leht Settings Links Databases Logout

Archive view Social view Personal view

Databases Forum

Social view > Databases > All databases > Database > Search result

### Jämaja abielud 1908-1926

<< < 1 2 3 4 5 6 > >> 1-50/271

Laulatuse kuupäev	Peigmehe eesnimi	Peigmehe perekonnanimi	Pruudi eesnimi	Pruudi perekonnanimi	Lehekülg EAA.3131.1.48	Saaga kaader EAA.3131.1.48
13.01.1908	Tiidrik	Vakkum	Miina	Sepp	27	18
03.02.1908	Laas	Pakkurt	Mari	Tamm	27	18
03.02.1908	August Wilhelm	Ader	Triin	Kaasik	28	18
03.02.1908	Tiidrik	Tõnn	Ewa	Remm	27	18
03.02.1908	Mihkel	Õigemeel	Triin	Poobus	27	18
10.02.1908	Johan	Toomus	Miina	Krull	28	18
10.02.1908	Jüri	Ley	Ann	Rand	28	18
10.02.1908	Mart	Suurhans	Triin	Wänt	28	18
10.02.1908	Tiidrik	Timmermann	Ann	Remm	28	18
24.02.1908	Predik	Rand	Mari	Wõrk	28	18
24.02.1908	Hindrik	Poobus	Ewa	Ankur	28	18
24.02.1908	Mart	Sooär	Triin	Matrus	28	18
02.03.1908	Willem	Wekmann	Mari	Helde	28	18
27.04.1908	Alfred	Lehmann	Anna Sophie Meta Helene	Nielsen	28	18
02.06.1908	Jüri	Toomus	Ann	Põlda	28	18

Figure 4. Example of a database made available by one user to the general public.

## Conclusions

Experience of the National Archives of Estonia says that users should be involved in time consuming processing of digitized materials for adding searchable data to digitized images. If the public archives give convenient tools to users for that, volunteers from the user community are eager to start producing and publishing data that adds on to the digitized images. The role of the archives in the near future should be digitizing their documents according to popularity and listening to the users' needs for providing good tools for archival fans for creating added value.

## **Franco LIBERATI, Maria Teresa TANASI**

### **Optical supports for digital preservation - problems and prospects**

**Abstract.** Today, archives and libraries are involved not only in digital projects, but also in definition of new policies in order to guarantee a long-term preservation of digital objects.

Digital preservation is considered a process that requires use of the best available technology and related procedures. In particular, information must be intact and readable from storage media; contents have to be accessible and interpretable; standard formats and migration plans must be developed. Digital data are stored in magnetic and optical supports which have different characteristics and life expectancy. National and international scientific committees promote standards and technical strategies to extend the useful life of digital media and protect them from degradation and technological obsolescence.

In this paper structure, technology, and degradation processes of common optical discs (CD, DVD, and Blu-Ray Disc) for digital preservation are described, with particular attention to Holographic Versatile Disc (HVD), an innovative technology which offers a storage density more capability than the other optical media. Furthermore, internal and external factors that can attempt the integrity of supports and data such as instability of components, environmental factors, and uncorrected handling are discussed. Finally, standard storage conditions and care for long-term preservation are reported.

**Keywords:** optical storage, holographic data system, recording materials, digital preservation device.

#### **Introduction**

In the last years organizations involved in preservation of digital information need to high reliability systems. Data can be stored on each medium that can represent their binary values (bitstream), such as magnetic or optical media. It is important to have knowledge of the different media, of particular software and hardware equipments for access and storage, and of conditions requirements for preservation.

Optical discs, due to their easy of use, large capacity and low costs, are considered supports for storing digital information. In 1982 the Compact Disc becomes the most common media for recording data. Ten years later, Digital Versatile Discs, increasing the capacity, had preferred. In 2007, Blue Ray Disc provided 25 GB.

Anyway, the lifetime of these supports, in other words the period of time in which the information is stored in safety, is matter of studies in all over the world. Data are vulnerable to loss and corruption; in fact, optical media are sensitive to heat, humidity, pollutants or can fail because of faulty reading/writing devices.

After some years, optical discs change their capacity, features, logical and physical format, and, consequently, hardware/software systems. In order to contrast digital obsolescence, contents must be copied periodically on new media and formats (refreshing and migration) [1].

Today, holographic supports are a new technology that promises to revolutionize the storage systems (500GB). In the past, the realization of holographic system has been discouraged by the lack of availability of suitable components, the complexity of holographic multiplexing strategies, and the absence of recording materials with satisfying optical storage requirements. Recently, new studies and researches have rekindled the interest to this technology.

#### **Structure and technology**

Optical discs, that use laser technology for storing and retrieval data, can be classified as follow: Compact Disc (CD), Digital Versatile Disc (DVD), Blu-Ray Disc (BR), and Holographic Versatile disc (HVD) [2][3][4].

CDs, DVDs and BRs consist of same basic materials and layers, but they are differently manufactured. There are many kinds of optical discs; the attention will be focused only on –ROM (read only memory) and –R (recordable) [5]. –ROM and –R discs have a multi-layer structure (Fig.1).

The substrate is a polycarbonate which provides the transparency useful for laser to reach the reflective and data layers. It also offers the necessary depth to maintain laser focus, and, at the same time, enough strength to remain flat.

The data-layer contains digital information: in –ROM discs, data are “pressed” in the reflective/substrate layer (molding process); in –R discs, data are written by a high-power laser which changes chemical structure (burning process) of an organic dye (cyanine, phthalocyanine, azo based). DVDs and BRs can have one or two data-layer.

The reflective-layer is a metal which reflects the laser beam to the photosensor. Three types of reflective metals are normally used: aluminium, silver, and gold. The photosensor transforms optical into electronic signals and, by means to an analogue-to-digital converter (ADC), digital information is reconstructed.

A very thin lacquer is applied to protect the disc from exposure to the environment (protective-layer). An optional label is useful as top layer for graphics design and logos.

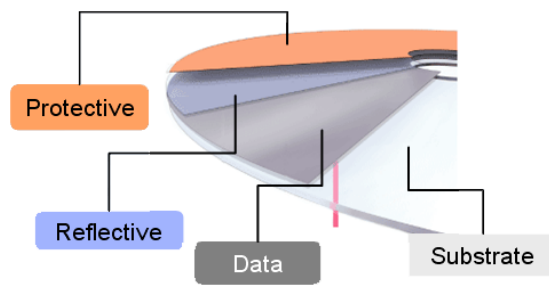


Fig.1 - Cross section of an optical disc

In all kinds of optical discs, data are marks (pits) impressed on the flat surface; the area between two pits is called land. An optical disc contains a track of pits arranged in a continuous spiral running from the inner circumference to outer (~5 Km). The drive reads marks on the track using a laser which measures the amount of light that gets bounced back from it. Areas with pits reflect the light less strongly than land areas. When photosensor detects a switch pit/land or land/pit, the system reconstructs the digital pulses. The pits on the data-layer are the physical manifestation of a complicated encoding process including multiplexer, interleaving, parity, error correction, modulation (EFM) [6][7].

CDs offer storage capacity 0.7 GB about, DVDs 4.7 GB for each data layer, and BRs, which use a blue-laser and achieve a spot size of a few hundred micrometers, provide 25 GB for each data layer (Fig.2).

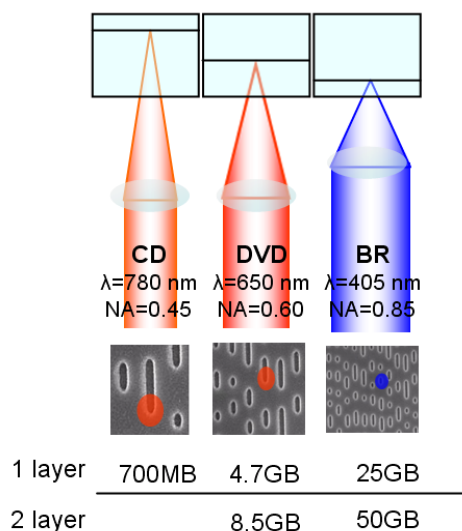


Fig.2 - Capacity, numerical aperture (NA), wavelength ( $\lambda$ ) in optical systems

Holographic Versatile Disc (HVD), that use an innovative technology, is composed by a recording-layer between two substrates (polycarbonate) with in the middle a dichroic mirror that reflects the blue-green light and allows the red light to pass through in order to gather servo information. The servo monitors the position of the read head over the disc (Fig.3).

Recording-layer materials are divided in two classes: inorganic photorefractive crystals and photosensitive organic polymers [8][9][10]. Two optical techniques for recording data in holographic systems are used: two-axis (angle multiplexing) and collinear (shift multiplexing).

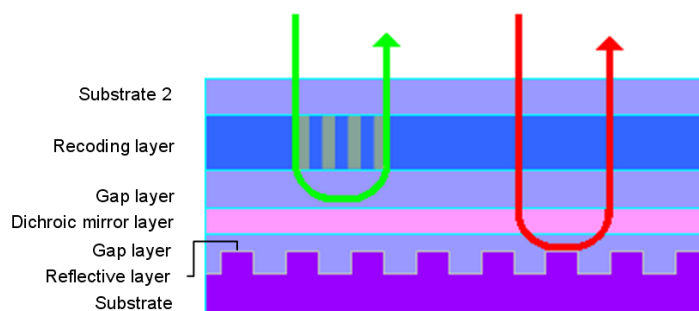


Fig.3 - Cross section of an HVD

During the HVD writing process, in the two-axis technique, binary data are disposed in a bi-dimensional organization (page).

A spatial light modulator (SLM), or page composer, translates page into an optical pattern, called image, where ones and zeroes are represented as opaque (black) or translucent (white) areas; each area is also called pixel. Based on liquid crystals, the SLM offers a valid contrast and rapid switching between black/white states. At the moment, the page composer is structured as a 1024x1024 pixels matrix (pixel size ~15-20 micrometers).

Once the image is created, a single laser beam is split into two: information beam, which is directed toward the SLM, and reference beam, which is directed, using lens and light deflectors, into recording-layer. When the information beam passes through the page composer, portions of the light are blocked by the opaque areas of the image, and portions pass through the translucent areas. In this way, the information beam carries the image and when the reference beam rejoins on the same axis, a pattern of light interference, the hologram, is recorded in a light sensitive medium. By varying the reference beam angle, the wavelength, or the media position, many different holograms can be recorded in the same volume of material. This process of superimposed holograms, called multiplexing, yields the enormous storage capacity.

In the HVD reading process, the reference beam is incident on the medium under the same conditions used for recording and it produces a diffracted beam representing the image. The optical information is revealed by a detector array (CMOS or CCD) which allows extraction of the page from the measured intensity pattern. Then, the signal enters into the threshold, error correction and demodulation circuits; finally, the calculator can process the bitstream.

In collinear technique, reference and information beams come from the same SLM.

HVD capacity is 300 GB (Fig.4) about and the collinear strategy is used in order to guarantee storage information with simple and minimal devices.



Fig.4 – HVD with (right) and without cartridge (left)

Technical details of optical discs are reported in table 1.

Optical discs	CD	DVD	BR	HVD
Maximum Capacity (GB)	0.7	8.54	50	1000
Data rate (Mb/sec)	0.15	1.35	36	1000
Wavelength	780nm	650nm	405nm	500nm
Numeric Aperture	0.45	0.65	0.85	0.65

Table1 - Technical details optical discs

## Degradation

Each layer of optical discs can degrade. The polycarbonate is a very stable polymer and it degrades slowly respect to the other layers.

The data-layer, in –ROM discs, is coincident with the reflective-layer; in –R discs, high temperatures and high humidity accelerate the deterioration process of organic dye. Also prolonged exposure to natural or artificial light can increase the degradation of data-layer, altering the chemical and optical properties of dye. Phthalocyanine seems to be the most stable.

The degradation of the reflective-layer depends on material: aluminium is subject to oxidation in contact with oxygen, pollutants and high humidity more than the other metals; silver reacts with sulphur dioxide; gold is very stable. The effects of oxidation are loss of reflectivity and, then, loss of readability.

The protective-layer has an unknown permanence due to not declared chemical formulation.

One of the aspects of physical deterioration is the different dimensional changes of layers in consequence of thermo-hygrometric conditions fluctuations. The outer layers are more vulnerable than the inner because they are subject to mechanical damages.

Scratches can attempt to integrity of discs, obstructing the correct read/write operations and, consequently, corrupting the data. Figure 5 shows two images of a CD with opportunely caused scratches. The effects on substrate and reflective layers are analyzed by equipment for quality tests of optical discs.

As regards HVDs there is not exhaustive information about their components and systems failure.

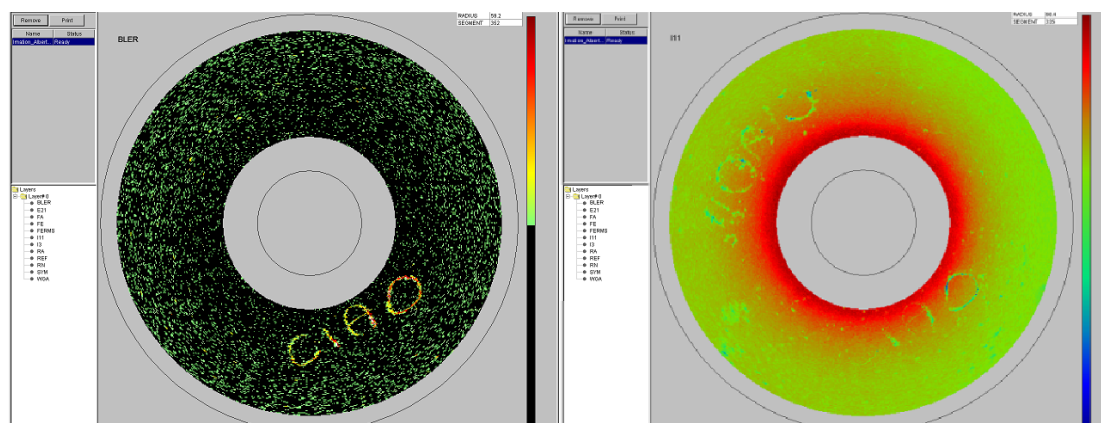


Fig.5 - Substrate (left) and reflective-layer (right) of a CD

## Storage condition, care and handling

ROM discs can be reliable for many decades if stored at suitable conditions, while CD–R, DVD±R and BR–R are at risk after just few years. Furthermore, degradation is expected over time, but some strategies can be taken in order to slow down it.

Generally, useful life of optical discs can be increased by storing at low temperature and low relative humidity, without fluctuations, minimizing the pollutants contents, avoiding the exposure to artificial and natural light, choosing proper shelves and boxes [11].

Storage temperature and relative humidity ranges recommended by ISO 18925 [12] are in Table 2.

HVDs present a protective cartridge in order to minimize effect of fingerprints and dust, but no information in order to guarantee a long-term preservation are reported in the scientific literature.

	Temperature	Relative Humidity
CD	<23	20%-50%
DVD	<23	20%-50%
HVD	?	?

Tab.2 - ISO 18925, Imaging materials - Optical disc media - Storage practices

## Conclusion

Some standards describe methods for estimation of optical discs life-expectancy [13][14][15]; generally, it is possible to extend their life applying appropriate storage conditions, proper care and handling. In addition, frequent refreshing and migration are necessary in order to preserve the recorded information and to cope with the technological obsolescence.

About the new optical supports, HDVs, it is possible affirm that they should become a candidate of next generation storage media but, at the moment, there is not exhaustive scientific literature about their systems, logical and physical structure, degradation, and their use in preservation field.

## Reference

- [1] K. Lee, O. Slattery, R. Lu, X. Tang, and V. McCrary, The State of the Art and Practice Digital Preservation, Journal of Research of the National Institute of Standards and Technology, Volume 107, Number 1, January–February 2002.
- [2] ECMA-130, Data interchange on read-only 120 mm optical data disks (CD-ROM), 2nd Edition, Switzerland, June 1996.
- [3] DVD Consortium, DVD-R for General, Part 1:Physical Specifications Ver .2.1, 1998.
- [4] Blu-ray Disc Founders, White paper, Blu-ray Disc Format, General, August 2004.
- [5] Optical Storage Technology Association, Understanding CD-R & CD-RW Technology, January, California, USA, 2003.
- [6] ISO 9660, Information processing - Volume and file structure of CD-ROM for information interchange, 1988.
- [7] G. Sharpless, An Introduction to DVD Formats, Deluxe Global Media Services Ltd, 2003.
- [8] K. Buse, Holographic Recording Medium, Optical processing and computer, SPIE-International Society for Optical Engineering, 1998.
- [9] A. B. Samui, Holographic Recording Medium, Recent Patents on Materials Science, Vol. 1, No. 1, 2008.
- [10] Bayer AG, Long Life for data, Bayer research Magazine n°18, 2007.
- [11] F. Liberati, G. Marinucci, M.T. Tanasi, Digital Preservation of Magnetic and Optical Support - Problems and Prospects, MATCONS, 2009.
- [12] ISO 18925, Imaging materials - Optical disc media - Storage practices, 2002.
- [13] ECMA-379, Test Method for the Estimation of the Archival Lifetime of Optical Media, 2007.
- [14] ISO 18921, Imaging Materials – Compact Disc (CD-ROM) – Method for Estimating the Life Expectancy Based on the Effect of Temperature and Relative Humidity, 2002.
- [15] ISO 18927, Imaging Materials – Recordable Compact Disc System – Method for Estimating the Life Expectancy Based on the Effect of Temperature and Relative Humidity, 2002.



## **Luciana GUNETTI, Eleonora LUPO, and Francesca PIREDDA**

### **Designing digital formats for cultural production and exploitation: from accessibility to use value**

Contents, languages and technologies for participative (on line) knowledge repertoires

#### **Abstract**

This paper aims to bring a theoretical, methodological and phenomenological contribute concerning the interpretation by design culture on the ways of producing shared knowledge and culture and enhancing Cultural Heritage in the digital environment. In fact, design strategies, skills and techniques can be applied bringing a design driven innovation in Cultural Heritage digital exploitation. In particular, the experimentation of contents, languages and technologies, represents the most original approach in articulating a systemic model of shared knowledge accessibility and production available to the user.

According with this approach, the modalities of sharing and using culture and knowledge repertoires in the digital space respond to the complexity that characterizes both the cultural system and the communication one. Design culture proposes exploitation models and tools able to translate this complexity in contents and languages and to turn the available technology in virtuous devices enabling representation and access.

The systemic design approach basically introduces the direction of producing, using, managing, experiencing cultural repertoires on line, working both on the archetype of the catalog and new collaborative formats, giving shape to a wider concept of accessibility: from making "available" the Cultural Heritage to providing the opportunity for diverse community of users to use it in practice.

New formats are therefore designed, formats capable of responding to the emerging communication practices: solutions such as visual timeline are integrated with multimedia, video and with collaborative and customised tools, making them accessible and usable by the various communities of users (stakeholders such as professionals and researchers, but also the large public).

We analyzed these critical nodes through a phenomenological mapping of virtuous experiences and examples, able to identify the potentialities of Web 2.0 as a platform for an integrated communication system, which is able to re-orienting Cultural Heritage valorization towards social practices and converastions. Designing new paradigms of shared knowledge and culture production and use, we (as researchers and institutions) can move from the simple concept of accessibility to that of "use value".

**Keywords:** Design driven innovation, Use value, Storytelling, Performance, Place

#### **"New" Cultural Heritage on line and dynamics of "use value" design driven**

What is Cultural Heritage made of? It's commonly agreed, that, even if Cultural Heritage appears to be fixed and immutable, the concept has evolved by time. How human cultural artifacts become Cultural Heritage is a dynamic process, because value is not a technical quality embedded in forms and processes, but in the way they are integrated in the social lifestyles and patterns. The Cultural Heritage is the result of social relations, and increases its sense the more it is recognized and incorporated in the collective conscience of a community, in other words, "practiced" (1) in its "use value" (2).

Cultural processes require complex times of negotiation and settlement longer than the ones experienceable by a community, and it's necessary to split them in phases in order to make them synchronic and acceptable by people. For example, according to Dorfles (3), «art is a changeable reality whose meaning differs depending from time, and can be identified with myth, religion, society, technology»: this leads to different interpretations and fruition modalities, corresponding to the user, the context and the time. The processes of "genetic coding" of Cultural Heritage is not neutral: something appears unquestionably worth of cultural value, only under the beliefs and the socio-cultural constructions of an age. Consequently it's necessary a precise "investment" (for instance an enhancement project) to deliberately underline a particular content as valuable to the community: it is a specific will of social construction of a community Cultural Heritage. This is undoubtedly a selective and elective process of social production and reproduction of values and meanings, that ends with a distinctive collective attribution (1).

In the contemporary society, the digital environment appears to be the most receptive context in enabling and incorporating the expression and legitimating of new heritage forms. In fact, it includes a wide repertory in consistence and typology of "new" Cultural Heritage forms and processes: from archives of digitalized tangible artifacts, to digital libraries of cultural expressions, catalogues of new forms of cultural production, and repositories of local knowledge, they all document the co-existence of the different formats that cultural identities can assume in the web. Databases and digital repositories has been explored in the last 20 years as

the most recognised model of knowledge and cultural contents archive, but in the network age, the active role of the user, together with the obsolescence of data and the interoperability of formats, require to rethink this conceptual model in a more participative “locus” for the building and exchange of collective and visual identity of a territory and its community. Looking at the more recent examples, it is always more and more evident that the digital environment has changed into a “place” that facilitates the social and collective construction processes of the value of new heritage forms, and not only a “space” to store them. In this perspective, participative processes become cultural expressions too, and sometimes Cultural Heritage themselves.

In our opinion, this transformation has not been spontaneous: it has been consciously led and shaped by communication design, according to the emerging of a communication paradigm shift from availability to accessibility, from usability to participation, and as a response to the complexity of the contemporary cultural production system too. So, in the last 10 years the design of digital formats for Cultural Heritage enhancement has been addressed to experiment languages, technologies, collaborative and sharing tools, to enable those cultural negotiation and legitimating processes, apart from building a collective and shared memory: in other words, to play and act the heritage beside than document it. In this sense, communication design supports the dynamics of transformation of the Cultural Heritage value from “value per se” to “use value”.

In particular, in the digital environment, the “use value” of Cultural Heritage relies on the capacity of design to enhance and make accessible the Cultural Heritage as a system and as a process for new different uses and users. The digital models and tools more design driven are the ones that apply the strategic and communicative potentiality of design to enhance and visualize the Cultural Heritage in a “re-usable” way, connecting its physical aspects with the digital ones. The first use value enabled by the design approach is generated by the exploitation on line of the heritage systemic nature, underlining the context and place where it has been generated from (from the physical localization to the natural, territorial, environmental, cultural and immaterial conditions which determined the “form” of the heritage and oriented its development), context that impacts and suggests new opportunity of fruition and further dissemination. The second use value enabled by design is arisen by the explicitation on line of the heritage process nature, enriching in its tangible elements with intangible aspects, like abilities, skills, narrations, performances and procedures, useful for its innovative production and re-production.

The Cultural Heritage contents, in this systemic approach, are managed and processed by communication design as “products” enjoyable by the final user (experts or generic) and usable for the production of new cultural contents, or educational purposes.

From the following case study analysis, it will be evident that the new digital formats are designed as devices that empower the user by suggesting and enabling opportunity of practice, re-use and re-contextualization of Cultural Heritage, structuring in the web participative repertoires and tools for the bottom–up production and experience of knowledge and culture.

### **Communication design for digital Cultural Heritage contents direction**

Turning every form of Cultural Heritage into an open resource by setting up digital integrated management systems is no longer the sole aim of communication design.

Its contribution, in terms of analytical and design tools needed to define one or more design models enabling the transformation of any cultural production into an open resource, is no longer based on cataloguing, on thesauri, on indexing, but focuses on the dialogue between digital atlases – complex and fluid communication systems based on information display – and the development of dense, localised systems, i.e. analogical atlases, based on narration, performance and places. If the places traditionally connected with knowledge in the field of the Cultural Heritage – let us call them nodes – are analogical, one may not think of managing large amounts of data and materials, of having all of the Cultural Heritage in one network. Hence the process started by Communication Design should go the other way round, i.e. it should not be based on the quantity, but on the quality of the cultural mediation that is made possible by new technologies (databases, thesauri) and, even more importantly, new languages (information design, video, etc.). The relationship between the analogical and the digital, characterised by a mutual exchange, makes the communities of Cultural Heritage users dynamic, by connecting them to each other through Narrations, Performances and Places – seen as the new ‘hotspots’ of the analogical/digital dialogue.

Today the enhancement of this heritage, both material and immaterial, in the digital and analogical environments is not carried out only by means of digital archives, online collections or online museums, but also through processes like narrations, performances and places, which enable the individual and social production and re-production of Cultural Heritage knowledge.

The Cultural Heritage system can use Design to try and put in place this process, starting from the metaphor of Aby Warburg’s meta-discursive and discursive atlas (early 20th century) in which he takes the ancient knowledge and images collected in his Library of the history of culture and displays them in a place like the

Mnemosyne atlas, up to the Actor-Network Theory (4) which takes shape in the cartography of controversy. We may consider the two methods as fundamental research practices for the development of new Cultural Heritage systems, both of which are based on the generation of atlases into which converge various representations by the overlapping of several different levels. Be it an analogical (Warburg) or a digital (Latour's controversy sites) atlas, it is necessary to construct the toponymy of the maps of contents of the atlas – in our case concerning the Cultural Heritage – as an ethnologist would do, describing every object and every social fact as a Network. The transition from the Mnemosyne Atlas to the Actor-Network Theory makes it possible to explore the ways in which past, present and future communities develop and maintain the connections between individuals and groups by means of narrations/mediations, performances/processes and places/systems managed by old and new languages (theatre, art, cinema, video, web). Studying all the actors (human being, technological artefact, institutional body, legal norm, etc.) who collaborate, more or less directly, to the creation of a (material or immaterial) piece of the Cultural Heritage is not enough, because everything depends on the type of action linking the various actors. On the one hand is the "existing" or "ready-to-use" Cultural Heritage; on the other the Cultural Heritage "under construction", from a state of "fact" or "artefact" to the "worknet", depending on the action of a vast network of actors. The work, movement, flow (action) that is generated is always an actor and, because actors and networks are two faces of the same reality, the search of new languages for the Cultural Heritage can no longer go into the direction of the digital alone, but it must also, above all, re-orient itself towards the analogical.

If we focus on the tree design-oriented actions/nodes of this digital-analogical dialogue a few questions are raised:

- Analogical/digital narrations. How can a community of users for a certain category of Cultural Heritage tell stories and define its own identity and link it to other identities? Investigation methods include storytelling, memory and the life story of the objects. These are cases in which strategies are applied to tell these stories outside archives and museums using digital technologies and new languages, e.g. the oral history of DHS (Design History Society) recording reminiscences, memories and experiences of the art community within the community of design history. The role of narration is meant as a strategy for emotional sustainability in a contemporary community of users and can potentially give birth to collaboration projects.
- Analogical/digital performances. How do the identities of the communities of users for the Cultural Heritage represent and manifest themselves? The performances of individual and collective multiple identities are considered, developing the Actor-Network Theory and the performativity of the Cultural Heritage, to build new spaces of expression, be them analogical or digital.
- Analogical/digital places. How do the communities of users act within and without space boundaries to preserve and create places for interaction and sharing? What roles do the communities play in the creation of places? The study of communities connected by way of digital networks, spaces and places will make it possible to identify the possibilities of innovation, developing competences and bringing about a greater social inclusion. For example, will it be sufficient to rest on cultural districts, producers of social and cultural inclusion in a territory and foundation for the building of a collective (analogical and/or digital existence of the Cultural Heritage)?

Being interconnected, all the three actions bring about the construction of atlases which need new tools and new representation formats, new devices. The change takes place in language: for example, thesauri as languages need to present contents in the form of narrations and performances through languages which are alternative to digital interfaces, e.g. oral story-telling, theatre and artistic performances. Conversely, places (cultural districts) need to present their contents – material, immaterial and environmental - by integrating them into cultural services targeted to users and to the development of relevant production chains using the new digital languages (video and collaborative tools).

### Case study framework and interpretation

Below are therefore proposed case studies of each of the three actions (storytelling, performance and place) capable of triggering a virtuous relationship between analog and digital nature of Cultural Heritage, respectively: Oral History project by the Design History Society ([www.designhistorysociety.org/projects/oral\\_history/index.html](http://www.designhistorysociety.org/projects/oral_history/index.html); [www.vivavoices.org](http://www.vivavoices.org)) and Telling Lives by BBC ([www.bbc.co.uk/tellinglives](http://www.bbc.co.uk/tellinglives)) (5); on one hand the Digital Library by Sardinia District ([www.sardegnaDIGITALlibrary.it](http://www.sardegnaDIGITALlibrary.it)), because of the ability to collect different forms of expression and modes of representation of the identity of the Sardinian culture; on the other hand Cultura Italia ([www.culturaitalia.it](http://www.culturaitalia.it)) and the database of the Powerhouse Museum in Sydney, Australia ([www.powerhousemuseum.com](http://www.powerhousemuseum.com)), for the performative research by the user and his interaction with online resources; Monti TV, a Web TV in Roma ([www.montitv.it](http://www.montitv.it)) and the project Changing Linz and Wikimap Linz by the AEC - Ars Electronica Centre

([www.aec.at](http://www.aec.at); <http://wikimap.hotspotlinz.at/de/index.php>) (6); in the end, transverse to the narratives and places, we report the project Storymapping by the Center for Digital Storytelling ([www.storymapping.org](http://www.storymapping.org); [www.storycenter.org](http://www.storycenter.org)). These cases were selected because they are representative of what is currently available online for access to Cultural Heritage and because of their potentiality for the use value, as it was described in previous sections, and for the communication language adopted. In particular, they allow you to link the Cultural Heritage with the dimension of time and history (timeline) or with the space and the reference area (mapping), while relations between actors in the community and the culture are developed through audio-visual narratives and the collaborative tools of Web 2.0.

Unfortunately, we don't have enough space to devote to each case a detailed analysis, but we can summarize in the table below some of the key features emerged.

		Case studies							
		DHS	Telling Lives, BBC	Sardegna Digital Library	Cultura Italia	Powerhouse Museum	AEC	MontiTV	Story - mapping
<b>Communication aims</b>	<i>Documentation</i>	√		√	√	√	√	√	√
	<i>Education</i>		√		√	√			√
	<i>Participation</i>		√			√	√		√
	<i>Collaboration</i>	√							
	<i>Promotion</i>			√	√	√		√	
<b>Users</b>	<i>Experts</i>	√		√	√				
	<i>Common people</i>		√	√	√	√	√	√	√
<b>Use value as:</b>	<i>Storytelling</i>	√	√						√
	<i>Performance</i>			√	√	√			
	<i>Place</i>						√	√	√
<b>Contents</b>	<i>Tangible artifacts</i>			√	√	√			
	<i>Cultural production</i>	√	√	√	√			√	
	<i>Local knowledge</i>		√	√		√	√	√	√
<b>Languages</b>	<i>Visual timeline</i>		√						
	<i>Visual mapping</i>				√		√		√
	<i>Video and multimedia</i>	√	√	√	√	√	√	√	√
	<i>Collaborative tools</i>		√				√		√
<b>Technologies</b>	<i>Streaming on line</i>		√					√	√
	<i>Library (On demand)</i>	√		√	√	√		√	√
	<i>Integrated devices</i>			√ (podcast)		√ (podcast)	√		√ (mobile)

We can focus on just certain elements with the aim of deriving the key factors in the proposal of design-oriented model for the cultural production and sharing.

The relationship between the scale of good (contents) and treatment (languages) is particularly evident in cases where the audiovisual and multimedia are used to enhance the narrative and emotional dimensions, reproducing knowledge and practices through audio-visual shot and, at the same time, producing new goods derived from the original ones (the document itself) and able to maintain and pass on new media that oral dimension typical of the cultural and tele-visual tradition. Orality is therefore not only a research methodology (cf. DHS), but also a documentary connotation of educational and popular communication. Moreover, Cultural Heritage visualization in relation to historical and geographical context, can convey information and widespread cultural knowledge from each single artifact to its relationship with the actors and the territory. The metaphors of the timeline and the map provide the user with a space of memory (7) triggering the performative dimension of interaction and the articulation of pathways for personal research and enjoyment. Digital technologies and the Web platform provide tools for georeferencing, while they encourage community participation and the production of content, on the other hand they tend to conform visual solutions and forms of representation.

In this regard it is useful to refer to the idea of Web as platform (8). Besides the technological feature of the distribution platform, it is necessary to consider the complexity of the environment in which it operates and the features of the Internet medium: the user is able to manage the information through a set of services, architectures of participation and collective intelligence. The scalability of digital content will allow the



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

dissemination and use beyond the individual device access, as the Web was a single big software upstream of all devices and common to all nodes in the network. Starting from this premise, then, you can think of a new metaphor for interfaces and models of knowledge management: it's no more a personal desktop but a place to exchange stories.

Finally, it is important to distinguish between a communication addressed to a community of interests and practices composed of experts and a very large or local community, that can access to Cultural Heritage without expertise and may or may not refer to a specific geographic area. Indeed, the goals of communication are in close relation with the characteristics of users, as well as the tone and style of project is therefore consistent with both aspects. Communication design will identify, therefore, adequate communication solutions both in terms of languages and technology of fruition.

## Conclusions

The systemic approach of design proposes, therefore, the cultural district as a system of goods and actors and as a communication system: within media convergence is useful moving towards integrated communications strategies that leverage the Net in order to articulate the Cultural Heritage communication into different formats and devices, that can produce value by multiplying the ways and contexts of use (from access to use). While it is true that the present condition is that the "always on", the Web is everywhere and represents an incredible potential for interaction, participation and collaboration (this order is a progressive path, from a simple access to an increasing specialized use). The cultural production traditionally understood as "high", faces diffuse and bottom-up practices, resulting in fertilization process of the cultural system and dialogue between actors in the district. In this sense, therefore, there cannot be cultural district which does not correspond to an integrated communication system. The design culture is proposing a participatory paradigma which is founded on one hand on mapping of cultural assets for their contextualization in historical, geographic and symbolic sense; on the other on the use value, understood as a re-appropriation and re-production of the goods themselves through the acquisition of interpretative tools and dialogue. Communication design is thus able to integrate the strategic dimension to the forms of expression more suited to the process of translation of knowledge, consistent with the specificity of the actors involved and oriented to the strengthening of knowledge networks.

## References

- [1] Toscano M. A., Per la socializzazione dei beni culturali, in Sul Sud. Materiali per lo studio della cultura e dei beni culturali, Jaca Book, Milano, 2004
- [2] Montella M., Valore e valorizzazione del patrimonio culturale e storico, Electa, Milano, 2009
- [3] Dorfles G., Le oscillazioni del gusto. L'arte oggi tra tecnocrazia e consumismo, Skira, Milano, 2004
- [4] Latour B., Reassembling the Social: An Introduction to Actor-Network-Theory, Oxford University Press, Oxford, 2005
- [5] Piredda F., Design della comunicazione audiovisiva. Un approccio strategico per la "televisione debole", FrancoAngeli, Milano, 2008
- [6] Kuka D., 'Linz Changes. Under the Urban Microscope', in Stocker G., Schopf C. (edited by), Ars Electronica 2008. A New Cultural Economy. The Limits of Intellectual Property, Hatje Cantz Verlag, Ostfildern, 2008, pp. 103-107
- [7] Yates F., L'arte della memoria, Einaudi, Torino, 1972
- [8] O' Reilly T., What Is Web 2.0. Design Patterns and Business Models for the Next Generation of Software, 09/30/2005, <http://oreilly.com/Web2/archive/what-is-Web-20.html>



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

## **Christoph MÜLLER, Anna WEYMANN, Bertram NICKOLAY, and Rodrigo Luna OROZCO DE ALENCAR**

### **Digitisation of library material: caught between user demands and preservation?**

#### **Abstract**

At the intersection between the continuously growing demand for digital information and the necessary preservation of cultural heritage, digitisation is desirable – and maybe soon unavoidable – for many libraries, archives and other institutions in the field of information science.

Especially smaller and medium-sized academic specialised libraries face the challenge of digitally preserving their unique, heterogeneous collections of different materials and formats and at the same time satisfying the demanding academic needs of their users.

A solution to the problem is sought by a team of interdisciplinary expertise. Working on a conceptual study, the library at the Ibero-American Institute (IAI) Berlin collaborates with the technology experts of Fraunhofer-Institute for Production Systems and Design Technology (Fraunhofer IPK) and arvato direct services Wilhelmshaven GmbH.

The project is characterised by a comprehensive collection and analysis of all factors that are crucial in connecting digital preservation and user demands in the best way possible. These findings shall help conceptualise beneficial, innovative and flexible technical solutions and workflows for automated digitisation of two-dimensional printed cultural heritage, and will lead us away from the dilemma of being caught between user demands and preservation.

**Keywords:** digitisation, library, technology, preservation, user demands

#### **A Conceptual Study for Comprehensive Digitisation Enterprises**

The Ibero-American Institute is an interdisciplinary centre that combines research, culture and information: the institute employs and supports international scholars and regularly hosts cultural exhibitions and events; its library is Europe's largest specialised library on Latin America, Portugal, Spain and the Caribbean. The users of the IAI are national and international scholars as well as students. In its capacity of a special collections library, the IAI acquires material on special subjects as thoroughly and exhaustively as possible. As a consequence, the collections are in large parts unique as well as characterised by various materials, conditions, shapes and formats. What is more, special collections libraries have an archival function and mostly only buy one copy. Consequently, in theory, every item sooner or later is due for preservation, and thus desired to be digitised.

Due to its users, acquisition strategy and embedment in science and research, the IAI combines features of academic libraries, archives and other knowledge institutions and can therefore function as an exemplary institution in a study that focuses on finding beneficial, innovative and flexible technical solutions and workflows for automated digitisation of two-dimensional printed cultural heritage.

In order to generate such a concept and to profit most from the digitisation of the IAI's collections, the interests of both the material as well as the users of the library have to be acknowledged, evaluated and respected as much as possible. For this endeavour, a team of traditionally separated sectors combines their competence. Personnel of the IAI contribute their knowledge about library and information science, collections, user needs and every-day workflows. Fraunhofer IPK, among the world's leading experts in the field of automated virtual reconstruction of destroyed documents, and scanning services provider arvato services contribute to the evaluation of the technical status quo with expertise in the fields of digitisation, handling of original material and technical requirements and engineering. This expertise is complemented by experiences learned about in good practice reports, interviews with representatives of German digitisation centres that collaborate with libraries, as well as a survey conducted among selected specialised libraries, archives and other, similar institutions. Furthermore, certain standards that have already arisen in the still young field of digitisation, as well as legal restrictions for Germany, will be collected and considered for the representation of a generic digitisation workflow.

#### **Collection Overview**

Among the first facts to be collected – with the help of a questionnaire conducted in the summer of 2009 – were those about what materials are held by specialised libraries, academic libraries, archives and other knowledge institutions. Since our study focuses on two-dimensional material only, this list lacks items such as audio-visual media that of course also form important parts of these institutions' collections.

Still, it turned out to be quite a panorama of different media: monographs, bound and unbound journals, newspapers, loose-leaf-collections, posters, sheet music, folded and unfolded maps, microfilms, microfiches, photographs (paper, slides, glass plate negatives), postcards, single documents, press-clippings, files, time-tables, brochures, blueprints, certificates, official gazettes, periodicals (proceedings etc.), patents, correspondence, diaries, medieval manuscripts, papyri, portraits, autographs, inherited special collections, art prints, sketches etc.

As far as the IAI's collections are concerned, they consist of about 1,000,000 monographs, newspaper and journal issues, about 300 bequests as well as other special materials collections, which include

- about 900.000 press-clippings
- about 200.000 microforms
- about 72.000 maps (topographical maps, city maps, roadmaps, historical maps and thematic maps, e.g. geology, highway systems, land use, settlement studies, languages, borders, botany) of which about 6.600 were produced between 1851 and 1945
- about 70.000 manuscripts (correspondence, notes)
- about 60.000 photographs (plus about 22.000 slides)
- about 10.000 glass plate negatives
- about 3.800 posters
- about 2.200 postcards.

For the purposes of the study, the IAI's collections are to be characterised as thoroughly as possible. Due to the high number of bound material (monographs, journals and newspapers) and owing the fact that they form, after all, the typical group of library material, the collections at IAI were divided into bound media and special media, i.e. bequests, maps, posters, press-clippings and images. As far as the first, 'traditional' group is concerned, it was decided to take a sample of 500 items and to note their characteristics in detail. The special materials were examined in single groups and their most important attributes noted (material, number of items, size and format, storage, damage, indexing).

### **Decisions for digitisation – influencing factors**

As has been established, any digitisation endeavour is, or rather should be, characterised by considering the various factors of preservation, user demands as well as the environment of the library.

#### **Material**

There are some media within the IAI's collections that qualify for preservation more than others, for example monographs and journals with acidic paper (about 15%). Also newspapers suffer from this lack in paper quality. Here even more, the paper literally dissolves when touched – some papers cannot even be taken out of the bundle. Furthermore, many items in the bequests (letters, scrap paper, photographs, and notes) are extremely fragile because of low-quality materials used and years of improper storage. The same can be said about the glass plate negatives: the layer of gelatine protecting the carbon motif is starting to come off and, moreover, they have been and still are unsuitably stored, so that their own weight is likely to crash them. Out of the photographs, many are bleached out or darkened, and on some of them a chemical reaction slowly renders the motif beyond recognition. As far as preservation is concerned, these would be the materials preferred for digitisation at the IAI. Apart from the mere scanning process, a digital removal of damages is of great interest.

Not necessarily damaged, but exceptionally rare are the IAI's bequests. Deceased scholars, some of them among the first Europeans to do research on Latin-America and its indigene cultures, have let their notes, scripts, drawings, photographs etc. to the IAI. On the one hand, this material draws scholars from all over the world to the institute, since most of the material is new to the scientific world. On the other hand, the collections are often in a bad condition, completely mixed up and not indexed at all – in other words, they do not qualify for material to be used in a library. Digitisation of these special collections would consequently mean both – preservation as well as preparation for scientific research. The presence of a digital copy would protect these materials from further handling; the originals could be stored away properly and only be offered for special research purposes.

While the general overview and the questionnaire showed that a wide range of materials exists, the sample of 500 revealed that even within a relatively homogeneous group (i.e. bound media), there is a wide range of different attributes that need to be considered – especially when having an automated process in mind: different paper qualities within one item (thickness, acid, size...), damages (yellowing, mould stains, tears, bends, acid, dirt, water stains...), aperture angle, attachments that are folded and bound within, print shining through, handwriting, gothic print or abnormalities in the layout (tables, landscape format, different fonts, irregular foliation).

## Users

At the IAI, users with a scientific interest form the major group. These users, most of them employed researchers or external scholars, work with all the IAI's materials, especially however with the bequests, of which they often are among the first to ever do research. On the other hand, many students of Latin-America related subjects use the library, since its collections are far wider ranged than those of the university libraries. These users mostly work with journals and monographs. For the project, IAI scholars and librarians established the requirements of both user groups regarding digitisation. To the scientific work of the scholars, it is important that...

- ... the digital copy is as authentic as possible (colour, fonts, notes should all be as in the original)
- ... all digital copies (also those of manuscripts) support full-text search
- ... (especially visual) media, e.g. maps, support additional functions such as digital navigation and connection with other material
- ... fragile materials are preserved
- ... rare materials and unpublished collections (if already properly indexed) are preferred in order to have them more easily and widely accessible
- ... the digitised material is presented in online platforms, since it would make the collections known, connect scientists and thus be fruitful to research.

Other users, mostly students and the interested public, consider being of importance:

- the content (as opposed to the shape, condition, feel...)
- a full-text search
- authenticity and integrity of the digital images
- a fast and easy access, since they often have a very limited time frame, preparing for an oral presentation, a report, a paper, a thesis
- excerpts: they often merely want a chapter, an article, sometimes only a few sentences and are happy to leave the heavy book in the library and only extract what they need, e.g. on portable storage devices or via e-mail
- a simultaneous presence of one work for when there are several users interested in the same subject (e.g. exam preparations)

## Library

Apart from considering the demands of the users and the necessities regarding the materials, it is just as important to ask the librarians their opinion about the essential needs to satisfy both sides' demands as well as possible.

From the librarians' point of view...

- ...the digitisation process should be adaptable and integrable to the current book processing at the library, because only perfectly integrated does the process save time and further allow the library to keep relying on their own expertise regarding formal and subject indexing
- ... the process should interfere with the daily work (especially circulation and return) as little as possible
- ... there must not be any violations of copyrights
- ... automated formal and subject indexing as well as quality control during the scanning process would be ideal.

## Standards and Legal Questions

It is advisable the digitisation process follow certain standards. In Germany, the Praxisregeln Digitalisierung by the DFG (German Research Foundation) serve as a benchmark in this case. They regulate criteria such as factors influencing what to be scanned, file formats, generating full texts, organisation of the resulting metadata, authenticity, image quality (colour, size), storage (short and long-term), data exchange, integrity or presentation. Even though these standards are only binding for projects funded by the DFG, these rules cover many points and are constantly improved by practical users. Our study will consequently attempt to respect these rules in the conceptualisation of technical solutions. Digitisation can, of course, only be realised with material that does not fall under copyright. In general, the act of digitising is, just as the paper copy, an act of duplication and therefore only allowed under certain circumstances. For example, the point in time when a publication does no longer fall under copyright differs from country to country (in Germany, 70 years after the author's death). In general, attention has to be paid regarding who gets which materials where, in what number and for what purpose.



## Conclusion and Next Steps

Digitising library material represents an intersection between user demands and preservation. This is especially true for specialised academic libraries and similar institutions, holding collections of various materials and formats. Digitisation improves the institutions' services in terms of a faster, easier access to documents that are fully searchable, and at the same time helps preserve rare and damaged originals. Different materials and their conditions as well as requirements of different user groups and the library will have to be considered and brought together, integrated into existing workflows, acknowledging established standards and legislation. All these factors have to be considered in the conceptualisation of new technology, and the above mentioned demands from the various perspectives set quite high expectations for the digitisation system, which should allow:

- flexible automatic digitisation of all printed materials in different formats
- interactive quality control
- identification and indication of expired copyrights
- an excellent image quality, suitable for OCR (Optical Character recognition) including manuscripts and gothic print
- easy generation of structured metadata through layout analysis modules (automated indexing)
- at least one master copy, and some commercial surrogates for publication, one representing the original and one full-text searchable version
- integration of the process into the existing workflow at the institution, including personnel, finances, logistics, presentation
- careful handling of sensitive material
- long-term preservation of the digital files
- and of course a reasonable price to be viable for the target institutions.

In the second phase of the project, which ends in July 2010, a technology monitoring will be performed in order to identify new solutions and its capabilities, as well as possible opportunities for further development and research. The Fraunhofer IPK and arvato services specialists with the help of expert interviews will define comparison criteria for technological approaches that attempt to satisfy the demanding requirements of a digitisation process. This is meant to find the optimal implementation of adequate technology. The framework for this phase is provided by the information gained through the characterisation of the IAI library, the conducted survey and the collected experiences of practical users of digitisation techniques. Furthermore, Fraunhofer IPK and arvato services will put together their knowledge about virtual reconstruction of destroyed documents to complement this assessment process.

Digitisation offers many benefits for our cultural heritage as well as our libraries and their users. As it turns out, this implies a wide range of requirements and an evaluation of which technical solutions are needed in order to fulfil them. After having collected all demands, our study will conceptualise a solution that is beneficial and flexible in terms of technology and workflow, and which puts us in the position to say that when it comes to digitising printed cultural heritage, we are not caught between user demands and preservation.

## References:

- [1] Praxisregeln Digitalisierung by the German Research Foundation (DFG):
- [2] [http://www.dfg.de/forschungsfoerderung/wissenschaftliche\\_infrastruktur/lis/download/praxisregeln\\_digitalisierung.pdf](http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/download/praxisregeln_digitalisierung.pdf)

## **Laura PECCHIOLI, Fawzi MOHAMED ,and Marcello CARROZZINO**

### **ISEE: retrieve information in cultural heritage navigating in 3D environment**

#### **Abstract**

As the whole field of preserving, documenting and studying the Cultural Heritage is interdisciplinary, and the way in which information is managed is not homogenous. Moreover the objects belong often to the real world and present a 3D component, not easy to represent using only two-dimensional approach. Storage, organisation and retrieval of such information in real time is challenging and commonly not very well structured. Very often the only unifying entity is the "object", which the information is related to, so an effective management of related data still represents a serious problem. 3D visualisation simulates spatial reality, allowing the viewers to more quickly recognise and understand what they see in the real world. Cultural heritage draws together several different professions. Furthermore, the relationship between the conservation managers, who are often unfamiliar with current documentation techniques, and the providers of the information, who tend to be highly technical practitioners without expertise in cultural heritage, is not easy to handle.

We present a new method to access spatial information through the interactive navigation of a synthetic 3D model, which reproduces the main features of a corresponding real environment. The information is ranked with a novel measure of the relevance, that depends on the position/orientation in the 3D space, allowing users to retrieve significant information. To give access to a larger audience, the method is accessible through an intuitive and user-friendly interface on normal Web browsers.

The system has been applied to case studies related both to outdoor and indoor environments. Actually the developments are relative as an interactive smart guide.

In particular we believe that an intuitive interaction in real time and in the context makes more accessible the information, and can help users in being more active, and learn in interesting ways.

**Keywords:** interactive 3D interface; relational database; gaussian; spatial relevance; overlap; XVR 3D engine.

#### **Introduction**

The Cultural Heritage normally refers to objects in the real world with a 3D component and often requires a uniform treatment to massive heterogeneous data. To keep in account these issues, the research started to focus on 3D representations. Nevertheless, using a 3D environment allows for a closer adherence to the real world (preserving location related data) and permits to respect the spatial relationships among different components. Our aim has been to develop a new approach to access and manage information, paying particular attention to cultural assets data management. This approach, called ISEE ("I see"), will be based on "interactive 3D models", because an interactive interface allows for a more natural behaviour, where the user can move freely and find the sections he/she is interested in. ISEE should be able to provide retrieving information by just looking around in a 3D environment, as moving and looking at the world is the main modality we use to gather information from it.

#### **State of art**

Today, in the field of Cultural Assets, it is very important to continuously research new ways to represent and query data. Normally one has to deal with information from different sources and formats, and with a lot of data produced in a short amount of time. Another problem is the communication between who provides recording, documentation and information management tools, and the professionals in cultural heritage management who use them. The ICOMOS/ISPRS Committee for Documentation of Cultural Heritage (CIPA Heritage Documentation) conducted a series of workshops between 1995 and 1999 to understand this incompatibility. Often conservation managers are unfamiliar with current documentation techniques, while the providers tend to be highly technical practitioners without expertise in cultural heritage [1]. One possibility to communicate the structure of a piece of information is to visualize it using a graphical representation. "Information visualization" is a wide field interdisciplinary in nature and represents one important process to transform and represent a large variety of data.

Internet has changed the way we organise data, but an integrated management method for cultural heritage ICT applications is still not available. The Virtual Reality (VR) or the Augmented Reality (AR) technologies are able to reconstruct 3D models of ancient culture, making them accessible to modern-day users [2]. This can be useful for members of the general public, but for specialists working in the fields of archaeology or restoration, this approach is not necessarily so useful and can even be misleading. 3D interaction is an intuitive paradigm

for a majority of people and additionally can convey more information. It is clear that the problem of moving from 2D to 3D is complex. The use of virtual reality can help to understand and to manage the real world, but the transition between these realities is not easy and not always the result satisfies the expectations. Often one can see a 3D model with a high level of detail and it's an excellent work. But it's not always possible to read the information while navigating within the model. It's necessary to decide "what must have the priority".

Actually, Web applications as ISEE are increasingly used, because they allow access from any computer, while keeping a centralised repository of information. The quality of their interface has practically reached parity with desktop applications.

## Motivation

We have chosen an interactive interface, because we want to involve the user: interacting one can learn more. We think that an intuitive interaction, such as looking at the environment where information is contextualised, is one of the solutions closest to normal human behaviour. Our intention is to create an environment that enables the co-operation and the exchange of knowledge among users. The interface, we propose for ISEE, enables the user to explore a three-dimensional space, where the objects are geographically referenced, and retrieve the related information. In particular professionals from the field should benefit from a greater level of freedom to manage and manipulate the information retrieved, with tools specific to their field. They will be able to insert more detailed data and to decide presentation and retrieval modes.

## The method

Our approach started considering the requirements for an adaptive and intuitive interface to access information. Thinking about it we had the idea of using the simple action of seeing as "a common language" to query and insert information. To achieve this, one has to define for each "view" the region on which the user is focused in that moment, that we called View Zone (VZ). At this point one can either be interested to recover the relevant information about that zone, or add new information about it. The complexity dictated by the type of data is simplified in a few easy moves (navigating an environment and looking around).

We decided to treat approximated zones to represent the objects in 3D space. In the method our information is associated with regions of the space, which we called Information Zones (IZs). The Information Zone does not have to coincide with a 3D object represented in the model, but they might be just a part of it or include many objects at the same time.

To define the region in a precise way that a computer can understand we used "3D gaussians": this is a function which assigns a value to each point of the space (it can be seen as a fog, more dense in the centre, and less dense on the periphery), which can be used to adequately describe the concentration of information. The interactive 3D viewer (VZ) and the IZ are approximated with a normalized 3D gaussians, and this provides to have a symmetric treatment. It allows us to use the interactive 3D viewer to visually insert the IZ of a piece of information (authoring), or to jump immediately to the view related to some information (retrieval). An innovative aspect is the definition of spatial relevance of information. A ranking calculates our relevant information and depends from the View Zone and the location of the information. The measure of relevance is depending on the spatial relationships between VZ and IZ (Figure 1). Intuitively, the relevance of information should be "maximal" (relevant) when its Information Zone (IZ) coincides with the View Zone (VZ) (Figure 2), decreasing when they are far apart. An IZ that has a size comparable to the current VZ is probably more interesting than an IZ that has a size very different from the current VZ.

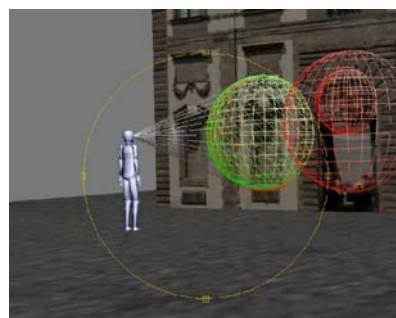
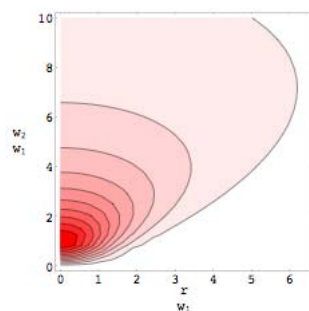


Figure 1 Contour plot of the overlap of two normalized gaussians with width "w1" and "w2" and a distance "r" between the two centres. Darker means a higher overlap.

Figure 2 Third person view of VZ and IZ from the side, the VZ coincides with an IZ and has maximum overlap. The VZ is represented as a green sphere, the IZ as red spheres, and the view cone is gray.

Using as technique of accessing information our overlap, based on the distance and the size of the VZs and IZs, is more accurate respect of other systems based on distance (e.g. GoogleMaps) or on the selection (3D interfaces, games etc.). These methods typically use the distance from the user not from where he/she is looking at as extra simplification. In an interactive 3D model “the level of zoom” depends on the distance of the object looked at, and in the same scene one can have different levels of detail.

Moreover the system can manage with high densities of data, because its usage provides an extra means to filter the information reduces the amount of information retrieved.

### The case studies

The prototypes developed in this work represent Web applications, where the user can explore in intuitive way a model and to discover the information linked to it. In order to visualize and interactively navigate the model on the Web, we used the XVR technology [3], jointly developed by PERCRO Scuola Superiore Sant’Anna in Pisa (Italy) and VRMedia s.r.l.. The 3D model is downloaded on the user client as soon as the user accesses the Web page. As soon as the download is finished, a first list of information automatically appears, presenting on the top the data most relevant for the zone the user is currently looking at. The structure of the archive implemented so far is quite simple. The information is registered in a file system (in xml, jpeg, tiff etc.) and stored as meta-information in the relational database (MYSQL) to have a fast access in real time (query, add etc.). A nice consequence of storing II information also as file is that normal tools for the automatic indexing of files could be used in the future to index the meta-information and document files to allow full text searches [4] even if they do not take into account temporal and spatial information. The last version of ISEE can upload files in kml, a format standard of Google Maps.

The development of the method started from a first case study: the crypt in St. Servatius in Quedlinburg (Sachsen-Anhalt, Germany), part of the World Heritage List (UNESCO). It presented different types of information related to the restoration and a cloud of points from a 3D scanner. The crypt represents one of the largest painting cycles of the 12th century in Germany [5][6][7]. The research and the work developed by Prof. H. Leitner and his students of the Hochschule für Bildende K nst of Dresden and is work in progress too. In particular it has been realized using GIS format extensively as a documentation tool, with the “base map” consisting of high-resolution rectified georeferenced photographs. The crypt can represent a prototype for sharing, query and add the information among professionals figures or common users (Figure 3).

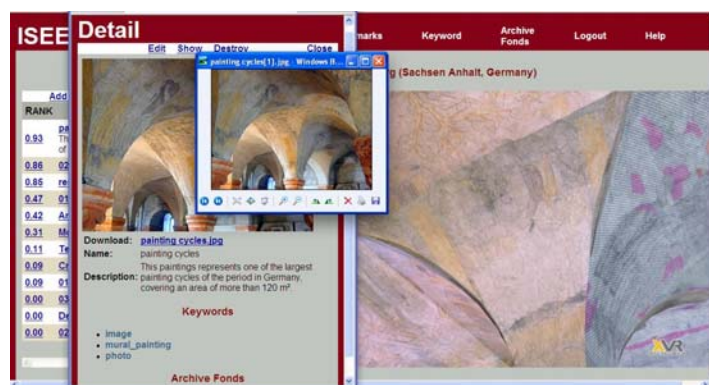


Figure 3 The interface of the crypt of St. Servatius: the mural paintings and the maps of the state of conservation in the same view.

### Developments

We are working for a use of a Web application supporting all browsers and to provide ISEE as city sightseeing. In the past the method had been already applied in real world with good results, using a GPS Compass (Vector CSI Wireless), providing 2D heading and positioning data connected to a laptop in Piazza Napoleone in Lucca, Italy [8]. The actual system on smart device works by a similar approach of the web application. The data itself is gathered and stored using a REST interface to the ISEE Web server (Figure 4a-4b). For efficiency reasons the interaction on the device is mostly 2D.



Figure 4 a) The actual development on mobile device b) The application ISEE in Berlin

## Conclusions

The interactive visualization makes information more accessible and improves the user experience. The work presented provides an intuitive and user-friendly interaction for accessing, inserting and modifying information in a 3D space. The method is suitable to all categories of users, both professional and non professional, because it is based on the simple action of navigating the 3D space and retrieving the information. Moreover the information is associated with regions of the space, and thus pre-processing of 3D models to subdivide them in suitable logical elements is not needed. New pieces of information can be inserted in the same way in which they are queried, just by looking. The use of extended zones allows us to use a ranking algorithm with superior performance than rankings based only on the distance. The proposed ranking algorithm matched the intuitive expectation of the users, as was verified with a formal usability test that was performed at completion of the work. The method we propose is intended to: allow easier information handling; use only simple and standard formats, in order to facilitate communication and exchange; provide the option of detailing the information source.

## References

- [1] [http://www.getty.edu/conservation/field\\_projects/recordim/index.html](http://www.getty.edu/conservation/field_projects/recordim/index.html)
- [2] B. A. Doug, E. Kruff, J. La Viola and I. Popyrev, 3D User Interfaces, Theory and Practice, Addison-Wesley, USA, 2005.
- [3] M. Carrozzino, F. Tecchia, S. Bacinelli, and M. Bergamasco, Lowering the Development Time of Multimodal Interactive Application: the Real-life Experience of the XVR project, In: Proceedings of ACM SIGCHI International Conference ACE, November 19 – 22, Valencia, Spain, 2005, pp. 270-273.
- [4] L. Pecchioli, F. Mohamed, A method to access the information through an interactive 3D virtual environment, In: Entwicklerforum Geoinformationstechnik- Junge Wissenschaftler forschen – Technische Universität Berlin, Institut für Geodäsie und Geoinformationstechnik, Berlin, Deutschland, (Eds) Shaker Verlag, 2008, pp.119 – 129.
- [5] Gosslau, Friedemann and Radecke, Rosemarie, Die Stiftskirche zu Quedlinburg, Eine Führung durch den romanischen Sakralbau und den Domschatz, (Eds.) Convent-Verlag, Quedlinburg, 1992.
- [6] H. Leitner, La conservazione delle pitture murali nella Cripta di San Servatii a Quedlinburg, In Arcadia Ricerche Eds., Proceedings of Bressanone, Scienza e Beni Culturali, XXI - Sulle pitture murali, Bressanone, Italy, 12-15 July 2005, pp. 233-240.
- [7] L. Pecchioli, F. Mohamed, M. Carrozzino, ISEE: accessing information navigating in a 3D virtual Environment. The case study of the crypt in St. Servatius in Quedlinburg, Saxony-Anhalt (Germany), In: Web Portal - Architectural Image-Based-Modeling, 2009.
- [8] L. Pecchioli, F. Mohamed, M. Carrozzino, H. Leitner, Accessing information through a 3D interactive environment, In: Proceedings of ICHIM07 Digital Cultural and Heritage, Toronto, Ontario, Canada, 2007. (<http://www.archimuse.com/ichim07/papers/pecchioli/pecchioli.html>)



Empowering users: an active role  
for user communities

INTERNATIONAL CONFERENCE  
Florence 15<sup>th</sup> - 16<sup>th</sup> December 2009

## SATELLITE EVENTS

### TUTORIAL

#### **LONG TERM PRESERVATION OF DIGITAL ASSETS: BASIC CONCEPTS AND PRACTICES**

Monday 14th December

The event brought together international experts who developed a one-day full immersion tutorial about issues related to long term preservation of digital objects. The tutorial started setting the scene about current initiatives and approaches, then it gave participants an understanding of the key digital preservation issues and decisions to be taken during the lifelong cycle of a digital archive. Some clear concepts, recommendations and “to do” list of things were presented. The major challenges and the most prospective solutions were introduced, even if findings in this field are not so mature. The experts defined needs and experiences about the specific cultural heritage sector, providing the audience with some technical recommendations about standards on digital archives, metadata, digital formats, strategies and criteria to certify tools and practices, check risks and help to take the right decisions for preservation planning. After lunch, the session started with an interactive hands-on work, where concrete experiences and practical tools developed by some of the most important European projects were presented and demonstrated. Target audience were librarians, archivists, museum curators, students, researchers and professionals in the sector of digital archives management, digital libraries, Internet applications and multimedia content creators.

### TUTORIAL

#### **DUBLIN CORE - BUILDING BLOCKS FOR INTEROPERABILITY**

Thursday 17<sup>th</sup> December

This Tutorial describes the Dublin Core Metadata Initiative, the history of the organization from a group of interested experts in 1995 to a formal organization with the legal incorporation in late 2008, and outlines the strategic directions and collaboration with other metadata initiatives over the years and the strategic directions that DCMI will be pursuing in the near future. After a brief outline of the organizational history of DCMI with presentation of the communities, task groups, processes, committees and operating rules, the tutorial will provide a general introduction to Dublin Core metadata, technical trends in the Dublin Core community over the past decade, and alternative approaches to descriptive metadata in the "Dublin Core" style. The second part of the tutorial reviews implementation technology alternatives using HTML, XML, and RDF, including new techniques such as embedded RDF a metadata in support of structured search.

Alongside the traditional paradigm of metadata interoperability on the basis of pre-coordinated agreements on natural-language definitions and specific data structures, the new paradigm of Linked Data offers a flexible framework for coherently merging diverse types of metadata on the basis of a shared underlying data model. The tutorial covers methods for expressing controlled vocabularies as Linked Data such as Simple Knowledge Organization System (SKOS).