Implications of the Web Semantization on the Development of Digital Heritage

Dr. Gustavo Urbano Navarro  
*Universidad Nacional de la Patagonia Austral. Unidad Académica San Julián, Argentina,*  
gusunavarro@gmail.com

**Abstract**  
The general objective of this document lies in the construction of a network-interoperable digital repository (DR), of open access for the filing, preservation and distribution of documentary material—local press and photographs of the first half of XX century concerning towns in the Patagonia region (Puerto Deseado, Puerto Santa Cruz, Comandante Luis Piedrabuena y Puerto San Julián), for the materialization of its available use within the academic environment, as well as in the secondary and tertiary levels of the educational system, and the community in general. Its result will consist of developing an experimental prototype focused on technical, descriptive and structural processes of the DR (metadata schema), and its interoperability.

**Author**  

1. Memory of the Patagonia Austral (www.koluel.org)  
The general objective of this document lies in the construction of a network-interoperable digital repository (DR), of open access for the filing, preservation and distribution of documentary material—local press and photographs of the first half of XX century concerning towns in the Patagonia region (Puerto Deseado, Puerto Santa Cruz, Comandante Luis Piedrabuena y Puerto San Julián) for the materialization of its available use within the academic environment, as well as in the secondary and tertiary levels of the educational system, and the community in general. Its result will consist of developing an experimental prototype focused on technical, descriptive and structural processes of the DR (metadata schema), and its interoperability.

Thus, the intention is for users to retrieve the stored material, by browsing the DR content through the search interfaces. The problem arose in the Patagonia region in response to the absence of a sustained policy for the long-term preservation of the historical heritage.
As a result of this, researchers do not have, in many cases, the possibility of access to collections of data or sources existing in municipal libraries, records and local documentation centers, since the contributions from the interior regions to the national history are increasingly significant. In our historiography there are lots of generalizing works supported by empirical studies carried out in and from the central area, without devoting too much study to realities and dynamics.

Therefore, the proposal refers to the development of such a repository taking into account the specific context of the historical heritage and the scientific legacy of Patagonia, the preservation and management of digital records, the data guardianship, as well as the widespread and maximization of visibility, use and impact of the regional scientific output in the international community.

That is to say, it has a double objective: on one hand, to preserve those contents and to enable the access and use either for teaching purpose or for academic research, encouraging the local specificity; and, on the other hand, to provide feedback of such research and give support to the electronic publications in the area of Social Sciences in the institutions of the region. Consequently, the creation of this repository intends to retrieve in digital files the regional cultural heritage, aiming to increase its visibility and impact.

2. Introduction

2.1 Research and Development Lines

This paper is structured around two convergent axes:

1. The digital preservation of the social historical memory and the cultural heritage of the region: management of digital files for long-term access and data curation; and
2. The development of an experimental prototype focused in technical processes and in its interoperability with other technological platforms.

3. Achieved/Expected Results

In recent years, information and communication technologies have provided new resources for the management of information, offering additional services for using documentary materials, apart from consultation and organization through the catalogues that digital libraries made available, enlarging and renewing their services, and opening their domain to other types of information resources. Hence, the creation of an e-infrastructure appears as a specific solution, which consents to collect, provide access to and share educational resources, having at its disposal a content storage system that is easily integrated and communicated with other operative systems (McLean & Lynch, 2003).

For this reason, e-infrastructures started to take a stand as important tools the function of which is to protect the resources, making them available to be shared with other applications, thus facilitating the content stream. This means that they include, apart from storage systems, toolsets that are useful for the reusing process. Although there are portals of the content administrator type, or collaborative work environment, which could certainly meet some of the requirements offered by repositories, from the technological point of view, one of the core problems lies in the fact that the export and import of bibliographic data is usually limited in those cases. In this way, they configure new and complex systems that enlarge the abilities of resources turning them into tools for learning and research.
Municipal libraries and documentation centers of the different towns in the Patagonia region have documentary materials, many of which are deteriorating on account of the lack of financing for the maintenance and sustainability of projects for their preservation.

The main objective of the project is to provide a possible answer to these problems, through the creation of a new structure for the processing of information, thus enabling the data retrieval and contextualization and its integration into flexible networks of knowledge, as well as with social networks.

4. Background

In the early 90s, the first initiatives of open e-infrastructure of specialized documents appeared with the scope of facilitating access to content, which was restricted to those who could afford them. One of the main drivers of the creation of these digital repositories was the Open Access movement in the United Kingdom. The free availability of its content means that any user may read the documents therein contained, without requiring either subscription or registration, that is to say that its availability is free, and the only restriction upon the distribution and reproduction is to grant authors the control on the integrity of their work and the right to be properly quoted and recognized (Budapest Open Access Initiative, 2002).

By the end of 2008, the European Commission launched the “Europeana” project, an e-infrastructure that does not constrain its collection to texts, but includes all kinds of cultural objects, such as: texts, photographs, videos, maps, manuscripts, paintings, newspapers and historical documents on record, i.e. the contribution of the repositories from all over Europe.

The creation of this digital object repository comes up as a solution to facilitate the compilation, the preservation, and the diffusion of differently supported documentary sources, that is it does not limit its collection to texts, but it also includes any kind of cultural objects, such as: texts, photographs, videos, maps, manuscripts, paintings, newspapers and historical documents on record and oral file of voices, providing a content storage system which is integrated and communicated with other systems, enabling the stream of knowledge; and carrying out the following intersectional objectives:

- Strengthen the contribution of culture to support sustainable development;
- Promote the role of culture in the interaction and dialogue among different cultures in the development policies aiming to social coherence and reconciliation;
- Improve the universal access to information and knowledge.

5. Development

During 2011, the team started to think about developing an electronic platform based on specific software. In order to define this platform, some documents that gave details of the comparison among Fedora, Dspace and Eprints, were previously read.

We threw ourselves directly to Fedora (Flexible Extensible Digital Object Repository Architecture), which was originally developed by researchers of Cornell University as an architecture to store, to administrate and to accede to digital content under the shape of digital objects inspired in the framework of Kahn and Wilensky (Kahn and Wilensky Framework). The project called Fedora
Repository Project, implements Fedora’s abstractions in a robust open-source software. It provides central repository services under the shape of web services with well-defined APIs. Moreover, it provides a collection of services and applications such as search, OAI-PMH, messaging, customers and more. But the problem of Fedora is that its interface is not so friendly; consequently, it was thought to use some development including Fedora, and so it was decided to test Islandora.

Islandora is an open source project underway at the Robertson Library at eSciDoc fedora the University of Prince Edward Island. Islandora combines the Drupal and Fedora software applications to create a robust digital asset management system that can be used for any requirement where collaboration and digital data stewardship, for the short and long-term, are critical.

There are a number of initiatives under the umbrella of the Islandora project, all developed using the innovative Islandora software suite, including:

- VRE (Virtual Research Environment) - a collection of customized Islandora sites used by researchers at UPEI and elsewhere to steward research data. More information about VRE is available from the links on the project's web site.
- IslandArchives.ca - a number of digital collections created by UPEI and partners, providing access to a wide range of digital documents and materials.
- Repository-In-A-Box - a unique solution for building Institutional Repositories and as seen with UPEI's own IslandScholar site.

Especially taking into account that the Islandora project provided for the framing of repositories on newspapers, images and voices, different tests were carried out by virtue of Koluelfs objectives. After several tests, the decision was not to use Islandora since previous knowledge on Drupal was required and this fact required great dependence with the Islandora group. Once the testing was finished, it was decided to use much of 2011 in testing other products under the Fedora interface. Namely Hydra.

Hydra is a repository solution that is being used by institutions to provide access to their digital content. Hydra provides a versatile and feature-rich environment for end-users and repository administrators alike. The project gives like-minded institutions a mechanism to combine their individual repository development efforts into a collective solution with breadth and depth that exceeds the capacity of any single institution to create, maintain or enhance on its own. The motto of the project’s partners is “if you want to go fast, go alone. If you want to go far, go together.” Hydra is an ecosystem of components that lets institutions deploy robust and durable digital repositories (the body) supporting multiple “heads”: fully-featured digital asset management applications and tailored workflows. Its principle platforms are the Fedora Commons repository software, Solr, Ruby on Rails and Blacklight. And above all, it works in free software. Unlike Islandora, Hydra is a great framework and it is being developed.

After performing all tests with Hydra, the next step was to work with eSciDoc.

- eSciDoc is a platform that contains Fedora-commons inside its structure and contains several applications.
- eSciDoc is a system targeted at research organizations, universities, institutes, and companies interested in eScience-aware knowledge and information management.
- eSciDoc enables you to publish, visualize, manage, and work with data artifacts (or objects). Objects include both publication data and research data across disciplines.
• eSciDoc provides a generic infrastructure and specialized solutions within the context of research questions. It integrates existing solutions and implement new ones.
• eSciDoc addresses aspects of data reliability, data quality, data curation, and long-term preservation. It covers the whole lifecycle of objects.
• eSciDoc supports semantic relations between objects.

The eSciDoc system is designed as a service-oriented architecture (SOA) implementing a scalable, reusable, and extensible service infrastructure. Application and discipline-specific applications can then be built on top of this infrastructure. The heterogeneity of the envisioned solutions in addition imposes an efficient handling of different kinds of content.

The service-oriented architecture fosters the reuse of the existing services. Other projects and institutions may reuse an eSciDoc service, either remotely or locally, thus becoming one building block with a broader e-Science infrastructure. At the same time, the SOA approach of eSciDoc comes with other advantages.

Instead of a complex and monolithic application, the eSciDoc service infrastructure can be seen as a set of loosely coupled services, which can be specified and implemented independently. This allows for an iterative implementation strategy for services. First services may already be implemented while others are still in their design phase. Based on feedback from early adaptor users ("pilots"), new services can be easily added, thus fulfilling user expectations in a more timely and user-driven manner.

The core technology used to implement the services is based on Java and XML. Instead of building the infrastructure “from a scratch”, the eSciDoc team chose to integrate existing open-source components as much as possible. eSciDoc services in general provide both SOAP and REST style interfaces. This allows for further development of solutions without constraining the selection of the programming languages, thus accelerating their implementation and enabling the involvement of various developer groups. Even simple scripting and "Web 2.0"-style mash-ups are supported.

The eSciDoc service infrastructure groups its services into three service layers: basic services, intermediate services and application services. The software used under eSciDoc platform was PubMan. PubMan supports research organizations in the management, dissemination and re-use of publications and supplementary material. It can be used out-of-the-box within the eSciDoc infrastructure. Being open source, it will be customized and extended for institutional- and discipline-specific needs.

Finally, a last test was carried out with a Content Management System CMS and Plone was chosen. Plone lets non-technical people create and maintain information for a public website or an intranet using only a web browser. Plone is easy to understand and use—allowing users to be productive in just half an hour—yet offers a wealth of community-developed add-ons and extensibility to keep meeting needs for years to come. Blending the creativity and speed of open source with a technologically advanced Python back-end, Plone offers superior security without sacrificing power or extensibility.

The Plone community is an incredibly diverse group that bridges many types and sizes of organizations, many countries and languages, and everything from technical novices to hardcore programmers. Out of that diversity comes an attention to detail in code, function, user interface and ease of use that makes Plone one of the top 2% of open source projects worldwide. (Source: Ohloh)

Plone’s intellectual property and trademarks are protected by the non-profit Plone Foundation. This means that Plone’s future is not in the hands of any one person or company. Thousands of websites across large and small businesses, education, government, non-profits, sciences, media and publishing are using Plone to solve their content management needs. This is supported by a global network of over 300
solution providers in more than 50 countries. Looking for a hosting provider to host your Plone site for you? You can find a list of providers and consultants on plone.net.

We are very proud to be known by the company we keep. Organizations as diverse as NASA, Oxfam, Amnesty International, Nokia, eBay, Novell, the State Universities of Pennsylvania and Utah, as well as the Brazilian and New Zealand governments—all use Plone.

Plone is open on many levels. It runs on Linux, Windows, Mac OS X, FreeBSD and Solaris—and offers a straightforward installation to get you up and running in minutes. It has been translated into more than 40 languages, and is developed with an unflinching emphasis on usability and standards compliance. Need a CMS that integrates with Active Directory, Salesforce, LDAP, SQL, Web Services, LDAP or Oracle? Plone does. Need to be sure your website is accessible? Plone meets or exceeds US Government 508 and W3C’s WAI-AA standards.

Worried about security? As an open source product, a large number of developers frequently scrutinize the code for any potential security issues. This proactive approach is better than the wait-and-see approach in proprietary software that relies on keeping security issues a secret instead of resolving them outright.

Based on Python and the Zope libraries, Plone has a technological edge that has helped it attain the best security track record of any major CMS (Source: CVECVE). In fact, security is a major reason why many CMS users are switching to Plone.

The market is full of open source content management systems, so it is important to do your homework before choosing one for your organization. Remember that a simple CMS may work out great to start with, but lead to problems with scaling or migration when you need more capability than it can provide. At the other end of the spectrum, a powerful CMS can be so difficult to learn and maintain that it never gains acceptance to users. Make sure the CMS you choose meets your needs today without compromising future growth.

In the end, it was decided to use Plone for the Koluel project. The decision was taken on account of its friendly interface, its translation into Spanish among many of the available languages, the integration with OAI and the possibility to program interfaces for cataloguing and classifying documents under Dexterity. Dexterity was created to serve two audiences: Administrators/integrators, and developers.

For administrators and integrators, Dexterity offers:

- The ability to create new content types through-the-web;
- The ability to switch on/off various aspects (called "behaviors") on a per-type basis;
- Improved collaboration between integrators (who may define a type's schema, say) and programmers (who may provide re-usable behaviors that the administrator can plug in).

For developers, Dexterity promises:

- The ability to create content types more quickly and easily, and with less boilerplate and repetition, than what is possible with Archetypes or plain CMF types;
- Content objects with a smaller runtime footprint, to improve performance;
- Types that use the now-standard zope.interface/zope.schema style of schema, and more broadly support modern idioms that sit a little awkwardly with Archetypes and its ilk.
As far as Koluel is concerned, work was accomplished jointly with the Zest software company, especially with Maurits van Rees in order to develop a product that uses all Dublin Core fields. The product is available in github and can be freely used.

References


Brown, Alejandra M; M. Alejandra González y Marisa G. Velazco Aldao. s.f. Proyecto piloto de desarrollo de la Biblioteca Digital de Tesis y Disertaciones del Instituto Balseiro. (Inédito)


