A First National Italian Register for Digital Resources for Both Cultural and Scientific Communities (Communication)

Maurizio Lunghi

Fondazione Rinascimento Digitale, Firenze, Italy
lunghi@rinascimento-digitale.it

Abstract. In this paper we present an Italian initiative, involving relevant research institutions and national libraries, aimed at implementing an NBN Persistent Identifiers (PI) infrastructure based on a novel hardware/software architecture. We describe a distributed and hierarchical approach for the management of an NBN namespace and illustrate assignment policies and identifier resolution strategies based on request forwarding mechanisms. We describe interaction and synergy with the 'Magazzini Digitali' project for the legal deposit of digital contents just launched by the Italian Ministry of Culture. Finally, we draw some conclusions and point out the future directions of our work.

1 Introduction

Stable and certified reference of Internet resources is crucial for 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) and Europeana, 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. A trustworthy solution is to associate to any digital resource of interest a Persistent Identifier (PI) that certifies its authenticity and ensures its accessibility. Actually some technological proposals are available, 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 from the scientific community, leaded by the National Research Council (CNR) through its Central Library and ITC Service. We have designed a hierarchical

M. Agosti, F. Esposito, and C. Thanos (Eds.): IRCDL 2010, CCIS 91, pp. 47–53, 2010.

distributed system, in order to overcome the criticalities of a centralised system and to reduce the high management costs implied by a unique resolution service. Our approach implies a central node responsible for the NBN:IT top-level Italian domain, and lower-level nodes each responsible for managing one of the Italian sub-domains (NBN:IT:UR, NBN:IT:UR:CNR, NBN:IT:FRD, etc.). The number of levels within this hierarchy is virtually unlimited. Only the nodes at the lowest level harvest metadata from the actual repositories and create NBN identifiers. The upper level nodes just harvest new NBN records from their child nodes and store them within their databases. In this way each node keeps all the NBN records belonging to its sub-domain. It is easy to see that within this architecture the responsibility for name creation/resolution is distributed and information about persistent identifiers is replicated in multiple sites, thus providing the necessary redundancy and resilience for implementing a reliable service.

2 Persistent Identifiers Standards

The association of a Persistent Identifier (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.

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.

The Document Object Identifier system (DOI) 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) 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 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) 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) 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.

The National Bibliographic Number (NBN) 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. 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.

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

3 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 was released demonstrating the viability of the hierarchical approach. 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). At the beginning of 2009 the Italian National Research Council (CNR) developed a second prototype.

Objectives

The project aims at:

- creating a national stable, trustable and certified register of digital objects to be adopted by cultural and educational institutions;
- allowing an easier and wider access to the digital resources produced by Italian cultural institutions, including material digitised or not yet published;
- 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;

- extending as much as possible the adoption of the NBN technology and the user network in Italy;
- creating some redundant mechanisms both for duplication of name-registers and in some cases also for the digital resources themselves;
- 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.

The proposed architecture (see Figure 1), introduces some elements of flexibility and additional features. 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, IT:FRD, etc.) to second-level naming authorities. Sub-domain responsibility can be further delegated using a virtually unlimited number of sub-levels (e.g.: 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 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.

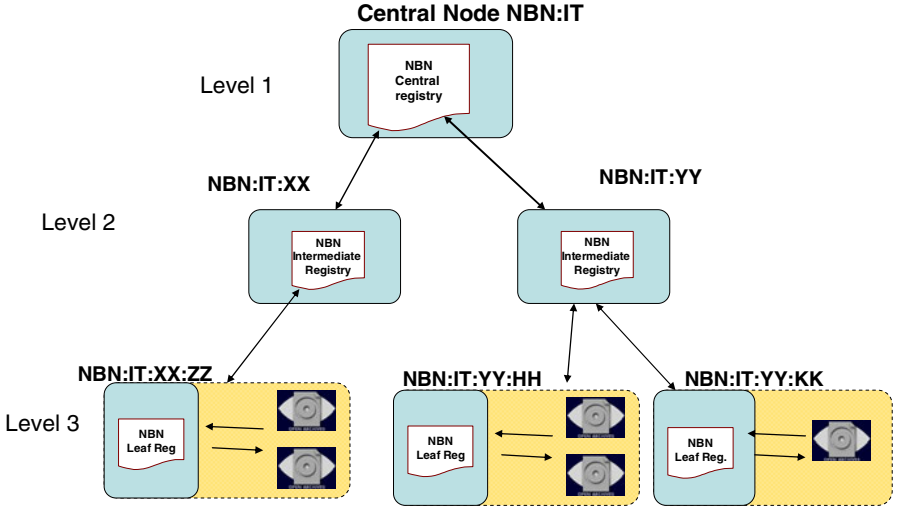

Fig. 1. The multi-level distributed architecture

This implies that every node can resolve every NBN item generated within the NBN:IT sub-namespace, 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.

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.

Guidelines

The policy should define rules for:

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

4 Magazzini Digitali/Digital Stacks

The Digital stacks project, established in 2006 by the Fondazione Rinascimento Digitale and by the Biblioteca Nazionale Centrale di Firenze, now relies on an infrastructure based on two main deposit sites (managed by the Biblioteca Nazionale Centrale di Firenze and by the Biblioteca Nazionale Centrale di Roma) and a dark archive (managed by the Biblioteca Nazionale Marciana, Venezia).

The name of the project Magazzini Digitali (Digital Stacks) intentionally recalls the term used to refer to the stacks of legal deposit libraries. In most aspects digital stacks are comparable to conventional ones: digital resources must be preserved for the long term; digital stacks grow as new resources are added; modification and deletion is not an option; it is impossible to predict the usage frequency of stored digital resources; and it is likely that some resources will be seldom or never be used.

The aim of the project is to set up an infrastructure based on a long term framework. Taking into account the fact that component failures are the norm rather than the exception, the infrastructure is based on data replication (different machines located in different sites) and on simple and widespread hardware components, non vendor-dependent, that can easily be replaced (just simple personal computers).

The infrastructure does not rely on custom or proprietary software but is based on an open source operating system and utilities (widespread acceptance means less dependencies).

The infrastructure has been developed by the Ministry of Culture in order to offer the first service for legal deposit of digital contents in Italy. The experimentation just launched started with the Doctoral thesis and some Universities already joined the test bed, obviously we put together with deposit also the PI assignment of the digital resources and so we encourage the Universities to adopt the appropriate policy and install the free software to become second level agencies to generate NBN names for their own resources.

5 Conclusions

The development of a strong policy for persistent identifiers of digital resources from both cultural and scientific communities is very important and poses a structural element for our information society future. Moreover the Italian development of a NBN register has been original and innovative in respect to what other European countries have done in this area. In parallel the Magazzini Digitali project set up a national infrastructure to for legal deposit of digital resources offering for the first time such a strategic service for user communities. The synergy between the two projects promises a serious and robust approach for long term vision of these digital resources management and use.

References

1. Hakala, J.: Using national bibliography numbers as uniform resource names. RFC3188 (2001), http://www.ietf.org/rfc/rfc3188.txt
2. Kunze, J.: The ARK Persistent Identifier Scheme. Internet Draft (2007), http://tools.ietf.org/html/draft-kunze-ark-14
3. Lagoze, C., de Sompel, H.V.: The Open Archives Initiative Protocol for Metadata Harvesting, version 2.0. Technical report, Open Archives Initative (2002), http://www.openarchives.org/OAI/openarchivesprotocol.html
4. Workshop, D.C.C.: on Persistent Identifiers. Wolfson Medical Building, University of Glasgow(June 30-July1, 2005) http://www.dcc.ac.uk/events/pi-2005/
5. ERPANET workshop Persistent Identifers, University College Cork, Cork, Ireland (Thursday June 17 - Friday June 18, 2004), http://www.erpanet.org/events/2004/cork/index.php
6. Dublin Core Metadata Initiative. Dublin Core Metadata Element Set, Version 1.1., http://dublincore.org/documents/dces/
7. Bellini, E., Cirinnà, C., Lunghi, M.: Persistent Identifiers for Cultural Heritage, Digital Preservation Europe Briefing Paper (2008), http://www.digitalpreservationeurope.eu/publications/briefs/persistent_identifiers.pdf
8. Bellini, E., Lunghi, M., Damiani, E., Fugazza, C.: Semantics-aware Resolution of Multi-part Persistent Identifiers. In: WCKS 2008 Conference (2008)
9. CENL Task Force on Persistent Identifiers, Report 2007 (2007), http://www.nlib.ee/cenl/docs/CENL_Taskforce_PI_Report_2006.pdf

10. National Library of Australia, PADI (Perserving Access to Digital Infornmation) Persistent Identifiers (2002), http://www.nla.gov.au/padi/topics/36.html#article
11. Relationship Between URNs, Handles, and PURLs Library of Congress, National Digital Library Program, http://lcweb2.loc.gov/ammem/award/docs/PURL-handle.html