Maurizio LANCIA, Brunella SEBASTIANI, Roberto PUCCINELLI, Marco SPASIANO, Massimiliano SACCONE, Luciana TRUFELLI, Emanuele BELLINI, Chiara CIRINNÀ, Maurizio LUNGHI

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, leaded 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/edlproject/
[2] Europeana 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
[13] Persistent URLhttp://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

[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/VI/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.
[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 Identifers, 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
[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