JLIS.it

Trust and persistence for internet resources

Maurizio Lunghi, Chiara Cirinnà
Emanuele Bellini

Introduction

Internet radically changed our way of working, communicating, living, producing and accessing information, interacting with institutions and bodies, buying things and managing resources. Now everything is available on an open and flexible infrastructure, often freely accessible to all the users: contents are usable by many services tailored to the user requirements. The web has probably been the killer application for the internet. In the past few years, the web moved from a web of documents towards a web of data where information is no more packaged in fixed documents but is available in a de-structured way and usable in a more flexible way by users. The recent developments on the web witnessed the emergence of the semantic web technologies and the linked open data[1] approach, associated with an increasingly large amount of data available for publishing and connecting structured data on the web. Linked data best practices, supported by W3C,[2] are now ready to be endorsed

[1] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

[2] W3C - http://www.w3c.it.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5494

by a relevant number of data providers, leading to the creation of a global data space - the web of data. Unfortunately, the LOD 5 stars[3] are mainly oriented towards the usability and standardization of data published on the web without considering the trustability and persistence of the data and the URI used to refer to them. In fact the objective of the LOD approach seems to be oriented to make a huge number of data on the web accessible in a non-proprietary format (e.g. CSV instead of Excel) and to link these data to other datasets (e.g. Genomes[4] or DBpedia[5]) to disambiguate content and to provide a context. However, in some cases, and especially in the cultural and educational domains, besides retrieving the needed data or their relations, it is also equally important to get information about their authenticity, integrity and provenance. Systems for certification using PIs for digital objects, for authors and for institutions can be of great help in order to refine the quality of information retrievable from internet and to largely increase its usability and the development of potential new services. This paradigm based on the identification and interconnection of data offers solutions to many of the actual library issues, like enhanced web searching, authority control, classification, data portability and disambiguation. In the web of documents identification and trust were provided by web sites and institutions supporting them, in the web of data they are integrated in the single piece of data. The evolution of this paradigm is increasingly important in a vision for the long term curation of the digital resources.

[3] http://www.w3.org/DesignIssues/LinkedData.html.
[4] http://www.geonames.org/ontology/documentation.html.
[5] http://en.wikipedia.org/wiki/DBpedia.

Requirements for the long term curation of digital resources

Presently the number of scientific and cultural heritage digital resources made available on the internet throughout digital library applications is constantly growing and it is now crucial to guarantee persistency, authority, reliability and wide dissemination of resources while supporting their long term curation. One of the main requirements to tackle this issue is to adopt credible and PI systems within the life cycle of these resources. A PI should be assigned only to resources that are stable, significant for the related user community and suitable with the scope of the identification system. A number of initiatives, standards, technologies are available, but it may be difficult for an institution to understand which of these are more appropriate for their digital objects. The PI technologies help make stable the reference to a digital resource, even if it is well-known that persistency isn't only a technical issue. In fact these technologies are not obviously reliable per se, no technology can exist indefinitely or guarantee services without a trustable organization and clearly defined policies. In our vision PI systems are meant as the available technology plus a trustable organization and precise policies for digital preservation, implemented by the managers of the related user community. The concept of persistence moves from the commitment of an institution/registration authority to a commitment of the entire user community served by PI. A PI system can be considered as a contract between the final users and the service-providers responsible for the implementation and maintenance of the PI-service and the functionality of the system. From this point of view, the persistence of a PI depends also on the commitment of the community that promotes and uses the identification system for their own resources. This happens when

the standard adopted is effectively oriented to the community requirements and the authority in charge to manage the system is recognized by the community itself. It is well known that the structural instability of simple URLs (e.g. domains no longer available) and related resources (relocation or updating) is one of the main issues that prevents the use of internet as a trustworthy platform for the research and the dissemination of digital contents. The current use of the simple URL approach used as persistent digital object identifier brings many and documented risks in a long term vision not only for retrieval and access of resources but also with respect to the loss of reference to the digital documents or the lack of guarantee of authority and provenance. These risks affect:

a) the cultural heritage and research domains, preventing the implementation of reliable citability services, research evaluation, digital preservation, access, etc.,

b) the business domain, preventing the use of purchase services provided on these objects,

c) the public domain (e-gov), slowing down the dematerialization process of public administrations.

It is clear that the problem is not only to face the HTTP 404 error, but it is moving towards identification systems able to support authority, reliability, preservation, certification, exploitation and wide dissemination of these resources. A trustworthy solution is to associate a trusted PI with the digital resources.

The challenge of trust

Trust, broadly speaking, concerns the assessment and management of the risks perceived by each actor entering into a relationship. In

other words, "trust entails risk". According to the ISO definition, the risk can be defined as the combination of the probability of an event and its consequences (ISO/IEC Guide 73). There are a number of events with bad consequences that could occur during the lifetime of the PI service, with different degrees of probability but all with high costs in case of failure. Examples of these risks are:

a) failing to determine the initial and recurring costs and the pricing of service (risks associated to the financial sustainability),

b) adopting technologies no longer available (risk associated to the standards adoption),

c) the object identified is no longer available on the network (risk associated to the agreement between content and service providers),

d) to lose the support of the community (risk associated to the community mandate), etc.

These factors can determinate the decrease (lowering) of trustworthiness in the PI service by the content provider and affect the dissemination and exploitation of digital resources. The various digital repositories store intangible objects and entities and make them available to users through telematics networks: we access our bank account as well the hospital or the municipality for official documents, we download tons of files and chat with avatar actors. But who certifies the identity of actors and guarantees our privacy? How can we rely on the authenticity of the documents we download? And also how can we trust the institute issuing an 'official' document? What is the risk if we cannot demonstrate that a document is not valid for our expected purposes? Which are the risks? A good amount of trustworthiness is necessary to live in this virtual and artificial world. A PI service must address at least the following core requirements:

1. global uniqueness: the PI is clearly part of a name domain and it is unique and associate to a unique resource.

2. persistence: it refers to the permanent lifetime the significant properties of an identifier, for example, it is not possible to reassign the PI to other resources or to delete it.

3. resolvability: it refers to the possibility of retrieving information regarding a resource or to access it directly on the internet.

Currently, there are different technologies and standards for the implementation of PI systems, but there isn't a general agreement on their adoption, often because some of these systems were born as technical solutions, without the support of the community of users who need specific levels of PI services. Systems like the PURL or Cool URIs(Berners-Lee) have considerable advantages in supporting the web of data implementation thanks to the their immediate dereferenceability through the protocol HTTP, but on the other hand, there are several limitations due to the fact that their persistence is not guaranteed in principle by an independent and trustable third party. It is well known that the Cool URI approach to persistence is based on the URL design. This approach, even if it is considered a best practice for the implementation of the semantic web in general and linked data in particular, is mainly based on technical solutions. The basic assumption is that a correct design of the URI should reduce the need to change them in order to ensure their stability over time. An example of this best practice is to avoid the explicit extension of web pages as .php or .asp so that changes in technology implementation do not affect the URI form (e.g. from PHP to ASP). In this perspective, the persistence is based uniquely on the commitment of individual institutions establishing a trusted relationship directly with the final users, without the mediation of a third party. Unfortunately, it is well known that the commitment

of a single institution is no longer sufficient to ensure neither long term persistence of URLs nor the trustworthiness of the resources in terms of provenance, authenticity, integrity, conservation, and so on. In practice resources are moving on the network, they can be changed or deleted due to a multitude of factors that cannot always be predetermined or regulated by the content management policies of institutions or governed by best practice techniques. A typical case occurs when an institution runs out because it has been absorbed by another institution, or it is suppressed, or simply its official name has changed. In these cases, the digital objects can be renamed to be adapted to the workflow of the new institution, or transferred to other institutions, or at worst deleted because they are no longer relevant to institutional goals. It is clear that all these actions can cause the breaking of the old URLs independently of how they were built. This may not be a problem if the institution does not handle scientific, cultural or administrative resources but it becomes a critical issue if these changes affect institutions like scientific datastore, libraries, archives, governmental dataset, and so forth. In these cases, for example, bibliographies based on simple URL or even cool URI referring to resources that were present in the archives of these institutions, can no longer be used to check the scientific work or to calculate bibliometric indexes. Another critical issue is related to the connection of datasets which have been updated several times. In such cases, it may be difficult or even impossible to verify the validity of the scientific outcome presented in a related paper. What is most critical, however, is the impossibility to implement systems to check the authenticity, provenance and integrity of these resources because of the absence of a third party able to guarantee the association name - resource. In this scenario, most benefits of a wide access to linked dataset are dissolved by the lack of their reliability.

NBN:IT service as a support of trust LOD

To tackle the challenge of trust in LOD, a possible solution could be to adopt a URN based PI solutions.[6] Presently, to implement a PI system, the main approach is to separate the identification from the localisation of the resources. As shown above, Tim Berner Lee advises that adopting clear and stable policies and implementation guidelines is sufficient to manage the persistent identification of resources on the internet. Even if this suggestion is reasonable and appropriate in some domains, it is evident that we cannot delegate this responsibility to each institution, in particular in the scientific and cultural heritage domain for two main reasons:

1. many institutions fail to decide the approach and the strategy to be adopted in terms of content selection, formats, naming, etc.;

2. many institutions fail to decide the approach and the strategy to be adopted in terms of content selection, formats, naming, etc.;

In any case, Uniforum Resource Identifiers (URIs) are widely used in the semantic web context to identify any type of resources or any real, digital, abstract, virtual object, trying to harmonise in a semantic vision all the user communities applications. For instance, to address this issue, the info-URI scheme[7] was developed by libraries and publishing communities for "URIs of information assets that have identifiers in public namespaces but have no representation within the URI allocation". It is clear that, in order to refer to a certified digital object in a trustable way, the use of URN or identifiers

[6] APARSEN DE22.1 Persistent Identifiers Interoperability Framework - http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=D22.1+Persistent+Identifiers+Interoperability+Framework.

[7] RFC 4452: http://info-uri.info.

that implements the RFC 1737 (Functional requirements for Uniform Resource Names) is today a best practice. The purpose of a URN is to provide a globally unique, persistent, location-independent resource identifier which can be used for the identification and access to the characteristics of a resource or for the access to the resource itself. The URN specification is part of the IETF family of specifications encompassed by the URI framework. This framework also includes URLs, which specify both a protocol and a location in order to give access to resources on the web. IANA is the registration authority for URN namespaces. URNs are designed to enable heterogeneous namespaces mapping onto a URN-space, and therefore enable the reuse of well-known identifiers. 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. But single implementations of namespace, like the URN-NBN or the DOI, offer a resolution-service available on internet. The NBN namespace, as a namespace identifier (NID), has been registered and adopted by the Nordic Metadata Projects but is being separately implemented by individual systems with no reference implementation which enable the coordination of information sources. In fact, several national libraries have developed their own NBN systems within national projects; several implementations are currently in use, each with different descriptive metadata or granularity levels. According to this, it is clear that the PIs, cannot support the LOD trustworthiness successfully. The NBN-Italy service supports at least three levels of persistence:(Bellini et al., "The National Bibliography Number Italia (NBN:IT) Project. A persistent identifier

supporting national legal deposit for digital resources")

1. *Persistence of the identifier NBN.* If the resource is no longer available online, the URN identifier will be maintained (e.g. as proof that at some point that resource has existed);

2. *Persistence of the association URNs and URLs.* It is a commitment that ensures that in the long term URN is resolvable (which leads at least to an address of URL type). The accessibility to the resource is not guaranteed but is assured the access to the the so-called "Tombstone" if the resource is no longer available on the network (e.g. "This ebook is no longer on the market");

3. *Persistence of the resource referenced by NBN.* Ensuring long-term existence and accessibility the resource referenced by URN. This is the level of persistence of NBN made possible thanks to the storage (statutory or voluntary) at the national libraries and authoritative description of the national bibliography.

Thanks to these levels of service, NBN-Italy names represent a clear added value if used in the LOD architectures to support the trustworthiness of the assertions (RDF triple). This proposal goes towards the integration of the LOD and PI systems, by exploiting the on-going initiatives and projects as outlined in the next paragraph.

Next steps: Den Haag Manifesto 2.0 and Florence Agenda

The forthcoming event "Cultural Heritage On Line 2012" that will be held in Florence in December 2012 aims to improve and make effective the "Den Hague Manifesto" through the union of several ongoing related initiatives, projects and stakeholders like: APARSEN

NoE,[8] Datacite,[9] EPIC,[10] and PersID[11]/URN-NBN, W3C5, Knowledge Exchange,[12] and so forth. Two of the major objectives that we are going to achieve are:

1. a review of the Den Haag manifesto and its improvement towards the 2.0 version.

2. the definition of a Florence Agenda to define a common strategy for a Trusted LOD implementation

Den Haag Manifesto 2.0

In the recent developments some initiatives are merging the open approach of the linked open data and the potentiality of the semantic web with the added value of identification, authenticity, and provenance offered by the PI systems. The Knowledge Exchange organised a seminar[13] on persistent object identifiers inviting various current practices to compare services and explore future cooperation and convergence. This seminar took place on 14-15 June 2011 at the DANS offices in The Hague and was hosted by PersID, SURF foundation and DANS. Three major players in the persistent object identifiers area, Datacite/DOI, EPIC/Handle and PersID/URN-NBN, informed each other about recent developments, shared user experiences and discussed trends and policies. In break-out sessions participants discussed the benefits and challenges in operating multiple

[8] APARSEN - http://www.alliancepermanentaccess.org.

[9] Datacite - http://www.datacite.org.

[10] European Persistent Identifier Consortium (EPIC), http://www.pidconsortium.eu.

[11] PersID- Building a persistent identifier infrastructure, http://www.persid.org.

[12] Knowledge Exchange http://www.knowledge-exchange.info.

[13] Knowledge Exchange, http://www.knowledge-exchange.info/Default.aspx?ID=440.

PI systems and the relation of PIs to linked open data communities: there was a clear interest in connecting the PI systems to the linked data standards. This led to the "Den Haag Manifesto"(DHM), which outlines a series of concrete actions to join the PID and Linked Open Data communities. FRD has participated in the working group to define opportunities for collaboration between LOD and PI systems. During the meeting a sort of "cultural gap" between the LOD and the PI community came up. The major differences concerned the concepts of identification, persistence and trustworthiness. In fact, the LOD approach is strongly oriented to the representation of the information flow on the web. In this view the resource can change over time according to the workflow of the publication. For instance, a dataset can be updated on the web several times while its URI can remain the same. With an opposite vision, the PI domains are more oriented to identify stable resources managed by systems of trusted digital repositories. During the work we tried to identify the main characteristics of the IP systems that can be imported in LOD. The results of this first assessment was the definition of a 5- point manifesto that morally committed the institutions working in the domain of PI and LOD to ascertain their possibility of integration. The points raised are:

1. A PIs can be an http URIs including content negotiation.

2. Using LOD vocabularies for diagram elements.

3. Identifying a minimum set of common elements across space identifiers in scholarly (examples are DOI kernel metadata, DataCite kernel, etc.).

4. To use 'same as' to help PI interoperability.

5. To use PIs for subjects and objects in the RDF triples.

Since then, the DHM is used as the basis for a co-ordinated approach to identifier issues across the PI and LOD communities, but starting from these points, the DHM has to be revised, specified and extended according to present trends and solutions. Moreover, it has to be supported by a shared agenda able to guide the forthcoming LOD and PI implementations, in order to have harmonized and interoperable solutions: the Florence Agenda.

A proposal for a Florence Agenda

Presently FRD is leading a specific work package (WP22) that is dealing with PIs interoperability and LOD within the APARSEN EU project. The APARSEN is a Network of Excellence of 34 institutions and is co-funded by the European Commission in order to fight the fragmentation of digital preservation of scientific records in Europe. In the first year the WP22 developed a reference model for interoperability of PI existing systems. The work started with identifying some basic user requirements for identifiers for digital objects, persons and bodies, then some criteria for trusted PI systems have been agreed. Finally an interoperability framework has been proposed where any trusted PI system can expose its data through a shared schema; the model proposes an ontology for interoperability of PI systems in line with the LOD approach. The Italian NBN initiative follows the same flow. The NBN project is leaded by the Italian legal deposit[14] consortium that has defined some criteria and guidelines to assign the PI. This defined workflow in conjunction with the commitment of the national libraries of Florence, Rome and Venice that manage such service assure the level of trust to the PI generated that, through its reuse in the LOD domain, enables the T-LOD implementation. The Florence Agenda aims to identify some

[14] www.depositolegale.it.

milestones, guidelines and criteria that can be adopted by the PI and LOD communities to cooperate to build a more reliable web of data.

References

Bazzanella, B., et al. *Persistent Identifiers Interoperability Framework – Alliance for Permanent Access to the Records of Science Network APARSEN*. 2012.

Bellini, E. and M. Lunghi. *Persistent Identifiers for cultural heritage*. Briefing paper – digitalpreservationeurope (DPE - EU project). 2007.

Bellini, E., et al. "Persistent Identifier Distributed System for Digital Libraries, Information Technology". *World library and information congress: 75th IFLA general conference and council*. Milan, Italy, 2009.

Bellini, E., et al. iPres2008 conference proceedings. 2008.

——. "Semantics-Aware Resolution of Multi-part Persistent Identifiers". *Emerging Technologies and Information Systems for the Knowledge Society*. Springer Berlin-Heidelberg, 2008. 413–422.

Bellini, E., et al. "The National Bibliography Number Italia (NBN:IT) Project. A persistent identifier supporting national legal deposit for digital resources". *JLIS* 3.1. (2012). (Cit. on p. 383).

Berners-Lee, Tim. "Cool URIs don't change". *W3C Design Issues* 1. (2009). <http://www.w3.org/Provider/Style/URI.html>. (Cit. on p. 380).

CENL. *CENL Task Force on Persistent Identifiers, Report 2007*. 2007.

——. *The National Libraries Resolver Discovery Service (RDS) - CENL Recommendation*. 2007.

Daigle, L., et al. *URN Namespace Definition Mechanisms*. 1999.

DPE. *Winer Dov, Persistent Identifiers systems in the Public Administration sector, Digitalpreservationeurope (DPE) Briefing Paper*.

Hakala, J. *Using National Bibliography Numbers as Uniform Resource Names "NBN". RFC 3188, 2001*. 2001.

ISO 3297:1986: Documentation – International standard serial numbering (ISSN). Geneva: International Organization for Standardization,

Luhmann, N. "Trust: a mechanism for the reduction of social complexity". *Trust and Power*. Wiley, 1979. 4–103.

Lynch, C., C. Preston, and R. Daniel. *Using Existing Bibliographic Identifiers as Uniform Resource Names*. 1998.

Masinter, L. and K. Sollins. *Functional Requirements for Uniform Resource Names (RFC 1737)*.

Moats, R. *URN Syntax*. 1997.

NESTOR. *Catalogue of criteria for assessing the trustworthiness of PI systems*.

NISO/ANSI Z39.56-1997 Serial Item and Contribution Identifier. Baltimore, MD: National Information Standards Organization,

NISO/ANSI Z39.9-1992 International standard serial numbering (ISSN). Baltimore, MD: National Information Standards Organization,

NISO/ANSI/ISO 2108:1992: Information and documentation - International standard book number (ISBN). Geneva: International Organization for Standardization,

Sollins, K. *Architectural Principles of Uniform Resource Name Resolution (IETF RFC 2276)*.

Wilson, Frank. *Think Paper 11: Trust and Identity in Interactive Services: Technical and Societal Challenges*.

MAURIZIO LUNGHI, Fondazione Rinascimento Digitale.
Lunghi@rinascimento-digitale.it

CHIARA CIRINNÀ, Fondazione Rinascimento Digitale.
Cirinna@rinascimento-digitale.it

EMANUELE BELLINI, Fondazione Rinascimento Digitale.
Bellini@rinascimento-digitale.it

Lunghi, M., C. Cirinnà, E. Bellini. "Trust and persistence for internet resources". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5494, p. 375–390. DOI: 10.4403/jlis.it-5494. Web.

ABSTRACT: Internet has changed our way of working, communicating, living, producing and accessing information, everything available on an open and flexible infrastructure accessible to all the users mainly free of cost. However in some cases, it's not only important to find information but also having information about its authenticity, integrity, provenance and relations with other pieces of information. Systems for certification using URN technology like the persistent identifiers for digital objects, for authors and for bodies can extremely help in order to refine the quality of information retrievable from Internet and to increase largely its usability and potential development.

KEYWORDS: NBN:IT; Persistent identifier; National Bibliography Number — Italy

---

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15