briefing paper

digital preservation europe

Identifier interoperability

Resources of interest in digital networks originate from a wide variety of sources, and may carry identifiers from different established public schemes, official standards, de facto schemes, or private cataloguing numbering. A key step in facilitating preservation, re-use and exchange of information is to enable users to re-use these identifiers (and their associated data) across different applications. Such interoperability of identifiers encompasses not only technical aspects of interoperability but consideration of the purpose and community of use of the identifiers.

Interoperability

Interoperability is the ability of independent systems to exchange meaningful information and initiate actions from each other, in order to operate together to mutual benefit. In particular, it envisages the ability for loosely-coupled independent systems to be able to collaborate and communicate. Identifiers are lexical tokens that denote things participating in these systems; a referent is the thing that is identified by an identifier.

A resource can be part of more than one domain, and can be identified by different systems, so it is necessary to guarantee interoperability between different identification systems as well as implementations based on the same namespace. Identifier interoperability is necessary for purposes such as:

- Metadata interoperability (since metadata is a relationship which somebody claims to exist between two referents);
- The creation of standard mechanisms for the expression of relationships between the referent of different standard identifiers;
- The creation of services common to more than one system, e.g. discovery of "related content" items; compiling multimedia objects, etc.

Several such use cases for identifier interoperability have been explored in both the ISO TC46SC9 identifier interoperability work and the RIDIR project.

Identifiers assigned in one context may be encountered, and may be re-used, in another place or time without consulting the assigner. Even if a resource is assumed to be part of only one domain, once it is identifiable it may be adopted independently in another domain (and possibly with undisclosed modifications). Crucially for interoperability, the context and assumptions made on assignment of an identifier may not be known to someone else encountering and using an identifier. For example, one system may take for granted that its scope is abstract work entities; another may assume only concrete realisations of an abstract work. Where the independent systems are known to each other they may agree to provide supporting information on such assumptions; but where they are not known to each other interoperability must be ensured by other measures.

Three sorts of interoperability can be distinguished:

- Syntactic interoperability. The ability of systems to process a syntax string and recognise it (and initiate actions) as an identifier even if more than one such syntax occurs in the systems.
- Semantic interoperability. The ability of systems to determine if two identifiers denote precisely the same referent; and if not, how the two referents are related.
- Community interoperability. The ability of systems to collaborate and communicate using identifiers whilst respecting any rights and restrictions on usage of data associated with those identifiers in the systems.

These three form dependent layers: community interoperability is only possible if semantic interoperability is ensured; semantic interoperability is only possible if syntactic interoperability is ensured.

Syntactic interoperability

Syntactic interoperability may be ensured if two systems follow the same technical specifications for processing an identifier string, where the scope of the likely identifiers to be encountered is reasonably predictable. In certain cases, rules may exist for directly incorporating an identifier from one scheme in the syntax of another scheme.

However interoperability may be wide ranging, making it difficult to anticipate the likely scope: identifiers may be encountered beyond web identifiers, e.g. network telecommunication and broadcasting schemes, and other globally-unique identifiers such as International Standard Book Numbers (ISBN) not originally designed for digital use. Registry schemes and assumptions such as protocol dependence must be defined or discoverable if the identifier is to be used in more than a simple catalogue listing (e.g. as in the Dublin Core scheme field "Resource Identifier", DC element syntax: DC.Identifier. There is no single registry of all identifiers. Many, but not all, identifiers of interest in a networked environment may be registered as URI schemes, but some schemes may be private or limited in their availability. Unique registry namespaces, akin to DNS domains, are part of the URN specifications (though not widely implemented). The info URI scheme was developed within the library and publishing communities (specifically, in conjunction with the development of the OpenURL standard) because of the need to specify common public namespaces as URIs (as pure identifiers: that is, to identify, not retrieve, de-reference, locate, etc.). The aim was to define URIs to reference information assets that have identifiers in public namespaces but no representation within the URI allocation – for example, LCCNs.

Some identifiers are purely abstract "denoting" tokens (names); others embody assumptions about the use to which the identifier will be put, such as resolution to retrieve, de-reference, locate, etc. Such assumptions may be deep; for example in the URI specification, it is assumed that URIs will be resolved, and the network path of the URI for resolution is implicitly DNS based: there are no real provisions to include systems that are not DNS based. Where identifiers explicitly include or implicitly assume specific protocols, proxy mechanisms (which translate one protocol to another) may need to be provided to ensure syntactic interoperability.

Semantic interoperability

Semantic interoperability deals with an obvious but difficult problem: even if two identifier strings can be syntactically processed alongside each other, how does a system know what the terms from another system mean? If A says "owner" and B says "owner", are they referring to the same thing? If A says "released" and B says "disseminated", do they mean different things? For effective interoperable management of entities:

- a unique identifier must be associated with a description of the referent entity, using a structured set of elements that provide information about that entity (that is, an identifier must be associated with some structured metadata to be interoperable); and
- the only way of unambiguously deciding if one term means the same as another, irrespective of what it is called, is by sharing a single frame of reference. A structured ontology (an explicit formal specification of how to represent the entities that are assumed to exist in some area of interest and the relationships that hold among them) with an underlying model that allows the generation of consistent new relationships, and a method of recording the agreement between the parties whose terms are included in it.

Two leading ontology initiatives that allow such comparisons in a shared frame of reference are the CIDOC conceptual reference model and the family of applications derived from the <indecs>-based semantic interoperability project (such as ONIX). These two have much in common, and some attempts are being made to investigate areas of commonality with library activities such as RDA. A joint initiative to develop a common framework for resource categorization has been launched. Ontologies are not yet in widespread use for fully automated transactions (as foreseen in the semantic web), but are in use in most serious multimedia metadata and messaging schemes to provide a basis for unambiguous, extensible, and precise definition of terms.

Community interoperability

Identifier schemes may carry rights and restrictions on usage of data associated with those identifiers. An identifier registry authority will need to consider on what basis it is able to collaborate with other schemes, or make its data public; even if this is syntactically and semantically possible there may be barriers to open interoperability. The assignment and use of a particular identifier may have obligations regarding data ownership, data quality, data maintenance, governance, and participation requirements; these restrictions may apply in both commercial and non-commercial settings.

Semantic interoperability using mapping to a common ontology framework will necessitate a bilateral agreement between two schemes to confirm the precise intent of each others identification (or if unilateral, a note that the mapping is therefore unconfirmed by one of the parties); this provides an opportunity also to consider the community obligations of such mappings.

Each identifier registry has an obligation to its community of users a to ensure that its data is accurate; it cannot therefore rely on someone else's metadata over which it has no quality control. Each identifier registry also needs to ensure that its standard is implemented through a business model: metadata has business value which provides support for registries to implement their standard; a registry cannot therefore be expected to hand over metadata to someone that it has no business relationship with.

References

DPE Briefing paper "Persistent Identifiers for Cultural Heritage"
http://www.digitalpreservationeurope.eu/publications/briefs/persistent_identifiers.pdf

RIDIR project (Resourcing IDentifier Interoperability for Repositories)
http://www.hull.ac.uk/ridir/

<indecs> project (Interoperability of data in e-commerce systems). Metadata Framework: Principles, model and data dictionary.
http://www.doi.org/topics/indecs/indecs_framework_2000.pdf

Identifier Interoperability: A Report on Two Recent ISO Activities.
D-Lib magazine, April 2006
http://www.dlib.org/dlib/april06/paskin/04paskin.html

The RDA/ONIX Framework for Resource Categorization.
D-Lib magazine, Jan/Feb 2007
http://www.dlib.org/dlib/january07/dunsire/01dunsire.html

CIDOC Conceptual Reference Model
http://cidoc.ics.forth.gr/index.html

“Info URI” registration scheme
http://info-uri.info/registry/docs/misc/faq.html

Digital Object Architecture
Kahn R.E. & Wilensky R. "A Framework for Distributed Digital Object Services".
http://www.cnri.reston.va.us/cstr/arch/k-w.html

If both agree, a bilateral agreement can be drawn up which specifies the nature of the collaboration (for example, the appropriate registration authorities may agree to share or compare the values and updating processes for accompanying metadata). If the two do not agree, there cannot be an obligation of interoperability. Identifier schemes should encourage such collaboration by providing clear guidance on rights and obligations; these are often requirements of formal standardisation processes.

Persistence and interoperability

Persistence is closely related to interoperability: persistence is “interoperability with the future”, i.e. the independent systems able to exchange meaningful information and initiate actions from each other are separated by time.

The DPE Briefing Paper on “Persistent Identifiers for Cultural Heritage” explores the requirement and implications for persistence in identifiers in more depth. Some identifier schemes may be established for particular valid but relatively short-term needs(e.g. streaming subscription video on a social network); others focus on persistence and preservation, with a concomitant commitment to maintaining registry schemes, and metadata.

An application designer will need to consider the benefits of basing an application on particular schemes, and avoid where possible schemes where the design is not in accordance with his own fundamental aims. URLs are often cited as the most common “identifiers” (e.g. DC: Identifier definition: “String or number used to uniquely identify the resource. The default is the URL to the resource”), although in fact a URL is an identifier of a location: as a consequence of the most common model of single redirection to one URL, the two are easily confused and the link between identifier and referent is not direct, and so easily broken. URLs have low barrier to entry and use, but low expectation of persistence. Identifier schemes using mechanisms to supplement the URL process (PURL, ARK, N2T) or avoid it (Handle, DOI) are preferable.

A mechanism which is specifically designed to support interoperability has been developed: in the Kahn/Wilensky Digital Object Architecture, referents (as “digital objects”) carry with them metadata and links to repositories. These digital objects do not replace existing formats and data structures, but instead provide a common framework for encapsulating those formats and structures, allowing them to be uniformly interpreted and thus moveable in and out of various heterogeneous information systems and across changes in systems over time. There are a few implementations of this but it has not yet been widely adopted (though the Handle identifier, part of this architecture, is itself widely used).

Lessons for implementers

- Avoid re-inventing the wheel: if it appears that you need to devise a new identifier scheme, examine whether the problem can be avoided by re-using existing identifiers.
- If a new scheme is needed, consider if an existing protocol or identifier registry can be harnessed to implement your scheme.
- Register your scheme with an appropriate public namespace declaration.
- Provide easy links for semantic mapping by specifying a well-formed metadata scheme and publishing it.
- Consider the community and business implications for others who may need to use your scheme.
- Provide clear guidance on rights and obligations of use of your scheme.
- Adopt identifiers with a mechanism for ensuring persistence

Norman Paskin, *Tertius Ltd*
n.paskin@tertius.ltd.uk