LCC Identifiers workstream
Identifiers specification

Version 0.2 | November 2012

These are the identifier specifications recommended for LCC. Detailed support for the recommendations is provided in the attached appendix.

1. **Identification is essential**
   - Each entity which needs to be recognised distinctly in the network should have at least one public identifier.
   - Types of entities here include those identified in the LCC Rights Reference Model: *Party, Creation, Place, Context, Right, RightsAssignment, Assertion, RightsConflict*.
   - In particular, this includes each item of **content** which needs to be recognised (at whatever level of granularity is required), and each **person or organization** who is recognised as (or claims to be) a contributor or rightsholder of content (an "interested party").
   - A public identifier is not necessarily humanly readable: "public" means that it is accessible to people or machines within the network.

2. **Identifiers should be resolvable**
   - A **resolvable** identifier in the digital network is one that enables a system to locate the identified resource, or some metadata or service related to it, elsewhere in the network.
   - Because the World Wide Web is the dominant network using the Internet, then it is a minimum requirement to support that, and a potential requirement to support other networks. This in effect recognises the URI as a practical common framework for global digital content identification. First class identifiers may still be used where appropriate and expressed as URIs where necessary. The URI syntax can incorporate existing standard or proprietary identifiers while remaining globally unique, and much technology already exists for recognising and resolving URIs in various ways. Resolution is essential, and on their own many existing ID standards (being pre-digital in origin, such as ISBN, ISRC and ISWC) don't natively support this but require a URI "prefix".

3. **Identifiers should be capable of multiple resolution**
   - An identifier should be capable of being resolved to more than one location for different types or instances of metadata, typically to find least one basic description and one statement of rights.
   - Multiple resolution of an identifier should be possible without special knowledge except for the ability to communicate using standard technical protocols.
   - Multiple resolution therefore requires some basic and extensible standard "typing" of resolution so that different services (in this case, different metadata types) can be automatically located. This approach is common and usually implicit within proprietary closed systems but is not yet generally recognised as an inevitable requirement of open linked data.

4. **Identifiers should be accessible**
   - Content identifiers should be embedded within the item of content or its message sidecar during interchange; or in metadata on webpages to support resolution to various services.
   - Either or both approaches are useful for different purposes.

5. **Identifier registration should be under well-defined registry operations and policies to ensure persistence.**
   - Adequate supporting descriptive and rights metadata should be declared under some method of governance (a registry or registration procedure) which ensures its ongoing maintenance and authority, in locations to which the identifier may resolve, using defined service types.
   - Accountability for persistence can only be ensured through a governed registry arrangement, where there are also provisions for maintaining metadata after the original Asserter is defunct, dead or otherwise unwilling to accept responsibility. This does not necessarily mean a central repository of metadata, but it requires a registration procedure supporting identification. Mechanisms are needed to minimise instances of several Parties issuing identifiers for the same content, where creation or original publication is shared. Mechanisms are also required for dealing with duplication (the issue of more than one identifier to different entities).

6. **Metadata associated with an identified entity should be published in standard form**
   - Metadata associated with an identified entity should be published in extensible and interoperable syntactic formats (such as RDF, JSON or XML) using formalised schemas with defined elements and using controlled vocabularies wherever possible.
   - There will always be a diverse range of metadata schemas for different sectors and functions. LCC provides for the semantic mapping of any well-formed schema to another, but cannot compensate for poorly-defined or ambiguous source data.

7. **The asserter of rights metadata should be identified.**
   - Authoritative rights metadata associated with an identifier should be formally “asserted” so that its provenance is clear. The "asserter” is not necessarily the same Party as the metadata source: an intermediary such as a collecting society or licensee may declare or pass on metadata for which the authority lies with a Party further up the supply chain, such as an author or publisher. An intermediary may therefore legitimately publish conflicting metadata from different asserters: this is quite common for example in collecting societies.

All of the above are qualified by the statement that this is an idealised strategic goal. The LCC project is charged with the task of describing how well or badly current activities measure up to this, and specifying where work needs to be done.

The following is also recommended as best practice (not a mandatory requirement):

- **Dynamic attributes of the identified entity should not be embedded into the identifier string itself.**
  This does not imply avoiding structure in the identifier string (for example, assigning a set of prefixes to one agent, which may then create its own unique namespace by further qualification).

The following are related recommendations for additional follow-up by the LCC project or its successor:

- **The option of supporting or developing a revived "info URI" registration scheme should be investigated by the LCC project.**

- **The Vocabulary Mapping Framework (VMF) should be used for mapping in any following LCC implementation / demonstration projects.**

- **The Vocabulary Mapping Framework (VMF) should be mirrored or expanded to cover needs for similar mappings for other entities, and attention be given to active maintenance and a governance structure of VMF.**

Appendix to LCC Identifiers specification

Contents

1 Introduction

1.1 LCC Identifier workstream

The LCC Project Plan defined (in section 6.1) the Technical Deliverables of the identifiers workstream as follows:

"Unique and persistent identification – of content, parties (people and organizations) and of rights, agreements and other rights related entities – inevitably lies at the heart of any scheme for the management of data in the rights data supply chain. If a piece of content is to be securely linked to the service where rights can be cleared, then both the content itself and, the service and the parties involved need to be properly and uniquely identified. There are many existing identifiers for content and parties, not least those that are governed by ISO standards, but not all of them yet meet the requirements for implementation in a rights infrastructure of the type envisaged. There are also known gaps in the current identification infrastructure for all rights-related entities.

Issues to be taken into account include:

- access to identifiers and associated disambiguation metadata
- granularity of identification
- uniqueness mechanisms and simplicity of application
- persistence and stability
- portability to different contexts
- resolvability
- governance

Deliverable: A set of requirements for identification to provide a uniform approach to accessing rights data (for both people and machines), with an analysis of the extent to which existing standard identifiers meet these requirements and of where gaps exist, and recommendations on how best those gaps might be filled. "

1.2 Approach adopted

The task has two top level questions:
(1) What are the entities in the LCC Rights Reference Model (RRM) which require identification?
(2) What is the functionality required of these identifiers for use in the Reference Model?

We have adopted a combination of two approaches:

- *top-down* underlying principles of identification. (**Section 2**) Generic design of identifiers appropriate for use in LCC, based on identifier principles and stated RRM requirements;
- *bottom-up* survey of currently available identifier implementations (**Section 3**).

In each case we do not need to re-invent the wheel and can use, or build on, some earlier proven analyses and surveys; there is a large body of material available on identification; our focus is on identifiers relating to content which may be used on digital networks.

Of these two approaches, the top-down functionality considerations (where the key issues appear to be *interoperability* and *resolvability*) are the more important part, and the bottom-up collation of existing identifier schemes plays a supporting role, since while specific entries can always be added or removed from a catalogue of recommended or widely used existing schemes without affecting parts of the catalogue, changes in or violations of fundamental principles will affect the whole.

Following this analysis, we present as **section 4** "The Multimedia Identifier Network", a summary of the integrated "big picture" for identifiers in the proposed environment of digital rights exchanges.

1.3 Terminology

Unless otherwise stated or the context clearly implies otherwise, in this document we use "identifier" to mean a system of syntax specification, with an active registry plus governance and operations procedures, for a set of referents.

Other uses of the term are common: argument about definitions can lead to the worst sort of scholasticism and in particular we avoid essentialism (the view that there is one inherently correct definition for a term). However, it is both valid and necessary to understand what different people mean by use of apparently the same term "identifier", since it is widely used and may be ambiguous. It is recommended to qualify the term to indicate the sort of usage that is intended where ambiguity may otherwise occur. It can be used to mean for example:

- a registry (a database of some existing identifiers with referents of some commonality of type or purpose)
- a namespace (a logical and extensible group of referents of some commonality of type or purpose; which may or may not have one or more registries (e.g. URIs))
- a syntax specification (e.g. ISO specification of how an ISBN is constructed as a 13 digit string)
- a specific string ("the ISBN for the following book…")
- a framework for generating subordinate namespaces (as in URN, URL, DOI..)
- a system (a specification plus governance and procedures, plus active registry; as in ISBN, DOI, etc)
- implicit binding – e.g. through DNS etc. – with no explicit referent specification (e.g. URL)

1.4 Technical and social infrastructure

As far as possible the "set of requirements for identification to provide a uniform approach to accessing rights data" has been cast as technology-neutral. There is, however, one exception since it is necessary to assume some level of implementation: an LCC-conformant identifier should be Internet Protocol compatible, as the digital network to which LCC applies substantially relies upon is the Internet[1]. We have avoided recommendations at a higher technology layer – for example, http content negotiation on the web - so as to provide recommendations which can accommodate changes to adjacent "layers" and be useful for multiple access streams (web, mobile, XML, etc.).

Issues of identifier persistence and registry management are largely matters of social infrastructure, though assisted by technology tools. We make only basic recommendations on this layer. If the LCC

[1] Internet" refers to the global information system that --
(i) is logically linked together by a globally unique address space based on the Internet Protocol (IP) or its subsequent extensions/follow-ons;
(ii) is able to support communications using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite or its subsequent extensions/follow-ons, and/or other IP-compatible protocols; and
(iii) provides, uses or makes accessible, either publicly or privately, high level services layered on the communications and related infrastructure described herein." (http://www.cnri.reston.va.us/what_is_internet.html#xv). What Is The Internet (And What Makes It Work) - December, 1999: Robert E. Kahn and Vinton G. Cerf

project requires these areas to be expanded we suggest that there be iteration with the Governance workstream to produce consistent recommendations.

1.5 Sources and logistics

This document includes substantial portions of text re-purposed or edited with permission from contributions developed as part of the LCC or related projects by DOI, CNRI and others. References to the LCC Rights Reference Model ("RRM") are to draft version 0.2, September 25 2012.

An initial draft v 0.1 was issued to identifier workstream members and other technical workstream leads, plus selected external experts, for comments. Version 0.2 is issued for the LCC plenary on Nov 13th 2012. A final document for sign-off alongside the final RRM will be made available at the end of the LCC project, anticipated to be January 2013.

2. Underlying Principles of Identification

2.1 indecs

As our starting point we adopt the principles of identification delineated in the indecs ("interoperability of data in e-commerce systems") project, part funded by the European Community Info 2000 initiative and by several organisations representing the music, rights, text publishing, authors, library and other sectors in 1998-2000, which has since been used in a number of metadata activities. The indecs Metadata Framework document "Principles, model and data dictionary"[2] is a summary. indecs provided an early analysis of the requirements for metadata for e-commerce of content (intellectual property) in the network environment, focussing on semantic interoperability. It built on a simple generic model of commerce (the "model of making") which shares its underpinnings in the contextual approach of the RRM. This foundation work has been developed and built on over the last decade in several content industry specifications which are aligned with the LCC approach including:

- RDA/ONIX Framework for Resource Categorization;
- Vocabulary Mapping Framework for major bibliographic and cultural heritage standards;
- DDEX (Digital Data Exchange) music industry messaging and data dictionary applications;
- ONIX (Online Information Exchange) standards for the use of publishers in distributing digital metadata about their products;
- Digital Object Identifier System metadata schemes;
- ISO/IEC 21000-6 (MPEG) Rights Data Dictionary (RDD);

The approach also has much in common with, and can be mapped consistently to, the CIDOC Conceptual Reference Model (CRM), an ontology for cultural heritage information, and the Functional Requirements for Bibliographic Records (FRBR) model in the library world.

We have not found any other underlying statements of principle which are not already encompassed in indecs or which meet the requirements of the LCC. Other proposals we have reviewed include:

- ISO TR 21449[3]: now outdated and does not add anything to the LCC analysis;

[2] The <indecs> metadata framework Version 2.0, June 2000: G. Rust & M. Bide.
http://www.doi.org/topics/indecs/indecs_framework_2000.pdf

[3] ISO/TR 21449, Content Delivery and Rights Management — Functional requirements for identifiers and descriptors for use in the music, film, video, sound recording and publishing industries

- URN Functional requirements (also now outdated and being reviewed in the light of developments since their original inception[4]);
- URI principles (see below under "Resolution") (also under potential review).

indecs proposed four principles as key to the management of identification:

- *The principle of Unique Identification:* every entity should be uniquely identified within an identified namespace.
- *The principle of Functional Granularity:* it should be possible to identify an entity whenever it needs to be distinguished
- *The principle of Designated Authority:* the author of an item of metadata should be securely identified.
- *The principle of Appropriate Access:* everyone requires access to the metadata on which they depend, and privacy and confidentiality for their own metadata from those who are not dependent on it.

indecs also produced a useful *definition of metadata*:

- *An item of metadata* is a relationship that someone claims to exist between two referents (entities).

The indecs framework stresses the significance of relationships, which lie at the heart of the indecs analysis and of the LCC aims. It underlines the importance of unique identification of all entities (since otherwise expressing relationships between them is of little practical utility). Finally, it raises the question of authority: the identification of the person making the claim is as significant as the identification of any other entity. indecs was therefore a significant step in recognising the major improvements needed in what we call below (section 4) the Multimedia Identifier Network which are essential for the success of rights information exchange.

2.2 Intelligence and identifier structure

Many disciplines over the years have learned that embedding attributes of the identified entity into the identifier string itself can produce a fragile identifier, subject to malfunctioning and misunderstanding, when those attributes can change over time. Location is the simplest example of this: any identifier that is location based will break whenever the identified entity is moved. Ownership is another, making it impossible for an entity to change hands and still keep its identifier. Keeping that sort of intelligence out of an identifier is good design and we recommend as *best practice* (rather than a requirement) for identifiers used in LCC the following:

- *The principle of unintelligent numbering*: Do not embed dynamic attributes of the identified entity into the identifier string itself.

Identifiers with no embedded attributes derived from, or dependent on, another entity are also sometimes called first class identifiers.

Note that this is not the same as defining an identifier specification "that contains no embedded intelligence". To fully understand this issue requires teasing apart a 'no intelligence' statement, as it depends on what intelligence is meant. Using structure in the creation of an identifier (for example, assigning a set of prefixes to one agent, which may then create its own unique namespace by further

[4] http://datatracker.ietf.org/wg/urnbis/charter/

qualification – e.g. the ISBN system) or in the resolution of an identifier on a network (for example, assigning an internet protocol such as http: to precede the identifier string) is not equivalent to embedding attributes of the identified entity into the identifier. Changes to the entity do not impact this more mechanical structure used to add flexibility to identifier creation and resolution. When identifiers are assigned on a federated model, it appears essential to include this level of structure (whilst still being able to avoid embedding attributes of the identified entity into the identifier string itself): all the existing federated global identification standards of which we are aware (e.g. DOI, GS1, Media Access Control [MAC address], Internet Protocol [IP and IPv6], EIDR) use a structured solution in order to allow maximum possible flexibility of local members in the allocation of identities. We have not, in contrast, been able to find any examples of unstructured global federated identifiers. It appears that successful federated global identification standards have found it necessary to adopt a structured numbering assignment system[5].

For completion, we note that Affordability (the ability to generate an identifier from content-in-hand: "a situation where an object's characteristics imply its functionality and use"[6]) in identifiers does not necessarily generate the same fragility as embedded intelligence: the object can have its identifier created (or recreated) from its invariant properties. However this is applicable only to unique physical objects (or unique digital objects in the form of hash signatures[7]) and unlikely to apply to LCC applications (most of which will rely on identification of abstractions such as concepts, classes, etc.).

2.3 Persistence

For long term interoperability identifiers must be persistent by at least return returning a "tombstone" message such as "this identifier refers to X which has since been removed due to Y" (cf. ISBN for out of print book titles). Identifiers used in the LCC should have well defined and public registry operations and policies likely to ensure persistence.

Persistence is the consistent availability over time (persistence has been called "interoperability with the future"), of useful information about a specified entity: it is ultimately guaranteed by social infrastructure (through policy) and assisted by technology. The aim should be to not shoot oneself in the foot by adopting inappropriate technology choices which will then restrict the best possible social infrastructure to maximise persistence: the principle of unintelligent numbering is clearly one such technical step. Other steps include managed metadata and indirection through resolution which allows reference to an entity to be maintained in the face of legitimate, desirable, and unavoidable changes in associated data such as organization names, domain names, URLs, etc. ; and governance steps to facilitate persistence in the event of registry demise (e.g. by orderly transfer of records).

The key social infrastructure necessary to ensure persistence is a registry, together with clear policies and procedures for how identifiers in the registry are assigned and managed. Further long term persistence requires continuity planning: governance consideration of the future of the registry in the event of the registration authority being unable or unwilling to continue. In the case of ISO

[5] A recent joint Engineering study for the proposed FSB Legal Entity Identifier (http://www.financialstabilityboard.org/publications/r_121024.pdf) by Loughborough University, Tahoe Blue, FIX, GS1, CNRI, and IDF examines the issue of structure in identifier assignment and its role in federation mechanisms in great detail. (the detailed study is currently restricted in circulation).

[6] http://www.usabilityfirst.com/glossary/term_66.txl

[7] e.g. the proposed URI scheme "Naming Things with Hashes" http://tools.ietf.org/html/draft-farrell-decade-ni-10

identifiers, a generic ISO Registration Authority agreement has recently[8] been substantially revised with input from some LCC participants (notably those of ISO TC546/SC9) and provides a minimum set of requirements providing some reassurance to the user community that assigned identifiers will be maintained. However these minimum requirements are not sufficient to plan a full implementation. The most widespread persistent identifier in use by the content sector, the DOI System, has developed a series of persistence requirements (together with governance and operational policies) which may serve as a suggested model for other registry authorities seeking to provide the same level of continuity[9].

2.4 Internet use of identifiers

2.4.1 Resolution

By resolution we mean going from the identifier to information about the identified entity or to the entity itself, using a network-based system. This is sometimes called de-referencing[10]. In current practice, the main focus of LCC work is currently on the WWW layer built on the underlying internet. That in turn uses the http (hypertext transfer protocol) and related developments, running on top of the Domain Name System (DNS) layer for resolution. DNS was never intended to be a persistent identifier system, and it has some fundamental issues relating to scalability, persistence and security when used for that, so it would be inadvisable for LCC to ignore protocols other than http, which may become increasingly important through mobile devices, etc.

We specify URI as a general concept as an identifier common format in which identifiers should be expressible as a pragmatic choice; http URIs are predominant on the Web. However some areas of content linkage may rely on http more than others: for example, Skype, Facetime, e mail, most instant messaging, etc. are non-http. Most mobile apps probably use http to exchange data (but there is really no easy way to tell, since the app hides everything). It is likely that most apps that display information that could be on a web page are using http (since much of the composition and display engine is already done as a combination of http and html). We can further distinguish between native apps and mobile web: a user can download a specific app (for e.g. an iPad) or can take any given web page and make an icon of it: they both look like apps on the screen but the web page needs connectivity and can only do whatever the web stuff can do; by contrast the 'native' app can anything it is programmed to do (though budgets may dictate a specific path for content providers who have to consider Apple, Android in many varieties, Microsoft, etc.). In theory the mobile web in HTML5 will be “write once run everywhere” but so far the native apps have the advantage and the lead; they can access things like the camera and other apps. Security is also an issue: in general security is easier to manage with a dedicated app rather than relying on whatever the web browser and web site give you.

---

[8] 2011. ISO state that “the generic RAA template is available upon request. There are 67 RAAs among ISO’s 19000 standard so this is not something that we put on our website. ISO Committees should only establish RAs for exceptional cases” (source: ISO Central Secretariat, 25 Oct 2012)

[9] www.doi.org: see in particular http://www.doi.org/doi_handbook/6_Policies.html#6.5

[10] We note also that the term “resolution” is used in some areas (but not in LCC) to denote what we would call disambiguation: e.g. OYSTER (Open sYSTem Entity Resolution: http://sourceforge.net/p/oysterer/home/Home/) “is an entity resolution system that supports probabilistic direct matching, transitive linking, and asserted linking”; the term “resolution” here (resolving conflicting data records) is not the same as “resolution” as used in network de-referencing. Both disambiguation (ensuring that we identify each unique entity, and associate a record for each identified entity) and network resolution (deploying the unique identifiers to look up the current state of the record) are necessary parts of an identification system; but need to be distinguished.

A technical definition is in IETF RFC 3404: identifier resolution is "a process by which an identifier string is employed to access its associated object and/or descriptive information about the object (metadata). This usually involves one or more intermediate mapping operations". More usefully, resolution is the process in which an identifier is the input — a request — to a network service to receive in return a specific output of one or more pieces of *current information* (state data) related to the identified entity (e.g., a location URL): that is, the associated state data may be dynamic (change over time) yet still be associated with the identifier. Multiple resolution (as in the Handle System[11]) is the return as output of several pieces of current information related to an identified entity: specifically at least one URL plus defined data structures. These may be configured so as to return only the most appropriate value for the given context[12], and thus multiple resolution is one option for facilitating contextual management of identifiers.

Note that we distinguish the referent (the thing that is identified by an identifier) from the result of a resolution request: resolution may return the referent (or more likely an instance or representation of it as a digital object), but more often will return some data about the referent.

2.4.2 URI

Uniform Resource Identifier (IETF RFC 3986) provides an extensible means for identifying a resource within the World Wide Web. Each URI begins with a scheme name that refers to a specification for assigning identifiers within that scheme; each scheme's specification may further restrict the syntax and semantics of identifiers using that scheme. The commonly seen http:URI is only one such scheme among some 75 defined (and a further 100 or so "provisional") URI assignments[13] forming a broad church of mainly technical protocols (mailto, ftp, telnet, file etc.) with little relevance to linking of content, with a few exceptions.

The URI specification defines (1) an implementation to access a location on a file server, commonly accessed using the http protocol though other protocols are allowed; (2) a syntax for referencing, through which e.g., ISBNs can be specified as URIs. The network path of the URI is implicitly DNS based; original URI specifications that assume the URI to be opaque have been overtaken by practical usage which assumes that the initial URI parser will look for meaningful characters (such as dot and slash).

The use of URIs as identifiers that don't actually identify network resources (for example, they identify an abstract object, or a physical object) was recognised as an unanswered problem in RFC 3305. This usage is important in any semantic application. To address this, the info URI scheme[14] (see further discussion 2.4.6 below) was developed by library and publishing communities for "URIs of information assets that have identifiers in public namespaces but have no representation within the URI allocation". OpenURL[15] adopts it and was a key the motivation for it. InfoURI registrations can be made by anyone, not necessarily the authority for a particular namespace.

[11] www.handle.net
[12] For an example using DOI, see http://www.doi.org/doi_handbook/5_Applications.html
[13] http://www.iana.org/assignments/uri-schemes.html
[14] IETF RFC 4452: http://info-uri.info
[15] OpenURL is a mechanism for transporting metadata and identifiers describing a content item (typically a text publication) for the purpose of context-sensitive linking through a local link resolver.

URIs may be used as "abstract" URIs (under the namespace “tag:” as an example[16]) for semantic web uses (RDF, some ontologies); therefore it is possible for any identifier to be cast as a URI, though whether this is useful will depend upon context of use.

2.4.3 URI in relation to URL and URN

There is commonly some confusion and misunderstanding about the term URI and related terms, which is entirely understandable given the historical ambiguity and confusion in their use. RFC 3986 (2005) aimed to end this by stating that a URI can be classified as a locator, a name, or both. In this view, the term URL refers to the subset of URIs that, in addition to identifying a resource, provide a means of locating the resource; the term URN has been used historically to refer to both URIs under the "urn" scheme (RFC 2141) which are required to remain globally unique and persistent even when the resource ceases to exist or becomes unavailable, and to any other URI with the properties of a name. RFC 3986 requires that the terms URL and URN be deprecated. This brings a uniformity to the technical treatment of all URIs; however the risk of confusion remains, from:

- cited documents which rely on earlier, now superseded, statements of the position;
- the use of one simple top level term (URI) may hide useful distinctions which some users, e.g., librarians, may wish to make between a unique name and a location, for example when a named resource is available at multiple locations;
- considerations of how widely used non-web identifiers (such as ISBNs, RFIDs, social security numbers, etc) relate to URIs, which can lead to:
- confusions re identifier, representation, and access mechanism;
- lack of appreciation of identifier usage outside the WWW;
- use for non-digital referents; and
- the requirement to perceive the web as only part of the Internet and the Internet as only part of information.

In the view now considered by RFC 3986 to be obsolete, URIs have two subclasses: URN (identifying names) and URL (identifying single locations). In the RFC 3986 view, web-identifier schemes are all URI schemes, as a given URI scheme may define subspaces; some of these may be access mechanisms (e.g., "http:") whilst others may be namespaces (e.g., "urn:").

W3C state: “The vulnerability of any digital material to unexpected or unintended changes in Internet domain name assignment, and hence to the outcome of domain name resolution, is widely recognised. The fact that domain names are not permanently assigned is regularly cited as one of the main reasons why http:URIs cannot be regarded as persistent identifiers over the long term”. [17]

2.4.4 Possible revision of URI specification

A recent post[18] to the W3C URI list by Larry Masinter (a long–term member of the W3C Technical Architecture Group and one of the co-authors of the URI syntax RFC 3986) proposed creating a new RFC that “obsoletes 3986 (URI) with a document that combined it with 3987 (IRI, Internationalized Resource Identifier, a generalization of URI allowing the use of Unicode), reverts to the "URL" name, and gave updated parsing advice”; he also posits the possibility of “removing any basis for support of using http URLs to "mean" abstractions or people”, on the grounds that there is confusion over “whether *http://larry.masinter.net#the_person* could identify, locate, or name me rather than a paragraph of my home page”; and “including URN”. It seems that the confusion between a referent and what an item resolves to is still not sufficiently appreciated. Any such URI re-definition is unlikely

[16] IETF RFC 4151: http://www.rfc-editor.org/rfc/rfc4151.txt

[17] Domain names and persistence: Report on a W3C workshop: Henry S. Thompson, Jonathan Rees, January 2012: http://www.w3.org/2001/tag/2011/12/dnap-workshop/report.html

[18] Nov 2, 2012: http://lists.w3.org/Archives/Public/uri/2012Nov/0000.html

to happen in the near future; such a move would appear to be a significant change in the development of W3C's approach to URL.

2.4.5. URN

Uniform Resource Name (RFC 2141, 1997) is a specification for defining names (identifiers) of resources for use on the Internet. In this RFC locations are assumed to be independent of names. URN resolution is still an active topic of discussion, and has active use, especially in the library community (e.g. for treatment of National Bibliography Numbers as URN in RFC 3188). RFC 2141 defines (1) a formal registration process as a urn namespace, and (2) accompanying specifications to implement a series of functional requirements for such namespaces. Existing identifiers may thereby be specified as a URN: e.g. an ISBN as *urn:isbn:9789521061547*; such identifiers may be implemented using a specially written URN plug-in and resolved to URLs: functionally this gives nothing beyond that achieved by coherent management of the corresponding URLs.

Currently URN is under review: an IETF Working Group, "Uniform Resource Names, Revised", has undertaken the task of reworking and updating the key URN RFCs (the so-called "URN-bis" process"), including RFC 2141, which date from 1997-2001, to reflect the URN implementation experience gained since that time. Proposed changes include updating the syntax specification, a formal IANA registration for the 'urn' URI scheme, revised URN examples, and updated descriptions of how URNs are resolved based on current practices. The outcome of this revisiting of the URN scheme is currently awaited[19].

URN architecture assumes a DNS-based Resolution Discovery Service (RDS) to find the service appropriate to the given URN scheme. However no such widely deployed RDS schemes currently exist: browsers cannot action URN strings without some additional programming in the form of a "plug-in". These carry no guarantee of ready interoperability with other deployments, which may require a different plug-in for each implementation and may use conflicting data approaches. Therefore most existing URN implementations embed the URN as a http URI which contains the URL of the relevant resolution service (e.g. for the URN form of the ISBN shown above, resolved via the Finnish national URN service http://urn.fi, the actionable form of the URN is http://urn.fi/URN:ISBN:978-952-10-6154-7). There is no global service aware of national and/or regional URN resolution services, but there are some proposals to provide one (e.g. http://www.persid.org).

The set of URNs, of the form "urn:nid:nnnnnn", is a URN namespace. ("nid" is here a URN namespace identifier, neither a "URN scheme", nor a "URI scheme.") The official IANA list of registered NIDs at http://www.iana.org/assignments/urn-namespaces lists 40 registered NIDs; however many of these are not widely used as URNs, including some content identifiers (e.g., ISSN, ISBN).

URN registration currently requires an additional layer of administration for defining a URN namespace (e.g. the string urn:doi:10.1000/1 rather than the simpler doi:10.1000/1) and redirection to access the resolution service,

2.4.6 Info URI

The "info" URI initiative was launched in 2003 *"to fill a requirement for using identifiers on the Web that derived from public namespaces but that had no canonical URL form".* Info URI was originated in

[19] Latest drafts, including a reworking of the specifications for ISBN and NBN as URN, were published in October 2012 at http://datatracker.ietf.org/wg/urnbis/

2003 by NISO[20] and became IETF RFC 4452[21]. According to that RFC "3.3. Maintenance of the "info" Registry: The public namespaces that may be registered in the "info" Registry will be those of interest to the communities served by NISO, and therefore NISO is committed to act as Maintenance Authority for the "info" Registry and to assign a Registry Operator to operate it."

In May 2010, the "info" URI Registry (info-uri.info/) posted this notice: "When work on the "info" URI scheme began, the W3C 'Architecture of the World Wide Web' (2004) had yet to be published, and the currently emerging framework for Linked Data was scarcely in its infancy. *Using the HTTP protocol for both access and persistent identity can be seen to be problematic in certain respects, although it has the undeniable virtue of requiring no additional registration infrastructure*[22]. Also, the need to guide and validate registrations of "info" URI namespaces created an approval process bottleneck that is inimical to the rapid and flexible progress that is seen to be the hallmark of the Web. The Linked Data idiom is currently ascendant, and accommodates both resource resolution and identification, which is different than the simple "info" premise of URI identification alone. This approach to resource identity is likely to conform more closely to evolving practice. For these reasons, it has been deemed appropriate to close the registry to further "info" namespace registrations. The "info" registry will continue to be supported for the foreseeable future, although prudent adopters should consider migrating their resource identity requirements towards mainstream Web practices over the long term."

The registrants of info URI weren't consulted about that statement; not all would agree with it, since "Linked Data" alone is not sufficient to establish a trustworthy industry-standard data exchange. A significant advantage of applying Linked Data principles and technologies to identifier-registered material is that it is 'data worth linking to': it is curated, value-added, data, which is managed, corrected, updated and consistently maintained by registration authorities and agencies. It is also ideally persistent, so avoiding 'bit-rot'. In practice, the quality of Linked data implementations is only as good as the data you are linking to, and the meaning and contextualisation of the link you use. The LCC system should enable "curated data", i.e. consistent, managed, linking so you can link to other "quality data" with confidence, while still using the standard Linked Data technologies. There are still many first class identifiers (ISBN, DOI, ISRC, social security numbers, etc.) which might need to be referenced by internet applications (first class in this case also means independent of any protocols used to resolve it). A list of registered info schemes[23] contains several well-known ones: the info scheme allows them to remain as first class identifiers, whereas expressing them in a http URL enforces fragility through use of the domain name system. It seems a pity that all these existing schemes have lost the easy ability to reference first class identifiers (the info URI scheme and registry still exists but clearly is deprecated). The only proffered alternative is to have each of the identifier schemes register as its own URI scheme, which surely was not the intent. An opportunity appears to exist to work with NISO on actions underway to help with this problem: for example, to request NISO to appoint another info URI registrar, and the LCC governance umbrella might suggest the

[20] NISO press release 28 Nov 2005
http://www.niso.org/news/pr/view?item_key=4b8a9e2d84fe28e5559d725eb6acd6fd9b1eb53d
[21] http://www.ietf.org/rfc/rfc4452.txt
[22] Italics added. Viewed from within the world of http, as in the statement quoted above, all first class identifier must all become second class identifiers - because the world is only http. If you accept that premise, then `all http's become first class because the "http://" namespace is immanent (e.g., if ISBN were invented now, it presumably would face claims that the syntax has to be something like ""http://www.isbn-international.org/1234561234567". There exists a case of actively used non-http resolution (Handle), which is enough to demonstrate the falsehood of that claim to http universality; and there exists a set of internet protocols allowing other resolution mechanisms to be invented.
[23] http://en.wikipedia.org/wiki/Info_URI_scheme

overall LCC future body as a candidate. We recommend that the option of including a revived info URI scheme is investigated by the LCC project.

2.5 Identifier interoperability

2.5.1 Types of identifier interoperability

To use identifiers in an integrated model we need to facilitate interoperability. At least three types of interoperability can be distinguished[24]:

- *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. This is fairly straightforward and trivial (the "bar code reader" level of interaction)
- *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. This is dealt with in the LCC metadata workstream.
- *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. This is the level of business interoperability: identifiers may well use the same syntax and the share the same semantics, but if the associated metadata has been costly to collect and manage there may be legitimate barriers to making this freely available. For the identifier workstream we have been agnostic as to open access/availability/commercial/paid for access of data: in the indecs principle of Appropriate Access, it is not for the workstream to specify what is "appropriate". However practical use of resolvable identifiers requires that some minimal set of associated metadata should be available to facilitate third party use[25].

2.5.2 Identifier interoperability schemes

Several initiatives focusing on aspects of identifier interoperability have been noted:

(1) The 2011 *Den Haag Manifesto* on persistent identifiers (PIDs) and Linked Open Data (LOD)[26] aimed to provide a base set of commonality among common persistent identifier schemes:

- Make sure PID's can be referred to HTTP URI's including content negotiation
- Use LOD vocabularies, for schema elements
- Identify the minimum common set of schema elements across identifiers in scholarly communication space.
- Use same-as relations to help PID interoperability across PID systems/schema's
- Work with the LOD community on simple policies/procedures to improve persistence of HTTP URI's.

However, the content community sees a very high need for interoperability at the semantic and community level in applications such as the Linked Content Coalition, but little demand for PID interoperability at the syntactic level (applications gathering information from URN, PURL, ARK, DOI

[24] "Identifier Interoperability": http://www.doi.org/factsheets/Identifier_Interoper.html;
[25] For an illustration of this in action see the DOI System concept of the DOI Kernel at http://www.doi.org/doi_handbook/4_Data_Model.html#4.3.1
[26] http://www.ncdd.nl/blog/?p=144

etc.), and hence we place a low priority on this issue. The simplistic view that "same as" relations will suffice is inadequate for LCC needs. The Den Haag manifesto has had little practical impact.

(2) APARSEN (The Alliance for Permanent Access to the Records of Science Network) is currently developing a *Persistent Identifier Interoperability Framework* which aims to build on the Den Haag Manifesto. It has been reviewed by one of the LCC technical leads who has offered substantial comments and suggestions, particularly the development of use cases and an invitation to consider collaboration with LCC. At present this focusses on Persistent Identifier interoperability at the syntactic level (applications gathering information from URN, PURL, ARK, DOI etc.), and has little relevance to interoperability at the semantic and community level in applications such as the Linked Content Coalition.

(3) The Corporation for National Research Initiatives (CNRI), developer of the Handle System, is developing an open source *Digital Object Based Interoperability Platform* (in collaboration with the Alfred P. Sloan Foundation)[27]. This is focussing initially on two different use cases, both outside the immediate scope of LCC (science data, and financial entity data), but the underlying principles may be useful for future LCC applications, as this will offer an open source suite for a distributed registration system linking to data and services across multiple existing information management systems, and thus enabling software clients to navigate and query multiple systems without detailed knowledge of those systems.

Of particular note in the context of resolution of identifiers (specifically multiple resolution), the CNRI project will build and deploy one or more data type registries, including information about services. The type registry would contain metadata about a certain data type as well as metadata about available services that could be used to process data of a certain type. The combination would allow either humans or machines to encounter data of a certain type, consult a type registry to understand the structure of the data so as to be able to parse it and to find relevant processing services, e.g., visualization. This approach is common and usually implicit within proprietary closed systems but is not yet generally recognised as an inevitable requirement of open linked data. This type registry would provide one means of supporting multiple resolution, by adding basic and extensible standard typing of resolution so that different services (e.g. different metadata types) can be automatically located.

(4) The WIPO *International Music Registry project* includes the aim of "Standard coordination, with the objective to promote the enhancement and coordination of standards associated with music metadata and identifiers"[28]. Such a global, standards based framework through which different industry initiatives in content identification and rights management data systems could interoperate is unlikely to be effective if confined to the music industry; nor does it make sense to re-invent existing efforts; therefore LCC is engaged with this initiative with the aim of collaboration and avoiding duplication. At present this WIPO effort seems to be at the level of a policy initiative rather than a detailed technical framework, so there is potential for such collaboration.

2.6 Co-reference and mappings

A potential problem arises in co-reference: the occurrence of multiple or inconsistent identifiers for a single resource. "Much of the Semantic Web relies upon open and unhindered interoperability

[27] Sloan Foundation press release – reference to be added
[28] http://www.internationalmusicregistry.org/portal/en/news/2012/news_0011.html

between diverse systems; the successful convergence of multiple ontologies and referencing schemes is key. However, this is hampered by the difficult problem of co-reference…"[29]

- Multiple identifiers for a single resource may not be fatal within a given system, even if inefficient, providing that a link recognising the multiples as equivalents can be established; but failure to establish such a link would result in "misses" for any attempt to return comprehensive rights information from different systems.
- A more troubling problem arises if such an equivalence is claimed but is not in fact in existence (i.e., if entity A and B are claimed to have the same referent, but in fact they do not. This problem of whether A is "really is the same as" B, has been a recurring feature of content identifier discussions over many years; the "same as" relation is contextual[30].
- Further complicating the issue, "compound objects" can be not only a package of several distinct separable things put together for convenience (an online Learning Object, for example, with a package of text, music, video); the things may overlap, or not be physically divisible instances. A given instance of an object may therefore encapsulate several related identifiers of different entities inherent in the intellectual property it represents, any of which might be exemplified in the object, e.g. (1) a pdf text file may simultaneously be an embodied instance of an abstract "work"; a particular publication edition of that work; and a specific format identifier; (2) a book is simultaneously an inseparable embodiment of an ISBN object and an ISTC object and a bar code object (and possibly in e formats a pdf, a file, etc). Unless explicitly declared, data interpreted from an identifier that appears to be apparent (e.g. obtained by resolution from a found identifier, or from a physical object in hand) may be ambiguous as to what is "identified".
- Identification might be asserted for the same referent by multiple sources. To dis-entangle any conflicting claims and reconcile these multiple assertions, the sources need to be traced; hence the asserters need to be clearly identified.

Some identifier frameworks offer the ability to express an existing identifier in the syntax, or as a "same as" metadata link to another system: for example: ISBNs may be expressed as GS1 bar codes[31]; ISO identifiers may be expressed as DOIs[32]. This confers both the advantage of being able to embody an equivalence, but the danger of embodying an incorrect equivalence – registries may not capture sufficient information with a specific registration, or may not publish detailed scope and application rules to enable re-use with confidence. Semantic interoperability analysis is the only way to establish this; failure to ensure true co-reference will result in irretrievable conflict in the systems' ability to interoperate. "Same as" relationships are clearly insufficient for a full articulation of rights, as defined in the RRM.

Accurate comparison of terms used in descriptive metadata to determine relations of creation identifiers, and mappings between them, is possible through the Vocabulary Mapping Framework (VMF). VMF is a downloadable tool, originally developed with funding from the Joint Information Services Committee (JISC), currently voluntarily hosted and administered by the International DOI Foundation (IDF) under the guidance of an independent multi-stakeholder Advisory Board. It provides support for semantic interoperability across communities by providing extensive and authoritative mapping of vocabularies from content metadata standards and proprietary schemes.

---

[29] Glaser, Hugh, Lewy, Tim, Millard, Ian and Dowling, Ben (2007) On Coreference and the Semantic Web. http://eprints.soton.ac.uk/265245/

[30] e.g. "On Making and Identifying a "copy"": http://www.dlib.org/dlib/january03/paskin/01paskin.html

[31] http://en.wikipedia.org/wiki/International_Article_Number_%28EAN%29#Bookland

[32] http://www.doi.org/factsheets/DOIIdentifiers.html; http://www.doi.org/factsheets/ISBN-A.html

VMF is an expansion of the existing RDA/ONIX Framework into a comprehensive vocabulary of resource relators and categories, and currently comprises a superset of those used in major standards from the publisher/producer, education and bibliographic/heritage communities (CIDOC CRM; DCMI; DDEX; DOI; FRBR; MARC21; LOM; ONIX; RDA). It is not intended as a replacement for any existing standards, but as an aid to interoperability, whether automatic or human-mediated. It includes mappings of terms from code lists or allowed value sets in the existing standards to the RDA/ONIX vocabulary, enabling the computation of "best fit" mappings between any pairing of standards. Subject to the terms of the VMF licence, VMF may be freely used to map and transform controlled vocabularies whether for commercial use or otherwise; and to inform the content of controlled vocabularies.[33] The support of the existing communities, plus the underlying use of the same contextual approach used in the RRM, makes VMF an obvious choice as a tool for LCC work such as a following Rights Data Integration project. VMF might be mirrored or expanded to cover needs for similar mappings for other entities. If VMF becomes more active, it will need active maintenance, and thus a more developed governance structure.

3. Currently available identifier implementations

3.1 Types of entities to be identified in the RRM

The LCC Rights Reference Model includes a list of entities to be identified – three well known ones (*Party, Place, Creation*) and five specific rights entities, the definition and use of which LCC will be pioneering (*Context, Right, RightsAssignment, Assertion, RightsConflict*).

From The LCC Rights Reference Model version 0.2: Table 11: RRM Entity Types

Attribute Type	*Definition*	*Examples*
Party	A human or other animate being (real or imaginary), or a legal person or organization capable of playing a role as an agent in a Context.	*John Smith, Coldplay, Microsoft Inc, Warner Music, the Boston Symphony Orchestra, Shrek*
Creation	Something made, directly or indirectly, by a human being(s).	*The textual work "Moby Dick"; a particular printed edition of "Moby Dick"; Mozart's 22nd Symphony; a photograph; the film Star Wars; a fragment of dialogue from "Star Wars"*
Place	A localizable or virtual place.	*Belgium; San Diego, CA; 15 High Street, Woking, Surrey, UK; Everywhere; johnjsmith999@hotmail.com; 020-8567-1047; Account No 1245265; Lat. 32o27', Long. 65o 88'; Outside London; Next to Jim's desk; www.anysite.org/thispage; Room 101, BBC Television Centre*
Context	An intersection of Time and Place in which Entities may play Roles.	*Earth during the Triassic Period; Europe in the Middle Ages; 1958 in Philadelphia; From 5.45pm to 7.13pm on May 5th, 2005 in Studio 1, Abbey Road Studios, London; 2006-06-0614:26 at www.anysite.org; Paying a*

[33] http://www.doi.org/VMF/index.html

		license fee; Having breakfast at Tiffany's; Somewhere, Sometime; Here and now; Always and everywhere; Writing an article; Owning a car; Publishing a journal
Right	A State in which a Party is entitled to do something in relation to a Creation, as a consequence of a law, agreement or policy.	*"Party A controls all rights in Creation C"; "Party A may copy, keep and view Creation C; but not on a computer of Type T and only after Payment P has been made by Party A to Party B"*
RightsAssignment	A decision as a result of which a Right come into existence.	*"Party A delegates control of European rights in Creation C to Party B"; "Party A permits Party B to make printed copies of Creation C"*
Assertion	A claim made about the truth or falsehood of a statement.	*A statement by Party A that it is true that Party B controls rights in Creation B; a corporate RightsPolicy granting user access privileges to people on certain management grades.*
RightsConflict	A State of disagreement or dispute over a Right.	*"Party A and Party B both claim Rights for Creation C in Germany"*

Also within the RMM are controlled vocabularies for *Categories* and *Times*: controlled vocabularies do not require new identifiers as a key *per se* (though many of the same principles apply) but where standards for these are available they need to be recognised and used appropriately, and so we mention these below.

3.2 Identification of Creations

Creations are the class of entity where identification standards and procedures are best understood and established. In the digital world, this results from two different yet converging trends: (a) the launch in the 1960s of the ISBN, and subsequent ISO family of related supply chain focussed identifiers of specific types of content; (b) the popularisation in the 1990s of digital location referencing through hypertext linking (the WWW).

3.2.1 ISO identifier schemes

A main group of content identifiers comes from ISO, through ISO TC46/SC9 (Information and Documentation). The list of SC9 standards[34] includes (dates are of the latest revision):

- ISO 2108:2005 International Standard Book Number (ISBN)
- ISO 3297:2007 International Standard Serial Number (ISSN)
- ISO 3901:2001 International Standard Recording Code (ISRC)
- ISO 10957:2009 International Standard Music Number (ISMN)
- ISO 15706-1:2002 International Standard Audiovisual Number (ISAN) Part 1 work identifier
- ISO 15706-2:2007 International Standard Audiovisual Number (ISAN) Part 2: version identifier
- ISO 15707:2001 International Standard Musical Work Code (ISWC)
- ISO 21047:2009 International Standard Text Code (ISTC)
- ISO 26324:2012 Digital object identifier system[35]

[34] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_tc_browse.htm?commid=48836&published=on

- ISO 27729:2012 International Standard Name Identifier (ISNI)
- ISO 27730:2012 International Standard Collection Identifier (ISCI)

These standards all have (or will have on next revision) a defined set of descriptive associated metadata. Unfortunately each set is independent of the other, with no common underlying data model, so mapping through a tool such as VMF is necessary to ensure interoperability. All need to be supported in LCC, though to implement them expressed as URIs may require additional steps by some of the registries.

3.2.2 Other creation identifiers

The ARROW[36] project, "a tool to facilitate rights information management in any digitisation project involving text and image based works" developed "ARROW infrastructure [which] allows streamlining the process of identification of authors, publishers and other rightsholders of a work, including whether it is orphan, in or out of copyright or if it is still commercially available"). As part of the project ARROW developed an inventory or "map of standards[37] with relevance to the ARROW project". This includes in its scope standards both for identifiers and for related themes (commercial messaging; conceptual models; metadata (generic, library, and rights); search; and technical protocols). Contributors included several of the current LCC technical workstream participants, with a one- or two-page data sheet for each standard. The last edition is relatively recent (2010); while it is not (we believe) being updated, so lacks more recent data (e.g. notably on EIDR, the entertainment industry registry[38]), it is still highly useful. We do not propose to repeat the ARROW analysis here but direct readers to it as a source.

3.3 Identification of Parties

Note: Party identification has relevance both to a Party as a primary entity (and hence an automated rights data supply chain expressing licensing etc.) and to identification of Asserters of other metadata. However it is authorisation rather than party identity which is the critical issue in asserting other metadata, where there is a need for each registry to have a governance mechanism to prevent multiple or ambiguous registration and provide (as far as possible) definitive, authorised registration. Cases such as orphan works may throw up problems of assertion, but projects such as ARROW address this[39].

The unique identification of Parties is the basis of an automated rights data supply chain. Party IDs are needed to identify creators, rights holders, licensors, licensees , users and others: they are the "alpha and omega" of the supply chain, allowing rights holders and users to be linked – imagine an online retail or banking system without a user login and password and the value of a Party ID is clear.

---

[35] Note that unlike the other SC9 standards listed "The scope of the DOI system is not defined by reference to the type of content (format, etc.) of the referent, but by reference to the functionalities it provides and the context of use" (ISO 26324, Introduction)

[36] http://www.arrow-net.eu/

[37] D4.4 State of the art and guidelines on applicable standards Edition.2 (July 2010) http://www.arrow-net.eu/sites/default/files/D4_4_State%20of%20the%20Art%20and%20guidelines_edition2.pdf in containing page: http://www.arrow-net.eu/resources/arrow-project-public-reports-deliverables.html

[38] www.eidr.org

[39] "ARROW infrastructure allows streamlining the process of identification of authors, publishers and other rightsholders of a work, including whether it is orphan, in or out of copyright or if it is still commercially available": http://www.arrow-net.eu/

Within proprietary systems, Parties are routinely issued with IDs for rights management and trading of all kinds. However, there is no generally established standard for Party IDs for rights holders, and only one real success story.

Parties play roles across sectors – for example, John Lennon was a composer, lyric writer, musical performer, actor, producer, artist, illustrator, text author, poet and photographer, among other things. Therefore if there is not going to be a single global Party ID for all interoperability (which there won't be) then various IDs must be authoritatively mapped. There are several initiatives worth noting as a basis for building a network of party identifiers in LCC. Several of these inherit ideas from the Interparty project[40], a spin-off from the indecs project.

The registration and identification of some abstract works is dependent on Party IDs. The administration of the ISWC, for example, is dependent on the CISAC IPI code. A party cannot get an ISWC for an abstract musical work unless its creators are all identified by IPI codes – otherwise anyone could go along and register "I love you" by "John Smith". This is one of the questions for registries for creations: in the absence of a governance mechanism for authorising and assigning the identifiers (similar to that for IPI, discussed below) how do agencies prevent multiple and ambiguous registrations? The same is true for Rights: without Party IDs, a Rights ID would be crippled.

3.3.1 The IPI code

Among the BIEM/CISAC collecting societies is there an established and ubiquitous Party ID (the IPI code[41], formerly the CAE number), and for over thirty years it has formed the basis of the relative success of international collaboration on licensing and royalty distribution within collecting societies and publishers for musical works (and to a lesser degree certain other CISAC-administered rights).

IPI has a number of features which explain its success, first in governance:

- An IPI code is allocated by the society of which a party is a member – this provides excellent verification of identity (linked directly to the party's commercial interests) and more or less removes the risk of duplication.
- The IPI registry in Switzerland records the society of each Interested Party so that the ID is extremely useful as the default for royalty payment ("I'm not sure of the song, but I know it was written by Paul McCartney")
- All societies have online access to the IPI registry

And in structure:

- It is an "unintelligent number"
- It is a "name ID" – each different name, pseudonym or alias has its own ID, and these are linked to a single underlying "Party ID"
- Pseudonym links are confidential and known only to those two whom a party wishes them known (there is one case of more than 100 pseudonyms of the same person)

IPI has weaknesses. It doesn't deal well with out-of-copyright and orphan works. Because (for example) Beethoven is not a member of a CISAC society, no-one has the recognised authority for uniquely identifying his works. It was suggested in the 1990s that societies "adopted" public domain creators on the basis of nationality, gave them IPI codes and oversaw the identification of their works, but this has not happened systematically, which is what is needed. The number of confusing and ambiguous "registrations" of public domain or arranged public domain works is correspondingly very large.

[40] http://www.interparty.org/
[41] http://www.ipisystem.org/

3.3.2 Activity in other sectors

In text, there has been nothing comparable to the IPI code: the ISNI (see below) is being introduced as the standard.

In music, performers have developed their own identifier (through the International Performer Database Association (IPDA) but plan to adopt ISNI. The labels are looking at options including but not limited to ISNI.

For still images there is no standard. Party Identification is one of the issues to be tackled by CEPIC within the proposed LCC/RDI project.

In the audiovisual sector there is no formal standard, though EIDR[42] now issue party identifiers (as DOIs) to audiovisual producers.

In the early 1990s there was discussion about opening up the IPI system to all, but it never got going because of political/commercial concerns, understandable when different groups of rightsholders were discussing collaboration. However, after a protracted process, there is now a promising ISO standard in ISNI.

3.3.3 ISNI

The ISNI (International Standard Name Identifier: ISO 27729:2012)[43] standard recently ratified was driven originally by the text publishing sector but backed by others including CISAC and the performers' associations (the International Performers Database Association). ISNI was developed as a standard for a "name" identifier for public parties "involved throughout the media content industries in the creation, production, management, and content distribution chains". OCLC, the US not-for-profit library co-operative, is managing the global registry database, and there will be multiple registration agencies. To date there are two (Bowker and Ringgold) who are respectively dealing with creators (predominantly in the text domain) and institutions. Both are just getting going. ISNI is focussed on identifying creators, not rightsholders:

*"…new ISO standard that will finally allow users to definitively identify contributors, across all forms of content. The **International Standard Name Identifier (ISNI)** is an ISO-certified global standard for the identification of contributors to creative works." (from the Bowker website).*

However, the standard says "*An ISNI can be assigned to all parties that create, produce, manage, distribute or feature in creative content—including human beings, legal entities (such as a company), or fictional characters"* which clearly embraces rights management. Bowker confirms this, so ISNI can be a Rightsholder Identifier. ISNI is being established as an interoperable identifier: a core part of its function is to map other standard or proprietary identifiers. CISAC societies, for example, will not abandon the IPI code, but IPI codes will be mapped to corresponding ISNIs.

ISNI has particular issues with verification and duplication. Unlike the IPI code, ISNIs will not be registered by a single method, pre-validated and de-duplicated by unique society membership criteria. Any organisation can, in effect, apply for ISNIs for any parties in which it has an interest – for example, a publisher or society registering all its authors. Data quality management and de-duplication is therefore a critical issue. ISNI is tackling this by having a single global database at OCLC, and building its initial database substantially from library authority records from the VIAF (Virtual International Authority File)[44] which enables the database to store a large amount of supporting

[42] Entertainment Identifier Registry: A universal unique identifier for movie and television assets www.eidr.org
[43] http://www.isni.org
[44] www.viaf.org

metadata (especially linked works) to support unique identification. “Registration” of ISNI will be as much about mapping to existing ISNIs as it will be about creating new ones – quality control is paramount, and drawing on centuries of bibliographic work and expertise is a wise and necessary step (very good to see the bibliographic and publishing communities collaborating in a major way on data issues for the first time).

ISNI is a “name number” which uses the same successful approach to pseudonyms as the IPI code, described above.

Because of its approach to authority data, ISNI is likely to have better success than the IPI code in dealing with unique identification of public domain creators (and by extension, supporting orphan work identification).

At the outset ISNI will be biased to the text and musical works/performance sectors, but there is no systemic barrier to other sectors participating. Not everyone is necessarily convinced or committed yet, and there are cost issues (as there were in the early years of DOI) which may be a problem for some.

3.3.4 NISO Institutional Identifiers Working Group

NISO (US National Information Standards Organisation) established an I2 Working Group[45] “to develop a robust, scalable, and interoperable standard for identifying a core entity in any information management or sharing transaction-the institution. The I2 Working Group did extensive community needs assessment with the publishing, library and repository use sectors”. With the emergence of ISNI, NISO reached an agreement to use ISNI for institutional identification, and I2 contributed further recommendations to the ISNI-IA that were incorporated into the ISNI standard. The I2 Working Group is now “finalizing a Recommended Practice, expected to be published in the next few months. This document will provide information on a profile that can be used by appropriate Registration Agencies to apply ISNI to institutions”. It remains to be seen how well this proposed profile fits into the bigger picture, but the fact that I2 teamed up with ISNI rather than creating yet another standard is commendable.

3.3.5 ORCID

ORCID, the Open Researcher and Contributor ID initiative, was established in 2010 and launched its service in October 2012[46]: “ORCID is an international, interdisciplinary, open, and not-for-profit organization created for the benefit of all stakeholders, including research institutions, funding organizations, publishers, and researchers to enhance the scientific discovery process and improve collaboration and the efficiency of research funding. ORCID aims to solve the name ambiguity problem in scholarly communications by creating a registry of persistent unique identifiers for individual researchers and an open and transparent linking mechanism between ORCID, other ID schemes, and research objects such as publications, grants, and patents”

ORCID was seen as a possible alternative to ISNI by some, but Bowker (as lead ISNI registration agency) and ORCID have now met agreed that they are complementary. ORCID is a specialized ID and should be mapped to ISNIs.

3.3.6 Legal Entity Identifier

ISO 17442 Financial Services – Legal Entity Identifier is to be launched in 2013[47] and is under development. The stated scope of LEI is on institutions holding financial assets, for the financial

[45] http://www.niso.org/publications/newsline/2012/wgconnectionoct2012.html#bi2
[46] http://about.orcid.org/news/2012/10/16/orcid-launches-registry
[47] http://www.financialstabilityboard.org/publications/r_121024.pdf

services sector. If implemented and extended this might play a role as a Party ID in rights agreements, but again LEI is a specialized ID and could in theory be mapped to ISNI.

3.3.7 Commercial/open source IDs

"Global" party identifiers are emerging as potentially powerful features in linked data in the likes of Google and Wikipedia. Google's new linked data initiative means that they will in due course have millions of party identifiers. However, these are effectively proprietary systems identifiers whose governance is not accountable.

3.4 Identification of places

We have assumed we are dealing here with geographical places, but note that in mobile networks end nodes such as telephone numbers might logically be defined as places[48], (as might technically IP addresses, though it is highly unlikely that this would find a use in LCC). This is not a hypothetical consideration – e.g. as is noted below in connection with GLN, geographical locations and the entities found there are often used interchangeably, with consequences for persistence and interoperability.

The wide range of examples given in the RRM has relatively few globally applicable standards:

- ISO 3166-1 standard country codes ("Codes for the representation of names of countries and their subdivisions") is probably the best known and established. It defines three sets of country codes:
    - ISO 3166-1 alpha-2 – two-letter country codes which are the most widely used of the three, and used most prominently for the Internet's country code top-level domains (with a few exceptions).
    - ISO 3166-1 alpha-3 – three-letter country codes which allow a better visual association between the codes and the country names than the alpha-2 codes.
    - ISO 3166-1 numeric – three-digit country codes which are identical to those developed and maintained by the United Nations Statistics Division, with the advantage of script (writing system) independence, and hence useful for people or systems using non-Latin scripts.

  ISO 3166-1 is widely used, implemented in other standards and used by international organizations. It is not the only standard for country codes (other country codes used by international organizations are partly or totally incompatible with ISO 3166-1) but appears to be the most likely basis for LCC use in e.g. defining national licensing territories.

- The Standard Address Number (ANSI/NISO Z39.43) is a unique identification code for each address of an organisation in the publishing supply chain it is administered by RR Bowker and in use widely in the USA though less so elsewhere. For an overview see a recent article in ISQ[49].

- The Global Location Number (GLN) is part of the GS1[50] supply chain system of standards (which also includes bar codes). GLN is broader in application than SAN, and is also used to identify legal entities (hence GLN crosses over into party identification). The GS1

[48] This is an example of the logical but initially somewhat confusing fact that all names are addresses in a namespace.
[49] The Use of the Standard Address Number (SAN) in the Supply Chain. Louise Timko. Information Standards Quarterly Summer 2011: Vol 23 No 3. www.niso.org/apps/group.../SP_Timko_SAN_isqv23no3.doc.pdf
[50] http://www.gs1.org/

> Identification Key is used to identify "physical locations or legal entities" in a hierarchy consisting of a GS1 Company Prefix and subsidiary location reference. Locations identified with GLN may be a physical location such as a warehouse or a legal entity such as a company or customer or a function that takes place within a legal entity. It can also be used to identify something as specific as a particular shelf in a store. Some physical supply chain and accounting systems may use GLN and these may need to interface with LCC in back office functions.

There will of course be many proprietary or internal place "standards" used in internal sales information systems etc., plus national address zip codes etc., GPS locations, etc. which will have application in specific territories for deeper sub divisions, which may need to interface with rights systems in any future automated "rights world". It is not possible to make general recommendations on these other than at the top level of following the principles of identification set out above.

It is worth noting that several of the examples given in Table 11 of "place" are not precise, nor do they necessarily need to be. Recalling the indecs definition of metadata as linking two referents, an unambiguous piece of metadata has to relate to precise enough things - referents - at each end of a link; e.g. the example given "Next to Jim's desk" (i.e., free form text, not in a defined registry) might be a perfectly precise enough referent as a localised description, but not if dealing with a geographically defined licence. This point applies to all entities, but for creation, parties, etc. the need for precision will normally lead to more specific registry identification.

3.5 Identification of rights entities

We are not aware of any international or national standards for identification of the types of rights entity which LCC has delineated in the RRM: *Context; Right; RightsAssignment; Assertion; RightsConflict.*

Related, we note that a proposed European Legislation Identifier (ELI) standard[51] was outlined in EU Council Document no. 17554/11 (metadata describing the document was posted on the EU official document register, but the full text of the document itself was not made public). Our understanding is that this will be used to identify laws, which is beyond the immediate scope of LCC, but as it relates to rights (without identification of the law on which a right is based, all that can be done to organize Rights is to establish a RightsType classification), this initiative should be noted. There appear to have been few public developments over the past year. In December 2011, a slide presentation about the European Legislation Identifier was made public. There is considerable interest in this document in the legal informatics community, particularly since new efforts, such as OASIS LegalDocumentML, are underway to harmonize legislative information systems across national boundaries.

3.6 Times

Times are assumed be taken from controlled vocabularies within LCC.

The most commonly used standard for time is *ISO 8601 "Data elements and interchange formats – Information interchange – Representation of dates and times"*[52] which provides an unambiguous and well-defined method of representing dates and times, so as to avoid misinterpretation of numeric

[51] http://legalinformatics.wordpress.com/2012/03/07/european-legislation-identifier/
[52] Latest edition 2004 (first published 1988): http://www.iso.org/iso/catalogue_detail?csnumber=40874

representations of dates and times, particularly when data is transferred between countries with different conventions for writing numeric dates and times.

ISO 8601:2004 is applicable whenever representation of dates in the Gregorian calendar, times in the 24-hour timekeeping system, time intervals and recurring time intervals or of the formats of these representations are included in information interchange. It includes calendar dates expressed in terms of calendar year, calendar month and calendar day of the month; ordinal dates expressed in terms of calendar year and calendar day of the year; week dates expressed in terms of calendar year, calendar week number and calendar day of the week; local time based upon the 24-hour timekeeping system; Coordinated Universal Time of day; local time and the difference from Coordinated Universal Time; combination of date and time of day; time intervals; recurring time intervals.

ISO 8601:2004 does not cover dates and times where words are used in the representation and dates and times where characters are not used in the representation.

Note that there may still be complexities in the implementation of ISO 8601: ISO 8601 is referenced by several specifications, but the full range of options of ISO 8601 is not always used. For example, the various electronic program guide standards for TV, digital radio, etc. use several forms to describe points in time and durations; the ID3 audio meta-data specification also makes use of a subset of ISO 8601.[53]

On the internet ISO 8601 is used in a profile of the standard that restricts the supported date and time formats to reduce the chance of error and the complexity of software. IETF RFC 3339 (" Date and Time on the Internet: Timestamps") defines a profile of ISO 8601 for use in Internet protocols and standards, and begins with the observation that "Date and time formats cause a lot of confusion and interoperability problems on the Internet". The more complex formats such as week numbers and ordinal days are not permitted and the RFC has minor technical deviations from the ISO specification; LCC implementers will need to note this restriction.

3.7 Categories

Categories are assumed to be taken from controlled vocabularies within LCC.

The RRM uses a **Category** Attribute (RRM, v0.2, section 4.2 and especially Table 5: Logical model of a Category) as a fully controlled data value denoting a classification, role or association of an Entity (for example, *Use Type=Play*). Category values may be any term from any code list, taxonomy or controlled vocabulary. There are myriad such lists (some are more useful than others[54]), and LCC should be open to use of any of them, though it is essential that the code list being used be precisely specified. We have not found any list of controlled vocabularies, or standards relating to classification identification; a comprehensive single "meta-catalogue" registry (catalogue of catalogues) does not exist. LCC should be able to use any classification, and will populate an identification of the classification through the mandatory elements shown in RRM Table 5 .

Categorisation has a long history through e.g. library classification. A useful analysis of principles can be found in the book by E. Svenonius[55]

[53] http://en.wikipedia.org/wiki/ISO_8601_usage

[54] For a memorable discussion see J.L.Borges, "The analytical language of John Wilkins", in Jorge Luis Borges, 'Other inquisitions 1937-1952'; 1964 (ISBN 0-292-76002-7).

[55] Elaine Svenonius: The intellectual foundation of information organization. Cambridge, Mass: MIT, 2000 (6th printing 2009) ISBN: 9780262512619 0262512610

3.8 Links

Links are not one of the primary entities in the RRM, but we note some current activities in this area as they are clearly relevant to LCC.

Conceptually the idea of a link identifier is important as we are beginning to see a whole class of "predicate identifiers" coming into use, without a full recognition that this is what they are. In ISO TC46 these include the ISSN-L (which defines a link between two related ISSNs) and the ISNI (probably).

ISO have recently issued a ballot to review a new TC46/SC9 Committee Draft standard, *"ISO/CD 17316, Information and documentation — International standard document link (ISDL).* This standard specifies the International standard document link (ISDL) identifier for the identification of links between objects. These objects may be media resources or more abstract items such as times or places." This is a development from a Chinese initiative which was specifying a specific link (for use with a proprietary pen technology and a printed mark to resolve to a URL – in essence turning a piece of print into a hyperlink) which has now been generalised. Members of the LCC technical workstreams have offered comments and feedback on the proposal, which currently seems to have critical problems but which are not hard to fix. In its current form ISDL would not be usable by LCC, but it is possible that a revised version might map well (or even mimic) the *logical model of a Link at RRM* section 4.6 Table 10. The name "International standard document link (ISDL) identifier" is inappropriate, as it is not linking only documents but resources of any kind (it can be used to link times to times, places to places etc. as specified).

3.9 General purpose identifier system: DOI

The Digital Object Identifier [DOI®] system[56] (ISO 26324) provides a technical and social infrastructure for the registration and use of "*persistent interoperable identifiers for use on digital networks"*. It is a general purpose identifier system (i.e. it may be applied to any of the entities above) specifically developed for the content industries with the aim of rights management at the forefront (though not the only application), initiated by the publishing community in 1998 and since adopted by other sectors for persistent unique identification of objects of any type. It places special emphasis on persistence and on semantic interoperability.

DOI is an acronym for "digital object identifier", meaning a "digital identifier of an object" rather than an "identifier of a digital object". It has so far been widely adopted for the identification of creations in some content sectors, notably the scholarly publishing, scientific data, and entertainment industries, with over 63 million DOIs assigned to date. The DOI system implements the Handle System[57] (a persistent identifier system which runs alongside, but does not require, DNS and is Unicode compliant) and the indecs Framework; a governance and management body oversees a federation of Registration Agencies providing DOI services and registration, and is the registration authority for the ISO standard (ISO 26324).

The DOI system may be used with existing standard identifiers such as ISBN[58], (either by inclusion in DOI metadata and/or in a DOI syntax)[59], or DOIs may be assigned to entities which are not otherwise already identified. The DOI system complies with the proposed LCC specification.

---

[56] Digital Object Identifier system: www.doi.org
[57] Handle System: www.handle.net . The Handle system provides "efficient, extensible, and secure resolution services for unique and persistent identifiers of digital objects," and may also be used for non-digital referents.
[58] DOI System and the ISBN System: http://www.doi.org/factsheets/ISBN-A.html
[59] DOI System and Standard Identifier Schemes: http://www.doi.org/factsheets/DOIIdentifiers.html

4. The Multimedia Identifier Network

In the digital age, automated transactions are dependent on the existence and linking of unique identifiers. For example, online shopping and banking, credit card transactions, air travel, logistics, online banking, email and social networking all work only if complex sets of identifiers and authorizations are registered and then later recognized in both manual and automated operations. In the online world, chains of linked identifiers have replaced, or run in parallel with, physical distribution infrastructure. Parties (people and organizations), products and processes (such as invoices, agreements and deliveries) all have unique identities with which the most complex, fast and detailed tracking and reporting is possible, but without which transactions often immediately become impossible.

Rights management (which includes licensing, usage reporting and settlement) is a set of processes that is critically dependent on networks of identifiers, but is one in which the absence of shared identifiers, and of reliable links between them, is at times severe. Many processes still rely upon manual recognition of ambiguous names, and making major improvements in this **Multimedia Identifier Network** is essential for the success of rights information exchange and of copyright licensing in general.

A **shared identifier** is one that is recognised by two or more different Parties or processes within the network. This may be a standardized ID, such as an ISBN or IPI Number, or it may be a proprietary ID such as an Amazon ASIN which is nonetheless published for use by others. Identifiers which are not shared, such as an organization's own private database IDs, cannot be part of the Multimedia Identifier Network until they are mapped at some point to at least one other identifier in the chain. For example: an audiovisual producer identifies a version of an audiovisual work with an internal ID of "ABC 12345", and then an ISAN is issued for the work of "1724-23B8-1246-0000-2-1M5Z-1456-T". Those two numbers must be mapped to one another as denoting the same Creation before the internal ID can have access to the Multimedia Identifier Network and the information which may pass through it. The more widespread, accessible, transparent and secure those mappings are, the more highly automated the network can become.

It is not essential that everyone uses the same identifiers (although the more that happens, the simpler the processes are), but it is essential that identifiers are securely mapped to one another so that they can be automatically translated.

There are a range of things which need to be identified to complete the Multimedia Identifier Network for rights management.

- Identifiers for **Parties,** the people and organizations who are creators, Rightsholders and Users of content. Some sectors (such as musical works) have established and successful shared Party IDs, but many sectors have none at all.

- Identifiers for **Creations**, the content itself. Many sectors have established standards, but some (such as images) lack these, and the granularity of identification may be inadequate (for example, not being able to identify chapters, clips or parts).

- Identifiers of a whole range of types, categories and concepts identified in **controlled vocabularies**. For example, rights verbs (eg "play", "adapt", "stream"), types of tools and media referred to (eg "mobile device", "television", "academic journal) and relators and roles (eg "author", "rightsholder", "part of") are typically defined in different controlled vocabularies and used in standard rights statements and licences. A term in a controlled vocabulary is an identifier in its own right, because it is used to refer

consistently to the same thing and is unique in its domain. However, no authorized mappings exist for most of the term identifiers used in different content and rights metadata schemes, so translation from one concept to another is arbitrary and only partially automated, and cannot be relied on in a wider network environment.

- Identifiers of **Time** and **Place**. ISO identifier standards exist for times and for territories which are globally recognised. Imagining a network without these standards give some indication of why the lack of shared standards for other entities causes such problems.

- Identifiers for **Rights** and **RightsAssignments**. There are very few public identifiers for these. How much they are needed, and how much they can be replaced by combinations of the identifiers above is yet to be established.

These identifiers provide the material for the basic building blocks of the Multimedia Identifier Network: **Links**. A Link is a statement in which two Identifiers of Entities are connected by a Relator. For example, the statement:

"John Smith is the author of John's Smith's book of jokes"

can be translated into three parts of a Link:

"John Smith" "Is Author Of" "John's Smith's book of jokes"

which in the Network may look like:

xyz/1234567 abc:author ISBN:9781273856633

Different types of Link are defined by the role or Relator term used. The most important types of Link for the Multimedia Identifier Network are set out in the LCC Rights Reference Model.

The Multimedia Identifier Network is an example of **Linked Data**, which is at present a rapidly growing Web phenomenon. However, it is not necessarily an example of **Open Linked Data**, because links between identifiers in the network need not all be made public to do their job. Nor is this simply a Web phenomenon: Linked Data is just the latest manifestation of a very old and successful method of representing data for automation.

In closing, we note the following from the computer scientist and science fiction author Vernor Vinge, in his 2006 novel of the near future "Rainbows End" (set in 2025), who envisages a world in which automated exchange of data is the norm, though still beset by problems we can recognise: *"They couldn't get any video, and … had suffered a 3030 error. (Xiu had looked that up; "3030" was a catchall code for a system deadlock caused by licensing conflicts.)"*

Document log

Editor: Norman Paskin, Linked Content Coalition Identifiers Workstream leader,
n.paskin@tertius.ltd.uk

Date	Version	Changes
25 Oct 12	0.1	• Initial draft issued for comment to identifier workstream members, technical workstream leads, and invited experts.
7 Nov 12	0.2	• Added sections on identification of time, categories, legislation • Restructured to move recommendations to be the initial section, other sections from v.0.1 now become a supporting appendix. • Expanded comments on mobile internet . • Added information on ELI • Added information on NISO I2 under identifiers of parties • Added WIPO IMR work in section on interoperability frameworks • Added further discussion of URN and URI registered assignments at section 2. • Added DOI in new section on general purpose identifiers in section 3 • Revised wording on ISTC • Extensive minor wording changes and corrections