Principles of identification

Version 1.0, April 2013

Editor Norman Paskin, Identifier Workstream Leader, n.paskin@tertius.ltd.uk

This specification comprises the LCC recommendations for the design and use of identifiers within the digital network in the digital rights supply chain. Detailed support for the recommendations is provided in the attached appendix.

The eight recommendations here are presented as a model of best practise for supporting the highest level of automation, trust and accuracy within the supply chain and network. They are not "mandatory" in the sense that none of them is legally or systematically enforceable for all identifier types, and failure to comply will not normally block the supply chain entirely, only make it more time-consuming, labour-intensive and error-prone.

1. Identification is essential

- Each entity which needs to be recognised distinctly in the digital network should have at least one public or shared 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 digital network.

2. Identifiers should not contain dynamic or confusing "intelligence"

- In general, "**dumb**" identifiers (that is, identifiers whose characters or elements have no intended meaning or referent) are preferable as they avoid some obvious pitfalls, but a limited "intelligence" can be safe and useful, and on occasion essential.
- Information about the **type** of the identifier is normally safe and useful: for example, prefixing an ISBN with "ISBN".
- Information about the **issuer** and **date of issue** of the identifier is best kept out of the identifier itself if possible in human-readable identifiers of content, as it is easily and commonly misinterpreted to refer to the owner or publisher of the *content* and its date of creation or publication. However, many identifier standards incorporate one or both of these references so they are often a *fait accompli*, and so the onus is on the parties or systems using them not to make false inferences.
- **Persistent information** about the entity (that is, information that should not change) should **not** be encoded within the identifier, because (a) like all metadata, it may be interpreted differently in different contexts and (b) it may be found to be incorrect at a later date. All such information should be declared as metadata, to which the identifier may resolve. Some existing identifier standards (such as V-ISAN) do encode

or imply metadata about the referenced entity (for example, that it is of a certain type or has certain properties) and so again this must simply be managed as well as possible.

- **Dynamic information** about the entity (that is, time-limited metadata such as status codes or rights ownership) should **never** be encoded in an identifier.
- These principle apply to **digital fingerprints** or binary identifiers created automatically from the digital structure of an item (for example, for recognising specific images or audio tracks) as well as identifiers created independently. Digital fingerprints do of course encode information about the entity they identify, but these are not-human readable and provided the content they identify is not itself dynamic this "intelligence" is normally benign and of course of enormous value for content recognition.

3. Identifiers should be resolvable

- A **resolvable** identifier in the digital network is one that enables a system to locate the identified resource, or some information about it (such as metadata or a 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 the Web, and a potential requirement to support other networks. This in effect recognises the URI as a primary practical common framework for global digital content identification. Non-URI identifiers may still be used where appropriate but should be expressible as or within 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".

4. 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: for example, to find least one basic description and one statement of rights.
- Choices in multiple resolution may be made by human beings or by machines following rules.
- Multiple resolution should be capable of managed change as data sources change: flexible resolution is essential to allow legacy and proprietary systems to interact.
- Multiple resolution of an identifier should be possible without special knowledge except for the ability to communicate using standard technical protocols.
- Multiple resolution requires a basic and extensible standard "typing" vocabulary of resolution so that different services (in the example given, 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.

5. Identifiers should be accessible

- Content identifiers should be accessible to users by (e.g.) embedding them where possible within the item of content or its message sidecar during interchange; making relevant information available in metadata; or embedding identifiers on webpages to support resolution to various services; and so on.
- Different approaches are useful for different purposes; the aim should be to provide accessible persistent identification.

6. Identifier registration should be under well-defined registry operations and policies.

- "Linked Data" technologies alone are not sufficient to establish a trustworthy industry-standard data exchange. The identifier-registered material must be 'data worth linking to': curated, value-added, data, which is managed, corrected, updated and consistently maintained by registration authorities and agencies. The LCC specifications should enable "curated data", i.e. consistent, managed, linking enabling links to other "quality data" with confidence, while still capable of using existing Linked Data technologies.

- Adequate supporting descriptive and rights metadata should be declared along with a registered identifier **to support discovery and avoid ambiguity**. Metadata should be registered under some method of governance (a registry or registration procedure) to ensures its authority and its ongoing maintenance in locations to which the identifier may resolve, using defined service types. Metadata about an entity is commonly declared by more than one party, and registry procedures can therefore provide ways of identifying the asserters of particular items of metadata and facilitate the resolution of conflicts within different metadata declarations.
- **Trust** in the accuracy and persistence of identifiers and their supporting authoritative metadata is critical. 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 an entity) and ambiguity (the issue of the same identifier to two or more different entities).
- Trust is necessary for several steps in identification:
    - that the identifier is for the thing you believe is being identified;
    - that the resolver you are using is the resolver you think you are talking to;
    - that the resolver you think you are talking to is actually the right resolver for the job;
    - that the data in the resolver relates to the thing you are asking about;
    - that the data in the resolver has been put there by a party with authority to do so;
    - that data in the resolver hasn't been subverted since it was registered.

7. 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. The specification and definitions of the schemas and vocabularies should be freely available to those needing to interpret the metadata.
- Standards may be public, formal, de facto or proprietary: there will always be a diverse range of metadata schemas for different sectors and functions (and this trend is likely to continue to increase.
- The semantic mapping of any well-formed schema to another (as described in the LCC specifications for metadata interoperability) cannot compensate for poorly-defined or ambiguous source data.

8. 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 provider or publisher of the metadata or the rightsholder: for example, an intermediary such as a collecting society, agent or licensee may create or pass on metadata on behalf of a party further up the supply chain, such as a creator or publisher: it is the party on whose behalf the metadata is declared who is the asserter or authority. An intermediary may therefore legitimately and necessarily publish conflicting metadata from different asserters on occasions, especially about rights ownership. It is a requirement for any metadata aggregator to have policies and methods for managing conflicting data from different sources, and on occasions from the same source.

The following are related recommendations for follow up by the LCC successor organization:

- The Vocabulary Mapping Framework (VMF) should be used for mapping metadata (terms and schemas).
- The Vocabulary Mapping Framework (VMF) should be mirrored and/or expanded from its current coverage to cover needs for similar mappings for other entities, and attention be given to active maintenance and a governance structure of VMF.
- Develop a scheme and methodology for associating a given existing non-internet registry scheme with a URI and associated structured metadata.

Appendix to "Principles of identification"

Version 1.0, March 2013

Editor Norman Paskin, Identifier Workstream Leader, n.paskin@tertius.ltd.uk

1 Introduction

1.1 LCC Identifier workstream

The initial 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 approach was through two top level questions:

(1) What are the entities in the LCC Rights Reference Model (RRM)[1] which require identification?

(2) What is the functionality required of these identifiers for use in the Reference Model?

The working group (25 participants at the initial project launch, later supplemented by a further 5 additions) adopted a combination of two approaches:

- *top-down* underlying principles of identification (**Section 2** of this paper): generic design of identifiers appropriate for use in the Digital Identifier Network, based on identifier principles and stated RRM requirements;
- *bottom-up* survey of currently available identifier implementations (**Section 3** of this paper).

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 (and how far they meet these functionality requirements) 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.

A separate LCC paper *The Digital Identifier Network* provides a summary of the integrated "big picture" for identifiers in the digital Rights Data Supply Chain.

The topic of identification of content is a large one with extensive published discussion. This paper does not attempt to be a comprehensive encyclopaedia of all issues, but to highlight the key issues of relevance for LCC. We have laid out a set of requirements for identification to provide a uniform approach to accessing rights data (for both people and machines), and a high level overview of the extent to which existing standard identifiers meet these requirements and of where gaps exist. We have not provided a detailed matrix evaluating every potential identifier system against the eight recommendations: each sector will need to consider how its selected identifiers meet the requirements and proceed accordingly. The next phase of LCC work should consider if such an overall matrix would be useful and should be compiled: crucially the value of such a matrix would lie in it being maintained, and therefore continuing governance and updating would be an issue.

[1] See the document *The LCC Rights Reference Model.*

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 so 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: as the Digital Identifier Network the digital network to which LCC applies substantially relies upon is the Internet an LCC-conformant identifier should be Internet Protocol compatible, as the digital network to which LCC applies substantially relies upon is the Internet[2]. 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.).

[2] 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

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.

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.

An initial draft v 0.1 was issued in October 2012 to identifier workstream members and other technical workstream leads, plus selected external experts, for comments. Version 0.2 was issued for the LCC plenary on Nov 13th 2012. This final document (version 1.0) accompanies the other deliverables at the end of the LCC project March 2013.

2. Underlying principles and issues of identification

2.1 indecs

As the 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"[3] 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, proven, and built on over the last decade in several significant content industry specifications which are aligned with the LCC approach, for example:

- 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.

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

We have not discovered any further underlying statements of principle which are not already encompassed in indecs or which meet the requirements of the Digital Identifier Network. Other proposals we have reviewed include:

- *ISO TR 21449*[4]: now outdated and does not add anything to the LCC analysis;
- *URN Functional requirements* (also now outdated and being reviewed in the light of developments since their original inception[5]);
- *URI principles* (see below under "Resolution"; also under potential review).
- *Dublin Core*: devised as a metadata set for searching for bibliographic resources on the internet, this has been called "fifteen terms in search of a data model". From the beginning its scope was limited; it is of some value for managing basic descriptive terms, but even there its limitations in terms of vagueness and ambiguity cause some serious problems (e.g. arbitrary distinction of "dc:creator" and "dc:contributor" which will be interpreted quite differently by different users, or the extreme vagueness of "dc:date"). Very few serious content metadata standards developed since Dublin Core have built on it, in both the content creator/publisher world (ONIX, DDEX, PRISM, PLUS etc.) and recent major bibliographic developments (FRBR and RDA).

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 also of the LCC's remit. 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

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

therefore a significant step in recognising the major improvements needed in the Digital Identifier Network[6] which are essential for the success of rights information exchange.

Independently, but wholly consistent with the indecs principles, the ontology expert John Sowa has noted that "Identifiers must be associated with sufficient metadata to specify (1) the permissible string of bits for an the identifier, (2) the naming scheme that determines how those bits are resolved to some entity, and (3) the ontological assumptions for determining how to interpret anything that may be found by this process", and has also provided a concise but incisive analysis of fundamental issues of identification on the Web [7].

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*.

This principle has also been recommended by W3C for URIs: "*Good Practice: Resource metadata that will change SHOULD NOT be encoded in a URI…*"[8] Note that this principle 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 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

[6] See the LCC Document "The Digital Identifier Network", published simultaneously with this document.
[7] John Sowa, at http://ontolog.cim3.net/forum/ontolog-forum/2007-04/msg00030.html; see also the in depth analysis in his book *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA, 2000. (summary at http://www.jfsowa.com/krbook/)
[8] The use of Metadata in URIs. TAG Finding 2 January 2007 http://tinyurl.com/ydd9yf

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[9].

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"[10]) 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 these are applicable only to unique physical objects (or unique digital objects in the form of hash signatures[11]) and are of no direct value in the identification of abstractions such as works, concepts and classes in which rights may exist.

2.3 Persistence

For long term interoperability identifiers must be persistent in the Digital Identifier Network, 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) when resolution is attempted. Identifiers 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 identifiers, a generic ISO Registration Authority agreement has recently[12] 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

[9] A recent paper on the FSB Legal Entity Identifier examines the issue of structure in identifier assignment and its role in federation mechanisms in great detail: *Braswell et al., Response to the Financial Stability Board's Request for an Engineering Study on the Best Approach to Managing the Structure and Issuance of Legal Entity Identifiers (LEIs)* (October 7, 2012): available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2197269

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

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

[12] 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)

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 used in the content sectors, the DOI System, has developed a series of persistence requirements (together with governance and operational policies) which may serve as a model for other registry authorities seeking to provide a comparable level of continuity[13].

2.4 Internet use of identifiers

2.4.1 Resolution, content management, and access methods

Identifier resolution is the process of going from an identifier to information about the identified entity and in some cases the entity itself. Identifiers that can be resolved over the Internet are sometimes described as 'actionable' and resolution is sometimes also called de-referencing[14]. In current practice, the main focus of LCC work is currently on the use of http (hypertext transfer protocol) built on the underlying internet. That in turn uses the http (hypertext transfer protocol) and related developments, generally 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 persistence and security when used for that[15] [16]. Protocols other than http may become increasingly important through mobile devices, etc.: "On the internet, web pages are only one of the many kinds of traffic that run on its virtual tracks. Other types of traffic include music files being exchanged via peer-to-peer networking, or from the iTunes store; movie files travelling via BitTorrent; software updates; email; instant messages; phone conversations via Skype and other VoIP (internet telephony) services; streaming video and audio; ….and there will undoubtedly be other kinds of traffic, stuff we can't possibly have dreamed of yet, running on the internet in 10 years' time"[17].

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. Of particular interest for content linking is the growth of mobile access: a reputable survey claims that mobile devices already account for 13% of all internet traffic; in 2012, 24% of all online shopping on "black

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

[14] 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.

[15] DARPA: "New Arch:Future Generation Internet Architecture"; D Clark et al. http://www.isi.edu/newarch/iDOCS/final.finalreport.pdf

[16] John Naughton: "Is it time for the internet to get the plumber in?". The Observer, 13 January 2013. http://www.guardian.co.uk/technology/2013/jan/13/internet-needs-to-get-rebuilders-in

[17] John Naughton: "The internet: Everything you ever need to know". http://www.guardian.co.uk/technology/2010/jun/20/internet-everything-need-to-know

Friday" (23 November) in the US was done via mobiles (up from 6% two years ago); and that in May 2012 mobile internet traffic in India overtook PC-based traffic[18]. Most mobile apps probably use http to exchange data but there is really no easy way to tell, since the app hides everything; in addition, mobile devises use technology which is less open than the web[19]. 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 (less open technology) have the advantage and the lead; they can access things like the camera and other apps and the advantage that security is easier to manage with a dedicated app rather than relying on what the web browser and web site give you.

2.4.2 Resolution and internet protocols

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[20]) 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[21], and thus multiple resolution is one option for facilitating contextual management of identifiers.

Note the distinction of 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.

---

[18] Mary Meeker: 2012 KPCB Internet Trends Year-End Update (Dec 03, 2012): http://www.slideshare.net/kleinerperkins/2012-kpcb-internet-trends-yearend-update

[19] http://www.guardian.co.uk/technology/2012/dec/09/smartphones-boom-bad-for-internet

[20] www.handle.net The Handle System was designed as a resolution system for digital objects and it serves as a level of indirection to any sort of current state data that you care to associate with the object through the identifier resolution mechanism. The Handle System provides a way to use DNS and URLs for identifiers, which simultaneously provides an identifier that can be resolved without using DNS and URLs, if you choose to use it like that. Most uses of the Handle System involve DNS, either as a way to get common web browser clients to communicate with handle servers (e.g. http://dx.doi.org/10.1037/0003-066X.59.1.29 or as the current state data returned from that resolution (e.g. http://psycnet.apa.org/?&fa=main.doiLanding&doi=10.1037/0003-066X.59.1.29).

[21] For an example using DOI, see http://www.doi.org/doi_handbook/5_Applications.html

It is important to understand the role, and limitations, of current internet resolution deployments especially the Doman Name System in relation to identifier management. This: *www.acme.com* is a domain name, which DNS resolves to an IP address, while this *http://www.acme.com/BigChart* is not a domain name: it is a URL, invented for hyperlinking. It relies on DNS resolution as the first step to find the IP address for an http server. DNS is an excellent resolution mechanism for domain names. This does not make it a resolution mechanism of any kind for other names or identifiers until you add something else. So using DNS and URLs for identifiers requires that you design some approach to using them consistently and coherently. In the same way that DNS and http URLs have not replaced databases but give you an easy way to reference databases, they will not replace well-structured identifier systems but can give you an easy way to reference those identifier systems.

2.4.3 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[22] 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; the formal URI specification that allows the URI to be opaque following the scheme name, e.g., 'http:' or 'mailto:', has been generally 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[23] (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[24] adopted it and was a key the motivation for it. InfoURI registrations can be made by anyone, not necessarily the authority for a particular namespace.

URIs may be used as "abstract" URIs (under the namespace “tag:” as an example[25]) 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.

---

[22] http://www.iana.org/assignments/uri-schemes.html
[23] IETF RFC 4452: http://info-uri.info
[24] 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.
[25] IETF RFC 4151: http://www.rfc-editor.org/rfc/rfc4151.txt

2.4.4 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 of 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”. [26]

2.4.5 Possible revision of URI specification

A recent post[27] 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

[26] 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
[27] Nov 2, 2012: http://lists.w3.org/Archives/Public/uri/2012Nov/0000.html

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 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.

2.4.6 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[28].

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

[28] 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/

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,

2.4.7 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 2003 by NISO[29] and became IETF RFC 4452[30]. 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. 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."

Viewed from within the world of http, as in the statement 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". We note that there exists a case of actively used non-http resolution (Handle), and there exists a set of internet protocols allowing other resolution mechanisms to be invented.

[29] NISO press release 28 Nov 2005
http://www.niso.org/news/pr/view?item_key=4b8a9e2d84fe28e5559d725eb6acd6fd9b1eb53d
[30] http://www.ietf.org/rfc/rfc4452.txt

2.4.8 Linked data and content identification

The adoption of URI in the LCC identifier specification conforms to the W3C Linked Data principles[31]. LCC takes the view that linked data needs to go further: linking is only as good as the quality of the data being linked to. LCC builds on the basic principles of linked data to address other issues such as the quality and typing of the values returned. URIs can be resolved to retrieve metadata about a content item, transaction, rights agreement, etc.

In the W3C Linked Data summary, it is noted that "an opportunity to make data interconnected… limits the ways it can later be reused in unexpected ways. It is the unexpected re-use of information which is the value added by the web….. Of course, this means that you have to get your data right, so it can be used in a reliable and automated way, as you write." LCC is about such *reliable* and *automated* use of information: to see the Web and other networks behave as far as possible in the reliable way that a single database does so that transactions can be made across it automatically and with confidence, using the Digital Identifier Network as a virtual database.

"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 infoURI schemes[32] 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 is unfortunate that all these existing schemes have lost the ability to reference easily a first class identifier (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. It is worth noting the fundamental issue of internet-based content identification, as analysed by the ontologist John Sowa[33], and his conclusion:

- "For physical objects, names are not unique because two different objects can have the same name.
- However, the laws of physics guarantee that no two physical objects can fill the same physical volume at the same time. Therefore, space-time coordinates can serve as unique identifiers.

[31] http://www.w3.org/DesignIssues/LinkedData.html
[32] http://en.wikipedia.org/wiki/Info_URI_scheme
[33] John Sowa, at http://ontolog.cim3.net/forum/ontolog-forum/2007-04/msg00030.html

- But we still have controversies between those who claim that terms such as "vase" and "lump of clay" represent only one individual at any given space-time location and those who claim that they represent two distinct individuals.
- The URLs and URIs of the WWW are based on a naming scheme that ultimately resolves to physical devices. It guarantees that an identifier will determine a unique storage location at a given point in time[34].
- However, the policies of the WWW and of each domain on the WWW permit the same identifiers to be resolved to different physical locations at different times.
- The nature of data allows multiple copies to be replicated at different locations very quickly, and it allows the same location to contain different data at different times.
- Those same issues make it very difficult to generalize a naming system designed for data to a naming system for physical entities and vice versa.
- These characteristics imply that the URIs of the WWW are important for certain kinds of resources, but they are just one scheme among many other "universal" schemes, such as social-security numbers, ISBNs, geographical co-ordinates, DUNS numbers, etc."

An opportunity appears to exist to take action to help with this problem: to develop a scheme and methodology for confidently and predictably associating a given existing non-internet registry scheme with a URI and associated structured metadata (the DOI system provides a clear example). The URN scheme and infoURI scheme, each devised to provide in part a solution, seem to have gained little practical uptake and traction in this space, and a complementary effort developed in the context of a clear user-led need such as LCC would be of benefit.

2.5 Federation of Identifier Systems

The governance of any identifier standard must consider the level and type of centralization desired in the system, ranging from a single monolithic registry to a more de-centralized federated system. Such considerations necessarily touch on both organizational/political considerations as well as network architecture issues. The creation or *minting* of an individual identifier (i.e., the number or code itself as opposed to the reference data) may be most efficiently carried out by a federated set of registrars.

[34] Although not destroying the main argument, it should be noted that this point is not precisely true, although it is an approximation which most users would accept (and was closer to the truth in 2007): the domain name piece of a URL may point to multiple IP addresses, which roughly correspond to multiple 'unique storage' locations at a given point in time (although to add to the complexity, that is also a little fuzzy as a given physical server can easily be the end point for routing to multiple IP addresses). The Sowa analysis is still very useful in considering the Internet as a collection of connected devices, but it continues to get more complicated; and this reinforces the point that identifiers require a specific dedicated mechanism beyond DNS.

2.5.1 Federation

Federation is not a precisely-defined term, even within the context of the Digital Identifier Network. Generally speaking, it describes an organizational structure somewhere in between a single entity or system and a set of completely independent entities or systems. Multiple entities or systems "federate" in order to jointly achieve some set of goals or functions while still maintaining some level of independence of action and governance. They do this by agreeing to co-operate with each other at some level, typically through the use of shared protocols or standards. The global telephone system is an example of this. There are many local and national telephone systems that work together, sometimes in relative ignorance of the details of each other's existence, by following a common set of technical protocols. The Internet can similarly be thought of as a set of local networks agreeing to a common address scheme (IP addresses) and implementing common network communication protocols such as TCP and UDP.

2.5.2 Federated identifier creation

Many global identifier systems use federated registrars to create valid and unique identifiers without the need to consult a central authority in each case. This is commonly done by sub-dividing the identifier space in some fashion and assigning or allocating the sub-divisions to various registrars. There are many examples of this approach being used successfully, some of which (Ethernet MAC addresses, IP addresses, GS1 product identifiers, and credit/debit card accounts) are outside the remit of LCC but provide working examples following the same principle. In the area of content management, the ISBN, ISRC, ISAN and ISWC are all managed through national registrars who assign numeric codes to content creators or publishers. In the Digital Object Identifier (DOI) system, an implementation of the Handle System, prefixes are allocated to organizations that then create identifiers by appending suffixes to those prefixes, but the distinctions between DOI registries are not based on nationality or territoriality but on the types of content or service being offered.

A critical issue in the federation of the issuing of identifiers is to protect, as far as possible, against the issue by of identifiers to the same content by two or more registries (and, by extension, the registration of duplicated descriptive or rights metadata). How this is done in any particular case will depend on the characteristics of the content and the sector affected: for example, the issue of unique ISWCs is generally guaranteed by its being tied to the collecting society membership of the creators (which in turn is made available by the management of another federated identifier, the IPI number for parties), whereas ISBN, ISAN or ISRC relies primarily on the necessary integrity of the internal processes of registering organizations (it is fundamentally in the interests of, say, a book publisher to ensure that a published edition has a single ISBN).

There is a particular problem with issuing identifiers for, and then preserving the uniqueness of identity of "orphan" and public domain works (those for whom there is no, or no known, rightsholder or responsible party). For each global standard, it becomes necessary for the registry or federation of registries to devise a sound method of enabling such content to be identified in a standard way.

There are numerous advantages to the approach of minting identifiers on a global scale by subdividing the space and distributing the authority to create the identifiers to a collection

of collaborating parties while still guaranteeing identifier uniqueness. Most of these advantages stem from the ability of the federated registrars to mint ids without consulting a central authority each and every time.

2.5.3 Federated identifier resolution

Identifiers may or may not be actionable within a single system or multiple systems, and those systems may or may not be tightly connected to whatever approach is used to create the identifiers. ISBN, as an example, was created for supply chain management and came into widespread use before ubiquitous network availability. The system for id minting was not associated with a system for resolving the ISBN to the sort of standardized data that the Digital Identifier Network requires. Many such systems sprang up but the tightly federated effort has pretty much been restricted to the minting of identifiers: it is thus actionable in some cases but not uniformly or consistently so across the entire collection of ISBNs.

Newer and global-scale identifier systems are more likely to include a federated resolution approach in addition to federated issuance. In a federated system, resolution is typically a multi-stage process. The resolution information will typically be distributed across multiple systems (controlled or used by the federates) and client software must first discover which of these to query. A common approach is for the identifier to be structured, or subdivided, such that client software knows which federate to query, or how to find out which federate to query, typically by the inclusion of a registry code within the larger identifier. In that sense, federated issuance and federated resolution fit together well. In the Domain Name System (DNS), for example, a set of root servers that are known to all DNS clients contain data that redirect clients to the appropriate lower-level DNS servers in a hierarchical fashion, going from right to left to ask, for example, the com server where to go for example.com and asking example.com where to go for www.example.com, and so on. The Handle System is more likely to have a two-level approach, with a set of root servers redirecting all queries that begin with a certain prefix, e.g., 10.1037 to a given set of servers to resolve, e.g., 10.1037/0003-066X.59.1.29.

The combination of a method for locating the federates responsible for a given subset of the overall namespace and an agreed-upon group of protocols enables the federating organizations to be addressed as a single virtual system. Client software does not need to know the location of every possible server ahead of time and can find the resolution data as it is needed, including in servers and systems that were only recently added to the federation. This gives a great deal of flexibility. Further, the federated systems need only agree on providing those services that they are required to provide in common. Each of the federated systems can provide additional services to their constituencies, possibly increasing overall efficiencies. This has been the experience of the International DOI Foundation, in which growth has come through the various registration agencies agreeing to provide basic DOI resolution in common while each separately provides a customized set of services to their customers.

Approaches to federation vary across systems, depending in part on community requirements and in part on the age and legacy constraints of each system but, as in the case of identifier issuance, the advantages of the federated approach all derive from a common characteristic - some level of independence from a central authority. There are two important aspects to this:

- *Federating existing systems.* An advantage of federation is that existing systems can be included without seriously disrupting their current operations. Whatever functionality is required for each federate can be layered on top of, or selected from, existing functions, thus reducing the need for new efforts and leveraging the proven reliability of existing systems.
- *Organizational Independence and Scalability.* Creating a global system by federating a set of local systems makes it easier to reach global scale. Domains and regions can be integrated by adding another federate without dictating how that federate must be created, funded, managed, etc. As long as the global level functions are met, e.g., provide data or services at a certain level of accuracy and timeliness, the underlying structures need not concern the global system. This also introduces more diversity of organizational and technical expertise and experimentation, making it less likely that the centralized system will stagnate.

2.5.4 Network architecture issues

There are a number of issues in the consideration of federated/centralized structures which relate to the architecture of the networks on which the identifiers will be used.

The first is *distributed computing* over networks. This is commonly associated with Internet-based federations (the web could be considered a very loose federation) but "distributed" does not equal "federated". Both centralized and federated systems can be physically distributed to avoid problems of single points of failure in either hardware or connectivity and can use redundancy to improve reliability.

Another issue of particular importance to identifiers is persistence (discussed in more detail earlier in section 2.3): will an identifier created today still support its intended function five, ten, or fifty years from now? The identifier technology and network architecture are important here in that they should provide the tools for persistence, e.g. keep meaningful semantics out of the identifier string so that changes in location or ownership don't break the identifier, as happens so often with URLs. But these tools will work only if there is an organization dedicated to their application and committed to making the identifiers work over time. Federations do provide strength in numbers and in that sense are likely to provide better organizational persistence than any single organization.

Finally, open architecture is key to long-lived and extensible systems. Today's Internet is the outstanding example of that approach. The protocols and standards used in connecting new or existing services to the Internet are widely and freely available, and any organization that wishes to provide a new service that can be interconnected to other users and services over the Internet can do so with minimal barriers to overcome. This has allowed it to survive and prosper in the face of enormous technical change. An open architecture Digital Identifier Network, using public interfaces for both input and output, will be much more likely to find wide-spread support and corresponding growth and stability. Again, this issue can be considered separately from centralization/federation, but open architecture and federation fit together well.

2.6 Identifier interoperability

2.6.1 Types of identifier interoperability

To use identifiers within the Digital Identifier Network we need to facilitate interoperability. At least three types of interoperability can be distinguished[35]:

- *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 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, or where it is commercially or otherwise confidential to a restricted audience, there may be legitimate barriers to making this freely available. The LCC identifier workstream has 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[36].

2.6.2 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)[37] 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.

[35] "Identifier Interoperability": http://www.doi.org/factsheets/Identifier_Interoper.html;
[36] 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
[37] http://www.ncdd.nl/blog/?p=144

However, the content community sees a very high need for interoperability at the semantic and community level within the Digital Identifier Network, but little demand for PID interoperability at the syntactic level (applications gathering information from URN, PURL, ARK, DOI etc.), and hence the LCC places a low priority on this issue. The simplistic view that "same as" relations will suffice is inadequate for the Digital Identifier Network. 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 within the Digital Identifier Network.

(3) The Corporation for National Research Initiatives[38] (CNRI), developer of the Handle System, is developing an open source *Digital Object Based Interoperability Platform* (in collaboration with the Alfred P. Sloan Foundation[39]). 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.

2.7 Co-reference and mappings

A problem arises with *co-reference* in the Digital Identifier Network: the occurrence of multiple or inconsistent identifiers for a single resource. "Much of the Semantic Web relies upon open and unhindered interoperability between diverse systems; the successful

[38] http://cnri.reston.va.us/
[39] Alfred P. Sloan Foundation http://www.sloan.org/

convergence of multiple ontologies and referencing schemes is key. However, this is hampered by the difficult problem of co-reference…"[40]

- 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 results 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[41].
- Further complicating the issue, "compound objects" may 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. For example, (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 straightforward (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.

All the above issues are real current problems, to greater or lesser degrees.

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[42]; ISO identifiers may be expressed as DOIs[43]. This confers both the advantage of being able to embody an equivalence, but the danger of embodying an incorrect equivalence which cannot then be rectified if a registry has not captured sufficient information with a specific registration. Effective processes to discover incorrect co-reference, and to amend data effectively when it is discovered, are essential for a registry. "Same as" relationships are clearly insufficient for a full articulation of rights, though fine for simply dealing with co-reference as long as the co-reference is correct.

---

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

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

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

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

2.8 Compliance

Content identifiers should be accessible to users, whether by being embedded within the item of content or its message sidecar during interchange, or published in metadata on webpages to support resolution to various services. Either or both approaches are useful for different purposes. We cannot solve the problems of rights and licensing without *consistently applied* identification systems. Both approaches assume that the identifier is the correct one, (i.e. has not been corrupted deliberately or accidentally by someone that one doesn't recognise the need for this). Compliance with identifier and metadata requirements, in particular preventing the removal of identifiers and metadata from content, has been identified as an important issue by the Hooper Report[44], which notes that some sectors need less work in terms of standards (in the sense that the standards already exist) but more in terms of compliance. In other words, using embedded identifiers works for some applications but not others. The current LCC Identifier workstream views compliance as outside its remit, but it is likely to be an important part of the LCC implementations (RDI and especially the Copyright Hub).

The book industry standards body Editeur compared best practice, (un)available identifiers and compliance risks in four media sectors (books, film & TV, music, photography) in a report[45] as part of the Linked Heritage project. The question of in-band vs. sidecar communication is a particular issue in digital photography, where the supply chain is somewhat different from that in the other three sectors. Much comes down to the degree of control or trust around the messaging used: the LCC has a role to play in reinforcing this point and so assisting in making Linked Data applications more authoritative.

Without some kind of protected "layer" of trust, either through the protocol, the application, or certification of compliance, transactions of value may be compromised. This is widely understood but not always provided for. URIs may be resolved using HTTP, or optionally HTTPS can be used to provide a layer of security (trust).

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*), one other general entity (*Context*), and four specific rights entities, the definition and use of which LCC is pioneering (*Right, RightsAssignment, Assertion, RightsConflict*).

From The LCC Rights Reference Model v1.0: Table 2: RRM Entity Types

EntityType	*Definition*	*Examples*
Party	A human or other animate being (real or imaginary), or a legal person or organization capable of playing a	*Tom Brown, Coldplay, Microsoft Inc, Warner Music, the Boston Symphony*

[44] Hooper Report: "Copyright works: Streamlining copyright licensing for the digital age", July 2012. http://news.bis.gov.uk/Press-Releases/Hooper-Report-Industry-should-lead-on-new-Copyright-Hub-67dd5.aspx
[45] http://www.linkedheritage.org/getFile.php?id=283

	role as an agent in a Context.	*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; TomjBrown999@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 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 comes into existence.	*An agreement in which Party A delegates control of European rights in Creation C to Party B; A license in which Party A permits Party B to make printed copies of Creation C; a corporate RightsPolicy granting user access privileges to people according to their employee roles and grades.*
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*
RightsConflict	A State of disagreement or dispute over a Right.	*Party A and Party B both claim Rights for Creation C in Germany*

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. 65^{o} 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 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"*

The RRM acknowledges one other Entity Type for which Identifiers are critical (Time), and one other set of essential identifiers (Category Values),

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 TC46 identifier schemes

A main group of content identifiers comes from ISO, through ISO TC46/SC9 (Information and Documentation). The list of SC9 standards[46] 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[47]
- ISO 27729:2012 International Standard Name Identifier (ISNI)
- ISO 27730:2012 International Standard Collection Identifier (ISCI)

Note that the ISNI is a Party, not a Creation, Identifier and is described more fully in section 3.3.

These standards all have (or will have on next revision) a defined set of descriptive associated metadata. However each metadata set is independent of the other, with no common underlying data model or common vocabularies, so the mapping of these through a tool such as VMF is necessary to ensure effective and extensible interoperability. Many of these are not yet expressible as URIs in a standard way and this may require additional steps by some of the registries. The ISO identifier registration authorities have held informal group discussions on collaboration re interoperability and re "identifier integrity" (trust issues re registration), but no formal steps have resulted.

3.2.2 ISO TC46 Identifier schemes reviewed by content type

Intellectual content is often categorized in four broad groups: music, text, audiovisual and still images. While this is a rough and ready approach which causes problems when pushed too far, it is a useful way to review the status of development of creation identifiers.

First though the distinction needs to be noted between abstract **works** and their **manifestations,** and the individual **items** which are distributed around the network. These distinctions are described elsewhere in the indecs and FRBR data models, but they have a particular significance for creation identifiers. None of the standard IDs listed above apply to *individual* physical or digital items (such as copies of a printed book, or a digital file): they are all identifiers of manifestations or works, which represent classes of items, The ISBN, for example, does not identify an individual printed book, but the entire **class** of books which form a specific published edition, each copy of which is considered to be an instance of the

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

[47] 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)

same manifestation. The same is true for ISRC and ISMN. A particular user such as a library may of course wish to assign a further identifier to their own copy of a manifestation, for various reasons, but there is no ISO standard for these.

Most of the other identifiers identify an abstract **work** - the underlying content which may be realised in any number of different manifestations. So the novel "Moby Dick" is a single abstract work which may be manifested in many different physical or digital editions: the work will be identified with an ISTC, while the manifestations may attract ISBNs or ISRCs (or both) according to their attributes.

Works and manifestations are different kinds of abstractions. A work is a single creation which may have any number of manifestations, while a manifestation is class of functionally identical items which typically originated with a single item which may then have been replicated any number of times. The work comes into existence along with its first manifestation, but the two are distinct and are commonly subject to different rights and may have different rightsholders.

3.2.2.1 Music/Audio

The ISWC (International Standard Musical Work Code, ISO 15707:2001)[48] was the first clearly recognized widespread application of an abstract work identifier, as a unique, permanent and internationally recognized ISO standard number for the identification of musical works . For example, the first ISWC "T-000.000.001-0" issued in 1995 to the song "Dancing Queen" identifies the song written by Andersson/Andersson/Ulvaeus, as distinct from any specific performances, recording, scores, arrangements, etc. made by Abba or any other party. Those "manifestations" will have other identifiers such as ISRC, ISBN or ISMN appropriate to their type.

In principle, the ISRC can be applied to audio content of any kind, including radio programmes or webcasts, but as yet there is no significant use beyond “traditional” commercial recordings.

3.2.2.2 Text publishing

More recently a corresponding concept for text-based works (ISTC = International Standard Text Code, ISO 21047:2009) has been standardised. The ISTC is a numbering system for the unique identification of text-based works; the term “work” can refer[49] in ISTC to any content that is predominantly text-based appearing in conventional printed books, braille books, audio-books, static e-books or enhanced digital books, as well as content which might appear in a newspaper or journal. As with the ISWC, it identifies the underlying content and is not dependent on the manifestation of that work. For example, in the case of John Smith, author of "John's Smith's book of jokes", the following base identifiers may be used:

[48] http://www.iswc.org/

[49] The term “work” must be used with care, as it may have different applications and implications in e.g. legal copyright discussion than in standards application.

- ISNI, to uniquely identify the author John Smith
- ISBN, to identify a particular manifestation of "John's Smith's book of jokes", and
- ISTC, to identify the content of "John's Smith's book of jokes" which may appear in other manifestations.

While a combination of all three (ISNI, ISBN and ISTC) may give a complete identification of the elements of a particular manifestation, the basic elements of creator and content may be separately and unambiguously identified by the ISNI and the ISTC.

Note that the ISTC, as with other Creation identifiers, may be applied at any level of granularity, so if necessary individual jokes in John Smith's book may have their own unique ISTCs. That may become necessary, for example, if specific jokes were reproduced in another collection.

There are two other standard and globally established work identifiers in the text publishing sector: the ISSN for serials/journals, and the DOI, which may be used to identify anything but whose largest application to date is for journal articles at the work level through the Registration Agency Crossref[50].

3.2.2.3 Audiovisual

Audiovisual works have two established standard identifiers: ISAN (including its derivative the V-ISAN) and the more recent EIDR identifier, which is an implementation of the DOI. In late 2012 the registration authorities of both agreed on a collaborative approach which would enable ISANs and EIDR-IDs to link and interoperate, which exemplifies the fact that it is not necessary for all parties to adopt the same standard identifier type provided they are "shared".

3.2.2.4 Still Images

At this point the most significant gap in the set of standard Creation identifiers is for still images (including photographic works): there is no standard. Initiative on this has been taken in recent years by the PLUS Coalition[51], and definitive work with the aim of reaching a globally-acceptable identifier and registry standard is to be undertaken by a number of parties under the leadership of the European picture libraries consortium CEPIC[52] within the Rights Data Integration project[53].

[50] www.crossref.org
[51] www.useplus.com/
[52] www.cepic.org/
[53] www.cepic.org/tags/tags/rights_data_integration

3.2.3 Other (non ISO TC46) creation identifiers

The ARROW[54] 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[55] 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[56]), it is still highly useful. We do not propose to repeat the ARROW analysis here but direct readers to it as a source.

3.2.4 Links between Identifiers

At the heart of the LCC, and the Digital Identifier Network itself, is the need for expressing standardised relationships between standardised identifiers. Between creations, these are generally of four kinds:

- "same as" links - ID1 denotes the same things as ID2
- "part" links - the entity denoted by ID1 is a part of the entity denoted by ID2
- "version" links - the entity denoted by ID1 is some kind of adaptation of the entity denoted by ID2
- "abstraction" links - the entity denoted by ID1 is an abstraction of the entity denoted by ID2

The last three of these link types has its counterpart ("whole", "source", "manifestation") when the link is looked at in the other direction.

A multimedia work (such as a website, for example) is likely to contain a large number of "parts", which in turn may be subject to relationships of any of these types. Rights may exist in any of these "part" creations, and the management of rights in the Digital Identifier Network is therefore critically dependent on the accuracy and accessibility of the links between them. If a website contains video clips, music, still images and a variety of text, then it may represent a manifestation of any number of ISANs, EIDRs, ISRCs, ISWCs, ISTCs, DOIs and (as yet unstandardised) image identifiers. At present these connections are

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

managed in partial, unauthorised and often opaque[57] ways, and the goal of LCC is to see these connections much more efficiently declared and managed for the benefit of all.

A necessary step towards this is to establish standard "relators" for the various Link types which can be used or mapped across all sectors, and this should be an important part of the ongoing work of the LCC.

3.3 Identification of Parties

The unique identification of Parties is the basis of an automated rights data supply chain. Party IDs are needed to identify creators, publishers, rightsholders, licensors, licensees, users, asserters and parties in rights conflicts: 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. The indecs model of "people make stuff, people use stuff, people do deals about stuff" underlines the simple primacy of parties: everything begins with a party, and without robust public or shared party IDs the foundations of the Digital Identifier Network are seriously compromised.

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 rightsholders, and to date only one real success story.

Parties also 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 no 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 within the Digital Identifier Network. Several of these inherit ideas from the Interparty project[58], a spin-off from the indecs project.

The identification of a Party has three common layers:

1. the identification of a unique human being or organization
2. the identification of different *names* by which a human being or organization is known
3. the identification of different *personae* or *aliases* adopted by a human being (or, less commonly, an organization).

One Party may have any number of names and personae which may need unique identification according to local functional requirements. For example, the performer known as David Bowie is a single human being with several names (including *David Bowie* and *David Jones*) and personae (including Ziggy Stardust). Each of these may require unique

[57] Of course, it is not always necessary for links to be "public", and at present many of them are established within the private databases of organizations such as publishers with interests in some of the content. The indecs principle of Appropriate Access applies here.
[58] http://www.interparty.org/

identification according to the purposes to which data is being put[59]. Some standards such as ISNI and IPI support this granularity.

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[60], 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 don't know the identity 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 formal 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

[59] The distinctions between different names, personae/aliases and roles played are "soft" and complex and the drawing of a line between them will be done in different ways by different parties. For example, is "Cliff Richard" just another name for the person originally known as "Harry Webb", or is it a different persona? Is "Ali G" a persona of the actor/comedian known as Sacha Baron Cohen, or just a role occasionally played by him? Is the fictional character of Winston Churchill depicted in a film the same person as the human being who was Prime Minister of the UK? and so on. There are ultimately no "right" answers to these questions and the LCC is concerned only that whatever criteria are applied by one party or sector can be mapped as accurately as their semantics allow to the criteria used elsewhere. As with creations, this requires "link" relators.

[60] http://www.ipisystem.org/

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: this parallels the "orphan works" problems everywhere.

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.

Elsewhere 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, although the PLUS Coalition has begun to issue IDs to registering Parties. 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[61] 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)[62] 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.

[61] Entertainment Identifier Registry: A universal unique identifier for movie and television assets www.eidr.org
[62] http://www.isni.org

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)[63] which enables the database to store a large amount of supporting 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. ISNI appears however to be currently "the only game in town" with a fundamentally sound methodology.

3.3.4 NISO Institutional Identifiers Working Group

NISO (US National Information Standards Organisation) established an I2 Working Group[64] "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

[63] www.viaf.org
[64] http://www.niso.org/publications/newsline/2012/wgconnectionoct2012.html#bi2

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[65]: "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 may be mapped to ISNIs like other sectoral Party IDs.

3.3.6 Legal Entity Identifier

ISO 17442 Financial Services – Legal Entity Identifier is to be launched in 2013[66] and is under development. The stated scope of LEI is on institutions holding financial assets, for the financial 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, and with social/communications media IDs (such as Facebook, Skype and Twitter) becoming increasingly important for networked identity.

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 and so their potential role in rights management is highly questionable without further authorization or warranty. Other self-issued social media IDs like those of Facebook and Skype suffer the same problem. Self-issued party IDs are self-evidently subject to very little governance, though it may be feasible, for example, for a person to map their own social media ID against their ISNI at some point if there is value in it (that is, if that ID is used elsewhere in accountable rights transactions).

Wikipedia IDs (and those from other large indexes like the Library of Congress Subject indexes) have more potential value to the Digital Identifier Network, as the IDs are not self-issued and there are editorial governance controls.

---

[65] http://about.orcid.org/news/2012/10/16/orcid-launches-registry
[66] http://www.financialstabilityboard.org/publications/r_121024.pdf

There are some pockets of potentially re-usable identifiers in specific sectors: for example, IMDB[67] for AV contributors (actors, directors, etc.) is semi-curated (user-contributed, but reviewed by staff before being accepted on the site) and is thus somewhat different from Wikipedia.

3.3.8 WebID

The W3C provides a specification for Web ID, "a way to uniquely identify a person, company, organisation, or other agent using a URI"[68]. The specification of WebID has been worked on since 2005, the latest specification being 2011[69]. The WebID page notes that "since you aren't a Document, a Web Page URL cannot be used to construct an Identifier that uniquely identifies you. It cannot be the Naming mechanism used by other Web users to accurately reference you. A Web ID looks similar to a home page URL, but it specifically identifies Entity You of Type: Person. Typically, the definition of Type: Person, comes from a vocabulary or ontology or data dictionary. One such vocabulary is FOAF, which is the basis of this effort."

As implied by the use of FOAF (Friend of a Friend project[70]), WebID focusses on social networking and does not have significant uptake in a structured way across content industries. Some social networking sites assign a WebID to participants automatically; some of these sites export (some of) the data which the participant has put into them. It is normally a subset -- perhaps just the social graph (i.e., who knows whom on the site). This is of very limited use beyond the site since the metadata may be uncontrolled and not mapped to a fuller content and/or rights ontology.

3.4 Identification of places

In the RRM, a place is defined as "a geographical or virtual place", and so includes not only any physical location but anywhere that a creation, party or data may be located or referenced, including the places or nodes identified by telephone numbers, URLs, IP addresses, email addresses or bank accounts). 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.

For the wide range of examples given in the RRM has relatively few globally applicable standards for physical locations:

- 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:

[67] http://www.imdb.com/
[68] http://www.w3.org/wiki/WebID
[69] http://www.w3.org/2005/Incubator/webid/spec/
[70] http://www.foaf-project.org/ : "FOAF defines an open, decentralized technology for connecting social Web sites, and the people they describe"

- 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[71].
- The Global Location Number (GLN) is part of the GS1[72] 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 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.
- AFNOR XP Z44-002-1997 code for the representation of names of historical countries[73]

  is important for archives and may be used to increase the value and correctness of historical descriptive metadata.

Standards exist ubiquitously for virtual locations, as by definition they are normally unlocatable without a unique identifier. For example, the following all operate under effective global identification systems:

- telephone numbers (ITU governance)

[71] 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
[72] http://www.gs1.org/
[73] http://www.freestd.us/soft/339586.htm

- email addresses, URLs, IP addresses (ICANN governance)
- bank sort codes/account numbers (industry bodies governance)

among others.

There are of course 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 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.

3.5 Identification of rights entities

We are not aware of any international or national standards for identification of three types of entity which LCC has delineated in the RRM: **Context, Assertion** and **RightsConflict**.

3.5.1 Identifiers of Rights Assignments

There are many proprietary identifiers of **Rights Assignments** (Licenses and Policies). There is some work in rights and rights assignments in the audiovisual sector, though the two are usually jumbled together – the assignment describes the right, rather than having a reference to the right. For example Avails[74] provides information about the time, location and business rules relating to offering an asset; MovieLabs in conjunction with others has developed metadata definitions for content recognition metadata, including but not limited to digital fingerprint[75].

In the music sector, the DDEX consortium[76] of leading media companies, music licensing organisations, digital service providers and technical intermediaries has standardised the format in which information is represented in XML messages and the method by which the messages are exchanged between business partners. These standards are developed and made available for industry-wide implementation. DDEX, as mentioned earlier, is consistent with the indecs approach of a contextual ontology (data model) with defined entities requiring identification.

A proposed European Legislation Identifier (ELI) standard[77] was outlined in EU Council Document no. 17554/11 (metadata describing the document was posted on the EU official

[74] http://movielabs.com/md/avails/
[75] http://www.movielabs.com/crmd/
[76] http://www.ddex.net/
[77] http://legalinformatics.wordpress.com/2012/03/07/european-legislation-identifier/

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 in some cases (Copyright Law, for example) are RightsAssignments according to the RRM and may therefore be referened in rights declarations. There appear to have been few public developments over the year since a slide presentation about the European Legislation Identifier was made public in December 2011. 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.

3.5.2 Identifiers of Rights

In the image sector the PLUS Coalition is in the process of implementing a public "Asset Claim" identifier which denotes the LCC **Right** entities (it has corresponding identifiers for Creation, Party and RightAssignment). Whether a more generally applicable Right ID or Rights Assignment ID will emerge or be required will to some extent be dependent on the success of the LCC in introducing its Rights model into the Digital Identifier Network.

It seems unlikely and unnecessary that a general Context ID will ever be required: there are many different specialized, proprietary Context IDs in use within the Rights Data Supply Chain (including License IDs, Usage IDs, Invoice Numbers and identifiers of any kind of performance). Whether any of these require a more widely used standard is not evident at this point.

3.6 Times

If all types of entity had identifier standards as robust and widely established as Times, most of the challenges of the Digital Identifier Network would have been met.

The most commonly used standard for time is *ISO 8601 "Data elements and interchange formats – Information interchange – Representation of dates and times"* [78] which provides an unambiguous and well-defined method of representing dates and times, so as to avoid misinterpretation of numeric 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.

[78] Latest edition 2004 (first published 1988): http://www.iso.org/iso/catalogue_detail?csnumber=40874

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.[79]

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 and controlled vocabularies

Category values (as defined in the RRM) are a particular kind of Identifier critical to the success of the Digital Identifier Network.

The RRM defines 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*). The category has two basic elements: the **Category Type** (eg *Use Type*) and the **Category Value** (eg *Play*) which may be any term from any code list, taxonomy or controlled vocabulary. There are myriad such lists (some are more useful than others[80]), and any of them may be used within the Digital Identifier Network,. Any value in such a list is an Identifier, as it must be unique within its namespace and it denotes a defined[81] entity or concept.

Individual values of identifiers in a code list or controlled vocabulary should be clearly defined and its management under the control of a recognised authority or registry. A comprehensive single "meta-catalogue" registry (catalogue of catalogues) does not exist.

A Category Value may denote any kind of entity or concept, and so straddles the whole range of entity types. There are many controlled vocabularies for every entity type defined in the RRM. In general, Categories represent classes or types of things (for example, Party Type, Right Type, License Type, Format), but a controlled vocabulary may also be used for identifying individual entities (such as Territories or Languages) where these are of limited and manageable scope, and where there is obvious value in the existence of a public identifier.

---

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

[80] 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).

[81] Standards of definition of controlled vocabularies and code lists vary enormously, and a vocabulary which simply uses controlled names without textual definition or description will be more open to ambiguity and abuse, but its values are still identifiers, even if the supporting metadata for them is inadequate.

Categorisation has a long history through e.g. library classification (though it dates back to Aristotle, whose methods are still generally used). For an analysis of principles see the book by E. Svenonius[82].

3.7.1 Mapping of controlled vocabularies

Because Category Values may be minted and deployed by anyone, their accurate mapping is critical to the success of the Digital Identifier Network. In general, mappings are done on a one-to-one, proprietary and as-needed basis, typically to enable one party to translate the values from an incoming message into values that its own system can recognize. This happens within organizations with multiple information silos (and therefore different vocabularies) as well as across organizations.

Mappings are not always precise, because the values recognised by one vocabulary may not be fully mirrored by those in another. It is also not uncommon for data to have to be restructured, as a single element in one system may be represented by a more complex set of identifiers in another. Within the rights data supply chain in the wider Digital Identifier Network there is are two further dimensions to the vocabulary mapping problem.

First, **authority**. Within a network, a party may be reliant on mappings carried out by an unknown third party: how can these be trusted, and how are they being maintained?

Second, **scale.** Many different vocabularies need to be mapped to many others. The number is increasing all the time, and the vocabularies themselves are changing and growing increasingly quickly in response to change (ONIX, for example, has more than 100 different code lists and issues revisions at least twice a year).

An obvious solution to these issues is the existence of "hub-and-spoke" mapping processes, where many different vocabularies can be mapped to single "hub" vocabulary, supporting many-to-many translation. For this to work, the hub vocabulary must be richer in structure than all of the vocabularies to be mapped. The **Vocabulary Mapping Framework** (VMF) was created for this purpose 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 is a tool for semantic interoperability across communities by providing extensive and authoritative mapping of vocabularies from content metadata standards and proprietary schemes. 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 some of the vocabularies 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.. Subject to the terms of the VMF licence, VMF may be freely used to

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

map and transform controlled vocabularies whether for commercial use or otherwise; and to inform the content of controlled vocabularies.[83]

VMF has not been extensively tested and used yet, but the support of several 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 and perhaps the Copyright Hub. If VMF becomes more active, it will need active maintenance, and thus a more developed governance structure.

3.8 Links

Primary entity identifiers provide the material for the basic "building blocks" of a Digital Identifier Network: Links (discussed in section 4 below). We note some current activities in this area that 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)* which states that " this proposed 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* in the RRM. 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

One identifier system already mentioned at 3.2.1 does not fit any single one of the categories of identifier of referent types listed above, since it is a general purpose identifier system (i.e. it may be applied to any of the entities above). The Digital Object Identifier [DOI®] system[84] (ISO 26324) provides a technical and social infrastructure for the registration and use of "*persistent interoperable identifiers for use on digital networks"*. It was 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

[83] http://www.doi.org/VMF/index.html
[84] Digital Object Identifier system: www.doi.org

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 65 million DOIs assigned to date. The DOI system implements the Handle System[85] (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[86], (either by inclusion in DOI metadata and/or in a DOI syntax)[87], or DOIs may be assigned to entities which are not otherwise already identified. The DOI system complies with the proposed LCC specification.

[85] 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.
[86] DOI System and the ISBN System: http://www.doi.org/factsheets/ISBN-A.html
[87] DOI System and Standard Identifier Schemes: http://www.doi.org/factsheets/DOIIdentifiers.html