Principles of identification

Version 1.1, April 2014

Editors Norman Paskin (n.paskin@tertius.ltd.uk), Godfrey Rust (godfrey.rust@rightscom.com)

An identifier is a name which is unique within its type and domain. This document comprises the LCC recommendations for the design and use of identifiers within the digital network in the content and rights supply chain. Detailed support for the recommendations is provided in the attached appendixes.

These recommendations are presented as a model of best practise for identification to support the highest level of automation, interoperability, trust and accuracy within the network. They are not *mandatory* in the sense that none is legally or systematically enforceable for all identifier types, and failure to comply will not normally block the supply chain entirely, but make its operation more time-consuming, labour-intensive and error-prone. "The digital network" here includes, but is not limited to, the internet.

1 Entities: what should be identified?

Public, persistent identification of key supply chain entities is essential.

- Each entity which needs to be recognised distinctly in the digital network should be assigned **at least one persistent public identifier** so that it may be denoted unambiguously wherever that is required or useful. The entity denoted by an identifier is known as its **referent.**
- A **public** identifier is one that is accessible and recognisable by people or machines within the digital network.
- Key entities which require identifiers include each item of content ("**creation**") which needs to be recognised (at whatever level of granularity is required), and each **party** (person or organization) who is recognised as, or claims to be, a contributor or rightsholder of content or an asserter of metadata.
- It is desirable for there to be a single standard public identifier for each entity, but where multiple public identifiers exist it is sufficient that they be linked ('mapped') in a way that enables one identifier to be automatically 'translated' to another.

2 Structure: what form should an identifier take?

- The assignment of an identifier always involves some pre-determined **general structure**, and some element of **specific value assignment**. The general structure may be as simple as "a ten-digit number" or may have a number of distinct components with different functions, such as a URI prefix, a date of issue, issuer code and check character. The specific value elements may be determined by something as simple as the sequential issue of a number from a range, or as complex as the generation of a digital "fingerprint" derived from the binary structure of the referent. It is normal but not essential that the assignment of an identifier is an automated process.

An identifier may have multiple "designations".

- The same identifier may take multiple forms or **designations** to fulfil different functions (for example, the ISBN has had three different designations, as a human readable 10-digit code

© Linked Content Coalition 2014

or “ISBN-10”, as a 13-digit barcode-compliant European Article Number or “ISBN-13”, and as an internet-resolvable Digital Object Identifier or “ISBN-A”).

- Because public identifiers in the digital network should be resolvable (see under “Deployment” below), and because the World Wide Web is the dominant network using the Internet, then any identifier in the digital network such should be expressible as a **URI** (Uniform Resource Identifier). The URI syntax can incorporate existing standard or proprietary identifiers (by adding a URI-compliant prefix to an existing identifier string) while remaining globally unique. Many existing ID standards, being pre-digital in origin[1], do not support internet resolution in their original format, and so an identifier may have a URI designation in addition to its original (for example, the ISBN-A example given above).

An identifier should not contain dynamic or confusing “intelligence”.

- In general, ‘**dumb**’ identifiers (that is, identifiers whose characters or elements have no intended meaning) are preferable as they avoid the risks of misinterpretation and change, but a limited ‘intelligence’ can be safe and useful, and on occasion essential.
- Encoding 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[2] 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 established identifier standards incorporate one or both of these references so they are often a *fait accompli*, and the onus is on the parties or systems using them not to make false inferences.
- **Persistent information** about the referent (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. However, some established identifier standards encode metadata about the referent (for example, that it is of a certain type or has certain properties) and so this must be managed as well as possible.
- **Dynamic information** about the referent (that is, time-limited or contextual metadata such as status codes or rights ownership) should **never** be encoded in an identifier.

3 Assignment: how should an identifier be issued?

An identifier should be issued under well-defined registry procedures and policies.

- A registry operates **a set of procedures and policies** for issuing identifiers. A registry may or may not manage a physical database or **repository** of identifiers. The governance of a registry may be established through a standard (as with ISO identifier registries) or it may be proprietary. A registry should establish trust in the accuracy and persistence of its identifiers and their supporting core metadata.

[1] For example, ISBN, ISRC, ISWC.
[2] The issuer and date of issue of the identifier is not, of course, the same as the issuer and date of issuer of the referent.

© Linked Content Coalition 2014

- The **scope** of the type(s) of referent for a type of identifier should be explicit in the registry procedures.
- An identifier should be assigned by a party **with appropriate authority** to make an accurate and unique identification of the referent.
- An identifier should be **unique within its type and domain**. The domain of a public identifier will normally be unrestricted and so it should be globally unique within its type.
- An identifier should be **persistent**, and once issued should *under no circumstances* be re-assigned to another referent, even if the original referent never came into existence or ceases to exist.
- An identifier should be assigned **at the earliest practical point** in the supply chain in which the referent comes into existence, before third party metadata or associations with it are established.
- Registry provisions should minimise instances of **co-reference** (the issue of more than one identifier to referent, often because of shared creation or ownership of content) and the more serious problem of **ambiguity** (the issue of the same identifier to two or more different referents), and deal with the resolution of these issues when they arise.

An identifier should be supported by metadata for discovery and disambiguation.

- An identifier should be associated with sufficient "core" descriptive metadata to enable its referent to be discovered and **unambiguously recognised.** Registry will therefore normally be associated with some form of metadata repository(s), but there may also be any number of metadata repositories associated with an identifier which are maintained by other parties independently of the original registry.
- Core metadata should be registered under a defined method of **governance** (a registry or registration procedure) to ensure its authority and its ongoing maintenance in locations to which the identifier may resolve, using defined service types.
- **Persistence** should be ensured through registry provisions for maintaining metadata after the original issuer or asserter is defunct or dead or otherwise unwilling to accept responsibility for it.
- Core metadata associated with a referent should be published in extensible and **interoperable syntactic formats** (for example, XML, RDF-TTL or JSON) using formalised schemas with defined elements and using controlled vocabularies wherever appropriate.

Registry procedures should be trustworthy.

- Registry procedures should ensure that users can **trust** that (a) the identifier is for the entity which they believe is being identified, (b) that the core registry metadata has been asserted by a party with appropriate authority and (c) that the core registry metadata has not been subverted since it was registered.

4 Deployment: how should an identifier be used?

An identifier should be accessible.

- Content identifiers should be **accessible** to users (including people and computers) by (for example) embedding them where possible within the item of content or its message sidecar during interchange, including them in public metadata or embedding them on webpages to

© Linked Content Coalition 2014

support resolution to various services. Different approaches are useful to meet different requirements: the aim should be to provide accessible persistent identification.

An identifier should be resolvable.

- A **resolvable** identifier in the digital network is one that enables a system to locate the referent, or some information about it (such as metadata or a service related to it) elsewhere in the network.
- Resolution of an identifier should be possible without special knowledge or proprietary tools except for the ability to communicate using **standard technical protocols**.
- Resolution should be capable of **managed change** as data sources change (avoiding "link rot" on the internet): flexible resolution is essential to allow legacy and proprietary systems to interact.

An identifier should be capable of resolution to multiple locations.

- An identifier should be capable of being resolved to more than one location ("**multiple resolution**") for different types or instances of metadata: for example, to find an example of the referent content, a description and a statement of rights. Choices in multiple resolution may be made by human beings or by machines, following rules.
- Any number of **repositories** of content or metadata may be accessed through resolution of the same identifier.
- Multiple resolution requires a basic and extensible standard **typing** vocabulary of resolution so that different services (for example, different metadata types) can be automatically located using standard protocols.

Identifiers of the same entity should be interchangeable.

- Where they exist in the network, multiple public identifiers with the same referent should be **mapped** in an accessible way so that they can be "translated" and substituted when necessary, whether automatically or by manual lookup. This is essential for entity types such as parties[3] which are always likely to have multiple public identifiers.

Resolution procedures should be trustworthy.

- Resolution procedures should ensure that users can **trust** that (a) the resolver being used is the one expected, (b) that the resolver being used is the right one for the task, and (c) that the data resolved to relates to the entity that is being asked about.

Appendices

This **Principles of Identification** document is supported by two Appendices:

- Appendix 1, **Identification in the digital content network,** follows the structure of the LCC **Principles of Identification** document and elaborates the recommendations.
- Appendix 2, **Identifier implementations in the digital content network,** provides an overview of the main current implementations of identifiers relevant to linked content.

[3] The ISNI (International Standard Name Identifier for public identities of parties) operates explicitly as a "mapping" identifier, following this approach.

© Linked Content Coalition 2014

Appendix 1

Identification in the digital content network

This document follows the structure of the LCC **Principles of Identification** document and elaborates the recommendations made there.

The *indecs* principles of identification

The following existing four principles of identification are adopted from the indecs metadata framework[4] and endorsed by LCC:

- *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

Prologue: Digital networks

Network technology

With the creation of the internet, and notably the WWW use of it from 1994, interactive applications changed dramatically over the last two decades. Major web applications emerged with significant scale increases in:

- The number of concurrent users and access points to data (via PCs on the web and later on mobile devices);
- The amount of data collected and processed, as it became easier and increasingly valuable to capture all kinds of data, including unstructured or semi-structured data;
- Cloud based services, outpacing relational database technology since relational databases are essentially architected to run on a single machine which is a mis-match for linked content, networked and fully cloud-based services[5].

The hypertext model that was selected for the Web, http/html, operates at the file level; so denoting a “document” not by a first class identifier but by a substitute (its address on a server) constrained the (web) identifier world to the use of the Domain Name System. DNS was designed for ease of redirection at the server level of IP addresses for delivery of packets of data; to get to some specific point within a file using http requires a further second mechanism dependent on the file address.

[4] http://www.doi.org/topics/indecs/indecs_framework_2000.pdf

[5] See e.g.”Why NoSQL: Three trends disrupting the database status quo” http://info.couchbase.com/WhyNoSQLWhitepaper.html

© Linked Content Coalition 2014

This hard-wires intelligence into an identifier string in a single parent domain, with implications for persistence and relationships to other identifiers.

It was then recognised (drawing lessons from non-network identifiers such as the "information and documentation" identifiers of ISO TC46/SC9) that identifiers should be first class objects that is, have an identity independent of any other item, including any protocols used to resolve the identifier, and so be free to have relationships which could be dynamic, i.e. fully contextual. A piece of content could then be identified independently of the server where it was (currently) to be found – avoiding the most common cause of lack of persistence, the infamous "404 not found" or link rot.

The logical distinction of *reference* and *resolution* (i.e. a referent is not necessarily the result of resolution) was not always appreciated, which led to much confusion in early web discussions of naming and addressing (see also 1.3 above). DNS provides resolution of IP packet addresses (URL) but is not ideal from the point of view of managing names as references (URI, URN) or from the point of view of the levels of persistence, scalability and security required in content naming.

In the 1990's, recognition of these issues led to proposals[6] for managing access to digital information not via the addressing of the component bits (packets of data), but as *digital objects*, a data structure for identifying and organizing information for access over a communication network. Adding to this the concept of *stated operations* (that is, defined types of operations that may be performed on a digital object) allowed for the possibility of identified objects to be distributed across networks, available for multiple uses in various ways, whilst maintaining complete independence of the underlying physical packets and wiring. This is therefore a level of abstraction from the underlying digital network to an information network view.

Digital networks and information networks

At the level of commerce and intellectual analysis, we are concerned with referents of all forms, notably abstractions. But at the level of technology, we are concerned simply with digital bits and how these are processed. Therefore persistent identifiers whose referent is of any form (digital, physical or abstract) are used in architectures whose sole focus is on digital objects ("bags of bits") which are processed in the same way. The link between the two views is provided through abstraction: an identifier may have a referent in any form, yet use an underlying digital technology to process and deliver either (1) a digital object which is a representation of (some aspects of) a physical object – e.g. a painting as discussed above; or (2) a digital object which provides sufficient information for representation for the purposes of the application – e.g. a data page about a person or a work.

Fig 1a shows the well-known hourglass model of the internet[7] . This model views an IP address as the "hourglass waist" or central common switching point between the layers above and below. The architectural aim is to keep this waist as thin and efficient as possible and not burden or ossify it with

[6] Managing Access to Digital Information: Cross-Industry Working team, 1997: http://www.xiwt.org/documents/ManagAccess.html

[7] See http://everything2.com/title/Hourglass+model

© Linked Content Coalition 2014

application detail (in the layers above) or technical implementation (in the layers below)[8], thereby allowing a single mechanism as the internet[9] (the TCP/IP protocols)[10]. Fig 1b[11] is an analogous view from the perspective of information management rather than switching technology. This views persistent identifiers as the common interchange core, the waist layer of the hourglass, for value added services in the layers above and data sources in the layers below. Analogies may be drawn with the earlier hourglass model of the internet: there are benefits in keeping this waist slim, and not burdening identifier systems with data sources or added value services which are logically separable. In each case, some technologies may choose to bundle some of these layers into packages for ease of use, but this should not compromise the ability of others who choose not to do such bundling. In implementing services one can choose to place functions in various design components (e.g. the services managed in a resolution system versus those pointed at by the resolution system, or having services centralized or distributed). The hourglass is a useful reminder of the design principle of keeping core identification as simple as possible.

[8] Steve Deering: "Watching the Waist of the Protocol Hourglass": http://www.iab.org/wp-content/IAB-uploads/2011/03/hourglass-london-ietf.pdf

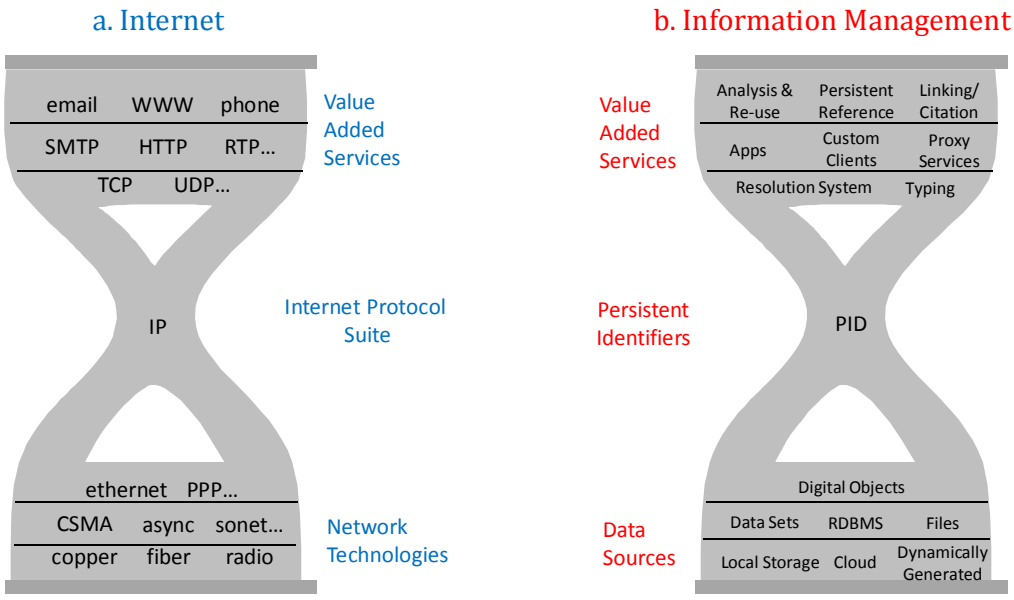
[9] "What Is The Internet (And What Makes It Work)": Dec 1999; Robert E. Kahn and Vinton G. Cerf http://www.cnri.reston.va.us/what_is_internet.html

[10] " "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." Resolution of US Federal Networking Council (cited in Kahn &Cerf op.cit.).

[11] The Information Management model was created by Larry Lannom, CNRI.

© Linked Content Coalition 2014

Fig 1: Hourglass Models



Corporation for National Research Initiatives

There is one notable difference between the two hourglass considerations. The waist of the hourglass in the "information management" model (persistent identifiers) is populated by a variety of identifier schemes from different sectors; whereas the waist in the "internet" version (IP) has only the single precise Internet Protocol specification. There have been attempts to define what level of interoperability or even fungibility might be possible among persistent identifier schemes to provide a single waistglass pinch point. However the requirements of interoperability required in information management (not only syntactic but semantic and community considerations) make this a much harder problem than is usually appreciated.

LCC: Ten targets of a digital identifier network

In 2013 The Linked Content Coalition (LCC) project was formed to scope what was needed to enable the digital network to function effectively for those who are creating, publishing and using content under any business model (or none). In 2014, the LCC (now a permanent consortium of standards bodies from all content sectors) published the following:

The effective operation of the digital content market relies needs the establishment of a global **identifier network** in which parties, creations, rights and usages are identified and linked in the internet in a way that enables the automated discovery of rightsholdings, and the licensing and reporting of usage.

The Linked Content Coalition has identified what it understands to be the essential elements of this network, and sets out below **ten targets** for data standards which, if fully implemented, would

© Linked Content Coalition 2014

provide the necessary infrastructure. Most, though not all, of these are partly in place at the beginning of 2014. A primary role of the LCC is to promote their completion.

1. **A global Party ID "hub".** Rightsholders and "Asserters" should be identified with an identifier linked to the ISNI "hub".

   A Party is a person or an organization (this includes different "public identities" of parties, such as pseudonyms adopted by creators). Unambiguous identification of Rightsholders and those who assert Rights declarations is most basic building block of the rights data network. The ISNI (International Standard Name Identifier) is a relatively new ISO standard identifier which can be used as an ID in its own right, but whose main role is to be a global "hub" to which different IDs for the same party to be linked together so that they can be automatically matched to or substituted for one another in systems when necessary. ISNI does not therefore *replace* other IDs, but enables them to interoperate with one another.

2. ***Creation IDs for all.*** Creations of all types should be identified to any required level of granularity.

   Public identifiers, supported by minimum metadata, are essential for Creations of all types in which rights are asserted (physical and abstract works as well as digital, because rights in all these are assigned in the digital network). Identifiers are needed at whatever level of granularity (sets, parts, fragments or derivations) specific rights are assigned for. Not all types of Creation have public ID standards, and those which do are not all as fully implemented as needed.

3. ***Right IDs.*** Content rights should be identified distinct from, but linked to, the Creations to which they relate.

   A "Right ID" which identifies a Right as a distinct data entity, separate from the Creation(s) it applies to and the agreements or policies which bring it into existence, is the most significant gap in the network's data. Because rights data is changeable, it cannot be reliably embedded into digital content itself, but should be accessible separately via linked identifiers.

4. ***Resolvable IDs.*** Identifiers should have a URI form which may be persistently and predictably resolved to multiple services within the internet.

   A resolvable identifier 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. Some identifiers, such as DOI and EIDR, are already resolvable, but many standard IDs do not yet have an expression in a URI format.

5. ***Linked IDs.*** "Cross-standard" links between identifiers should use interoperable terms and be authorised by interested parties at both ends of the link.

   Where one Creation (for example, a sound recording identified by an ISRC) has a dependent relationship with another (for example, a musical work which it contains, identified by an ISWC) then the vocabulary term describing that relationship should be standardised in some public schema, and it should be possible for Creators or Rightsholders of either of the identified Creations to agree or dispute the validity of the link under some registry procedure.

© Linked Content Coalition 2014

6. ***Interoperable metadata.*** Standard content and rights metadata schemas and vocabularies should have authorised, public mappings which enable terms and data to be automatically transformed from one standard into another.

   As with other identifiers[12], it is neither possible nor necessary for everyone to use the same schemas and terms, although the more common usage there is, the better. What is needed is for authoritative mappings (authorised by those who govern the schemas) available as services supporting automated "translation" of metadata.

7. ***Provenance of rights data.*** The provenance ("Asserter") of Rights declarations should be made explicit.

   In a distributed data network like the internet, the provenance of rights declarations must be explicit if systems or users are to be able to trust it (or not). The Asserter of a statement of Right may or may not be the same party as the Rightsholder. Without the ability to identify the Asserter of a Right (with or via an ISNI), there is no basis for secure automated identification of Rights in the network, or for the identification and management of conflicts (see target 9).

8. ***Digital Content Declarations.*** Anyone should be able to make standardised, machine-interpretable public statements about Creations and the Rights and permissions which apply to them.

   Using the elements described in 1-7 above, Rightsholders and their agents require a means by which any Party can simply identify and describe themselves, their content and their Rights in a Web or other network environment. This is especially useful for the huge volume of "direct-to-Web" publishing which now takes place, but can be applied by anyone. The standard should be built into services which support the publication and management of content and related IP in the network.

9. ***Dispute management*****.** Conflicts between public Rights declarations should be automatically identifiable so that their resolution can be managed.

   Conflict or dispute management has always been an important task for CMOs (collective rights management organizations) because they receive conflicting rights claims from different Parties. Where rights data moves out into the more "open" linked data, the same issues occur, but will be on a larger scale and not always under control of a single organization. Standard ways are needed of identifying and tracking these. *[link].*

10. ***Linked fingerprints*****.** Digital "fingerprints" should be mapped to registered Creation identifiers.

    Proprietary digital content recognition systems[13] (for example, Content ID, Picscout, Soundmouse and Attributor) provide the means for a variety of functions, including the tracking of digital usage. Linking these IDs to registered Creation identifiers ensures that such functions can be fully integrated with the rights data network.

[12] Note that terms in controlled vocabularies are identifiers, as they are unique names within their domain and type. When expressed as URIs they just become more identifiers in linked data.

[13] For example, proprietary systems such as Content ID (video), PicScout (images), Soundmouse (audio) and Attributor (text)

© Linked Content Coalition 2014

5 Entities: what should be identified?

Public, persistent identification of key supply chain entities is essential.

5.1 Identity is functional

The main reason for assigning an identifier is to separate things which are the same as each other from things which differ from them. If two things are different they require separate identifiers. To avoid the paradox of identity[14] we must add the qualification "…which differ from them *for some purpose*", and establish the criteria being used to make the distinction.

Identity is therefore never **absolute** ("A is the same as B") but **contextual** or functional ("A is the same as B *for the purpose of* C").

For example: two pencils on a table have been taken from the same packet of newly purchased pencils. For the purposes of writing, or of re-ordering more pencils by quoting the manufacturer's number printed on the side of each pencil, the two are *fungible* (that is, indistinguishable) and do not need to be separately identified. But if one pencil has been handled by a criminal whose fingerprints on it are crucial evidence of his presence at the scene of the crime, then for the purpose of forensic analysis the two pencils are no longer indistinguishable and need to be separately identified (for example, by adding an evidence tag).

In a second example, from information management, two editions of *Robinson Crusoe* may be indistinguishable for the purpose of a textual citation of the Defoe work (that is, they have identical content) even if one is an e-Book version and the other a leather bound presentation copy, but they must be separately identified for purposes of ordering a replacement copy.

Describing the contextual identity issue, the indecs project stated (as one of its four principles key to the management of identification) the *Principle of Functional Granularity*: "It should be possible to identify an entity whenever it needs to be distinguished". We might add "…for a particular purpose".

In everyday language the term "context" is often used loosely, but when applying identifiers the nature of the context should be definable in terms which can be analysed with precision, so that where necessary rules can be applied to determine granularity. The Linked Content Coalition's data model is based, like other interoperability initiatives, on a precise contextual model derived from the indecs project, where a context is defined as an intersection of *time* and *place*, in which *entities* may play *roles.* Fig 2 provides a summary of such a model. Each entity in a context needs an identifier.

[14] Liebniz's law of identity, or the indiscernibility of identicals ("X is identical with Y if and only if every property of X is a property of Y and every property of Y is a property of X"), leads to the conclusion that "Roughly speaking, to say of two things that they are identical is nonsense, and to say of one thing that it is identical with itself is to say nothing at all." (Ludwig Wittgenstein)

© Linked Content Coalition 2014

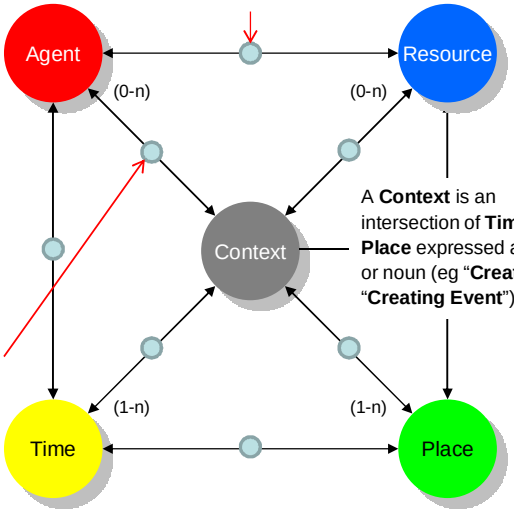
Fig 2: **Context Model**

This example is based on the **<indecs>** Context Model, which informs the design of the LCC and other content sector data standards.

Each Context creates relationships between every pair of Entities in them. These **Roles** are where metadata comes from (eg "**IsCreatorOf**")

A **Resource** is an Entity playing a Role in a Context other than that of Time or Place.

Many Events have **Agents** playing active roles which are the causes of the Event, which are often shown as being primary Entities in a Context, although logically they are types of Resource.



A **Context** is an intersection of **Time** and **Place** expressed as verb or noun (eg "**Create**" or "**Creating Event**").

The **Roles** which an Entity plays are defined by the Contexts in which it plays them (eg "**Creator**" is a type of Agent role).

A **Time** is a point (**Timepoint**) or portion (**Period**) of sequential time. Timepoints may be of any granularity (for example, of a nanosecond or a century).

A **Place** is a point (**Placepoint**) or portion (**Extent**) of Localizable ("geographic") or NonLocalizable ("virtual") space. Placepoints may be of any granularity (for example, a microdot or a continent).

Contexts and meaning

Meaning arises through Events. Metadata describes relationships between Entities, all of which occur through Events, or from the States which follow them. All Events and States are Contexts.

Non-contextual metadata (for example, relationships between Creators and their Creations, or Creationbs and their Time or Place of Creating) are often shown as direct relationships without reference to the underlying Context, but this Context is the key to understanding their meaning and origin.

In the Context Model every element of metadata is a "link" between two Entities. In physical models, many links are "denormalised" and treated as Attributes belonging to an Entity. The LCC Entity Model provides one example of the formal representation of this.

Even when two identifiers have been securely determined to have the same referent, this does not necessarily imply that the two identifiers are fungible. They may have *semantic interoperability* (that is, they denote the same referent), but the social infrastructure in which they are used may have restrictions, e.g. resolution to some metadata or access to a repository may be restricted to a community of registered users in one or both systems, and they may lack *community interoperability*.

5.2 Forms of Referent

A referent is the entity denoted by an identifier. Unique identification requires that each identifier has one and only one referent. A referent may have (and usually does have) more than one identifier.

Referents may be of any form:

- **physical** (for example, a book identified with an ISBN, a building identified with a postal address or a human being identified with a Social Security Number)
- **digital** (for example, a digital file identified with a URI or a virtual location identified with a URL)

© Linked Content Coalition 2014

- **spatio-temporal** (for example, an event such as a rights agreement identified with a licence number[15], or a insurance policy identified with a policy number).
- **abstract** (such as a song identified with an ISWC, or a theme identified in a controlled vocabulary or code list).

5.3 Types of Referent

There are several types of entity for which public identifiers are essential in the digital content network. The LCC Rights Reference Model (RRM)[16] defines eight types of entity of interest in the rights data network, of which three (**party**, **creation**, **place**) commonly have public identifiers. The RRM also defines a **right** as a distinct entity and proposes that a public identifier is needed for it.

Appendix 2 of the LCC Principles of Identification provides details of the status of public identifiers for parties and creations.

5.4 Classes and individual manifestations

Identifiers of manifestations are often applied not to individual items but to classes of creations. For example, an ISBN is applied to the class of books which are functionally identical as published editions, whereas an individual copy of the book to be located within a library or a second-hand bookshop will have its own distinct local identifier.

5.5 Controlled vocabularies as identifiers

Controlled vocabularies (sometimes known as “code lists” or “allowed value sets”) are essential groups of public identifiers in the digital network, although they are often not recognised as such. In part this is because they are often presented as “words” in natural language, but a controlled vocabulary term is unique within the vocabulary and once it is given a namespace prefix to turn it into a URI (eg dc:creator, ddex:SoundRecording) then it becomes a global public identifier within the digital network. Controlled vocabulary terms are commonly to identify classes of entity.

5.6 Digital representations of non-digital material

It is one of the ironies of digital identifiers such as DOIs (Digital Object Identifiers) that they often do not actually identify digital objects. A common role for digital materials is as a representation or manifestation of a *non*-digital entity (a creative work, a person etc.), because on digital networks such physical or abstract entities cannot be manifested directly. In linked content we are dealing not only with physical resources but with abstractions and digital materials. Abstractions and digital

[15] A “License ID” is often casually thought to refer to the document containing the details of the agreement, but in LCC is it recognised as an identifier of an event in which agreement was reached or permission unilaterally granted for some right to come into existence. The License document is a record of that event, and of the specific right(s) which exist as a result of it. The document itself may have a separate identity (indeed, there may be different copies of the document, as of any document, with distinct identifiers). The same principle applies to all manner of agreements or policies.

[16] http://www.linkedcontentcoalition.org/#!lccframe/c4nz

© Linked Content Coalition 2014

resources are just as "real" as physical entities, even though intangible: they interact with physical entities through actions in commerce and intellectual analysis[17].

A clear example is a digital representation of a physical painting, enabling users to see the painting without visiting the gallery in which it hangs. This example illustrates an important concept: in making a representation, some aspects of the original context are inevitably lost (as with the process of identification, representation is done *for some purpose):*

> " It's easy to forget that the very idea of a digital expression involves a trade-off […] A digital image of an oil painting is forever a representation… A real painting is a bottomless mystery, like any other real thing. An oil painting changes with time; cracks appear on its face. It has texture, odor, and a sense of presence and history.
>
> Another way to think about it is to recognize that there is no such thing as a digital object that isn't specialized. Digital representations can be very good, but you can never foresee all the ways a representation might need to be used. For instance, you could define a new MIDI-like standard for representing oil paintings that includes odors, cracks, and so on, but it will always turn out that you forgot something, like the weight or the tautness of the canvas.
>
> The definition of a digital object is based on assumptions of what aspects of it will turn out to be important. It will be a flat, mute nothing if you ask something of it that exceeds those expectations. If you didn't specify the weight of a digital painting in the original definition, it isn't just weightless, it is less than weightless. A physical object, on the other hand, will be fully rich and fully real whatever you do to it. It will respond to any experiment a scientist can conceive. …. it is impossible to represent it to completion. A digital image, or any other kind of digital fragment, is a useful compromise. It captures a certain limited measurement … within a standardized system that removes many of the original source's unique qualities.."[18]

6 Structure: what form should an identifier take?

6.1 Common identifier forms

Standard identifiers in the content supply chain have commonly incorporated some or all of the following elements:

- the **type** of identifier (eg ISBN)
- a code for the **issuer** of the identifier
- a code indicating the **place of issue** (typically a territory) of the identifier
- a date or datetime indicating the **time of issue** of the identifier
- one or more **check characters** calculated by an algorithm to validate the integrity of the identifier

[17] An extensive discussion of this issue can be found in the work of Karl Popper, see e.g. Objective Knowledge: An Evolutionary Approach, 1972, Rev. ed., 1979, ISBN 0-19-875024-2

[18] Jaron Lanier, You Are Not a Gadget (2010), Ch.10

© Linked Content Coalition 2014

The advent of the Web, and with it the URI (Uniform Resource Identifier) in its various specialized forms (which include URL, URN and DOI) has led to a more flexible and open-ended approach to identification, with uniqueness being secured by a globally unique common prefix (such as a DNS domain name www.anything.com/ or a DOI registration agency prefix "10.0001/") with a user's local identification model providing the remainder or suffix of the identifier.

At the same time there has been a steady move away from human-readable identifiers to **machine-readable** strings, a growing number of which are assigned by, and only readable by, computers. Increasingly important among these are digital "fingerprints", derived from the binary structure of the content of digital objects[19]. This ability to generate an identifier from content-in-hand is known as **affordance**: "a situation where an object's characteristics imply its functionality and use"[20]. "Afforded intelligence" identifiers are not normally as fragile as intentionally 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[21]) and are of no direct use as identifiers of abstractions such as works, concepts and classes, or physical referents outside the digital network, such as people or buildings.

An identifier may have multiple "designations".

6.2 Multiple forms ("designations") of the same identifier

It is increasingly common for a single identifier to be expressed in more than one standard form for different purposes. The book identifier ISBN provides a representative example. Its original, "human readable" standard form (known as the "ISBN-10") was of a 10-digit number with the common prefix "ISBN" (for example "ISBN 817525766-0"). This then became expressible as a 13-digit European Article Number (known as the "ISBN-13") suitable for encoding as a printed and machine-readable "barcode" ("9788175257665"). More recently the same ISBN has become expressible as a Digital Object Identifier (DOI) (known as an "ISBN-A", or "actionable ISBN) using the standard DOI syntax in the form "10.97881/75257665" to enable it to be resolved in a digital network. All three identifier forms have the same referent. In LCC terminology, these are different **designations** of the same identifier.

An identifier should not contain dynamic or confusing intelligence.

[19] Exemplified by products and and services such as (for example) PicScout (images), ContentID (Google/Youtube), Soundmouse (audio) and Digimarc Attributor (text).
[20] http://www.usabilityfirst.com/glossary/term_66.txl
[21] e.g. the proposed URI scheme "Naming Things with Hashes" http://tools.ietf.org/html/draft-farrell-decade-ni-10

© Linked Content Coalition 2014

6.3 Meaning in identifier strings

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. On the other hand, it is not the simple case that identifiers should be "dumb", but care is needed in embedding and using meaning in identifier strings.

Meaning in identifiers may be classified in one of three way: **vital intelligence** (see 2.4), **"risky" intelligence** that may be best avoided as it may be easily misunderstood (see 2.5), and **dynamic intelligence** which is changeable and so should always be avoided (see 2.6).

6.4 Vital intelligence

Vital intelligence in an identifier relates primarily to the process of allocation of the identifier itself, to ensure its global uniqueness typically by assigning a **set of prefixes** to one agent, which may then create its own unique namespace by further qualification, as in 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).

When identifiers are assigned on a federated model (see 3.8), it is normally essential to include such a prefix. Existing federated global identification standards of which LCC is aware (for example DOI, GS1, Media Access Control [MAC address], Internet Protocol [IP and IPv6], EIDR) use such a structured prefix solution in order to allow maximum possible flexibility of local members in the allocation of identities. We have not, in contrast, been able to find any examples of unstructured global federated identifiers. It appears that successful federated global identification standards have found it necessary to adopt a structured numbering assignment system[22].

The prefix may designate the **type** of identifier in the string itself (as in http:) or the **issuer** (as in a DOI prefix such as 10.1000). In human readable identifiers these may be presented transparently (as with "ISBN" or "DOI" preceding the rest of the identifier), or it may be encoded more cryptically: for example the three-digit code "978" at the beginning of a 13-digit EAN (European Article Number) means that the remaining 10 digits represent an ISBN.

It is also common for identifiers to include one or more **check characters** calculated by an algorithm, with which the validity of a particular string can be checked. For example, the ISBN has a single digit check character in the last position, and this varies according to whether the form of the identifier is a ten character ISBN or a 13-digit EAN.

[22] 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

© Linked Content Coalition 2014

6.5 "Risky" intelligence

Identifiers commonly include information about the **issuer** and sometimes the **date of issue** of the identifier. For example, the book identifier ISBN contains a group of 2-8 digits which designate the publishing group and publisher issuing the ISBN to the referent, and the sound recording identifier ISRC contains a two-digit code designating the year in which the ISRC was issued.

Such information may clearly be useful, but it is also easily and commonly misinterpreted. The issuer of the ISBN may be assumed to be the owner or publisher of the book in question, but may only be acting on their behalf; and the year of issue of the ISRC may be assumed to be the year in which the recording was released, but of course because a great deal of older recorded material is digitized, this is commonly incorrect.

It is also preferable that **persistent information** about the entity (that is, descriptive or other 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. Identifiers commonly include information about the type or form of the referent, or (in the case of creations) information about links to associated creations (such as the original of which the referent is a version or excerpt).

All such persistent information is best declared as metadata and stored in a repository to which the identifier may resolve (see section 4).

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

6.6 Dynamic intelligence

Dynamic information about the entity (that is, information which may only be true for a particular time or in a particular context) should **never** be encoded in an identifier. This includes

- **location** information (such as a URL) for the referent (unless the referent is itself the location, of course);
- **status** information (such as the availability of a product); and
- information on ownership of, or **rightsholdings** in the referent, as this may change over time, place and for different purposes.

This principle has also been recommended by W3C for URIs: "*Good Practice: Resource metadata that will change SHOULD NOT be encoded in a URI…*"[23] Note that this principle is not the same as defining an identifier specification "that contains no embedded intelligence".

[23] The use of Metadata in URIs. TAG Finding 2 January 2007 http://tinyurl.com/ydd9yf

© Linked Content Coalition 2014

7 Assignment: how should an identifier be issued?

An identifier should be issued under well-defined registry procedures and policies.

7.1 Registries

A registry is a database or other information structure providing a definitive record of identifiers issued within a particular namespace, typically with associated data such as basic metadata about each referent, the registrant, and "audit trail" information about when and by whom the identifier was issued, and so on.

7.2 Scope of identifier registries

An identifier registry will have a defined scope, such that certain entities are within its scope and others outside it. For creations, in pre-internet days, when much material in libraries was physical in format (books, periodicials, CDs, tapes, etc.), classification was simple, on the basis of format (although even then mixed media items caused problems); and within this by e.g. subject (e.g. "books whose subject is poetry"[24]. The definition of scope may be translated into the presence of certain attributes of a referent, which define the associated metadata to be registered. This assists the practical development and operation of some registries, especially when their scope is the more traditional library material of physical objects or material which is expressed via physical objects. Typically, ISO TC46 SC9 identifier schemes arose from such library-like documentation considerations, and creation identifier standards (ISBN, ISRC, ISWC, ISSN, ISAN etc) have been

However, this approach is increasingly inadequate. Pre-internet, a few exceptional identifier schemes allowed for abstract classification of any entity (for example, OID). With the rise of the internet, where all entities becomes represented as digital objects ("bags of bits") which are processed in the same way, the limitation of traditional classification has become clearer and the scope of identifiers such as URI, URN, HDL or DOI is recognised as potentially arbitrarily wide. For practical reasons, limitations of registry scope must be introduced to provide a basis of social infrastructure, but this tends to determined by functionality rather than form, so (for example) DOIs are allocated by different registration agencies each of which have a defined coverage or scope normally based on application (e.g. "scholarly citation") rather than referent format.

Further adding to the problem of classifying by format, not only is the amount of data requiring identification growing rapidly, but the nature of data is changing as well: it is estimated that over 80% of data generated today is unstructured or semi-structured and so does not fit readily into a pre-formed classification.

The view that the world of information can be categorised into non-overlapping independent silos in one definitive manner is now recognised as mistaken. All classifications are arbitrary; since all

[24] Svenonius, E. (2000) The Intellectual Foundation of Information Organization. MIT Press, Cambridge, Massachusetts

© Linked Content Coalition 2014

classifications are a form of identification through grouping into identified classes, they are subject to the same choices of functional granularity as with single identification. There are no absolutes, just mechanisms and human understanding:

> *"These ambiguities, redundancies, and deficiencies recall those attributed by Dr. Franz Kuhn to a certain Chinese encyclopedia entitled Celestial Emporium of Benevolent Knowledge. On those remote pages it is written that animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance".*[25]

7.3 Assignment at the point of creation or derivation

At what point should an identifier be issued (or "minted")? The answer is "the earliest possible point", but this is not always straightforward.

For example, suppose a creation (a product of some kind) passes through a supply chain. Parties ("nodes") in this chain are denoted A, B, etc., and an identifier assigned at point X will be denoted i_X. In a simple supply chain, the product made by the producer at A is supplied to a consumer at C via a retailer at B, and the chain can be summarised like this: A → B → C. If the consumer is to manage the entity down this chain and reliably reference it by a unique identifier, the identifier used in the chain must be the identifier i_A, which is assigned at point A, because

- the product at A is supplied unaltered to C, so the same thing is present at A, B and C,
- the same product might be supplied via an alternative retailer D: A → D →C , so B can have no role in determining identity, and
- C may wish to cross-reference material supplied by both B and D.

So all parties in this simple supply chain should use i_A. A is the **point of entry** to the supply chain, which may or may not be the same thing as the **point of creation**, but in any case it is the earliest possible opportunity in this model. In fact, i_A may even have been issued even before the creation of the product, if there is a workflow at A which requires identification at the beginning of the production process.

Another party in the chain may of course decide to use an additional identifier for their own internal processes, and store it as a "mapped identifier" to i_A. This brings risk, though, as such local identifiers may be mis-used and "leak out": for external interoperability this is not efficient and creates a potential source of delay or breakage in the chain.

A level of complexity occurs when B creates and supplies a version or **derivation** of what is provided by A (for example, a different format, or a product containing some additional feature added by B,

[25] J.L Borges: "The Analytical Language of John Wilkins". See also other Borges discussions of metadata quoted at www.tertius.co.ltd.uk, "Orbis Tertius".

© Linked Content Coalition 2014

such as language subtitles). Now, when is a derivation something new? What A considers to be "the same thing" is not what C may consider to be "the same thing"[26].

If it is determined that the version is indeed a new entity, this becomes a subset of a more general case: when the chain is not of simple supply but of re-purposing or modification: A → E → C, where a party or machine at E processes the entity from A to create a new entity which is supplied to C (for example, A creates an audiovisual work and E adds subtitles to it). It is clear that here both i_A and i_E are necessary, with E being the **point of derivation** since A may be supplying the work to others to provide similar opportunities for value creation, and it is also clear that the relationship of i_A to i_E must be declared. C may be purchasing other, different, works derived from A by another supplier F (for example, with subtitles in a different language). This requires i_F, and a declaration enabling a distinction from i_E to be ascertained. Two chains may wish to determine if they are supplying the same customer C; and so on.

The point at which a new identity, and therefore a new identifier, comes into existence will vary according to the type and nature of the entity, and will be a matter of agreement, policy or custom. For example, the point at which minor changes and corrections in the text of a book warrant a new edition will be a matter of editorial judgment. Up to that point, re-publication of the amended book may be with the same ISBN, and after that point there will be a new ISBN.

It is necessary to have distinct identifiers at the point at which a new entity is made available, and the relationship to the contributing entities in a new creation must be declared in a usable way. It should be possible to describe how two different entities derived from a common underlying component (e.g. i_E and i_F) relate to each other. In all of these cases, the principle that identifiers should be assigned at the earliest possible point (where the new entities come into existence or into the supply chain) minimises error and maximises efficiency.

This is still much less complex than reality. In a linked content world, simple linear chains as these simple examples will intersect and form branching and reuniting networks. Material produced by one party may be re-used by third parties without direct knowledge of the originator. Standardised principles enabling unique identification and data models enabling objective description of relationships are essential. Even this is a simplification and there will be complications due (for example) to legacy systems, or there will be multiple identification for legitimate reasons (such as privacy).

Such chains are multiplying and may be subject to constant change. Many products which were once viewed as fixed resources are now more likely to be continuous resources with frequent changes (news websites, for example).

[26] Those working with identifiers will be aware of the common reaction from users faced with such complexities: "but we all know what we mean by…". Indeed, you may know what you mean, and I may know what I mean – but it does not follow from this that you mean what I mean.

© Linked Content Coalition 2014

7.4 Co-reference

Co-reference is the term for 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 convergence of multiple ontologies and referencing schemes is key. However, this is hampered by the difficult problem of co-reference…"*[27]

Co-reference is unavoidable for many entities, where there is not a single, clear "early as possible" point of assignment of an identifier. For example, people and organizations ("parties") are identified in many different systems in different domains and for different purposes. It is not practical to impose a single global party identifier.

However, it is also unnecessary. Co-reference can be effectively managed through **mapped identifiers**, where any number of different identifiers for the same referent are associated directly with one another in "same as" relationships. For example, the International Standard Name Identifier for parties (ISNI) operates on precisely this basis. Some ISNIs may not be used publicly, but will act as central identifiers for the purpose of many-to-many mappings. Identifiers a, b and c may all be mapped to a single ISNI, and are therefore mapped as the "same as" one another and can be automatically substituted for one other to enable data to pass from one domain to another, where different identification systems are used.

It is not uncommon for two or more identifiers *of the same type* to be assigned to the same referent in error. To deal with such cases, the registry procedures must support the deprecation of one identifier by its "merging" with another. The **merged identifier** does not cease to exist, but remains "mapped" to the active identifier so that any usage of the "merged" identifier can be substituted for usage of the active one.

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[28]; ISO identifiers may be expressed as DOIs[29]. This confers the advantage of being able to embody equivalence, but is also open to the risk of embodying 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.

7.5 False co-reference

Common language provides, and sometimes makes assumptions about, context which needs to be carefully translated into precise entity attributes and relationships when referents are abstracted to be computed. Without the underlying analysis, one may not be identifying the correct referent.

---

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

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

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

© Linked Content Coalition 2014

A pitfall arises if a clear delineation is not made of the difference between a datum, the symbolic name of that datum, the address at which the datum is stored, and the symbolic name of that address[30]. It is not always necessary to separately distinguish these (each may be reified if necessary; this is an issue of functional granularity), but confusing them when they are distinguishable leads to problems. For example, an address may be reified, which requires identification (naming), so that the reification then has a name; a name is an address in a name space called "addresses"; and this is at the heart of much confusion over naming and addressing in discussions of URN, URL, URI and discussions of their usage as persistent "identifiers".

7.6 Ambiguous identifiers

Ambiguity is much more serious that co-reference, as it may be systematically unresolvable once it has occurred. An **ambiguous identifier** is one that has been wrongly applied to two or more different referents in error. This happens most commonly when it is not recognised at the time of assignment that two things which were thought identical are in fact different (for example, two different people called "John Smith"). It may also occur in error when a system re-issues an identifier to a new referent without recognising that it is already assigned to another. There is no easy answer to this: the trail of identification must be reviewed and the correct identifiers assigned where possible.

7.7 Functional granularity in creations

Functional granularity means that "it should be possible to identify an entity whenever it needs to be distinguished for some purpose". For example, a book may have a single ISBN, but if individual chapters or illustrations are extracted from it and published separately elsewhere, they may require distinct identifiers of their own; or when a new edition or translation is produced, it will require a new ISBN; or if a 30-second audio or video clip is taken from a longer recording for promotional purposes it will require its own identifier as it is deployed in different systems.

Functional granularity applies to the identification of any entity type, but it is especially significant in creations, and become more so as digital content can be fragmented, aggregated and transformed in an ever increasing number of ways, so the examples given here are all of creations.

There is a set of common relationships which may result in the need of new identifiers in relation to existing creations: **parts** (3.6.1), **aggregations** (3.6.2), **fragments** (3.6.3), **derivations** (3.6.4) and **manifestations** (3.6.5). There is no theoretical limit on the smallness of a fragment or the enormity of an aggregation which may be identified if there is a reason for doing so.

7.7.1 Parts

A part is a distinctly identified component of something. The entity may have been assembled from pre-existing parts (for example, an online "learning object" with a package of text, music, video and

[30] This issue was memorably dissected by Lewis Carroll (pseudonym of the logician C.L.Dodgson) in *The White Knight's Song* which appears in *Alice in Wonderland*: see the discussion of it by Eric Walker at http://www.alice-in-wonderland.net/?school/alice1020.html

© Linked Content Coalition 2014

images) or it may be created as a whole in which parts are later identified (for example, the chapters of a book). The parts may be identified, described and managed separately, and become parts of other compound objects.

7.7.2 Aggregations

An **aggregation** is created out of several pre-existing entities, perhaps with new material added. The aggregation can then be represented, in part, by the relationships that exist between the distinc t identifiers of its components.

7.7.3 Fragments

A **fragment** is a resource that is subordinate to (or "contained within") another, primary resource. The fragment (such as an arbitrary clump of text or a 15-second excerpt from the middle of an audio track) is not inherently a first class object but instead its identity is defined as a subset of the primary resource: of course this may be identified as a new first class entity if there is a functional need.

A problem raised by fragment identifiers is the existence of an infinite set of possible ad hoc identifiers from one base primary resource (for example, time ranges in a video). In many cases today "fragment" is used in one specific sense: in http to refer to the piece of a URL that the server doesn't really know about and that the client hangs on to and then processes returned html to do the right thing. This is a function of the hypertext model that was initially selected for http/html, which operates at the file level; so to get to some specific point within a file using http requires a second mechanism. In the internet, fragment identifiers are well understood in principle, but not uniformly dealt with.

An example from music well illustrates the application of functional granularity to fragments. In a European music collecting society the royalty distribution rules were biased in favour of the number of different works used as background to a television program, rather than simply the duration of music, and it was therefore more rewarding to have six "musical works" of ten seconds each rather than one work of sixty seconds. As a result the standard practice of composers and publishers was to register large numbers of very short works with titles like "Man goes into room" and "Man goes out of room" rather than a single work "Background music from Man In A Room".

7.7.4 Derivations

Derivations cover all situations where a new creation is made by adapting, in some way, an existing creation. This covers all forms of versions, adaptations, arrangements, translations, transliterations, mashups, remixes, photoshoppings, director's cuts, edits, editions, and so on (and on). In the digital network the opportunity for new types of derivation grows constantly, and the volume of derived content along with it.

The question of "when do I need to assign an identifier to a new version" is the classic case of functional granularity. In principle, any change in the attributes of a creation may result in the identification of a new derivation, but the point at which a creator or publisher declares a "new"

© Linked Content Coalition 2014

derived creation is arbitrary, and will commonly be a matter of convention: it is functional granularity at its purest – "I say something is a new version when I say it is".

Each application space defines its own rules, which may be guided by legal requirements or commercial opportunities. For example, laws or regulations may state that even an addition of a comma in a statement, even if it does not affect meaning, requires resubmission as a new item, whereas such an addition in a news article may be viewed as a proofreading correction that does not merit new identification. This necessarily requires not only tools for semantic interoperability but agreements on community interoperability. In scientific publishing, version control is significant for definitive attribution, so this community has agreed some rules for definitive version identification and tools to assist in their implementation[31].

An example from music will serve to illustrate the profound difference that a commercial opportunity may make to granularity. A music collecting society database contains hundreds of copyrighted "arrangements" of the Christmas carol "Stille Nacht/Silent Night" even though the musicological significance of the "arrangements" are generally trivial, because as the original work is long out of copyright, the original creator will not likely to dispute the claim. The same database contains only a single versions of Lennon & McCartney's "Yesterday", despite the fact that there are many musicologically distinct versions extant, for the simple reason that the original creators/owners will not recognise derivations. Using musicological, rather than commercial, criteria as a basis would result in a quite different pattern of functional granularity.

7.7.5 Manifestations and abstractions

The relation of manifestation to abstraction (or "work") is the most important and problematic when dealing with "compound creations". Music provides a clear model in the distinction between the underlying **work** (for example, a "song" like "Yesterday") and a **manifestation** of it in the form of a performance, recording or sheet music.

Works and manifestations have different rights, and often different rightsholders (for example, the composer and publisher of a work, and the performer and record company producing the sound recording). Although perhaps most clear in music, the manifestation/work split occurs in all types of content. Whenever something new is created, both a manifestation and a work will come into existence, with their accompanying rights. For example, a writer creating a poem is simultaneously creating an underlying work (the "poem")and the first manifestation (the document or "fixation" on the page or computer). Any number of new manifestations and performances of the poem may take place afterwards, but there remains a single abstracted "work" until such time as someone (for example, the poet) decides that sufficient change has happened to it to declare a new version (fort example, when it is translated in to another language).

The "parallel worlds" of perceivable manifestations and abstract works are increasingly reflected in identifier standards (for example, music with the ISRC and ISMN for manifestations, and the ISWC for works; text with the ISBN for books and the ISTC for textual works). Perhaps even more surprisingly,

[31] e.g. CrossMark: http://www.crossref.org/crossmark/

© Linked Content Coalition 2014

all Digital Object Identifiers (DOIs) issued to date have been to referents w hich are abstract works rather than digital manifestations. The maintenance of the fixed relationship between a manifestation and the work(s) it contains is in itself a critical registry function, which may be regarded as essential core metadata, but is as yet not commonplace in existing registries.

7.8 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 (that is, the number or code itself as opposed to the reference metadata) may be most efficiently carried out by a federated set of registrars.

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.

7.8.1 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.

There are many 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.

© Linked Content Coalition 2014

7.8.2 Avoiding co-reference in federated registries

A critical issue in the federation of the issuing of identifiers is to protect, as far as possible, against co-reference (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).

7.9 Core metadata

An identifier should be supported by metadata for discovery and disambiguation.

It is normally essential that when an identifier is minted, sufficient metadata is registered to enable the identifier to be discovered, and for the referent to be uniquely identified. The essential metadata ill normally including “management” metadata which will include:

- The date/time of issue of the identifier
- The registration authority or party minting the identifier

and some basic information about the referent:

- At least one name of the entity
- At least one type of the entity

For creations, typical core metadata will include:

- A title of the creation
- The creator(s) or major contributor(s) to the creation, with their roles
- The date/time of creation
- The basic structural type of the creation (abstract, physical, digital, spatio-temporal)
- The mode(s) in which the creation is intended to be perceived (audio, visual, taste, smell, touch)
- The character of the content of the creation (words, image, music, data etc)
- A form or format of the creation
- A genre of the content
- One or more measurements (for example, duration, size in MB)
- Links to other creations from which this one is derived if it is a version, excerpt or manifestation
- Annotations describing unique characteristics for the purpose of disambiguation.

Several of these categories (such as structural type + mode + character) may be combined in a “creation type” category (for example, “video clip”). However, such a list can never be exhaustive,

© Linked Content Coalition 2014

because each type of creation may have specific attributes which are vital for identity or disambiguation.

7.10 Persistence

Once assigned, an identifier should never be re-assigned to another referent. The attributes of the referent may change while its identity, and therefore its identifier, persists. For example, a human being may grow older or undergo a gender re-assignment, but its identity persists. Similarly, a collection of creations may be added to or reduced, but their identity persists, or a territory's borders may change, but its identity persists.

The core metadata may also be added to or corrected, where it is found to have been incorrect. Persistence should also be ensured through registry provisions for maintaining metadata after the original issuer or asserter is defunct or dead or otherwise unwilling to accept responsibility for it.

7.11 Identifying "Orphan" and other unadopted entities

There is a particular problem with issuing identifiers for, and then preserving the uniqueness of identity of, entities for whom no-one in the network has a primary reason for taking responsibility. In the content and rights network, this includes "orphan" and public domain works (those for whom there is no, or no known, rightsholder or responsible party) and parties who are deceased or defunct and whose works are out of copyright.

There are two opposite risks: that no-one will identify them, or that several people will, in incompatible ways. 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.

8 Deployment: how should an identifier be used?

An identifier should be accessible.

8.1 Four components: registry, resolution, repository

Four components are logically needed for deployment of first-class identifiers on digital networks: a registry of the identifiers (see 4.2); resolution mechanisms to link the identifier to some data (see 4.3-4.6); repositories where data may be found (see 4.7), and interoperability between different identification systems (see 4.8).

8.2 Registry

Registries are described above (see 3.1).

© Linked Content Coalition 2014

8.3 Resolution

An identifier should be resolvable.

Resolution is the process by which an identifier is the input request to a network service to receive in return a specific output of one or more pieces of current information related to the identified entity.

8.3.1 Basic resolution

In the digital content network, to promote linking and discovery of content, identifiers should be resolvable to a minimum set of publicly declared data. This does not of course mean that all metadata about and entity should be public, and does not preclude the development of added value services built on such identifiers which are marketed to communities of registered users, but it does recognise the limitations of all identifier registries.

8.3.2 Resolution and reference

Resolution and reference are not necessarily the same (though in some special cases might be). There is only one referent for a given unique identifier, but there may be several resolution results for that identifier, and these may change over time.

In the special case where the identifier referent is “this specific file on this specific server” then this may indeed resolve to the referent. Resolution of domain name based URLs to an IP address is an example (though even this is complicated in real implementations due to the possibility of re-direction, caching, proxy servers, aliases, etc.).

8.4 Persistent resolution

For interoperability, identifiers must be persistent, 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 largely “dumb” 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).

© Linked Content Coalition 2014

The key social infrastructure necessary to ensure persistence is a registry (see 3.1). 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[32] been substantially revised (notably those of ISO TC46/SC9) and provides a minimum set of requirements providing some reassurance to the user community that assigned identifiers will be maintained, but 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[33].

8.5 Federated 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, for example, was created for supply chain management and came into widespread use before ubiquitous network availability. The system for identifier 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 assignment. 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 assignment 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.

---

[32] 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)

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

© Linked Content Coalition 2014

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.

8.5.1 Approaches to federated resolution

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, (for example, providing data or services at a certain level of accuracy and timeliness) then 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.

8.5.2 Network architecture issues in federation

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**: 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. These

© Linked Content Coalition 2014

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.

An identifier should be capable of resolution to multiple locations.

8.6 Multiple resolution

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). *Multiple* resolution is the return as output of *several* pieces of current information related to an entity, such as at least one URL plus defined data structures providing additional data, options, or allowing management. Linking one referent to several outputs, each of which could be separately managed through an identifier resolution system (that is, additional items can be added or changed at the registry) has clear advantages, and provides an obvious way to move from a monovalent system (where an identifier can only resolve to one thing: a chain) to a multivalent system (where multiple linked connections can be made: a network).

At its simplest, single resolution could return a list from which to make a manual choice (e.g. "further information is available from the following sources…"). However, this is not a scalable solution for an automated environment, which needs to enable processing without intermediate human involvement. For machines to make sense of a complex response, the response has to be structured such that machines can distinguish the alternatives and act on them as appropriate. This is probably best done with multiple returned values as opposed to a single complex value. There is a need at this point for a common understanding of the values and their consistent representation, shared between (a) the parties and communities involved in creating those returned values; and (b) the parties and communities involved in creating the services that automatically interpret and act on those values. This is fundamentally no different than the case of single resolution which is acted upon automatically (for example, redirecting a single identifier to a single location such that the user just sees the end result) but is more complex, extensible to added services, more scalable, and requires greater coordination around standards and best practices. Including content negotiation as one of multiple values, for example, a single identifier resolving to both (a) a single content page for a human user and (b) a structured set of metadata for that content for further processing, is a current common example of multiple resolution.

© Linked Content Coalition 2014

If the various resolution options are defined as type:value pairs (e.g. “type = URL”: “value=www.acme.com/123”), then automated and extensible management becomes possible, by choosing a type or types on resolution. This requires a standard set of defined types (and an extensibility mechanism to build further types), such that users could (a) adopt an existing type to provide a stated function, rather than invent their own; (b) be able to interrogate a list of existing types to determine the properties of a defined type. This technical infrastructure could then be further facilitated by a set of tools allowing appropriate management within a community of the type:values (*Designated Authority* and *Appropriate Access*), so that the common understanding of the values and their consistent representation can easily be implemented. For example, a given registry could take control of all the type:value records for a given referent; or these could be delegated with specific separate permissions to separate managers; or intermediate models can be envisaged , such as a rights repository and a content repository collaborating such that one identifier infrastructure can be used to serve two problems (e.g. embedded identifiers in media to “indicate copyright” and “‘identify the work for other reasons”).

8.7 Repositories

A repository is one or more information structures providing definitive data associated with the referent, possibly including representations or instances of the referent.

8.7.1 Repositories and registries

Repositories are logically separable from registries, though some applications may in practice combine elements of both: for example, a registry of identifiers will typically carry core metadata describing the referent of each identifier, whereas a repository may carry much more diverse and extensive metadata or content.

8.7.2 Resolving to repositories

Identifiers resolve to one or more repositories. Different applications may require additional contextual data and so require the identifier to be used with a different repository. In linked content applications, metadata may be needed from multiple repositories to provide definitive contextual information: for example, product information (formats, price etc.) is typically not stored in the same repository as rights data (permissions, licenses, licensors, royalty agreements, etc.) and typically these sets of data are under different management or different policies (privacy, public or private access, etc.)

There is a rough analogy here with a traditional library: Index/Catalog/Stacks = Registry/Resolution/Repository. For example, the Registry is the way to find an identifier when you don’t have it in hand, like the “Readers Guide to Periodical Literature”. The analogy is loose, however, because of the confusion between architectural components and functions and the natural inclination to make them match one-to-one. Registries may be independent, or may be constructed to use a specific resolution method; resolution and registry tools may sometimes overlap (for

© Linked Content Coalition 2014

example, the next version of the identifier tool the Handle System[34] (v8) will add reverse look-up so that each Handle Service can act as a mini-registry, searchable by the values in the handle records). The Digital Object Architecture implements this logical model (see Appendix 2 for details).

8.8 Interoperability

At least three types of identifier interoperability may be distinguished[35]:

8.8.1 Syntactic interoperability

Syntactic interoperability refers to 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 relatively straightforward and trivial: the "bar code reader" level of interaction.

8.8.2 Semantic interoperability

Semantic interoperability refers to the ability of systems to determine whether two identifiers denote precisely the same referent, and if not, how the two referents are related.

Identifiers of the same entity should be interchangeable.

An important tool for interoperability within the digital network is the mapping of **alternative identifiers** for specific entities. This is a commonplace process in computer systems to enable data from one system (which uses identifier I_A to refer to entity A) to be integrated into another (which uses identifier I_B to refer to entity A). The system records that $I_{A=} I_{B,}$ and other alternative identifiers may be added to the set.

For example, a rights management organization may have its own internal identifier for a particular creation, and a rightsholder of the creation may have their own identifier which they have notified to the organization and by which they wish to have royalties reported. The two identifiers are "mapped" within the organization's system, and it is therefore able to report royalties to the rightsholder using the appropriate identifier.

In some cases a **registry of mapped identifiers** is created to enable many-to-many mapping of alternative identifiers (the ISNI for party identities is a prominent example). In such a registry there is one "hub" identifier for to which all other alternatives are mapped. The registry can then provide services which enable people or machines to look up and "translate" from any identifier into an identifier of another specified type, making them effectively interchangeable. The "hub" identifier (for example, the ISNI) may or not be an identifier that is publicly used: its primary purpose is to support the mapping so that people of machines can substitute one for another for any purpose.

[34] www.handle.net
[35] "Identifier Interoperability": http://www.doi.org/factsheets/Identifier_Interoper.html;

© Linked Content Coalition 2014

8.8.3 Community interoperability

Community interoperability refers to 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 share the same syntax and 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. However, practical use of resolvable identifiers requires that some minimal set of associated metadata should be generally available to facilitate third party use[36].

[36] For an example of this in practise see the DOI System concept of the DOI Kernel at http://www.doi.org/doi_handbook/4_Data_Model.html#4.3.1

© Linked Content Coalition 2014

Appendix 2

Identifier implementations in the digital content network

This appendix 2 to the LCC **Principles of Identification** document provides an overview of the main current implementations of identifiers relevant to linked content. It summarises the conceptual frameworks for considering identifiers in the digital network (section 1); internet use of identifiers (section 2); and major implementations of identifiers for specific entities or groups of entities (section 3). It is recommended that this Appendix 2 be read in conjunction with the underlying principles discussed in *Appendix 1: Identification in the digital network*.

1. Frameworks

1.1 Data models for context

An underlying ontology-based analysis is a pre-requisite to express the full dynamic, contextual, nature of managing content.[37] Many different data models have been developed in different content sectors, some deeper than others. Linking content may bring these different sectors into the same application; mapping to a common model, as done within LCC, is the only way to precisely determine if material defined under one data model is "the same" as that from a different model when linking material from different sources. Identifiers are required for each entity; necessarily, these models are infinitely extensible.

In the 1990s three frameworks emerged which have provided the analytical and practical basis for the main metadata developments for media and content which are currently implemented. These were the Functional Requirements for Bibliographic Records report (FRBR) in the library world; the indecs metadata framework (indecs) among media/content providers; and the CIDOC Reference Model (CRM) for museums and archives. The process of FRBRization is underway across the library world, promising greatly improved discovery and access to items and collections. The indecs framework underpins the multimedia content standards of ONIX (from the text publishing domain), DDEX (music), DOI , LCC, and LCC implementations such as RDI. CRM is being introduced in its domain.

Developed independently of one another, these three reached some strikingly similar conclusions:

- They recognise non-material entities as key in content management.
- All three start from an analysis of the process by which things come into being, rather than the things themselves (the "model of making", in indecs terms).
- FRBR and indecs share core terms such as "expression" and "manifestation".
- CRM and indecs share a detailed modelling of events, developed independently.
- Ontologically, all three agree on the priority of relationships as the basis for metadata.

[37] J. F. Sowa, "Knowledge Representation: Logical, Philosophical and Computational Foundations" Brooks/Cole, Pacific Grove, CA, 2000. http://www.jfsowa.com/krbook/index.htm

© Linked Content Coalition 2014

However, these models also have some important differences, not least in the specific meaning attached to the names of terms they employ. Each was informed by different functional requirements, and so has evolved different mechanisms for dealing with the issues that seemed most important to them. Broadly, they are compatible, and effective integration of metadata from schemes based on them should be achievable, but they must be handled with care; notably terms like *performance, manifestation, fixation, expression* and *work* need to be carefully mapped to ensure they are being used compatibly. The approach taken by LCC is derived from indecs but encompasses the wider mapping necessary to interoperate with other frameworks.

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, has since been used in a number of metadata activities. The indecs Metadata Framework document "Principles, model and data dictionary"[38] is a summary[39]. 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.

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*[40]: 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[41]);
- *URI principles* (see below under "Resolution"; also under potential review).

---

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

[39] In addition a useful brief independent overview can be found at http://www.slaw.ca/2014/02/24/applying-the-indecs-model-to-interoperability-of-legal-data/

[40] 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

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

© Linked Content Coalition 2014

- *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.

The indecs framework document notes that “It is rare that any of these is fully realised; but the extent to which they are realised largely determines the ultimate usefulness and resilience of any given metadata schema in terms of its effective interoperability with other domains.”

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 therefore a significant step in recognising the major improvements needed in the Digital Identifier Network[42] 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

[42] See the LCC Document "The Digital Identifier Network", published simultaneously with this document.

© Linked Content Coalition 2014

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 [43].

1.2 Digital Object Architecture

The Digital Object Architecture is an implementation of the three-component logical model for implementation of first-class identifiers on digital networks described in Appendix 1: a registry of identifiers; a resolution mechanism to link the identifier to some data; and repositories where data may be found. ITU standard ITU-T X.1255 "Framework for Discovery of Identity Management Information" (2013)[44] lays out an architecture, including types and type registries, as the underpinning of the 'Framework' as a citable technical standard. It also includes a number of useful definitions. While the Recommendation is focused specifically on identity management information, it is applicable more generally to many different types of information in digital form.

The Digital Object Architecture (DOA)[45] on which ITU ITU-T X.1255 is based is a logical evolution of the internet's fundamental architecture[46]. It is a framework combining resolution, registry and repositories in an integrated approach and tools; using persistent, globally unique identifiers, as provided by the Handle System, it offers enhanced flexibility in how objects are stored, moved, replicated, and referenced[47]. The DOA provides a mechanism for the creation of, and access to, digital objects as discrete data structures with unique, resolvable identifiers. These digital objects provide a foundation for representing and interacting with information on the Internet. CNRI make available implementations of DOA components for download, installation, and use by any organization or community, as an open-specification and software. The Handle System is an implementation of the DOA resolution component. It is used by DOI (ISO 26324). The DO Registry[48] enables users to provide their own metadata schemas, after which objects are registered with their metadata and that metadata is indexed and made searchable (each such DO Registry is, in effect, a specialized index over a collection of digital material in one or more repositories). Some (but not all) DOI applications also use the registry component[49]. The DOA is logically independent of the underlying "wiring" DNS but fully compatible with it (e.g. DOA resolution may be mapped to DNS via proxy servers).

In 2014, the Digital Object Architecture will reach a significant juncture with a change in the administration of one of its key components, the Global Handle Registry (GHR). CNRI has maintained control over the administration of the GHR since it was first made available in the Internet by CNRI in 1994. Plans are now well underway to transfer overall administration of the GHR to the DONA Foundation, a non-profit organization based in Geneva. The Foundation will be responsible for

---

[43] 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/)

[44] available free of charge at http://www.itu.int/rec/T-REC-X.1255-201309-I; ITU announcement: http://newslog.itu.int/archives/137

[45] www.cnri.reston.va.us/papers/OverviewDigitalObjectArchitecture.pdf

[46] Kahn, Robert E. "The Architectural Evolution of the Internet". Corporation for National Research Initiatives, November 17, 2010. hdl:4263537/5044
(= http://www.cnri.reston.va.us/papers/Architectural_Evolution_Internet_17Nov10.pdf)

[47] See e.g. "Digital Object Repository Server: A Component of the Digital Object Architecture": Sean Reilly & Robert Tupelo-Schneck. D-Lib Magazine, January/February 2010, Volume 16, Number 1/2
http://www.dlib.org/dlib/january10/reilly/01reilly.html

[48] www.doregistry.org

[49] "Using the DOI System with Digital Object Registry technologies": www.doi.org/doi_handbook/5_Applications.html#5.7

© Linked Content Coalition 2014

determining the set of system administrators, for digitally signing critical system information, and for establishing the overall policies and procedures governing the GHR's operation. Multiple independent parties, which are authorized and credentialed by the Foundation, will be responsible for the distributed operation of the GHR[50].

2. Internet use of identifiers

2.1 Assumptions

As far as possible the "set of requirements for identification to provide a uniform approach to accessing rights data" called for by LCC 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[51]. 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.).

2.2 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[52]. 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[53] [54]. Protocols other than http may become increasingly important through mobile devices, etc.: "On the internet, web pages

[50] Interview with Dr Robert Kahn: http://itu4u.wordpress.com/2014/01/06/lost-something-on-the-internet-never-again-with-new-digital-object-do-architecture/

[51] 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

[52] 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.

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

[54] 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

© Linked Content Coalition 2014

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"[55].

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 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[56]. 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[57]. 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.

2.3 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[58]) is the return as output of several pieces of current information related to an identified entity:

---

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

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

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

[58] 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).

© Linked Content Coalition 2014

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[59], 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.

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 DOI system

The Digital Object Identifier [DOI®] system[60] (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 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 100 million DOIs assigned by the end of 2014. The DOI system implements the Handle System[61] (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[62], (either by inclusion in DOI metadata and/or in a DOI syntax)[63], or DOIs may be assigned to entities which are not otherwise already identified. The DOI system complies with the proposed LCC specification.

2.5 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

[59] For an example using DOI, see http://www.doi.org/doi_handbook/5_Applications.html
[60] Digital Object Identifier system: www.doi.org
[61] 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.
[62] DOI System and the ISBN System: http://www.doi.org/factsheets/ISBN-A.html
[63] DOI System and Standard Identifier Schemes: http://www.doi.org/factsheets/DOIIdentifiers.html

© Linked Content Coalition 2014

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[64] 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[65] (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[66] 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[67]) for semantic web uses (RDF, some ontologies); therefore it is possible for any identifier to be cast as a URI, though whether this is useful will depend upon context of use.

2.6 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;

[64] http://www.iana.org/assignments/uri-schemes.html
[65] IETF RFC 4452: http://info-uri.info
[66] 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.
[67] IETF RFC 4151: http://www.rfc-editor.org/rfc/rfc4151.txt

© Linked Content Coalition 2014

- 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". [68]

2.7 Possible revision of URI specification

A post[69] to the W3C URI list by Larry Masinter (a long–term member of the W3C Technical Architecture Group and one of the co-authors of the URI syntax RFC 3986) proposed creating a new RFC that "obsoletes 3986 (URI) with a document that combined it with 3987 (IRI, Internationalized Resource Identifier, a generalization of URI allowing the use of Unicode), reverts to the "URL" name, and gave updated parsing advice"; he also posits the possibility of "removing any basis for support of using http URLs to "mean" abstractions or people", on the grounds that there is confusion over "whether *http://larry.masinter.net#the_person* could identify, locate, or name me rather than a paragraph of my home page"; and "including URN". It seems that the confusion between a referent and what an item resolves to is still not sufficiently appreciated. Any such URI re-definition is unlikely 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.

2.8 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

[68] 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

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

© Linked Content Coalition 2014

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[70].

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[71] at lists 40 registered NIDs; however many of these are not widely used as URNs, including some content identifiers (e.g., ISSN, ISBN). URN registration currently requires an additional layer of administration for defining a URN namespace (e.g. the string urn:doi:10.1000/1 rather than the simpler doi:10.1000/1) and redirection to access the resolution service.

2.9 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[72] and became IETF RFC 4452[73]. 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

[70] 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/

[71] http://www.iana.org/assignments/urn-namespaces

[72] NISO press release 28 Nov 2005 http://www.niso.org/news/pr/view?item_key=4b8a9e2d84fe28e5559d725eb6acd6fd9b1eb53d

[73] http://www.ietf.org/rfc/rfc4452.txt

© Linked Content Coalition 2014

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.

2.10 Non-ASCII characters, internationalisation, Unicode

The issue of non-English characters and special characters in identifiers is a complex one which can only be briefly summarised here. In theory (and ideally, from the point of view of local language use), identifiers could incorporate any printable characters from the Universal Character Set (UCS-2), of ISO/IEC 10646, which is the character set defined by Unicode v2.0. The UCS-2 character set encompasses most characters used in every major language written today. In practice, the treatment of non-standard characters across Internet applications varies: because of specific uses made of certain characters by some Internet technologies (the use of pointed brackets < > in xml for example), there are effective restrictions in day-to-day use and special encoding may be required, which cannot always be guaranteed to be understood. Despite the proposed development of Internationalised URIs (IRIs), in practice the use of foreign language symbols cannot be guatrnteed to be widely supported.

Even an apparently trivial issue such as case sensitivity is not simple: DNS is not, the rest of URLs may or may not be (this depends on the server), Unix and PC/Mac file names differ (Microsoft Windows in general is not case-sensitive, Unix operating systems are always case sensitive). Mark-up language tags, etc. can all cause unexpected problems and one cannot guarantee that any particular piece of software will respect case sensitivity and not conflate two identifiers intended to be different. Some search engines and directories are partially case sensitive. Different web browsers may differ in case sensitive handling (web browser developers have advised that "authors should not rely on case-sensitivity as a way of creating distinct identifiers, unless they are designing solely for a truly standards-compliant browser").

This argues in favour of case insensitivity and simple alphanumeric (ASICII) characters being the safer, and more robust, option for future evolution and development of identifiers on digital networks. Note that even then, traps remain, e.g. names with leading digits may cause problems in certain applications.

2.11 Fragment identification

A fragment identifier is a string that refers to a resource that is subordinate to another, *primary* resource. The fragment is not a first class object[74] but instead its identity is defined as a sub-set of the primary resource. A problem raised by fragment identifiers is the existence of an infinite set of possible ad hoc identifiers from one base primary resource (e.g., time ranges in a video). And of course for most people today "fragments" is used in one specific sense (http) - the piece of a URL

[74] First class = "one that has an identity independent of any other item".

© Linked Content Coalition 2014

that the server doesn't really know about and that the client hangs on to and then processes the html returned to get there or do the right thing (this is a function of the hypertext model that was initially selected for http/html – it is at the file level so to get to some specific point required a second mechanism). In the internet, fragment identifiers are well understood in principle, but not uniformly dealt with[75]: among proposals of particular interest are:

- IETF RFC 5147 *"URI Fragment Identifiers for the text/plain Media Type"*. http://www.rfc-archive.org/getrfc.php?rfc=5147 "This memo defines URI fragment identifiers for text/plain MIME entities. These fragment identifiers make it possible to refer to parts of a text/plain MIME entity, either identified by character position or range, or by line position or range. Fragment identifiers may also contain information for integrity checks to make them more robust". RFC 5147 proposes a fragment identifier for text/plain documents based on character and line positions and ranges within the document using the keywords "char" and "line": e.g. http://example.com/document.txt#line=10,20 identifies lines 11 through 20 of a text document. Hence it has more affordance[76] than the ISMC proposal, but is more limited as it deals only with text. RFC 5147[77] is therefore not identical in scope, but somewhat similar in concept to the idea of the ISMC.

- W3C has a draft specification for Media Fragments: http://www.w3.org/TR/media-frags/ – this is restricted in two senses: (1) it specifies only use of http; and (2) the specified addressing schemes apply mainly to audio and video resources - the spatial fragment addressing may also be used on images. The Media Fragments 1.0 specification, still a working draft, specifies the syntax for constructing media fragment URIs and how to handle them when used over the HTTP protocol. The syntax is based on the specification of particular field-value pairs that can be used in URI fragment and URI query requests to restrict a media resource to a certain fragment. Because of its restrictions, this W3C draft does not appear to be directly relevant to ISMC, but as it will no doubt be widely promoted it would be helpful to make clear the differences if ISMC goes forward.

- The Handle System deals with potentially infinite fragments by introducing a delimiter, with the base as a registered handle [an identifier of the primary resource], and defining a transformation on any possible tail. The *template handle* construction makes use of <template> tags in XML-structured handle values. When a server receives a resolution request for a handle which is not in its database, it determines if there is template for constructing the handle values; if so the server looks up the base handle (i.e. the part before the delimiter) and adds the part after the delimiter from the template XML <value> tags

[75] http://en.wikipedia.org/wiki/Fragment_identifier

[76] Affordance = "the ability to generate a syntactically correct identifier from content-in-hand".

[77] RFC 5147 is a "Standards track" RFC from April 2008, but as far as I can tell it's actually no more developed than an "informational" RFC and so has no particular special standing. Unlike ISO, the RFC process has many "standard track submissions" that are never taken further. I cannot find any evidence of RFC 5147 being adopted or supported. The RFC Standards track is not a particularly rational process: TCP/IP, for example, never was a standard and it is used trillions of times every day. RFC 5147 purports to update 2046, which is the MIME standard from 1996 and its still listed in Proposed Standards despite the fact that it is used in every http header every day.

© Linked Content Coalition 2014

defining the handle values of the result. Hence infinite fragments can be managed as they are created, through templates built on the primary resource. It is possible that MPR codes could be optionally managed "behind the scenes" in this way but it is probably not part of any standard.

2.12 Linked data

The adoption of URI in the LCC identifier specification conforms to the W3C Linked Data principles[78]. 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 in linked data by internet applications (first class in this case also means independent of any protocols used to resolve it). A list of registered infoURI schemes[79] 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[80], and his conclusion:

- "For physical objects, names are not unique because two different objects can have the same name.

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

© Linked Content Coalition 2014

- 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.
- 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[81].
- 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.

2.13 Identifier interoperability schemes

Several initiatives focusing on aspects of identifier interoperability are noted:

(1) The DOI System has a focus on ensuring interoperability both with other DOI applications and with non-DOI identifiers.[82]

(2) The 2011 *Den Haag Manifesto* on persistent identifiers (PIDs) and Linked Open Data (LOD)[83] 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

[81] 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.

[82] See DOI Handbook, 2.7 Relationship between the DOI system and other ISO identifier schemes and 2.8 Relationship between the DOI system and other (non-ISO) identifier schemes

[83] http://www.ncdd.nl/blog/?p=144

© Linked Content Coalition 2014

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

However, the content community sees a very high need for interoperability at the semantic and community level 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.

(3) 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. However 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.

(4) The Corporation for National Research Initiatives[84] (CNRI), developer of the Handle System, is developing an open source *Digital Object Based Interoperability Platform* (in collaboration with the Alfred P. Sloan Foundation[85]). 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.

The capability of resolving an identifier to more than one location or repository is gradually becoming recognised as an inevitable requirement of open linked data. There are work-arounds to this problem such as content negotiation on the web, but usually ad hoc per implementation; 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 (based on different metadata types) can be automatically located: work on this approach is under way under the auspices of the

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

© Linked Content Coalition 2014

Research Data Alliance and other efforts. Specific typing would enable a common resolution approach for specific applications, e.g. a type to openly make a "Digital Content Declaration".

2.14 Compliance tools

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 , 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 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. Entity 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 role as an agent in a Context.	*Tom Brown, 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*

© Linked Content Coalition 2014

		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. 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 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;*

© Linked Content Coalition 2014

		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

© Linked Content Coalition 2014

A main group of content identifiers comes from ISO, through ISO TC46/SC9 (Information and Documentation). The list of SC9 standards[86] 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[87]
- 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 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

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

[87] 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)

© Linked Content Coalition 2014

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)[88] 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[89] 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:

- 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.

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

[89] 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.

© Linked Content Coalition 2014

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[90].

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[91], 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[92] within the Rights Data Integration project[93].

3.2.3 Other (non ISO TC46) creation identifiers

The ARROW[94] 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[95] 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[96]), 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

---

[90] www.crossref.org
[91] www.useplus.com/
[92] www.cepic.org/
[93] www.cepic.org/tags/tags/rights_data_integration
[94] www.arrow-net.eu/
[95] 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
[96] www.eidr.org

© Linked Content Coalition 2014

- "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 managed in partial, unauthorised and often opaque[97] 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[98], a spin-off from the indecs project.
The identification of a Party has three common layers:

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

[97] 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.
[98] http://www.interparty.org/

© Linked Content Coalition 2014

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 identification according to the purposes to which data is being put[99]. 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[100], 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

[99] 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.
[100] http://www.ipisystem.org/

© Linked Content Coalition 2014

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[101] 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)[102] standard recently ratified was driven originally by the text publishing sector but backed by others including CISAC and the performers’ associations (the International Performers Database Association). ISNI was developed as a standard for a "name" identifier for public parties "involved throughout the media content industries in the creation, production, management, and content distribution chains". OCLC, the US not-for-profit library co-operative, is managing the global registry database, and there will be multiple registration agencies. To date there are two (Bowker and Ringgold) who are respectively dealing with creators (predominantly in the text domain) and institutions. Both are just getting going. ISNI is focussed on identifying creators, not rightsholders:

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

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

ISNI has particular issues with verification and duplication. Unlike the IPI code, ISNIs will not be registered by a single method, pre-validated and de-duplicated by unique society membership criteria. Any organisation can, in effect, apply for ISNIs for any parties in which it has an interest – for example, a publisher or society registering all its authors. Data quality management and de-duplication is therefore a critical issue. ISNI is tackling this by having a single global database at OCLC,

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

© Linked Content Coalition 2014

and building its initial database substantially from library authority records from the VIAF (Virtual International Authority File)[103] 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[104] "to develop a robust, scalable, and interoperable standard for identifying a core entity in any information management or sharing transaction-the institution. The I2 Working Group did extensive community needs assessment with the publishing, library and repository use sectors". With the emergence of ISNI, NISO reached an agreement to use ISNI for institutional identification, and I2 contributed further recommendations to the ISNI-IA that were incorporated into the ISNI standard. The I2 Working Group is now "finalizing a Recommended Practice, expected to be published in the next few months. This document will provide information on a profile that can be used by appropriate Registration Agencies to apply ISNI to institutions". It remains to be seen how well this proposed profile fits into the bigger picture, but the fact that I2 teamed up with ISNI rather than creating yet another standard is commendable.

3.3.5 ORCID

ORCID, the Open Researcher and Contributor ID initiative, was established in 2010 and launched its service in October 2012[105]: "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 agreed that they are complementary, and further discussion of common aims is understood to be under way. ORCID is a specialized ID and may be mapped to ISNIs like other sectoral Party IDs.

---

[103] www.viaf.org
[104] http://www.niso.org/publications/newsline/2012/wgconnectionoct2012.html#bi2
[105] http://about.orcid.org/news/2012/10/16/orcid-launches-registry

© Linked Content Coalition 2014

3.3.6 Legal Entity Identifier

ISO 17442 Financial Services – Legal Entity Identifier is to be launched in 2013[106] 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.

There are some pockets of potentially re-usable identifiers in specific sectors: for example, IMDB[107] 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"[108]. The specification of WebID has been worked on since 2005, the latest specification being 2011[109]. 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[110]), 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)

[106] http://www.financialstabilityboard.org/publications/r_121024.pdf
[107] http://www.imdb.com/
[108] http://www.w3.org/wiki/WebID
[109] http://www.w3.org/2005/Incubator/webid/spec/
[110] http://www.foaf-project.org/ : "FOAF defines an open, decentralized technology for connecting social Web sites, and the people they describe"

© Linked Content Coalition 2014

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:
    - 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[111].
- The Global Location Number (GLN) is part of the GS1[112] 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

[111] 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
[112] http://www.gs1.org/

© Linked Content Coalition 2014

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[113]

  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)
- 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[114] 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[115].

[113] http://www.freestd.us/soft/339586.htm
[114] http://movielabs.com/md/avails/
[115] http://www.movielabs.com/crmd/

© Linked Content Coalition 2014

In the music sector, the DDEX consortium[116] 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[117] was outlined in EU Council Document no. 17554/11 (metadata describing the document was posted on the EU official document register, but the full text of the document itself was not made public). Our understanding is that this will be used to identify laws, which 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"* [118] 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.

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

© Linked Content Coalition 2014

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.[119]

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[120]), 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[121] 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.

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[122].

---

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

[120] 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).

[121] 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.

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

© Linked Content Coalition 2014

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 map and transform controlled vocabularies whether for commercial use or otherwise; and to inform the content of controlled vocabularies.[123]

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

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

© Linked Content Coalition 2014

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).

© Linked Content Coalition 2014