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

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.

- 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

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.