E-Citations : actionable identifiers and scholarly referencing

Norman Paskin
Director, The International DOI Foundation
(n.paskin@doi.org)

This document discusses the role of "actionable" identifiers such as the Digital Object Identifier (DOI) in enabling scholarly citations in a digital environment. Citation is a sub-set of the general wider concept of linkage, but an interesting one for two reasons: it is a practical example being worked on today, and it demonstrates that linkage only between digital entities is insufficient for scholarly citation. It also points the way to future extensions of the concept of linkage, with new uses being made of Internet resources, and tools which can support these new patterns of citation.

1.   Science needs persistent links

The scientific literature represents the "minutes of science", the intellectual heritage of the research process. It is accessible in a reliable repeatable way: scientific communication is founded on dependable links between articles (i.e. references to earlier work). When I access a link, I must get the "same thing" as when anyone else now does, or did in the past. These links, called citations, are unique strings that unambiguously identify another article. These unique strings are identifiers; they may not look like numeric identifiers, and there is no single, uniformly accepted way of citing (no standard "syntax" for such identifiers), but humans can interpret the various formats of "bibliographic information" in common use to find the unique reference.

Note that this process is historical, and not dependent on "cyberspace". The current linkage system has been in operation for many decades, and is implemented on an infrastructure of paper and social institutions (libraries, publishers, archives, and abstracting and indexing services).

It would be advantageous to have the "minutes of science" accessible electronically. Building a system that embodies these citations using unique persistent identifiers, actionable on the Web, would be an excellent implementation of the idea of "Uniform Resource Identifiers" at the basis of the conception of the World Wide Web[1]. I will refer to the citation links that would be implemented in such an electronic environment as E-citations[a].

2.   Articles and "papers"

Scientific articles are also called "papers" - for good reason. The name indicates the historical close tie between the intangible[b] work embodied in the article and its physical manifestation (a set of printed pages). Many of our ideas about citations were formed in "paper space" as

---

[a] A term suggested by Joel Baron.
[b] The word abstract is often used for this concept, in the sense of "existing as a mental concept; opposite to *concrete*". However I've avoided that word here because of the risk of confusion with abstract in the sense of "a short summary of an article" (which is of course also important in scientific publication).

opposed to cyberspace. Citations are traditionally to another physical manifestation (journal, volume, issue) - the syntax of the identifier is semantically grounded in the print model. Historically there was no need to distinguish the paper from the intangible work: one was inseparable from the other; citing the paper (physical form) was the same as citing the work (intangible).

The distinction between the intangible work and a manifestation of it may seem arcane: but the advent of digital manifestations (electronic publishing) in addition to paper journals made an operational functional distinction obvious. A publisher has a production line, editing articles for publication, and the publisher will use a number to identify the entity he is processing. In print-only days, that process resulted in a single published entity (a printed manifestation). With digital publishing, the production line will bifurcate at the end of the editorial process to produce two (or more) entities (e.g. a printed article, and an on-line PDF or HTML file). Those two entities are related: they are "the same article". The number that is on the production line item and carries forward into both the published entities (telling that they have something in common) is the identifier of the "intangible work". (This was the origin and intent of the Publisher Item Identifier[2]). A more formal analysis (such as that of INDECS[3]) would say the Work is an abstract Creation, made of concepts, whilst the published entities are examples of physical Creations (manifestations) and are made of atoms and/or bits. (Separately, the music world has dealt with these issues for many years: Beethoven's 5th symphony as an intangible work, manifested in manuscripts, performances, recordings, etc.).

So we now need to think not of one space (the physical, paper space) but of three "spaces" in which "the same" articles appear:
• Information space = the work as intangible entity (ideas)
• Cyberspace = digital manifestation (electronic, made of bits)
• "Paper space" = physical manifestation (cellulose and ink, made of atoms)

If we had started with a clean sheet, knowing this in advance, we would probably be using a naming scheme that centred on the intangible work, and then cascaded down to paper and digital manifestation systems. But of course we started from paper only; our citation model is to journal, issue, page.

Some attempts have been made to create digital manifestations using file names that mimic the paper paradigm. However, if this is done without recognising the common link - the intangible entity- this creates two worlds that don't interconnect. Analysis of reference linking shows that the citation to the intangible work is a key component in the process of moving from an instance of a citing document to an instance of the cited document.[4]

3.   Citations need to recognise intangible, digital and paper space

For citation purposes, a scientist wants all references to his "work" to be counted irrespective of manifestation format. It is also essential to the scientific record that a paper published simultaneously on paper and electronically is "counted" as one single contribution to science, not two. What is important here is the intangible work (and hence, an identifier of it), which identifies the work irrespective of its various manifestations (and their locations).

Identifiers that only deal with the "cyberspace" world won't solve the problem of citations (i.e. the creation of a seamless means of navigating citation links) unless every single printed document of the existing literature of science is digitised and every reference therein is converted to the corresponding digital manifestation pointer. This seems unlikely to happen immediately, particularly for material that has already appeared in printed form. What is needed is a common scheme that can deal with entities of all forms (intangible, digital, and

physical): a framework able to identify, distinguish and relate all of these manifestations and the corresponding intangible work. Sometimes we need to identify all manifestations as "the same thing" (for citations); yet for other purposes we need to be able to distinguish manifestations (typically, to access a paper copy via a library or a digital copy via the Web). This is an example of the problem of "When are things different from each other?" Of course, any two items differ in some way (or they would not be two items). But it is useful to group items as "the same" for some purposes, at some times: for example, the ISBN of a book groups all copies of the same printed edition as one "book" title[c].

Since sometimes we need to group things in one way, and at other times in another way, we cannot use just one simple number. Instead we must use a network of identifiers; or supporting metadata about the entity (well-formed metadata is no more than a network of identifiers, of metadata entities).

4.   Identifiers in cyberspace

Identifiers can be used to link an identified entity to its location (e.g. on the Web). However, a problem with using identifiers in cyberspace is ensuring that they are persistent (persistence being an absolute requirement for E-citations). Two general approaches are presently being suggested to making identifiers persistent on the Internet:

(a) Don't change the URL. Design the URL to be an unchanging persistent label for the resource at a maintained location.
(b) Assign a name to the entity, and use redirection: accept that URLs may change, so assign a name ("URN") (which does not change) and a mechanism of resolving this to the URL (purl.org and redirect.net are examples of such services).

There are ongoing debates about which approach - (a) or (b) - is best. A scientist doesn't care which of these methods underlies any system he uses, provided it works. What is inhibiting this from happening is that there are currently problems with both methods, so *neither* works.

A problem with (a) is that the genie is out of bottle: people do change URLs, and there may be legitimate reasons for such changes (e.g., the sale of a journal from one publisher to another; archiving; etc.). In addition, there may legitimately be multiple URLs corresponding to one entity (discussed further below). A problem with (b) is that book-marking in browsers uses not the reference "mid point" (the name) but the end point - the URL one ends at; when this changes, the book-marked link is broken. (Neither of these exhausts the lists of difficulties with each approach, but they are sufficient to make either unusable for E-citations).

5.   Technology can help

It has been correctly stated that persistence is a social issue, not a technology issue. If people did not change URLs, solution (a) would work. But it is probably difficult to enforce this now - and there would be other problems resulting (as discussed in the next section: multiple instances). Technology can however help us fix our social issue by solving the problem of (b); it's possible to devise technical schemes that don't have this problem. To be usable the technical fix must build on what we have now, not be a theoretical architecture requiring replacement of existing infrastructures.

---

[c] This is an oversimplification, even though it serves to illustrate the concept; a single book may have multiple editions, and hence multiple ISBNs (e.g. hardback and paperback editions)

The Crossref Consortium of several major scientific publishers[5], using some of the techniques developed earlier in a prototype by the International DOI Foundation[6] membership, has proposed the use of the DOI as an actionable identifier, as an implementation of mechanism (b) described above. This assigns an identifier to the "intangible Work" entity, and uses resolution to locate a manifestation instance. It thus uses the concept of an "actionable identifier" (a managed name allocation process, an Internet resolution scheme, and accessible metadata about the entity) to provide persistent single redirection. The potential for DOI technology to use multiple resolution (which links a DOI to many potential resolution points), adding other value-added features, provides an extensible scheme which could be used to address some of the related issues such as multiple instances and version control that are described in the remainder of this paper.

6.   Multiple Instances

We recognise that it can be useful to group items as "the same" for some purposes, at some times. For example, all copies of the printed US edition of "Weaving the Web" (many thousands) are treated for some purposes as one "Book ". When I refer to Tim Berners-Lee 's book, page 43, I refer to the set of all copies. But when I actually want to read it, I need to get a particular instance (a copy); for example the copy bearing my own name and signed by the author is a specific instance of the edition.

In the digital world there are already many cases where the same article is legitimately available from more than one content provider (i.e. at more than one location, as, for example, at mirror sites). That is exactly analogous to the case of multiple copies of a book. I may find a citation (to an intangible work, or set of all manifestations) and need to find one specific instance among many possible ones. There may be many copies for legitimate reasons[7]; yet all are the "same".

The general approach of linking through "naming services" as in (b) above (each article is given a name - some form of identifier - which can be converted into a URL at the time of use through a resolution service which maintains a "database" (or more generally, a repository) of identifier-to-URL relationships) adds a potentially valuable feature. If the database holds one-to-many relationships, it can embody the relationship of one class to many instances. The identifier of the entity could then be resolved to all locations; or to one, chosen from the available options. The fact that URLs are unable to deal (alone) with the necessary expression of relationships and multiple instances is not then a problem. This is the approach taken in the Digital Library Architecture[8]. Other approaches, compatible with the identifier approach but adding specific software tools, have also been announced recently (e.g. LinkBaton from Openly[9], and the SFX tools now being developed by ExLibris[10]).

The use of schemes that recognise instances and classes in a data model is a commonplace among one community which has to deal with these issues in practical terms: the library community. Recently the IFLA study on Functional Requirements of Bibliographic Records[11] has examined the underlying principles, an approach taken further in the recent INDECS[3] activity.

7.   Metadata framework

Since we cannot use just one simple number to express all possible uses, we must use a network of identifiers, with supporting metadata about the entity. Metadata to be used by computers must be unambiguous, or "well-formed"; that is, each element of metadata is itself an identified entity. Hence a network of identifiers and metadata is in fact simply a network

of identifiers of entities, some of which relate to other entities by having a specific relationship with them.

This approach is exemplified in the INDECS activity (though acceptance of one specific approach is not necessary for acceptance of the general principle). A set of metadata about each identified entity, available to all users and conforming to an overall generic interoperable metadata framework, enables applications that offer functionality beyond a simple persistent identifier. Metadata about the identified entity can interoperate with metadata from other sources (e.g. about context) to construct services and transactions.

There are many approaches to metadata sets for articles (or elements of the scientific record). Recent proposals include a descriptive set for preprint publications (the Open Archives Metadata Set)[12], a descriptive set for version control for an article[13], and a set to describe any published article (devised for use with DOI)[14]. If these are mapped to an overall conceptual model, interoperability can be achieved among community-specific metadata vocabularies (an aim of the Harmony project[15] as well as of INDECS). It seems essential to use such a mapping scheme to avoid the generation of a Tower of Babel from these differing proposals, especially if one takes into account the possibilities of linking to non-document types. There is a clear task here for industry bodies and communities to agree guidelines and to liaise with the bodies maintaining those data dictionaries and conceptual models which gain widespread acceptance.

8.   Versions and naming (granularity)

Another, but related, practical problem has become acute with the advent of electronic publishing. This is the question of how to name different "versions" of an entity.

- When papers are published digitally, it becomes easy to create updated versions (amended, corrected, additional information, etc.).
- When papers are published simultaneously on paper and digitally, the two may differ. The digital version may have additional supporting material in the form of three-dimensional images or videos, active links, etc. that cannot be accommodated in the conventional two-dimensional paper version.

What criteria do we use to indicate that two entities are no longer "the same paper", but that one is a new distinctly different version of the old? The term "version" is loosely defined. What counts here are not differences in manifestation appearance (rendering or format); we have no problem saying that a PDF file and an HTML file of a paper are "the same paper", even though a bit-by-bit comparison would show marked differences. What is important for the scientific record is whether one thing differs from another in terms of the intellectual content, and if so, what is the relationship between them. For example, does one (the conventional printed version) contain a subset of the information contained within the other (an electronic version with additional supporting data)?

There is no automated procedure that could apply here to determine distinctness, other than the simple insistence that any substantive change requires that a new version must be declared. But what constitutes a 'substantive change'? Does the addition of a comma, or the correction of a typographical error, creates a distinct new version? The affirmative answer is the approach taken in some jurisdictions for legal documents, yet this pedantic approach seems unhelpful for practical scientific record purposes. It could work, if a highly automated and rigorous approach was taken, in which every version was related to earlier versions by means of relationships (subset of, corrected version of, new edition of...). Yet a sensible scientific author (and publisher) would recognise that there is a difference in kind between the

correction of a missing comma for stylistic purposes only, and the correction of "g." to "mg." in a clinical dosage specification, though both are single character changes.

This issue of "granularity" (at what level of detail do we distinguish?) is a matter of social judgement, not technology. INDECS has coined the useful phrase of *functional granularity* as one of its guiding principles: "*an entity needs to be identified only when there is a reason to distinguish it*". Book publishers will be familiar with this form of judgement in everyday practice: the distinctions between a "reprint", a "new edition", a "corrected reprint", a "revised reprint", etc. are not hard and fast, but matters of tradecraft judgement (often influenced by marketing considerations as much as scholarly ones). Digital technology cannot expect to re-invent 500 years of print experience and social constructs overnight. However, what might be very useful is for agreement to be reached on some practical conventions for best practice in identifying versions. A start has been made in recent STM/ICSU discussions that attempt to reach consensus on what constitutes publication in terms of preprints and "published" formats. This could be extended to other granularity issues.

9.   Future extensions of citation behaviour

With the rise of digital publishing, we are seeing new opportunities for linkage and hence for citation. Future scientific communications may well add other aspects of citation, using other tools to record "the minutes of science". Links may no longer be solely to printed text documents or to their digital equivalents, but e.g. to embedded multimedia[16]. We have seen the start of this changed behaviour in the development of electronic "preprints" (such as the Los Alamos "xxx" service) and their integration into the sociology of scientific communication. Early attempts by publishers to cope with preprints have seen them treated as essentially "versions" of a later paper (see above). However other linkage forms may require entirely new forms of citation. For example, it is presently possible to link from a paper to a database of scientific images, where the image may be dynamically chosen as to resolution, technical specification, source, etc., as is proposed in the BioImage Database Project[17], a European initiative for a database of multidimensional biological images. BioImage is working with the International DOI Foundation and INDECS to see how DOIs and metadata can best be used for such linkages.   Similar links made to other databases (e.g. of biological sequence or macromolecular structures), or to software tools used to expedite scientific research, seem inevitable in future forms of reporting.

The concept of such linkages between resources on the Internet (of which E-citations are a small part) is central to the conception of the Web. The World Wide Web Consortium (W3C) is putting much effort into architectures and tools that enable this to happen, and into resources that enable computers rather than humans to determine the links. The Resource Description Framework (RDF) provides the focus for W3C's work on metadata.   RDF's notion of a resource is a URI-identifiable entity . The URI specification[18] provides a broad definition of resource as "anything that has identity". Therefore RDF description services (a concept now being discussed, but not yet finalised) could be used to expose descriptions of *any* entities (objects, resources, things) that can be identified using URIs. Familiar examples include an electronic document, an image, a service (e.g., "today's weather report for Los Angeles"), or a collection of other resources. Not all resources are "network retrievable", since human beings, corporations, and bound books in a library can also be considered as resources[d].   However, from the point of view of citations, such definitions seem to be too wide ranging to be immediately meaningful. In contrast, the DOI specifically limits itself to dealing with items of intellectual property. Admittedly this is a huge universe, and the definition could (and probably will) lead to DOIs for services, although the initial applications are to static entities.

---

[d] URI specification section 1.1

As the identifiers used in such linkages are themselves used to create new services (from simple look up of core metadata, to more sophisticated services created from context-sensitive metadata about users and transactions as well as content), additional functionality will be required of them. The "actionable identifier" (the combination of managed naming, resolution service, and accessible metadata) can deliver a wider range of actions if options or alternative choices can be built into applications. One way of easily doing this is by multiple resolution (as proven in the Handle[19] technology underlying the DOI). This allows the identifier of an entity to be associated (in the resolution system) with many pieces of "state" data about the entity. Those data types could be things like locations (URL), or pointers to sources of associated data or specified services. One or more of these data types could be the location of specific metadata sets, rights statements, etc., to form the basis of specific services associated with the DOI; therefore the actions associated with the "actionable identifier" can extended to many different services.

It's important to note that such additional data, associated with an identified entity via the resolution system, need not be determined by the DOI assigner. Instead they could arise from a different source. In fact, a metadata record could reside in a completely separate location from an instance of the entity itself, and the resolution mechanism would serve as a pointer to it; this is entirely in conformance with the description of the RDF model and its syntax specification[20][e].

10.   Recommended actions

Given that much of the intellectual and technical framework for dealing with actionable identifiers as citations is still developing, what can we do now to aid progress? Some actions relate to functional design, and some to technical implementation.

It would be useful if the scientific community (in the widest sense, that is including those involved in information dissemination) produced some provisional guidelines on scientific reporting; as well as the current discussion of what constitutes a publication, these should be extended to cover when two reports should be distinguished from each other (preprints, versions, etc., i.e. the issue of functional granularity). Guidelines would also be helpful in avoiding a myriad of different metadata sets for different purposes, by agreeing a common data model from which specific (yet interoperable) sets could be derived. These guidelines would start from the functional requirements of science, but make use of metadata activities such as the INDECS framework to express the precise distinctions agreed upon.

Technical implementations of any identifier system used with citations should strive to follow what we might call the Hippocratic oath; "first do no harm": wherever possible, they should allow future extensibility and, by design, enforce the minimal possible proscription on future

---

[e] Appendix B of the RDF model and syntax document notes the following:
"Descriptions may be associated with the resource they describe in one of four ways:
1. The Description may be contained within the resource ("embedded"; e.g. in HTML).
2. The Description may be external to the resource but supplied by the transfer mechanism in the same retrieval transaction as that which returns the resource ("along-with"; e.g. with HTTP GET or HEAD).
3. The Description may be retrieved independently from the resource, including from a different source ("service bureau"; e.g. using HTTP GET).
4. The Description may contain the resource ("wrapped"; e.g. RDF itself).
All resources will not support all association methods; in particular, many kinds of resources will not support embedding and only certain kinds of resources may be wrapped".

development.   Even if the intent of the application is not to embrace all the concepts described here, they should recognise that such future extensions may be built on the foundations they develop.   For example, the Crossref consortium is unlikely to produce in its first release a mechanism which deals completely with appropriate copy identification, future citation forms, and so on; but by using DOIs with defined genres (metadata elements), it can make use of an interoperable and extensible metadata framework which allows such further development. In this way the initiative is unlikely to find that at some future point earlier decisions have precluded the extension to new, as yet undefined, forms.   Similarly, technical solutions for resolution, metadata declaration etc should use open standards, as far as possible without having to wait for conclusions from the various IETF and W3C discussions of the precise nature and characteristics of URNs and URIs, or RDF schemata.

Acknowledgements

I thank Joel Baron (formerly of the New England Journal of Medicine) and Professor David Shotton (University of Oxford) for helpful comments made after reviewing earlier versions of this document.

Comments and criticisms should be sent to the author (n.paskin@doi.org)

Note

This paper was first published on the web site of the International DOI Foundation (www.doi.org)

© 2000   The International DOI Foundation, Inc.

References

[1] Berners-Lee, T et al: World Wide Web: The Information Universe.
(in: Electronic Networking: Research, Applications and Policy, Vol 1 No2 (1992)
Meckler, CN, USA)
(pdf copy at: http://www.w3.org/History/1992/ENRAP/Article_9202.pdf)

[2] Paskin, N: Information Identifiers.
Learned Publishing Vol 10 No 2 pp 135-156 (April 1997)
(available at www.elsevier.nl/homepage/about/infoident/)

[3] Interoperability of Data in E-Commerce Systems.   http://www.indecs.org

[4] Reference Linking for Journal Articles. Priscilla Caplan & William Y. Arms
D-Lib magazine, volume 5 No 7/8 (July/Aug 1999)
http://www.dlib.org/dlib/july99/caplan/07caplan.html

[5] Crossref www.crossref.org

[6] Digital Object Identifier www.doi.org

[7] P. Caplan & D. Flecker: Choosing the Appropriate Copy.
Digital Library Federation Architecture Committee. http://www.niso.org/DLFarch

[8] http://www.xiwt.org/documents/ManagAccess.html

[9] http://www.openly.com

[10] http://www.exlibris-usa.com

[11] IFLA Study: Functional Requirements of Bibliographic Records.
http://www.ifla.org/VII/s13/projects.htm

[12] Open Archives Metadata Set.
http://www.openarchives.org/sfc/sfc_oams.htm

[13] AAAS/ICSU Press Working Group: Defining and Certifying Electronic Publication in Science.
(proposal to STM Annual meeting, Frankfurt, 12 October 1999); (available at http://www.stm-assoc.org)

[14] http://meta.doi.org/doi-x-reflinks_v1-0.PDF

[15] http://www.ilrt.bris.ac.uk/discovery/harmony/index.htm

[16] Mckiernan, G. "Embedded Multimedia in Electronic Journals"
Multimedia Information and Technology, vol 25 (4), pp. 338-343 (1999)

[17] www.bioimage.org

[18] http://www.ics.uci.edu/pub/ietf/uri/rfc2396.txt

[19] www.handle.net

[20] RDF Model and Syntax specification. http://www.w3.org/TR/REC-rdf-syntax