

ICSTI Seminar: Digital Preservation of the Record of Science (Feb 14/15 2002), to be published by IOS Press

Digital Object Identifiers

**Dr Norman Paskin
The International DOI Foundation
P O Box 233
Kidlington
Oxford, OX5 1XU
UK**

**Tel: +44 (0) 1865 843798
Fax: +44 (0) 1865 843446
Email: n.paskin@doi.org**

© IDF 2002

Introduction

My thanks to Kurt Molholm and his ICSTI colleagues for the invitation to present DOI at a meeting on the digital preservation of the record of science. Whilst I am not an expert on the topics of archiving and preservation, I believe that many of the issues which we have encountered in developing the Digital Object Identifier (DOI) system are common to preservation discussions; and many of the communities represented at this conference, such as the STM publishing community, the scientific communities, and the metadata communities, have been very active in collaborating with us. Issues of persistence and the solutions to it - such as separation of the problem into components, and the recognition that persistence is fundamentally a *social infrastructure* rather than a *technology infrastructure* problem - have been fundamental design principles of information identifiers that influenced DOI.¹ Equally, the importance of unique identification has been recognised by the digital preservation community²

1 Interoperability

The connection between preservation and DOI work lies in interoperability. When we consider how we might ensure the digital preservation of the record of science, one logical approach is to create a central, single, authoritative place of deposit and management of such data; a new Library of Alexandria. For the second time in history, people are laying plans to collect all information; the move to digital technologies for storage, access, and co-ordination of distributed efforts makes this possible.³ Yet there is one fundamental difference in this modern approach: a virtual library implies that rather than one source and mechanism, there will be many. Therefore, the various components contributing to, or using, this archive must interact with each other in a structured way: they must interoperate.

In addition to the spectrum of social interoperability considerations⁴, there are at least six different types of technical interoperability:

- Across media (such as books, serials, audio, audiovisual, software, abstract works, visual material)
- Across functions (such as cataloguing, discovery, workflow and rights management)
- Across levels of metadata (from simple to complex)
- Across linguistic and semantic barriers
- Across territorial barriers
- Across technology platforms

Interoperability in the face of legitimate change has been a theme of the DOI work. The problem of preservation is when the dimension of change is time: "how do we interoperate with the future?"

2 DOI, Handle and indecs

2.1 Overview of DOI

DOI is a tool for naming "content objects" as first class objects in their own right, with a mechanism to make these names actionable through "resolution". In this way DOI offers persistent managed identification for any entity. But that alone is not enough: managing resources interoperably requires appropriate metadata. Creating a mechanism to provide a description of what is identified in a structured way allows services about the object to be built for any purpose. The IDF has outlined, and is actively developing in more detail, a standard way of not only doing this, but also mapping to existing metadata standards such as ONIX⁵ (for product information), Dublin Core⁶ (for resource discovery) and so on, allowing each community to bring its own identifiers, descriptions and purposes into play. Finally, wrapping these tools into a social and policy framework, through the DOI Registration Agency federation, allows the development of DOIs in a consistent quality-assured way across many sectors, opening the possibility of managing multimedia objects seamlessly.

It is not the aim of this presentation to give details of the DOI system: these may be found elsewhere⁷. However, in order to understand the key issues of interoperability relevant to preservation, it is necessary to understand the principles which DOI is built on: (1) the level of indirection (separating the name from the particular instance addressed) offered by *resolution*; and (2) the use of well-formed *metadata* to describe the objects identified and so offer appropriate hooks for third-party services about these objects to be constructed in a reliable interoperable manner.

DOI has used as reference principles and implementations the Handle System⁸ for resolution and the indecs framework⁹ for metadata. It is possible that other approaches could be substituted for these but this would not alter the fundamental concepts.

2.2 Resolution

A DOI is a name (identifier) for an entity in a network environment. Entities identified by a DOI may be of any form, including both tangible entities ("manifestations") and abstractions (sometimes called "works"). Resolution is the process of submitting an identifier of an entity to a network service and receiving in return one or more pieces of current information related to the identified entity. In the case of the Domain Name System (DNS), as an example, the resolution is from domain name, e.g., www.doi.org, to a single IP address, e.g., 132.151.1.146, which is then used to communicate with that Internet host. In the case of the DOI, using the Handle System as a reference implementation, the resolution is from a DOI, e.g., 10.1000/140, to one or more pieces of typed data: e.g. URLs representing instances of (manifestations of) the object, or services such as e mail, or one or more items of metadata; the Handle system resolution of one identifier to multiple data is called a "multiple resolution" mechanism. Resolution can be considered as a mechanism for maintaining a relationship between two data entities; an item of metadata is a relationship that someone claims exists between two entities: therefore, such metadata relationships between entities may be articulated and automated by resolution.

Using multiple resolution, a DOI can be resolved to an arbitrary number of different associated values: multiple URLs, other DOIs, or other data types representing items of metadata. Resolution requests may return all associated values of current information, or all values of one data type; these returned values might then be further processed in a specific "client" software application. At its simplest, the user may be provided with a list of options; more sophisticated automated processes would allow for the automated choice of an appropriate value for further processing.

Resolution provides a mechanism for persistence of URLs, by interposing a level of managed redirection. The lack of persistence in identification of entities on the Internet is a commonplace. Even the most inexperienced of users of the World Wide Web rapidly becomes familiar with the "404 file not found" message that means that a specified Web address cannot be found – the URL for that web page cannot be resolved. For example: "One of the web sites I maintain is the Lisweb directory of library homepages. Every week, I run a link checker that contacts each page to see if it is still there, and every week about 20 sites that were in place seven days before have vanished. Across the Internet, the rate at which once-valid links start pointing at non-existent addresses -- a process called "link rot" -- is as high as 16 percent in six months. That means that about one sixth of all links will break¹⁰." Another example in writing this paper, I consulted two articles of interest that I had printed for perusal from current web sites in recent weeks (one from The Economist and one from The Guardian); on writing this paper and checking the URLs, one had changed to an archival URL (The Guardian article) and The Economist article, initially free, has reverted to one accessible only to subscribers. This demonstrates that not only *location*, but also other relevant properties like *access rights*, may change and need to be considered in managing persistence.

2.3 Metadata

A metadata system designed for stability needs to be multimedia, multi-functional, multi-level, multilingual, multinational and multi-platform. Such an approach is said to be well formed.

<indecs> (Interoperability of Data for Electronic Commerce Systems)^{11, 12} was a project that with the backing of DGXIII of the European Commission, brought together as partners and affiliates a global grouping of organizations with an interest in the management of content in the digital environment.

The <indecs> project was created to address the need, in the digital environment, to put different creation identifiers and their supporting metadata into a framework where they could operate side by side, especially to support the management of intellectual property rights. <indecs> was a time-limited project, which finished its work early in 2000. Its output is highly regarded and its analysis has been adopted in a number of different implementations. The IDF, together with EDItEUR, is responsible for managing and further developing the output of indecs, and in a consortium¹³ to build a Rights Data Dictionary - a common dictionary or vocabulary for intellectual property rights named <indecs2RDD> to enable the exchange of key information between content industries and ecommerce trading of intellectual property rights. Work done by IDF in developing the DOI Namespace (a data dictionary for DOI use, based on indecs) was used as input to indecs2RDD. This data dictionary has been accepted as the basis for the ISO MPEG-21 rights data dictionary: it is in effect a deepening and extension of the original indecs work, and is now being actively developed further.

What does it mean for metadata to be “well formed”? There are only two types of metadata that can be regarded as well formed¹⁴.

- Labels: the names by which things are called (of which “titles” are a subset). These are by their nature uncontrolled and broadly uncontrollable. Identifiers are a specialized type of label, created according to rules, but names nevertheless. The fact that they are created in accordance with a prescribed syntax makes them less prone to ambiguity than other types of label and therefore more readily machine-interpretable than completely free-form labels.
- All other metadata (if it is well formed) needs to be drawn from a controlled vocabulary of values, which are supported by a data dictionary in which those values are concisely defined. This means that the values in one metadata scheme (or in one “namespace”) can be mapped to those in another scheme; this mapping may not be exact – where two definitions in one scheme both overlap with (but are not wholly contained within) a single definition in another, for example. However, the use of a data dictionary avoids the sort of ambiguity that is inherent in natural language, where the same word may have very different meanings dependent on its context. Where precision of meaning is essential, human beings can clarify definition through a process of dialogue. This is not generally the case with computers.

The need for something like indecs has arisen from the growth of the digital world but in theory could have been created in a non-digital, non-network world, since indecs is essentially a general ontology, independent of any digital network – it is not in other words in any way tied to the Web in preference to other mechanisms. indecs will be implemented on the internet and other networks through implementing things such as DOI services using it, linking resolution and metadata.

The mapping between different metadata schemes may be more or less exact. It may also involve considerable loss of information or no loss of information at all. It is obviously advantageous to achieve as close a mapping as is possible; this is most easily achieved between schemes that share a common high-level data model. The <indecs> data model underlies all DOI metadata. The same analysis underlies ONIX International, rapidly becoming widely accepted as the metadata dictionary for the publishing industry internationally. Similar developments are now occurring in other media sectors (through e.g. the adoption of indecs by MPEG-21).

Fundamental requirements defined within the indecs project and used within DOI are:

- Unique identification: every entity needs to be uniquely identified within an identified namespace;

- Functional granularity: it should be possible to identify an entity when there is a reason to distinguish it;
- Designated authority: the author of metadata must be securely identified;
- 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> data model was devised to cover the same field of endeavour as the DOI – all types of intellectual property (“creations” in <indecs> terminology). The fundamental principles it defines are however applicable to any data representation. It is an open model, which is designed to be extensible to fit the precise needs of specific communities of interest. It was also designed to be readily extensible into the field of rights management metadata, the data that is essential for the management of all e-commerce in intellectual property. The <indecs> analysis asserts that it is essential for the dynamic data necessary for the management of rights to be built on a foundation of the rather more static data that identifies and describes the intellectual property, and that these two layers of metadata can easily interoperate with one another. A core concept of indecs and DOI is that in fact there is no logical separation of rights metadata from many other metadata; indecs2RDD is in fact a deepening and extension of the fundamental indecs model which has been widely endorsed. The extension of indecs2RDD on the basis of digital rights management does not imply in any way a model which is only applicable to “commercial” data; indeed the metadata tools we are building are highly relevant to public data, and in the indecs model a transaction can be free of any charge but still follow the fundamental model of usage.

The adoption of the <indecs> metadata model gave DOI metadata a firm basis in an intellectual analysis of the requirements for metadata in a network environment that has been tested in real world applications. It will provide easy interoperability with other metadata schemes constructed using the same analysis, and a basis for interoperability with metadata schemes based on alternative analyses.

However, it does not greatly matter to the DOI whether the <indecs> analysis and developments based on its framework come to be widely used for the management of intellectual property on the Internet (although we believe it will be very helpful if they do.) What matters to the DOI at this stage is whether DOI metadata itself provides a good basis for the management of intellectual property entities in the network environment. We are convinced that good data models, based on rigorous analysis, will be essential for this purpose. For example, we see libraries as likely be looking to IFLA's Functional Requirements of Bibliographic Records work as a basis; FRBR maps well to <indecs>. Data dictionaries and transfer protocols based on the <indecs> analysis are already being implemented in commercial contexts.

All a DOI needs is a few kernel elements and a map to a consistent data model. We use an underlying model as a way of guaranteeing that those few elements are useful when people want to extend on them. The reason for using the <indecs> model is that it is alone in having demonstrated its extensibility to real-world transactions, through rights management. DOI implements the indecs dictionary, creating a mechanism to provide a description of what is identified in a structured way and allowing services about digital content objects to be built for any purpose.

The IDF guidelines provide a standard way of doing this, and hence a means of mapping to existing standards such as ONIX, Dublin Core and so on, allowing each community to bring its own identifiers and descriptions into play. Wrapping these tools into a social and policy framework, through the Registration Agency federation, allows the development of DOIs in a consistent quality-assured way across many sectors, opening the possibility of managing multimedia objects seamlessly.

3 Digital object architecture and preservation

Since persistence is interoperability with the future, for any persistent object we must be able to:

- Uniquely identify it

- know what it is (describe it precisely)
- provide a means of actioning it (accessing it, or some service about it)

Preserving the access to the information (not just within an archive, but interoperably) is crucial. While digital archives are getting a good deal of attention, they are being used mostly for “data preservation” rather than retrieval and active research. This is analogous to preserving the stone-chiselled hieroglyphics on Egyptian obelisks in the British Museum. However, no Rosetta Stone is yet being constructed as a means for deciphering the data. Because of the linguistic issues involved, “semantic preservation” is tough enough even if users know whether the data were written in ASCII, UTF-8, EBCDIC or some other digital code used for formatting.

DOI’s resolution tool, the Handle system, is based on a wider concept, the Digital Object Architecture. As DOI develops, we are taking more lessons from the digital object architecture analysis, which has as its goal a framework for managing digital (information) objects^{15,16}. A simple summary of the key principles of the architecture might be: to manage an object, give it a name and “talk to it”; don’t worry about where it is and don’t worry about what it’s made of, as these details can be considered secondary characteristics. This enables the architecture to rise above details of application versions and content formats.

The Digital Object Architecture components include the *Handle System* (used by DOI) which, through resolution, allows one to go from a name to attributes. This provides a fundamental indirection system for Digital Object management on the net. Of course, as we will see later there is “no free lunch” and an added layer of infrastructure, which must be managed, is a corollary. The other main Digital Object Architecture component is the *repository system* (not as yet used by DOI). This provides access across systems, space, time, and frees digital content from constraints of specific technologies. Again, there is “no free lunch” and an added layer of infrastructure, which must be managed, is a corollary.

The repository approach to interoperability^{17,18} is modularisation: separating raw data (byte stream) from data types (interpretation), and separating type definition from type implementation. The mechanism is extensible (new types can be created on demand) and new components can be made accessible (with controls) over the network, using a standard repository access protocol (RAP). Repositories can be used to access the “correct” version of the associated software, so that e.g. an object written in Word is readable even in the future under Word version of 2030, etc.

As yet, the DOI system has not investigated or implemented a Repository approach for any of its applications, though it seems certain that there could be some useful application to be developed here; DOI has brought the indecs approach to metadata to digital object management, and the prospect of integrating this with the full digital object architecture is attractive. Other applications of the Handle system are investigating the Repository approach, notably the Defense Virtual library (in which the Defense Technical Information Center (DTIC), the Defense Advanced Research Projects Agency (DARPA), and the Corporation for National Research Initiatives (CNRI) are developing a pilot digital library implementation, building on Digital Object Architecture research. This Defense Virtual Library is establishing a framework in which DTIC can build future network-based services and collections, and is developing a testbed in which to further develop and refine Digital Object Architectures.)

The Digital Object Architecture was designed from the start as an interoperability mechanism (as was indecs), with the concepts of modularity and extensibility to accommodate change. Metadata is tightly bound to the object through defined content types, which maintain “intent of use” for complex objects; these can self-described and registered, can include source code; implementations can change over time w/o changing the type; new types can be attached to old data. Repositories are implemented in hardware-independent language, and Repository source code is available from CNRI.

In many ways, the technology aspects of such implementations are minor: policy and organizational discipline are key and testbeds and research should inform policy development.

The usual strategies for digital preservation focus on how to ensure that the file is readable in some time in the future, ranging from 500 years (from our perspective, an unanswerable question) to 5 years (a feasible question), and the 3 main strategies are:

- Keep every component: preserve everything including the hardware and software (problems of cost and the requirement to finance an "Institute of Ancient Computing Machinery"),
- Emulation: emulate the old machinery/software (this raises problems in the long term when emulations themselves require emulation, i.e. problems of "Russian doll" emulations) and
- Migration: translate from one form to another (problems of managing the correct information)

Migration and similar strategies seem to be the more productive options for the future, since at least they move the problem to one of managing bits rather than atoms. They allow us to separate components and interpose levels of redirection to take account of changes in components or attributes. One does not need to adopt the full repository approach to see the benefits in principle: there are a number of ways one could envisage a persistent identifier being linked/update to take account of serial upgrades etc, where each upgrade is precisely specified by well-formed metadata: and also DOI multiple resolution could offer "default" pointers to archives etc.

DOI can help institutions manage archives by making access to the material easier: it is easy when you have just one Word file to preserve, less easy when you have 5 million, and the DOI and related techniques such as repositories allow you to access and perhaps re-create the object in a translated form, with associated metadata for the management.

4 Issues of preservation and identifier mechanisms

In their excellent 1999 study on Digital Preservation¹⁹ (from which I quote heavily in this section), Bide et al noted several issues where identification, in the sense of a well-structured system with appropriate indirection and metadata, can offer potential solutions.

A major issue they described is that of "what [precisely] is going to be preserved". In checking for deposit eligibility, libraries will find that electronic publications are not "registered" anywhere, which means that a complete listing of all electronic publications even for a national area does not exist. Nor is there a standard identifier that is widely used to identify electronic publications (as the ISBN and ISSN are used for print publications), and which could be used to compile such a listing. Further, scientific articles may be now referenced in several different versions²⁰.

A second, deeper, problem comes from considerations of precise structured metadata. Deposit has conventionally involved consideration only of "manifestations" of works²¹. These physical manifestations – particularly books – have been susceptible to relatively straightforward unique identification. For over a quarter of a century, almost all the books deposited at the British Library will have had an ISBN, a way of identifying a specific manifestation. The identification of periodical publications is somewhat more complex, since the ISSN is in reality a work identifier; the introduction of standardised unique identification of individual issues (and articles within issues) is much more recent, following the introduction of the Serials Item and Contribution Identifier (SICI). Nevertheless, a standard, unique identifier is widely used that can (for example) discriminate between two serial publications with the same name. With printed publications, it is thus possible for relatively unambiguous communication to take place between library and publisher. As we move in the direction of electronic products, particularly online publications, the situation becomes more complex. It is broadly true that offline digital products, particularly those, which are included in the terms of the BL deposit scheme, follow much the same pattern as printed publications in terms of identification. They are almost certain to have either a print-related identifier like an ISBN or a Unique Product Code/EAN barcode. Matters become further complicated with online products. Different ISSNs are mandated for the electronic version of a printed serial publication (even if the two are fundamentally and logically identical). While it may often make sense for different manifestations of the same intellectual content to have different identifiers, the underlying abstract work, the journal

itself, is the same. This makes decisions on what is and is not identical in intellectual content (and what to archive) even more difficult to discern; a fundamental requirement for well-formed metadata is the use of an ontology (“an explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them”) which describes such relationships²². In the digital environment, there can be a profusion and different, related, manifestations of an underlying work, as John Sowa puts it: “Computers multiply entities faster than Ockham’s Razor can shave them.”

Once inside an archival deposit system, it is arguable that the matter of unique identification becomes one for the libraries alone (in the sense that in a closed universe, which we expect a library deposit system to be, unique identification is a matter only for those who manage the system). Nevertheless, careful thought shall be given to the ways in which unique identifiers are used at the point of deposit, particularly to the extent that they may be used in the future as finding and location aids – i.e. interoperability considerations. The long-term value of the deposit archive will depend, to some extent at least, on the approach taken to identification. Bide et al conclude: “In this connection, we find the approach being taken by the International DOI Foundation persuasive. This work, which is closely related to the work of the <indecs> project, proposes that a limited kernel of metadata should be deposited alongside every registered DOI. The kernel metadata will be supplemented and qualified for different genres of content. An approach similar to this, in which a minimal, but tightly defined set of metadata is expected to be deposited by the publisher, would appear to us to be a realistic approach”.

Digital publication also allows the publication and interchange of smaller (and hence many more) components – whereas in the physical world a book is transacted as a whole, digitally its component chapters may be manipulated as independent objects, perhaps with no information as to context. The ISBN working group are setting up a sub-group to look at how ISBN can be extended to include fragments, especially of digital publications, given the failure of BICI to take off, and both EDItEUR and IDF will be participating. This issue of granularity is one which indecs has analyzed closely.

A further issue, which seems certain to cause much debate is the issue of copy protection and related licensing rights. Dan Bricklin, inventor of VisiCalc, has told how the VisiCalc software, and the ability to study how an early PC program was designed, might have been lost to the public forever if Bricklin’s original company, Software Arts, had not turned out to have a solitary unprotected version²³. That one copy became the download on what is now the most popular part of Bricklin’s website. “Copy protection will break the chain of formal and informal archivists who are necessary to the long-term preservation of creative works” says Bricklin²⁴. One need not go so far as Bricklin and other advocates like Laurence Lessig²⁵ that “copyright doesn’t work in the digital age” to recognise that this is a practical issue for preservation, and one which seems to get relatively little consideration in discussions of the problem of archiving. Copy protection per se is not a feature of an identifier mechanism like DOI (deliberately); rather, DOIs are “hooks” by which many different copy-protection mechanisms and other rights management features might connect for interoperation.

5 Identification of datasets

There is an application of DOIs now under active discussion, which has a particular relevance for the topic of digital preservation of the record of science. This is the allocation of a unique identifier (DOI) to a “dataset”.

DOIs are now widely used for the identification of scientific articles (and their citation electronically)²⁶, which form the backbone of the peer-reviewed record of science. Through the CrossRef consortium, in which 101 publishers are collaborating at the time of this seminar, over 4.3 million DOIs have been allocated so far to scientific articles, including extensive back files^{27 28}. In addition to the benefits of persistent resolution, and defined metadata and services, there are some instant benefits in interoperability where none existed previously: for example, publishers do not need to re-allocate legacy identifiers but are able to incorporate whatever they are using; even if one publisher uses Publisher Item Identifiers (PII) and another uses Serial Item and Contribution Identifier

(SICI), the resulting DOIs can be used interoperably to create links in the CrossRef database. CrossRef is now considering application of DOIs to other scientific publication types, including conference proceedings, encyclopaedia entries, and book chapters and is working with IDF, ONIX and ISSN among others.

Scientific articles record underlying data, in part. Often this is available not within the article but as extensive "supplementary data" either for peer-review or in published form. Increasingly however data leads a life of its own in archives such as gene sequence databases. At this stage, the term "dataset" in a DOI context has not been precisely analysed and defined, and it is likely that there are several different interpretations which will need to be specified and reconciled (though the concepts of indecs certainly allow this exercise to be structured appropriately). For example, one approach views a dataset as the set of data associated with a scientific *publication*; other approaches view it as the set of data associated with an *experiment*; and still another possibly as a (growing) set of definitive values of answers to a defined *problem* (rather like a definitive table of data values in a standard reference book). The ontology needs careful work if we are to avoid confusion; and each interpretation has implications for the owners and originators of the information captured in the dataset. The potential for a "tidal wave" of data from recent automation-aided experiments requires that we get the management of such data right.

In the last few months, the International DOI Foundation has considered several separate approaches to such DOI allocation, and these illustrate different approaches:

- IUCOSPED (the IUPAC-CODATA Task Group on Standardization of Physico-Chemical Property Electronic Data Files)^{29 30} invited IDF to participate in an IUPAC workshop in 2001. The concept of standard electronic files (SELFs) for physico-chemical datasets was discussed, and concluded that it is clearly an area of potential application for DOI and of particular interest to STM publishers and related bodies, especially because of the criterion the group established that a "dataset" to be valid must have been published as part of a formal article. ICSTI supplied some funding for an initial investigation in this area with Dr Henry Kehiaian of the University of Paris.
- IDF is currently working with our member the European Molecular Biology Organisation, through its E-BioSci activity³¹ on DOIs for biological information sets. EMBO has supplied funding for this activity by joining IDF as a member, and another IDF member Ingenta (www.ingenta.com) is a commercial partner. Some aspects of this work, which has just begun and is a 3 year programme, will have lessons applicable to other data types (e.g. physico-chemical data sets). Although this work has only just begun, it is based on some earlier DOI work on biological images, and the technical metadata work has a close connection with the indecs work (one of the key consultants worked on both).
- A working group on "Citation of Primary Data" of CODATA and the DFG (German Science Foundation) is seeking to establish a DOI profile to publish and to cite primary scientific data independently from literature publication. The initial approach of the working group is to establish a primary data DOI Application Profile and to install a DFG-supported registration agency in Germany. For such applications, the DOI application profile (metadata set) core must be defined in such a way as to capture all the relevant data to uniquely and precisely define the set of data; and a social infrastructure for quality control set up, via a registration agency.
- A fourth potential application (IUPAC spectrographic data) has been brought to my attention at this symposium.

DOI rules are agnostic as to whether or not there is a related formal publication entity corresponding to a dataset, although the initial implementations and indeed focus of DOI have been on intellectual property in the form of copyrighted materials, and it is easier to see how this applies to articles than to databases. Specific communities and applications may define rules about this. The SELF concept uses identifiers for *documents* (bibliographic identifiers) and for *parties* (such as publishers). The BioImage project (now taken up in E-BioSci) used identifiers for *experiments* and *tools*³². DOI (and ISO) have

developed identifiers at the bibliographic level, now being extended to the abstract work level, and the DOI application used by CrossRef can be the identifier of the publication. Identifiers for other entities like data sets and parties are less well established, but the overall indecs framework provides a framework in which to do this.

It is not yet clear how these various approaches relate, and whether it would be advantageous to keep them independent or to co-ordinate some collaboration through IDF. This is an issue which will be discussed at the ICSTI meeting following this seminar. The International DOI Foundation (IDF) would be very interested in working with the scientific and publisher communities on the concept of persistent identification of datasets, if funding for a project can be made available (possibly widening the scope and sharing the costs with other disciplines). One problem is convincing someone to fund the work, since initially publishers and libraries do not seem to see it as mainstream to their concerns, and it is (as ever) difficult to fund fundamental infrastructure work of this kind. Perhaps organisations such as CODATA, ICSTI, and UNESCO can help here.

6 Persistence of DOI

6.1 Overview

If DOIs are proposed as a tool for managing persistence, it is fair to ask what makes a DOI itself persistent. The DOI Handbook discusses persistence in the chapter on "Policy". It may seem odd for "persistence" (permanence) to be discussed in a chapter on policy, rather than on of the sections on technology. There is a simple reason for this: persistence is ultimately guaranteed by social infrastructure (policy); persistence is fundamentally due to people, and technology can assist but not guarantee.

Identifiers must persist in the face of legitimate change. There are legitimate, desirable, and unavoidable reasons for changing organisation names, domains etc. One aim of naming entities/resources is to avoid tying an entity name to a domain name, or any other piece of variable metadata (which led to practical difficulties with the domain name system when trademarks associated with the domains caused conflicting claims)³³. The entity can be persistently named as a first class object irrespective of its location, owner, licensee, etc. Distinguishing names from locations is essential for E-commerce. It is trivially true that "all names are locations" (in a namespace), but practically, most people worry about spaces like URLs, and that's the wrong level. Naming entities as first class objects³⁴, rather than locations, enables better management of multiple instances of an object, for example.

Persistence is something we are familiar with in the physical world: ISBNs for out of print books can still be useful, and we would want any identifier scheme to outlast the entities being identified. Persistent identification alone is a good enough reason to adopt identifiers such as DOI which provide a means by which potential customers can find your digital offering even if a "broken link" URL of a retailer or other intermediary intervenes.

Technology can help with persistence. Using DOIs, only one central record, which is under the control of the assigner, needs to be changed in order to ensure that all existing DOIs which are "out there" in other documents can still resolve correctly: a redirection (resolution) step enables management in the redirection directory, thereby ensuring that one change can be picked up by many users, even if they are unaware of the change. But to manage the data in the directory takes effort, time, incentive, etc. - either you do that locally (using tools such as purl, managing a service yourself at your own site) or as a global service (DOI, which ensures that third party uses can be made of the identifier interoperably). In the case of DOI management of data is a service role (and hence also business activity) for registration agencies. We can learn from other activities like bar codes, ISBNs, and other data systems. People aren't free, so there's a cost to this, and just like the physical bar code system, the DOI aims to be a self-funding operation. DOIs won't be appropriate for many things, and some people won't feel this people cost merits the reward. But we do think DOIs are a viable solution for adding value to enable content management of intellectual property on a large scale.

DOI is an implementation of URN (Uniform Resource Names) and URI (Universal Resource Identifier) concepts, and can be formalized within these frameworks³⁵. The aim of each is to allow persistence of naming irrespective of other characteristics.

The central DOI resolution system is managed to ensure that persistent names can be resolved to non-persistent attributes such as location. One of the problems with the World Wide Web today is that once an object is moved, anyone searching for that object may encounter an error message. This is because URLs identify a location, not an object. The DOI, by contrast, specifies an object, not a location. Each DOI is registered in the Handle System and can be resolved to at least one location somewhere on the Internet. When an object is moved, all a rights holder has to do is re-point the DOI to the new location and the object can be found once more; any external party accessing the DOI does not need to know of the change and will be taken transparently to the object. The DOI system is designed to enable registrants to make up-to-date changes easily and consistently, and to monitor errors. Additional tools (such as workflow implementations) are already being developed by outside parties, and more will follow.

6.2 Persistence of the resolution technology

One of the key issues for the IDF in implementing Handle (HDL) as the technology for DOI was: how do we know that HDL itself is going to be persistent; will HDL be around in 5 years/50 years? There are both social and technical measures, which are relevant.

The HDL system is an open standard, so anyone can build/use one; but it relies of course on the top level Global Handle Registry to be in place somewhere (just as e.g. the internet Domain Name System assumes there will always be a root server and directory around somewhere). CNRI has a commitment to funding and maintaining these; were that to fail, there are enough large scale implementers of handles to ensure that it will be "picked up" by someone (e.g. Library of Congress, the US Dept of Defense, IDF, etc.). The Global Handle Advisory Committee, containing representatives of major handle users and stakeholders, exists to safeguard the future of the Handle System; IDF has a seat on the GHAC.

At the technical level we can take steps in improving resilience of the infrastructure, mirroring machines to insure against power outages etc.: the normal things one would do to improve technical system reliability. Key Handle infrastructure is placed with a professional hosting company with resources to ensure 24x7 cover. Further steps to make the system more persistent from an organizational point of view are under discussion, largely influenced by IDF requirements.

6.3 Persistence of the identified object

Just as there are legitimate, desirable, and unavoidable reasons for changing organisation names, there may be equally legitimate, desirable, and unavoidable reasons for declaring that an entity identified as a DOI is "no longer available". For example, a major publisher's policy on article withdrawal of electronic products states: "under exceptional circumstances, an article must be removed from an electronic product due to legal obligations on behalf of the Publisher, owner, copyright holder or author(s); or on moral or ethical grounds if an article with an error, or with results/statements has been found inaccurate and could be potentially damaging".

The DOI system provides a mechanism for managing this process. At minimum, a DOI registrant is free to have the DOI resolve to a response screen indicating that the identified entity is no longer available. This in itself will be very useful (consider for example that ISBNs for books which are out-of-print are still useful). A response such as this is certainly more useful than "404 not found". Beyond this, a publisher or RA is free to define its own policy: it may be useful to develop a default or fall-back mechanism for certain Application Profiles of DOIs, whereby DOIs which are no longer available through the original distribution channel of their registration are re-directed to an archival source, or to a standard source of data with an indication of the reason for the withdrawal. Since DOI can be associated with multiple options, services can be constructed on the basis of Handle data types (and potentially indecs metadata entities); these services could include archiving options, which could either be global (a default option; a service provided by the

assigning DOI agency e.g. CrossRef); or could be a local implementation choice made by a library or other organization, combining global DOI resolution with local tools such as OpenURL (already demonstrated³⁶)

There may be specific rules for this developed within a DOI Application Profile; or there may be some generic rules, which can be devised for all DOIs. It is clear that there are many different reasons for such "out of print" digital objects: for example, an old version replaced by a new: the publisher response to a query about an old edition DOI could be a creative marketing approach (i.e., "you have requested information on an early edition which is now superseded by.." or, ".." it has been superseded by the new edition but you can still obtain the older version from xyz [second hand dealer or old source]...")

IDF has concluded that it would be premature to determine a one-size-fits all mechanism; it is likely to be a result of functional requirements for the particular DOI, including commercial issues. However this is an issue where we welcome active contributions and suggestions.

6.4 Stability and invariance of the associated metadata

A principle IDF policy is that DOI kernel metadata be stable and persistent. This needs to be considered in relation to situations where the appropriate data may change: for example the very common situation where, when a commissioned work, or a planned publication, is first registered in a database, it has only a working title and possibly even a proposed author who may then turn the project down. From an archive point of view, this raises interesting issues of preserving "early drafts", which in the physical worlds may be of great interest to later scholars.

There are three logical ways of handling this:

1. Make it an absolute rule that if the metadata changes in any way at all, a new DOI must be registered (it may be questioned whether this is enforceable);
2. Allow some kernel elements (title, primary agent) to have one or more superseded values as well as their current value. "Stable" would then mean that the content of an element, once entered, could not be changed, but its status as "current" or "superseded" could; and "persistent" would mean that superseded values would never be deleted.
3. Adhere strictly to the literal interpretation of "stable" as "invariant" and "persistent", so that if (say) a title changes after DOI registration, the registered title in kernel metadata cannot be changed, but the extended metadata managed by the RA would carry the definitive title and link it to the registered title.

Each demands certain disciplines from registrants and RAs; and each has different implications for the possible use of the kernel metadata. This may be an issue which will be dealt with at the Application Profile level; working policy at present is that the precise interpretation of this is something that registration agencies may want to agree, but the *principle* is that the intention upon registration is that the *kernel metadata is not likely to change*. While we acknowledge that mistakes and updates occur in the real world we are setting a high standard to encourage registrants to get it right. All registration agencies will want to allow errors to be corrected, so we should not assume a difference between theory and practice. For this reason we adopt the term "stable" rather than "invariant".

7 DOI and archiving initiatives

Since the inception of the DOI we have been engaged in constructive dialogue with a number of library and archiving activities. DOIs were recommended in the British Library BNB report on "Digital Preservation" cited earlier. They feature heavily in other studies such as the initiative on persistence of the Australian National Library³⁷. The recent Library of Congress Action Plan³⁸ includes an action (3.7) to " Evaluate feasibility of assigning a persistent identifier or a naming system on an international scale; develop and promote guidelines for shared resolving system" for which IDF has offered to act as a collaborator. DOIs have also been presented at meetings of the Council of Directors Of National Libraries and the Conference of European National Librarians, in the context of

their potential for use as National Bibliography Numbers and at CIMI³⁹ (Consortium for the Computer Interchange of Museum Information).

There is a widespread recognition of the advantages of assigning identifiers; and a widespread misconception that an abstract specification (like a URN or URI) actually delivers a working system rather than a namespace that still needs to be populated and managed. A common problem we have encountered in discussions of archiving is the desire to adopt a system at no cost. It is inescapable that a cost is associated with managing persistence and assigning identifiers and data to the standards needed to ensure long-term stability. This is because of the need for human intervention and support of an infrastructure. It is accepted that assigning a library catalogue record, for example, will typically cost anything up to \$25. Assigning an ISBN or ISSN or National Bibliography Numbers will also have costs, even if these are not paid directly by the assigner. Although a DOI is free at the point of use, there is a small fee to an assigner for creating a DOI (a few cents). This is because we have deliberately chosen to make the DOI a self-funding (though not for profit) system. Our task now is to show that DOI offers value for money as a tool which archives can use - rather as most libraries find it easier and simpler to buy shelving systems rather than build their own, even though the basic materials would be far cheaper.

If adding a URL “costs nothing” (which itself ignores some infrastructure costs), why should assigning a name? It is indeed possible to use any string, assigned by anyone, as a name – but to be useful and reliable any name must be supported by a social as well as technical infrastructure that defines its properties and utilities^{40,41}. URLs for example have a clear technical infrastructure (standards for how they are made), but a very loose social infrastructure (anyone can create them, with the result that they are unreliable alone for long term preservation use as they have no guarantee of stability let alone associated structured metadata). Product bar codes, Visa numbers, and DOIs have a tighter social (business) infrastructure, with rules and regulations, costs of maintaining and policing data – and corresponding benefits of quality and reliability (When a credit card is presented, we can be reasonably certain that the number is valid, and has been issued only after careful correlation with associated metadata by the registrant). It does not necessarily imply a centralised system – it may be a distributed system (like domain names), but it must have some form of regulation.

Such regulation of infrastructure for a community benefits all its members; funding the development of it is often a problem, and there is no “one size fits all” solution to how this should be done. But finding a workable model for the development of an infrastructure can yield obvious benefits. In 18th century Britain one simple change, the turnpike system, by requiring travellers rather than local parishioners to pay for road building, resulted in the establishment of an elaborate road network and a drastic reduction in journey times. There are many modern examples – 3G telephone networks, railways – which are struggling with the right model for supporting a common infrastructure. The Internet was largely a creation of central (US) government; the product bar code, a creation of a commercial consortium. The IDF has chosen as its model the concept of Registration Agencies, based on market models like bar codes and Visa rather than on centralised subsidy: these Agencies effectively hold a “franchise” on the DOI: in exchange for a fee to the IDF, and a commitment to follow the ground rules of the DOI system, they are free to build their own offerings to a particular community, adding value services on top of DOI registration and charging fees for participation. CrossRef is one proven example in text publishing; we expect to see other variants on this theme develop.

At the outset of the DOI development, a very simple business model was introduced whereby a prefix assignment is purchased for a one-off fee. A fee was introduced not to cover actual costs, but to recognize the fact that some charging for DOIs would be a necessity. IDF introduced a charge of \$1000 for allocation of a prefix allowing unlimited number of DOIs to be constructed using that prefix. The charge is one-off and entitles the registrant to an infinite number of suffixes; there is no annual fee, though we reserve the right to vary this at a future date; there is no limitation placed on the number of DOI prefixes that any organization may choose to apply for. It was recognized at the outset that this fee structure was a starting point but would be insufficiently flexible for the long term.

We are now in a process of migration to the long term aim of a wide variety of potential business models, using third part registration agencies, in recognition of the fact that such a simple model is not a "one size fits all" solution. The direct prefix purchase route is still an option, but our intention is that eventually all future DOIs will be registered through one of many Registration Agencies, each of which will use one or more defined DOI Application Profiles, and each of which is empowered to offer much more flexible pricing structures. The pricing structures and business models of the Registration Agencies will not be determined by the IDF; each RA will be autonomous as to its business model. Business models for these agencies could include, but not be limited to, cost recovery via direct charging based on prefix allocation, numbers of DOIs allocated, numbers of DOIs resolved, volume discounts, usage discounts, stepped charges, or any mix of these; indirect charging via cross subsidy from other value added services, agreed links, etc. The IDF will place minimal constraints on the business models offered by RA's, and enter into discussion on practical implementation of any of these.

Can DOIs be made available at "no charge"? Yes, if the costs of doing so can be met from elsewhere.

(a) IDF itself is willing to allocate a DOI prefix free of charge to organizations or limited experimental non-commercial uses (please contact us if you wish to apply for this);

(b) The business model includes two separate steps: a business relationship between IDF and an RA (the "franchise fee"); and a business relationship between an RA and a DOI registrant (the "registration fee"). The two are not directly connected; this enables the RA to offer to registrants any business model whatever, which suits its needs. This could include assigning DOIs without charge. Hence DOIs can be used in both commercial and non-commercial settings, interoperably. However, the franchise fee in such an example cannot be zero; this would immediately undercut any commercial use, and it would not provide any financial support for the operation of the system itself. Like any other piece of infrastructure, an identifier system (especially one which adds much value like metadata and resolution) must be paid for eventually by someone. So an organization could, if it wished, assign DOIs freely (registration fee zero to registrants) and subsidize this added-value service by paying a franchise fee to IDF from a central fund, as an acceptable cost for supporting the service.

8 References

- ¹ Paskin, Norman; Toward Unique Identifiers; *Proceedings of the IEEE*; 87 (No.7) July 1999; pp. 1208-1227
http://www.ieee.org/organizations/pubs/pub_preview/PROC/87proc07_toc.html
- ² Jones, Maggie; Beagrie, Neil; Preservation Management of Digital Materials: A Handbook; The British Library; October 2001
- ³ Kahle, Brewster; Prelinger, Rick; Jackson, Mary E; Public Access to Digital Material
<http://www.dlib.org/dlib/october01/kahle/10kahle.html>
- ⁴ Arms, William; Hillman, Diane; Lagoze, Carl; Krafft, Dean; Marisa, Richard; Saylor, John; Terrizzi, Carol; Van de Sompel, Herbert; A Spectrum of Interoperability; D-Lib Magazine; January 2002
<http://www.dlib.org/dlib/january02/arms/01arms.html>
- ⁵ ONIX Product Information Standards.
<http://www.editeur.org/onix.html>
- ⁶ Dublin Core Metadata Initiative.
<http://dublincore.org/>
- ⁷ DOI Handbook
http://www.doi.org/handbook_2002/index.html
- ⁸ The Handle System
<http://www.handle.net>
- ⁹ DOI Handbook, Appendix - An Introduction to the <indec> metadata framework
http://www.doi.org/handbook_2002/appendix_5.html
- ¹⁰ Dowling, Thomas; One Step at a Time; NetConnect; Fall 2001, page 36.
- ¹¹ Rust, G; Bide, M; The <indec> Metadata Framework: Principles, model and data dictionary; 2000.
<http://www.indec.org/project.htm#finalDocs>
- ¹² Rust, G.; Bide M. indec Summary Final Report; 2000
<http://www.indec.org/project.htm#finalDocs>
- ¹³ IDF Announcements :
<indec>2rdd adopted as MPEG-21 baseline technology
<http://www.doi.org/news/020114-DRM.html#consortium>
Major organizations to develop digital rights management (DRM) standard
<http://www.doi.org/news/011101-DRM.html>
- ¹⁴ Rust, G; Metadata: The Right Approach, An Integrated Model for Descriptive and Rights Metadata in E-commerce; D-Lib Magazine
<http://www.dlib.org/dlib/july98/rust/07rust.html>
- ¹⁵ Kahn, R.E and Wilensky, R; A Framework for Distributed Digital Object Services; 1995; Reston, VA: Corporation for National Research Initiatives (CNRI).
<http://www.cnri.reston.va.us/home/cstr/arch/k-w.html>.
- ¹⁶ Cross-Industry Working Team; Managing Access to Digital Information
<http://www.xiwt.org/documents/ManagAccess.html>
- ¹⁷ Active Digital Object Repository Architecture (ADORA)
http://www.handle.net/wkshp_000920/ADORA-DOI.ppt
- ¹⁸ Bianchi, Christophe; Petrone, Jason; Distributed Interoperable Metadata Registry; D-Lib Magazine; December 2001
<http://www.dlib.org.ar/dlib/december01/blanchi/12blanchi.html>
- ¹⁹ Bide, M; Potter, E J; Watkinson, A; Digital Preservation: an introduction to the standards issues surrounding the deposit of non-print publications; September 1999
<http://www.bic.org.uk/digpres.doc>
- ²⁰ International Working Group; Definition of Publication in the Electronic Environment; July 2000
<http://www.alpsp.org/define2.pdf>

-
- ²¹ Svenonius, Elaine; The Intellectual Foundation of Information Organization; MIT Press; 2000
- ²² Sowa, John F; Knowledge Representation: Logical, Philosophical and Computational Foundations; Brooks/Cole; 2000
- ²³ Lillington, Karlin; Sentries at the gate; Guardian; Thursday December 20 2001
<http://www.guardian.co.uk/Archive/Article/0,4273,4322913,00.html>
- ²⁴ Bricklin, Dan ; Copy Protection Robs The Future; www.bricklin.com
<http://www.bricklin.com/robfuture.htm>
- ²⁵ Lessig, Lawrence; Code and Other Laws of Cyberspace; Basic Books; 1999
- ²⁶ Paskin, N; E-Citations: actionable identifiers and scholarly referencing; Learned Publishing Volume 13, July 2000, pp. 159-166
<http://www.catchword.com/alpsp/09531513/v13n3/contp1-1.htm>
- ²⁷ CrossRef Web Site
<http://www.crossref.org>
- ²⁸ Brand, Amy; CrossRef turns One; May 2001.
<http://www.dlib.org/dlib/may01/brand/05brand.html>
- ²⁹ IUPAC Task Group on Standardization of Physico-Chemical Property Electronic Data Files (IUPAC-CODATA Project)
http://www.iupac.org/projects/1999/024_1_99.html
- ³⁰ CODATA Task Group Web Site
<http://www.codata.org/2000tg.html>
- ³¹ E-BioSci - a European platform for access and retrieval of full text and factual information in the Life Sciences
http://www.e-biosci.org/E-BioSci_overview.html
- ³² Lindek, Steffen; The BioImage Metadata Framework; International DOI Foundation
<http://www.doi.org/topics/metadata.rtf>
- ³³ WIPO Domain Names
<http://ecommerce.wipo.int/domains/>
- ³⁴ Kahn, R.E and Cerf, V.G; What is the Internet (And What makes it Work); Internet Policy Institute; December 1999
http://www.internetpolicy.org/briefing/12_99_story.html
- ³⁵ Paskin, Norman; Neylon, Eamonn; Hammond, Tony; Sun, Sam; Uniform Resource Identifier (URI) scheme for Digital Object Identifiers (DOIs); Internet Draft: draft-paskin-doi-uri-00.txt (February 2002)
- ³⁶ Caplan, Priscilla, Oren Beit-Arie, Miriam Blake, Dale Flecker, Tim Ingoldsby, Laurence W. Lannom, William H. Mischo, Edward Pentz, Sally Rogers, Herbert Van de Sompel; Linking to the Appropriate Copy: Report of a DOI-Based Prototype.
<http://www.dlib.org/dlib/september01/caplan/09caplan.html>
- ³⁷ Dack, Diana; National Library of Australia; Persistent Identification Systems - Report on a consultancy; May 2001
<http://www.nla.gov.au/initiatives/persistence.html>
- ³⁸ Bibliographic Control of Web Resources: A Library of Congress Action Plan
<http://lcweb.loc.gov/catdir/bibcontrol/actionplan.html>
- ³⁹ Consortium for the Computer Interchange of Museum Information
<http://www.cimi.org/>
- ⁴⁰ Brin, David; The Transparent Society: Will Technology Force Us to Choose Between Privacy and Freedom?; Addison-Wesley; (1999) 384 pages
- ⁴¹ Seely Brown, John; Duguid, Paul; The Social Life of Information; Harvard Business School Press; 2000