# The Digital Object Identifier initiative: current position and view forward

*DOI discussion paper (version 3)*

**CONTENTS**



**Context/Status:** This paper represents a summary of work in progress and is subject to amendment at any time. *It is posted to invite comments and criticisms*. This version has been compiled following comment on earlier versions. Further amended versions may be posted to the DOI web site at *http://www.doi.org* and may include changes of policy or proposed direction in light of comments and criticisms received, which are welcome and should be sent to *n.paskin@doi.org* or to the DOI discussion mailing list at *discuss-doi@doi.org*. Nothing in this document should be construed as a commitment to permanent policy of the International DOI Foundation. This paper may be reproduced and distributed for the purposes of comment, providing that it is reproduced in total including this statement.

Author: Norman Paskin (n.paskin@doi.org): see acknowledgements section
Date: 14 August 1998
Version: 3
Location: posted at http://www.doi.org

**Acknowledgements**

I am grateful to all the participants in recent DOI workshops and discussion lists for their help with some of the concepts presented in this document, and in particular to Bill Arms, Leslie Daigle, Larry Lannom and Godfrey Rust.

**Summary/abstract**

Sections 1-4 introduce a necessary common vocabulary and summarise work to date; sections 5 is a scope statement; section 6 focuses on likely future development; sections 7-11 on practical issues of implementation. References are given in [square brackets].

There exists a commonly accepted architecture model for managing information as digital objects (section 1.1); a component of this architecture is the process of resolution (1.2). There is no commonly accepted equivalent model for intellectual content in general (irrespective of medium) and the move to digital content management requires a standard vocabulary to be defined (2.1) (including Creations, Works, Packages and Objects) which makes a distinction in particular between the identification of abstract works as in citations, for which standards are still in development (2.2) and the identification of tradeable digital Objects manifesting those works (2.3); the DOI can be used to provide a unique identification scheme useable with this data model. It is also necessary to give special consideration to identification of entities which have a special relationship to those works, such as abstracts (2.4) and dynamic collections of works (2.5). In implementing a workable scheme it is necessary to recognise that some Creations are separably identifiable but related and this must be discernible; a key problem which remains is the conflicting practical requirements of a Work identifier needed for citations, and an Object identifier needed for trading (2.6).

The DOI initiative has received widespread support (3.1); the initial implementation uses Handle technology (3.2) in a restricted subset, with few limitations on scope (3.3). It is intended that the system be developed further in a parallel tracks approach which will gradually evolve an interoperable framework from early implementation experiments (3.4). The DOI is a system offered in conformance with existing and evolving standards from both the content/information community (4.1) and the technology/Internet community (4.2).

The scope of the DOI is now defined as digital services offered against Content, irrespective of whether the Content is digital or non-digital. This has the implication that whilst all digital Objects may have a DOI, not all DOIs relate to digital Objects (5).

In order to develop the DOI further and offer DOI-based services, a necessary distinction is drawn between content, services, and mechanisms (6.1). We make a new distinction of “level 1" DOIs (using a single data type value, as in the current implementation) and “level 2" DOIs (using multiple resolution values, necessary for more sophisticated uses) which will require more sophisticated client tools (6.2). Metadata is necessary to be associated with each entity assigned a DOI, in order to create useful services (6.3). Level 2 DOIs can be nested to create some services (6.4). Services could be provided against DOI identifiers in several ways which are not yet defined in detail (6.5); practical steps to achieve these services require definition of the services desired by users, introduction of level 2 DOIs, appropriate support tools, definition of DOI metadata set(s), and a means of grouping related identifiers (6.6).

The DOI system is already in use and will continue to evolve; some guidelines can be given to facilitate common approaches: assign DOIs to Creations, not Resources; assign separate DOIs

to separate but related Creations; treat services as a response page requiring manual user intervention (possibly in a standard form) pending development of automated services; do not assume that every portion of an existing Web Presence should be associated with a DOI; and do not confuse what a DOI identifies with what it resolves to (7).

The paper briefly discusses the assignment of DOI (8), usage (9) and business model (10); each of these will need further expansion. Section 11 provides a list of additional action points arising from the discussion document.

## Introduction

Standard ways exist or are in development for managing Internet *Resources*, such as Uniform Resource Locators (URLs), Names (URNs), and metadata (Resource Description Framework, RDF). These resource mechanisms provide an infrastructure for managing resource discovery and distribution, but not a sufficient framework in which to manage *intellectual content* and the rights which accompany that content, such as access rights and copyright. Publishing - in whatever medium - is the distribution of intellectual content and concomitant rights management (e.g. royalty payments to authors and composers) in any medium. E-commerce of intellectual content - *digital publishing* - requires content management with a variety of associated services to manage access and other rights. Importantly, persistence (permanence of naming) is a requirement; an identifier should persist longer than the object it identifies [Paskin2]). Managed Web distribution is only one component of the required architecture - necessary, but not sufficient.

The DOI initiative [DOI], launched in October 1997 following a prototyping phase [Rosenblatt] aims to develop a common mechanism to enable intellectual content management to be integrated with Internet technologies. The DOI initiative brings together two communities: the digital technology-oriented community, devising digital library architectures and appropriate technical solutions; and the content-oriented community which views "being digital" as one of several possible mechanisms of publishing.

The emphasis of the present paper is on issues addressed in recent discussions and workshops including scope and guidelines for application of DOI. Other issues such as business models are dealt with only briefly and will be expanded in later documents. It will be seen that the DOI is ambitious in scope and leads to facing a number of inescapable issues (such as the difference between a citation and a purchase) which are crucial to advancing digital publishing. The International DOI Foundation will attempt to provide answers to these questions.

## 1. The technology viewpoint: Digital Objects

### 1.1. Digital Object architecture

Information can be captured and manipulated digitally: intellectual content can be embodied in coherent collections of bits, or Digital Objects. The DOI builds on the Digital Object concept. A Digital Object is a meaningful piece of data: "a data structure whose principal components are digital material, or data, plus a unique identifier for this material" [Kahn/Wilensky]; "not merely a sequence of bits or symbols...it has a structure that allows it to be identified and its content to be organized and protected..." [XIWT]; "a Document-like Object", according to the Dublin Core activity [Caplan]; a Knowledge Object (KNOB), [Kelly]. There has been substantial progress in defining architectures for digital object structures, e.g. digital libraries, repositories [Arms], URIs, and improved mechanisms for Digital Object access which resolve to multiple data types (such as URLs). In such an architecture a client (user computer) interfaces with three distinguishable entities:

- Resource discovery tools (search engines, metadata databases, catalogs, etc);
- Resolution System: inputs an entry related to a specific Digital object and returns some data about that object, such as location;
- Repository/Collection: the home of the Object from where it may be retrieved.

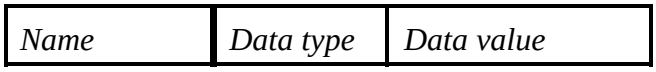
Digital Objects are a meaningful sub-set of Internet Resources (as in URL, URN, etc). A "Resource" is ill-defined: "a file in a directory, [which] can exist on any machine on the network, can be served via any of several different methods, and might not even be something as simple as a file: URLs can also point to queries, documents stored deep within databases, the results of a finger or archie command, or whatever." [URLguide].

The DOI initiative makes use of the infrastructure offered by these architectures and tools, and specifically uses the Handle resolution system [Handle]. However it is not synonymous with those activities as will be explained below: in particular the "Digital Object Identifier" should not be considered to be applicable to all "Digital Objects" in the Kahn/Wilensky sense, nor restricted to only digital entities.

### 1.2 Resolution

Identifiers can be assigned to entities such as Digital Objects; to make them usable with the technology architecture we need to invoke the process of resolution. Resolution means the initial step of using an identifier to do something useful -- sending the identifier (e.g. DOI) to some system to get back the next bit of information that you need to obtain the item or associated service. It is likely that a client would want to consider all pieces of information coming back from that resolution system to enable a service, including obtaining some version of the item itself. This process is not specific to Handle implementation (the implementation used for DOI), but can be generalized to all resolution services, of which the Handle System is the primary operational DOI example.

A general model of resolution, as an aid to discussion, is the following:




which will be interpreted as: sending *Name* as input to the resolver returns as output *data value* of a particular *data type*. A data type, in this context, means a template which the data value is "slotted into" and which then provides a context for the interpretation of that value. For example, the number 0262193736 is a value, which becomes useful information if slotted into the data type "ISBN", enabling its interpretation as the book "Internet Dreams" by Mark Stefik.

This becomes useful in a system of information or Creation identifiers in two ways:
- the Input may be related to a specific identifier ("related to" means that it may be a Creation identifier, or denote such)
- the Data type + value may be many things, including a location (URL) or another identifier

(thereby providing a link between identifiers of input and output), including another Name (allowing nesting).

The model shown above reflects the initial implementation of the DOI in the Handle system using one data type (URL) and one value of that data type; in the general resolution model however, "input" could resolve to a table of multiple data types and data values:

| *Input* | *Data type 1* | *Data value 1* |
|---|---|---|
| | *Data type 1* | *Data value 2* |
| | *Data type 2* | *Data value 3* |

*etc.*

In the Handle system, data types are registered in the Handle record as numerical values corresponding to defined data types; values >65535 are open to be user-defined (e.g. by the IDF community).

The initial announcement of a resolution process as a basis for DOI provoked some concern, in that it was felt that the DOI assigner could control the meaning of an identifier by controlling resolution (at the extreme, the only thing you could do with a DOI was send it to the publishers' resolution system and they could change the meaning of the thing from one day to the next). This is not so: the intention of the DOI initiative is that the identifier can be used independently of the global resolution system, e.g. in a local resolution system such as a library system access catalog. Local resolution of global identifiers would be necessary in examples such as the resolution of a DOI which relates to a Creation which is already held locally (e.g., an article in a site-licensed digital journal collection; which is a very real likely use of such a system).

## 2. The content viewpoint: Creations

### 2.1 Vocabulary for discussing content

The digital technology community takes as its starting point all digital mechanisms, and views intellectual content mechanisms as a sub-set. In contrast, the intellectual content community takes as its starting point all creative works, and views digital mechanisms as a sub-set: from the standpoint of creation or dissemination of intellectual content, "digital" is one of many possible carrier mechanisms, but an increasingly important one. While the digital world has necessarily worked with defined and well-structured concepts, the content world has not (until now) found it necessary to be so rigid: standard numbering (of books, serials, and recordings) and product bar codes have been useful but there is no *widely accepted data model* defining all creative and publishing acts, necessary in placing these in a digital world. Consequently a *consistent vocabulary* is not clearly defined (e.g. "version" can mean many things) and is now necessary to avoid confusion in the DOI development.

A useful framework which offers systems analysis thinking for intellectual content forms is the analysis originally developed for the CIS (Common Information System), implemented for music [Hill] and since generalised [Rust], and we have adopted this framework. At the heart of intellectual content is creativity; this analysis starts by defining *Creations*: "products of human imagination and/or endeavour in which rights may exist", of four types:

*Work:* abstract: made of concepts and ideas (e.g. a composition)
*Package*: physical: made of atoms (e.g. a book)
*Object:* digital: made of bits (e.g. a file)
*Performance:* spatio-temporal: made of actions (e.g. a broadcast)

Creations may be related to each other:

*manifestation:* different-Creation-type-relationships. Creations expressed in differing types e.g. a print article (*Package*) and a digital file (*Object*) manifesting (related to) the same *Work.* Two manifestations are two Creations.
*derivative:* same-Creation-type relationships; as described below, both a *version* and a *component* are derivatives.

Formal relationships between Creations are defined by Links:

| *Link* | *Definition* | *Example* |
|---|---|---|
| Component | A Creation which is contained or manifested within another Creation. (Converse is a *composite:* A is a *component* of B, B is a *composite* including A) | Aria in opera, sample in recording, track on album, poem in anthology, illustration in book, article in journal, audiovisual clip on CDROM |
| Version | A Creation derived from another Creation of the same type, with or without new elements being added. Two versions are two different Creations. | Arrangement, adaptation, translation, different technical format, paperback edition, remix, film edit |
| Reference | A Creation whose content refers explicitly to the content of another Creation | Abstract, review, annotation, advertisement, blurb |

These terms will be used in these precise senses in this paper. For publishers to whom these concepts are new, a few examples may be helpful. The relation between any two Creation types may not be 1:1 and can be many:many (n:n) in all cases: some examples are:

1 Object: n Works *(A publisher is selling an Object which is a compilation of several Works merged into one file)*

n Objects: 1 Work *(A publisher is making available different electronic formats of the same Work as independently purchasable Objects)*

1 Package: n Objects *(A publisher is making available a CD Rom which contains multiple digital files each of which is available on the Net as an Object)*

| | |
|---|---|
| 1 Work: n Packages | *(Different publishers produce their own ISBN of "Canterbury Tales")* |
| 1 Work: n Works | *(A work with many referenced or component works; e.g. a journal article; an anthology)* |

The consequence of this analysis is that *no single identifier is capable of serving all purposes*; an identifier for a Package will have different requirements than the corresponding Work. A bookseller needs to separably distinguish the different ISBNs relating to a given title, whereas a reader may wish to consider all ISBNs of the same Work as indistinguishable; therefore E-commerce will require a *network of related (linked) identifiers*. The linkage between identifiers could be by direct referencing (resolving one identifier to another) or by metadata: for example, metadata which lists derivatives of the same Creation.

## 2.2. Work identification in publishing

Work identification is not a new concept for publishers and users of information though it has not been made explicit in the past. A *citation* is a familiar example where the Work, rather than manifestations in different Creation types, is relevant (and DOI has been seen as potentially a key tool in citatiion). Scientific publishing has traditionally used a reference to a Package identifier (printed article) to stand for the Work (citation), which was adequate when only one manifestation of the Work existed. The migration to electronic publishing is resulting in multiple manifestations of the same work (typically as a printed article and as an electronic file), leading to incompatible citation uses: there is no longer one manifestation which can stand for the work; for example, the references:

*Learned Publishing, 1997, Vol. 10 No. 2, pp 135-156;*

and *http://www.elsevier.nl/homepage/about/infoident*

refer to the same Work, but this is not obvious without knowing some additional information (metadata).

When specifying equivalence across Creation types - such as equivalent citations of different formats and manifestations - it is the underlying abstract creative work (the defining common entity which the other Creation types are manifestations of) which is the important common factor. What is needed to create unique citations (links) is an identifier of the Work. An International Standard Work Code is currently under development by ISO [ISWC] (in the CIS model the ISWC serves as a Work identifier in Music but has not yet been formally extended to other media). The ISWC, or something like it either directly from metadata, or computed indirectly from some canonical form of e.g a SICI is necessary to interchange citations regardless of medium: we would like a resolution system that resolved from ISWC-like information to the various related manifestations or vice-versa (a good business opportunity for some A&I services perhaps). Therefore there should be one Work identifier, with many Object identifiers used to identify the related Object manifestations (similar to the current position with Package identifiers e.g. ISBN).

For publishing, work identifiers are not in widespread use. The SICI identifies particular versions

of Packages (an article in two different formats has two different SICIs; it might be possible to define some canonical version of a SICI which would do this, but it is not currently covered in the SICI specification). A mechanism of Work identification which is currently available to text publishers is the Publisher Item Identifier, PII (not a formal standard but a pragmatic working mechanism open to anyone to construct from the basic definition document [PII]). Although not explicitly stated in the original PII description (because the CIS analysis was not then available), it is clear that PII is intended to serve as a Work identifier; the same PII remains with the article no matter what format it is published in or where it is in the publication chain. There was always a difference between SICI and PII in this respect [PII]. (However, because the PII was an informal consortium initiative it has not been formalised sufficiently to guarantee that every usage conforms to this aim, and PII was allowed to be used, and probably mis-used, by whoever wanted it). Essentially the PII parallels the ISWC work in a separate field. It is not proposed here that PII be the only mechanism; equivalent mechanisms might be found elsewhere e.g. production tracking system numbers. Work identification standards are seen as important to the success of the DOI (because of its potential use for citations), and therefore IDF is participating in the definition of the ISWC currently under development under ISO TC46/SC9 and will include PII in its discussions.

### 2.3 Object identification in publishing

Identifiers for Packages are well established; examples include ISBN, ISSN, ISMN, ISRC, ISAN. The SICI standard derived from ISSN can be used as a package identifier. There are no established accepted standards for the unique persistent identification of Digital Objects or for the subset of them which are Creations. The allocation of a DOI Handle to Objects solves this problem. Assigning a DOI becomes the assignment of such an identifier to the Object. This may prove to be a key role for the DOI; but as described below the scope of DOI is required to be larger than Objects and therefore the relationship is one-way: *whilst every Digital Object could have a DOI, not every DOI identifies a Digital Object.*

Consider those DOIs assigned to Objects. By analogy with equivalent package identifiers there is a requirement for those DOIs to differentiate the multiple versions of the same Work (i.e. a separate DOI for each Object which is a version). The approach taken by e.g. ISBN in the physical publishing world is that different formats have different identifiers because they identify versions which are to be distinguished for practical purposes: the ISBN of a hardback and softback are not explicitly related, they are identifiers of different Package versions of the same Work linked via ISBN agency metadata. Adopting this choice for those DOIs, two different formats (e.g. a pdf and a HTML representation of the same article) need to be separably identifiable and potentially have different rights associated with them. Therefore two formats would be two Objects, and have two different identifiers (DOIs). A practical difficulty arises in that it is much easier to generate new versions of an Object than a Package, and therefore the number of potential related Objects manifesting the same Work could be very large. The only workable definition of sameness is therefore automated bit-wise comparison (but conversely, new versions could potentially generate their own identifiers and related metadata automatically). This would allow easy ordering/differentiation of the correct version without use of human intervention

to specify; and would allow automatic generation of metadata on creation of a new format. Note that DOIs assigned to such Objects would *not* immediately be useful as citation tools, for which a separate but related Work identifier is necessary.

The issue of assigning different identifers to different format versions (Objects) is a separate issue to that of whether the format should be explicitly declared in the identifier, e.g. using a digital format extension code such as ".pdf". The approach of DOI to date has been to allocate "dumb" numbers with intelligence carried in metadata.

The conflicting requirements of a Work identifier (section 2.2 ) needed for citations, and an Object identifier (section 2.3 ) needed for trading, pose specific problems for the development of the DOI, which is required to serve the needs of both. Therefore it is essential to establish a vocabulary and mechanism which can distinguish and accommodate all types of Creations. In fact the emergence of digital media presents this problem not just to the DOI but to other systems which have up to now managed non-digital media and are now "bending" the analysis presented here: e.g. ISBNs (conceived as Package identifiers) are being allocated to electronic versions of books; ISSNs (more like Work-set identifiers) are being assigned to different electronic and print versions of a serial but with some confusion as to formats such as pdf; SICI has elements which allow it be used to identify digital media as well as physical media. It appears that a logical interoperable data model and vocabulary framework for intellectual property would be helpful to all such efforts.

### 2.4 Identification of abstracts

The issue of assigning DOIs to abstracts is of particular interest to the secondary services such as the abstracting and indexing community. Abstracts should be separably identifiable from the parent article:
1. If an abstract is not the author's original, but is either re-worded or has some additional data added (such as links to other services), the abstract will be a new Creation in its own right and therefore separably identifiable in any manifestation.
2. If an abstract is excerpted without change from an article, it is still separably identifiable by mechanisms such as SICI's Derivative Part Identifier, resulting in a new SICI. If this were included as a DOI suffix, a new DOI results.
3. Metadata would indicate that the relationship of the Abstract Creation to the Article Creation was that of a specific type of link, the component.

Note that the separate *identification* of an abstract is irrespective of the issue of rights ownership (e.g. copyright) of the abstract.

### 2.5 Dynamic Data sets

Dynamic data sets present additional problems. It is not clear yet if the same analysis applies to these; it may seem at first glance that for example a journal fits the example given above (1

Object: n Works - *A publisher selling an Object which is a compilation of several Works merged into one file)* but it seems arguable that this is true only for a fixed extent of publication (e.g. a volume or issue) and that an electronic journal itself should not be identified by an "Object identifier" on the same basis as an article; only its components should, the journal itself being a service or entity associated with the super-set of those Objects. Ongoing publications probably require identifiers of a different order: this is not a new concept, but the need to clearly define such entities in a digital environment is raising some interesting challenges. The ISSN is an existing identifier of (ongoing) serials and therefore a starting point for expansion to the digital world, and the concept of "Ongoing Entities" as Publications currently being discussed in the ISSN community as a potential next-generation identifier would define serials as only one of several classes of ongoing publication. "Ongoing Entities" includes serials (as now defined), series (both numbered and unnumbered), multipart sets and collections that are not complete as first issued, loose-leafs, and electronic resources that are intended to be added to/updated for some time (e.g., databases, Web sites, etc.) [Reynolds]. If an ISSN were to be incorporated into a Creation Identifier used for digital purposes (e.g. DOI), therefore specifying a dynamic Creation, would this be acceptable and what problems would it raise (for example, as to persistence)? An analogous issue is raised by the suggestion of assigning a single identifier to an evolving work (if so, how is the historical development captured, or is this just ignored, etc.). It seems likely that the information community will wish to transact dynamic works and therefore we must arrive at useful solutions for how the DOI Foundation can assist in this: the DOI community therefore needs to engage in this discussion about such dynamic entity identifiers. The music and entertainment rights industry may have some useful input, as the problem is analogous to the treatment of e.g. rights for ongoing broadcast series.

### 2.6 Relationship of different Creation types and publishing formats

In order to manage the various Creation types (and have services based on them), we need to unambiguously specify them individually by unique identifiers. For some Creation types, accepted international standards for information identifiers exist [Paskin] and are used in commercial transactions: e.g. ISBNs for books (Packages). Publishing currently uses mixed media: in the move from print (Packages) to digital (Objects) publishing there are many instances where the relationship between an existing Package (for example a printed article) and a digital Object (e.g. the same article in HTML) needs to be recognised. The general issue raised is that publishers are now beginning to publish the same citable work in multiple formats and versions. How should the DOI initiative deal with this?

In the discussions to date there have been suggestions to explicitly recognise these relationships by incorporating some intelligence into the DOI Handle. The DOI syntax allows use of existing numbering schemes. This has a consequence that it is possible to incorporate as a DOI suffix a number which is an existing identifier of a specific Creation type. An initial suggestion was to incorporate the SICI identifying the existing Package (for example a printed article) into the DOI of the digital Object which is the same article in HTML. At the current stage of STM publishing, it is argued, reference to the paper "original" is useful. Incorporating the SICI as intelligence is an essentially redundant, but potentially convenient, means of embedding within one

Creation type (the Object) an element of metadata (akin to the Relationship field of Dublin Core) specifying the other Creation type (the printed journal Package) manifesting the same Work. This initially seems logical; however including a SICI implies a 1:1 relationship of Package and equivalent Object. This is not true: both are likely to have multiple formats (e.g. the Object will currently typically be both pdf and SGML-related formats [Kasdorf]). What is actually important is not the relation of the Object to the Package, but the relation of the Object and the Package to the Work. The uniting factor of the printed article and the displayed HTML is that each manifests the same underlying intellectual content, the Work. The Work may have manifestations in multiple Creation types, and multiple versions of each Creation type. The number of versions of Objects is likely to increase (e.g. XML; versions having active embedded applets versus versions with inactive links; etc.).

The DOI needs to reflect the relationship between related Creation types. The choice is whether to declare the relationship explicitly, using a syntax such as *10.1000/(WORK)abcde.*. where *(WORK)* is an identifier of the Work. The problem with this approach is that there are many differing requirements for any DOI. It seems more practical to declare the relationships implicitly, via metadata (using a field such as Relation of DC), and for the DOI handle per se to remain a "dumb" number. This would require input of a DOI to reveal metadata about the object, either as a separate service or as one resolution result in a multiple resolution model. The option of a standard "metadata response page" as one possible result of a resolution request appears attractive and will be pursued by the Foundation.

A key problem which remains is the conflicting requirements of a Work identifier needed for citations, and an Object identifier needed for trading.

## 3. DOI initiative

### 3.1 Support for the DOI initiative

The International DOI Foundation began to recruit member organizations from mid-March 1998. By 1 July 1998, 23 organizations were Members. The number continues to grow [Members]: significantly, both in Membership and the recently elected governing Board there is wide international representation, and a broad spread of interests such as technology companies (e.g. Microsoft); professional publishers (currently the majority category); music industry (including CIS system members); and author and copyright agencies, representing a unique achievement in bringing such a wide range of digital technology and intellectual content interests together in a practical development activity.

Although the W3C (World Wide Web consortium) is dealing with many issues of interest to the intellectual content world, it is broader in scope and has relatively few intellectual content members. The DOI initiative takes a different - but entirely complementary - stance from W3C's focus on the much larger universe of Internet Resources. It also differs from - and again is complementary to - the Cross Industry Working Team's approach (again a body with few content providers) of Managing Access to Digital Information using Digital Objects [XIWT].

The development of the DOI system is a practical initiative; the system is currently operational and can be used now, and it is intended to roll out some prototypes of extended functionality later this year. DOI has implications for standards development; IDF is maintaining links with relevant standards organisation activities such as ISO TC46, NISO, and IETF, and DOI development will proceed in close cooperation with them.

### 3.2 Initial implementation

Although the resource mechanisms of the Internet conceptualise Uniform Resource Identifiers as a means of accomplishing some elements of Object management (both names of Objects = URNs, and their locations, URLs) only URLs have been widely implemented; a syntax has been defined for URNs. One of the aims of the DOI initiative was to put in place an identifier scheme for Objects which could be readily used by the content industries and which could be used as a URN; another aim was to facilitate digital trading (these two aims are not synonymous). The aim of the DOI system is however not to be only a URN system, but to be a system which is flexible enough to use in a variety of environments.

The DOI initiative adopted an architecture including a resolution mechanism (relating an identifier to various services or actions, including the action of location); metadata about the item identified; an administrative agency to manage the real-world business process of identifier assignment and management; and finally, an authority controlling the DOI namespace and defining policies. The relationship of these components can be schematised as a core technology (three component system of DOI + resolver + DOI metadata); a surrounding set of activities concerned with administration; and an outer layer of policies which govern all of the technical operations and administration and control the overall namespace of DOIs. The number of different components lead to confusion in terminology, compounding the confusion about scope. The confusions included *Component terminology* (was "DOI" the identifier input, the resolution system, or the administrative system?) and *Scope* (was a DOI an Identifier of a Digital Creation = Object, or a Digital use of a Creation Identifier?)

The current resolution mechanism used by the DOI is the Handle system devised by CNRI [Handle]. Since that mechanism allows resolution to multiple data types, and the selected type can be determined programmatically, services associated with the DOI handle (the name, within the DOI namespace, of the Creation being referred to) can be intelligently allocated as a client/server transaction: of course, the client must identify the types of data that it can deal with and the server must identify what is being returned, i.e. a protocol agreed. A Handle can return multiple data types but does not directly support service requests (e.g. "tell me the format of this entity"): service requests are dealt with as separate data types, or as arguments built on data types. Handle can be used in conformance with the URN syntax.

The authority controlling the DOI initiative is a not-for-profit Foundation (the International DOI Foundation, IDF); administrative and metadata systems are being put into place under the guidance of the Foundation as part of an ongoing development. Funding for the Foundation is from organizations with an interest in creating such a system; eventually a cost-recovery operation

of DOI system for the administrative agencies is envisaged. It is not necessary to be a member of the Foundation in order to use the DOI system.

The DOI Handle syntax conforms to the URN syntax, according to which it would be acceptable to create a URN namespace for DOIs, so that "DOI URNs" would be:
*urn:doi:<DOI-as-it-stands-now>* i.e., when appearing as a URN, the "urn:doi:" is required to distinguish the identifier as such. Some issues remain to be finalised within the syntax (currently under review by NISO) such as whether non-allowed characters (such as <brackets> which appear in identifiers such as SICI) will be included. If so, the URN syntax will still be followed, so that "DOI URNs" would be: *urn:doi:<doi-as-it-stands-now-with-special-chars-replaced>*, i.e. the DOI remains as the DOI syntax would require, with the understanding that URIs (still) have syntax restrictions which require certain characters to be escaped in order to not confound parsers.

### 3.3 DOI prototype rules and consequences

At the launch of the DOI, two decisions were taken for prototype applications which have carried through into the initial implementations: although not formally codified, the effect was to impose the following unwritten rules:

*(Rule 1)* Few restrictions were placed on use, so as to encourage experimentation and define what was needed for the next steps in a very pragmatic way. It was openly allowed that a publisher could allocate a DOI to anything he found to be useful; and that DOI could resolve to any URL. As a result, prototype uses [DOI gallery] included locating (resolving to) directly an Object, and also a variety of digital services: typically digital access points, to either digital Objects (access screens, order screens etc), or to resources acting as access to non-digital Creations (order forms for books). (Location is in itself a service, but one which is trivially and readily available via URL and therefore not usually considered).

*(Rule 2)* Although the Handle technology was being used, tight restrictions were placed on the extent of the Handle technology deployment: only one data type (URL) was allowed, and only one instance of that data type; hence, one DOI = one URL. The Handle technology has the capability of resolving to multiple data types but this was not used in the DOI prototype.

Both rules were perfectly reasonable starting points. But it is now clear that they need to be refined and evolved. The consequence of a combination of the results of (1) and continuance of (2) led inevitably to trying to make one URL access point do everything. It also then led to confusions, since the Handle input (the DOI number, on the left hand side of the resolution model) was seen as synonymous with the (single) URL output (on the right hand side of the resolution model): sending that input led to that URL which is therefore (wrongly) "what the DOI identifies".

a. One DOI could resolve to one URL which was a service, rather than an Object: e.g. resources acting as access to non-digital Creations (order forms for books). This lead to the mistaken view

that *"this DOI identifies this service";*
b. Another DOI could resolve to an Object. This led to the view *"this DOI identifies this Object"* and therefore (a seductively easy step) "this DOI identifies this Creation in all manifestations"- even more confusing because some uses had this intention whilst others did not; most had not addressed the issue of multiple manifestations.

The confusion is clarified if we allow resolution (conceptually) to be to multiple data types and data values. Then the Handle input (the DOI number) is no longer synonymous with a single output and it becomes easier to talk of the DOI number relating to a Creation (on the "left" of the resolution model) and resolving to outputs which might be services or other Creation types (on the "right" of the resolution model). The fundamental cause of these confusions is now seen as a failure to separate content (identification) from services (resolution result)

### 3.4 Development path of the DOI system

DOI is not a closely-defined standard. The allocation of a prefix to an organization does not currently prescribe the scope of any application of the resulting DOIs, other than to application in the broad area of intellectual property content and according to the stated terms and conditions for prefix-holders. It is not the intention that the application of DOIs by prefix-holders be proscribed or "policed"; under the current implementation approach this would in any case be difficult to do. However guidelines and best practice statements are required.

Thus far the DOI has progressed by a practical "individual" approach: DOIs are allocated by each individual organization and the applications and resolutions are determined by that organization alone. However as we move towards a more sophisticated Handle implementation with multiple resolution, it is necessary to define some standards which will allow interoperability of applications and encourage third-party application use. This would suggest a "data model" approach, which enables applications built on a common vocabulary. The "individual organization" approach encourages speed and flexibility of implementation, and valuable experimentation, and is akin to prototyping. The "data model" approach is slower but more readily guarantees interoperability and is more akin to a standard application development route from requirements definition through to application building.

The IDF is encouraging a parallel tracks approach to development which will allow the best of each: "individual" implementation is supported, whilst in parallel there will be planning of an interoperable framework (which will learn from the results of the individual approaches), offered for more sophisticated future uses (such as level 2 DOIs described below). This will result in a core common vocabulary both for services and for differentiating content types (so that it is clear what is meant by "version", for example). In information systems terms, we hope to stimulate a "rapid application development/ evolutionary prototyping" approach, in which applications built using the "individual" thinking should influence, and be compatible with, more sophisticated "data model" approaches.

## 4. Standards activities and the DOI

The DOI is not a formal standard, but is a system which is offered in conformance with related standards. It is the intention that DOI develop in conformance with related evolving standards from both the content/information viewpoint and the technology viewpoint

### 4.1 Standards from the content / information community

DOI should be interoperable with other systems via International Standards such as those of ISO for identifiers of creation types (ISBN, ISSN, ISMN, ISRC, ISAN, ISWC, etc.): e.g. a result from a DOI resolution which specifies an ISBN should be usable in an ISBN context. The International DOI Foundation (IDF) is now a member of NISO (National Information Standards Organisation) which is both mandated by ANSI as the official US ISO body in this area, and also has a wider international role in information standards. ISO/TC 46 /SC 9 is currently discussing two standards (ISAN, ISWC) and IDF is to be represented in the Working Group on ISWC which as noted in the discussion on Creation types is considered especially relevant. The syntax of the DOI Handle has been accepted as a work item by NISO and is currently under discussion.

The Foundation is an affiliate in the EC-funded INDECS (Interoperability of Data in E-Commerce Systems) activity (which also includes some IDF members), and is in active discussion with the ISSN Centre. IDF will also participate in the forthcoming IMPRIMATUR forum on Standards for the Intellectual Property Businesses. A summit meeting on metadata is currently being planned by a team representing DOI, CNI, Dublin Core, UKOLN, Data Definitions and INDECS.

The adoption of DOIs for Objects creates a solution for the identification of Creations which are Digital Objects. We will consider putting this through formal standardisation, e.g. through ISO via the current NISO work item on DOI syntax.

### 4.2 Standards from the technology / Internet community.

The principle technology standards activities relevant to DOI are those of resource addressing, in particular the URN (Uniform Resource Name) and URI (Uniform Resource Identifier) concepts. Standards discussions of Internet related issues take place in both W3C and IETF. IDF has not joined W3C at present, since we are unable to make the necessary commitment of practical involvement, but is represented by our technology partner CNRI and two of our current members (Microsoft, Elsevier) as active members.

It is intended that the DOI implementation using Handle be interoperable with Internet infrastructure standards. Handle has been submitted as an Internet Draft to IETF [HandleID] and is fully documented in terms of both resolution and administration protocols [Handledoc]. Handle System namespace is based on Unicode 2.0 which therefore allows (potentially) a wide range of characters to be used, although in current implementations (which are by proxy servers to "translate" into the currently supported http protocol standards) it is restricted to the ASCII

character set supported by URL syntax.

We will consider whether it might be helpful to develop a separate IETF RFC on the specific DOI, once some of the questions described in this paper have been answered.

There has been some confusion about whether or not Handles, and hence all current DOIs, are URNs. This results from differing terminology and paths of the URN development efforts: "URN" is defined either generally (*URI that has an institutional commitment to persistence, availability, etc.;may also be a URL e.g. PURLs* [W3cAddress]) or precisely (*a URI scheme defined in URN Syntax (RFC2141) and its related specifications* from URN Working Group of the IETF). Historically, the general notion of URN naturally preceded the specific work of the IETF URN working group. Handles were one of the first URN implementations (the Handle System largely preceded the work of the IETF group): Handle clearly fits the generally defined functional requirements of URNs (persistence and availability) and so a Handle, and hence all current DOIs, are examples of "generic" URNs. The URN Working Group has adopted a framework approach which could accommodate a variety of resolution systems; the URN syntax specifications call for each URN to consist of the prefix "urn" followed by a namespace identifier followed by a namespace specific string (*urn:nid:nss*). URN resolution would consist of two steps: using a top level Resolver Discovery Service to identify a resolution service for *nid* and then using that service to resolve *nss*. The various name spaces could each have their own resolution services, which could change over time. Currently there is no widely available end-to-end URN resolution service; the Resolver Discovery Service is currently built experimentally using NAPTR which uses DNS. The question of whether or not Handles and/or DOIs could become specific URN namespaces (e.g. *urn:hdl:10.123/456* or *urn:doi:10.123/456*) cannot yet be answered as the URN namespace registration procedures are still in draft form; only time will tell if this becomes a viable and attractive option. The IDF is neutral in these matters and will support whatever framework leads to clear interoperability and ease of use.

Handle appears to have some advantages over the specific URN syntax in terms of internationalization: URN mandates ASCII encoding, while Handle mandates UTF-8 encoding there is no significant difference between the two syntaxes when used within the protocols but when used for user input, UTF-8 is more efficient and more acceptable for non-ASCII alphabet users; however handle protocols are not yet widely supported in browsers. A Handle System Resolver exists as an extension for web browsers (NetScape 3.0+. IE 3.0+, for PC and Mac with Unix to be developed soon) enabling them to talk directly to the Handle system (avoiding the http proxy server step), downloadable with a clickable license [CNRI]. The IDF will support efforts to extend this functionality to browsers as a tool which would enable efficient DOI resolution.

It is also possible that URN specification and definition will become less important and that attention may focus in favour of a URI specification, for which there are a number of proposals, all in draft stage, which appear to be converging. The basic proposal is to have number of alternative trees of URI types, e.g:
*1) IETF* - http, ftp, mailto, telnet, etc. RFCs are required; probably protocols, although not rigid, in case someone can come up with something at that level that happens to be something else (DOI would probably not fit here). At this level these are assumed to be public standards not owned by

anyone other than the IETF process.
*2) Vendor* - use "vnd.vendor-name" as a prefix, e.g., vnd.microsoft:<some scheme>. A variation on this is to skip the vnd and to reverse domain names instead, e.g., org.DOI:<some specific DOI>. The examples are all some commercial operation, e.g., com.worldcom:, but DOI could fit here. No standards track RFC required, although informational RFC encouraged.
*3) Personal* - use any name, with prefix prs:
The assumption behind the categorization is that the schemes that qualify for the IETF Tree will be universally supported in web browsers and other common clients, while those in numbers 2 and 3 are intended for increasingly specialized audiences. It is assumed is that IANA, or some similar organization, would keep the registry. Until there is a clearer indication of the likely direction of Internet standards (e.g. as a result of the recent DNS namespace controversy), it does not seem advisable for IDF to make specific proposals, but we will continue to monitor these discussions.

The IDF is also monitoring discussions in the MPEG community (MPEG 4 and MPEG 7) and XML community, and is meeting with representatives of the RDF initiative of W3C [Miller], both in reference to potential metadata standards.

## 5. Scope of the DOI initiative

The focus of the DOI initiative is on Content. The purpose of the DOI initiative was to provide enabling technology for trading in digital publishing. In this context the term "trading" is used here to indicate any transaction (commercial or non-commercial) which is required in the information (= intellectual property, or content) community. A question which has not been dealt with adequately up to now is whether DOI is a digitally-enabled identifier of objects; or an identifier of digital objects. The answer is both; this is a key concept which expands the DOI beyond any other service dealing with digital information only.

In trying to establish guidelines for what should be the scope of the effort (what content or information should, and what should not, be persistently and reliably identified), consensus has been reached [DOIarch] on the following:

a. The focus of the effort is on content (Creations), rather than general E-commerce: *the scope of DOI is confined to intellectual content.*

b. A useful distinction is between primary Creations which are persistent products issued by a publisher (e.g. articles), and related peripheral items (e.g. order form for the article) of the content provider. The primary Creation should have an information identifier; but the order form is an incidental instantiation of a service associated with a primary entity, and it is that primary entity which is the reason for the entire exercise. An information identifier focuses on the primary product, and is intended to be of uses in enabling services to be offered for those primary products: *the DOI is intended to identify intellectual content, not services associated with the content (but these services may be pointed to by the DOI resolution mechanism).*

c. "What the DOI identifies" is not the same as "what the DOI resolves to". This is not easy to

see in the 1DOI=1URL model, but recall the wider resolution model in which the Handle input (the DOI number) is conceptually not synonymous with a single output. DOI number relates to a Creation (on the "left" of the resolution model) and resolves to outputs which might be services or other Creation types (on the "right" of the resolution model).

d. A limitation only to digital manifestations whilst ignoring non-digital (physical and abstract) creations is unhelpful and inappropriate: *the scope of what DOI identifies is defined by meaningful content, not by digital manifestation*. However it is likely that digital manifestations (Objects) will assume increasing importance over time.

e. Since Creations are manifested in different types, the same primary product content (Work) may be manifested in both digital and non-digital forms. Each manifestation (including the abstract Work) is capable of being traded digitally (e.g. a non-digital book may be ordered digitally; a digital article may be accessed digitally; rights in an abstract work may be sold; an abstract work may form a citation), and therefore *each manifestation must be identifiable separately*.

In the community of interests which supports the International DOI Foundation there is a requirement for both "Digital Trading of Creations" and "Trading Digital Creations". ***Digital Trading of Creations*** requires systems which enable trading on open networks (i.e., the Internet, rather than private EDI networks); the provision of services which lead to transactions about Creations, of any type. For example, making available a digital access point which refers to a specific ISBN (physical book) and leads to a digital transaction resulting ultimately in the purchase of that physical book by placing an order on-line. Also included under this context is "digital trading of digital Creations" (Objects); for example, making available a digital access point which refers to a digital Object in HTML format and leads to a digital transaction resulting in access to that Object via a browser. Digital trading of Creations is a subset of Digital Trading in general; that is, of E-commerce, and will have shared requirements. Trading of Creations requires unambiguous persistent identifiers of each Creation. These exist for non-digital Creations (e.g. ISBN, SICI, ISAN, ISRC, etc;). Therefore when we are trading items such as books, it would be sensible to use existing identifier schemes such as ISBN as the identifier which is the label for the entity being transacted, and which are interoperable with other services. ***Trading of Digital Creations (Objects)*** refers to trading of that subset of Creations which are Objects, i.e. manifested digitally, and requires unambiguous persistent identification of each Object. Identifiers for non-physical Creations (Objects, Works) are less pervasive than for physical Creations (articles, books) and accepted international standards do not exist for the whole class.

There are two different requirements for these two different activities. ***Digital Trading of Creations*** requires a system which takes as *input* a Creation Identifier and responds with *output* of a service (transaction) related to that Creation. It therefore requires a mechanism to take input and deliver output implemented on the Internet: resolution. ***Trading of Digital Creations (Objects)*** requires development of an accepted standard for a Creation Identifier applicable to this class of Creations which are manifested digitally (Objects); this need not have a specific implementation. However when the DOI system is applied to Objects (i.e. a DOI Handle is assigned to an Object), an identification scheme is created which serves this purpose (and which

could also be used outside the DOI system as a naming scheme).

It is worth adding three notes to help in the interpretation of some of the concepts outlined up to now:
(1) "Digital Objects" of the technology schema are not equivalent to the "Objects" of the Creation content schema. The technology concept of Digital Object is very broad and there seems no way to differentiate usefully at that level between a journal article (Creation) and an order form (a service function). While it is reasonable to say that publishers should assign persistent identifiers to their primary output but not to their peripheral items, there is no way to tell those apart at the structural level. So we must distinguish by role or function. The concept of "Creation" then becomes very useful in implying a higher level function description: "products of human imagination and/or endeavour in which rights may exist".
(2) The distinction of Internet Resource, Digital Object and Creation is not a wholly clear one. They are overlapping but not mutually exclusive sets. The definitions used here are insufficient for legal definition, or for indefinite applicability, though sufficient to clear the way forward for progress in our limited area of concern at present. It could be argued that even an order form is "product of human imagination and/or endeavour in which rights may exist", and advertising certainly is a Creation. Therefore we should not aim to remove entirely the original working definition (or non-definition) of DOI scope, which was that a DOI could be applied to whatever its assigner wished; but we should distinguish Creation (DOI allowed) from service (DOI disallowed). Examples help to show the distinctions: a simple web order form to enter credit card details is a Resource (it exists on the Web); it is also a Digital Object (it exists on the Web, and it is certainly a meaningful set of bits, and it can certainly have some unique identifier); but it is not a Creation (by analogy with the package world, a printed book has an ISBN, a printed order form does not). An order form is a Resource which is a component of a *digital service* performed on an Creation. A journal article on the Web is a Resource, a Digital Object, and a Creation. A printed book is a Creation, but not an Object.
(3) It might be argued that digital services could be artificially treated as Objects in their own right, and have a DOI assigned to them: the equivalent of giving an ISBN to a printed order form for a book. However this then uses DOI (as an identifier) to identify both Objects and Resources offering a service performed on that Object (rather as if the early days of ISBN had allowed ISBNs to be assigned to absolutely any physical object, from a private notebook to the bookshop building). This is a confusion from the 1DOI=1URL model, and it is worth stating again that the Handle input (the DOI number) is conceptually not synonymous with a single output: DOI number relates to a Creation (on the "left" of the resolution model) and resolves to outputs which might be services or other Creation types (on the "right" of the resolution model)

## 6. Development of DOI-based services

### 6.1 Distinguishing content, services, and mechanisms

The focus of the DOI initiative is to enable services for the management of intellectual content in digital form. Irrespective of whether this is interpreted as *Digital Trading of Creations* or *Trading of Digital Creations (Objects)*, a critical distinction which has become apparent from the

prototype work is the need to separate content, services and protocols:

- content must have an identifier specific to its Creation type; the DOI must offer to the resolution service an identifier which denotes a Creation.
- services: actions offered when a DOI is resolved by a specific service (e.g. resolve to a location; display the metadata; buy the Object; etc). Services may be viewed as the verbs of the E Commerce language, now being defined in initiatives such as DPRL [Xerox];
- protocols or mechanisms to implement specified services. Protocols could include defined (existing or new) data types within Handles, and mechanisms such as scripts associated with existing data types (e.g. URLs).

Identifiers will be permanent (persistent); the services offered may evolve over time; and the mechanisms implementing the services may change over time.

Services can be considered as actions to be performed with reference to content (a Creation); the Creation is denoted by the DOI. If these actions are described by the verbs in an E-commerce language, then the prototype DOI system limitation of 1:1 means that our DOI vocabulary effectively has only one verb: "get" - the DOI is associated with one URL and obtains that URL via a http proxy using the http "get" command. In essence then the DOI prototype offers little more than other "get" redirection services for persistent URLs [PURL]; to achieve the full potential we need to offer mechanisms which can implement a variety of services.

At each level (content, service, mechanism) there is a degree of independence from the next layer; i.e., it is not sensible to let the currently available services influence the identifier too closely nor the current mechanisms influence the services too closely. We should not build a service because the technology permits it, but instead first think of what services are needed and then look for mechanisms which can supply them. In the next stages of DOI development we will be producing a list of services which we need, and then we will be looking for mechanisms to implement them. If the current version of the Handle system and associated tools can do the implementation, that mechanism will be used; if not either we will need to find alternatives or (more likely) new tools, procedures, etc. will be specifically developed to fit our implementation request. DOI is not by definition tied to Handle as a mechanism, but Handle was designed as a flexible infrastructure able to accommodate such requests. Both CNRI and the DOI Foundation are keen to work with open standards and to integrate our efforts with other efforts.

### 6.2 "Level 1" and "Level 2" DOIs

This terminology is introduced here as a shorthand for distinguishing the possible DOI uses of the Handle (or other resolution system). There is a "quantum leap" in the possible application in moving from one resolution result to several resolution results.

*Level 1* refers to the model where one DOI = one URL, i.e. the Handle record is restricted to one instance of one data type (URL). This is the current implementation. Typically the DOI resolves to a response screen from which is offered a choice of services. The services are not offered within the DOI system as such, but the DOI system leads to the starting point for those services, which are then chosen by human intervention of the user. The DOI leads to the front door. Level

1 DOIs are limited to the "get" paradigm, where the assumption is that *identifier = hyperlink = click on it = get it*. This resolution system and identifier system allows only one action: returning an Object. It is therefore difficult to group objects, differentiate versions, formats, etc. The DOI does at least allow persistence of the identifier resolution, a significant improvement on URLs which are unreliable in current uses with intellectual property such as journal articles (e.g as many as 50% of URLs are inactive [Ford/Harter]), but not a significant improvement on other solutions such as PURLs [PURL].

*Level 2* refers to the model where one DOI = multiple data types, i.e. the handle record holds one or more types and/or type instance. This is the expansion to the full Handle capability referred to above. It allows the ability to resolve to one or several points; in principle this could allow automation of the services offered against the Creation to which the DOI refers; the DOI leads not just to the front door, but also to inside the building. This allows the "services paradigm": complex resolution; selection from a group of values associated with one DOI; services such as "describe" (metadata), "purchase", etc. It requires an ambitious program of development but offers significant advantages if successful.

It is entirely possible (and likely) that level 1 and level 2 DOIs would co-exist in the DOI system, both globally and within individual prefix-holder applications. Level 1 DOIs exist and are usable now without additional software support. Level 2 DOIs require the definition of services and the development of application mechanisms, probably with additional software requirements, but offer much greater potential (beyond the "get" paradigm and persistent URL implementation, moving to a "services" paradigm).

Level 1 DOIs are likely to be progressed by the "individual" development approach, with some minor benefits from a data model framework. Level 2 DOIs require a more rigorous data model of the type we are now beginning to deploy.

### 6.3 Metadata and look-up / search services

When a DOI is assigned, we need to collect some metadata about the Creation denoted by the DOI: at the least, it will be necessary to have a minimum set of metadata which enables a service of: "given this data I have about the Creation, tell me its DOI". The DOI Foundation sees the definition of that metadata set as an urgent task. Relevant work is going on in several places: a metadata working group with EDItEUR; a prototype metadata service from one DOI prefix owner (Wiley); a draft proposal for a base metadata set from NFAIS (National Federation of Abstracting and Information Services); and planning by proposed Directory managers such as ISBN International with whom IDF has signed a letter of intent.

It is clear that once such metadata has been established it could serve many useful purposes, not least in linkage between related Creations (e.g. between a Work manifested in one form and in another form; or between a Creation and its Reference links), and that therefore there are many opportunities for devising added-value services and working with existing abstracting and information services. Current thinking focuses on a "bare bones" DOI metadata set such as

assigner, Creation type and information identifier (if referring to a non-Object), enumeration (journal, volume, issue, page, and possibly first author), format.

It is useful to make a distinction between “look-up” and “search”. To provide a fragment of information (e.g., a piece of metadata) and request an identifier is a search; to provide a complete key (e.g., an ISBN) and request a corresponding DOI is a "look-up". The distinction is important because the two different kinds of services are very different in practice. There may be one or several look-up services in order to provide a whole system (e.g. the Domain Name System is used to look up addresses for a particular machine name). Search services tend to be more successful as "value added" services (e.g. searching for a company name in Alta Vista). DNS does not work as a search service; conversely AltaVista is poor as a look-up service (retrieving many hits when you only want one).

It is not possible to define all the metadata about a specific Creation (since it is potentially infinite). Some (but not all) metadata collection can be automated. All metadata must be well-structured to be re-usable in a variety of contexts. It is unlikely that we will agree on a single namespace authority for each possible element of data (and thus metadata); but it is achievable to require that each metadata element follow a well-formed syntax and namespace, and explicitly declare its namespace authority so that a standard mapping could be invoked for “metadata cross-walks” or mappings from one data set to another. DOI metadata elements should be well defined and follow existing standards (i.e., namespaces).

Outside some specialised areas no mechanism is available to attach metadata to Objects. The Resource Description Framework RDF [Miller] offers an answer to this but there will be issues of how fast will it be implemented; as RDF uses XML as its interchange format, metadata implementations will require a widespread update of the installed browser base. RDF is however already a useful way of system analysis of a set of metadata and the IDF is in close contact with the RDF initiative.

### 6.4 Nesting DOIs

In building potential services and interconnected identifiers, note that one DOI may refer to others. A single DOI for a work could resolve to a piece of metadata plus the component parts, e.g. a DOI for a conference proceedings which referenced 28 DOIs for papers. This concept, used in conjunction with level 2 DOIs, may also be useful in a possible design which solves the key issue of conflicting requirements of a Work identifier (section 2.2 ) needed for citations, and an Object identifier (section 2.3 ) needed for trading. It would be possible to nest identifiers such that a Work identifier resolved to component Object identifiers (manifestations):

*(see figure next page)*

| *Name* | *Data type* | *Data value* |
|---|---|---|
| DOI of Work | DOI | DOI of Object (pdf format) |

| | | |
|---|---|---|
| | DOI | DOI of Object (HTML format) |
| | | etc. |

### 6.5 Providing services against identifiers

There are multiple possible designs for implementing services against identifiers. The goal of the IDF will be to define the vocabulary and syntax for those services such that they could work across various designs, perhaps at the same time but more likely over time. This is an ambitious program which requires some considerable work to define the alternatives in operational detail, and to define the optimal path from one stage to another.

This section explores some of the options. Assume you want to associate multiple things, such as services, with a single identifier. Further assume that you have a separate resolution system and a separate generic "repository" system (which could be, for example, web site). There are two choices, in terms of system architecture (the choice should be transparent to the user and we should be able to switch from one to the other over time without changing the syntax of the identifier):

1. All the possible services are known to the resolver; the client software must know how to ask the resolver for the right service. The answer to a resolution query is a direct pointer off to some repository, which the client only needs to understand in some elementary fashion (e.g. it can use "standard" web programming).

2. The resolver does not know about the separate services; it only knows where to go to ask about all the services. The client has to understand how to take that simple answer and turn it into the more complicated service request to the repository.

This could be done, in a somewhat crude fashion, with web sites. Assume a known vocabulary of services and an agreed upon syntax, e.g. *pdf@10.123/456* is a request for a pdf file related to a DOI identifier for a Work. The Handle system contains a single (or multiple parallel) URL(s) for 10.123/456 with a value of *http://pub.com/cgi-bin/doi-services* (parallels would be e.g. *pub1.com, pub2.com,* etc.) and the client knows enough to take that and turn it into another URL of form *http://pub.com/cgi-bin/doi-services?doi=10.123/456&option=pdf* which it sends off as an http get. The organization that ran pub.com would have to have built a script or program of some kind (called doi-services) that parsed the arguments and knew enough to send back a pdf file (or a message to the effect that there is no pdf file and pdf@10.123/456 doers not exist).

An alternative to creating and implementing such a list of services *de novo* is to adopt an existing repository structure. CNRI has developed the Handle system as an abstraction of the process just described, both at the client/server interaction level and, not apparent here, at the data storage level. Handle becomes part of a digital object architecture including a repository scheme. This uses a protocol RAP (Repository Access Protocol) and a data type DLS (Digital Library Service) inside the Handle system (this is parallel to http and URL) which can provide these services. In

this case, method 2 assumes that the client knows the service it wants and it knows a protocol (mechanism) for accessing the particular repository (RAP); a query obtains the location of the repository (Digital Library Service) from the Handle system and then interacts using RAP to isolate a PDF version.

### 6.6 Practical steps for development of services

The way forward is therefore to further evolve each of the initial starting points of the prototype: - have the DOI be applicable to any Creation (but restricted to only Creations);
- deal with the services required to be performed on the Creation not by the DOI syntax but by other means, the *resolution* mechanism (i.e., lift prototype implied rule (2)) and metadata.

Fortunately, the use of the Handle implementation as the current DOI resolution mechanism means that lifting the restriction on single data types is possible (moving from “level 1" to “level 2" implementation). The DOI implementation of Handle extended to the full Handle functionality does however require appropriate tools designed to deal with the extended functionality, which is an important decision point in the DOI development. In keeping with the policy of separating services and mechanisms, we will define services ("verbs") independently of the Handle system until we understand what is needed, and then see if that maps on to the existing Handle implementation (i.e., determine what we want to do and then how to do it, instead of observing what can be done with the current tools, and determining which features could be useful in the current context).

The recognition of services and a variety of mechanisms needed to deploy them, and the current implementation status, suggests the following practical way forward for the DOI initiative:

- defining a finite but growable list of standardized services defined for the purposes of the “DOI community”. It will be helpful if we agree on an initial list of standard “verbs” for our initial vocabulary of services which can then be expanded in a controlled manner;
- expanding the DOI mechanism to resolve to not one URL but multiple URLs, where URLs are an appropriate mechanism to provide a service (e.g. “locate” from a list of possible locations); in doing this we recognise that having clients pick from multiple resolutions is not a trivial task to do well.
- expanding the DOI to resolve to other data types where the Handle and other data types provide appropriate mechanisms to deliver the required service. These could be:
  (1) additional standard data types (e.g. e mail)
  (2) additional data types already in use in a specific community but not yet widely accepted as a formal standard (e.g. the Repository Access Protocol, RAP)
  (3) newly defined data types, as allowed in the Handle (data types >65536) which the DOI community agrees to introduce on a pragmatic or prototype basis as consortium standards and could perhaps later become official standards. The IDF could consider defining some other data types, such as "metadata record": the mechanics of Handle (DOI) resolution consists of clients sending a message to the system asking for all data records with specified data types; the default is all data records. To access e.g. metadata record, the client would ask the system for all data records of type "metadata record”. This

would require clients which are able to understand the new data types intelligently, and thus possibly some form of plug-in built from e.g. the Handle client library. It will be helpful if we agree on an initial limited list of data types for our initial services implementations, which can then be expanded in a controlled manner. This would provide an incentive for application builders to create DOI-useful client software;

- making available plug-in browser support tools (such as the CNRI Handle System Resolver plug-in [CNRI]) which provide improved capabilities for deployment of services beyond the http protocol. (The DOI system will always be designed to be usable on commonly deployed protocols, but specific plug-ins might be offered to improve on that basic level of service functionality.)
- encouraging the development of further browser tools to give "intelligent" client capability for use with level 2 DOIs.
- defining a common standard for metadata to describe DOI-identified entities.
- defining a means of associating related identifiers (e.g. of a Work and its manifestations)

With a standard list of services defined, we can talk about how to signal which service the client wants, and we can then talk about how that service is supported for a given DOI. In the context of URN work some basic services have been identified which seem to be useful for resolving URNs: requesting a list of URLs known to be associated with a given URN, metadata associated with a given URN, etc [URNres]. These are not the full list of services required for e-commerce but serve as an illustration of how to deal with the problem.

The IDF will initiate work items which focus on defining the items above. It will also expand the current DOI application to the full Handle capability of multiple data types, and promote the development and deployment of prototype client software for intelligent manipulation of resolution results.

## 7. Guidelines for present DOI assignment

We wish to promote the use of the DOI to identify Creations. Undoubtedly we will wish to mesh Creations with the greater world of Digital Objects/Resources (just as in the package world, a publisher will produce both books and order forms, catalogues etc. for those books). The latter can be "services" for those packages; but only in the digital world does it become necessary to make this explicit and rigorously define such services: in the analogue world we can live with without worrying about such definitions. Note the corollary of this: just because a publisher is using DOIs and DOI-based services does not mean he should use them for everything; he should not replace all his URLs by DOIs.

Given that the DOI system is still evolving; and that this analysis indicates the need for expansion to other and multiple data types before services can be adequately constructed, it is not surprising that no final recommendations can be made which will suffice for all future developments. With one DOI resolving, at present, to only one URL, what should publishers who wish to experiment with DOI systems consider as a sensible way forward? The following guidelines are offered as preferences which if followed should fit with likely expansion of the DOI system:

- 1. Assign DOIs to *Creations*, but not to all Resources or Digital Objects, and not to Services
- 2. Assign DOIs to *Objects* and to *non-Digital Creations: Works and Packages*.
- 3. Follow the data model and terminology outlined in this paper, e.g. assign different DOIs to different digital format versions of the same Work (which are different Objects).
- 4. The current resolution of the DOI may be directly to the Object, or to a URL which acts as a "shop front" to a Creation; that is, currently services are treated as DOIs resolving to Resources which act as intermediates (order forms, access forms, etc.), requiring manual intervention by the user. (Guidelines for response screens may be a possible activity for the DOI Foundation at this stage). At a later stage these resolutions to intermediary URLs may be replaced by additional data types, other DOIs, or other URLs to accomplish a full range of automated services.
- 5. Do not assume that every existing URL of your digital Web presence must be "replaced by" (i.e. be resolved to from) a DOI, since much of this web presence may be Resources not Objects. Consider the advantages of each mechanism and the likely usage and persistence of the entity you wish to offer on the web.
- 6. Do not confuse what the DOI identifies with what it resolves to: to help in this, adopt the model of "level 2" multiple resolution in your thinking, even if at present we are only assigning one resolution value.

The development of additional functionality for the DOI (defining a list of required services with specifications; expanding the current DOI application to the full Handle capability of multiple data types, and promoting the development of prototype client software for intelligent manipulation of these data types) are priorities.

## 8. Who assigns a DOI?

Terms and conditions for assignment of DOIs are available on the DOI web site [Terms] and include the conditions that DOIs may only be assigned to items where the assigner has the right to use such Content in the DOI system; this is not the same as ownership since such rights may be assigned e.g. as part of a service offering by a third party . The Issue of "ownership" of a DOI was raised during the May workshops: who has the right to assign a DOI? It has been proposed that the DOI service should in fact be called a "Publishers DOI" [Caplan2] because it originated in a particular application area related to managing rights and permissions. In fact the DOI should be no more publisher-centric than the ISBN service. ISBNs are assigned by publishers for books which they publish, even if the content is copyright of another party; the publisher has been assigned specified rights of publication and ISBN assignment is part of that publication process. Similarly, DOIs relate to Creations which are Objects in which rights pertain; the rights are initially entirely with the Creator, but certain rights may be licensed or granted to a publisher for digital publication, and as part of that process assignment of a DOI would take place. Thus the question of who owns a DOI is no different to who "owns" an ISBN.

There is a general need for persistence that goes beyond objects for which the DOI application

is primarily designed; for example, the thousands of user-created web pages on a campus-wide information system all have URLs, and they all move around. It has been assumed that they are very unlikely to be assigned DOIs or managed through the DOI application, but there is no reason why they should not be; and indeed there are some advantages to having a fully interoperable content management schema for both "internal" and "external" documents. Libraries and communities looking for persistence in other contexts might want to turn to Handle technology per se, the DOI implementation of handle, or other components or lessons from the overall DOI solution.

An issue which remains to be resolved is that if a Creation is licensed for publication in more than one way (e.g. to two publishers), there may well be more than one DOI assigned to a work. Does this matter, any more than a single book may have multiple ISBNs?

## 9. Who uses a DOI?

As with other identifiers, once a DOI is assigned it can be referred to (used) by anyone. The assigning publisher will wish to use it. However, it would also be possible for intermediaries to construct services which are based on DOI which might offer services such as "is this DOI available from your service? If so under what terms (etc.)", just as book wholesalers build ISBN into their systems for books. There would be no costs levied by the International DOI Foundation or agencies for the use of DOIs which were already publicly available.

## 10. Business model outline

The intention of the International DOI Foundation is to assign one or more agencies which would be responsible for administrative aspects of the DOI system, such as allocation of prefixes, registration of DOIs, and policing the maintenance of allocated DOIs. An initial intent is for the international group of ISBN agencies represented by ISBN International to act as one such agency [ISBN]. The operation of such agencies is to be on a cost-recovery basis, which requires a fee to be levied at some point in the working mechanism to meet the necessary costs both of these agencies and the technical infrastructure of the Handle implementation. Analogous to ISBN allocation by those agencies which are commercial entities, the costs are to be borne by the assigner rather than by users.

An initial fee structure of $1000 (as one-off payment) for the allocation of a prefix was introduced early in the project to indicate that the service would not be free to assigners, and to provide a barrier to prevent frivolous experimentation. The level of this fee will be reviewed, to accommodate both large and small potential users.

In addition some early modelling of costs suggested that it may be appropriate to introduce a fee on an annual basis for each DOI held in the service, to cover the costs of maintaining appropriate resolution infrastructure and management etc. It is recognised that this fee must be small enough to allow assigners to use potentially many thousands of DOIs without undue inhibition on grounds

of costs, and indeed it has been set at zero initially. The level of such fees may vary with each agency and possibly the business model would differ among different intellectual property types in line with accepted practice and the particular infrastructure of an industry, so that e.g. professional information publishers may operate on a different basis to the music industry. This need not matter providing that each conforms to an overall set of DOI rules for interaction. At present the International DOI Foundation is in discussion with potential agencies and these will need to generate their own business model proposals.

There will be unavoidable costs for the DOI system. A standard on its own will not give us an implementation, and an implementation costs money to someone somewhere; the task is to make it efficient.

## 11. Conclusions: the way forward

The International DOI Foundation defined a moving list of ten work areas as a result of workshops held in May 1997 [Workshop]: the clarifications outlined in this paper help to extend and deepen that list, and contribute towards answers for the questions relating to ownership, scope, valency, metadata and syntax.. The way forward adds to this list the following (some of which are refinements of existing work items):

- expansion of the DOI technical core to the full Handle capability;
- assignment of DOIs to Creations, not Resources, whilst still allowing resolution to Resources;
- development of a defined (extendable) list of services for Objects, beginning from the point of view of the required usage ("verbs");
- prototype implementations of mechanisms to accomplish those DOI-mediated services, including additional browser functionality (and efforts to ensure widespread support for this);
- development of a controlled (extendable) list of possible additional data types for use within the DOI Handle resolver, and experimentation with them to construct service mechanisms;
- definition of a well structured (controlled namespaces) minimal DOI metadata set for a look up service, and implementation of a corresponding mechanism;
- definition of a well structured (controlled namespaces) extendable DOI metadata set to support the construction of services;
- development of guidelines for response screens in cases where this would facilitate implementation of services;
- promotion of clearer use of terminology of the various components of the DOI initiative, and clear distinction between Creations, services, and service mechanisms..
- liaison with ISO TC46 re work identification (ISWC) in relation to PII and the fundamental position of work identification in the DOI.
- liaison with the ISSN Centre re the development and applicability of identifiers for dynamic entities.

Note that already on the list of actions, not discussed here, are issues such as the development and

deployment of a “ping” test for DOIs; a service to distinguish local cached / site licensed DOI holdings from external services; etc. (See May workshop report for details).

**References/links**

For explanation of undefined abbreviations, refer to DOI glossary http://www.doi.org/IDGLOSS.html


[Arms] Arms, W., et al 1997
An Architecture for Information in Digital Libraries
D-Lib Magazine, February 1997
http://www.dlib.org/dlib/february97/cnri/02arms1.html

[Caplan] You Call It Corn, We Call It Syntax-Independent Metadata for Document-Like Objects: Priscilla Caplan.
The Public-Access Computer Systems Review 6, no. 4 (1995): 19-23.
http://info.lib.uh.edu/pacsrev.html

[Caplan2] DOI or Don't We? : Priscilla Caplan
The Public-Access Computer Systems Review 9, no 1 (1998)
http://lib-04.lib.uh.edu/pacsrev/1998/capl9n1.htm

[CNRI] CNRI Handle System Resolver
http://www.handle.net/resolver/index.html

[DOI] Digital Object Identifier system home page
http://www.doi.org

[DOIarch] Archive of DOI discussion list 1998
http://www.doi.org/mail-archive/discuss-DOI/maillist.html

[DOIgallery] DOI Phase 1 prototype participants Gallery
http://www.doi.org/gallery/tour.html

[Ford/Harter] The Downside of Scholarly Publishing: Problems in Accessing Electronic Journals through Online Directories and Catalogs
Charlotte E. Ford & Stephen P. Harter
College and Research Libraries, 59 (1998), 335-346

[Handle] The Handle System overview
http://www.handle.net/overviews/hs-version4.html

[Handle doc] Handle Resolution Protocol Specification/ Handle Administration Protocol Specification

http://www.handle.net/documentation.html

[HandleID] Sun, S.X. 1997
Handle System: A Persistent Global Naming Service.
Overview and Syntax
http://www.handle.net/draft-ietf-handle-system-01.html

[Hill] The Common Information System: Keith Hill
http://www.doi.org/workshop/minutes/CISoverview/index.htm

[ISBN] International DOI Foundation and ISBN International Sign Letter of Intent to Pursue an Agreement: Statement of Intent
http://www.doi.org/news.html

[ISWC] ISO/TC 46 /SC 9 Working Group 2, International Standard Work Code (ISWC)
http://www.nlc-bnc.ca/iso/tc46sc9/iswc.htm

[Kahn/Wilensky] A Framework for Distributed Digital Object Services
Robert Kahn & Robert Wilensky
http://www.cnri.reston.va.us/home/cstr/arch/k-w.html

[Kasdorf] SGML and PDF- why we need both
http://www.press.umich.edu/jep/03-04/kasdorf.html

[Kelly] The Role of A&I Services in Facilitating Access to the E-Archive of Science: Maureen Kelly
http://www.icsti.nrc.ca/icsti/forum/fo9711.html#role

[Members] List of DOI Foundation members
http://www.doi.org/idf-member-list.html

[Miller] An Introduction to the Resource Description Framework: Eric Miller
D-Lib magazine, May 1998
http://www.dlib.org/dlib/may98/miller/05miller.html

[Paskin] Information Identifiers
Learned Publishing, 1997, Vol. 10 No. 2, pp 135-156;
http://www.elsevier.nl/homepage/about/infoident/

[Paskin2] Towards Unique Identifiers
Proceedings of the IEEE, Special Issue on "Identification and Protection of Multimedia Information" (In press)

[PII] Publisher Item Identifier

http://www.elsevier.nl/homepage/about/pii/

[PURL] Persistent Uniform Resource Locators home page
http://purl.org

[Reynolds] ISSN and Seriality: Regina Reynolds
ISSN discussion document, unpublished

[Rosenblatt] The Digital Object Identifier: Bill Rosenblatt
http://www.press.umich.edu/jep/03-02/DOI.html

[Rust] Metadata: The Right Approach - An Integrated Model for Descriptive and Rights Metadata in E-commerce. Godfrey Rust
D-Lib Magazine July/August 1998
http://www.dlib.org/dlib/july98/rust/07rust.html

[Terms] DOI terms and conditions document
http://www.doi.org/terms.html

[URLguide] http://www.ncsa.uiuc.edu/demoweb/url-primer.html

[URNres] IETF Draft March 1998: URI Resolution Services necessary for URN Resolution
http://www.ietf.org/internet-drafts/draft-ietf-urn-resolution-services-06.txt

[W3Caddress] Naming and Addressing: URIs
http://www.w3.org/Addressing

[Xerox] Digital Property Rights Language
Mark Stefik: Letting Loose the Light: Igniting Commerce in Electronic Publication; in: Stefik (ed) Internet Dreams, MIT Press, 1997
stefik@parc.xerox.com

[XIWT] Managing Access to Digital Information: An Approach Based on Digital Objects and Stated Operations. Cross-Industry Working Team
http://www.xiwt.org/documents/ManagAccess/ManagAccessTOC.html