DOI® System Information Kit

International DOI Foundation
July 2011
Overview of DOI® System

A DOI name is an identifier (not a location) for an entity on digital networks. It provides a system for persistent and actionable identification and interoperable exchange of managed information on digital networks.

Unique identifiers are essential for the management of information in any digital environment. Identifiers assigned in one context may be encountered, and may be re-used, in another place (or time) without consulting the assigner, who cannot guarantee that his assumptions will be known to someone else. To enable such interoperability requires the design of identifiers to enable their use in services outside the direct control of the issuing assigner. The necessity of allowing interoperability adds the requirement of persistence to an identifier: it implies interoperability with the future. Further, since the services outside the direct control of the issuing assigner are by definition arbitrary, interoperability implies the requirement of extensibility. Hence the DOI System is designed as a generic framework applicable to any digital object, providing a structured, extensible means of identification, description and resolution. The entity assigned a DOI® name can be a representation of any logical entity.

The DOI System is built using several existing standards-based components which have been brought together and further developed to provide a consistent system: the entire system has recently been accepted for standardisation within ISO (ISO TC46/SC9). The DOI System was developed as a cross-industry, cross-sector, not-for-profit effort managed by an open membership collaborative development body, the International DOI Foundation (IDF) founded in 1998. The DOI System is in widespread use, e.g. for scientific primary publishing, in government documents and in data. DOI names need not be explicitly declared (though this may be useful): e.g. in a web context a DOI name may be used in a http form as a URL (through a proxy server), whilst retaining the advantages of managed persistence. The DOI System may be used to offer an interoperable common system for identification of data.

DOI System components

The DOI System provides a ready-to-use packaged system of several components:

- a specified standard numbering syntax;
- a resolution service (based on the existing Handle System®);
- a data model incorporating a data dictionary (based on the indecs Data Dictionary); and
- an implementation mechanism through policies and procedures for the governance and application of DOI names.
DOI name syntax

The DOI name syntax is a standard for constructing an opaque string with naming authority and delegation (NISO Z39.84, DOI Syntax). It provides an identifier “container” which can accommodate any existing identifier: e.g. 10.1234/NP5678, 10.5678/ISBN-0-7645-4889-4 and 10.2224/2004-10-ISO-DOI are all valid DOI name syntax. The DOI name has two components, the prefix and the suffix, which together form the DOI name. The portion following the “/” character (the DOI name Suffix) may be an existing identifier. The portion preceding the “/” character (the DOI name Prefix) denotes a unique naming authority. There is no limitation on the length of a DOI name.

A DOI name may be assigned to any item of intellectual property, which must be precisely defined by means of structured metadata. The DOI name itself remains persistent through ownership changes, and unaltered once assigned.

A prefix is assigned to an organization that wishes to register DOI names; any organization may choose to have multiple prefixes. Following the prefix (separated by a forward slash) is a suffix (unique to a given prefix) to identify the entity. The combination of a prefix for the Registrant and unique suffix provided by the Registrant avoids any necessity for the centralized allocation of DOI names.

An existing standard identification system number such as ISBN may be incorporated into a DOI name, by using this as the suffix, if the registrant finds it convenient to do so (it is of course recommended that precisely the same entity be identified by the two systems). The DOI System is not alone in being one that can incorporate existing identifiers: for example, physical bar codes can be used to express ISBNs.

DOI name resolution

Resolution is the process in which an identifier is the input (a request) to a network service to receive in return a specific output of one or more pieces of current information (state data) related to the identified entity: e.g. a location (such as URL) where the object can be found. Resolution provides a level of managed indirection between an identifier and the output. The resolution component allows redirection on a TCP/IP network from a DOI name to associated data. Initial applications have been resolution to a single location (URL), providing a tool for persistence (since even if a URL is changed, the DOI name still functions and redirects to the new location). However more useful resolution may be to multiple associated data such as multiple locations, metadata, common services, or to extensible assigner-defined data. Applications of the DOI System using multiple resolution are now increasingly in use. The resolution tool used in the DOI System is the Handle system. This conforms to the functional requirements of the URI and URN concepts, and has many advantages over other mechanisms including global scalability, full Unicode character support, and security.

The Handle System implementation in the DOI System has been supplemented by expanded technical infrastructure and features specific to DOI System
applications. Handle multiple resolution allows one entity to be resolved to multiple other entities; it can therefore be used to embody e.g. a parent-children relationship, or any other relationship, and is therefore suitable for describing relationships of objects (data sets). Handle per se deliberately has no pre-existing constraints to define a framework to express relationships (analogy: spreadsheet software): The DOI System is an application of Handle which adds this constraint for a specific purpose of content management (analogy: a spreadsheet application). In the DOI System the constraints are defined through metadata grouping the entities, using a semantically interoperable data dictionary.

**DOI® Data Model**

The DOI Data Model consists of a data dictionary and a framework for applying it. Together these provide tools for defining what a DOI name specifies (through use of a data dictionary), and how DOI names relate to each other, (through a grouping mechanism, Application Profiles, which associate DOI names with defined common properties). This provides semantic interoperability, enabling information that originates in one context to be used in another in ways that are as highly automated as possible.

The DOI System uses an interoperable data dictionary built from an underlying ontology. The data dictionary component is designed to ensure maximum interoperability with existing metadata element sets; the framework allows the terms to be grouped in meaningful ways (DOI® Application Profiles) so that certain types of DOI names all behave predictably in an application through association with specified Services. This provides a means of integrating the features of Handle resolution with a structured data approach. DOI names need not make use of this data model, but it is envisaged that many will: any DOI name intended to allow interoperability (i.e. which has the possibility of use in services outside of the direct control of the issuing Registration Agency) is subject to DOI® Metadata policy, which is based on the registration of terms in the iDD.

A data dictionary is a set of terms, with their definitions, used in a computerized system. Some data dictionaries are structured, with terms related through hierarchies and other relationships: structured data dictionaries are derived from ontologies. An ontology combines a data dictionary with a logical data model, providing a consistent and logical world view. It differs from the traditional taxonomic approach to knowledge representation in that it does not follow a rigid/parent child hierarchical structure (terms may inherit meaning from more than one parent) and a more complex relationship is maintained.

An interoperable data dictionary contains terms from different computerized systems or metadata schemes, and shows the relationships they have with one another in a formal way. The purpose of an interoperable data dictionary is to support the use together of terms from different systems. The IDF is the Registration Authority for one such dictionary, the ISO/IEC MPEG-21 Rights Data Dictionary, and is the co-developer of a wider indecs Data Dictionary which includes this and is used by DOI names.

**DOI System implementation**

The DOI System is implemented through a federation of Registration Agencies which use policies and tools developed through a parent body, the International
Key facts on Digital Object Identifier System

Key concepts and abbreviations
- DOI = Digital Object Identifier
- IDF = International DOI Foundation (operating and governing organisation): www.doi.org
- RAs = DOI Registration Agencies (= members of IDF offering the system to customers who wish to assign DOI names)

Status: operational system
- Foundation launched to develop system in 1998. First applications launched 2000
- Currently used by c. 4,000 naming authorities (assigners) e.g., 3,300 publishers, EU documents, science data sets, etc.
- Over 50 Million DOI Names assigned to date
- Over 210,000 DOI name prefixes (naming authorities within the DOI system)
- Around 100 million DOI resolutions per month
- DOI names have been assigned by 12 RAs (past and current)
- Well established in professional information sector; best known applications are CrossRef (www.crossref.org) and DataCite (www.datacite.org)
- Initial applications are simple redirection
- More sophisticated functionality available e.g. multiple resolution, data typing, "Application Profiles"
- International Standard (ISO: in publication)

Scope
- Digital Identifier of an Object (not "Identifier of a Digital Object")
- Object = any entity (thing: physical, digital, or abstract)
  - Resources, parties, licences, etc.
- Digital Identifier = network actionable identifier ("click on it and do something")
- Initial focus on entities was documents/media e.g., articles, data sets
  - Now also moving into parties and licences
  - Extending to other sectors, e.g., Movie industry? Financial sector? Music? Newspapers?
- Extensible by design (as e.g., URI): not intended as a publishing-only solution (digital convergence)
- International – e.g., in 2007 appointed China RA

What it does
- provides an actionable, interoperable, persistent link:
  - actionable - through use of identifier syntax and network resolution mechanism (Handle System®)
  - persistent - through combination of supporting improved handle infrastructure (registry database, proxy support, etc) and social infrastructure (obligations by Registration Agencies)
  - interoperable - through use of a semantically interoperable data model and grouping mechanisms.

Governance
- IDF = operating and governing organisation
- Provides the social infrastructure
  - e.g., obligations for persistence, back-up, in event of failure, etc
- Proven model: successfully transitioned the management of persistent identifiers between different registrants and between different RAs
- US "Not for profit" open membership (with membership fee)
- Federation of Registration Agencies makes up significant part of the IDF (possibly 100% eventually)
- Elected Board, working groups (including RA Working Group)
- No full time staff (contracted managing agent, outsourced functions)

Business model
- IDF receives membership fees from RAs, contracts technical operator
- RAs are members of IDF and pay a fixed fee per year
- Costs of operating the system are divided across RAs so that IDF is cost-neutral
- Assigners are customers of RAs
- RAs might have their own existing numbering scheme, existing communities etc. – which can be integrated with a DOI Application, not replaced by it (e.g. ISBN)
- RAs are autonomous independent bodies. They offer services to assigners using DOI names
  - RAs’ business model with their customers is entirely autonomous
  - RAs only obligation to IDF is a licence/operating agreement
RAs may choose to put DOI names "under the hood"
- Inspired by bar code model, ISBN etc.: assigner pays
- Some RAs are commercial; others are themselves member communities (e.g., CrossRef)

**Technical infrastructure**
- **Handle System**: persistent identification in digital networks (devised by TCP/IP co-inventor)
- **Indecs**: principles of contextual ontology data model for associated metadata("interoperability of data in e-commerce systems"), implemented in Vocabulary Mapping Framework (http://www.doi.org/VMF/index.html)
- Both used elsewhere: aim was not to re-invent the wheel

**Standardisation**
- ISO 26324, Information and Documentation — Digital Object Identifier System: currently in publication.
- URI (within info-URI scheme)
- Mechanism for, and emphasis on, enabling re-use of other existing identifier schemes e.g. ISBN: see http://www.doi.org/factsheets/DOIIdentifiers.html

**Documentation**
- Website http://www.doi.org/
- Factsheets – recommended for coverage of topics in detail: http://www.doi.org/factsheets.html
- Summary articles: http://www.doi.org/about_the_doi.html

**Origin**
- 1996 proposal from the three major international publishing trade associations to develop infrastructure for digital publishing; they brought together expertise in numbering content (the ISO standard ISBN) and expertise in digital network technology (CNRI)
- Need in the digital supply chain for an equivalent of the analogue bar code: migration from analogue to digital networked content cannot rely on URLs as identifiers (e.g. due to "linkrot": "404 not found")

**Relation to other schemes**
- Strong focus on interoperability and on working with existing and new schemes
  - Technical, syntactic and semantic interoperability
  - Use of DOI names by multiple RAs (through APs and Services)
- Involvement with key activities in the content sectors: ACAP, ONIX, ARK, PURL, info URI, URN, Open URL, GSI, MPEG-21, IETF, RDA, DCMI, FRBR, ITU-T Idm, ICSTI, CENDI, HSAC, indecs, CONTECS-DD (etc!)
- Adopt existing proven components
- Note potential confusion: the term "identifier" can mean several different things – not always clear like-for-like comparison (e.g. URI v ISBN...)

**Intellectual property considerations**
- IDF owns DOI®, a registered trademark for the system
- IDF does not have any patents (or patent applications) on DOI System
- IDF collectively licences appropriate technology from suppliers on behalf of members (CNRI Handle System global license, data dictionary tools, etc.)
- All RAs must sign RA agreement re use of DOI System
- Optional legal "Community of Interest" agreement to enable sharing of information on relevant patent issues – findings confidential to signatories
- IDF is a participant in related semantic vocabulary work e.g. CONTECS-DD, Vocabulary Mapping Framework

June 2011
DOI Foundation (IDF). The IDF is the governance body of the DOI System, which safeguards (owns or licences on behalf of registrants) all intellectual property rights relating to the DOI System. It works with RAs and with the underlying technical standards of the DOI System components to ensure that any improvements made to the DOI System (including creation, maintenance, registration, resolution and policymaking of DOI names) are available to any DOI name registrant, and that no third party licenses might reasonably be required to practice the system standard. DOI name resolution is freely available to any user encountering a DOI name.

The DOI System has the flexibility to deliver identification and resolution services that fulfil the requirements of any application domain. However, these don't come "in a box" since someone needs to build the specific social and technical structures to support the particular requirements of a community (such as scientific data). The rules about what is identified, and whether two things being identified are (or are not) "the same thing", are made at a lower level: in a specific application of the DOI System. This is a role of Registration Agencies. This provides an identification system of enormous flexibility and power while hugely increasing the importance of an explicit structured metadata layer, since without this the identifier essentially can have no meaning at all outside a specific application.

The IDF provides implementation through agreed standards of governance and scope, policy, to define “rules of the road”. It also provides a technical infrastructure (resolution mechanism, proxy servers, mirrors, back-up, central dictionary) and a social infrastructure (persistence commitments, fall-back procedures, cost-recovery (on a self-sustaining model), and shared use of the system. The IDF is not a standards body, but a central authority and maintenance agency. The IDF is already the appointed registration authority for the ISO/IEC MPEG 21 Rights Data Dictionary, and is proposed as the registration authority for the DOI System within ISO TC46/SC9. IDF delegates and licenses authority to use the system through Registration Agencies, each of which can develop its own applications and use the DOI System in “own brand” ways appropriate for their community.
The International DOI Foundation (IDF) is a strong supporter of the Handle System® (http://www.handle.net/) and believes it to be the best infrastructure component available today for managing digital objects. That is why the DOI system uses the Handle System. Handles by themselves are necessary but not sufficient for the function of the DOI System, a complete framework for managing intellectual content and facilitating electronic commerce. It is sometimes asked: “If the DOI System is based on the Handle System, why not cut out the middleman and use handles alone?” The answer is simple: “DOI® names are more than handles”.

The Handle System is only one component of the DOI System
The Handle System provides a general-purpose global name service enabling secure name resolution over the Internet, designed to enable a broad set of communities to use the technology to identify digital content independent of location. The DOI System utilises the Handle System as one component in building an added value application, for the persistent, semantically interoperable, identification of intellectual property entities:
• The DOI System provides a ready-to-use system of several components: a specified numbering syntax, a resolution service (based on the Handle System), a data model system (including the indecs Data Dictionary), and policies and procedures for the implementation of DOI names through a federation of Registration Agencies.
• One component of the DOI System is the Handle system, and its implementation in the DOI System has been supplemented by expanded technical infrastructure and features specific to DOI System applications.

Persistence
Handle System software may be implemented by anyone who agrees to basic licensing terms; there is no requirement that a user’s implementation be persistent. The Handle System technology provides persistence if used with appropriate social infrastructure. The International DOI Foundation (IDF) builds on the technical infrastructure of the Handle System a social structure guaranteeing persistence. Persistence is a function of organizations, not of technology; a persistent identifier system requires a persistent organization and defined processes:
• The International DOI Foundation (IDF) provides a federation of Registration Agencies. Dependency on any one RA is removed since the IDF mandates that should any RA go out of business, its DOI names subsequently will be “homed” in another RA; and that should the IDF cease to exist, the RAs will be able to continue to operate the DOI System.
• IDF mandates that its RAs provide processes to ensure that DOI names guarantee persistence.
• The IDF is designed to be a persistent organization through self-funding by balancing operational income (from a small fee per DOI name assigned) against costs of ongoing infrastructure provision and development.
Several major public companies stand behind the IDF as initial funders and/or Board members; the composition of the Board and IDF membership may change but the organization continues.

**Consistency**

Handle System protocols ensure interoperability for resolution purposes among a diverse set of implementations. At the application level, there is no requirement that consistent rules must be in place for multiple applications. The DOI System adds such a requirement:

- The IDF determines rules to which all DOI names must adhere (what they may be applied to, restrictions on arbitrary and temporary assignment, restrictions on removal from the system, etc).
- DOI names are managed through a Directory Manager who ensures and implements quality assurance processes in assignment.
- A DOI® API (application programming interface) for the Handle System defines consistent ways of accessing and managing DOI names, Application Profiles, and Services.
- The consistent use of the DOI name prefix and numbering syntax provides numbering interoperability in the intellectual property sector, and brand recognition of what the number is for.
- The optional data model component provides semantic consistency for true interoperability.

**Ease of use**

The Handle System license does not include ongoing technical support; it typically is installed and managed by technical staff. The DOI System provides a turn-key application:

- The DOI System employs staff who manage the development and operation of the system, its relations with users, outreach to standards and other communities, and resolution of problems.
- The DOI System employs a Directory Manager who provides support to RAs in registering and managing DOI names, and who provides guidance, troubleshooting, statistics reporting, advice to and liaison with Registration Agencies and their customers.
- The DOI® Handbook embodies policies, procedures and guidelines for Registration Agencies and application developers developed by the DOI System federation of agencies, guaranteed by detailed legal agreements.

**Expressing Relationships**

Multiple resolution allows one entity to be resolved to multiple other entities; this can be used to embody relationships, e.g. a parent-children relationship, or any other relationship. The Handle System technology allows this; the DOI System provides a framework to achieve it:

- Multiple resolution is a feature of the Handle System technology, but the Handle System per se (deliberately) has no pre-existing constraints to make a useful framework (think of it as like spreadsheet software): the DOI System is an application of the Handle System which adds this constraint (think of it as like a spreadsheet template already prepared to receive data).
- In the DOI System the constraints come from the metadata which defines the entities, which is the data dictionary approach: hence IDF's role in MPEG-21 RDD and the indecs Data Dictionary. That enables one to express relationships. See the related factsheet "DOI® System and Data Dictionaries".

**Technical infrastructure**

The Handle System provides a resolution service shared by all Handle System implementations. The DOI System adds dedicated improved technical infrastructure:

- The Handle System consists of Global root servers, local handle servers, clients, and proxy servers. The scalable global root infrastructure enables users who install local handle services to interoperate with the root and each other, depending on permissions that have been established. The DOI System adds its own dedicated...
expanded infrastructure including replication servers for those IDF RAs that operate local handle services for their DOI names, secondary sites, mirrored servers and proxy servers housed at a secure commercial hosting service facility.

- The Handle System license provides the reference implementation of Handle System, the database component of which was not designed to scale above a few million handles. The DOI System employs a much more robust database implementation capable of scaling to any number of handles.
- The DOI System Directory Manager provides technical infrastructure configuration and performance checks to ensure evolutionary growth of the DOI System.

Semantic interoperability

Handles (including DOI names) will be resolved by the Handle System, but there is no requirement in the Handle System for declaring what is being identified, or for ensuring semantic interoperability across several identified resources. The DOI System adds this facility, specifically designed for its area of applications, which is now being implemented and will be a feature of advanced applications of the DOI System:

- The DOI System provides a kernel of structured data upon which extended metadata schemes can be built if required, and a means of precisely specifying the entity identified.
- The DOI System provides an optional tool to map an existing metadata scheme through a structured standard ontology, thereby ensuring semantic interoperability so that DOI names from different sources may be used as the key in building multi-component media objects or managing multiple assets.
- The DOI Data Model embraces both a data dictionary and a framework for applying it. The data dictionary component is designed to ensure maximum interoperability with existing metadata element sets; the framework allows the terms to be grouped in meaningful ways (DOI System Application Profiles) so that certain types of DOI names all behave predictably in an application through association with specified Services. This provides a means of integrating the features of handle resolution with a structured data approach.
- IDF maintains the indecs data dictionary, the underlying tool for semantic interoperability, which integrates with the standard ISO MPEG21 Rights Data Dictionary (RDD). The IDF is also the chosen candidate to become the MPEG Registration Authority to manage the RDD.

Development activities

The Handle System provides upgrades of the global general-purpose naming service. The DOI System adds to this resources for active development of the DOI applications and advanced features:

- Through its working groups and technical support staff, the International DOI Foundation provides ongoing development support and shared resources for the community of DOI name users.
- Through use of the DOI System in commercial systems, individual Registration Agencies have an incentive to allocate their own resources to the development of new features, or to collaborate with other RAs to develop features that may be shared with the wider DOI name user community.

Costs

Costs for developing Handle System implementations and DOI System applications are not directly comparable since they relate to different things:

- To provide a comparable service to the DOI System, a Handle System implementation would need to add the features listed above as these are not provided as part of the general-purpose software. The costs of implementing handles include appropriate Licence costs for the Handle System plus internal costs necessary for creating one or more of these features in a working implementation with specific rules and services. The costs of implementing DOI names include all these as turnkey features, not simply DOI name assignment as a handle, and vary
depending upon the application provided through Registration Agencies’ value-added services. The costs of assigning DOI names therefore vary from one Registration Agency to another depending on the business model chosen; DOI names may be free or charged for as part of a service offering.

- Use of the Handle System for any production operation requires a Licence from CNRI. The IDF already has a Licence for use of the Handle System, together with the ability to sub-license this for DOI name assignment to all Registration Agencies. Use of the DOI System therefore does not require a separate Handle System License.

**Governance**

*Governance of the Handle System and the DOI System are independent:*

- Policies and planning for the continued evolution of the Handle System are the responsibility of an evolving Handle System Advisory Committee representing major users and interested parties, established by CNRI, the organisation which developed the Handle System. The IDF is a Member of that Committee and so has a voice in the governance of the Handle System.

- The DOI System is managed by the International DOI Foundation, an independent, not-for-profit, open membership organisation. The IDF has an elected Board and nominated working groups which oversee all aspects of the development and implementation of the DOI System. CNRI is not a member of the IDF but provides services to it under commercial agreements. These agreements also ensure continuity of the resolution service provided to IDF.

This is one of a series of DOI® factsheets. To see the latest version of this factsheet online, and to see the other factsheets go to: [http://www.doi.org/factsheets.html](http://www.doi.org/factsheets.html)
DOI® System and Internet identifier specifications
Version 2.3

A standard represents an agreement by a community to do things in a specified way to address a common problem. Whilst the DOI community has developed the DOI System, it has also ensured conformance with relevant generic external formal standards. This factsheet discusses those relevant in the Internet communities (IETF and W3C). There has been considerable debate here on the issue of generic standards for naming objects.

Comparing generic identifier standards

A DOI® name differs from commonly used Internet pointers to material such as the URL, because it identifies an object as a first-class entity, not simply the place where the object is located. A DOI name also differs from identifiers such as the International Standard Book Number (ISBN), International Standard Recording Code (ISRC), etc., because it can be associated with defined services and is immediately actionable on a network.

The comparison of persistent identifier approaches is difficult because they are not all doing the same thing. Imprecisely referring to a set of schemes as 'identifiers' doesn't mean that they can be compared easily. Similarly, when any two technologies (e.g., two web browsers) are compared, the criteria used for comparison must be defined.

URI, URL, and URN

Historically there was ambiguity and confusion in the use of these terms. RFC 3986 (2005) aimed to end this by stating that a URI can be classified as a locator, a name, or both. In this view, the term URL refers to the subset of URIs that, in addition to identifying a resource, provide a means of locating the resource; the term URN has been used historically to refer to both URIs under the "urn" scheme (RFC 2141) which are required to remain globally unique and persistent even when the resource ceases to exist or becomes unavailable, and to any other URI with the properties of a name.

RFC 3986 requires that the terms URL and URN be deprecated. This brings a uniformity to the technical treatment of all URIs. However the risk of confusion remains, from:

- cited documents which rely on earlier, now superseded, statements of the position;
- the use of one simple top level term (URI) may hide useful distinctions which some users, e.g., librarians, may wish to make between a unique name and a location, for example when a named resource is available at multiple locations;
- considerations of how widely used non-web identifiers (such as ISBNs, RFIDs, social security numbers, etc) relate to URIs, which can lead to:
  - confusions re identifier, representation, and access mechanism;
  - lack of appreciation of identifier usage outside the WWW;
  - use for non-digital referents; and
  - the requirement to perceive the web as only part of the Internet and the Internet as only part of information.

In the view now considered by RFC 3986 to be obsolete, URIs have two subclasses: URN (identifying names) and URL (identifying single locations). In the RFC 3986 view, web-identifier schemes are all URI schemes, as a given URI scheme may define subspaces; some of these may be access mechanisms (e.g., "http:"), whilst others may be namespaces (e.g., "urn:").
There are strong arguments against all URIs being expressed forever as http protocol strings: see a good summary on the IETF URI Review mail list at http://www.ietf.org/mail-archive/web/uri-review/current/msg00978.html

**URI**

Uniform Resource Identifier (RFC 3986) provides an extensible means for identifying a resource within the World Wide Web. Each URI begins with a scheme name that refers to a specification for assigning identifiers within that scheme; each scheme's specification may further restrict the syntax and semantics of identifiers using that scheme.

URI specification defines (1) an implementation to access a location on a file server, commonly accessed using the http protocol though other protocols are allowed; (2) a syntax for referencing, through which e.g., ISBNs can be specified as URIs. The network path of the URI is implicitly DNS based; original URI specifications that assume the URI to be opaque have been overtaken by practical usage which assumes that the initial URI parser will look for meaningful characters (such as dot and slash).

The use of URIs as identifiers that don't actually identify network resources (for example, they identify an abstract object, or a physical object) was recognised as an unanswered problem in RFC 3305. This usage is important in any semantic application. To address this, the info URI scheme (RFC 4452: http://info-uri.info) was developed by library and publishing communities for "URIs of information assets that have identifiers in public namespaces but have no representation within the URI allocation". OpenURL adopts it and was a key the motivation for it. InfoURI registrations can be made by anyone, not necessarily the authority for a particular namespace. DOI is registered in the infoURI scheme.

**URN**

Uniform Resource Name (RFC 2141) is a specification for defining names (identifiers) of resources for use on the Internet. Locations are assumed to be independent of names. URN resolution is still an active topic of discussion, especially in the library community (e.g. for treatment of National Bibliography Numbers as URN in RFC 3188). RFC 2141 defines (1) a formal registration process as a urn namespace, and (2) accompanying specifications to implement a series of functional requirements for such namespaces. Existing identifiers may thereby be specified as a URN: e.g. an ISBN as urn:isbn:9789521061547; such identifiers may be implemented using a specially written URN plug-in and resolved to URLs: functionally this gives nothing beyond that achieved by coherent management of the corresponding URLs.

URN architecture assumes a DNS-based Resolution Discovery Service (RDS) to find the service appropriate to the given URN scheme. However no such widely deployed RDS schemes currently exist: browsers cannot action URN strings without some additional programming in the form of a “plug-in”. These carry no guarantee of ready interoperability with other deployments, which may require a different plug-in for each implementation and may use conflicting data approaches. Therefore most existing URN implementations embed the URN as a http URI which contains the URL of the relevant resolution service (e.g. for the URN form of the ISBN shown above, resolved via the Finnish national URN service http://urn.fi, the actionable form of the URN is http://urn.fi/URN:ISBN:978-952-10-6154-7). There is no global service aware of national and/or regional URN resolution services, but there are some proposals to provide one (e.g. http://www.persid.org).

The set of URNs, of the form "urn: nid: nnnnn", is a URN namespace. ("nid" is here a URN namespace identifier, neither a "URN scheme", nor a "URI scheme.") The official IANA list of registered NIDs at http://www.iana.org/assignments/urn-namespaces lists 40 registered NIDs; many of these are not widely used as URNs (e.g., ISSN, ISBN).
DOI is not registered as a URN namespace, despite fulfilling all the functional requirements, since URN registration appears to offer no advantage to the DOI System. It requires an additional layer of administration for defining DOI as a URN namespace (the string urn:doi:10.1000/1 rather than the simpler doi:10.1000/1) and an additional step of unnecessary redirection to access the resolution service, already achieved through either http proxy or native resolution. If RDS mechanisms supporting URN specifications become widely available, DOI will be registered as a URN.

**URL**

Uniform Resource Locator (RFC 1738) is a location on a file server in the WWW; redefined in RFC 3986 as "a type of URI that identifies a resource via a representation of its primary access mechanism (e.g., its network "location"), rather than by some other attributes it may have". In this view "URL is a useful but informal concept" (RFC 3305). In practice, it identifies a single location, and therefore is widely used incorrectly as a (mutable) identifier of the resource at that location (so the same resource at two URLs would have two URL "identifiers"). This bad practice arose from the failure to distinguish name and location in early WWW development. Adding to the problem, URLs carry semantics of the Domain Name they are based on and are therefore unsuitable as opaque identifiers; they may also be contextually qualified. URLs are pervasive as the foremost mechanism of location specification throughout the WWW, but less useful outside it.

Attempts to circumvent the problem of using URLs as citable identifiers by developing persistent identifier alternatives are well documented (PURL, DOI, ARK, etc.).

A DOI name may be represented as a URL (http string) by prefacing the string http://dx.doi.org/ to the DOI of the document (e.g., to resolve the DOI name 10.1000/182, enter into a browser the address: http://dx.doi.org/10.1000/182). Web pages or other hypertext documents can include hypertext links in this form.

**DOI functional requirements**

The DOI system is designed to fulfil several additional functional requirements which offer significant advantages in generic naming, notably:

- Neutral as to implementation. DOI allows but does not require http or other protocols. The design principle is that DOIs are not specific to the web or any other implementation (e.g., information may be delivered in non-web platforms such as PDAs). DOI is designed to be applicable in any environment on the Internet (the global information system linked by a globally unique address space based on the Internet Protocol (IP) using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite).
- Flexible as to implementation. The DOI system has been designed around a data and transaction model that can work in a wide variety of environments. The current implementation works well with, but does not require, http or other web protocols, and can be used in any environment on the Internet (the global information system linked by a globally unique address space based on the Internet Protocol (IP) using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite).
- Granularity of naming and administration at the object level. Allows but does not mandate coarser level granularity tools such as domain names. Specifically, DOI resolution in native resolver form does not require the use of the DNS (Domain Name System): the DNS administrative model argues against using it as a general-purpose name system and has well-recognised problems of security and updating.
- Neutral as to language/character set. Compatible with, but not restricted to, the ascii character set. DOI names can use the Unicode capability of the Handle System to develop DOI names in Japanese, Chinese, etc., characters. The current DOI syntax restricts initial implementations to ascii simply for ease of adoption, but is intended to be widened (backward compatibility) to Unicode in a future revision.
- Multiple resolution to typed data offers the possibility of expressing semantic relationships.
• Social infrastructure providing persistence through organizational backup, data integrity measures, etc.

Other internet persistent identifier schemes

The Handle System is a technology specification for assigning, managing, and resolving persistent identifiers for digital objects and other resources on the Internet. It is the underlying resolution component for the DOI System. The Handle System is the most appropriate persistent identifier management system for the DOI System: see the related DOI factsheet “DOI System and the Handle System” (http://www.doi.org/factsheets/DOIHandle.html). There are several other Internet persistent identifier mechanisms proposed by individuals or organisations, having various emphases on social infrastructure or technology. There are several studies of persistent identifier management sustainable infrastructure and services available, such as the PILIN project (http://www.pilin.net.au/Closure_Report.pdf).

A persistent uniform resource locator (PURL) is a Uniform Resource Locator (URL) that does not directly describe the location of the resource to be retrieved but instead describes an intermediate (more persistent) location which, when retrieved, results in redirection (e.g. via a 302 HTTP status code) to the current location of the final resource. PURLs are said to be “an interim measure, while Uniform Resource Names (URNs) are being mainstreamed, to solve the problem of transitory URIs in location-based URI schemes like HTTP”. (http://en.wikipedia.org/wiki/PURL).

Extensible Resource Identifier (XRI) is a scheme and resolution protocol for abstract identifiers. While the Handle System is focused on the secure administration and resolution of identifiers into handle records, XRI is more concerned with defining properties and semantics of identifiers to allow for extensible namespace resolution, segmenting of identifiers, identifier cross referencing, and semantics for accessing resources. The Handle System uses its own protocol over udp, tcp and http; XRI uses its own XRDS over http or https. Handles could be implemented in XRI as a internal resolution system within the XRI resolver, or as a registered XRI Service End Point (SEP). See also http://en.wikipedia.org/wiki/XRI.

Archival Resource Key (ARK) is a Uniform Resource Locator (URL) that provides a multi-purpose identifier given to information objects of any type. ARKs contain the label ark: in the URL, which sets the expectation that the URL terminated by '?' returns a brief metadata record, and the URL terminated by '??' returns metadata that includes a commitment statement from the current service provider. (http://en.wikipedia.org/wiki/Archival_Resource_Key).

Last revised May 2010

This is one of a series of DOI factsheets. To see the latest version of this factsheet online, and to see the other factsheets, go to: http://www.doi.org/factsheets.html.
DOI® System and standard identifier schemes
Version 2.1

This factsheet discusses the relationship between the DOI System and other standard identifier schemes (also known as registries of identifiers).

Identifier Registries

An identifier registry is a compilation of unique identifiers, with some information on each item so identified, registered through an organization which maintains it. The registry follows a syntax specification for the identifiers (typically a formal standard), and the agency provides a means of registering identifiers.

Usually the agency focuses on a particular field of interest and the items registered are of one content type, so the resulting registry achieves scale as the recommended, definitive, or most widely used list of identified entities of that content type. A good example is the ISBN for books (ISO 2108:2005; the ISBN registration authority is www.isbn-international.org).

Less commonly, sometimes a standard identifier syntax is defined without an implementation to build a registry (e.g. ANSI/NISO Z39.56 - Serial Item and Contribution Identifier, SICI). Without a single registry these schemes are less likely to be comprehensive and are not as useful for interoperable applications.

Differences between DOI System and most standard identifier registries

While the DOI System is a registry, it has significant differences compared to traditional single content-type registries:

- **Purpose**: The purpose of an identifier registry is to manage a given collection of identifiers; the primary purpose of the DOI System, on the other hand, is to make a collection of identifiers actionable and interoperable, where that collection can include identifiers from many other controlled collections.

- **Coverage**: A registry aims to be definitive and comprehensive across its content type; the DOI System is not intended to be a comprehensive identifier registry of all items falling potentially within its scope, but any such item may be registered as a DOI name.

- **Scope**: The scope of a standard registry is defined and fixed by content type (e.g. books and ISBN). The scope of the DOI System is potentially any resource involved in an intellectual property transaction; hence the coverage of DOI names is extensible (actual use expands continually as new areas of application are created): "The scope of the DOI system is not defined by reference to the type of content (format, etc) of the referent, but by reference to the functionalities it provides and the context of use". (ISO 26324 FDIS, Introduction)

- **Granularity**: The granularity of a content registry is typically defined and fixed by content type. The DOI System may be applied at any desired level of functional granularity, which may be modified by creating supersets or sub sets (including related types).

Registries may be of many different types, and are used in many applications, so it is difficult to generalise across all cases. This factsheet focuses principally on illustrating how the DOI System works with standard, accepted and widely implemented identifier registries in the information and documentation sectors, typified by those in ISO TC46/SC9.
The DOI System does not replace other schemes

“The DOI name does not replace, nor is an alternative for, an identifier used in another scheme” (ISO 26324 FDIS, Scope). Example:

- 10.97812345/99990 is a DOI name (an identifier in the DOI scheme); it cannot be validly submitted to an ISBN point-of-sale ordering system, or converted to a GS1 bar code for use as an ISBN bar code; it does not conform to the ISBN syntax.
- 978-12345-99990 is an ISBN (an identifier in the ISBN scheme). It cannot be validly submitted to a DOI name resolution service; it does not conform to the DOI syntax.

However both identifier strings have the same referent.

Recognition of standard identifier schemes in the DOI System

The DOI System explicitly recognises other schemes. The ISO DOI specification (ISO 26324) sets out the specifications for recognising existing schemes. At minimum, the DOI Kernel metadata must record the fact that another registry identifier exists. Additional optional steps are possible, including:

- a consistent way of including the other scheme in the DOI syntax;
- a business relationship to facilitate this, by collaboration between the IDF and the relevant registry. Where such collaboration is agreed, new potential may be unlocked: the ISBN-A application is an example of the linkage of DOI names to an existing registry.

The DOI System is designed to assist identifier interoperability (see the other factsheets on this issue: "Identifier Interoperability"; "DOI System and Internet Identifier Specifications"; "The ISBN System in Relation to the DOI System").

IDF has supported the development of new single content type registries where none exist (e.g. ISTC, ISNI); these can often provide a useful means of constructing a suitable DOI name syntax for such entities, and provide a specific community focus for applications. IDF does not advocate developing a new single content type registry where one already exists: multiple registries may confuse users, and a new registry is unlikely to be an economically successful entrant where an established registry already exists.

Other (non-DOI) schemes exist which can be used to build persistent identifiers (e.g. URN; URI; ARK; PURL; XRI, etc.). Only the DOI System mandates that other standard identifier schemes must be explicitly recognised and noted as part of metadata and or syntax. Hence only the DOI System explicitly promotes the use of identifiers from other accepted schemes.

An illustration

Consider as an example the most widely used DOI service, CrossRef: "CrossRef's specific mandate is to be the citation linking backbone for all scholarly information in electronic form. CrossRef is a collaborative reference linking service that functions as a sort of digital switchboard." (www.crossref.org)

CrossRef assigns DOI names to "scholarly information". “Scholarly information” is not one homogenous type of information entity. In the digital world, people can and will cite anything. The main items cited are articles, but scholarly information can include other things besides articles. Consider an example of a music journal citing a sound recording as a reference: CrossRef cannot use the existing ISO sound recording scheme (ISRC) alone, so CrossRef assigns a DOI name (an ISRC is not resolvable to the CrossRef database; a
DOI name does not resolve to the ISRC database. "The DOI name does not replace, nor is an alternative for, an identifier used in another scheme": ISO 26324 FDIS, Scope).

If an ISRC for the item does not exist, the new DOI name for that recording has no relation to the ISRC system. If an ISRC for the item exists, CrossRef records the ISRC as part of the DOI metadata. Optionally, if this is likely to occur often, IDF agrees a consistent way of including ISRC syntax in DOI names. Also optionally, CrossRef establishes a business relationship to facilitate this.

The same is true for other information objects, and in other DOI applications. It is also true between DOI applications – there may be “overlap” (and yet no interference with proprietary applications, without agreement), e.g. CrossRef assigns DOI names to cited items, while DataCite (www.datacite.org) assigns DOI names to datasets, which might be cited items. Interoperability between registration agencies on the basis of the DOI infrastructure assures that these applications are compatible, but cannot access full data from other applications unless so agreed.

De facto identifier registries created by the DOI System

There are some examples where DOI applications in a particular sector have become useful as a de facto registry of a particular content type where none previously existed. Examples are CrossRef (scholarly articles) and DataCite (scientific data sets). The development of a DOI application then provides a registry as a spin-off from the application.

When a single accepted registry or standard identifier does not exist, many different schemes may be in use. The creation of a DOI name then provides a means to quickly make these interoperable in the DOI System: e.g. CrossRef publishers might use PII, SICI, or internal schemes, but the resulting DOI names using these syntaxes in the DOI suffix become interoperable in CrossRef:

- Registrant using PII: doi:10.2345/S1384107697000225

All these DOI names have similar behaviour in the DOI System.

Multimedia objects and identifier registries

Digital multimedia objects will contain heterogeneous material, and have significant relationships across multiple content types. To deal with this, identifiers need to be interoperable at various levels: syntax, semantics, and community. Registries need to interact with Internet identifier specifications¹ and with other registries in a structured manner: this was recognised in SC9's 2004 Technical Report on functional requirements for identifiers and descriptors across content industries², and the 2005 SC9 Registration Authorities work on Identifier Interoperability³. Data relationships between objects in different registries can be managed through DOI applications: see the factsheet Managing Data Relationships.

Last revised: April 2010
Notes

1 Some ISO/TC46/SC9 registries have developed specifications for use with URN: ISSN (RFC 3044) ISBN (RFC 3187) and ISAN (RFC 4246). These do not preclude other uses in URIs.

2 ISO/TR 21449:2004, "Content Delivery and Rights Management: Functional requirements for identifiers and descriptors for use in the music, film, video, sound recording and publishing industries" discusses identification and description schemas for intellectual content and products to express the relationships involved in production, distribution, and rights management. The functional requirements for identifiers and descriptors set out in ISO/TR 21449:2004 are centred on intra- and inter-industry business transactions relating to production, distribution, and rights management in the content industries (i.e., the music, film, video, sound recording and publishing industries).

3 The SC9 identifier interoperability working group report is summarised as part 2 of http://www.dlib.org/dlib/april06/paskin/04paskin.html
Factsheet: the ISBN system in relation to the DOI system

ISBN and ISBN-A

The ISBN (International Standard Book Number) is a 13-digit identification number and system, widely used in the international book trade for over 35 years and assigned through a network of international ISBN Registration Agencies. ISBNs are used to identify each unique publication whether in the form of a physical book or related materials such as eBooks, software, mixed media etc.

The DOI® System offers a persistent actionable identifier for use on digital networks. A “DOI name” refers to the syntax string within the “DOI System”. The ISBN–A (“the actionable ISBN”) is a service powered by DOI®, in which an existing ISBN is expressed in the DOI System.

Why express an ISBN through the DOI System?

ISBN is adequate for trading physical products, but cannot easily take advantage of new possibilities digital networks offer, e.g.:

- Combining a publication’s identity with a preferred quality-controlled Internet destination where fuller descriptive information, additional or related content or e-commerce options can be provided;
- Variable licensing dependent on the user;
- Locating the optimal fastest URL for downloading large files;
- Dynamic tracking of digital-product sales.

As the Internet develops, new technologies are likely to offer further opportunities to publishers (e.g. peer to peer, resource sharing, additions to VoIP applications).

Publishers’ internal systems and bibliographic agency databases use ISBN as the key identifier. Introducing additional identifiers adds complexity, cost and potentially confusion; it therefore makes sense to continue using the ISBN as a principle identifier for digital publications, rather than add a new number. By including the ISBN in the DOI syntax in a standard way, a DOI is obtained which is derived from an ISBN. The ISBN International agency and the International DOI Foundation have agreed a way of doing this.

About ISBN-A

ISBN–A (“the actionable ISBN”) is a DOI name derived from an existing ISBN, by including the ISBN in the syntax string of the DOI.

- By definition, an ISBN–A identifies the same referent as that ISBN. It incorporates, but does not replace, the corresponding ISBN. The referent is determined by the ISBN agency.
- ISBN-As do not automatically exist for every ISBN; they exist only once the agency has registered them in the DOI System.
- The ISBN–A and the ISBN are used in different systems for different purposes:
  - An ISBN on its own cannot be resolved in the DOI System. It must be expressed and registered as an ISBN–A.
• ISBN-As are only assigned by DOI Registration Agencies which are also ISBN agencies (if they choose to offer this service).
  o Publishers may obtain DOI names from other DOI-RAs if they so wish.
• The purpose of creating a ISBN-A is to make an existing ISBN useful in a DOI application.
• Publishers, registration agencies, bibliographic databases and web-based systems only need one number-string (the ISBN) to manage and utilise a corresponding DOI name, so development of new applications for the book supply-chain can be made efficiently.
• The DOI metadata model enables provision of compatible and mutually supportive information to that provided from the ISBN system (such as bibliographic information in ONIX for Books standard format.)

Syntax of ISBN-A

The ISBN-A is constructed by incorporating an ISBN into the allowed DOI syntax:
Example: 10.97812345/99990

The syntax specification, reading from left to right, is:
• Handle System DOI name prefix = "10."
• ISBN (GS1) Bookland prefix = "978." or "979."
• ISBN registration group element and publisher prefix = variable length numeric string of 2 to 8 digits
• Prefix/suffix divider = "/
• ISBN Title enumerator and checkdigit = maximum 6 digit title enumerator and 1 digit check digit.

Note:
• the total length of the "Bookland prefix", the "ISBN registration group/publisher prefix" plus the "ISBN Title enumerator and checkdigit" will always equal 13 digits.
• the check digit from the ISBN remains unchanged; the DOI System prefix addition does not affect the check digit, nor is this used by the DOI System.

Examples of applications

• ISBN-A resolving to a managed web page service providing descriptive detail about the book e.g. publisher, title, author, subject and product description, cover image, cataloguing data. Publishers could further customize their pages with hyperlinks they control.
• ISBN-A associated by the publisher with an ad hoc set of relevant information and services (dynamically modifiable by the publisher at any time) to be displayed to the final user whenever a title assigned with an ISBN-A is cited and referenced over the Internet. Using DOI multiple resolution features, the ISBN may be turned into a one-click gateway to enhance the “book experience”, e.g.
  o to convey in a single access point reading samples, press releases, interviews, reviews;
  o to provide direct purchase from selected on line retailers;
  o to provide the reader with a choice of alternative formats of the same title;
  o to build a reader-community by linking to the ISBN-A widgets, blogs, social networks add on tools, podcasting;
  o to redirect to print on demand, rights clearance services and better exploit the long tail effect.
• ISBN-A services are directly controlled by the publisher even when titles are hosted by third party platforms, e.g. on Google Book Search once users discover a title within Book Search, they get access to ad hoc information and services related to that title, selected and updated by the publisher himself, simply by clicking on the ISBN-A.
• ISBN-A services may also collaborate with other, non-ISBN, DOI services. The International DOI Foundation will encourage this, to facilitate synergies between different services:
  - Generically: DOI names may be grouped into Application profiles; any single DOI name can be a member of one or Application Profiles; RAs might therefore agree to offer ISBN-A registration which would enable the DOI name to participate in several application profiles and services offered by different RAs.
  - By individual publishers: e.g. a publication using a DOI name in a citation linking service may also use an ISBN-A enabled service by resolving the citation to a ISBN-A

• Structured relationships may be expressed through appropriate use of the DOI System, e.g. a ISBN-A might resolve to related ISTC data about the work and provide some automated data expressing the link.

DOI collaborations with other identifier systems

There are other similar examples of identifier system collaboration: notably the ISBN may be expressed as a bar-code to be processed by store electronic systems. The value of each system is enhanced if they work together.

The guiding principles for referencing other identifier schemes within the DOI System are to maximise utility to potential users, and to maximise efficiency among registration agencies. The draft ISO Standard for the DOI System states that "if an object or class of objects identified by DOI name(s) is already within the scope of another ISO standard identifier, the DOI registration authority shall consult with the registration authority of the other ISO identifier to agree a mechanism for the inclusion of the character string of the identifier within the DOI syntax. The character string of the other ISO identifier shall be integrated into the DOI syntax, unless the relevant Registration Authority of that standard identifier indicates otherwise, or another integration mechanism (such as referencing as part of DOI metadata) cannot be agreed".

The ISBN-A is the first example of such an agreement in action.

Last updated: 13 July 2011
Fact sheet: identifier interoperability
Version 1.0

Resources of interest in digital networks originate from a wide variety of sources, and may carry identifiers from different established public schemes, official standards, de facto schemes, or private cataloguing numbering. A key step in facilitating preservation, re-use and exchange of information is to enable users to re-use these identifiers (and their associated data) across different applications. Such interoperability of identifiers encompasses not only technical aspects of interoperability but consideration of the purpose and community of use of the identifiers.

Interoperability
Interoperability is the ability of independent systems to exchange meaningful information and initiate actions from each other, in order to operate together to mutual benefit. In particular, it envisages the ability for loosely-coupled independent systems to be able to collaborate and communicate. Identifiers are lexical tokens that denote things participating in these systems; a referent is the thing that is identified by an identifier.

A resource can be part of more than one domain, and can be identified by different systems, so it is necessary to guarantee interoperability between different identification systems as well as implementations based on the same namespace. Identifier interoperability is necessary for purposes such as:

- Metadata interoperability (since metadata is a relationship which somebody claims to exist between two referents);
- The creation of standard mechanisms for the expression of relationships between the referent of different standard identifiers;
- The creation of services common to more than one system, e.g. discovery of “related content” items; compiling multimedia objects, etc.

Several such use cases for identifier interoperability have been explored in both the ISO TC46SC9 identifier interoperability work and the RIDIR project.

Identifiers assigned in one context may be encountered, and may be re-used, in another place or time without consulting the assigner. Even if a resource is assumed to be part of only one domain, once it is identifiable it may be adopted independently in another domain (and possibly with undisclosed modifications). Crucially for interoperability, the context and assumptions made on assignment of an identifier may not be known to someone else encountering and using an identifier. For example, one system may take for granted that its scope is abstract work entities; another may assume only concrete realisations of an abstract work. Where the independent systems are known to each other they may agree to provide supporting information on such assumptions; but where they are not known to each other interoperability must be ensured by other measures.

Three sorts of interoperability can be distinguished:

- Syntactic interoperability. The ability of systems to process a syntax string and recognise it (and initiate actions) as an identifier even if more than one such syntax occurs in the systems.
- Semantic interoperability. The ability of systems to determine if two identifiers denote precisely the same referent; and if not, how the two referents are related.
- Community interoperability. The ability of systems to collaborate and communicate using identifiers whilst respecting any rights and restrictions on usage of data associated with those identifiers in the systems.
These three form dependent layers: community interoperability is only possible if semantic interoperability is ensured; semantic interoperability is only possible if syntactic interoperability is ensured.

**Syntactic interoperability**
Syntactic interoperability may be ensured if two systems follow the same technical specifications for processing an identifier string, where the scope of the likely identifiers to be encountered is reasonably predictable. In certain cases, rules may exist for directly incorporating an identifier from one scheme in the syntax of another scheme.

However interoperability may be wide ranging, making it difficult to anticipate the likely scope: identifiers may be encountered beyond web identifiers, e.g. network telecommunication and broadcasting schemes, and other globally-unique identifiers such as International Standard Book Numbers (ISBN) not originally designed for digital use. Registry schemes and assumptions such as protocol dependence must be defined or discoverable if the identifier is to be used in more than a simple catalogue listing (e.g. as in the Dublin Core scheme field “Resource Identifier”, DC element syntax: DC.Identifier. There is no single registry of all identifiers. Many, but not all, identifiers of interest in a networked environment may be registered as URI schemes, but some schemes may be private or limited in their availability. Unique registry namespaces, akin to DNS domains, are part of the URN specifications (though not widely implemented). The info URI scheme was developed within the library and publishing communities (specifically, in conjunction with the development of the OpenURL standard) because of the need to specify common public namespaces as URIs (as pure identifiers: that is, to identify, not retrieve, de-reference, locate, etc.). The aim was to define URIs to reference information assets that have identifiers in public namespaces but no representation within the URI allocation – for example, LCCNs.

Some identifiers are purely abstract “denoting” tokens (names); others embody assumptions about the use to which the identifier will be put, such as resolution to retrieve, de-reference, locate, etc. Such assumptions may be deep; for example in the URI specification, it is assumed that URIs will be resolved, and the network path of the URI for resolution is implicitly DNS based: there are no real provisions to include systems that are not DNS based. Where identifiers explicitly include or implicitly assume specific protocols, proxy mechanisms (which translate one protocol to another) may need to be provided to ensure syntactic interoperability.

**Semantic interoperability**
Semantic interoperability deals with an obvious but difficult problem: even if two identifier strings can be syntactically processed alongside each other, how does a system know what the terms from another system mean? If A says “owner” and B says “owner”, are they referring to the same thing? If A says “released” and B says “disseminated”, do they mean different things? For effective interoperable management of entities:

- a unique identifier must be associated with a description of the referent entity, using a structured set of elements that provide information about that entity (that is, an identifier must be associated with some structured metadata to be interoperable);
- and
- the only way of unambiguously deciding if one term means the same as another, irrespective of what it is called, is by sharing a single frame of reference. A structured ontology (an explicit formal specification of how to represent the entities that are assumed to exist in some area of interest and the relationships that hold among them) with an underlying model that allows the generation of consistent new relationships, and a method of recording the agreement between the parties whose terms are included in it.

Two leading ontology initiatives that allow such comparisons in a shared frame of reference are the CIDOC conceptual reference model and the family of applications derived from the <indecs>-based semantic interoperability project (such as ONIX). These two have much in common, and some attempts are being made to investigate areas of commonality with library activities such as RDA. A joint initiative to develop a common framework for resource categorization has been launched. Ontologies are not yet in widespread use for fully automated transactions (as foreseen in the semantic web), but are in use in most serious
multimedia metadata and messaging schemes to provide a basis for unambiguous, extensible, and precise definition of terms.

**Community interoperability**
Identifier schemes may carry rights and restrictions on usage of data associated with those identifiers. An identifier registry authority will need to consider on what basis it is able to collaborate with other schemes, or make its data public; even if this is syntactically and semantically possible there may be barriers to open interoperability. The assignment and use of a particular identifier may have obligations regarding data ownership, data quality, data maintenance, governance, and participation requirements; these restrictions may apply in both commercial and non-commercial settings.

Semantic interoperability using mapping to a common ontology framework will necessitate a bilateral agreement between two schemes to confirm the precise intent of each others identification (or if unilateral, a note that the mapping is therefore unconfirmed by one of the parties); this provides an opportunity also to consider the community obligations of such mappings.

Each identifier registry has an obligation to its community of users a to ensure that its data is accurate; it cannot therefore rely on someone else’s metadata over which it has no quality control. Each identifier registry also needs to ensure that its standard is implemented through a business model: metadata has business value which provides support for registries to implement their standard; a registry cannot therefore be expected to hand over metadata to someone that it has no business relationship with. If both agree, a bilateral agreement can be drawn up which specifies the nature of the collaboration (for example, the appropriate registration authorities may agree to share or compare the values and updating processes for accompanying metadata). If the two do not agree, there cannot be an obligation of interoperability. Identifier schemes should encourage such collaboration by providing clear guidance on rights and obligations; these are often requirements of formal standardisation processes.

**Persistence and interoperability**
Persistence is closely related to interoperability: persistence is “interoperability with the future”, i.e. the independent systems able to exchange meaningful information and initiate actions from each other are separated by time.

The DPE Briefing Paper referenced below on "Persistent Identifiers for Cultural Heritage" explores the requirement and implications for persistence in identifiers in more depth. Some identifier schemes may be established for particular valid but relatively short-term needs (e.g. streaming subscription video on a social network); others focus on persistence and preservation, with a concomitant commitment to maintaining registry schemes, and metadata. An application designer will need to consider the benefits of basing an application on particular schemes, and avoid where possible schemes where the design is not in accordance with his own fundamental aims. URLs are often cited as the most common “identifiers” (e.g. DC:Identifier definition: “String or number used to uniquely identify the resource. The default is the URL to the resource”), although in fact a URL is an identifier of a location: as a consequence of the most common model of single redirection to one URL, the two are easily confused and the link between identifier and referent is not direct, and so easily broken. URLs have low barrier to entry and use, but low expectation of persistence. Identifier schemes using mechanisms to supplement the URL process (PURL, ARK, N2T) or avoid it (Handle, DOI) are preferable.

A mechanism which is specifically designed to support interoperability has been developed: in the Kahn/Wilensky Digital Object Architecture, referents (as “digital objects”) carry with them metadata and links to repositories. These digital objects do not replace existing formats and data structures, but instead provide a common framework for encapsulating those formats and structures, allowing them to be uniformly interpreted and thus moveable in and out of various heterogeneous information systems and across changes in systems over time. There are a few implementations of this but it has not yet been widely adopted (though the Handle identifier, part of this architecture, is itself widely used).
Lessons for implementers

- Avoid re-inventing the wheel: if it appears that you need to devise a new identifier scheme, examine whether the problem can be avoided by re-using existing identifiers.
- If a new scheme is needed, consider if an existing protocol or identifier registry can be harnessed to implement your scheme.
- Register your scheme with an appropriate public namespace declaration.
- Provide easy links for semantic mapping by specifying a well-formed metadata scheme and publishing it.
- Consider the community and business implications for others who may need to use your scheme.
- Provide clear guidance on rights and obligations of use of your scheme.
- Adopt identifiers with a mechanism for ensuring persistence.

Resources

DPE Briefing paper "Persistent Identifiers for Cultural Heritage"

RIDIR project (Resourcing IDentifier Interoperability for Repositories)
http://www.hull.ac.uk/ridir/


Identifier Interoperability: A Report on Two Recent ISO Activities.
D-Lib magazine, April 2006
http://www.dlib.org/dlib/april06/paskin/04paskin.html

The RDA/ONIX Framework for Resource Categorization.
D-Lib magazine, Jan/Feb 2007
http://www.dlib.org/dlib/january07/dunsire/01dunsire.html

CIDOC Conceptual Reference Model
http://cidoc.ics.forth.gr/index.html

"Info URI" registration scheme
http://info-uri.info/registry/docs/misc/faq.html

Digital Object Architecture
http://www.cnri.reston.va.us/cstr/arch/k-w.html
The indecs framework

Version 1

The DOI System uses work based on the indecs Framework as the basis for its semantic interoperability model. This factsheet is provided as background information on indecs.

indecs project

indecs (an acronym of "interoperability of data in e-commerce systems"; written in lower case) was a project part funded by the European Community Info 2000 initiative and by several organisations representing the music, rights, text publishing, authors, library and other sectors in 1998-2000, which has since been used in a number of metadata activities. A final report and related documents were published; the document "Principles, model and data dictionary" is a concise summary.

indecs provided an analysis of the requirements for metadata for ecommerce of “content” (intellectual property) in the network environment, focussing on semantic interoperability. Semantic interoperability deals with the question of how one computer system knows what the terms from another computer system mean (e.g. if A says “owner” and B says “owner”, are they referring to the same thing? If A says “released” and B says “disseminated”, do they mean different things?).

indecs built from a simple generic model of commerce (the "model of making"): a model of the life cycle of any kind of content or intellectual property from conception to the final physical or digital copies. The top-level model is summarised as “people make stuff; people use stuff; and (for commerce to take place) people make deals about the stuff”. If secure machine-to-machine management of commerce is to be possible, the stuff, the people and the deals must all be securely identified and described in standardised ways that machines can interpret and use. Central to the analysis is the assumption that it is possible to produce a generic mechanism to handle complex metadata for all different types of content. So, for example, instead of treating sound carriers, books, videos and photographs as fundamentally different things with different (if similar) characteristics, they are all recognised as creations with different values of the same high-level attributes, whose metadata can be supported in a common environment.

The indecs Framework

The indecs analysis supports interoperability of at least five different types:

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

The indecs project developed a framework, described in detail in the final project documents, within which such interoperability could be achieved. Indecs proposed four principles as key to the management of identification:

- The principle of Unique Identification: every entity should be uniquely identified within an identified namespace.
- The principle of Functional Granularity: it should be possible to identify an entity whenever it needs to be distinguished
- The principle of Designated Authority: the author of an item of metadata should be securely identified.
• The principle of Appropriate Access: everyone requires access to the metadata on which they depend, and privacy and confidentiality for their own metadata from those who are not dependent on it.

Indecs also produced a useful definition of metadata:
• An item of metadata is a relationship that someone claims to exist between two entities.

The indecs framework stresses the significance of relationships, which lie at the heart of the indecs analysis. It underlines the importance of unique identification of all entities (since otherwise expressing relationships between them is of little practical utility). Finally, it raises the question of authority: the identification of the person making the claim is as significant as the identification of any other entity.

Use of indecs

The indecs framework does not presuppose any specific business model or legal framework; it can accommodate any (e.g. it can be used to describe transactions of copyrighted, open source, or freely available material).

The framework has been developed further as a generic ontology-based approach dealing with defined types of entity and attribute, and the relators that link them within a contextual model structure (where context is defined as an intersection of time and place, in which entities may play roles). Its main use to date has been in applications of commercial transactions of content and in some library-related applications. Examples of applications using this approach include:
• RDA/ONIX Framework for Resource Categorization
• ISO/IEC 21000-6 (MPEG) Rights Data Dictionary (RDD)
• DDEX (Digital Data Exchange) Music industry messaging and data dictionary applications
• ONIX (Online Information Exchange) standards for the use of publishers in distributing digital metadata about their products
• Digital Object Identifier System metadata schemes.

Other developments are continuing, notably through the semantic engineering tools and services marketed by Rightscom under the Ontologyx brand. The approach also has much in common with the CIDOC Conceptual Reference Model (CRM), an ontology for cultural heritage information, and the Functional Requirements for Bibliographic Records report (FRBR) in the library world.

Mapping of terms

Different models of the life cycle of content may have important differences, not least in the specific meaning attached to the names of terms they employ. FRBR, indecs and CRM were each informed by different functional requirements, and so evolved different mechanisms for dealing with the issues that seemed most important to them. Broadly, they are compatible, and effective integration of metadata from schemes based on them should be achievable, but they must be handled with care. As an example: the terms abstraction, manifestation, item and expression are often used in considering content life cycles (e.g. a sound recording is the expression of a musical work during a recording session at a particular place and time, and is distinct from, say, the master tape made, which is a manifestation). These were dealt with in indecs, but may have slightly different meanings in other schemes. Such an analysis of meaning of a term from a scheme is possible in indecs by mapping the precise definitions into further terms with precise definitions within the indecs Framework. Indecs and other frameworks continue to be developed and refined through the process of implementation.
The indecs project [http://cordis.europa.eu/econtent/mmrcs/indecs.htm]


4 MPEG-21 Rights Data Dictionary (ISO/IEC 21000-6) [http://iso21000-6.net/]

5 Digital Data Exchange (DDEX) [www.ddex.net]

6 ONIX (Online Information Exchange) [http://www.bisg.org/onix/index.html] [http://www.editeur.org/onix.html]

7 Digital Object Identifier system [www.doi.org]

8 [www.rightscom.com]


Managing data relationships using DOI® resolution
Version 1.0

Data as relationships

Managing data implies managing relationships between entities: "A has the relationship B to C" (e.g.: "Daniel Defoe is the author of Robinson Crusoe"; "Article Y is chapter of Book X"; "this book has dollar price 25..."): more concisely, "A isBofC"1. The DOI® System can ensure such data are persistent and interoperable.

A, B and C all need to be precisely identified for automation2, though relationships can be expressed loosely ("my car is blue") or precisely ("the car registration number ABC1234 has colour paint code BG45678"), depending on the need: any automated or interoperable application needs more precision.

- A and C in A isBofC may or may not already have identifiers in one or more registries: existing identifiers can be used in a DOI name3, or a DOI name can be minted as a de novo identifier string.
- Standard relationship types (values of B in A isBofC) from all the main content-related identifier and metadata schemes are part of the Vocabulary Mapping Framework4, so a relationship may be specified precisely and interoperably.

A, B and C may all be assigned DOI names if required. In a simpler case, A has a DOI name, and B and C (not necessarily uniquely identified) are part of the resolved record of that DOI name.

Relationship between entities is often called metadata5. In managing content, as an example, metadata includes any of authorship, provenance, rights positions, pricing, ownership, distributor, aggregator and licensee data, production information and identification of how, when and the context of content use. Some metadata may be confidential and proprietary (companies may make a business from providing it); other metadata may be usefully made public to provide "hooks" for access.

Relationships can be static or dynamic. Static relationships can be published without further concern. Dynamic relationships (where the current value of C may vary, or the number or type of relationships B may vary) need support to be persistent: this can be provided by value-added services. A classic example is "item A has URL C" (which if not managed leads to "404 linkrot", lack of persistence). Once the relationship is made public, the assigner cannot always control its use: one of the current main uses of the DOI System is to provide persistence in this URL relationship through managed redirection (the assigner does not need to patrol every mention of the URL).

Resolution can be used to express relationships.

- Resolution is the process in which an identifier is the input — a request — to a network service to receive in return a specific output of one or more pieces of current information (state data) related to the identified entity: e.g., a location (URL). In the DOI System, the data is structured in type-value pairs.
- Multiple resolution is the return as output of several pieces of current information related to a DOI referent: at least one URL (though possibly several), and defined data structures assisting in management; the ability to "get metadata about" the DOI referent in structured interoperable form provides considerable added value.

URL resolution locates an item or arrives at a managed destination page with further links to be selected. DOI name resolution can provide more information, which clients can process, thereby managing the relationship links in the resolution rather than at the
destination. URL resolution is a one to one relationship; a DOI name resolution offers the option of one to many (multiple resolution). This may be:

- One DOI name to many URLs. When the entity A is available at several URLs, a DOI name can record all, and provide all or the most appropriate of these. This is currently in use with some DOI applications.

- One DOI name to many other data types. One or more of the entries in a DOI handle record could be used to express relationships (e.g. data type = URL, Value = http; or e.g. data type = "relationship", value = "Chapter of" (simplified examples). Since a value might also be a DOI name, these can be nested to express any level of complex relationships AisBofC (e.g. this DOI name is a chapter of that DOI name).

- A standard grouping mechanism for treating similar DOI names and similar DOI System services as classes is available through the DOI data model of Application Profiles and Services. A standard way of expressing relationships is available in the Vocabulary Mapping Framework, also denoted by DOI names.

Application considerations

Where relationships are managed in a silo application, controlled by one managing body, considerations of interoperability may be irrelevant. But when that application needs to link to others, or be exposed so as to enable requests from other applications to provide a service, interoperability becomes important.

Interoperability is the ability of independent systems to exchange meaningful information and initiate actions from each other, in order to operate together to mutual benefit. The context and assumptions made on assignment of an identifier may not be known to someone else encountering and using an identifier, so data about the referent needs to be easily available.

The majority of DOI applications do not currently use metadata relationships in this structured interoperable way. There will always be a need for simpler implementations that don't carry the full interoperability load (silo applications); they may be in the majority for a long time yet. But the increasing use of linked applications, and the value of a comprehensive rights and permissions management infrastructure, implies it would be prudent to manage data so as to allow a ready transition to such an interoperable common framework.

Separation of internal data and systems and the exposure of that data to the outside world is standard information management practice. DOI names offer a way of exposing data and associated relationships to others in a standard form, based on granularity analysis (what assets get separately identified) and interoperability in a standard fashion. Just as the use of DOI names as persistent pointers provides a value-added layer on top of changeable URLs, the use of DOI names to connect assets managed by multiple organizations can provide a value-added layer on top of information management silos built by individual organizations, to reduce transaction costs of mobilizing and using assets.

Application design, therefore, requires decisions on identifier granularity and commercial value of expressing relationships; these may be business decisions (e.g., with e-books, it might be useful to express the relationships between an original work and all its published versions: the problem isn't whether to use an ISBN or DOI name, but to agree the level of granularity at which e-books need to be identified and who should provide the identifiers and manage the data given that many of the large publishers are only prepared to assign ISBNs to the generic .epub file.)

Last revised April 2010
Each relationship of the form "A is B of C" can also be expressed as "C is B of A" (e.g. Robinson Crusoe has author Daniel Defoe): i.e., any piece of data may be "metadata" for another piece of data. There will be multiple relationships about any entity (AisBofC, AisDofE, etc.)

See the indecs project: http://en.wikipedia.org/wiki/Indecs_Content_Model.


Vocabulary Mapping Framework: http://cdlr.strath.ac.uk/VMF/.

For simplification, we have omitted from this discussion the provenance of the relationship statement ("who says that AisBofC?"). For most purposes it is sufficient to allow this to be implicit; in the DOI System, it is implied by the right to manage the DOI record, but for an application where this was a direct concern it could be made explicit as another relationship.


This builds on the extensible data typing mechanism of the Handle System®.


As with physical resources: "all standard formal property documents are crafted in such a way as to facilitate the easy measurement of an asset’s attributes. If standard descriptions of assets were not readily available, anyone who wanted to buy, rent, or give credit against an asset would have to expend enormous resources comparing and evaluating it against other assets — which also would lack standard descriptions. By providing standards, Western formal property systems have significantly reduced the transaction costs of mobilizing and using assets.” (H de Soto, The Mystery of Capital, 2000)
DOI® System and data dictionaries
Version 3.1

A data dictionary is a set of terms, with their definitions, used in a computerized system. Some data dictionaries are structured, with terms related through hierarchies and other relationships: structured data dictionaries are derived from ontologies. An ontology combines a data dictionary with a logical data model, providing a consistent and logical world view. It differs from the traditional taxonomic approach to knowledge representation in that it does not follow a rigid/parent child hierarchical structure (terms may inherit meaning from more than one parent) and a more complex relationship is maintained.

An interoperable data dictionary contains terms from different computerized systems or metadata schemes, and shows the relationships they have with one another in a formal way. The purpose of an interoperable data dictionary is to support the use together of terms from different systems.

The DOI system uses an interoperable structured data dictionary. The IDF’s Data Dictionary (IDD) is a structured ontology, in which meaning, once it has been defined, can be passed on from one term to another by logical rules of association such as inheritance and opposition. In recognition of the great diversity and complexity associated with multimedia content, it is designed to represent as many different specialisations of meaning as its users require, and to show their relationships in a structured way in order to support the mapping and transformation of terms between different schemas and systems.

The IDD is currently in use as an internal tool. However it is built on exactly the same principles as the public ISO/IEC 21000-6 Dictionary, the MPEG-21 RDD (www.iso21000-6.net), for which the IDF is the ISO Registration Authority (IDD and RDD are instantiations of the same methodology). The ISO/IEC 21000-6 dictionary is designed as a component of MPEG-21 to support the MPEG REL (Rights Expression Language), but is based on a generic methodology. IDF also intends to promote the adoption of ISO/IEC 21000-6 as the registry for metadata semantics for standard content identifiers, and to this end IDF is investigating how the IDD and RDD can converge.

The DOI® Data Model embraces both this data dictionary and a framework for applying it. The data dictionary component is designed to ensure maximum interoperability with existing metadata element sets; the framework allows the terms to be grouped in meaningful ways (DOI® Application Profiles) so that certain types of DOI® names all behave predictably in an application through association with specified Services. This provides a means of integrating the features of Handle resolution with a structured data approach. DOI names need not make use of this data model, but it is envisaged that many will: any DOI name intended to allow interoperability (i.e. which has the possibility of use in services outside of the direct control of the issuing Registration Agency) is subject to DOI® Metadata policy, which is based on the registration of terms in the IDD.

A Data Dictionary creates semantic compatibility

IDD exists to solve an obvious but difficult problem: how does one computer system know what the terms from another computer system mean? (If A says “owner” and B says “owner”, are they referring to the same thing? If A says “released” and B says “disseminated”, do they mean different things?) The data dictionary provides a way of
describing relationships between terms, and confirming agreement about this, so that A or B (or anyone else) can make use of one another’s metadata with confidence and in a highly automated way.

- It may be assumed that A knows what he means, and B knows what she means; but they may be assuming totally different concepts from each other. This is true of any term: concepts (e.g. “depression” as understood by the mental health, economics, and meteorological communities), roles (e.g. “publisher” as understood by music, newspaper and book industries), and physical formats (e.g. “folio” as understood by the bookkeeping, legal, and printing communities).
- The only way of unambiguously deciding if one term means the same as another, irrespective of what it is called, is by sharing a single frame of reference: a structured ontology (an explicit formal specification of how to represent the entities that are assumed to exist in some area of interest and the relationships that hold among them) with an underlying model which allows the generation of consistent new relationships, and a method of recording the agreement between the parties whose terms are included in it.
- Mapping terms from one scheme to another is not always straightforward. Terms may be expressed in different parts of speech and tenses, and meanings are often “contextual” (e.g. the same term “Identifier” in one place may mean “Product Identifier”, and in another place mean “Party Identifier”, within the same scheme). The IDD is designed to support these levels of complexity and contextuality.
- Whilst there are many ontology approaches, few address the semantic interoperability requirement; IDD uses the most well-developed method for this.

Data dictionaries are necessary for efficient interoperability

Metadata interoperability means enabling information that originates in one context to be used in another in as automated a way as possible. Information in one context will typically use a metadata scheme appropriate to that industry, sector or company. Using this in another context, which may use a different metadata scheme, requires semantic mapping and transformation of terms across the two metadata schemes.

- Whilst “crosswalks” can be constructed to compare terms in any two metadata schemes, the total number of such crosswalks grows much faster as the number of schemes grows linearly (N schemes require (N/2)(N-1) mappings). The existence of one dictionary reduces this to N mappings, one for each scheme.
- Bilateral agreement between dictionary and scheme ensure that the existence of agreed mapped terms enables extensibility – mapping to another scheme - without reference to the originators of each scheme. Such mappings will increasingly be computable and thus automated.

IDD is one component of the DOI System

The DOI System provides a ready-to-use system of several components: a specified numbering syntax, a resolution service (based on the Handle System), a data model system (including the indecs Data Dictionary), and policies and procedures for the implementation of DOI names through a federation of Registration Agencies.

- One component of the DOI System is the IDD. Its implementation in the DOI System has been supplemented by expanded technical infrastructure and features specific to DOI® applications.
- The IDD is the repository for all data elements and allowed values used in DOI® metadata declarations; it is the heart of the DOI® Data Model process. The IDD enables the definition and ontology of all metadata elements to be available to all RAs, and provides the necessary mappings to support metadata integration and transformations required for data interchange between RAs who require it.
- The functions of the IDD are to support:
  - interchange of metadata between Registration Agencies using standard messaging (RMDs);
  - automated use of Kernel metadata declarations;
  - interoperability between Application Profiles by common semantics for DOI® Services and ResourceTypes.
**IDD is not unique to the DOI System**

**DOI** development aimed to use existing or developing standards, rather than develop unique tools. The IDD has a long history which began around the same time as the DOI System, and is used in several major activities.

- The IDD is built using methodology from the indecs (interoperability of data in e-commerce) framework, an influential multimedia metadata project from 1998-2000 backed by groups from the content, author, creator, library, publisher and rights communities, which pioneered a model of event-based metadata as a solution for integrating rights. Indecs in turn drew on earlier work from the library community (FRBR) and music community (CIS).
- Subsequent versions of the methodology have been used as the basis for the DOI System, for the MPEG-21 Rights Data Dictionary (RDD), and heavily influence the current development of messaging systems for the publishing industry (ONIX) and music industry (MI3P).
- The methodology for the development of such data dictionaries was initially codified during the development of the MPEG 21 Rights Data Dictionary by the CONTECS:DD consortium (which included the IDF).
- The International DOI Foundation (IDF) and EDItEUR (the International Group for electronic commerce in the book and serial sectors) harmonise ONIX and DOI® metadata through the use of this common data dictionary (and welcome collaboration with others adopting a similar approach).
- The methodology has been validated against the W3C ontology language OWL-DL.
- The methodology for constructing interoperable Data Dictionaries which underlies IDD is in use commercially as Ontologyx. Ontologyx is a comprehensive data meta-model, an analytical tool and an ontology; a computing platform to articulate the tools is under development. In some ways Ontologyx has the same relationship to the DOI System as the Handle system: it is a component, but also used in other ways outside the DOI System.

**IDD is neutral as to business model**

The semantic analysis underlying the IDD is independent of any implementation model.

- It was fundamental to indecs (despite “e-commerce” in its name) that it had no inherent commercial model, and it remains so for all the work that has followed it. It is just as critical to be able to say "this is not subject to copyright" as to say the opposite; one of the problems any "non-commercial" person or organization has is to be able to state that something is freely available and under what circumstances. A broad ontology supporting rights expressions must be able to support any kind of expression of any kind of right, agreement or licence or any terms or none. Most organizations have the need for both freedom and protection of intellectual property in different contexts. The IDD is not solely a tool for intellectual property as "commercial property" but is neutral as to the intellectual property regime being used.

**IDD does not mandate one metadata scheme**

Since the aim of the IDD is to facilitate mapping between schemes, it does not mandate one scheme. There are minimum requirements of DOI® Registration Agencies that must be followed in the DOI application to ensure that the metadata can indeed be mapped into the IDD:

- An RA must be capable of producing a Kernel Metadata Declaration for each DOI name, using a small set of standardised terms from the IDD.
- Metadata exchanged between RAs supporting DOI® services should be exchanged using an agreed message format, a DOI® Resource Metadata Declaration (“RMD”)
- Proprietary terms (data elements and values) used by RAs in Kernel and Resource Metadata Declarations should be registered in the IDD.
- RAs are otherwise free to use any metadata schemes for gathering, storing or disseminating metadata.
IDD provides authority

Every term entered into the IDD carries information on its status as to origin and mapping agreement

- If a mapping from scheme A to IDD is reciprocally agreed with the governance authority of scheme A, then the dictionary can embody an assured mapping which will enable users of the dictionary to interpolate mappings from their own schemes, through IDD, to scheme A and know that this will be considered authoritative by scheme A. Such mappings will be dynamically updated as new versions of schemes are made available.
- Any RA contributing terms to the IDD can specify who is allowed to see or specify their own terms.
- Any public terms are accessible to all IDF agencies; e.g. ONIX, DOI® terms from the kernel and RMD, and the MPEG21 RDD.

IDD construction

Users need not understand the underlying concepts and construction of the IDD.

- It is no more a requirement to know the details than it is for the design of a web page to require one to read all the underlying internet protocol RFCs.
- A fundamental role of the IDF with the IDD is to provide assurance to users that the work has been peer-reviewed, tested in practical implementations, and is based on sound principles. The IDD structure has been tested in several implementations, public and private, including the MPEG 21 RDD evaluation, the DOI System, and mapping to the W3C ontology language OWL-DL.
- Some key features of the IDD methodology are:
  - Extensible and granular: the ontology extends its core “Context Model” to whatever level of detail and granularity is required.
  - Multiple, different, specialized views are available: these include the Resource Model, based on ten core data elements, and the Rights Model, based on a set of specialized Contexts.
  - Local terms: an RA can add all its own local data elements and names into the ontology, and use only those terms it needs. It can include different terms from different internal systems and map them together.
  - External terms: it incorporates external and standard schemes such ISO territory, currency and language codes, and sector specific external schemes, allowing them to be treated seamlessly alongside local terms.

IDD and the MPEG-21 Rights Data Dictionary (ISO/IEC 21000-6)

The IDD is currently in use as an internal tool. However it is built on exactly the same principles as the public ISO/IEC 21000-6 Dictionary, the MPEG-21 Rights Data Dictionary (www.iso21000-6.net), for which the IDF is the ISO Registration Authority

- IDD and RDD are instantiations of the same methodology. The ISO/IEC 21000-6 dictionary is designed as a component of MPEG-21 to support the MPEG REL (Rights Expression Language), but is based on a generic methodology. IDF also intends to promote the adoption of ISO/IEC 21000-6 as the registry for metadata semantics for standard content identifiers, and to this end IDF is investigating how the IDD and RDD can converge.
- As currently specified, all terms in the RDD are mapped into the IDD; that is, RDD is one of the authorities specifying terms within IDD. RDD is therefore a sub set of IDD. It is conceivable that some future RDD terms might be added to the RDD which are not within IDD; the two Data Dictionaries would then overlap and share some common terms.
- The RDD Registration Authority will establish an automated web based look-up system. It is anticipated that access will be granular, with levels of privilege being established to ensure that those with authority to access the Dictionary are able to view what is appropriate while private TermSets, if they exist, are kept confidential.
DOI® System Applications

Version 1.1

The DOI® System has the flexibility to deliver identification and resolution services that fulfil the requirements of any application domain. However, these don't come "in a box". Someone needs to build the specific social and technical structures to support the particular requirements of a community. The rules about what is identified, and whether two things being identified are (or are not) "the same thing", are made at a lower level: in a specific application of the DOI System. This is a role of Registration Agencies. This provides an identification system of enormous flexibility and power – while hugely increasing the importance of an explicit structured metadata layer, since without this the identifier essentially can have no meaning at all outside a specific application.

In designing a DOI System application several questions need to be addressed, including:

- What are we identifying with this DOI® name?
- What are we resolving to from this DOI name?
- What, if any, explicit metadata are we making available through this DOI name?
- How will the cost of providing the application using the DOI System be met?

What are we identifying with this identifier?

This deceptively easy question is one of the most difficult encountered in all discussions about identifiers (but the one most commonly overlooked) and an answer is often much more difficult than it might at first appear. The problem is compounded by the closely related question of when two things are "the same thing" (the answer to which is entirely contextual).

- The DOI System has a core strength: a DOI name can be used to identify anything within its very broad "creation" scope (which may also sometimes be perceived as a weakness, since it sometimes makes it very difficult to explain!). This means it can be used widely, and enable interoperability. The rules about what is identified, and whether two things being identified are (or are not) "the same thing", are made at a lower level: in a specific application of the DOI System. Attempting to extend the use of the identifier without explicit metadata risks a breakdown, since the implicit metadata needs to be made explicit to allow such interoperability away from the original context. If that breakdown happens, the "identifier system" will appear to have failed; when what has failed is the design of the appropriate implementation.

- For example: imagine an Application A which enables a user to purchase an e book in a specific format, pdf; that book, in that format, needs to be identified (separately from say the paperback physical format or the...
audio book format) for the purposes of a purchase using Application A. Further up the rights chain, the rights in that book will have probably been managed at the abstraction level (identifying the work, irrespective of all purchasable formats which are derived from that abstract work). An Application B elsewhere (which A may not be cognisant of, and vice versa) for the purposes of rights clearance could consider all formats to be the same. Application A and B will be able to interact only if each is precise about what it is identifying; if they are not, then mixing the two (in order to generate e.g. royalty payments in a third application C) may result in chaos through misunderstanding of what is being identified in each case.

- The means by which the assigner defines what entity is being identified is to specify the properties of the entity in terms of an underlying ontology. Assigning DOI System kernel metadata is a convenient starting point for doing so and ensures conformance without the need to be an expert in the underlying ontology. Ultimately, to ensure that a DOI name can be applied with any necessary degree of precision to any entity we use a model of making (from which is derived the indecs Data Dictionary: see “DOI® System and Data Dictionaries”) which deals with all possible entities through their context: the events which result in their formation (the interaction of agent, resource, time and place), and which therefore can relate two entities through the events of derivation of one from another.

**What are we resolving to from this identifier?**

A DOI name can resolve to anything. At minimum it will resolve to a URL, but there may be multiple URLs or multiple other data types returned on resolution. The issue is therefore not “what can a DOI name can resolve to” in a generic way, but rather what it does resolve to in any specific context. It depends on what the assigner wants it to resolve to in a particular application. The solution lies in a specific application domain rather than in the DOI system.

- Note specifically that what a DOI name identifies may not be what the DOI name resolves to. This initially may seem counter-intuitive but is obvious when one considers that (1) resolution may return multiple things; or (2) a DOI name may identify e.g. an abstract work, which by definition can only be perceived through a derived manifestation; or (3) a DOI name may identify a person, who is more than some set of data returned on resolution.

- To return to the previous example: an identifier of a work may well resolve to a specific manifestation (CrossRef DOI names are an example), yet this does not mean that another different manifestation of the same work would need a different DOI name.

**What metadata are we making available through this DOI name?**

Most DOI names are not yet used for widespread interoperability, but are used within specific applications. They do not need to reveal explicit structured metadata and have no defined (externally readable) Application Profile (they are formally called “zero AP” applications).

- Such existing applications include:
  - Using a DOI name which resolves to one (base) URL. The assigner ensures that what is resolved to is a single URL, and that this is appropriate for that application; but makes no guarantee that it is suitable for other applications. No explicit metadata is necessary as this is carried inside the application. (Example: CrossRef).
Using DOI names which resolve to multiple resolution data, but have no defined Application Profile. The assigner ensures that what is resolved to is a multiple set of data; that this data is appropriate for that application; but makes no guarantee that it is suitable for other applications. No explicit metadata is necessary as this is carried inside the application. [examples: many of the CDI applications using Multi-link™]

Interoperability is enabling information that originates in one context to be used in another in ways that are as highly automated as possible. Using DOI names for full interoperability implies that DOI names assigned in one context may be encountered, and may be re-used, in another context (or application) – without consulting the assigner. Interoperability of this sort is the case in the illustrative examples of applications A, B and C given earlier. Hence if the DOI name is designed to support interoperability, the metadata describing what the DOI name refers to must be explicit: the information necessary must be discoverable from the DOI name. The DOI System allows this in the form of an Application Profile (see below).

This does not imply that assigning DOI names which can be used interoperably requires the explicit declaration of all metadata the assigner collects; but it does imply that some (at minimum a kernel) is declared explicitly. This will prevent the confusion of e.g. work and manifestation cited in the earlier example (A,B,C).

Interoperability of DOI names
Any DOI name which is intended for interoperability – that is, which has the possibility of use in services outside of the direct control of the issuing Registration Agency (RA)- is subject to DOI System metadata policy, to achieve two objectives:

1. To ensure that metadata held by different RAs is not fundamentally inconsistent, and
2. To ensure that an efficient and extensible means of interchange exists for transporting metadata between RAs (and in future other service providers).

- DOI names are grouped through an Application Profile (AP) - a group of DOI names with some shared characteristic(s). This grouping will be used to connect DOI names to Services which are common to the group. A Service is a specific transaction that can be performed on or with the DOI name. Each DOI name is associated with one or more Application Profiles, and each AP is associated with one or more defined Services.
- The DOI® Data Model defines the way in which DOI names, Application Profiles, and Services relate. APIs have been built which allow application programmers to use the data model, using DOI® APs and Services within their applications.
- One of the services, possibly the only one for some DOI names, is the provision of kernel metadata for each DOI name. This is associated with AP 1. Other sets of metadata may also be available for some DOI names and this, as with other services, would be known through the inclusion of a given DOI name in an AP and the association of that AP with the given service.
- Following DOI System metadata policy and making available a kernel of declared metadata allows added value additional information (such as non-public metadata) to be built on top of the DOI name with confidence that it can be used interoperably as and when required.
- The provisions above do not apply to DOI names registered under the “Zero AP” which are not designed for open interoperability.
How will DOI names assigned in an application be paid for?

A cost is associated with managing persistence and with assigning identifiers and data to the standards needed to ensure long-term stability, because of the need for human intervention and support of an infrastructure. The DOI System operates on the basis that such costs are borne by the assigner of the DOI name. The way in which these costs are recouped depends on the application.

- DOI names are assigned on behalf of registrants (content owners and other bodies) by approved Registration Agencies (RAs). These RAs contribute to the IDF to support the infrastructure (technical and social) which underpins the DOI System.

- The three principle areas of cost are incurred in
  - Number and metadata registration (maintenance of resolution destinations; declaration of metadata; validation of number syntax and of metadata; liaison with the IDF registry)
  - Infrastructure (resolution service maintenance, scaling and further development; customer guidance and outreach; marketing; administration)
  - Governance (common "rules of the road"; further development and support of the system)

- RAs are free to establish their own business model for the allocation of DOI names. The services offered by a DOI RA will include more than simple provision of a DOI name: these value added services may include data, content or rights management. DOI RAs may also choose to collaborate with others in business agreements and services, as well as partake of the generic facilities offered by interoperability features of DOI names.

- There is no single business model applicable to all DOI RAs, and consequently no single answer to the question of how a DOI name is paid for and what it costs.

This is one of a series of DOI System factsheets. To see the latest version of this factsheet online, and to see the other factsheets go to:

http://www.doi.org/factsheets.html
Value Added by the DOI® System
Version 1.2

This factsheet provides a bullet-point overview of the value added by the DOI System. It aims to provide a useful checklist of the functionality of the DOI System and its value compared to other systems which provide some or all of the functionality provided by DOI® names.

For more detail on certain aspects of the DOI System underlying technologies, see the accompanying DOI System factsheets.

Principles of comparison
The value of the DOI System derives from three functions:

- Persistent identification
- Network actionability
- Semantic interoperability

Value is added by each of these three. Comparison with alternative systems depends on which of the functions (none, one, or more) is provided by the compared system.

Design
A system for persistent, semantically interoperable, identification of intellectual property entities on any digital network.

- Persistent Identifier: identifies the intellectual property entity itself rather than its location. Managed directory system relates to services such as location, enabling management of digital objects for long-term access. If ownership of the entity or the rights in the entity change, the identification of that entity does not change; the responsibility for managing the DOI name may change, but not the DOI name itself.
- Actionable Identifier - a user can use a DOI name to do something; facilitates simple location finding but also much more complex applications.
- Interoperable Identifier: allows use in services outside the direct control of the issuing assigner.
- Neutral as to technical implementation: designed to interoperate with past, present and future technologies. Assumes only the existence of Internet connectivity (not necessarily its underlying transport mechanisms).
- Neutral as to sector or media type: can apply to any format of physical, digital or abstract entities such as resources, parties, licences, abstractions

Standards based
The DOI System utilises existing components in building an added value application.

- Uses existing standards based components, which are also in use elsewhere – advantages of widespread use and support
• System as a whole to be an ISO standard (in process in TC46SC9): value of validation through an external standards process.

Syntax
Standard NISO Z39.84 syntax
• Standardised through NISO: provides branding recognition and ease of application.
• Unicode compliant.
• No length or character restrictions.
• No semantic requirement, but can incorporate semantics where appropriate and meaningful for other related applications.
• Can incorporate existing identifiers (from other schemes) where appropriate

Resolution
Uses Handle System®
• The Handle System provides a general-purpose global name service enabling secure name resolution over the Internet, designed to enable a broad set of communities to use the technology to identify digital content independent of location.
• In use in many other applications
• Standardised as IETF RFCs
• Separates identity of underlying digital objects from location; ensures persistence by resolving the DOI name to a current associated value such as a URL.
• Resolution may be to multiple pieces of data (multiple resolution).
• Supports administration granularity down to the level of individual DOI names
• Allows internationalized character sets; supports non-ASCII native characters
• Secured resolution service: supports client/server authentication, service integrity, and confidentiality, public key infrastructure if required.
• Architecture of the Handle System is flat, scaleable, and extensible: number of steps needed to resolve is independent of number of handles in existence
• Logically central, but physically decentralized
• Supports Local Handle Services, when desired
• Handle resolutions return entire “handle records” or portions thereof
• Handle records are also digital objects
• Handle servers are certificated with the system
• Handle records are signed by the servers
• DOI names are handles whose primary prefix is “10”; provides recognition
• Enhanced browser support available for direct handle access: handle plug-in, and available e.g. for Adobe Acrobat as plug-in
• Uses proxy servers for unenhanced browsers
• Types are resolvable in the Handle System
• Types may be created dynamically
• Types may be locally named, mapped into bit strings without semantics
• Runs on 24/7 secure extensible servers

Data structure
Multiple resolution of handles allows one entity to be resolved to multiple other entities; this can be used to embody relationships, e.g. a parent-children relationship, or any other relationship. Handle technology allows this; the DOI System provides a framework to achieve it.
• Allows precise specification of exactly what is being identified.
• Data model available for optional specification of DOI name records, grouping in application profiles and services for re-use, shared and/or common use.
• Data dictionary available, based on (and consistent with) ISO MPEG-21 RDD standard and other <indecs> activities; consistent with systems such as ONIX and Grid;
• Data Dictionary standardised (MPEG21) and in use elsewhere
Does not mandate one metadata scheme; allows use of any existing metadata scheme whilst enabling mappings between application areas consistently and production of XML messages etc.

Allows declaration of metadata to any level of functional granularity.

Technical Infrastructure
The International DOI Foundation funds, on behalf of all its members and DOI System users, certain aspects of technical infrastructure underlying the DOI System.

- Redundant distributed servers, both handle servers and any proxy or application servers required, plus backup procedures.
- DOI System specific development and implementation of proxy features (e.g. open URL functionality)
- Central DOI System Directory, staffed and outsourced with backup and reliability guarantees
- Service agreement for management and implementation of shared technical resolution infrastructure
- Data dictionary shared with ONIX (and potentially others), staffed and outsourced
- Uses improved Berkeley database compared to standard Handle System release; benefits of shared expertise, community of users with knowledge and service agreement for support.

Social infrastructure

- The Handle System technology provides persistence if used with appropriate social infrastructure. The International DOI Foundation (IDF) builds on the technical infrastructure of the Handle System a social structure guaranteeing persistence. Persistence is a function of organizations, not of technology; a persistent identifier system requires a persistent organization and defined processes
- Not-for-profit neutral organisation open to anyone to join
- Aligned with standards setting organisations, to promote level playing field
- Strength in numbers and economies of scale: for both technical and social infrastructure, leverage joint efforts of a community.
  - Facilitates applications to be developed & deployed
  - Facilitates Metadata registries to be created and maintained
  - Encourages third-party value-added providers
- Shared managed use of Trademark and intellectual property: branding, ease of concept marketing
- Shared outreach to other standards communities and potential user communities
- Shared research and development as necessary.
- Participates in defining wider Persistent Identifier space through Handle System governance and development
- Common policies and procedures for assignment of DOI names; provides rules and mechanisms for implementation;
- Certification process for “Qualified” Registration Agencies
- Scalable extensible franchise model to bring in new RAs
- Agreements to take over orphaned identifiers, distributed shared expertise in various areas e.g. proxy functionality, and potentially a number of agreements on how to structure applications.
- Shared licensing and purchasing agreements for technical services

Business potential
The DOI System enables registrants to incorporate persistent, semantically interoperable, identification of intellectual property entities as part of services providing added value

- Providing identification, management and metadata services
- Enabling third-party value-added capabilities: allows individual registration agencies to provide their own added value functionality on top of an agreed common infrastructure
- Selling infrastructure technology
- Helping organizations manage their own information better & offer new types of services
- Stimulating access to "surface information" and "embedded information" with appropriate access controls and conditions of use
- Necessary infrastructure for digital policy enforcement (including DRM)

This is one of a series of DOI System factsheets. To see the latest version of this factsheet online, and to see the other factsheets go to: http://www.doi.org/factsheets.html