Digital Object Identifiers for scientific data

Dr Norman Paskin
International DOI Foundation
Oxford OX2 8HY UK
n.paskin@doi.org

Paper presented at 19th International CODATA Conference, Berlin, Nov 10 2004

Abstract

The Digital Object Identifier (DOI) is a system for identifying content objects in the digital environment. DOIs are names assigned to any entity for use on Internet digital networks. Scientific data sets may be identified by DOIs, and several efforts are now underway in this area. This paper outlines the underlying architecture of the DOI system, and two such efforts which are applying DOIs to content objects of scientific data.

DOIs provide persistent identification together with current information about the object. The system is managed by the International DOI Foundation (IDF), an open membership consortium including both commercial and non-commercial partners, and has recently been accepted for standardisation within ISO. Several million DOIs have been assigned by DOI Registration Agencies in the US, Australasia, and Europe. DOI is a development combining several existing standards, notably the Handle resolution system and the indecs Data Dictionary. DOIs can be used for any form of management of data, whether commercial or non-commercial.

The DOI system has several components: a specified numbering syntax, a resolution service, a data model, and procedures for the implementation of DOIs. Any existing numbering schemes, and any existing metadata schemes, that provide an accepted numbering or descriptive syntax for a particular community or area of interest (such as formal ISO standards or accepted community practice) can be used within the DOI System.

Introduction

A Digital Object Identifier (DOI) is a name (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[1].

It has long been recognised that unique identifiers are essential for the management of information in any digital environment[2]. 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 DOI 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 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[3]: the entire system has recently been accepted for standardisation[4] within ISO (ISO TC46/SC9)[5]. The DOI 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)[6] founded in 1998. The DOI is in widespread use, with over 15 million DOIs assigned, from over 1000 naming authorities (allocators of DOIs). DOI forms a key feature of scientific primary publishing as part of the CrossRef system[7] (providing a pre-publication processing tool enabling cross-references to be persistent and not rely simply on URLs and bibliographic citation matching). DOIs are being adopted for use in government documents (such as EC, OECD, UK government, etc). In use, a DOI is a mechanism "behind the scenes", and need not be explicitly declared (though this may be useful): e.g. in a web context a DOI may be used in a http form as a URL (through a proxy server), whilst retaining the advantages of managed persistence. DOI may be used to offer an interoperable common system for identification of science data. Two current projects considered as examples here are the TIB project (on citation of primary data sets) and the Names for Life project (on biological taxonomy).

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 DOIs.

DOI syntax

The DOI 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[8]: e.g.

10.1234/NP5678
10.5678/ISBN-0-7645-4889-4 and
10.2224/2004-10-ISO-DOI

are all valid DOI syntax: the portion following the "/" character (the DOI Suffix) may be an existing identifier. The portion preceding the "/" character (the DOI Prefix) denotes a unique naming authority.

The word "identifier" can mean several different things: (1) labels: the output of numbering schemes e.g. "ISBN 3-540-40465-1"; (2) specifications for using labels: e.g. on internet URL, URN, URI (URI = Uniform Resource Identifier); or (3) implemented systems: labels, following a specification, in a system - e.g. the DOI system, which is a packaged system offering label, tools and implementation mechanisms. The DOI system is an example of identifier sense (3); it may include identifiers in sense (1) as a suffix and may also conform to identifier specifications[9] in sense (2).

DOI 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 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 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 DOI using multiple resolution are now increasingly in use. The resolution tool used in the DOI system is the Handle system[10] (IETF RFCs 3650, 3651, 3652). This conforms to the functional requirements of the URI and URN concepts, and has many advantages over other mechanisms including global scalability, security, and opacity[11].

The Handle system implementation in DOI has been supplemented by expanded technical infrastructure and features specific to DOI applications[12]. 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): DOI is an application of Handle which adds this constraint for a specific purpose of content management (analogy: a spreadsheet application). In DOI the constraints are defining 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[13]. Together these provide tools for defining what a DOI specifies (through use of a data dictionary), and how DOIs relate to each other, (through a grouping mechanism, Application Profiles, which associate DOIs 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 DOIs 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. DOIs need not make use of this data model, but it is envisaged that many will: any DOI 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[14]. 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.

DOI implementation

DOI is implemented through a federation of Registration Agencies which use policies and tools developed through a parent body, the International 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 components to ensure that any improvements made to the DOI system (including creation, maintenance, registration, resolution and policymaking of DOIs) are available to any DOI registrant, and that no third party licenses might reasonably be required to practice the DOI standard. DOI resolution is freely available to any user encountering a DOI[15].

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. This is a role of DOI 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[16].

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 DOI in “own brand” ways appropriate for their community.

DOIs and scientific data

An earlier presentation[17] outlined the rationale for assigning DOIs to data, outlined some early suggestions for applications, and summarised the DOI system as it stood at that time.

The identification of scientific data is logically a separate issue from the identification of the primary publication of such data to the scientific community in the form of articles, tables, etc. DOI is already the core technology for maintaining cross-references via persistent links between a citation and internet access to article, in the CrossRef system used by over 350 publishers representing the majority of STM articles. DOI is used in a pre-publication link builder and is now being extended to specialist uses such as constrained Google searching. Currently 9,000 DOIs per day are added to CrossRef; over 12 million DOIs are now registered with CrossRef, of which over 850,000 are assigned to books and conference proceedings.

The DOI as a long-term linking option from data to source publication is of fundamental importance. For example, in the Landolt-Börnstein series of tables of Numerical Data and Functional Relationships in Science and Technology, the automatic generation of 30,000 documents is to be done with DOIs (already assigned to book publications): in the online version of the LB collection, a new feature of all 2004 volumes will be improved metadata and cross reference linking enabled in this way[18].

Some projects or communities have developed their own identifier schemes, which may be useful for their own area. A recent example is the Life Science Identifier[19] developed by I3C/IBM: this provides a community agreement around a simple URN mechanism, which is non-generically extensible and non-globally resolvable but meaningful within the bio-informatics community for certain purposes. Such identifiers can be incorporated into a DOI to make them globally interoperable and extensible and take advantage of other features provided in the DOI system.

Recently two projects in particular have undertaken interesting DOI applications with science data: these are briefly described here as illustrative of two different areas of DOI use.

DOIs for scientific data sets: TIB project

DOIs could logically be assigned to every single data point in a set; however in practice, the allocation of a DOI is more likely to be to a meaningful set of data following the indecs Principle of Functional Granularity[20]: identifiers should be assigned at the level of granularity appropriate for a functional use which is envisaged.

The German National Library of Science and Technology (TIB), the world's largest library of science and technology participated in a project made possible by a grant from the Deutsche Forschungsgemeinschaft (German Research Foundation), to implement the use of DOIs to persistently identify scientific data sets. This follows from earlier work by a National Committee of CODATA, the Committee on Data for Science and Technology of the International Council for Science (www.icsu.org), resulting in an internal report "Concept of Citing Scientific Primary Data" in May 2002 recommending the use of DOIs. A continuation as a project for pilot implementation funded by DFG Oct 2003 to Oct 2005 at TIB (German National Library of Science & Technology) is co-ordinated

by the World Data Center for Climate (WDCC) at the Max-Planck-Institut für Meteorologie in Hamburg[21].

A team from several data centres, led by Dr Michael Lautenschlager at the World Data Centre for Climate, will focus on providing a means of publicly registering data sets with a persistent identifier and structured basic description. The pilot deployment will use geo-reference data (e.g. from observational stations, satellites, and climate models), but will be designed to be extensible to any scientific data. This use of DOI will provide for the effective publication of primary data using a persistent identifier for long-term data referencing, allowing scientists to cite and re-use valuable primary data. The DOI's persistent and globally resolvable identifier, associated to both a stable link to the data and also a standardised description of the identified data, offers the necessary functionality and also ready interoperability with other material such as scientific articles.

The key problem this project addresses is the reliable re-use of existing data sets, in terms both of attribution of data source (the proposed solution being to make data publications citable in a standard way as are articles through the Science Citation Index), and the archiving of data in context so as to be discoverable and interoperable (usable by others). The extensibility of the mechanism is provided by allocating DOIs for data sets, with associated metadata using a core management metadata (applicable to all datasets) and structured metadata extensions (mapped to a common ontology) applicable to specific science disciplines.

Application scenario[22]:
During her research for the World Data Center Climate (WDCC) Dr. Weather gains primary data about the weather in Hannover in the year 2003. Primary data is tested, evaluated, stored and administrated at the WDCC. Primary data is registered and allocated DOI at the TIB, with quality control of metadata, etc (e.g. the assigned data cannot be changed once allocated).
Dr Weather can now cite this with a resolvable DOI e.g

DOI:10.1594 /WDCC/W_Han_2003_MMB_2	
10.1594 (Prefix)	= TIB as the registration agency.
WDCC	= research institute.
W_Han_2003_MMB_2	= internal name of the Data

The DOI is resolvable directly, or via http as
http://dx.doi.org/10.1594/WDCC/W_Han_2003_MMB_2

Usage scenario 1:
Dr. Storm is reading publications from Dr. Weather in a journal and would like to analyse her data under different aspects. The DOI may be resolved to obtain the data set for use. In his publication "Comparison of the weather from Hannover and Miami" Dr. Storm cites Dr. Weather's data using its DOI, referring to the uniqueness and own identity of the original data.
Citation example:
Weather, 2003: "Weather in Hannover for 2003"
(doi: 10.1594/WDCC/W_Han_2003_MMB_2)

Usage scenario 2:
Mr. Nice is writing a paper about the sales figures of ice cream in Hannover in 2003, but he has no information about the weather. He searches via TIB central registration agency metadata search; the result is
doi:10.1594/WDCC/W_Han_2003_MMB_2
He resolves the DOI to find the data. The metadata refers him to the WDCC as publisher and data archive. In his paper he cites the data using the DOI.

DOIs for taxonomic data: Names for Life project

The aim of this project is "future-proofing biological nomenclature"[23]; it proposes DOIs as persistent identifiers of taxonomic definitions. A name ascribed to a given group in a biological taxonomy is fixed in both time and scope and may or may not be revised when new information is available. Change occurs (e.g. new species are recognised, species reassigned as the founding species of new genera; synonyms; species split into subspecies which later became separate species) resulting in changes of names, genera, families, classes, and relationships over time. When taxonomic revisions do occur, resulting in the division or joining of previously described taxa, authors frequently fail to address synonymies or formally emend the descriptions of higher taxa that are affected. DOI is proposed as a tool to manage a data model of nomenclature and taxonomy (enabling disambiguation of synonyms and competing taxonomies) using a metadata resolution service (enabling dissemination of archived and updated information objects through persistent links to articles, strain records, gene annotations and any other data).[24]

Whereas the different Codes of Nomenclature guarantee persistence of a formal name, the serial, cumulative nature of effective and valid publication allows the name to obsolesce in relation to the taxon it originally denoted. In contrast, it is the taxon itself that persists, and the granularity with which it is defined increases over time. The formal name provides an archival record of taxonomic definition only for a single point in time: the date of publication. A robust and persistent taxonomy requires taxonomic definition to be a maintained, networked resource, rather than a retrospective sequence of names and emendations. A commonly referenced terminology based on persistent, increasingly refined taxa is needed to replace or augment a static nomenclature that diverges over time from the taxonomy it initially denotes. This disjunction of nomenclature and taxonomy results in an accumulation of names of dubious value in the literature and databases.

The Names for Life project is developing a model for assigning DOIs to prokaryotic taxa as a test case. Though the definition of a taxon may be refined and its nomenclature redefined, the DOI will persist, leaving a forward-pointing trail that can be used to reliably locate digital and physical resources, even when a name may be deemed obsolete. Forward linking from a synonym to a record of the publication that asserts synonymy is especially important, as there is currently no mandatory mechanism for asserting and resolving names that become ambiguous. The model seeks to strengthen the association of names with taxa by using DOIs to track the taxonomic definition of a name over time. It is extensible to the level of individual genes within a given species. However, the real power of this method lies in the ability of DOIs to become embedded in the information environment, providing a direct and persistent link to the full record of taxonomic and nomenclatural revision and ensuring consistency and accuracy throughout online scientific resources. A DOI-based infrastructure for formally associating nomenclature with taxonomy enables a name to be used unambiguously and persistently, only one mouse-click away from a record of its current definition and historical development.

References

[1] Paskin, Norman. "Components of DRM Systems: Identification and Metadata" in E. Becker et al (eds) "Digital Rights Management: Technological, Economic, Legal, and Political Aspects in the European Union" in the series Lecture Notes in Computer Science (Springer-Verlag, 2003) pp. 26-61. http://www.doi.org/topics/drm_paskin_20030113_b1.pdf

[2] Svenonius, Elaine; The Intellectual Foundation of Information Organization; MIT Press; 2000

[3] The DOI Handbook
http://www.doi.org/hb.html

[4] Resolutions of the ISO TC46/SC9 meeting in Washington, document SC9N395

[5] ISO TC 46/SC 9, Information and Documentation - Identification and Description
http://www.collectionscanada.ca/iso/tc46sc9/index.htm

[6] International DOI Foundation, http://www.doi.org

[7] CrossRef, http://www.crossref.org

[8] DOI Factsheet: DOI and Numbering Schemes: http://www.doi.org/factsheets/DOIIdentifiers.html

[9] DOI Factsheet: DOI and Internet Identifier Specifications
http://www.doi.org/factsheets/DOIIdentifierSpecs.html

[10] Handle System: www.handle.net

[11] Dyson, Esther: "Online Registries: The DNS and Beyond". Release 1.0, September 2003
http://dx.doi.org/doi:10.1340/309registries

[12] DOI Factsheet: DOI and Handle http://www.doi.org/factsheets/DOIHandle.html

[13] DOI Factsheet: DOI and Data Dictionaries
http://www.doi.org/factsheets/DOIDataDictionaries.html

[14] Sowa, John F; "Knowledge Representation: Logical, Philosophical and Computational Foundations"; Brooks/Cole; 2000
http://users.bestweb.net/~sowa/krbook/

[15] IDF re-affirms DOI as an open specification. DOI news October 2004
http://www.doi.org/news/DOINewsOct04.html

[16] DOI Factsheet: DOI Applications
http://www.doi.org/factsheets/DOIApplications.html

[17] Paskin, Norman. "Digital Object Identifiers and Digital Preservation of the Record of Science". Proceedings of ICSTI Seminar: Symposium on Digital Preservation of the Record of Science, ICSTI, 14 - 15 February 2002, IOS Press, 2002.
http://www.doi.org/topics/020210_CSTI.pdf

[18] Dr. Rainer Poerschke, Editorial Director Landolt-Börnstein, personal communication

[19] Salamone, S: "LSID: An Informatics Lifesaver"
http://www.bio-itworld.com/archive/011204/lsid.html

[20] Rust, Godfrey; Bide, Mark (2000). "The <indecs> Metadata Framework: Principles, model and data dictionary."
http://www.indecs.org/pdf/framework.pdf

[21] Lautenschlager, M. and Sens, I. (2003) "Konzept zur Zitierfahigkeit wissenschaftlicher Primierdaten" Information – Wissenschaft und Praxis 54 (2003) 463-466

[22] "Pilot Implementation: Publication and Citation of Scientific Primary Data": Jan Brase and Michael Lautenschlager. Presentation to International DOI Foundation meeting June 2004
http://www.doi.org/idf-members/members_meeting2004/presentations/CorkLondon-PrimData-DOI-0604-v3.ppt

[23] Garrity, G. M.; Lyons, C. "Future-proofing biological nomenclature". Omics, 2003, Volume 7, Number 1, pgs. 31-33. This material was presented at: Workshop on Data Management for Molecular and Cell Biology Feb. 2-3, 2003 Lister Hill Center, NLM, NIH Campus, Bethesda, MD. A version of the paper as presented is at the workshop website: http://www.eecs.umich.edu/~jag/wdmbio/garrity.htm

[24] "Names for Life": Catherine Lyons. Presentation to International DOI Foundation meeting June 2004
http://www.doi.org/idf-members/members_meeting2004/presentations/n4_220604.ppt