



DOI and data dictionaries

Version 1.0

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 *ontologies*. An *interoperable* data dictionary contains terms from different computerized systems or metadata schemes, and shows the relationships they have with one another in some 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 *index Data Dictionary (iDD)*. DOIs need not make use of this, 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.

The index 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 "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 metadata system (based on the indecs Data Dictionary), and policies and procedures for the implementation of DOIs through a federation of Registration Agencies.

- One component of the DOI System is the iDD. Its implementation in DOI 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 metadata 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 DOI

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, 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 DOI, 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 DOI.

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. Of course, 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, 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.
- For those who need a detailed knowledge of the underlying methodology and construction, this is described in the DOI Handbook appendices. 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.

Physical access to iDD

An automated web based look-up system for the Dictionary is under development.

- Those authorised to access the system will be able to do so, while those not authorised are denied access.
- 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 Terms, if they exist, are kept confidential.

iDD and the MPEG-21 Rights Data Dictionary (RDD)

The ISO MPEG-21 Rights Data Dictionary is another notable data dictionary built on similar principles. It derives from work funded by IDF and others using the same methodology as the iDD. Consequently the two are closely related and fully integrated.

- 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 MPEG 21 RDD requirements in terms of management and availability for MPEG use are very similar to those of the iDD in relation to DOI implementation. For this reason, the IDF is currently proposed as the Registration Authority for the MPEG-21 RDD, and if accepted will subcontract management of the RDD and iDD to the same expert supplier, Ontologyx.
- RDD has a requirement that 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. A copy of the RDD Dictionary must be made available upon request to National Bodies of JTC 1 that are members of ISO or IEC, to liaison organizations of ISO or IEC and to any interested party.
- The MPEG Rights Data Dictionary provides the necessary semantic interoperability for use of rights expression languages and other tools.