Towards a rights data dictionary

Identifiers and semantics at work on the net

by Norman Paskin, Director, International DOI Foundation

Norman Paskin, Director of the International Digital Object Identifier Foundation, describes how internet technology and semantic definition systems are coming together to offer tools for multimedia e-commerce based on open standards.

Bits of computer code called 'digital identifiers' became an integral element of global commerce as soon as computer communication got seriously involved. In any real-world transaction, the parties involved know what is being traded because they can generally see it, touch it and, if needs be, ensure they are all in the same room with it. But online, transactions demand some way to ensure that everyone is talking about the same thing. This usually means some way of referring unambiguously to the subject of a transaction. Mark Bide noted recently in imi insights (Only Connect, April 2002), however, that digital communication – and by extension, the whole field of identifiers – is about more than just protocols and syntax like XML. To connect fully with the real world it also has to be about semantics. The significance of standardised semantics – a means of ensuring a degree of universal machine-to-machine understanding – is now firmly understood in the academic and theoretical domains and recognised in concept in the notion of the 'Semantic Web' but, in Mark's understated words, "porting that theoretical work into the rather messy real world of business will be a challenge."

The Digital Object Identifier (DOI) initiative is not alone in recognising the usefulness of unique identifiers for digital entities ('digital objects') in a network environment but it is a leading player. Importantly, using DOIs, names of such objects can be 'resolved' in order to create useful services. Resolution is the process of submitting an object's name to a network service and receiving one or more pieces of current data related to the entity identified by the name. Resolution ensures that a single name can be used to manage the object, even if any of those individual pieces of data about it change. For example, if a URL changes, the DOI will still 'resolve' to the updated URL, without the need to change bookmarks and other references. More interestingly, this data could be URLs or any other defined protocols - the object itself, locations or services about the object etc. (examples of such services have been created by the DOI Foundation). But, whatever the data, some structured, semantically meaningful metadata is needed to be associated with whatever is retrieved if the retriever is to make use of it in an automated environment.

The need for interoperable data in e-commerce systems was the basis for the 1998-2000 indecs project. Where DOI has taken a unique pioneering role is in adding to resolution the use of this semantic approach: linking a commercially useful application of resolution to intellectual property (the internet Handle System) with structured semantics based on the indecs model. Initial DOI implementations in the text sector are now being supplemented by increasingly sophisticated value-added tools for metadata management and multimedia content management. Critical to the web services model is the notion of 'the resource'. 'DOI Services' will be web services that exploit the power of DOI-named resources (objects) and the additional semantic information this confers.

The DOI effort was in the forefront of demonstrating the synergy between identifiers and semantics. Whilst showing the advantages of using resolution was an easy win (e.g. as used by journal publishers in CrossRef to avoid the problem of "linkrot" with URLs), demonstrating the role of semantics and providing a consistent set of tools to use, has been a hard challenge but one which now promises to yield a substantial prize. The International DOI Foundation (IDF) is one of the organisations which funded the original indecs framework and has continued to develop it further in partnership with an increasing range of organisations. The adoption of this work late last year as the basis for the MPEG Rights Data Dictionary, a fundamental part of the emerging ISO MPEG-21 Multimedia Framework, was highly significant. MPEG stands for Moving Picture Experts Group, an organisation originally conceived to standardise the compression schemes for digital images, both still and moving. Today,

its activities range far more widely. In particular the MPEG-21 initiative provides an overarching framework for the all-electronic creation, production, delivery and trade of all kinds of content and content combinations. The idea is to apply it to a wide range of content-based network initiatives including digital library development, broadcasting, music and video distribution, asset management, content filtering and, of course, all forms of electronic publishing. The MPEG-21 work is now on track to produce a definitive international standard in 2003 and useful early applications well before then.

Semantics antics

The original <indecs> analysis was a reference model. Fine and well, but practical implementations are required. So, in April 2001, IDF funded a feasibility study for the development of a Rights Data Dictionary (RDD) based on indecs. As a result, from mid-2001, a consortium of rights holder representatives and providers of services agreed to develop this in an initiative called <indecs>rdd. The <indecs> consortium represents major groups of rights owners and ancillary service providers. Currently its members are the International DOI Foundation (IDF), the Motion Picture Association of America (MPA), the Recording Industry Association of America (RIAA), the International Federation of the Phonographic Industry (IFPI), Enpia Systems, and Melodies and Memories Global (a subsidiary of Dentsu). The project is managed by Rightscom Ltd.

The fundamental design for such a tool was submitted to MPEG in December 2001 and subsequently selected as the baseline for the MPEG-21 Part 6 *Standard for a Rights Data Dictionary*. The consortium has continued to complete the standard specification. A secondary but important objective is to put the dictionary to work as an operational system. The dictionary will standardise several hundred terms as part of a structured semantic 'ontology', a schema setting out standardised rules (which computers can handle) governing the meanings and relationships among terms used in intellectual rights management and trading. This is good news: applications like DOI will be able to use these and the process of creating the dictionary will involve mappings to key metadata sets already used in e-commerce such as ONIX.

The need for a standard rights data dictionary

Rights management has to work in an open, distributed, multi-protocol computer environment. The huge amount of digital content now being traded – legally and illegally – requires an infrastructure for rights. The terms used in various 'rights expressions' which mediate the use of digital items – ownership statements, licenses, permissions, offers, requests and agreements – need to be unambiguously understood by computers. Together these terms are often called rights metadata.

For instance, if a license agreement states that a commercial *consumer* must *pay* a particular *fee* to *copy, play* and *keep* a particular *format* of *a digital file* in a particular time and place and that a *student* may do the same for a *reduced price,* all those terms (in italics) must be interpreted by a computer or user to mean what is intended by the licensor. To achieve such a level of unambiguous interpretation, there must be a common data dictionary of terms involved in rights. This is a common requirement in computing but in the area of rights management there are three problems which make it especially challenging.

Three problems

- First, rights are complex. Rights metadata can quickly become much more complicated than the simple license example given above. For example, all kinds of media, content and usage might be involved, including rights in underlying abstract works and ownership of rights often changes over time. A rights data dictionary must, therefore, be capable of supporting the simplest through to the most complex of rights expressions.

- Second, rights expressions will be mixed with other types of information. Agreements, offers and licenses may include any terminology taken from descriptive, legal or financial systems. A rights data dictionary must be broad enough to embrace terms from any other kind of metadata that might occur in a rights expression.

- Third, many dictionaries are already in use. Different market sectors, individual companies and organizations may have their own working dictionaries and schemes. Some deal with rights, some don't. Many groups will not want

or be able to change to a new dictionary or use a new one alongside the terms from their own namespace. Yet, because these groups are now all co-operating in common multimedia areas, some way of connecting them is essential. In other words, an effective rights data dictionary must allow the use of terms from existing and future namespaces.

The solution

The architecture for the <indecs> data dictionary has been developed over a number of years to cope with just these problems of complexity and interoperability. It combines the main elements of a data dictionary, a multi-lingual dictionary, an ontology and a thesaurus and is well suited to this task because:

- It has a powerful conceptual base, based on a strong and mature underlying data model, a core of several hundred terms to which any number of others may be added in a systematic way.

- It is highly structured. Every term has a unique identifier and a 'genealogy' that defines precisely and logically how it relates to others. Because the underlying model is very rich, it can accurately describe highly complex relationships between terms.

- It is inclusive. Any terms from other dictionaries can be added (by assigning a unique identifier and a genealogy). Other terms are not just 'extensions' or 'mapped' words. They become an integral part of the dictionary.

- It is highly granular, able to support terms at any level of detail, fragmentation or versioning required by users.

- Users can 'mix and match' terms. Because any 'mapped' scheme is part of the dictionary, terms from different namespaces can be combined to form rights expressions without loss of meaning.

- It supports 'transformations' to provide the underlying semantic tools to translate terms from one scheme to another in a highly automated way. This is critical to allow different metadata schemes to co-exist in the multimedia environment.

- It is legally neutral. <indecs>rdd does not define legal terms. It can be used to make rights expressions that draw on any existing legal definition, or none.

- It is business-model neutral. <indecs>rdd terms can be used to describe any situation in which any kind of rights are owned, managed, protected or used.

- It is not a Rights Expression Language (REL). A data dictionary is not an expression language (such as XrML, now adopted as baseline technology for the complementing MPEG-21 REL standard). An REL deals with the way in which terms are expressed in computer language. The dictionary defines the terms. An REL will use terms defined in an RDD.

- It has uses beyond rights. Because of its generalized model, <indecs>rdd can provide a comprehensive basis for metadata expressions and schemes for purposes other than rights – such as resource description, workflow management and event reporting. It could be used as a tool for the deployment of semantic based web services.

How <indecs>rdd will be used

Why would a company want use such a tool? Well, think of the reality of implementing today the sort of agreement mentioned above. <indecs>rdd is a tool which will be used in an automated way (often invisibly) to define precisely all the terms required in such an implementation and to help to create, transform and interpret such expressions. Using an open standard will, of course, confer the usual benefits of easy use by others, lower overall cost and conformance. It will provide a ready-made standard terminology for rights management. Organizations needing to create rights expressions or to enhance their existing metadata schemes, will be able to use <indecs>rdd as a source for terminology. Apart from providing a structured basis for metadata selection, it ensures interoperability with other compliant schemes from origination to end-use.

The dictionary will grow constantly as other schemes are mapped and so (as with anti-virus software) regular updates will be an essential component. It will support application software at all points in the 'content chain'. <indecs>rdd will

be available to support the making, transforming and interpreting of rights expressions.

Closing the circle: integrating identifier resolution and semantics

The DOI is designed to provide the technical means to deliver business services. A DOI could (in future) have at least one application profile (AP) and have descriptive metadata encoded according to a scheme using the dictionary. The DOI record (the resolution step) could contain one typed value that identifies the AP and one typed value that contains the descriptive metadata, either directly or by reference (as XML).

One useful way of dividing up potential applications in intellectual property transactions is suggested by the highest level of the indecs data model: 'people' do 'deals' about 'stuff'. Each of those three entities will need identifiers:

- So far, 'stuff' (i.e. things that are transacted such as documents, recordings, images etc) are the subject of DOI applications. It's pretty obvious that 'stuff' can be any creation, that is physical (a book: like an ISBN); a digital file; an abstraction (a Work); or a performance. That is, we are not restricted here to identifying digital objects that are content. The digital objects can be referents or services about an entity (more accurately: the data that is returned from resolution could be the object itself if a digital file or some information about the object if a non-digital entity). One can for example imagine a DOI identifying an abstract Work (e.g. 'Alice in Wonderland') which resolves to various sources of manifestations of that work (printed editions, etc). (http://www.doi.org/handbook_2000/enumeration.html#4.8).

- 'People' (in the wider sense of individuals, organisations, character names, etc) is a potential identifier implementation that is the topic of wide interest, including a recently launched one-year project (Interparty). The project, funded by the EU, is an initiative aimed at building the interoperability of party numbering systems – to provide a means of online, on-demand, checking of identities of parties and to specify and develop an exploitation plan and governance structure for a Directory of Parties.

- 'Deals' would be the identification of specific licences, agreements, specific transactions etc. This is already a requirement of the music industry and many likely e-commerce systems in the future. Again, using an open standard rather than proprietary systems makes sense.

'Deals' also takes us on to identifying other things which may technically be 'stuff' but which have some more fundamental role as instruments of value (in the same way that you can consider a stock certificate as 'a document' while, in fact, the more meaningful way of dealing with it is as a representation of a financial instrument). There is a fascinating paper by Kahn and Lyons on this subject.

So, is this the complete solution?

Semantics fully automated? Not quite. The nirvana of totally automated knowledge representation is still some way off. Moving up a layer, something is going to have to *interpret* and *act on* the metadata for DOI services, so some DOI client software that is asked to show or use some specific piece of metadata for a given DOI should be able to parse that instance or know someone who could. This is still a huge step but by providing a functioning framework of resolution and semantics, the stage is set for many such applications to be built.

As a leading researcher, John Sowa, has noted: "Knowledge representation is a multidisciplinary subject that applies theories and techniques from three other fields:

1. Logic provides the formal structure and rules of inference.
2. Ontology defines the kinds of things that exist in the application domain.
3. Computation supports the applications that distinguish knowledge representation from pure philosophy.

Without logic, a knowledge representation is vague, with no criteria for determining whether statements are redundant or contradictory. Without ontology, the terms and symbols are ill defined, confused, and confusing. And without computable models, the logic and ontology cannot be implemented in computer programs. Knowledge representation is the application of

logic and ontology to the task of constructing computable models for some domain".

One might add that to make all this real for businesses, a fourth component - a real, working system - is now within sight.

From the EPS Archive

Only Connect
(imi, April 2002)

Collection-level description: so whose metadata is it anyway?
(EPS Update Note, 7 May 2002)

DOI-EB: Metadata is the critical bridge
(EPS Update Note, 20 March 2001)

Author Information

Dr. Norman Paskin became the first Director of the International DOI Foundation in March 1998. Prior to this he worked for twenty years in the scientific publishing industry in both the United States and Europe, most recently as Director of Information Technology Development for Elsevier Science (1994-1998). He was actively involved in information identifiers issues for the scientific technical and medical publishing community, and has published several papers on this and related topics. Dr. Paskin is a member of the Board of Directors of the U.S. National Information Standards Organization (NISO). He can be contacted at n.paskin@doi.org.