Digital Rights Management

Technological, Economic, Legal and Political Aspects in the European Union

Last Revision: January 13, 2003

digital-
rights-
management.org

2.3 Components for DRM Systems

2.3.1 Identification and Metadata

Norman Paskin [1]

Abstract

Identifiers (unique labels for entities) and metadata (structured relationships between identified entities) are prerequisites for DRM. The term identifier can mean a label numbering scheme, specification, or fully implemented identifier system in a specific infrastructure. Implementations require a social infrastructure. In an automated environment, the entity being managed must be defined in a structured way, by means of attributes. Managed entities will often be abstractions, and the choice of which possible entities to distinguish as separable is not absolute but dependent upon function and context.

Interoperable DRM requires a persistent means of identification and structured description. Persistent identification can be aided by use of Internet technologies which allow indirection, separating names from attributes. Structured description requires an ontology framework, such as the indecs framework, which can support mappings using a managed data dictionary.

The Practical Significance of Identifiers and Metadata in DRM

As commerce has become increasingly less dependent on the physical presence of both buyer and seller, means of identifying things uniquely and describing them unambiguously have become more and more important. The use of computers in mediating some aspects of the trading relationship has further accentuated this requirement. The near-universal adoption of "unique identifiers" such as the ISBN or the UPC/EAN barcode has been a direct consequence (and a precondition) for the development of EDI (electronic data interchange) and electronic trading.

The Internet, as it becomes a medium for trading in intellectual property, drives us several steps further. The digital network linking trading partners has for the first time to embrace consumers rather than simply supporting business-to-business transactions. The identity of the things that can be traded becomes much less clearly delineated when they may be computer files rather than physical objects. Users no longer have to access "content" only in pre-packaged products – it becomes possible to provide them with the precise customized package

[1] Norman Paskin, International DOI Foundation, n.paskin@doi.org.

of content that they want (and which theoretically at least no one else may want). By the Internet we mean here the network of digital computers linked together by a globally unique address space based on the Internet Protocol (IP) or its subsequent extensions, able to support communications using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite or other IP-compatible protocols; and providing high level services layered on that infrastructure [2]; the World Wide Web is only one such manifestation. In addition, many identifiers and metadata will be used in private, EDI, or other networks: hence a sound design principle is application independence: identifier and metadata structures should be independent of any specific technical expression.

In digital rights management (which I'm defining broadly here as the management of any rights, including those of non-digital entities, through digital means), we use digital representations of resources, parties, licences and other entities (digital objects) to articulate a property system. One of the most important things a formal property system does is transform assets from a less accessible condition to a more accessible condition, so that they can do additional work. Unlike physical assets, representations are easily combined, divided, mobilized, and used to stimulate business deals. By uncoupling the economic features of an asset from their rigid, physical state, a representation makes the asset "fungible" - able to be fashioned to suit practically any transaction [3]. Digital objects may also directly represent value [4], though for current DRM purposes we are largely content to have DRM technologies work with normal currency mechanisms - concepts such as DigiCash, Beenz and the like have not (yet) found success.

The management of the myriad transactions implicit in such a complex network environment will only be possible if mediated by computer systems. This puts additional pressure on the requirement for unambiguous identification and description of the content through metadata. Persistent identification and description is a prerequisite for the management of intellectual property rights in the digital environment. Whilst identification of content is the most advanced area - perhaps because in many ways the easiest - the same principles apply to identification of all entities involved in rights transactions: parties, resources and agreements, as described in the indecs (interoperability of data in e-commerce) model of commerce [5]. The indecs framework has been widely recognised as a significant contribution to understanding metadata in the context of DRM, and the present article draws heavily on the indecs work and its implementation in the Digital Object Identifier [6], though the principles discussed, and conclusions drawn, are independent of any specific application.

[2] Kahn, Cerf (1999).
[3] De Soto (2000).
[4] Kahn, Lyons (2001).
[5] Rust, Bide (2000).
[6] DOI.

The Relationship of Identifiers and Metadata

Identifiers and metadata are two sides of the same coin. An *identifier* is an unambiguous string denoting an entity; an *item of metadata* is a relationship that someone claims to exist between two entities, each of which may have an identifier (and must, in an automated environment). These entities may include both objects and concepts: e.g. an item of metadata may be "this book has a cover coloured blue", and that blue may be specifically identified by a Pantone number; both the book and "blue" would be identified entities. *Entity* is a term used to mean simply something that is identified. The underlying idea, from the <indecs> project, is that nothing exists in any useful sense until it is identified.

An *ontology* is a tool which is able to structure relationships between entities; 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 [7].

Identifiers

An identifier is an unambiguous string or "label" that specifies an entity (something that is identified). Note that the term "identifier" has become rather overloaded and is used synonymously for several related concepts; discussed in more detail in section 5. In computer science terms, an identifier is a name; the entities named occupy a specific domain of application (the namespace) and are points in that namespace. "Naming is one of the most important and most frequently overlooked areas of computer science. In computing it is rumoured: everything is a naming problem" [8]. Once points in a name space are addressable, applications can be constructed which provide links (i.e. denote relationships) into the namespace or between points, to express metadata. Identifiers assigned to intellectual property entities would enable connections to be denoted (at an intellectual level and in practical terms for trading) between entities which are physically separated, which may be abstract properties, or are the product of separate authors etc.

The principal reason for assigning identifiers to points in a namespace is to realise that abstract namespace as a real digital environment (addresses in a network or computer system), which can then be readily manipulated. Information expressed in a digital manifestation is a Digital Object: "a data structure whose principal components are digital material, or data, plus a unique identifier for this material" [9]. "A digital object is not merely a sequence of bits or symbols...it has a structure that allows it to be identified and its content to be organized and protected..." [10]. These definitions capture the idea that a digital

[7] Sowa (2000).
[8] Irlam (1995).
[9] Kahn, Wilensky (1995).
[10] Cross Industry Working Team (1997).

object is a meaningful piece of data, reflected in other descriptions such as DLO (Document-Like Objects [11] or KNOBs (Knowledge Objects) [12].

From the standpoint of intellectual property or "content", an Object is a digital subset of a greater class of entities, Creations (products of human imagination and/or endeavour in which rights exist) encompassing in addition to digital objects, physical packages, spatio-temporal performances, and abstract works. Intellectual property - broadly, "works of human intellect or imagination" - can be formally defined in an ontology such as indecs, but where possible the analysis references definitions agreed by the World Intellectual Property Organization and related international treaties like the Berne Convention. These Creations may each have applicable namespaces, not all of which have digital realisations. From the standpoint of the Internet, a Digital Object is a Resource as specified in the Uniform Resource naming schemes.

Unique Identification

Uniqueness is the essential attribute of an identifier, which must be unambiguous in the defined namespace: a given identifier must specify (be bound to) one and only one object in that space. This does not imply that one object may have only one identifier (a one-to-one relationship), since a one-to-many relationship (an entity having several labels, each unambiguously specifying it) may be necessary in some contexts, and is likely in many DRM applications: as multimedia entities become more complex, or parties such as publishers operate in multi-media, multi-national environments, it becomes inevitable that they will acquire more and more domain identifiers, which may or may not require reconciliation. The question of whether - or how - different identifiers for the same entity should be reconciled is both practical and political. The multiple labels may be valid in different namespaces to guarantee interoperability (e.g. a sound clip within a multimedia scientific document may have one identifier within a music identification scheme, another identifier within a document archive); or the multiple identifiers may be within the same namespace, perhaps for pragmatic reasons beyond the abstract design of the namespace.

The indecs *Principle of Unique Identification* is that "every entity should be uniquely identified within an identified namespace". It is difficult to overstate the importance of this simple and commonplace principle. At one level it can be said that the basis of interoperable metadata is simply about the relationships of recognisably unique identifiers. In pre-digital bibliographic and commerce systems, effectiveness depends to a great extent on the robustness of their identification systems: the UPC/EAN product numbers, the ISBN book identifier and the CAE composer/author/publisher identifier are among the most successful identification systems in use in the world of content management; they form

[11] Caplan (1995).
[12] Kelly (1997).

the backbone of highly effective distribution systems in their respective industries.

In contrast, where unique identifiers for major entities do not exist or are poorly implemented within a domain, data management costs are higher - and simple, effective management systems difficult to develop. The absence of unique "party" identifiers for creators and publishers in the major content industries, the scarcely visible implementation of the ISRC for sound recordings, and the lack of a standard agreement or licence identifier in any copyright community, are each examples of gaps that are crippling for interoperability within a domain, let alone between traditional domains. Some of these gaps are now being filled: e.g. the InterParty project [13] is providing one way of approaching party identification, by investigating a framework to make existing party identifiers interoperable.

Multi-media, multi-lingual, multi-national, multi-purpose metadata also requires that unique identification applies at all levels, including the use of "controlled vocabularies" for values of properties such as measures, form and type. In truly well-formed metadata, the only "free text" properties of an entity are found in its names or titles; in some instances (for example, in trademarks and in the UK Actors registry Equity), even names may be protected to ensure their uniqueness in a given domain.

For wider interoperability, the most important properties of an identifier are uniqueness within a given domain; stability (identifiers should never be transferred to another entity); security, whether through protection by watermarking or encryption, and/or by internal consistency through the use of check digit algorithms; and the public availability of some basic descriptive metadata for the entity identified, without which the identifier has only limited use.

Identifiers as Numbering Schemes, Specifications, and Identifier Systems

We need to make an important terminology distinction at this point about the use of the word "identifier". As the use of numbering in digital networks has developed, the historical use of the word in this context has become expanded to the point where it is now used synonymously to cover several different things, all of which are useful but which actually carry different implications that need to be separated in a detailed understanding of practical DRM applications. It's important to understand the differences here; and to note that these are not mutually exclusive (one particular "identifier" may fit into one or all of these categories).

[13] The InterParty Project Web Site: http://www.interparty.org.

Identifiers as "Labels": The Output of Numbering Schemes

A numbering scheme is a formal standard, an industry convention, or an arbitrary internal system such as a an incremented production serial number etc., to arrive at a consistent syntax for denoting and distinguishing separate members of a class of entities [14]. The scheme is a specification for generating a number: this resulting "number" may include alphanumeric characters, but the accepted parlance is to speak of these as numbers (e.g. ISBN = International Standard Book Number). The intent is of establishing a one-to-one correspondence between the members of a set of labels (numbers), and the members of the set counted and labelled. The product of the process is enumeration, a cardinality judgement, and assigned numbers for each cardinal member. An example would be the ISBN, where a separate ISBN is assigned to each book edition. The numbering scheme may or may not be accompanied by some apparatus - for example, a registration agency and maintenance agency for the ISO TC 46 series of identifiers.

The important point here is that the resulting number is simply a label string (a "noun"). It does not of itself create a string that is actionable in a digital or physical environment (a "verb") without further steps being taken. It may be used (and probably will be used) in databases; or it may be incorporated into another mechanism later.

The most common standard numbering schemes of interest in DRM include those standardised by ISO [15]:

- ISBN: ISO 2108:1992 International Standard Book Numbering (ISBN) [16]
- ISSN: ISO 3297:1998 International Standard Serial Number (ISSN) [17]
- ISRC: ISO 3901:2001 International Standard Recording Code (ISRC) [18]
- ISRN: ISO 10444:1997 International Standard Technical Report Number (ISRN) [19]
- ISMN: ISO 10957:1993 International Standard Music Number (ISMN) [20]
- ISWC: ISO 15707:2001 International Standard Musical Work Code (ISWC) [21]

[14] Ehlers (1994).
[15] ISO TC49/SC9 - Information and Documentation - Identification and Description Standardization of information identifiers, description and associated metadata and models for use in information organizations (including libraries, museums and archives) and the content industries (including publishing and other content producers and providers).
[16] ISO 2108:1992.
[17] ISO 3297:1998.
[18] ISO 3901:2001.
[19] ISO 10444:1997.
[20] ISO 10957:1993.
[21] ISO 15707:2001.

- ISAN: Draft ISO 15706: International Standard Audiovisual Number (ISAN) [22]
- V-ISAN: Draft ISO 20925: Version Identifier for audiovisual works (V-ISAN) [23]
- ISTC: Draft ISO 21047: International Standard Text Code (ISTC) [24]

Whilst these ISO TC46 identifiers were originally simple numbering schemes, of late they have also begun to adopt the notion of associating some minimal structured descriptive metadata with the identifier. Also relevant are the ISO-affiliated NISO standards including:

ANSI/NISO Z39.84 The Digital Object Identifier [25]

Identifiers as "Infrastructure Specifications": Making Labels Actionable

"Identifier" is also sometimes used to mean a mechanism or syntax by which any label (as defined above) can be expressed in a form suitable for use with a specific infrastructure tool. This is sometimes known as creating an "actionable identifier" - meaning that in the context of that particular piece of infrastructure, the label can now be used to perform some action: e.g. in an internet Web browser, it can be "clicked on" and some action takes place.

Of particular relevance for DRM, the set of internet specifications known as Uniform Resource Identifiers (embracing URLs and URNs) provide mechanisms for taking labels and specifying them as actionable within the internet. These are discussed in more detail later in this paper - here we simply note the functionality that such systems are intended to provide. The same principles can apply in the physical as well as internet environment - for example by prefixing an ISBN with the EAN sequence 978 or 979, the ISBN becomes a UPC/EAN identifier expressible as a physical bar code symbol, or a radio-frequency tag, for use in the physical supply chain [26].

Importantly, note here that such "identifiers" do not mandate a way of creating labels, they merely accept any labels: hence if one does not have an existing numbering scheme, it will be necessary to adopt or create one in order to form URIs. A URI specification merely ensures that a label follows the rules to become actionable in an Internet environment: a specification is not an implementation, with all the other aspects that a fully functioning identifier system (see below) may require: URI may for example specify the syntax, and specify a recording registration procedure, but not create a managed environment (e.g. by which

[22] ISO 15706.
[23] ISO 20925.
[24] ISO 21047.
[25] ANSI/NISO (2000).
[26] Osborne (2002).

registrations are "policed"), or carry any specifications of metadata or policy (which I consider to be the hallmark of a full *identifier system*). Some identifier specifications of this form may have limited rules or requirements for implementation: so far this is limited to the URN specification including a proposed (not implemented) mechanism for resolution. The acid test one should ask of such a specification is: *what does specifying my label in this particular form get me, in practical terms, in a specific infrastructure?*

Identifiers as "Implemented Systems": Implementing Labels in an Infrastructure Environment

The UPC/EAN is an "identifier system" in the physical supply chain; a DOI is an "identifier system" in the digital supply chain. ISBNs for example become implemented in the physical supply chain through UPC/EAN bar codes or RfiD tags. This sense of "Identifier" denotes a fully implemented identification mechanism that includes the ability to incorporate labels, conforms to an infrastructure specification, and adds to these practical tools for implementation such as registration processes, structured interoperable metadata, and a policy/governance mechanism. Such a system is necessary for practical DRM applications; since DRM deals with digital entities, structured metadata will be an essential component of such a system. The DOI is one of the better developed, with several million DOIs currently in use by several hundred organisations.

Both ISO TC 46 and URN have published suggested lists of requirements for their identifiers - the first covering what I have called here "labels", the second what I have called "infrastructure specifications". I have summarised these elsewhere [27] and suggested that a practical *identifier system* (which builds on both concepts) for digital use (DRM) should assume a combination:

- Unique "dumb" identification: unambiguous simple identification (label assignment) of a defined piece of information; opaque strings, not hard-wired with any specific application intelligence;
- Well-formed metadata: defined namespaces and controlled values within those namespaces for each value of a metadata element, defined by inherent structure not by their function in a particular application. A means of expressing an ontology to facilitate interoperability in many different functional applications;
- Support for arbitrary levels of granularity;
- Multiple, co-existing, labeling schemes should be possible, including support of existing (legacy) schemes; groups of content owners with common interests should be able to devise their own schemes which should then be interoperable in an open framework; multiple (overlapping) identification of content must be allowable. This implies extensibility: the ability to add within a scheme a particular namespace that defines that element.

[27] Paskin (1999).

- Links to distributed metadata: dumb identifiers pointing to specific repositories for different pieces of data, relating to different functions e.g. copyright, trading, EDI; details of medium, version, format etc. conveyed as metadata;
- Distributed (cascading) administration responsibility: once below a certain level, no central agency permission needed to assign unique numbers (sublevels assigned by the owner of the higher level);
- Policy and governance process: a management structure design for the practical operation of the identifier registration and maintenance processes.

The three uses of the word "identifier" (*label, infrastructure specification, and implementation*) can become easily confused, since one particular string can be in more than one category. But to see why we need to be precise, consider the following statement:

> "*For use on the Internet, an ISBN label can become a URN specification; an ISBN label can be incorporated into a DOI, which is an implemented identifier system following the URI specification.*"

Replacing the more precise terms in this statement by the loose unqualified synonym "identifier" results in confusion:

> "*an ISBN identifier can become a URN identifier; an ISBN identifier can be incorporated into a Digital Object identifier, which is an implemented URI identifier*"

(true, but only on close textual analysis!).

Social Infrastructure and Costs

Creating an implemented identifier system for DRM is not a trivial task: it necessarily incurs some costs, in three principle areas:

- "label" registration; maintenance of resolution destination(s); declaration of metadata; validation of number syntax and of metadata; liaison with the registry; customer guidance and outreach; marketing; administration
- Infrastructure: resolution service maintenance, metadata registry maintenance, and further development
- Governance: common "rules of the road"; business model for cost recovery, development of the system

There is a widespread recognition of the advantages of assigning identifiers (labels); and of making these actionable; and a widespread misconception that an abstract infrastructure specification (like a URN or URI) actually delivers a working system rather than a namespace that still needs to be populated and managed. A common misperception is that one can have such a system at no cost. It is inescapable that a cost is associated with managing persistence and assigning identifiers and data to the standards needed to ensure long-term stability for DRM. This is because of the need for human intervention and support of an

infrastructure. Assigning a library catalogue record, for example, will typically cost anything up to $25. Assigning an ISBN or ISSN or National Bibliography Numbers will also have costs, even if these are not paid directly by the assigner. The most widespread model of recovering costs is from the assigner community: the DOI as an example is free at the point of use, but there is a small fee to an assigner for creating a DOI (a few cents) because the model chosen is that of a self-funding system (on the model of the UPC/EAN system).

Understanding identifiers in the digital world is fraught with such misunderstandings: "adding a URL costs nothing" (which itself ignores some infrastructure costs), "so why should assigning a name have a cost?" It is indeed possible to use any string, assigned by anyone, as a name; but to be useful and reliable any name must be supported by a social as well as technical infrastructure that defines its properties and utilities. URLs for example have a clear technical infrastructure (standards for how they are made), but a very loose social infrastructure: anyone can create them, with the result that they are unreliable alone for long term stable use as they have no guarantee of persistence let alone associated structured metadata. UPC/EAN product codes, Visa numbers, and DOIs have a tighter social (business) infrastructure, with rules and regulations, costs of maintaining and policing data – and corresponding benefits of quality and reliability. When a credit card is presented, we can be reasonably certain that the number is valid, and has been issued only after careful correlation with associated metadata by the registrant. It does not necessarily imply a centralised system: it may be a distributed system (like domain names), but it must have some form of regulation.

Such regulation of infrastructure for a community benefits all its members; funding the development of it is often a problem, and there is no "one size fits all" solution to how this should be done. But finding a workable model for the development of an infrastructure can yield obvious benefits. There are many modern examples (3G telephone networks, railways) which are struggling with the right model for supporting a common infrastructure. The Internet was largely a creation of central (US) government; the product bar code, a creation of a commercial consortium. Product codes, Visa numbers, and DOI for example use the concept of Registration Agencies, rather than relying on centralised subsidy. These Agencies effectively hold a "franchise": in exchange for a fee to the governing body, and a commitment to follow the ground rules of the system, they are free to build their own offerings to a particular community, adding value services on top of identifier registration and charging fees for participation.

Identifiers may of course be made available at "no charge", if the costs of doing so can be met from elsewhere (there is no such thing as "free", only "alternatively funded"). Like any other piece of infrastructure, an identifier system that adds value (like metadata and resolution) must be paid for eventually by someone. An organization could, if it wished, assign identifiers freely (registration fee zero

to registrants) and subsidize this added-value service by paying a franchise fee to the governing body from a central fund, as an acceptable cost for supporting the service.

Namespaces as a Way of Managing Identifiers

The development of domains or namespaces within the Internet has helped in the relaxation of pressure on the need for absolute uniqueness in the structure an identifier: URIs provide specifications for universal disambiguation that allow even common terms to assume unique, network-wide, status.

A namespace is a set of names in which all names are unique. While one is working within one namespace, uniqueness is by definition not a problem. A potential problem arises when two namespaces containing the same label (but for different entities) are made interoperable. This is the issue faced by e.g. merging of databases. Namespaces allow reference to each label in the form nid:nss (namespace identifier: namespace specific string), so that the full string includes both an identifier of the namespace and the specific string within that namespace. This is the solution adopted within URNs and by XML, which has popularised the concept over the past few years. XML namespaces provide a simple method for qualifying element and attribute names used in Extensible Mark-up Language documents by associating them with namespaces identified by URI references [28]. The XML namespaces recommendation works, but a number of underlying issues (e.g. validation) remain unclear [29]. Nevertheless XML is the de facto standard way of communicating data and highly advisable for any identifier/metadata scheme to make its elements available in this form.

However, we are far from having all DRM transactions automated, and although this is a logical solution if every transaction was fully and precisely specified, in practice if a particular community is working in one namespace, or using less formal methods, it will usually assume "nid" to be implicit - which brings problems when two namespaces need to be considered. A practical example is the author identified as "Joan Brady" - in fact, a different person in the "UK author namespace" (a Whitbread Prize novelist) and in the "US author namespace" (author of "God on a Harley"): in effect, these undeclared namespaces collide on an Amazon.com search, resulting in confusion and ultimately threats of litigation [30].

There is no fundamental logical difference between a "name" and "an address" - an address is the name of a location, i.e. a name in a namespace consisting of addresses (e.g. the URL namespace). But this does not mean that addresses can always be used as useful names: in DRM, a requirement is to manage entities (resources, parties etc) as "first class objects" - that is, named entities in their

[28] W3C (1999).
[29] Bourret (2000).
[30] Bide (1999).

own right - not via a property (location) which may vary independently of the entity.

Abstractions

In most cases when an intellectual property entity is identified, the entity being identified is not tangible, but an abstraction. Clearly this is the case when identifying abstractions such as the underlying work "Robinson Crusoe" which has many different manifestations as book editions, or " Eroica symphony" in many recordings, scores, and performance. Not as readily appreciated is that apparently "tangible" entities are also abstractions: e.g. the ISBN identifies not the copy of a book which you have in your hand, but the class of all such copies, an abstraction.

Abstractions need an ontology to make sense of them. More than one ontology can provide tools for dealing with any set of entities, but we need to be careful not to mix definitions from different ontologies without careful mapping: every schema has its own inherent contextual model and its elements are defined in those terms. For example, there is a fundamental difference in the way in which the library-derived FRBR model [31] defines the term "expression" and the way <indecs> defines "expression", but this is not to say that only one is right: each recognizes the entity that the other is calling "an expression" and wishes that the other had called it "foo". Mapping elements is a completely different and much more complex process than declaring data elements. The indecs/DOI/ONIX group, for example, can map more or less any other schema successful within their models, but we would not assume that any other schema would adopt the same definitions of (say) agent, resource or event. It has been well said that "there are more abstractions than are ever conceived of".

Identity and Sameness

A fundamental purpose of identifiers is to define when two things are "the same" and hence denoted by the same identifier. The intuitive meaning of "the same" needs some logical analysis if it is to be applied consistently for automation. The word 'same' is used sometimes to indicate similarity (qualitative sameness), as in '*Alice is the same age as Bob, and the same height as last year*' , sometimes to indicate that what is named twice should be counted once (numerical sameness), as in '*the morning star and the evening star are the same planet*'. The word 'identical' can also have the former sense (identical twins, identical dresses) as well as the latter; hence philosophers are liable to discuss both kinds of sameness under the label 'identity'. Qualitative sameness is a comparison of metadata: entity A and B share a relationship to entity C. Numerical sameness is a simple logical relation through comparison of identifiers, in which each thing

[31] IFLA (1998).

stands only to itself. "Although everything is what it is and not anything else, philosophers try to formulate more precisely the criteria by means of which we may be sure that one and the same thing is cognised under two different descriptions or at two distinct times" [32].

Numerical sameness leads to a trap for the unwary: if we say, "Two entities are the same if they have the same identifier," we seem to create a puzzle: how can they be two if they are the same? If identity is a relation it must hold either between two distinct things or between a thing and itself. To say that A is the same as B, when A and B are distinct, is bound to be false; but to say that A is the same as A is to utter a tautology. Different solutions have been found by different philosophers for this "paradox of identity". This may seem like remote philosophising, but in fact lies at the heart of practical implementations.

In determining whether A is the same as B, we find that ultimately nothing is the same as something else; however, it makes sense to consider that A is the same as B *for a defined purpose* (i.e. in a defined context). To give a practical example, a photocopy of this article is not the same as the original in some ways (it is printed on different paper stock, it is located in a different part of space, etc.); but it might be considered the same - a copy - *for the purposes of intellectual property* (it retains the typographical layout and semantic sense). Here, the attribute "paper stock" is irrelevant, the attribute "manifestation of the defined work X" is relevant, for the purpose of DRM. Whilst this seems almost trivial in a physical environment, where the purpose and context are intuitively understood even if not stated, in a fully automated digital environment the attributes and context are less intuitive. This is why it is difficult to translate intuitive concepts from the physical world into the digital; e.g. arriving at a definition of "to copy" in the digital environment makes no sense without a context. In recent MPEG-21 discussions, some technologists argued that there can be "no such thing as a digital copy" - A and B must differ because of the sequence in which their data representations are laid down on a hard disk, for example. Yet it clearly is nonsense to say that "the action of copying is impossible in the digital domain": this would undermine copyright law as rampant copying is patently occurring in practice. Hard disk sequencing is an irrelevant attribute *for the purpose of* IP law - though case law in this area is sparse - and similarly, in more traditional IP interests, photocopier technologists are not ideal intellectual property lawyers.

So it is meaningless to ask "Are A & B the same thing?" and only meaningful to ask "Are A & B the same thing for the purpose of....". Technically we do this by considering which attributes of A need to be retained in creating the replica B; some attributes are ignored, considered irrelevant for some defined purpose. A description is a set of properties that apply to a certain object: two incomplete descriptions denote the same object if they have an identifying property

[32] Kemerling (2002).

in common [33]; the descriptions are for a purpose, and the "identifying property" (or more likely set of properties) is the one by which we define that common purpose or context of the A and B comparison.

When we make statements we normally leave a great many attributes unstated because we assume general or specific knowledge on the part of our audience. However when we come to fully automated DRM, which relies on exchange between computer systems, we cannot expect that any inferences from "common knowledge" will be applied. We need to consider an entity as no more than the sum of its stated attributes. I may say you can copy my CD and its entire contents and sell it in a jewel box: exactly what kind of jewel box, and what the printing on the CD and the inlay says is irrelevant to the copy. It is a replica if the stated attributes are the same at whatever level of granularity is explicit. It may even be a copy if it is not a CD, if the only stated attribute I have given is "this recording". DRM will rely on the same principle as any other computer system: computers are dumb, and if something is not specified it cannot be taken into account.

The same principle of considering a comparison *relevant for some purpose* applies to the use of metadata in automated applications: we must sort the metadata into sets (application profiles) which are relevant for the particular purpose of that application. As Karl Popper elucidated, there is no neutral purpose-free "tabula rasa", always a purpose which is inherent in a particular act of perception [34]. The recognition that all considerations of identity require recognition of context is fundamental to the context model underlying the indecs Data Dictionary (which will be discussed later in this paper), in which all are things are ultimately part of events or situations, taking place in defined contexts.

Granularity

The paradox of identity is related to the concept of recognising granularity. Recognising sameness among a population, as we have seen, depends on choosing which particular set of attributes of a number of entities we consider relevant, and which are irrelevant, and ordering the population into sets defined by the relevant attributes for the purpose in hand.

Granularity refers to the level of content detail identified; and to this we must add again the qualifier "identified for a particular purpose". To take an example from text publishing, the ISBN [35] identifies the whole book; the BICI [36] identifies component parts of a book (e.g. chapters, sections, illustrations, tables). This may be enough for some uses but is clearly inadequate for others. If we

[33] Guarino, Welty (2000).
[34] Popper (1972).
[35] ISO 2108:1992.
[36] NISO (2000).

are to be able to identify all rights owners in a particular piece of content, that may require a far finer degree of granularity of identification, to the level of the individual illustration or quotation from another source. Similarly, if information is to be traded with customers at a level of granularity finer than the "chapter" or the "article", then publishers may have compelling marketing reasons for being able properly to identify and to keep track of what is being traded.

The level of granularity that may need to be identified becomes effectively arbitrary in a digital environment. This might suggest a requirement for relational identification where (like the BICI) smaller fragments are identified by reference to the larger "whole" from which they come, although this "intelligence" would have some drawbacks, not least in terms of the size and structure of the codes and a preferable route would be to express the relationship through readily accessible metadata. Considerations of granularity are fundamental to a logical analysis of DRM, and a key point is the purpose and context of the granularity choice.

Functional Granularity

The indecs *Principle of Functional Granularity* is that "it should be possible to identify an entity whenever it needs to be distinguished." When should an identifier be issued? In this deceptively simple question lies the most basic question of metadata: for which data is it meta-? Resources can be viewed in an infinite number of complex ways. Taking the indecs metadata framework document as an example, it has an identifier in the <indecs> domain: WP1a-006-2.0. But to what does this refer? Does it refer to the original Word document, or to a pdf version available on the Website? Or does it refer to the underlying "abstract" content irrespective of delivery format? If it refers to the Web document, is this also adequate as a reference to local copies that have been downloaded onto other computers or servers? The document's parts may require identification at any level (for example, section 2.2, or Diagram 14). If you wish to make a precise reference to a sentence from another document, you will need a more precise locator, and its nature will depend on whether your reference is intended to allow automated linking. As the document has been through many stages of preparation, how many different versions need to be separately recorded? Each of these requires the exercise of functional granularity: the provision of a way (or ways) of identifying parts and versions whenever the practical need arises.

The application of functional granularity depends on a huge range of factors, including the type of resource, its location in time and place, its precise composition and condition, the uses to which it is or may be put, its volatility, its process of creation, and the identity of the party identifying it. The implication of this is that a resource may have any number of identifiers. The same entity may be subjected to functional granularity across a range of views. The basic "elements" of a resource may be entirely different according to your purpose. Stuff may be analysed, for example, in terms of molecular entities (chemistry),

particles such as electrons, quarks or superstrings (physics), spatial co-ordinates (geography), biological functions (biology, medicine), genres of expression (creations), price categories (commerce), and so on. In the digital environment, stuff can be relatively easily managed at extreme levels of granularity as minute as a single bit. Each of these process will apply identifiers of different types at different levels of (functional) granularity in different "dimensions"; these may need to be reconciled to one another at a point of higher granularity.

Functional granularity does not propose that every possible part and version is identified: only that the means exists to identify any possible part or version when the occasion arises. Identification is not the same as mark-up, though if a section is distinguishable by some mark-up coding it will be subsequently easier to specify it as separately identified.

Conflicting Views of Granularity: Difference within Sameness

What is "the same thing" for one user, purpose, or context will be "two different things" for another. The two users may have different purposes in mind when they ask "are X and Y the same?"; and as we have seen, this question is implicitly "are X and Y the same for the purpose of...?" Failure to comprehend these different views (purposes) across a supply chain results in considerable friction. Some practical examples will illustrate this. For clarity, I refer in each case to two different users - the party who sees "the same thing" as X and the party who sees "two different things" as Y.

There has been much discussion (as yet not fully resolved) of this in the context of eBooks [37]: publisher X wishes to use one identifier (the ISBN) to refer to all technical formats of an eBook, since they are all "the same book"; yet supplier Y needs to distinguish different formats (a customer ordering one format wants that and no other). Some publishers have in fact suggested using the ISBN with some form of qualifier (or parameter) to do this; the International ISBN agency prefers to recommend different ISBNs for each format [38]. These are the two general approaches to recognising difference within sameness, each of which may be valid in some circumstances: a "*single identifier with qualifier*" or "*create new multiple fixed identifiers*".

The "single identifier with qualifier" approach is used in solving the "appropriate copy" problem in one application with DOIs [39]. The generalised case is that since an identifier is normally that of a class (an abstraction), it is assumed that each member of the class is equivalent; but in reality this may not be so in all contexts, and there are many instances when more than one legitimate copy is available, and some copies are not available, due to the context of the request. In

[37] Anderson Consulting (2000).
[38] ISO (2002).
[39] Beit-Arie (2001).

the appropriate copy example, publisher X allocates one identifier to an article; library user Y finds that because of local loading, aggregator databases, paper copies or mirror copies, she needs to distinguish copy one from another; in each of these cases, the address to which the identifier given by X should appropriately resolve depends on the location or affiliation (in general, the context) of the user Y who is making the resolution request. To solve this problem it makes sense to contextualise the use of the identifier by some tool such as OpenURL. A full analysis of any transaction, in the further work done using indecs for MPEG [40], shows that ultimately all transactions are contextual and can be expressed as an event or a situation; and a full analysis of the sue of identifiers will show that ultimately of course they are all used in some context.

The "create new multiple fixed identifiers" approach is shown in the emergence of the ISTC. New identifiers may be needed and require the creation of a new namespace if the namespace currently being used cannot satisfactorily include a new type of entity without disrupting the existing business. A good example is the identification of textual abstractions and the identification of their manifestations (books): ISBNs are in widespread use for identifying (separately) each different edition of e.g. Cervantes' Don Quixote. These are different (if customer Y orders the leather bound limited edition with illustrations by Dali, he is unlikely to be happy to receive the $1.50 Worlds Classics paperback edition). Yet authors agencies, rights organisations, and librarians X may all be interested in the general work and not concerned with specific editions for some purposes (a library reader wishing to find a copy of the work, for example). This led to the development (with the full collaboration of the ISBN agency) of a new identifier, the ISTC, which can be used to identify this entity (the textual abstraction) [41]. This example also usefully shows that it is not always the smaller granularity entities which the driver for the creation of new identifiers: in this case, a new identifier is required which may be related to "supersets" of ISBNs.

These two ways of dealing with "difference within sameness" are not always clear black-and white alternatives, and once again functional granularity will be the arbiter of which to use in which cases: is there a need to agree on a separate identification scheme (a new namespace), or can we live with the difference being defined by qualification after the identification step at a local level, which is not likely to be widely used across a supply chain? If the entities being finely differentiated are the object of commercial transactions across multiple partners, or are likely to be stored and used in communication to identify precisely the differentiated entity (rather than the unqualified entity), then I believe the separate new identifiers approach is likely to be optimal in the long term.

In each solution, the same logic applies: whether we refer to them as "a qualified identifier with two different qualifiers" or "two identifiers which have a relation"

[40] ISO/IEC 21.
[41] ISTC.

is semantics: “ISBN 1234” and “ISBN1234-as qualified-Z” are separate strings. They denote different entities, they must do otherwise there wouldn't be a need for two strings. It may well be that party X only needs the first, but if party Y has a need to deal with all these different transformations generated by X at a business level and needs to know the various sub “qualified” identifiers, then Y is going to end up having to store the [qualified] identifiers and treat them as static separate strings, i.e. separate identifiers - probably in a separate database because the particular numbering system X has used isn't sufficiently granular for Y's needs.

If entities need at some point to be differentiated for long-term purposes (which typically they do in any DRM chain for e.g. audit etc), then inescapably someone somewhere will be managing multiple identifiers [strings] with multiple metadata [as there are multiple entities] that have a defined relationship. This need not be a concern if that management is in an isolated internal database, but increasingly such data is becoming exposed to interoperability, the heart of DRM. Wherever this happens, this is easier to do by treating all differentiable entities as having fixed identifiers - persistent opaque strings with associated data - rather than some as derived by qualification. This allows a common mechanism for persistence, registration, and interoperability. There are many related identifier labels (namespaces) and no one can deal with all possible needs - this is why ISTC had to be added on top of ISBN, rather than overloading one system and asking it do two fundamentally opposing jobs; an identifier system or framework which can contain all these, such as DOI, is making more and more sense.

Intelligence in Identifiers

A dumb identifier is an opaque identifier string that serves solely as unique label and has no other inherent or implied meaning (synonyms: simple or insignificant identifier). An example is a manufacturing sequence number; a consortium of manufacturers may use this as an interoperable identifier by preceding each string with some means to guarantee uniqueness across originators. In text publishing, an early example was the PII (Publisher Item Identifier) [PII], simply a sequence number from an individual publisher (and incidentally a precursor of the ISTC; most PIIs are now used in the form of DOIs through the CrossRef implementation [42]).

An intelligent identifier is a string that has at least some segment capable of ready interpretation outside the identifier scheme to derive meaningful information (synonyms: compound or significant identifier). Intelligent identifiers which carry some information in their structure relating to the entity they identify, such as a format, date or producer code, are of some value in particular circumstances, but problems of ambiguity or volatility often render much of this

[42] CrossRef - Web Site: http://www.crossref.org.

apparent "intelligence" unreliable. A manufacturing sequence number that explicitly included as its opening string the year of manufacture would contain such intelligence. The SICI (Serial Item and Contribution Identifier) [43] contains substrings denoting elements such as date of publication, page number, etc. Intelligence is the insertion into the name syntax for one namespace of a string which has applicability in another namespace: it therefore creates a hard-wired link between the two entities in the two namespaces: i.e., metadata. Hard wiring is appropriate only if the relationship will never need to change, which is not always easy to guarantee (as the year 2000 problem amply demonstrated).

"Affordance" is the ability to enable construction of a unique identifier from examination of the physical manifestation (or some metadata record of it), rather than by reference to a central database of identifiers [44]. Affordance is therefore a counterpoint to the concept of intelligence: intelligence implies ability to derive, some element of metadata about the object, from the identifier; affordance implies the ability to derive the identifier from the object or metadata. Another term for this is computability: given the object instance, the identifier for a namespace may be computed. The SICI scheme allows a SICI code to be created by algorithm from known citations; while this could be done manually, it can be automated by algorithms [45]. This enables a user to retrieve citation records from various databases, and subsequently create the SICI code that could then be used to search more efficiently across multiple text databases to find the actual article. Given the variation and performance of search capabilities across multiple systems, an algorithmic key is more likely to find the document than a reformatted version of the initial query or bibliographic citation textual elements. For the SICI or other such access keys to be highly successful, more standardization of bibliographic citation data elements is needed; however, it seems to hold promise for locating a bibliographically denoted work from numerous different online resources and legacy systems.

Aids to Identifier Use: Readability and Check Digits

Readability refers to the design of identifier syntax in such a way as to aid interpretation by human inspection in an application. The design of the Internet domain name system is a clear example where simple IP addresses (numerical values) are associated with more readable or memorable strings (such as www.ibm.com); the price to be paid for this is literal, in that certain memorable or readable strings become much more valuable than others in a commercial context, although the underlying numbers appear to be of identical value. Readability can be assisted even in numeric, dumb, schemes: an example is the Publisher Item Identifier (PII) which consists of seventeen alphanumeric characters in a single string (e.g. S1384107697000225); for readability when the PII is printed

[43] NISO (1996).
[44] Green, Bide.
[45] Paskin (1999).

slashes, space and parentheses are added where necessary, to ease the reading of the code and divide it into segments each with a defined origin though not meaning (e.g.S1384-1076 (97) 00022-5). These additional elements are stripped out for machine readable use and/or reinstated on printing and do not form part of a machine-readable string or check-digit algorithm. Readability is important if an identifier will be entered by keyboard rather than automatically. Readability is not necessarily synonymous with intelligence (the DNS example uses intelligence, the PII example does not), though where an intelligent number is used readability will be enhanced by visually parsing into the component intelligent elements. Readability may also help in some limited cases of error correction (e.g. recognising that a string 3002 representing a year should really be 2002).

Identifier labels may contain a check digit: usually the last in the sequence within an identifier string, algorithmically derived from the preceding digits, rather than being part of the identifier itself. The aim is to ensure that if one digit is incorrectly transcribed, the check digit will change as an alerting mechanism, and that if two digits are incorrectly transcribed, the chance of their combined effect on the check digit cancelling each other out is minimised. Recalculation of the check digit from the body of the number, followed by comparison with the stated check digit, can be performed algorithmically at key points in processing. Note that this provides error detection, but not error correction. In a typical check digit algorithm, each digit is assigned a different weighting factor (ideally a prime number). Digits and their corresponding factors are individually multiplied and summed, the resulting sum divided by a prime modulus number, leaving a remainder being the check digit; using prime numbers minimises the chances of internal cancellation. Check digits occur in for example ISBN and ISSN numbers and in other contexts, e.g. bank account numbers; ISO has a recommended standard for check digits [46]. Check digits are typically of importance in an entry step (where identifiers have to be manually transcribed as input) and less important in a transmission step where error correction protocols such as packets (TCP/IP) are already in place, although their original introduction was to ensure consistency in both types of activity.

Internet systems have error correction in the transmission protocol, but not on entry: URLs (URIs) do not contain check digits. This may lead to the assumption that check digits are of less importance, in an Internet-enabled world, than had been assumed in earlier automation phases. Whether or not this is true depends to some extent on the consequences of an error slipping through: whether inputting an incorrect identifier generates an error message, or simply locates the wrong object. A message may be transmitted correctly, but contain incorrect initial input: e.g. omitting check digits in bank account numbers would not provide adequate error protection for most users.

[46] International Standard Data processing – Check character systems - ISO 7064:1983.

Resolution

Resolution is key to creating actionable identifiers from simple labels in a digital network, through implemented schemes. Resolution is a 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 related to the identified entity: e.g. a location (such as URL) where the object can be found. The technology supporting this capability is a *resolver*. In the case of the Domain Name System (DNS), as an example, the resolution is from domain name, e.g., www.doi.org, to a single IP address, e.g., 132.151.1.146, which is then used to communicate with that Internet host. In the Handle System [47] [48] [49] [50] [51], a well-designed and scalable resolution system designed by one of the originators of TCP/IP, the resolution is from a "Handle" to one or more pieces of typed data: e.g. URLs representing instances of the object, or services, or one or more items of metadata. Resolution can be considered as a mechanism for declaring a relationship between two data entities; an item of metadata is a relationship that someone claims exists between two entities: therefore, metadata relationships between entities may be articulated and automated by resolution.

In computer science terms, resolution is "adding a level of indirection" (sometimes called redirection): manipulating data via its address. Indirection is a powerful and general programming technique of processing data by maintaining a pointer to the current item and incrementing it to point to the next item, such as a new value. Providing that the performance issues of adding this extra communication step can be overcome, indirection is a very useful way of separating one into a relationship of two entities, which may then be separately managed - e.g. a name and a location. This then provides a mechanism for managing persistence of the name even if the location varies.

The concept of the URN (Uniform Resource Name) was introduced into the Internet to allow indirection, such as "N2L" (URN to URL) resolution. One of the earliest applications for DRM was the DOI for simple, single point resolution. Each DOI has at minimum a single URL to which it will resolve. This allows the location of an entity to be changed while maintaining the name of the entity as an actionable identifier. DOI is not alone in providing a solution to this problem. Other applications, for example PURLs (Persistent URLs), can provide this simple level of resolution. It has been argued - though increasingly this is a lost cause - that URLs can (in theory) themselves be used as a persistent identifier - that their use as a transient identifier is a social, not a technological, problem. However, this lack of persistence of the URL is only the first of many challenges that the DOI System was designed to manage.

[47] The Handle System - http://www.handle.net/.
[48] Handle RFCs - http://www.handle.net/documentation.html.
[49] Sun, Lannom (2002).
[50] Sun, Reilly, Lannom (2002).
[51] Sun, Reilly, Lannom, Petrone (2002)

Multiple Resolution

An identifier is a name for an entity; in the network environment, there may be many identical copies ("instances") of the same piece of content. A single identifier may be used to manage the existence of multiple "instances", or multiple metadata relationships, or multiple services, if the resolution step can offer linkage not simply from one identifier to a single piece of data (e.g. a URL), but to multiple data. The Handle System is such a multiple resolution technology (a URI and in conformance with URN, as discussed below). The need for multiple resolution if one is to construct any complexity is obvious if one envisages the resolution process as a set of connections between points in a logical space: univalent linkage (single resolution) offers very limited construction possibilities (simple chains); polyvalent linkage (multiple resolution) offers unlimited branching constructions.

The Handle System is used in e.g. the DOI, the D-Space project [52] and other systems [53]. Uniquely, by using the Handle System in combination with the indecs approach to metadata, the DOI system provides a full framework for identifiers to be articulated by means of resolution and interoperable metadata. The DOI System is also designed to manage much more complex DRM-related services than resolving to multiple instances of the same piece of content, such as accessing metadata about the entity that the DOI identifies. At its simplest, the user may be provided with a list from which to make a manual choice. However, manual choices are not a scalable solution for an increasingly complex and automated environment. The DOI will increasingly depend on automation of "service requests", through which users (and, more importantly, users' application software) can be passed seamlessly from a DOI to the specific service that they require.

Persistence

Critically for DRM, even if ownership of the entity or the rights in the entity change, the identification of that entity should not change. The responsibility for managing the identifier may changes, but not the identifier itself.

The lack of persistence in identification of entities on the Internet is a commonplace. Even the most inexperienced of users of the World Wide Web rapidly becomes familiar with the "Error 404" message that means that a specified Web address cannot be found - the URL for that web page cannot be resolved. Resolution offers a mechanism to assist, by assigning names rather than locations. But persistence is ultimately guaranteed by social infrastructure (policy); persistence is fundamentally due to people, and technology can assist but not guarantee.

[52] DSpace Web Site: http://www.dspace.org.
[53] Applications of the Handle System: http://www.handle.net/apps.html.

A URI should persistently identify a resource. A DOI (a URI with specific application in intellectual property plus added features) identifies a specific intellectual property entity, which may or may not be an Internet-accessible file, and ensures persistence through policy; a URL identifies a specific address on the Internet. These applications of identification are completely different. One identifies an entity; the other identifies a location (where a specific entity may or may not be found). The analogy is with the ISBN (which identifies the book) and the shelf-mark (which identifies the place where the book is to be found). When the location changes, the shelf mark changes - but the ISBN does not.

Identifiers must persist in the face of legitimate change. There are legitimate, desirable, and unavoidable reasons for changing organisation names, domains etc. One aim of naming entities/resources is to avoid tying an entity name to a domain name, or any other piece of variable metadata (a problem encountered in recent domain names/trademarks disputes). The entity can be persistently named as a first class object irrespective of its location, owner, licensee, etc. Distinguishing names from locations is essential for E-commerce. It is trivially true that "all names are locations" (in a namespace), but practically, most people worry about spaces like URLs, and that's the wrong level. Naming entities as first class objects, rather than locations, enables better management of multiple instances of an object, for example.

Persistence is something we are familiar with in the physical world: ISBNs for out of print books can still be useful. Persistent identification alone is a good enough reason to adopt identifiers such as DOI which provide a means by which potential customers can find your digital offering even if a "broken link" URL of a retailer or other intermediary intervenes.

Technology can help with persistence. For example using DOIs, only one central record, which is under the control of the assigner, needs to be changed in order to ensure that all existing DOIs which are "out there" in other documents can still resolve correctly: a redirection resolution step enables management in the redirection directory, thereby ensuring that one change can be picked up by many users, even if they are unaware of the change. But to manage the data in the directory takes effort, time, incentive, etc. – either you do that locally (using tools such as PURL, managing a service yourself) or as a global service (the DOI being such a service for intellectual property entities). In the case of DOI management of data is a service role (and hence also business activity) for registration agencies, an approach used in other activities like bar codes and ISBNs. People aren't free, so there's a cost to this, and just like the physical bar code system, the DOI aims to be a self-funding operation. DOIs won't be appropriate for many things, and some people won't feel this people cost merits the reward, but DOIs (or any other system which offer similar functionality) are a viable solution for content management of intellectual property on a large scale.

DOI is an implementation of URN (Uniform Resource Names) and URI (Universal Resource Identifier) concepts, and can be formalized within these frameworks. The aim of each is to allow persistence of naming irrespective of other characteristics.
In addition to persistence of the identifier, a fully operational service such as DOI has to consider also persistence of the resolution technology, persistence of the identified object (archiving and preservation); and stability and invariance of the associated metadata. These topics are beyond the scope of the present article and interested readers are referred to other discussions [54].

Internet Specifications for Identifiers

Ideally, to ensure efficient use across many DRM applications we should follow the *principle of application independence*: metadata structures should be independent of any specific technical expression. Identifier and metadata systems whose development is shaped by technical rather than semantic constraints will be less than optimal, but technological differences must be resolved at the point of interoperability, since they cannot be wholly anticipated at source; so we cannot always follow this principle in full. Internet usage of identifiers is of particular significance in DRM.

Uniform Resource Identification Specifications

URN (Uniform Resource Name) and URI (Uniform Resource Identifier) are specification schemes for persistent identifiers of resources in the Internet. Existing identifiers such as ISBN, ISMN, DOI etc may be registered as URI and URN schemes, to enable implementations to make use of the technical specification. URIs and URNs should therefore be considered as a "framework" for enabling identifiers to work in an internet environment, rather than as a competing system of identification to existing schemes such as ISO identifiers (as explained above, ISBNs are labels, and URI/URN are specifications for using those labels in a digital context.)

In order to make use of such specifications, an implementation mechanism must be put into place. It is important to distinguish two issues:

- The Internet specifications of "what is" a URN and a URI: these differ slightly from each other (see below);
- What this means for practical implementation: irrespective of internet specifications, to make use of persistent identification schemes in useful ways will usually require more than a simple technical implementation. Especially, policy and governance issues (such as scope, authority to issue), and control of assigned metadata (quality control, interoperability considerations, etc) will be important components in adding value in practical implementations (an "implemented identifier system" as described above).

[54] DOI 7.

Definitions and of the URN and URI concept are spread across a number of documents; the specifications are also continuing to evolve. "Naming and Addressing: URIs, URLs, etc" [55] provides an overview of W3C (World Wide Web consortium) materials related to Addressing. Recently (November 2002) the W3C has proposed a further "URI Activity" [56] to deal with remaining issues of URI and URN definition, documentation, and reconciliation. The URN concept was originally driven by the IETF; the URI concept by the W3C.

URI, Uniform Resource Identifier, is defined as "the generic set of all names/addresses that are short strings that refer to resources". In some publications from W3C, URI is also defined as "Universal Resource Identifier". A URI may be a pure name or de-referenced by any service; in the latter case, the namespace provides its own mechanism ("bootstrapping"). On its own, any URI specification is just a specification: it requires code distribution for any implementation. URI schemes are only intended to "address information spaces that are globally useful" [57]. URIs are not intended to rely on any additional network services. A software client either knows what to do with, e.g., ftp, or it does not: this is the key difference with the URN specification.

URN, Uniform Resource Name, is defined according to W3C in two ways: (1) as "an URI that has an institutional commitment to persistence, availability, etc.; (2) as "a particular scheme, urn:, specified by RFC2141 and related documents, intended to serve as persistent, location-independent, resource identifiers." Thirteen RFCs specify URN syntax, services, namespace registration process and technical implementation of URN resolution in the present Internet [58]. URN architecture [59] assumes an additional network service that would allow a client to deal with a previously unknown URN type, e.g. *urn:isbn*. Specifically, a DNS-based middle layer (RDS) is used to find the specific service appropriate to the given URN scheme. URN resolutions are then delegated to that scheme-specific resolution service. The original RDS mechanism proposed was NAPTR (Name Authority Pointer); more recently a variant of this, DDDS (Dynamic Delegation Discovery System) has been proposed. These are proposed DNS extensions that would use DNS to provide a regular expression for the namespace, e.g., turn urn:isbn:1234567890123 into http:// isbn.org/1234567890123. These have not so far been widely used in a production sense: there are no practical implementations of large scale. There may be identifier strings being laid down as specifications (fifteen URN namespaces have already been registered, including several ISO identifiers such as ISSN and ISBN, and National Bibliography Numbers, NBNs), e.g., *urn:isbn:123456789*, but at this point there is no appar-

[55] W3C: "Naming and Addressing: URIs, URLs, etc". Available at http://www.w3.org/Addressing/#19991.
[56] W3C (2001).
[57] Palmer (2001).
[58] URI.net web site: http://www.uri.net/.
[59] IETF (1997).

ent advantage to that over the simpler *isbn:12345678*. In neither case is there a readily available well known global resolution service. Implementations (most are in libraries and are based on NBNs [60]) rely on local distribution of specific plug-ins and know-how.

The DOI System implements the URI/URN notions to enable identifiers to be global persistent and actionable object names, with the added aim of doing this in a coherent way across a wide range of media types and identifier schemes. Name resolution is currently by two separate methods to reference DOIs on the Internet: as URIs (doi:10.123/456) and as URLs (http://dx.doi.org/10.123/456). Each string can stand on its own, as a pure unique name, or it can be resolved using some network service. Resolution of the URI form would require software not yet commonly found on users' desktops (but which can readily be supplied by means of plug-ins such as for the Handle System [61]). Resolution of the URL form requires a proxy or gateway service out on the network. Existing identifier schemes may use DOIs or adopt their own individual resolution scheme: if these individual schemes are successfully and widely deployed the identifier would then be usable as a persistent name for that namespace alone.

Persistent URLs (purls)

A PURL is a Persistent Uniform Resource Locator [62]. Functionally, a PURL is a URL. However, instead of pointing directly to the location of an Internet resource, a PURL points to an intermediate resolution service. The PURL resolution service associates the PURL with the actual URL and returns that URL to the client. The client can then complete the URL transaction in the normal fashion. In Web parlance, this is a standard HTTP redirect. PURL was devised by OCLC's Office of Research after participating in the IETF URI work. There is nothing incompatible between PURLs and the ongoing URN (Uniform Resource Name) work; PURLs satisfy many of the requirements of URNs using currently deployed technologies and can be transitioned smoothly into a URN architecture once it is deployed.

PURLs are all http based. This is both their strength and their weakness. When you send a PURL to a PURL server, you are sending a special URL to a web server via http, and the web server will send back a perfectly typical web server answer - all http. The difference is that there is a special PURL server or module linked to that web server that inspects the URL, looks at a table to see what it means today, and returns that. It is one level of indirection, just like a single value DOI or Handle, but it is all contained within a single server and that single server is permanently attached to a specific domain name: PURL servers don't know about each other. In some ways it is no different from the way DOI uses

[60] IETF (2001).
[61] "Handle System plug in". Available at http://www.handle.net/resolver/index.html.
[62] Persistent Uniform Resource Locator Web Site - http://www.purl.org.

a Handle proxy, dx.doi.org, which re-interprets DOI Handle queries into http (if DOI were never going to go beyond the proxy server approach and never make use of the multiple resolutions and data types, PURL would be a comparable technological component to the DOI's chosen Handle protocol. There are ways in which one might imagine PURLs being developed to provide an approximation towards multiple resolutions and multiple data types. Content negotiation has always been in http, but like most W3C considerations is oriented at attributes of the document in hand. The more you push this, from document centric things like "give this to me in German" to more "attributes" like "tell me about rights", the more tenuous the approach would become.

As PURLs are http, they are designed to be used only in the web: this may not be an obvious problem at present, but the development of many mobile and other platform technologies means that not everything that happens on the internet from this point forward will necessarily be an extension of the www protocols; nor will DRM solutions which are based on web-only techniques prove satisfactory to the content industry (URN and URIs by contrast can be implemented with other protocols). PURLs have been widely available for several years but are not widely implemented in commercial settings and do not provide a sufficiently sophisticated infrastructure for identification in relation to DRM (though to be clear, no one would claim that PURLs provide such a comprehensive facility; they are a useful tool for simple local persistence management).

DRM Identifier Implementations Require Metadata

In assigning an identifier to a single digital entity it is necessary to also provide some defining attributes if that identifier is to be widely useful. Identifiers are simply names: names that follow a strict convention and are unique if properly applied. Unique identifiers are particularly valuable in machine-mediated commercial environments, where unambiguous identification is crucial. Some identifiers tell you something about the thing that they identify – for example, since "ISBN" is the acronym of "International Standard Book Number", the identifier "ISBN 1- 900512-44-0" can reasonably safely be assumed to identify a book (always assuming that ISBN rules have been correctly followed). However, to find out which book it identifies, it is necessary to consult metadata – the identifier links the metadata with the entity it identifies and with other metadata about the same entity. Metadata is an integral part of making the identifier useful. Some of this metadata may be held in private systems (the publisher's warehouse system, for example) but some of it is more widely available (e.g., Books in Print).

If a digital identifier simply offers a system providing persistent single point location on the Internet (e.g. PURL), then metadata is not be essential to its function. However, for DRM uses, the identifier system must provide the basis for a full range of services relating to intellectual property in the network envi-

ronment: metadata becomes an essential component. It is easiest to discuss this concept by considering a specific example, the DOI, which has been designed specifically with DRM uses in mind. The DOI can identify any kind of intellectual property entity, and because it is by design an "opaque string", the user can tell nothing about what it identifies from just looking at the DOI: the user can access and inspect metadata related to the DOI, since the entity it identifies may not itself be open to direct inspection – it may be an abstract "work" or a performance. Metadata is needed because a number alone does not impart anything useful (like a telephone number without an attached name). To use the identifier we need some additional data, for example:

- what is the creation that is identified?
- does it have another identifier I might know (e.g., an ISBN?)
- does it have a name (title)?
- who are the parties responsible for its creation or publication?
- what sort of thing is it? (abstract, physical, digital or spatio-temporal),
- what is its mode? (visual, audio, etc.)
- does it belong to a particular application type (e.g., article linking)?

We cannot list "all metadata" associated with an entity (by definition impossible) but a limited "kernel", applicable to all DOIs and meeting these requirements, is the basis for extensions to specific purposes (Application Profiles), using the Handle system ability of multiple resolution as a tool [63]. Using the principles of interoperability defined by indecs, these Application Profiles can be defined in existing metadata schemes, where that makes sense for a particular user community (ONIX, SCORM, SMPTE, DC). A DOI application will use a particular set of metadata: we call this an Application Profile. If metadata is to be commonly accessible by applications, common format(s)/schemas must be used and registered. This implies a standard vocabulary or data dictionary for mappings to/from both the kernel and the wider application sets. Metadata permits both recognition of the entity that is identified by a DOI and its unambiguous specification; it also allows for the interaction between the entity and other entities in the network (and with metadata about those entities).

Well-formed Metadata; The <indecs> Framework

The analysis of the <indecs> project on interoperability of data in e-commerce systems [64] clarified the requirement for unambiguous "well formed" metadata. This does not propose that all metadata for intellectual property has to be managed in a single metadata scheme. It does though propose that all such metadata needs to be "well formed"; this will allow metadata developed in conformance to different schemes to interact or "interoperate" unambiguously. Without that interaction, different metadata schemes risk becoming the "trade barriers" of the future. There are only two types of metadata that can be regarded as well formed:

[63] DOI 5.

[64] <indecs> Web Site - http://www.indecs.org.

- Free-form labels: the names by which things are called (of which "titles" are a subset). These are by their nature uncontrolled and broadly uncontrollable. Identifiers (in the sense of section 5.1) are a specialized type of label, created according to rules, but names nevertheless. The fact that they are created in accordance with a prescribed syntax makes them less prone to ambiguity than other types of label and therefore more readily machine-interpretable than completely free-form labels.
- Metadata drawn from a controlled vocabulary of values, which are supported by a data dictionary in which those values are concisely defined. This means that the values in one metadata scheme (or in one "namespace") can be mapped to those in another scheme; this mapping may not be exact – where two definitions in one scheme both overlap with (but are not wholly contained within) a single definition in another, for example. However, the use of a data dictionary avoids the sort of ambiguity that is inherent in natural language, where the same word may have very different meanings dependent on its context. Where precision of meaning is essential, human beings can clarify definition through a process of dialogue. This is not generally the case with computers.

The mapping between different metadata schemes may be more or less exact. It may also involve considerable loss of information or no loss of information at all. It is obviously advantageous to achieve as close a mapping as is possible; this is most easily achieved between schemes that share a common high-level data model. The <indecs> data model underlies all DOI metadata. The same analysis underlies ONIX International [65], rapidly becoming widely accepted as the metadata dictionary for the publishing industry internationally. Similar developments are now occurring in other media sectors (e.g. the adoption of indecs by MPEG- 21).

Fundamental principles defined within the indecs project and used within DOI are:

- *Unique identification*: every entity needs to be uniquely identified within an identified namespace;
- *Functional granularity*: it should be possible to identify an entity when there is a reason to distinguish it;
- *Designated authority*: the author of metadata must be securely identified;
- *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.

The <indecs> data model was devised to cover all types of intellectual property ("creations" in <indecs> terminology). It is an open model, which is designed to be extensible to fit the precise needs of specific communities of interest. It was also designed to be readily extensible into the field of rights management

[65] EDItEUR.

metadata, the data that is essential for the management of all e-commerce in intellectual property. The <indecs> analysis asserts that it is essential for the dynamic data necessary for the management of rights to be built on a foundation of the rather more static data that identifies and describes the intellectual property, and that these two layers of metadata can easily interoperate with one another. <indecs> was a time-limited project, which finished its work early in 2000. Its output is highly regarded and its analysis has been adopted in a number of different implementations. The work has since been developed and further elaborated, and forms the basis for the ISO MPEG-21 rights data dictionary discussed below.

Simple metadata solutions, the most notable being the Dublin Core [66] developed as a means of encouraging resource discovery on the Web by having content creators declare any of a small core of 15 elements to their creations, do not follow these principles. The original aim of Dublin Core has been very much superseded by the remarkably effective "resource discovery" search engines such as Google, leaving a large amount of effort on metadata in search of a new area of application, and it unfortunately has been too tempting to divert this original effort into other applications which require considerably more complexity than resource discovery. "The Dublin Core, while far from perfect from an engineering perspective, is an acceptable standard for such simple metadata [but] efforts to introduce complexity into Dublin Core are misguided" [67].

Indecs provides an ontology (an explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them) for talking about Intellectual Property transactions and so will inform the creation of, or simply provide, the metadata terms for articulating practical DRM applications.

Without an ontology and structured framework, metadata terms and classifications become ultimately useless for anything other than the purpose the deviser had in mind, recalling the famous parable of Jorge Luis Borges [68]: "*These ambiguities, redundancies, and deficiencies recall those attributed by Dr. Franz Kuhn to a certain Chinese encyclopaedia entitled Celestial Emporium of Benevolent Knowledge. On those remote pages it is written that animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance*" ("The Analytical Language of John Wilkins").

[66] Dublin Core Metadata Initiative - http://dublincore.org/.
[67] Lagoze (2001).
[68] Borges (1999).

The indecs definition of metadata ("an item of metadata is a relationship that someone claims to exist between two entities") provides a concise paraphrase of much of the <indecs> framework. It 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.

Tools for Expressing Metadata Elements

The indecs framework is an abstract ontology, independent of medium and technology. Techniques are being developed which are appropriate for expressing such ontologies (structured data) on the web, notably RDF and TopicMaps. In the long term, the vision of "the semantic web" will require such ontologies and means of expressing them.

RDF, the Resource Description Framework [69], provides "a lightweight ontology system to support the exchange of knowledge on the Web" (the weasel word here is "lightweight" - for serious DRM applications, a lightweight approach may or may not be insufficient) - RDF is essentially a way of representing ontologies as attributes and relationships using XML.

The TopicMaps specification [70] provides a model and grammar for representing the structure of information resources used to define topics, and the associations (relationships) between topics, again using XML. Names, resources, and relationships are said to be characteristics of abstract topics, which have defined name, resource, and relationship. One or more interrelated documents employing this grammar is called a "topic map".

The ISO 11179 [71] standard for data elements provides a means of specifying basic aspects of data element composition, including metadata. The standard applies to the formulation of data element representations and meaning as shared among people and machines; it does not apply to the physical representation of data as bits and bytes at the machine level; nor does it speak to semantic mappings (ontologies), but if DRM identifiers and metadata are able to adopt ISO 11179 principles without disadvantage, there are obvious benefits in terms of making data widely available in a readily understood form. An ISO 11179 data element is composed of three parts:

- an object class: a set of entities
- a property: a peculiarity common to all members of an object class;

[69] W3C Web site - Resource Description Framework: http://www.w3.org/RDF/.
[70] TopicMaps.org Web Site - http://www.topicmaps.org/.
[71] ISO/IEC 11179.

– a representation, describing how the data are represented, i.e. the combination of a value domain, datatype, and, if necessary, a unit of measure or a character set.

The combination of an object class and a property is called a data element concept (DEC). ISO/IEC 11179 provides procedures and techniques for associating data element concepts and data elements with classification schemes for object classes, properties and representations and related tools such as the assignment of numerical identifiers that have no inherent meanings to humans, icons, etc.

Once a set of elements is precisely defined for a schema and readily available in some format such as XML, the schema can be used in interoperable applications.

Commercial tools such as Adobe's Extensible Metadata Platform (XMP) are now coming on stream [72] and promise to take the concepts of structured metadata and XML and provide a widespread means of applying them, though it remains to be seen how successful these become.

Interoperability

In the <indecs> framework, interoperability means *enabling information that originates in one context to be used in another in ways that are as highly automated as possible*. Commerce does not necessarily mean the exchange of money: any environment where creations are made or used employing electronic means is encompassed by commerce in this sense.

The information that needs to interoperate here is metadata: data of all kinds relating to creations, the parties who make and use them, and the transactions that support such use. The problems to be overcome are often as simple as the fact that a term such as "publisher" has a quite different meaning in two different environments which now need to exchange metadata; they are also as complex as the fact that a single creation may contain a hundred distinct pieces of intellectual property, the rights of which are owned or controlled by many different people for different purposes, places and times. Changes in the status or control of these rights, recorded in different and unconnected systems, will need to be capable of being communicated automatically in many different ways.

Types of Interoperability

Interoperability in e-commerce has many different dimensions. As traditional sectors and business models break down, organisations increasingly face the need to combine or access information that arrives in a variety of forms and that comes from a variety of sources. The creator of metadata about a piece of intellectual

[72] Rosenblatt (2002).

property will want to be sure that the accuracy and effectiveness of the information he creates (often at substantial cost) can survive intact as it negotiates a range of barriers. Automated DRM needs to support interoperability of at least six 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 linguistic and semantic barriers.
- Across territorial barriers
- Across technology platforms.

A good e-commerce metadata system therefore needs to be multimedia, multifunctional, multi-level, multilingual, multinational and multi-platform. Such an approach may be said to be well-formed.

The failure of interoperability in each of these dimensions can be seen as trade barriers to e-commerce interoperability. These barriers are not all yet generally critical, only because the volume of e-commerce traffic in intellectual property is relatively modest: yet we are now seeing an unprecedented explosion in the development of intellectual property metadata schemas. Listed alphabetically below are just some of the major initiatives where substantial metadata vocabularies, models, databases and/or interchange formats are currently being developed or deployed, showing the communities in which they currently operate or from which they were originated:

ABC [73]	(general ontology model)
CIDOC [74]	(museums and archives)
CIS [75]	(copyright societies)
Dublin Core [76]	(library originated, resource discovery)
GRID	(recording industry)
IFLA FRBR [77]	(libraries)
IMS [78]	(education)
International DOI Foundation [79]	(content industries)
IEEE LOM [80]	(education)
MPEG7 [81]	(audiovisual)
MPEG-21 [82]	(audiovisual originated)
ONIX [83]	(book industry)
P/META [84]	(audiovisual)
SMPTE [85]	(audiovisual)

These schemes, developing from different starting points, are all converging on the "barriers" we have identified. To some degree, each is finding that is has

[73] Lagoze/ Hunter (2001).

[74] International Committee for Documentation of the International Council of Museums (ICOM-CIDOC) - Web Site: http://www.willpowerinfo.myby.co.uk/cidoc/.

to become multi-media, multi-function, multi-level, multi-lingual and technology neutral. As convergence renders the traditional sector divisions increasingly meaningless, they will inevitably need to interoperate with one another substantially. In future, essentially the same metadata about, for example, a web document, may need to be handled within each of these schemes, and many more.

2.4 Creating Interoperability: Mapping Metadata

If two metadata schemes are in use and a DRM application needs access to both, then a mapping between them will need to be created. Mappings are concerned with meanings, not names; entities can have different names in different schemes, and the same word can mean different things in different schemes. Simple one-to-one mappings between schemes are commonplace; some mappings are very precise, and others loose. However, the more schemes come into play, the more one-to-one mappings will be required, each of which is costly in resources and likely to be less than adequate. With the rapid growth of metadata schemes this is becoming an increasing problem. When there are N schemes, there are (N/2)x(N-1) one-to-one mappings needed; this rapid growth in complexity can be eased by mappings through a central point or dictionary: each scheme then requires mapping once (N schemes require N mappings).

The emergence of the indecs Data Dictionary (iDD), as articulated in the MPEG-21 RDD, offers precisely such an extensible yet firmly grounded ontology for such a dictionary. It should be possible to create any required one-to-one mappings making use of the iDD ContextModel structure. The DOI's Metadata System is built on this basis: all terms used by DOI Application Profiles must be mapped into the iDD, establishing the relationship between a term and all other terms used by APs, and is the way in which semantic integrity is achieved. This is a painstaking process, but it is typically a once-off for each term or scheme, with subsequent maintenance required only when new terms are added, or amendments made. Mechanisms for modifying mappings, adding and deleting new Terms are provided for by the iDD, although of course the consequences of such

[75] International Confederation of Societies of Authors and Composers (CISAC) - Web Site: http://www.cisac.org.
[76] See above Fn. 66.
[77] IFLA (1998).
[78] IMS Global Learning Consortium, Inc - Web Site: http://www.imsproject.org/.
[79] International DOI Foundation - Web Site: http://www.doi.org.
[80] IEEE.
[81] ISO/IEC 7.
[82] ISO/IEC 21.
[83] BIC.
[84] Hopper (2002).
[85] Society of Motion Picture and Television Engineers - Web Site: http://www.smpte.org/.

changes can be serious. A mapped term becomes a part of the Dictionary. The iDD structure is capable of recognizing any number of contextual meanings, and as new ones are identified in the course of mapping, they are placed in their appropriate place in the dictionary and ontology.

The level of granularity described above is unnecessary if only two or three schemes are being mapped. However, the fundamental assumption underlying the iDD and the DOI Metadata System is that in time there will be many applications whose metadata requires integrating at various levels, whether simply at the DOI Kernel level or to support more complex searching and processing. Semantic integrity on such a scale appears unachievable without a central tool such as the iDD, for two simple reasons: precise mapping depends upon at least one of the mapped schemes having a rich underlying model in which to precisely locate the others' terms; and multitudinous one-to-one mapping schemes are unsupportable both economically and in terms of maintaining consistency.

A mapping cannot produce unambiguous or precise mappings if the terms used in the source scheme are themselves ambiguous or imprecise. iDD can accurately describe the ambiguity and leave the resolution to users. What iDD should be able to achieve is accurate mapping as far as the source data allows, producing considerably better results than a host of many-to-many mappings based on more limited models and varying techniques. The iDD contains the logic and data to support many kinds of processing, such as data transformations or the creation of scheme-to-scheme maps, but these will require the development of application software and business processes. Contextual mappings provide one of the necessary bases for semantic interoperability, but do not provide everything.

Mapping in this precise way is practically focussed on entities that can be clearly defined and have a role in the resource-based functions typical of current DRM applications. Mapping complex concepts is possible, but concepts like "digital rights management" are not currently consensually precisely defined; there is a majority view that it is digital management of rights, rather than management of digital rights, but beyond that "DRM is something to do with managing, something to with rights and something to do with the digital environment. But not necessarily" (Godfrey Rust). Focussing on what is practically definable through practical tools like the MPEG-21 RDD, rather than arguing about "what is" DRM as a whole, is likely to produce useful implementations.

MPEG-21 and Other Activity

The ISO/IEC/MPEG-21 standard multimedia framework activity [86] is one of the most promising practical developments in DRM, which has embraced a structured view of identifiers and metadata, specifically by using the indecs metadata

[86] SC29/WG11.

framework as a basis for well-formed structured metadata though the MPEG-21 Rights Data Dictionary. The details of this extensive standards effort are beyond the scope of this chapter, but it is useful to comment of the relationship of MPEG-21 to some of the concepts and efforts which have been discussed.

The MPEG-21 world consists of *Users* who interact with *Digital Items*. A Digital Item can be anything from an elemental piece of content (a single picture, a sound track) to a complete collection of audiovisual works: an MPEG "digital item" can be considered a sub-set of what DOI calls a "Digital Object". The specification of "identifier" in the MPEG-21 DII [87] is: "Digital Items and their parts within the MPEG-21 Framework are identified by encapsulating Uniform Resource Identifiers (URIs), into the Identification Description Scheme" - that is, it provides another "identifier specification", adopting URI, rather than a detailed specific implementation. Hence identifier implementations such as DOI which are specified as a URI can be used in MPEG-21 to identify Digital Items.

Whilst the framework for DRM rules for "consumption" specification by end user devices are laid down in MPEG-21 part 4 [88] [89], the full mechanism for expressing identified and described resources in a rights environment (essentially a messaging standard for permissions) requires the MPEG-21 part 5 "Rights Expression Language" (REL) - significantly influenced by and largely based on ContentGuard's Extensible Rights Mark-up Language, XrML [90]) - and the underlying MPEG-21 part 6 Rights Data Dictionary (RDD) standard [91] [92], each of which are in development at the time of writing. Two significant points should be noted:

- The "REL" is misleadingly named, from the point of view of the content industries - whilst very useful, its scope is restricted to "rights" which can be practically expressed as some *action* in a digital context, rather than *legal* concepts like "copyright" which have no direct executable equivalent; and hence it is rather more a "network privileges language" - does the user have the "right" to delete, install, execute, etc. (verbs such as copy are derived from the basic framework but are not root verbs.)
- The RDD is built on the basis of the indecs Data Dictionary (iDD) referred to earlier as a useful mapping tool, by a group of organisations representing both commercial interests and trade bodies across the content industries which sponsored a Consortium [93] to develop the indecs framework into a Rights Data Dictionary. Hence articulating the MPEG-21 RDD through a practical operating registration authority (which is necessary, since the dic-

[87] ISO/IEC 21 final.
[88] Koenen (1999).
[89] ISO/IEC 4.
[90] XrML Web Site: http://www.xrml.org/
[91] ISO/IEC 21 draft.
[92] Paskin (2001).
[93] DOI News.

tionary is by definition dynamic) will provide a common basis for mappings for DOI (which already sues the preliminary version) and other identifier system implementations in DRM.

Other DRM consortium standards activities have been launched in specific sectors, one of the most notable being the Open Mobile Alliance [94], whose standardisation work in "OMA Download" include both DRM (building on the Open Digital Rights Language proposal [95] submitted to W3C [96], which was rejected by the MPEG-21 review process) and the over-the-air delivery of generic content. OMA has the support of Nokia, a significant player in the mobile delivery of content.

In the commercial DRM market, a number of proprietary interests and solutions are currently being actively promoted: these include Microsoft (which is aligned with ContentGuard), IBM, Macrovision (a leading player in DRM for consumer media), and Sony and Phillips who have recently jointly acquired Intertrust. There are many other smaller companies developing technologies for securing digital media. Some of these can be seen as implementation layers on top of a standards framework such as MPEG-21; others adopt a non-MPEG approach (such as the use of ODRL by the Mobile Nokia). This has led some commentators to state that DRM standards will be driven by the victor in a commercial shoot-out, rather than it an industry trade association or standards committee [97]. Proprietary solutions suffer from the obvious problems of technology lock-in, obsolescence, and interoperability - despite which, it is certainly possible that one of these might become a de facto standard.

Whatever the solution or solutions which are chosen, it remains essential to have a logical and consistent application of identifiers and metadata in an underlying extensible framework (such as indecs) which can be used to map whatever solution seems to be the more popular to those solutions which are less popular.

Acknowledgements

Parts of this article are based on material from The DOI Handbook, including earlier contributions by Mark Bide (Rightscom Ltd.), Godfrey Rust (Data Definitions) and Laurence Lannom (Corporation for National Research Initiatives).

[94] Open Mobile Alliance Web Site: http://www.openmobilealliance.org/.
[95] The Open Digital Rights Language Initiative Web Site: http://odrl.net/.
[96] W3C (2002).
[97] Bulletin.

Literature

Anderson Consulting (2000):
Anderson Consulting (22.3.2000): "A Bright Future for eBook Publishing: Facilitated Open Standards". AAP Annual Meeting. Available at http://www.publishers.org/digital/dec2000anderson.ppt.

ANSI/NISO (2000).:
ANSI/NISO Z39.84 - 2000 Syntax for The Digital Object Identifier. Available at http://www.niso.org/standards/standard_detail.cfm?std_id=480.

Beit-Arie (2001):
Beit-Arie, Oren et al. (September 2001): "Linking to the Appropriate Copy: Report of a DOI-Based Prototype". D-Lib Magazine. Volume 7. Number 9. Available at http://www.dlib.org/dlib/september01/caplan/09caplan.html.

BIC:
EDItEUR and Book Industry Communication [BIC] (Nov 2000): "ONIX International Overview and Summary List of Data Elements". Available at http://www.editeur.org/onixfiles1.2/ONIX%20Overview%20R1.2.PDF.

Bide (1999):
Bide, M (January 1999): "Directory of Persons: Outline Specification"; EDItEUR ltd. Available at http://www.indecs.org/pdf/persons1.pdf.

Borges (1999):
Borges, Jorge Luis (1999): "John Wilkins' Analytical Language", (1942). translated in Weinberger, E. (ed.). Borges: Selected Non-fictions. Viking. New York.

Bourret (2000):
Bourret, R (March 2000): "Namespace Myths Exploded". XML.com. Available at http://www.xml.com/pub/a/2000/03/08/namespaces/.

Bulletin:
The Bulletin (20.11.2002): Seybold News & Views on Electronic Publishing. Volume 8. No. 8.

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

Cross Industry Working Team (1997):
Cross Industry Working Team [XIWT] (1997): "Managing Access to Digital Information: An Approach based on Digital Objects and Stated Operations". Available at http://www.xiwt.org/documents/ManagAccess.html.

DOI:
The DOI Handbook. Available at http://www.doi.org/hb.html.

DOI 5:
DOI Handbook Chapter 5. International DOI Foundation. Available at http://www.doi.org/handbook_2000/metadata.html#5.2.

DOI 7:
DOI Handbook Chapter 7. International DOI Foundation. Available at http://www.doi.org/handbook_2000/policies.html.
DOI News:
International DOI Foundation (Oct 2001): "DOI News". Available at http://www.doi.org/news/010925-indecs2.html.

EDItEUR:
EDItEUR: ONIX Product Information Standards. Available at http://www.editeur.org/.
Ehlers (1994):
Ehlers, Hans-Jurgen (1994): "Identification Numbering in the Book, Library and Information World". ISBN Review - 15. pp. 89-214.

Green, Bide:
Green, Brian/ Bide, Mark: "Unique Identifiers: a brief introduction". Book Industry Communication/EDItEUR. Available at http://www.bic.org.uk/uniquid.html.
Guarino, Welty (2000):
Guarino, Nicola/ Welty, Christopher (August 2000): "Identity, Unity, and Individuality: Towards a Formal Toolkit for Ontological Analysis". Proceedings of ECAI-2000: The European Conference on Artifical Intelligence. IOS Press. Amsterdam.

Hopper (2002):
Hopper R. (April 2002): "European Broadcasting Union Technical Review - P/Meta - Metadata Exchange Scheme v1.0". Available at http://www.ebu.ch/trev_290-hopper.pdf.

IEEE:
IEEE Learning Technology Standards Committee - Learning Object Metadata Working Group. Available at http://ltsc.ieee.org/wg12/.
IFLA (1998):
IFLA (1998): "IFLA Study Group on the Functional Requirements for Bibliographic Records - Functional Requirements for Bibliographic Records". Available at http://www.ifla.org/VII/s13/frbr/frbr.pdf.
IETF (1997):
IETF RFC 2141 (May 1997): URN Syntax. Available at http://www.ietf.org/rfc/rfc2141.txt?number=2141.
IETF (2001):
IETF RFC 3188 (Oct 2001): Using National Bibliography Numbers as Uniform Resource Names. Available at http://www.ietf.org/rfc/rfc3188.txt?number=3188.
Irlam (1995):
Irlam, Gordon (1995): "Naming".
Available at http://www.base.com/gordoni/naming.html.
ISO TC49/SC9:
Available at http://www.nlc-bnc.ca/iso/tc46sc9/.

ISO 2108:1992:
Available at http://www.nlc-bnc.ca/iso/tc46sc9/standard/2108e.htm.
ISO 7064:1983:
Available at http://www.iso.ch/iso/en.
ISO 10957:1993:
Available at http://www.nlc-bnc.ca/iso/tc46sc9/standard/10957e.htm.
ISO 10444:1997:
Available at http://www.nlc-bnc.ca/iso/tc46sc9/standard/10444e.htm.
ISO 3297:1998:
Available at http://www.nlc-bnc.ca/iso/tc46sc9/standard/3297e.htm.
ISO 3901:2001:
Available at http://www.nlc-bnc.ca/iso/tc46sc9/standard/3901e.htm.
ISO 15707:2001:
Available at http://www.nlc-bnc.ca/iso/tc46sc9/standard/15707e.htm.
ISO (2002):
ISO (November 2002): "Frequently Asked Questions about changes to the ISBN". Available at http://www.nlc-bnc.ca/iso/tc46sc9/isbn.htm.
ISO 15706:
Available at http://www.nlc-bnc.ca/iso/tc46sc9/isan/wg1n1.htm.
ISO 20925:
Available at http://www.nlc-bnc.ca/iso/tc46sc9/20925.htm.
ISO 21047:
Available at http://www.nlc-bnc.ca/iso/tc46sc9/wg3.htm.
ISO/IEC 4:
MPEG-4 ISO/IEC JTC1/SC29/WG1/ N4668 : Overview of the MPEG-4 Standard, March 2002. Available at http://mpeg.telecomitalialab.com/standards/mpeg-4/mpeg-4.htm.
ISO/IEC 7:
ISO/IEC JTC1/SC29/WG11 - Coding of Moving Pictures and Audio, MPEG-7 - Multimedia Content Description Interface. Available at http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm.
ISO/IEC 21:
ISO/IEC JTC1/SC29/WG11 - Coding of Moving Pictures and Audio, MPEG-21 Part 6 - Rights Data Dictionary. Available at http://mpeg.telecomitalialab.com/standards/mpeg-21/mpeg-21.htm.
ISO/IEC 21 draft:
MPEG-21 ISO/IEC Committee Draft 21000-6 Information Technology - Multimedia Framework - Part 6: Rights Data Dictionary. Aug 2002. Available at http://www.nlc-bnc.ca/iso/tc46sc9/identify.htm#MPEGRDD&REL.
ISO/IEC 21 final:
MPEG-21 ISO/IEC final Draft International Standard 21000-Part 3 - Digital Item Identification, ISO/IEC JTC 1/SC 29/ N5002 Sep 2002. Available at http://www.nlc-bnc.ca/iso/tc46sc9/identify.htm.
ISO/IEC 11179:
ISO/IEC 11179 Information technology – Specification and standardization of data elements. Available at http://www.diffuse.org/meta.html.

ISTC:
International Standard Text Code - ISTC - Draft ISO 21047. Available at http://www.nlc-bnc.ca/iso/tc46sc9/wg3.htm.

Kahn, Cerf (1999):
Kahn, Robert E. Cerf, Vinton G. (1999): "What is the Internet (And What Makes It Work)". Available at http://www.cnri.reston.va.us/what_is_internet.html.

Kahn, Lyons (2001):
Kahn, Robert E./ Lyons, Patrice A (2001): "Representing Value as Digital Objects: A Discussion of Transferability and Anonymity". D-lib. Volume 5 - Number 5. May 2001. Available at http://www.dlib.org/dlib/may01/kahn/05kahn.html.

Kahn, Wilensky (1995):
Kahn, Robert E./ Wilensky, R. (1995): "A Framework for Distributed Digital Object Services". Available at http://www.cnri.reston.va.us/home/cstr/arch/k-w.html.

Kelly (1997):
Kelly, Maureen C. (1997): "The Role of A&I services in Facilitating Access to the E-Archive of Science". ICSTI Forum. No. 26. pp1-4. Available at http://www.icsti.nrc.ca/icsti/forum/fo9711.html#role.

Kemerling (2002):
Kemerling, G (February 2002): "A Dictionary of Philosophical Terms and Names". www.philsophypages.com. Available at http://www.philosophypages.com/dy/.

Koenen (1999):
Koenen, Rob (February 1999): "MPEG4; Multimedia for our time". IEEE Spectrum. Vol. 36. No. 2. pp. 26-33. Available at http://mpeg.telecomitalialab.com/documents/koenen/mpeg-4.htm.

Lagoze (2001):
Lagoze, Carl (January 2001): "Keeping Dublin Core Simple: Cross-Domain Discover or Resource Description?". D-Lib Magazine. Volume 7. Number 1.. doi:10.1045/january2001-lagoze.

Lagoze/ Hunter (2001):
Lagoze, Carl/ Hunter, Jane (2.11.2001): "The ABC Ontology and Model". JoDI. Vol 2 Issue 2. Available at http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze.

NISO (1996):
NISO Standard - Serial Item and Contribution Identifier - ANSI/NISO Z39.56 - 1996 (Version 2). Available at http://www.niso.org/standards/resources/Z39-56.pdf.

NISO (2000):
NISO (August 2000): Draft Standard - Book Item and Component Identifier - NISO Press. Available at http://www.niso.org/pdfs/BICI-DS.pdf.

Osborne (2002):
Osborne, Andrew (2002): Business data in the supply chain; part 1: auto ID; "Does radio signal the end of the line for bar codes?". E.cominfo.net. Available at http://www.ecominfo.net/supplychaindata/index2.html.

Palmer (2001):
Palmer, Sean B. (2001): "New URI Schemes: 99% Harmful". Available at http://infomesh.net/2001/09/urischemes/.

Paskin (1999):
Paskin, Norman (1999): "Toward Unique Identifiers". Proceedings of the IEEE, 87 (no. 7) July 1999. pp. 1208-1227. Available at http://www.ieee.org/organizations/pubs/proceedings/intro.html.

Paskin (2001):
Paskin, Norman (June 2002): "Towards a Rights Data Dictionary – Identifiers and Semantics at work on the net". imi insights. Available at http://www.epsltd.com/IMI/IMI.htm.

Popper (1972):
Popper, Karl R. (1972): "Objective Knowledge: An Evolutionary Approach". Oxford University Press.

Rosenblatt (2002):
Rosenblatt, Bill (25.11.2002): "XMP: The Path to Metadata Salvation?". The Seybold Report. Vol 2. No.16. Available at http://wwwseyboldreports.com/TSR/subs/0216/html/contentman.html.

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

SC29/WG11:
SC29/WG11 N 4333: MPEG21 Technical Report TR 21000-1 2001 (2001-07-20): "Information technology - Multimedia framework (MPEG-21) - Part 1: Vision, Technologies and Strategy". Available at http://www.nlc-bnc.ca/iso/tc46sc9/mpeg21/wg11n4333.pdf.

De Soto (2000):
De Soto, Hernando (2000): "The Mystery of Capital". Basic Books.

Sowa (2000):
Sowa J. F. (2000): "Knowledge Representation: Logical, Philosophical and Computational Foundations". Brooks/Cole.

Sun, Lannom (2002):
Sun, Sam/ Lannom, Larry (July 2002): "Handle System Overview". CNRI. Available at http://www.ietf.org/internet-drafts/draft-sun-handle-system-10.txt.

Sun, Reilly, Lannom (2002):
Sun, Sam/ Reilly, Sean/ Lannom, Larry (July 2002): "Handle System Namespace and Service Definition". CNRI. Available at http://www.ietf.org/internet-drafts/draft-sun-handle-system-def-06.txt.

Sun, Reilly, Lannom, Petrone (2002):
Sun, Sam/ Reilly, Sean/ Lannom, Larry/ Petrone, Jason (July 2002): "Handle System Protocol (Ver 2.1) Specification". CNRI. Available at http://www.ietf.org/internet-drafts/draft-sun-handle-system-protocol-03.txt.

W3C (1999):
W3C (1999): "Namespaces in XML" - 14 January 1999. Available at http://www.w3.org/TR/REC-xml-names/.

W3C (2001):
W3C (Oct 2001): "Uniform Resource Identifier (URI) Activity Statement". Available at http://www.w3.org/Addressing/Activity.

W3C (2002):
W3C (Sept 2002): Open Digital Rights Language (ODRL) Version 1.1. Available at http://www.w3.org/TR/odrl/.

Dr. Norman Paskin

The International DOI Foundation - Washington & Geneva

Dr. Norman Paskin became the first Director of The International DOI (Digital Object Identifier) Foundation in March 1998. Prior to this he worked for twenty years in the scientific publishing industry in both the U.S. and Europe, in roles including editorial, management, and information technology development. 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.

The International DOI Foundation (http://www.doi.org) was established in 1998 to support the needs of the intellectual property community in the digital environment. The Foundation is supported by member organisations from a broad spread of interests such as technology companies, professional publishers.

Norman has led the DOI Foundation in its development of the DOI as a standardised identifier for the intellectual property communities (including text, music, images, and multimedia), which can work with existing identifiers and internet technology. He is actively involved with a range of related standards activities developments, and is responsible for the appointment of service providers for the efficient operation of the technology and business activities of the DOI system, and in engaging Foundation members in active involvement in defining policies and solutions.

For further information on the DOI initiative and the DOI Foundation, please see the DOI web site (www.doi.org).