Toward Unique Identifiers

NORMAN PASKIN

Invited Paper

This paper discusses the creation and use of unique identifiers for intellectual property. General concepts applicable to unique identifiers are defined and discussed [identifier, digital object, dumb and intelligent identifiers, readability, affordance or computability, multiple identification, resolution, metadata, persistence, granularity, derivatives (e.g., versions, formats, manifestations, and copies), check digits, and intermediate objects]. Requirements for unique identifiers are reviewed. Capacity issues for an identifier scheme and business issues (cost, antitrust considerations, and intellectual property rights) are explored. Technical and administrative issues of identifiers are discussed, with particular reference to the uses of identifiers, which necessitates intelligence within a system of unique identifiers (scope, protocol independence, multiple roles, fungibility, persistence, standards, and emerging structural metadata approaches). Two brief illustrations of failure in unique identifiers are given. The role of unique identifiers on the Internet is discussed with explanation of the architecture of uniform resource addressing, specifying the resource from a uniform resource name (URN), names and addresses, URN implementations, and a future digital object infrastructure. Brief examples of unique information identifier systems in music, text, and digital objects in general are discussed. Recommendations are made for actions to assist in the development of future identifiers.

Keywords*—*Costs of identifiers, digital libraries, digital objects, handle, identification, identifiers, information identifiers, information objects, intellectual property, intelligent identifiers, international standards, labeling, metadata, naming, persistence, resolution processes, standards, uniform resource naming, unique identification, World Wide Web (WWW).

I. Introduction

This paper explores general issues in the design and use of unique identifiers for intellectual property, especially with reference to digital manifestations. Unique identifiers are widely used to designate physical objects, assisting in trading (e.g., the Universal Product Code bar code system), and the extension of similar principles to digital and abstract entities is a prerequisite for digital commerce of rights and intellectual content. Although the design of unique identification schemes is a technical problem, it is also a business issue with implications for what is identified and how identified items are made available. For example, the scholarly information community has recognized the need for an improved identifier scheme as publication of text and illustrations in print converges with multimedia; scientific communication increasingly occurs via the exchange of multimedia data such as three-dimensional structures, interactive mathematical equations, graphs, video clips, and links to Internet sites. The scientific, technical, and medical (STM) publishing world has to assess both the mechanisms and the commercial implications for electronic interchange (both for research and for commerce), with information identifiers (unique identifiers applied to pieces of information) as enabling technologies [20], [37], [38]. The development of identifier systems is currently a fast-moving area, and this paper presents an overview as of mid 1998.

II. Concepts

We begin with a brief overview of concepts (some of which will be explored in depth in later sections). An identifier is an unambiguous label which specifies an entity. In computer science terms, an identifier is a name; the entities named occupy a specific domain of application, the namespace, and identify points in that namespace. "Naming is one of the most important and most frequently overlooked areas of computer science. In computing it is rumored: everything is a naming problem" [27]. Once points in a namespace are addressable, applications can be constructed which provide links (i.e., denote relationships) into the namespace or between points. 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 or are the product of separate authors.

The principal reason for assigning identifiers to points in a namespace is to realize abstract namespace as a real digital environment (addresses in a 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" [32]. "A digital

Manuscript received March 18, 1998; revised November 9, 1998.

The author is with The International DOI Foundation, Kidlington OX5 1XU U.K. (e-mail: n.paskin@doi.org).

Publisher Item Identifier S 0018-9219(99)04953-1.

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 . . .” [62]. These definitions portray a digital object as a meaningful piece of data, reflected in other descriptions such as document-like objects (DLO) [6] or knowledge objects (KNOB) [33].

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, spatiotemporal performances, and abstract works [48]. These may each have applicable namespaces, not all of which have digital realizations. From the standpoint of the Internet, a digital object is a resource as specified in the “uniform resource” addressing schema.

Uniqueness is the essential attribute of an identifier, which must be unambiguous in the defined namespace; a given identifier must specify (“point 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. 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 [e.g., the international science publisher Springer-Verlag gives both a German and U.S. International Standard Book Number (ISBN) for each of its publications].

A dumb identifier is an identifier string which serves solely as a unique label and has no other inherent or implied meaning or use (synonyms: simple or insignificant identifier). An example is a manufacturing sequence number; an example of this used as an information identifier is the Publisher Item Identifier (PII), which is simply a sequence number from an individual publisher (originator), preceded by a string designed to guarantee uniqueness to an originator.

An intelligent identifier is an identifier string which serves both as a unique label and also has at least some part which is capable of ready interpretation outside the identifier scheme to derive meaningful information (synonyms: compound or significant identifier). A manufacturing sequence number which explicitly included as its opening string the year of manufacture would contain such intelligence. An example of intelligence in an information identifier is the Serial Item and Contribution Identifier (SICI), which 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 namespaces.

Context is necessary to interpret both dumb and intelligent identifiers: a string such as 0 262 193 736 is truly dumb unless the context in which it must be interpreted is known. In computer science terms, an identifier is a namespace-specific string; context is given by denoting the particular namespace in which that string has validity, e.g., the ISBN namespace (in which the string 0 262 193 736 now has meaning, enabling its interpretation as the ISBN number 0262193736, denoting the book *Internet Dreams* by Stefik).

Readability refers to the design of an 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 Internet protocol (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 PII, which consists of 17 alphanumeric characters in a single string (e.g., S1 384 107 697 000 225); for readability, when the PII is printed, slashes, spaces, 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)00 022-5]. 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 SICI 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.

In the current multimedia world of both digital and nondigital access and use, some users will identify intellectual content by reference to physical manifestations of the content which may not carry an identifier affixed to them, so it may be desirable to develop numbering systems which have affordance [20], i.e., 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. Affordance is therefore a counterpoint to the concept of intelligence: intelligence implies ability to derive, from the identifier, some element of metadata about the object; 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. An example is the SICI scheme, which allows a SICI code to be created by algorithm from known citations; while this could be done manually, it can be automated by algorithms (e.g., SICIgen). This enables a user to retrieve citation records from various databases and subsequently create the SICI code which could then be used to search more efficiently across multiple text databases to find the actual article. Given the variation 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.

Resolution is a process in which a unique identifier is the input (a request) to a system to specify some specific output, e.g., a location [such as uniform resource locator (URL)] where the object can be found. The system supporting this capability is a resolver. The ultimate aim of Internet resolution design—specifying resource locations—is an architecture of resolution infrastructure that is multitiered and permits distributed information management at all levels (rather than centralizing information away from its "caretaker," which produces lags in updating and other errors). A registry is a mechanism which reveals what resolution systems are able to resolve the named object (i.e., denotes which systems have knowledge of that particular namespace).

Metadata are data about data; they are the most important and difficult concept to deal with when designing applications of identifiers. Metadata can be used to provide supporting data such as "descriptive metadata" (a description of the package which contains the digital object, of which a subset may be bibliographic data) and "rights metadata" (intellectual property business rules). For a piece of digital information, descriptive metadata would include information such as the format, number of pages, number of tables, version number, number of embedded items, etc. Rights metadata would include information such as copyright owner, license fees for defined usage, distribution rights, etc., and thus codify some of the business behavior of the object (analogous to the properties of "objects" in object-oriented computing environments). Metadata may be highly valuable as an analytic and commercial tool (several companies provide metadata and data mining tools for retailers, trading not in physical objects but in information about object transactions). The ability for metadata to be shared across many applications and contexts has led to the need for interoperable metadata, discussed below.

Persistence is permanence of naming, enabling unambiguous specification of entities for an indefinite period. It implies that the identification scheme should be neutral as to medium or carrier but relate to the "core" digital object content in whatever format it may be preserved or carried forward. In practical terms, no one designs unique identifiers for infinite lifetimes but for what is considered reasonable, and this may vary between participants. Librarians, publishers, lawyers, and computer scientists have greatly different notions of persistence: anything over a few years old is "ancient history" for most technologists, while librarians know that the work they do today will be used 100 years from now; these extremes present different perspectives on persistence. From an engineering perspective, there is no such thing as true persistence; there is only a designed lifespan, and a migration plan to a further system before that lifespan is reached [35].

Granularity denotes the degree of resolution or granule size of the information items identified. For example, an identifier could be applied to a complete article; for some purposes it is useful to identify individual figures or tables within that article, a finer level of granularity. Similar issues arise in the context of: identification of components and collections; a series or group; an item and parts of an item; a high-resolution, uncompressed image and a thumbnail image; a database; a record in the database; and an image of an item described by a record.

Versions, formats, manifestations, and copies are all relatively loosely defined terms used by different authors in different contexts. They relate to derivatives of an object in terms of time, medium, place, number, or iterative degree of intellectual content and need to be carefully and unambiguously defined if they are to be of real use (as discussed later in this paper). Derivatives may need to be separably identifiable for some purposes, yet grouped as one for other purposes. Therefore a single identifier may be incapable of carrying sufficient information to allow for this and a set of identifiers, related via metadata, required.

Identifiers 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 minimized. Recalculation of the check digit from the body of the number, followed by comparison with the stated check digit, can be performed automatically ("hashing algorithms") 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 minimizes the chances of internal cancellation. Check digits occur in, for example, ISBN's, ISSN's, and in other contexts, such as bank account numbers; the International Standards Organization (ISO) has published a recommended standard ISO 7064 for check digits [8]. 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 are already in place, although their original introduction was to ensure consistency in both types of activity. Many information identifiers currently under discussion are being considered for use on Internet systems, which have error correction in the transmission protocol, but not on entry; URL's do not contain check digits. This has led 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. Omitting check digits in bank account numbers would not provide adequate error protection for most users.

Intermediate objects or meta-objects are objects which are not themselves the end point of a specification by a unique identifier, but which may be a useful aggregation of information. Meta-objects are a cluster or set of metadata, treated as if they were themselves a first-class object (i.e., assigned a separate identifier). Bibliographic data, for example, could be considered an intermediate object (a catalogue record) or considered part of the "header" information of the object itself, i.e., part of the overall metadata. The term indicates that, conceptually, the intermediate object is viewed as a step on the way toward retrieving the object: it is an application of metadata. The intermediate object may be a "shop front" or "parts list" for a source of the object in various formats from a particular vendor. Technically, intermediate objects are no more difficult to specify than other objects, but the business issues associated with them (such as who has rights to specify to or from what) may be complex. Meta-objects may serve a useful purpose as a linking object specifying pointers to several related objects, and thus form an essential component of some identifier schemes [2].

III. Requirements

Is it possible to define basic requirements or desirable specifications for identifier design? Two sets of well-thought-out general requirements which represent a useful starting point for specific proposals are those of Thacker (ISO) and those of the uniform resource name (URN) implementers: the two have much in common, but each also highlights some specific useful issues. I recommend that any proposed identification scheme is compared against both of these templates, summarized below, and any areas of noncompliance re-examined. Some of these requirements have implications which are beyond the technical ones; for example, the recommendation for the administration system for an identification system to be decentralized but controlled (i.e., the responsibilities for maintaining the system should be parsed among a number of participant groups) raises a number of political issues regarding who runs the identification system and how the costs of running the system are apportioned. Although the identifier itself is a neutral thing, the applications using that identifier may have competing interests. The same holds for the role(s) of the parties involved in the assignment and administration of the identifier, which may have vested interests.

A. ISO TC46 Suggestions

A useful list of requirements was proposed by Thacker of the ISO/TC 46/SC 9 (ISO Technical Committee 46: Information and Documentation Standards) Secretariat [54]; it builds on earlier ISO considerations described by Ehlers [29].

1) *Uniqueness:* An identifier must be assigned to one object only and must never be reused.
2) *International:* The identification system must be international in scope
3) *Neutral:* Identifier should not be inextricably tied to a specific application; it should be able to function as a common identifier for a range of applications and interested parties and, therefore, should not carry any intrinsic "baggage" that is specific to the needs of a single application or interest group.
4) *Persistent:* Identifiers must have an unlimited lifespan even though the objects they identify may not.
5) Designed for *ease of use* in automated systems: identifier should have a prescribed syntax.
   Desirables:
   5.1) identifier should preferably be numeric or capable of conversion to numeric form;
   5.2) should preferably incorporate a check digit;
   5.3) component elements should be computer parsable.
6) *Granularity*: Identification system should accommodate a single or composite item as well as any "meaningful" individual units incorporated in a composite item. Components should be capable of being described or defined as single objects.
7) *Definitions (Scope)*: The scope of the identification system must be defined in terms of the type of objects to which the identifier may be applied and in such a way as to delimit its scope from any similar identification systems for other types of objects. The role(s) of the parties involved in the assignment and administration of the identifier must also be defined.
8) *Capacity*: Identification system must have the numbering capacity to cover the volume of objects defined within the scope of the system.
   Desirable:
   8.1) the construction of the identifier should include an element to refresh the capacity of the system at periodic intervals (e.g. a year of assignment identifier or some sort of administrative sequence identifier).
9) Places(s) to attach/embed identifiers and to record associated metadata. This also requires the development of directories linked to the identifier.
   Desirable linkages would be:
   9.1) mechanisms to resolve the identifier to a copy of the item it identifies;
   9.2) ways to link related materials;
   9.3) ways to link related identification systems where applicable.
10) *Administration System*: Preferably decentralized but controlled; responsibilities for maintaining the system shared among its participant groups, geographic regions, and/or interested parties.

11) Must meet the needs of a variety of interested parties: all parties involved in the document "food chain" should share the same identifier for an object throughout its life cycle and have a voice in the development of the identifier system (leads to the need for such identifiers to be developed within recognized standards-setting forums at international levels).

B. URN Requirements

The Internet Engineering Task Force (IETF) informational RFC 1737 [57] laid out functional requirements for URN's, identifiers applicable to the Internet. It also made recommendations about the form that such names might take. The finalized URN syntax (RFC 2141) adopted essentially all these recommendations.

1) *Global Scope:* A URN is a name with global scope which does not imply a location. It has the same meaning everywhere.
2) *Uniqueness:* The same URN will never be assigned to two different resources.
3) *Persistence:* It is intended that the lifetime of a URN be permanent. That is, the URN will be globally unique forever, and it may well be used as a reference to a resource well beyond the lifetime of the resource it identifies or of any naming authority involved in the assignment of its name.
4) *Scalability:* URN's can be assigned to any resource that might conceivably be available on the network.
5) *Legacy Support:* The scheme must allow support of existing legacy naming systems if they satisfy the other requirements described here.
6) *Extensibility:* Any scheme for URN's must permit future extensions to the scheme.
7) *Independence:* It is solely the responsibility of a name issuing authority to determine the conditions under which it will issue a name.

(Notice that these requirements focus on the URN, but make no assertions about the resource that it identifies; e.g., a URN may be globally unique and last forever without any guarantee that the resource identified by the URN is unique or permanent.)

IV. Capacity

In order to design an identification scheme of unique identifiers for a set of entities, we need to know how many unique entities there are (the potential size of the namespace). The identification system must have the numbering capacity to cover the volume of objects defined within the scope of the system, over the defined period of coverage, with the likely level of traffic. This may be simply a question of syntax (e.g., defining a fixed field length for allowable identifiers, being easier to program than floating field lengths with delimiters), or it may be a technical issue of capacity of the infrastructure to be used for processing the identifiers. It may be prudent to design the system to be able to refresh the capacity of the system at periodic intervals, e.g., including in the identifier an intelligent piece, such as year of assignment or an administrative sequence identifier, which allows selective archiving. It is not enough to count the current number of entities; there must also be consideration of what the rate at which new entities are created, bearing in mind that identifiers will not be reused even if the entity they refer to is no longer extant (e.g., ISBN's for out-of-print books are still used to identify that book and are not reassigned).

For example, in designing unique information identifiers for STM publishing, which represents a small subset of this total but one which is quite well defined and which makes a useful case study, the approach taken was to define the entities as scientific documents, i.e., journal articles. The Institute of Scientific Information (ISI) Journal Citation Reports Database carries around 670 000 documents per annum (from 6000 journal titles); assuming this represents about 75% of the significant material to identify (an informed estimate), and knowing that this figure is growing at 2.5% per year, at least 1 million documents per year need to be discernible in any comprehensive identification scheme covering STM materials. This assumes one version of each document; taking into account multiple year coverage, potential complications in electronic publishing (e.g., versions of a document existing on a server as a preprint, as a paper submitted for publication, as an accepted paper, and as a paper with comments from readers), and smaller granularity if components are to be identified, at least 10^{13} distinct identifiers are needed for STM publishing entities alone [37].

The estimated capacity of the identification scheme must then be related to the capacity of the proposed implementation technology. Digital identifiers use the Internet. It is not possible to determine the exact size of the Internet, where hosts are located, or how many users there are. There are, however, several sources of useful data such as Network Wizards [61] and the Co-operative Association for Internet Data Analysis [4]. It is sometimes assumed that the number of possible items which could be identified will in any case far exceed the number of possible URL's, necessitating other means of differentiation; there are two counter arguments to this. First, URL's are file structures which can be subdivided within a given server to an effectively arbitrary length. Second, the number of possible "locations on the Internet" will increase—the future generation IP IPv6 will use 128-bit addressing, which allows for 665 septillion (665 $\times 10^{24}$) addresses per square meter of Earth's surface [12].

V. Business Issues

A. Antitrust

Electronic commerce is becoming a significant economic activity, and the assignment of unique identifiers will be an essential component of a commercial digitally enabled activity, without which a commercial undertaking is severely handicapped. If the identifier assignment is controlled through one central mechanism, with alternatives being nonexistent or inferior, there could be a danger of this mechanism assuming a monopoly role and unfairly influencing trade. While this has not yet happened with unique identifiers for individual objects, discussions of possible mechanisms for URN assignment [and URN-compliant implementations such as the Digital Object Identifier (DOI)] in 1997–1998 were certainly affected by the perceived similarity of any central identification/naming registration and the root structure of the Internet and domain naming scheme of the Internet, which did encounter such legal difficulties in uniquely naming (identifying) Internet websites. Network Solutions, Inc., a private company, had an exclusive arrangement with the U.S. government to assign addresses that end in ".com," ".org," and ".net." Network Solutions' monopolistic arrangement with the National Science Foundation (NSF) drew criticism from Internet users when ".com" became a *de facto* "world class" commercial name. The arrangement became the subject of a Justice Department antitrust investigation and two lawsuits in federal court, leading eventually to proposals for a system of competitive domain-name allocation [13], [60].

B. Costs of Identifiers

A major issue which must be considered in any implementation of an identifier scheme is how the administrative work necessary to sustain the system, plus the development of further functionality or responses to other developments, is funded.

The development of an identifier scheme, as with any standards development, is a time-consuming and therefore costly effort. It may be possible to fund this by voluntary efforts (each participant organization paying its own way), such as the ISO committees and work groups, or in a more structured funding body such as W3C or the International DOI Foundation. However, at some point the identifier scheme will need to move to an operational phase, and there will be costs associated with registration, maintenance, and systems. The choices are then to have a scheme which recovers costs directly from its operations (linked in some way to number of identifiers issued or usage) or one which is subsidized for the benefit of the industry or community concerned. Ideally, an identifier scheme should be an open standard with a low cost barrier to entry; but there are no "free" identifiers, only "differently funded" identifiers where the cost to some participants in the chain may be zero.

C. Intellectual Property Rights

The 1996 Geneva WIPO Copyright Treaty [59] has been mistakenly cited as necessitating or encouraging the use of unique identifiers, generating "urban myths" such as "no legal protection can be granted to an object which is not uniquely identified" or "it is an offence to modify or remove any information related to intellectual property inside a digital file." Some clarification may be helpful.

The 1996 treaty was the first to specifically address electronic or digital copyright. The WIPO Treaty, however, has no teeth until and unless it is enacted in the national legislation of the treaty-signing countries. It is expected that because of the compromises made in Geneva during the final negotiations, where much of the detail was taken out in an effort to obtain consensus, different countries will interpret and implement the treaty differently. In the United States, discussion of the enabling legislation [24] raised a number of issues, in particular, liability of online service providers for copyright infringement; this issue could potentially implicate link providers and identifier directory functions in liability for infringement. It is not true that an object is only protected if it is uniquely identified: all objects or works are protectable under copyright law, whether or not they have an identifier attached to them. The treaty did note that national legislation should prohibit the modification or removal of copyright management information (Article 12 deals with protecting rights management information contained in headers) and further should discourage technical devices that would strip away this information as well (although legislation to prohibit circumvention by device is not essential in order to comply with the treaty). Clearly the drafters of the treaty thought that technical protection of copyright works was important; however, there are obvious problems in reliance on technical remedies, as technology can prevent infringing conduct but also inhibit noninfringing conduct. The treaty does not imply that technical protection or the use of a unique identifier is an absolute need. Tagging a file (or assigning an identifier) is useful both as a self-help measure and as a way of measuring or providing evidence that an infringement has occurred but has no independent legal effect. Unique identifiers such as the DOI will fit into the process of piracy prevention by enabling the online tracking of digital objects and should help to combat misuse, for example, by being suitably embedded and concealed within the digital object using a watermark.

VI. Technical and Administration Issues

A. Intelligence Within a System of Unique Identifiers

An identifier scheme must be usable so as to provide some information: applications must be built, using the identifiers in meaningful schemes. Therefore intelligence must exist at some point in the system, either in the identifier itself, the system which manages the identifiers, the associated store of metadata related to the identifiers, or (failing all else) in the user who is presented with the result of a query and must decide how to interpret it. An initial starting point in many identifier schemes is to attempt to create intelligent identifiers, which carry within the identifier structure some information which will be provide information to the user. There are problems with this approach.

First, note that some types of components in such an intelligent identifier could be "politically" or administratively problematic; e.g., year of publication (because it could be interpreted as relating to the lifespan of copyright in the object. A less controversial year element would be the year the identifier was assigned), producer, publisher, or other rights holder (because some parties may not want this information to be explicit and because the rights holder may change over time, thereby rendering this element meaningless), or country of origin of the object. More desirable (uncontroversial) intelligent components could be a code for the group or regional agency that assigned the identifier or a code for the type of object or identification system to which it belongs.

Second, intelligent components tend to be application specific, which can create problems for wider implementations. For example, it is possible to create an identifier which carries some intelligence as to format, but there are many potential uses for identifiers; while an intermediary may need to select by format, irrespective of intellectual content (and therefore find this scheme useful), a reader may wish to select an object by intellectual content, irrespective of format (and therefore find this scheme unhelpful). We cannot prescribe or predict the uses to which an identifier will be put, and therefore whether or not two objects are classed as resembling each other in some way is in the eye of the beholder and will need to be construed from object metadata, specified by the identifier, but not within the identifier. To include an intelligent code indicating similarity within two identifiers is to "hard wire" that defined similarity (and usage) between the two objects identified. Those favoring intelligent identifiers claim pragmatic considerations mean that it will be very useful if certain resemblances (the more common ones likely to be of use to the particular type of user the system is serving) can be easily found from the identifier itself, avoiding look-up tables.

Third, hard wiring a link to another namespace presents problems if at some point the rules of that second namespace change.

A possible alternative approach is to attempt a mechanism analogous to that of library "faceted classification" schemes [47], whereby a single identifier is broken into several fields, each with a classification element. Many of the problems encountered in library classification will be encountered in such an approach; faceted schemes for identifiers have been largely abandoned in favor of approaches which link the identifier to other elements, creating a network of components which can be used as a system to convey intelligence.

A usable identification system must therefore be more than a unique identifier. A system will consist of (at minimum): an identifier; a resolver which, given the identifier, specifies the object location; and a database which, given the identifier, specifies data about the object. (An example exhibiting all these components is the DOI). Intelligence can be constructed within a network of dumb identifiers by processes such as nesting (objects within objects, meta-objects) and programmatically controlled choice of multiple resolution options. There is increasing support for the construction of intelligent systems from networks of unintelligent components, rather than the use of intelligent identifiers; yet there is a persuasive argument for intelligent identifiers if they are known to be used only for predefined restricted usages. The more universal an identifier is intended to be, the greater the range of potential uses and therefore the less likely is intelligence within the identifier to be feasible. Thacker [54] made a sensible compromise recommendation that:

> ... desirable intelligent components would be: a code for the agency that assigned the identifier; [and] a code for the type of object or identification system to which it belongs [the namespace]. Serious consideration should be given to handling [other] application-specific requirements in other ways, such as by directories and databases of information linked to the identifier, or by add-ons or extensions to the identifier.

Given the increasing potential linkage and spread of individual identifier schemes in a digital environment, the best way to "future proof" an identifier scheme is to forego any intelligence within the identifer itself.

B. Scope

"Scope" defines the set of objects to which identifiers in a given namespace are to be assigned; it defines what are the distinguishable individual items within the set to be differentiated. The designer must consider whether there is a need to identify only a set of entities, or if granularity (component identification within the entity) must be considered. Since it is impossible to predict what level of granularity will prove useful at a future date, and given the impracticality of identification down to the lowest possible "atomic" level, the only practical course if components and composites to be allowed is to design an identification scheme which can identify to an arbitrary level of granularity (i.e., is extensible in construction) and has the potential capacity for this, but not to actually identify something (assign a public identifier) until it is needed, i.e., the principle of functional granularity [48]. Note that this is not the same as saying that one should not internally distinguish some entity until needed; at a mark-up level [e.g., within a standard generalized mark-up language (SGML) document], a very fine degree of granularity is advisable at the outset to ensure future flexible component reuse and to avoid the need for recoding at a later date, but the assignment of a public identifier to such granules will only be justified if the granule entity is tradeable separately.

The design will also need to consider what data model applies to the set of entities and what are the criteria by which two entities are specified as different for the purposes of this particular namespace. Objects may be different but related, and an additional consideration will be how such relationships are described. This is likely to be an issue which first arises in considering "versions." Typically, it is simple to identify a given object (or describe a domain or set of such objects, e.g., "scientific articles") to differentiate in perpetuity between that object and another object which is in the same set ("different but of the same sort"). The problem arises in trying to determine whether we should differentiate, by means of different identifiers, cases such as the following:

1) an article in pdf and the same article in hypertext mark-up language (HTML);
2) an article in HTML with highlighted links, and the same with "greyed out" links;
3) an article and the same article with incorporated corrections;
4) an article and a second version (edition) issued some time later;
5) an article and a summary of the same article;
6) an article with full figures and the same with thumbnail figures;
7) an article and a copy of that article at another location;
8) an article and a facilitating functional copy of that article at the same location (e.g., a cached copy), etc.

It is possible to think of cases where each would need to be distinguished. Should this be by intelligence of the identifier or carried as metadata? A logical analysis (such as those described below) will describe versions as one aspect of wider object relationships.

A third issue is the identification of entities which change in time and dynamic data sets, e.g., a set of data in response to a query with parameters of time and scope. This is increasingly likely in an interactive digital world and is being addressed both by traditional bibliographic groups such as ISSN, who are considering numbering of dynamic or "ongoing" entities [43] and by Internet considerations of "human friendly names" (in which the name "weather" would determine a different result depending on the location or user), though one might argue that these are examples not of names but of queries. As of yet, there is little that is useful to say on this topic.

The Common Information System (CIS) of music publishing is an example of an identifier system which defined terminology, and relationships between terms (data models) [23]. Rust [48] summarizes and develops the CIS data model; defining digital objects as a type of creation, he then discusses derivatives as same-creation-type relationships of two types:

1) a version relationship (an object made through the rearrangement of the elements of another object, with or without new elements being added);
2) a component relationship (one object is a subset of another: A is a component of B, B a composite including A).

In each case, any temporal precedence in such a relationship—one of the objects being defined as the "source which came first" —should be excluded: the original or source in a version relationship may not always be identified, and one may find examples of components preceding composites and vice versa. "Each object must be treated as having its own identity ... any system which relies on a family relationship will break down" [48].

In the same paper, Rust defines manifestations as different-creation-type-relationships. For the purposes of information identifiers, this category is of relevance in multimedia where creations may be expressed in differing forms.

The IFLA study on functional requirements for bibliographic records [25] has arrived at similar results; there are some differences in the analysis, but both show the essential need for a well-structured formal data model to provide clear answers to questions of scope and relationships which are otherwise intractable. If a common data model (or high-level data model allowing for further specifications in defined areas) could be agreed upon for intellectual property entities, interoperability of information identifiers and their metadata would be possible. This desirable aim is now being addressed in the European Commission (DG XIII) study interoperability of data in e-commerce systems (INDECS) [26], discussed below in the context of structural metadata.

Whatever the formal definition of scope of an identifier system in theoretical data model terms, we must recognize that in practice there will be attempts to misuse the identifier system or extend it beyond its original intent. An example is the ISBN system and "illegal teddy bears." Once the bookselling community had a numbering system which it found useful (the ISBN, scoped as applying only to books and book-like objects and applied mainly in identifying separate published books for tele-ordering), it found it efficient to use the same numbering system for inventory control; as book shops began to sell ancillary items such as teddy bears as a minor part of their range, they found it useful to allocate an ISBN to these for internal purposes. Whatever the rules of a system, there will be such attempts to expand it arbitrarily, and this must either be allowed or prevented by a central management function.

C. Protocol Independence

A protocol is a standard for communication between components (hardware and/or applications); examples are the World Wide Web's hypertext transfer protocol (http) and the Z39.50 client/server information retrieval protocol. If an identifier can be applied to, or used by, any communication process, irrespective of the specific protocol in use, it is protocol independent. Systems which can resolve identifiers to World Wide Web locations, while still fulfilling the requirements for identifiers such as persistence, etc. (unlike URL's), would be of tremendous value. There is a split in opinion in how this should be achieved: by relying on current protocols (giving a quickly implemented solution, at the risk of being locked into one protocol) or by building alternative protocols which more readily enable this (e.g., [22], at the cost of introducing a new piece of technology). There is a strong argument that protocol independence, allowing either approach, is a desirable requirement for naming schemes [11]. Separating naming from the resolution process solves several concerns:

1) persistence beyond specific locations (e.g., solving the "disappearing URL" problem);

2) different communities have different resolution needs;
3) allows resolution systems to evolve and be changed over time; it is tempting to design for IP's, the current dominant digital paradigm, yet it must be noted that widespread use of the World Wide Web is less than five years old;
4) serves to cleanly separate the issue of the technical architecture of the digital infrastructure from business issues of content providers and users who wish to build applications on this infrastructure; it is now recognized that many of the early efforts to specify URN's within IETF were slow to progress because of this complicated mixture of driving issues;
5) although most end users will use the World Wide Web to access data applications, a robust data application will have internal and support systems built with servers using other protocols, so solutions which are not dependent on one specific protocol are desirable.

Despite the ideal of protocol independence, there are those who argue that we must recognize that building identifiers to fit current protocols is a more practical and realistic option; URL's are not, in theory, protocol specific (http, ftp, telnet, etc.) but in practice, http is currently the widely used protocol. There are protocol independent information-naming systems available already (ISBN, ISSN, SICI, etc.) but these systems are not resolvable on the one global distributed network currently of greatest interest to electronic publishers: the World Wide Web. The argument in favor of reliance on DNS, http, and so on is that if these protocols are later superseded, it will be "simple" to translate protocol-bound names into whatever replaces it. "http://" is just a string of characters; in the meantime, it works, it is universally deployed, and it comes for free with every piece of web software. For example, persistent URL's (PURL's) [41] are a simple-to-implement method of indirect resolution suitable for a small-scale implementation, using an intermediate pointer which always remains associated with the end resource. The technology employed is simple: it will have to be changed at some point in the future, but the changes should not be difficult to accomplish, nor will they compromise the integrity of the relationship of names and their associated resources (at least for a small-scale implementation). Those who argue in favor of allowing protocol specificity claim that the difficult issue of naming is what goes after the protocol identifier, and that is what needs close attention; they claim that the cost of making names truly protocol independent is higher than adopting existing technology, and little identifiable practical benefit accrues from this additional expenditure (those costs will be borne by someone: publishers or libraries, or their patrons and customers; someone will pay the costs in the currency of competitiveness, of free exchangeability of names, and money).

At the least, I conclude from this ongoing debate that implementations of identifiers should be either limited in scope and scale, allowing protocol specificity, or (if they are intended to have wider applicability) be designed to be compatible with existing structures and protocols but independent and not specific to these, providing evolutionary continuity from current to future protocols.

D. Multiple Roles

Identifiers may be used as retrieval keys within a bounded set of information (the namespace) which is registered in resolvers and databases. Information identifiers may, however, be constructed from unique identifiers which have other "legacy" purposes too—other namespaces outside that bounded set—and it is important that the network discovery function, which identifiers have supported and potentially can enable, must not be forgotten or compromised in the developing electronic infrastructure. Numbers initially intended for one purpose have often assumed an expanded role for other applications; these roles include the internal management of documents by publishers, unique identification of works within international publishing commerce, retrieval of bibliographic records and citations for the development of library catalogue databases, and end users' means of finding a document. There are many variations of each, and many instances of overlap among them [40]. Simple identifiers with links to metadata support this multiple role of identifier strings.

E. Fungibility

Fungibility (interchangeability) is a critical component of naming systems. Many participants in an information chain will have need for assigning and using unique identifiers; if the identifiers assigned to various manifestations of intellectual works are to have any chance of persistence, they must be exchangeable among the classes of naming authorities; they cannot be the exclusive prerogative of publishers, governments, libraries, etc. Naming is not a technological issue, but we cannot expect the original naming authority of a resource to retain that responsibility forever; that responsibility must be transferrable in an open way, so that the most natural agency (and often multiple agencies) can manage names (but not necessarily the named resource) in a manner unencumbered by economic or technological constraints. Nonetheless, there is need for defined responsibilities and authorities if chaos is to be prevented. Essentially, the argument is for open systems, conforming to architectures, standards, and rules which are publicly available specifications.

F. Persistence

The design of identifiers should aim at persistence; an identifier must persistent longer than the entity it identifies (e.g., ISBN's for out-of-print books are useful). Persistence is ultimately a function of organizations, not technology. Technology alone will not guarantee persistence; no technical system can be infinitely supportable. At best, if well designed, it will have a finite lifetime and a well-specified process for migrating to a subsequent technology. As has always been the case, persistence in the electronic world is a function of the commitment of organizations that undertake to support longevity of access to information for reasons of economic interest (copyright owners, publishers, vendors, etc.), mission (libraries, archives, and museums as the guardians of the cultural, technical, and scientific heritage of societies), or legislative mandate (governments fulfilling their legislative responsibility). Each of these sorts of commitments has strengths and weaknesses; no one should expect more of any of them than is reasonable. Publishers are bought and sold, and their commitments will change accordingly. Cultural institutions change slowly and are resource constrained, though they are generally more stable than other agencies. Governments change, and their priorities will change to reflect the political exigencies of the time. Technologies are supportable only if there is a sufficient base to justify the supporting costs. Hence the digital library community must identify institutions of long standing that will take the responsibility for resolving institutionalized names into current names for the forseeable future [35].

G. Standards

It is desirable for identifiers to be ultimately codified within recognized *de jure* standards-setting bodies, preferably a forum with no vested interests in one particular application of the identifier. This helps to ensure the identifier meets the needs of a variety of interested parties involved in the object "food chain" (who should share the same identifier for an object throughout its life cycle and have a voice in the development of the identifier system), ensures neutrality, allows for internationalization, and is capable of extension. The main formal standards arenas for infrastructure and enabling technologies are ISO, the International Telecommunication Union (ITU), and IETF. Standards here are set by consensus from a group of potential users. The aim is publication of open standards, which can be adhered to at little or no cost to the user; the disadvantage may be a slow process of consensus. There is undoubtedly a tension at present between the traditional standards-setting activities under the ISO umbrella and the perceived need for more rapid standards evolution to allow technical progress, such as controlled by IETF and W3C and computer industry consortia [9].

Standards may be developed for specific areas of content at an earlier stage by industry bodies, consortia, or subgroups of standards bodies set up to serve these specifically, e.g., MPEG for audiovisual identifiers, CIS for music, etc. These may (as in the case of CIS) concentrate initially on closed user group standards, opening out to other industries or users at a later date, and may justify specific costs for implementation among that user group. W3C is a consortium which is, in effect, a standards-creating body by virtue of monopoly but without some of the governance structures of the formal standards bodies [19].

Finally, standards may emerge from a small group of active users promoting guidelines which may be *de facto* standards; this is in fact essentially what W3C (a consortium, not an open standards body) is doing, framed in the language of standards development, and it is able to do so because of the tremendous momentum behind its creation and the need for a body to set standards much more quickly than ISO allows. The risk of incompatibility immediately may be worth the more rapid progress which such an "evolutionary" approach can give, though at some point formal codification of the *de facto* standards may be necessary. An example is the evolution of HTML browser capabilities, which led to divergence followed by consolidation of the resulting, richer set of functionalities in a new expanded standard, eventually resulting in extensible mark-up language (XML). The Internet itself originated via *de facto* standards (TCP and IP) rather than approved ISO standards (OSI) [21]. The convergence of interests in a digital world leads to the need for identifier standards to have input and be recognized across both technical and administrative standards bodies.

H. Structured Metadata

The intelligence derived from an identifier system must lie with metadata rather than being embedded within intelligent identifiers if the system is to be extensible and used in many contexts. For a useful resource list of metadata discussions see [63].

A given entity to which an identifier is applied may have associated with it, in the identifier system, data which provide additional information, e.g., about its content, rights, etc. These metadata are potentially an infinite set. There is no such thing as "all of the metadata" for an entity, as someone may devise a system which uses a piece of associated data not previously considered and recorded in the identifier system. However, for certain uses it may be possible to define a limited set of commonly used metadata. It is desirable to define and structure this set so as to allow the maximum number of application uses of all or part of it. Given five core fields of metadata A, B, C, D and E, someone may wish to use $A+B+C$ for one use (service 1); another $A+D+E+C$ for service 2. Each will want to have the metadata fields be usable without "translation" into the application he is devising. We can consider the use of fields $A+B+C$ for service 1 to be an output from the core set of metadata $A+B+C+D+E$. We want to allow for multiple outputs from this set for various functional applications. The Internet creates a situation where metadata may be created once and remain in its original location and yet be accessed by everyone. Equally, we want to minimize the size of the overall core set and the work needed to compile it. So we want the input which creates the core set to be done once, in a well-defined way, which does not reflect a specific function but allows multiple functions. This can be done by designing the core metadata to reflect the inherent structure of entities and metadata, rather than its function in any one context or application [48].

So each element of metadata is a piece of data, an entity in its own right which may be the central concept in another application. For example, a person (entity) may have an element of metadata such as "address"; "address" may in turn be the central field in a geographical information system. Since any data piece can potentially be used anywhere else, the data must be well formed to allow this, rather than designed to be useful in only one functional application. Well formed (borrowing from the terminology of XML and similar approaches) implies in this context not only developed on the basis of inherent structure, rather than function, but related in a consistent manner.

1) Values of elements coded according to the rules of a defined namespace, or a controlled vocabulary (so a free text description is not normally acceptable, except in the case of necessarily "self-defining" or "autonomous" elements like title, but a standardized code such as standard address number is, since it enables ready reuse in other contexts).
2) Structured in relation to other entities by means of a formal data model, with entity/attribute relationships following well-defined rules. This enables any output design to incorporate the entities and attributes into their own model.
3) Using a high-level common data dictionary, data model, and set of standardized labels for metadata.

The analysis of structured metadata is beyond the remit of this paper (see especially the ongoing development of this concept under the INDECS Project [26] and the related Digital Object Identifer [14] which has adopted the INDECS approach) but briefly the approach is to define structural entities (e.g., place, party, creation, right, agreement); descriptive elements of entities (e.g., for creations, elements might include the identifier or label itself; controlled vocabulary descriptor labels, event, extent, relation) and types within these elements (e.g., within the element "extent" would be the types "spatial" and "temporal").

Examples of metadata models which are to some extent following this "well-formed" approach are found in, e.g., IFLA's Functional Requirements for Bibliographic Records [25] and the CIS for music works [23]. Such well-formed data models, describing structural entity attributes and relationships [34] can then be represented in the resource description framework (RDF) which uses XML (designed primarily as a means of extending generic mark up capability to the Web) as its interchange format. In this way, metadata which are well structured may be encoded for use on the Web, allowing interoperability of applications. RDF is a specification developed within the W3C metadata activity [36], [42] and defines a resource as any object that is uniquely identifiable by a uniform resource indicator (URI). The properties associated with resources are identified by property types, and property types have values. Property types express the relationships of values associated with resources; values may be atomic in nature (text string, number) or other resources (which then in turn may have their own properties). This triadic model of resources (=entities), property types (=attributes), and values (=attribute values) is a data model which enables an unambiguous method of expressing semantics (thus intelligence).

I. Other Metadata Initiatives

There have been many efforts to establish metadata structures for particular communities; examples include MARC records for libraries [58], Federal Geographic Data Committee for Geospatial Data [17], Consortium for the Computer Interchange of Museum Information [10], etc. These provide a tool for those communities, but in the absence of a common data model, interoperability between these "silo" applications is very limited and requires the construction of specific bilateral "crosswalks" (mapping transformations between the unrelated data models of each silo) with increasing complexity as the number of silos grows [53]. In a digital environment, this approach is no longer satisfactory.

One attempt to overcome this problem began by assuming that a common requirement for interoperable metadata would be resource discovery. The Dublin Core (DC) [15] initiative has now defined a core set of metadata for the single functional purpose of resource description: a 15-element metadata element set intended to facilitate discovery of electronic resources. The Core is described as three groups which indicate the class or scope of information stored in them: 1) elements related mainly to the content of the resource; 2) elements related mainly to the resource when viewed as intellectual property; and 3) elements related mainly to the instantiation of the resource.

Originally conceived for author-generated descriptions of Web resources, the DC has also attracted the attention of formal resource description communities such as museums and libraries. However, DC currently represents an output functional format as described in the previous section: an application of metadata for one purpose (resource description, akin to a cataloguing system), rather than the well-formed approach guaranteeing interoperability. For example, DC has two proposed separate fields ("author" and "contributor") which reflect a function in the particular instance concerned, yet the entity (person) concerned may be an author in one context and a contributor in another; this is not a structural distinction but a functional one. In the current "DC-simple" consensus, DC fields are not standardized as to content or mandatory in a well-structured form, and no common high-level data model is offered. However, mechanisms to do this are to be offered in the approach of "qualifiers," which provide further structure and define namespaces or controlled vocabularies for given elements. This approach is still developing, and there is not yet a stable consensus on the actual qualifiers, and on further possible groupings of the elements, to determine whether it can be made to conform to a well-formed structural data model. However, since the building of an interdisciplinary, international consensus around a core element set is stated to be the central feature of the evolution of the DC, it would be extremely desirable if this could be achieved. A common data set which was: 1) agreed as a structural, multi-application, set across the INDECS, DC, IFLA, CIS, DOI, etc. communities and 2) expressible in RDF would be an immense step forward in use of metadata on the Web (and therefore, intelligent use of unique identifiers).

The DC-simple (unqualified) set has claimed the following advantages: simplicity (it is intended to be usable by those with no experience of formal resource description models); "semantic interoperability" (a commonly understood set of descriptors); international consensus; and flexibility. However, this simplicity is reached at the price of not achieving full interoperability between applications, since they will arrive at their own conventions for element specification ("semantic interoperability" does not mean practical interoperability of the elements so defined). Defenders of the Core concept claim that although initially motivated by the need for author-generated resource description, the DC includes sufficient flexibility to encode the additional structure and more elaborate semantics appropriate to more formal resource description applications. Critics of the concept claim this is ingenuous; labeling a field semantically without defining a tight namespace or controlled vocabulary for it does not provide a working metadata model, and there is a danger that the elements defined from the point of view of one application will lead to conflicts in other applications. Crucially, the Core has yet to demonstrate that its "Qualifiers" approach leads to a consistent well-formed data model. However, there is a strong desire on the part of many concerned in these various initiatives to achieve consensus using the approach of adding qualifiers to the DC elements, and perhaps redefining or regrouping some elements.

DC "simple" (unqualified) has been proposed for standardization in its current consensus incarnation as a resource description metadata set for both IETF [45] and the National Information Standards Organization (NISO). It remains to be seen if the standardization processes lead to acceptance, and if so, what the implications are of having a standard set which lacks a well-formed data model. There is certainly a danger that such a set would be adopted unquestioningly by some application developers as the entity set for a structured data model, rather than starting from a structure approach and viewing the DC simple set as a simple interface (as it was originally conceived). At the time of writing, it seems too early to tell which way this argument will go. There are also attempts to apply XML and RDF specifications to describe and use the DC concepts, including: a specification of the metadata element set in RDF with examples of Core instances and extended schemes; tools for the user to create RDF and compatible DC metadata; and services to index RDF-compatible DC metadata and provide search interfaces. At the time of writing there is no clear consensus on these items.

There are also efforts to define a usable set of core metadata for another functional purpose, that of rights, and to articulate means of structured presentation of such metadata, e.g., the Stanford Framework for Interoperable Rights Management (FIRM) [18] and proprietary standards from companies working on copyright management software such as Digital Property Rights Language (DPRL), a Xerox language for expressing fees, terms, and conditions that govern the use of a digital work, intended to be read by computers acting as trusted systems [52]. These efforts are at an early stage, but because they are forced to consider rather more complicated relationships than just description, they are likely to require the structural approach for ultimate success.

J. Failures in Unique Identifiers

Some standards for identifiers may be well-formed but unsuccessful for other reasons (e.g., the withdrawn ISO 9115 standard *biblid* for unique bibliographic identification). Conversely successful, robust and mainly reliable systems may have their faults, the exceptions which prove (i.e., test) the rule. In the world of publishing there has been no more successful and widely adopted identifier than the ISBN, which enables millions of commercial transactions daily; equally, the latest version of SICI has been designed with much feedback and consultation and appears almost foolproof. Both are proven unique identifier systems, yet in both some failures occur, which are worth noting as they are often unreported, and as illustrations of how difficult it may be to reach a total solution.

The ISBN has been successful as a trading and distribution identifier, but according to some librarians a failure in terms of identifying unique bibliographic entities.

> French publishers, for example, routinely reuse the same ISBN on three or four unrelated titles released within the same year. Many publishers leave off the language digit (this was the norm for SBN, the earlier standard) or drop a digit or two from elsewhere in the ISBN, leaving ... less than 10 numbers ... Where the ISBN appearing on the piece differs from the copy, other than in the first (language) or last (check) digit, prefer the piece as source over the copy [7].

This failure is not technical but appears to be administrative.

The SICI scheme is prone to failure in some rare cases involving articles appearing on the same page and having similarly abbreviated titles. To deal with the multiple-article-per-page problem, SICI uses a "title code" of up to six characters, usually formed from the initial letters of title words. Different articles on a page can be usually distinguished by this title abbreviation. In principle, however, it is possible to have two or more articles with the same SICI title abbreviation and hence the same overall SICI code. A SICI title abbreviation also requires human judgment when the title contains symbols (e.g., #) which is a further possible source of ambiguity. These potential failures are technical, i.e., inherent to the design; the failure rate in the UnCover database is reported as three per million [5], small but enough to be problematic in mass automated systems.

VII. Unique Identifiers and the Internet

This section discusses identifiers in the context of one particular medium: the digital infrastructure forming the Internet.

Table 1
Uniform Resource Addressing

URI (Uniform Resource Identifier)	the generic set of all names and/or addresses that are short strings that refer to resources.
URL (Uniform Rcsource Locator)	the set of URI schemes that have explicit instructions on how to access the resource on the internet.
URN (Uniform Resource Name)	(1) a URI that has an institutional commitment to persistence, availability, etc.(may also be a URL e.g. PURL) (2) A particular scheme which is currently under development in the W3C and IETF which should provide for the resolution using internet protocols of namcs which have a greater persistence than that currcntly associated with internet host names or organizations. When defined, a URN(2) will be an examplc of a URI.
URC (Uniform Resource Citation, or Uniform Resource Characteristic)	A set of attribute/value pairs describing a rcsource. Some of the values may be URIs of various kinds. Others may includc, for example, authorship, publisher, datatypc, date, copyright status and shoe size: a set of fields and values with some defined free formatting.

A. Uniform Resource Addressing

Although interchange of digital object information in closed systems via electronic data interchange (EDI) has been carried out for many years, it is the rise of the Internet which has dramatically increased the need for standardized electronic interchange of digital objects in an open environment. There are a number of "uniform resource" schemes applicable to the Internet designed to provide open standards for resource labeling and location; some of these are specified in detail, others as yet only in outline.

Work on extending the various definitions and standards for uniform resource addressing was transferred recently from IETF to the W3C (World Wide Web consortium) [56] and some issues remain to be clarified. Table 1 presents the definitions current in September 1998.

The basic concepts are discussed under the W3C architecture domain [64]; nevertheless there remains some confusion in the explanation of these terms (for example, the W3C Addressing web site gives an apparent explanation of the entities referred to as URI's, of which URN's, and URL's are subsets, but then confuses the picture by referring to "URI's, a.k.a. URL's" and by noting, following the explanations, that "the problem with this picture is that...the standard-track specs don't reflect it..."). It is hard to avoid the impression that while the discussion of these concepts evolves in the hands of the cognoscenti, it is difficult for newcomers and outsiders to obtain a clear and consistent picture. What follows is therefore an attempt at an explanation of an evolving picture.

To begin with, even the term resources is rather vaguely defined, and it is hard to avoid unhelpful circular definitions such as "a resource is anything that has identity" [65]. The W3C architecture domain refers to resources as "documents, images, downloadable files, services, electronic mailboxes, and other resources" [56]. The starting point of the W3C explanation (though not, historically, of its evolution) is to note that "The web is an information space....URI's are the points in that space....URI's are short strings that identify resources in the web." The scheme defining URI dates from 1992, when the IETF standards body attempted to standardize schemes for resource naming which would avoid problems associated with reliance on the URL only (an application and concept which predated the URI). Two classes of URI were defined: the URL and the URN.

1) URL: A URL is the address of a resource on the World Wide Web; a compact representation of the location and access method for a resource available on the Internet, now defined as "the set of URI schemes that have explicit instructions on how to access the resource on the internet." URL's are conceptually protocol independent, and allowed protocols include telnet, ftp, gopher, etc; most people know these as forms using http, which enabled transfer of a great deal of information already existing in gopher and ftp sites onto web-based browsers (evolutionary continuity). This low barrier to entry meant that little planning took place to define mechanisms for persistence and exchangeability, hence the familiar problem of "404: not found" or broken links.

2) URN: A URN is a persistent identifier for a resource and is protocol independent [44]; the requirements for URN's were noted above. Again, the W3C official explanation is confusing; first two separate definitions are given: "(1) a URI that has an institutional commitment to persistence" (a rather strange definition since it refers not to technical structure but to use) and "(2) a particular scheme ... under development ... which should provide for the resolution ... of names which have greater persistence than that currently associated with Internet host names or organizations." To add even further to the confusion, it is then noted that "when defined, a URN (2) will be an example of a URI." (As an aside, this source demonstrates the problem of information identifiers very well, since I cannot be sure that this wording will either persist or even be consistent: parts of the web page referred to have been changed significantly since the draft version of this article was completed, and the wording here is taken from the web site "last revised 1998/10/20 05:43.") URN's and URL's are shown in the W3C explanation as overlapping, since some URN's may be URL's (e.g., PURL's). The diagram accompanying the W3C text shows the URI space as greater than just URN's and URL's, though no further explanation is given.

URN's and URL's are tools to locate points in information space. However, the point so located is a resource that needs to be described. The data structure of attribute/value pairs describing a resource was originally defined as the URC, a concept which has since not been developed to any great extent. Since in addition to the URN and URL information the URC could contain other data about the instance of the resource at a specific URL, the URC is perhaps better thought of as a metadata structure for the resource. By analogy with bibliographic mechanisms, some authors have referred to URC as uniform resource citation; however, I believe this to be a potentially misleading term, as a written citation is not as rich as the possible metadata architecture which could eventually emerge from this system. In fact, it seems that URC looks increasingly out of place as a resource discovery tool and should more properly be relegated to discussions of metadata architecture such as RDF.

The URN seems likely to become the most useful form of URI for intellectual content on the Internet, with URL's and URC's supporting the operations of a unique identifier system. To obtain a resource, an identifier is needed which resolves the URN to a URL (resolves a name to a location). URN's are capable of supporting existing (legacy) identification schemes, e.g., using existing bibliographic identifiers as URN's [46]. A number of efforts are attempting to implement working URN schemes [1]. The URN syntax [57] allows for multiple implementations (schemes) which conform to the syntax; the hope is that a small number of high-quality naming schemes will emerge and be accepted and used by the content industries. It is important to note that this activity is an ongoing effort which has not yet reached stability; because of this, potential users who manage large collections of digital information (e.g., publishers) have been reluctant to commit to using any form of URN during a period of flux, but there is increasing pressure from the market to provide persistent schemes for failsafe digital information commerce. Recent initiatives coupling efforts from the content industries with technical solutions (such as the DOI, based on the Handle URN implementation) offer the possibility of a bridge to enable progress to be made with real, large-scale implementations.

URN syntax has been defined (but is still subject to disputes in some areas of interpretation). URN defines a name as several components, shown in a simple hypothetical example URN, e.g., `urn:isbn:0670856053`.

1) The first component is the string "urn" announcing that this is a URN. (Opinions differ as to whether this common component will be needed and should be included.)
2) The second component is the namespace identifier (NID), which specifies which particular domain of names is being used. The namespace can be an existing (legacy) system, as in this hypothetical example using the ISBN.
3) The third component is the namespace specific string (NSS), which defines, within the given namespace (identifier system) being used, the unique label of the resource (object).

The URN, therefore, has the general syntax `urn:nid:nss`.

The particular URN naming and resolution scheme being used may optionally be shown, by including a scheme identifier (SI) which states the particular implementation scheme; for example, `urn:hdl:cnri.dlib/august95` is a URN which uses the Handle system (`hdl`) scheme [22]. In such a case, the NID:NSS string is made up of a unique naming authority (=NID) and unique identifier assigned by that naming authority (=NSS). Thus the general syntax of a URN in this type of scheme is `urn:si:nid:nss`. This URN syntax, therefore, conforms to Thacker's suggestion that intelligence within an identifier be minimal, by including in the identifier itself solely a code for the agency that assigned the identifier and a code for the type of identification system to which it belongs.

When used on the Internet at present, the URN (designed as protocol independent) has to be implemented using existing protocols. Typically, a proxy server is used which communicates with the world in a generally accepted language (protocol) such as http, and with the special URN tool in the language of that scheme. Thus, to the outside world the syntax appears "translated" into http, e.g., http://dx.doi.org/10.1000/123456789, which is a URN implementation using the Handle scheme, and within this the DOI (sequence beginning 10.1000...) as the namespace. It may be argued that this example is a URN implementation only in the broader sense (the first of the W3C definitions), and a full implementation of URN would look like urn:hdl:10.1000/... or urn:doi:10.1000/... and hdl or doi would have to be registered with IANA (if and when such a procedure exists) and further with some RDS somewhere.

B. Resolution

Resolution is the process of specifying a resource attribute (e.g., location) from a URN. Within a URI, there exist name or address strings relating to resources (Table 2). To specify from a resource name, a URN (which, recall from the previous section, is one type of URI) to the resource location (e.g., URL) a resolution process must be selected (step 1) and then implemented (step 2). The process is one of input of the name to the resolver, and output of the location from the resolver; the resolver is said to return the location (i.e., it returns the location to a user in answer to his query input).

Step 1 is to locate a resolver (which can map from URI's to the information about the resources the URI's identify: a database or some other mechanism); one approach suggested to make this process "transparent" is the Naming Authority Pointer (NAPTR), based on a new DNS resource record. This approach suggests using the existing global DNS infrastructure and has not been implemented; at present, users have to be directed to the resolver in some other way, such as via proxy. Eventually other approaches may be needed for reasons of scalability.

Step 2 is then to communicate with the resolver using a chosen protocol. The protocols supported by the particular resolver will be revealed by step 1. Most commonly at present, http is used. Thttp (=trivial http) is a specification which enables resolver requests and responses to be encoded as http, and can therefore be easily retrofitted to existing http servers [45]; however, other protocols can in principle be used, e.g., Z39.50 (recall that URN's are to be protocol independent), and development of these is likely to be an area of development. Note in this process of getting from a URN to the resource that the process is designed to be compatible with existing structures and protocols (DNS and http), but not specific to these; this gives evolutionary continuity from current to future protocols. There are proposals and implementations for alternative protocols such as Handle [22]. (The URN effort is explicitly designed to accommodate multiple identifier namespaces and resolution systems; the Handle System® comprises exactly one such case, with a very specific data and service model, resolution and administration protocols, a model for secured service, and so on.)

In the conceptual design of URN's, namespaces are to be registered and assigned unique NID's. Any resolution services associated with these namespaces require further registration with a resolution discovery system (RDS) which clients could use to begin, or discover, the appropriate resolution mechanisms.

C. Names and Addresses

In an early World Wide Web design document, Berners-Lee referred to "The Myth of Names and Addresses" [3]. His point was that World Wide Web links appear to be fragile and nonpersistent, not because of the architecture of the system, but because of poorly structured choices for http:// addresses (URL's) and management of these addresses. An extreme of this viewpoint is that there is no need for anything other than (well-maintained, well-structured) URL's, and appropriate redirection. (To say that all that is needed is URI's is, in effect, to side step the question until URI's are clearly defined and fully implemented.) In an ideal architecture this may be valid; nonetheless, the fact that there are low barriers of entry for creating URL's and few rules beyond the outline structure has indeed led to fragility and lack of permanence of http:// locations; attempts to add on to this "technical hacks," such as redirection mechanisms, do not appear to be efficient in the long term (e.g., one can imagine that over time a content which moved to several sites could create a lengthy zig-zag of redirection, much like physical mail being redirected on to several addresses). There are also issues of more concern to intellectual property owners, content providers, and users, not just technical fixes, such as how citations are recorded, i.e., as the end resource, or the "wrong" http:// starting point if a resource is moved? Experience has shown that most providers of resources are relatively poor at managing the infrastructure necessary to ensure persistence (e.g., ensuring that URL's are maintained and do not change), not necessarily because of incompetence but because resources may change ownership and location as part of business transactions.

It should be noted at this point that URL should not be considered synonymous with http://, since http is only one of several allowed URL protocols—others include ftp, gopher, and telnet—and it is logically possible to define other future URL types which are better structured as to naming permanence, but usually this is not what is meant in such discussions. It is also not at present clear what is the process for adding to the list of URL types maintained by IANA, although this is being worked on, and the Handle technology (for example) has been suggested as an additional allowed URL type (not yet accepted). I believe that creating unique identifiers for resources (not locations), and making these protocol independent, offers a structurally sounder solution for the problems of information-content management.

Finally, there is a danger of assuming that a grand design which encompasses all possible digital resources and transactions in the web "information space" will solve all the problems of information identification: it will not. Much information is still nondigital, and the world of intellectual property will continue to contain physical creations (e.g., books), abstract creations (works), and spatiotemporal creations (performances), as well as digital creations. Most rights transactions still relate to use of intellectual property in physical form. Perhaps the first step to bridge this gap is to provide digital services associated with nondigital creations.

D. Some URN Implementations

As described above, the URN syntax allows for multiple naming schemes. A URN namespace is now defined as an existing or new namespace that has a well-defined mapping to the URN syntax and that also has support systems for resolution using the URN infrastructure. These support systems may change over time. It is presumed that different URN namespaces will be proposed for identifying different types of resources and different assignment policies. Individual communities will establish resolution systems that are tailored to their needs, but there is nothing inherent in the identifier that precludes the use of other resolution systems [11]. Note that the owner of a namespace does not have to be the entity that manages the operation of the resolution service; that creation of a URN namespace implies a certain commitment (direct or delegated to another party) to provide the ability in perpetuity to resolve published URN's to resources.

Table 2
Use of Unique Identifiers in the CIS of the Music Industries

Resource type	Identifier tool
Work (Creation): Works Net* (Works Information database)	ISWC =International Standard Work Code* [ISWC]
Participants: Interested Party Information	IPI Number (successor to CAE = Compositeur, Auteurs, Editeurs file)
Agreements: Agreements Information file*	ISAC = International Standard Agreement Code*
Accompanying audiovisual materials: AV Index	ISAN =International Standard Audiovisual Number* [ISAN}
Physical manifestation: Sound Carrier and Recording Information File	ISRC = International Standard Recording Code [ISRC]

* indicates under development.

Some of the URN project implementors have worked together in attempting to reach agreement on the URN framework [1]; examples of URN implementations include: the Resource Cataloging and Distribution Service (RCDS), aimed at solving scalability and service issues; Handle System (discussed below); x-dns-2 and Path URN (both based on the Internet DNS); and work toward using Whois++ as an Internet Directory Service, including distribution of URN resolution data and maintenance responsibility in a global publishing environment.

Handles [2] are a promising implementation of URN's which have been put into use in a variety of projects. Handle technology is predicated on the ideas of Kahn and Wilensky's digital object framework [32]. Handles were conceived within the URN syntax but have since been submitted as a separate IETF draft [22], intended to become an IETF informational RFC, in recognition of their more extensive potential structure. Applications include: the DOI, an identification system for the publishing industry [14]; the Library of Congress's National Digital Library Program using handles to identify material in the library's own collections and promote the use of handles with other collections; and the Networked Computer Science Technical Reports Library, a cooperative project involving computer science research groups across the world, led by Cornell University, creating an open architecture using handles. Handle specification allows some key metadata of an object, notably rights data, to be held within the handle and thus travel with the object.

E. Unique Identifiers and a DOI

Identifiers of digital objects are the initial enabling step which allows the construction of digital economies, or business applications, based on managing access to digital information. These concepts are beyond the scope of this paper, but one underlying technical framework is described in the work based on digital objects and stated operations of Kahn and Wilensky [32], [62].

The digital property trust concept of Stefik [51] is more specifically focused on rights trading enabled by specified digital objects, building also on the Xerox work on languages designed for this purpose [52]. It proposes:

> . . .an organization that ensures the health of digital publishing and promotes a lively international commerce in digital works. In conjunction with consumers, publishers, creators, and platform vendors, it would set the standard for the evolving digital property language and issue digital certificates to conforming platforms. It would also maintain the master repositories and perhaps ensure security and financial transactions.

While both of these concepts describe a theoretical construct for a DOI, a recent practical step has been taken with the launch of the International DOI Foundation, a not-for-profit body which aims to support the needs of the intellectual property community in the digital environment by supporting, developing and governing, as a neutral trusted party, an enabling technology of identifiers of intellectual content (in all forms) linked to a technology for digital services [14].

VIII. Some Examples of Unique Information Identifier Systems

Two examples serve to illustrate that unique identification systems require both careful technical design and also the solution of a number of commercial application issues. Note that both of these examples use digital technologies but relate not only to digital intellectual property but other media. One of the better developed efforts in providing commercially usable identifier schemes is the CIS of the music industry [23], which has a long history of rights management and a need for automated interoperability. When complete, this scheme will encompass a range of identifiers for defined purposes, working within one data model to enable automatic tracking of rights. Table 2 summarizes the main components as an indication of the scope of this system. As of yet, the CIS is not Internet-enabled, though CISAC (the system's governing body) is actively investigating how the system might be opened up to related activities such as the DOI.

Another area where unique information identifiers have been developed is in identification of scholarly journal articles. Two initial approaches which illustrate different philosophies are the PII and SICI. The PII was promulgated by a group of STM publishers [39] as a dumb identifier for articles which could be allocated at a very early stage of the publication process; the intention was to stimulate development of systems using this concept, potentially extending its scope (e.g., components, books) and encompassing many potential uses; its intent was to identify the abstract work manifested in a particular published entity (a concept now being developed as the ISWC [31]). SICI [49] is an evolution of an earlier identifier principally aimed at computable bibliographic descriptions for cataloguing and resource discovery. Both provided a potential enabling technology or standard, deliberately without defining commercial mechanisms for their implementation. Neither specify defined metadata approaches, leaving this for specific applications. The two describe the same entity resource (article) using different schema, e.g,

PII : S0 361 923 096 003 310
and
SICI : 0361-9230(1997)42:⟨245:OaEoSR⟩2.0.TX;2-B

refer to the same article.

At the time when PII was devised the SICI (version 1) had very limited applicability to electronic articles; a recent revision (version 2) has considerably widened its potential applicability. The large degree of intelligence designed into the SICI conveys both advantages, such as some affordability, and disadvantages, such as relative unwieldiness. Approaches such as PII present a much simpler identifier, at the price of necessitating accompanying metadata and look-up schemes. Both are now possible component elements of an Internet-enabled wider scheme, the DOI.

The DOI, a Handle (URN) implementation, was initiated within the publishing industry by the Association of American Publishers (AAP), and has now been taken over by an international not-for-profit foundation with widespread support from a wide range of content and technology organizations [14]. The system uses an identifier syntax, a resolver directory, and associated metadata. Eventually, there may be separate directory managers in each country or operated for each applicable industry sector (publishing, photos, music, software, etc.). The identifier is capable of incorporating legacy identifiers (both existing standards and proprietary ones). The power of the DOI system results from its persistent, automatic routing system, combined with an administrative system supported by the content providers. Because digital content may change ownership or location over the course of its useful life, the DOI system involves a directory linking the current web address of the resource associated with that DOI. When an object is moved to a new server or the copyright holder sells the product line to another company, one change is recorded in the directory and all subsequent users will be sent to the new site. Information about the object (metadata) is maintained by the publisher (or potentially a specialist third party) in databases; it might include the actual content or the information on where and how to obtain the content or other related data. Potentially, the DOI will also serve as an agent for automated transactions such as buying a subscription, downloading a file, joining a forum, looking up a reference in a database, etc.

Identifiers for multimedia objects are discussed elsewhere in this issue. In a digital world, distinctions between the different types of digital objects dissolve; each is merely a "digital object." Identification schemes devised in one area of content will need to become interoperable with those from other areas. As such, an ideal enabling technology would provide a common framework which would allow the building of interoperable identification schemes.

IX. The Way Forward

Unique identifiers for future application to intellectual property (including its expression as digital objects) will take several themes for a solution. The end point of design is to create a system with intelligence; as we have seen, that intelligence can come from several points: the identifier; some specific functional metadata; or a system which establishes formal relationships between structured entities and well-formed metadata.

If an identifier scheme is being designed for very restricted use in a closed user group, with a guarantee that no other applications are needed, it may be feasible to adopt the approach of making the identifier intelligent to the extent that each carries its own metadata; the identifier system will be relatively simple to design but very limited in potential. If that is not possible, but the envisaged use is still one single application, it may be sufficient to use dumb identifiers plus some function-specific metadata (like DC simple for resource discovery); again, this may work well for that application if not for other uses. But it seems likely that both these scenarios are likely to be untenable in the future. Increasingly, metadata will be created once but used and accessed by anyone; the designer cannot know the applications to which it will be put. There is a spectrum of increasing flexibility and interoperability as move through these various designs from intelligent identifiers with no metadata, to full, well-formed, metadata. The way forward is to design systems on the following basis.

1) *Unique "Dumb" Identification:* Unambiguous simple identification of a defined piece of information; dumb identifiers, not hard wired with any specific application intelligence.
2) *"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. Using schemes such as RDF, metadata can be structured in an extensible way. A common means of expressing high-level data models for structural metadata to facilitate interoperability in many different functional applications.
3) Support for arbitrary levels of granularity.

4) Multiple, co-existing identification 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 (e.g., a sound clip within a digital object may be identified by a music identifier as well as being part of a document with another identifier). This point implies extensibility, i.e., the ability to declare within a scheme the particular namespace used for that element (as in the URN model).
5) *Links to Distributed Metadata:* Via simple 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.
6) *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); this requires a careful management structure design for the operation of the identifier registration process.

Common themes which seem to be emerging are structure and extensibility, which are defining schemes which can plug in to the infinite set of data and metadata by declaring their applicable entity/attribute schemas.

However, providing schemes which solve all the technical issues described here is a necessary but not sufficient condition for the success of a unique identifier system. To become an accepted and widely implemented scheme, business issues must also be dealt with. The history of earlier initiatives such as ISBN [29] demonstrate that several years may occur from the conception of a system to its commercial realization, even (as with ISBN) where the industries concerned fully support the creation of the system. In the case of Internet identifiers, for example, the current URN implementations leave open the option of many different naming authorities and schemes; while it is hoped that this will result in “a small number of high quality naming schemes” [1], there is no guarantee of this. A multitude of identifier schemes is hardly an improvement on the current situation; one can easily imagine disputes as to administrative systems for connection of different schemes, or jurisdiction. A major challenge in moving toward systems which are practical and universal is therefore to provide administrative solutions (or businesses) which solve political ownership issues (the antitrust and property rights issues) which meet the needs of a variety of interested parties, which provide guaranteed ability in perpetuity to resolve identifiers to resources, and which can be effectively administered so as to provide minimal barriers to common use. If all of these issues can be solved, there will be strong levels and support and commitment from those using the system.

It is worth noting that unique identification of intellectual content entities is only one step in the process necessary to create a digital rights trading infrastructure. In addition, we need to uniquely identify both the interested parties who are the participants in a transaction, and the agreement which governs the terms of the transaction of the entity between the parties (as is the aim of the CIS model). These three groups of identification have been described as the three legged stool of e-commerce.

As a first step, the initiatives from the technical infrastructure viewpoint (such as URN’s) must meet with initiatives from content providers. Since the owner of a namespace does not have to be the entity that manages the operation of resolution, it should be possible to partition responsibilities for a complete system among parties with appropriate expertise. So, for example, an organization like ISBN International (with expertise in namespace management) might work with a URN implementation by a technology company. The DOI is a notable attempt to provide one such solution, by bringing together a technical URN-compatible implementation with an administrative scheme having the support of the publishing and other content communities, aiming to resolve commercial business issues of the technical solution. The International DOI Foundation was established following the launch of the DOI (at the Frankfurt Book Fair, October 1997) to “support the needs of the intellectual property community in the digital environment” [14]. The Foundation has an intensive program of development of the DOI as a standardized solution for the intellectual property communities (including text, music, images, and multimedia), active involvement with related identifier discussions and systems and digital infrastructure developments, and the appointment of service providers for the efficient operation of the technology and business activities of the DOI system. A goal of the foundation is to engage members in active involvement in defining policies and solutions; membership is drawn from both technical and content communities, being open to “all who wish to participate in shaping the infrastructure for electronic publishing and information dissemination and ensuring the effective respect of copyright.”

Acknowledgment

Many of the ideas expressed in this paper have been suggested or influenced by discussion with the author’s colleagues in various activities relating to digital information identifiers. The author would like to thank in particular B. Arms of CNRI, M. Bide of Mark Bide Associates, L. Daigle of Bunyip Systems, L. Lannom of CNRI, and G. Rust of Data Definitions.

References

[1] W. Arms, L. Daigle, R. Daniel, D. LaLiberte, M. Mealling, K. Moore, and S. Weibel. (1996, Apr.). Uniform resource names: A progress report. D-Lib magazine. [Online]. Available WWW: http://www.dlib.org/dlib/february96/02arms.html.

[2] W. Y. Arms, C. Bianchi, and E. A. Overly. (1997, Feb.). An architecture for information in digital libraries. D-Lib magazine. [Online]. Available WWW: http://www.dlib.org/dlib/february97/cnri/02arms1.html.

[3] T. Berners-Lee. The myth of names and addresses (axioms of web architecture: 2). [Online]. Available WWW: http://www.w3.org/DesignIssues/NameMyth.html.

[4] Cooperative Association for Internet Data Analysis (CAIDA). Background CAIDA. La Jolla, CA. [Online]. Available WWW: http://www.caida.org//Caida/background.html..

[5] R. D. Cameron. (1997). Toward universal serial item names. [Online]. Available WWW: http://elib.cs.sfu.ca/USIN/USIN.html.

[6] P. Caplan, "You call it corn, we call it syntax-independent metadata for document-like objects," *The Public-Access Comput. Syst. Rev.*, vol. 6, no. 4, pp. 19–23, 1995. [Online]. Available WWW: http://info.lib.uh.edu/pacsrev.html.

[7] Cataloguers toolbox: Bibliographic control numbers. [Online]. Available WWW: http://www.mun. ca/library/cat/bibcontr.htm.

[8] *Data Processing—Check Character Systems*, International Standard ISO 7064, 1983.

[9] L. Chiariaglione. (1996). Communication standards: Gotterdamerung? [Online]. Available WWW: http://drogo.cselt.it/ufv/leonardo/paper/standardization.html.

[10] Consortium for the Computer Interchange of Museum Information. Introduction. [Online]. Available WWW: http://www.cimi.org/about/introduction.html.

[11] L. Daigle, R. Daniel, and C. Preston, "Uniform resource identifiers and serials," *Serials Librarian*, vol. 33, nos. 1–4, pp. 325–341, 1998.

[12] D. P. Dern and S. Mace, "The internet reinvented," *Byte*, pp. 89–96, Feb. 1998.

[13] U.S. Dept. of Commerce. (1998). A proposal to improve technical management of internet names and addresses. [Online]. Available WWW: http://www.ntia.doc.gov/ntiahome/domainname/dnsdrft.htm.

[14] International DOI Foundation. About the DOI. [Online]. Available WWW: http://www.doi.org/about_the_doi.html.

[15] Dublin Core Metadata Initiative. [Online]. Available WWW: http://purl.oclc.org/dc/.

[16] EAN International. European Article Numbering. [Online]. Available WWW: http://www.ccib.org.lb/ccib/ean.htm.

[17] U.S. Census Bureau. FGDC Subcommittee on Cultural and Demographic Data (SCDD). Working draft of the content standards for digital geospatial metadata: Thematic supplement for geospatially referenced cultural and demographic data metadata. [Online]. Available WWW: http://www.census.gov/geo/www/standards/scdd/CDsupplement.html.

[18] M. Roscheisen and T. Winograd. The FIRM framework for interoperable rights management. [Online]. Available WWW: http://mjosa.stanford.edu/˜roscheis /IMA/index.html.

[19] S. L. Garfinkel, "The web's unelected government," *Technol. Rev.*, vol. 101, no. 6, pp. 38–47, Nov./Dec. 1998.

[20] B. Green and M. Bide. (1997). Unique identifiers: A brief introduction. BIC (Book Industry Communication). [Online]. Available WWW: http://www.bic.org.uk /bic/uniquid.html.

[21] K. Hafner and M. Lyon, *Where Wizards Stay Up Late: The Origins of the Internet*. New York: Simon & Schuster, 1996, pp. 246–250.

[22] S. X. Sun. (1997). Handle system: A persistent global naming service (overview and syntax). [Online]. Available WWW: http://www.handle.net/draft-ietf-handle-system-01.html.

[23] K. Hill, "CIS—A collective solution for copyright management in the digital age," *Copyright World*, vol. 76, pp. 18–25, Dec. 1996/Jan. 1997.

[24] U.S. House, Bill US HR 2281.

[25] International Federation of Library Associations and Institutions. Study Group on Functional Requirements of Bibliographic Records. (1998). Functional requirements for bibliographic records. [Online]. Available WWW: http://www.ifla.org/VII/s13/projects.htm.

[26] INDECS. INDECS overview. [Online]. Available WWW: http://www.indecs.org/overview/overview.htm.

[27] G. Irlam. (1995). Naming. [Online]. Available WWW: http://www.base.com/gordoni/naming.html.

[28] ISO/TC 46 /SC 9 Working Group 1. International standard audiovisual code (ISAN). [Online]. Available WWW: http://www.nlc-bnc.ca/iso/tc46sc9/isan.htm.

[29] H.-J. Ehlers, "Identification numbering in the book," *Library Inform. World ISBN Rev.*, vol 15, pp. 89–214, 1994.

[30] *Documentation—International Standard Recording Code (ISRC)*, International Standard ISO 3901, 1986.

[31] International standard work code ISO 15707 (ISWC) ISO/TC 46/SC 9 working group 2. [Online]. Available WWW: http://www.nlc-bnc.ca/iso/tc46sc9/iswc.htm.

[32] R. E. Kahn and R. Wilensky. (1995). A framework for distributed digital object services. [Online]. Available WWW: http://www.cnri.reston.va.us/home/cstr/arch/k-w.html.

[33] M. C. Kelly, "The role of A&I services in facilitating access to the e-archive of science," *ICSTI Forum*, no. 26, pp. 1–4, Nov. 1997. [Online]. Available WWW: http://www.icsti.nrc.ca/icsti/forum/fo9711.html#role.

[34] J. Martin, *Strategic Data Planning*. Englewood Cliffs, NJ: Prentice-Hall, 1982.

[35] J. S. Miller. (1996). W3C and digital libraries. [Online]. Available WWW: http://www.dlib.org/dlib/november96/11miller.html.

[36] E. Miller. (1998, May). An introduction to the resource description framework. *D Lib Mag.* [Online]. Available WWW: http://www.dlib.org/dlib/may98 /miller/05miller.html.

[37] N. Paskin, "Information identifiers," *Learned Publishing*, vol. 10, no. 2, pp 135–156, 1997. [Online]. Available WWW: http://www.elsevier.nl/locate/infoident.

[38] ——. (1997). Digital information objects and the STM publisher. STM Annual Report. [Online]. Available WWW: http://www.elsevier.nl/locate/diginfo.

[39] American Chemical Society, American Institute of Physics, American Physical Society, Elesevier Science, IEEE. (1995, May). Publisher Item Identifier as a means of document identification. [Online]. Available WWW: http://www.elsevier.nl/inca/homepage/about/pii/.

[40] N. F. Pope. (1998). Identifiers' multiple roles: Complementary or contentious. [Online]. Available WWW: http://www.library.wisc.edu/libraries/issues/doiroles.98.html.

[41] OCLC. Persistent uniform resource locators. (1998). [Online]. Available WWW: http://purl.org.

[42] W3C: Resource description framework. [Online]. Available WWW: http://www.w3.org/RDF/.

[43] R. Reynolds. (1998). ISSN and seriality. [Online]. Available WWW: http://lcweb.loc.gov/acq/conser /serialty/issn.html.

[44] R. Moats. URN syntax. IETF RFC 2141. [Online]. Available WWW: http://www.ietf.org/rfc/rfc2141.txt.

[45] R. Daniel. (1997). A trivial convention for using HTTP in URN resolution RFC 2169 internet engineering task force. [Online]. Available WWW: http://www.ietf.org/rfc/rfc2169.txt.

[46] C. Lynch, C. Preston, and R. Daniel. (1998). Using existing bibliographic identifiers as uniform resource names. [Online]. Available WWW: http://www.normos.org/ietf /rfc/rfc2288.txt, 1998.

[47] J. Rowley, *Organizing Knowledge: An Introduction to Information Retrieval*, 2nd ed. London, U.K.: Ashgate, 1992.

[48] G. Rust, (1998, July). Metadata: The right approach, an integrated model for descriptive and rights metadata in e-commerce. D. Lib. Magazine. [Online]. Available WWW: http://www.dlib.org/dlib/july98/rust/07rust.html.

[49] *Serial Item and Contribution Identifier*, ANSI/NISO Z39.56-1996 (Version 2), 1997.

[50] SICI generator. [Online]. Available WWW: http://www.ep.cs.nott.ac.uk/˜sgp/sicisend.html.

[51] M. Stefik, "Letting loose the light," in *Internet Dreams: Archetypes, Myths, and Metaphors*, M. Stefik, Ed. Cambridge, MA: MIT Press, 1997, pp. 219–2537.

[52] ——, *The Digital Property Rights Language Manual and Tutorial, Version 1.08*, Xerox Palo Alto Res. Center, Palo Alto, CA, Feb. 3, 1997.

[53] M. St. Pierre and W. La Plant. (1998). Issues in crosswalking content metadata standards. [Online]. Available WWW: http://www.niso.org/crosswalk.html.

[54] J. Thacker, "Notes prepared for a meeting hosted by the National Information Standards Organization on June 18, 1997," unpublished.

[55] Uniform Code Council, Inc. ID numbers and bar codes. [Online]. Available WWW: http://www.uc-council.org/main/ID_Numbers_and_Bar_Codes.html.

[56] D. Connolly. (1998). Names and addressing: URI's. [Online]. Available WWW: http://www.w3.org /Addressing/Addressing.html.

[57] K. Sollins and L. Masinter. (1994). Informational request for comments: 1737. Functional requirements for uniform resource names. [Online]. Available WWW: http://ds.internic.com/rfc/rfc1737.txt.

[58] Library of Congress Network Development and MARC Standards Office. (1998). MARC standards. [Online]. Available WWW: http://www.loc.gov/marc.

[59] WIPO. WIPO copyright treaty, adopted by the Diplomatic Conference on Dec. 20, 1996, CRNR/DC/94. [Online]. Available WWW: http://www.wipo.org/eng /diplconf/distrib/94dc.htm.

[60] World Intellectual Property Organization Internet Domain Name Process. [Online]. Available WWW: http://wipo2.wipo.int/process/eng/processhome.html.

[61] Internet Domain Survey (monthly). [Online]. Available WWW: http://www.nw.com/zone/WWW /report.html.

[62] Cross Industry Working Team. Managing access to digital information: An approach based on digital objects and stated operations. [Online]. Available WWW: http://www.xiwt.org/homepage, 1997.

[63] T. Kuny, Ed. Digital Libraries: Metadata Resources. [Online]. Available WWW: http://www.ifla.org/II/metadata.htm.

[64] W3C Consortium. Naming and addressing: URIs, URLs. [Online]. Available WWW: http://www.w3.org/Addressing/.

[65] T. Berners-Lee, R. Fielding, U. C. Irvine, L. Masinter. (1998, Aug.). Uniform resouce identifiers (URI): Generic syntax IETF RFC 2396. [Online]. Available WWW: http://www.ietf.org/rfc/rfc2396.txt.

Norman Paskin received the Ph.D. degree in biochemistry from the University of Nottingham, UK.

He became the first Director of The International Digital Object Identifier (DOI) Foundation in March 1998. Prior to this, he was with Elsevier Science, where he held a number of editorial, management, and technology roles, most recently as Director of Publishing Technologies. He has worked for 20 years in the scientific publishing industry in both the United States and Europe. He was actively involved in information identifiers issues for the scientific, technical, and medical (STM) publishing community, and he has published several papers on this topic prior to his role with the DOI. His involvement with technology and information began with computer modeling using literature data in 1976. From 1994 to 1998, he was Director of Information Technology Development for Elsevier Science.