On the utility of identification schemes for digital earth science data: an assessment and recommendations

Ruth E. Duerr • Robert R. Downs • Curt Tilmes •
Bruce Barkstrom • W. Christopher Lenhardt •
Joseph Glassy • Luis E. Bermudez • Peter Slaughter

Received: 7 February 2011 / Accepted: 24 June 2011 / Published online: 16 July 2011
© Springer-Verlag 2011

Abstract In recent years, a number of data identification technologies have been developed which purport to permanently identify digital objects. In this paper, nine technologies and systems for assigning persistent identifiers are assessed for their applicability to Earth science data (ARKs, DOIs, XRI, Handles, LSIDs, OIDs, PURLs, URIs/URNs/URLs, and UUIDs). The evaluation used four use cases that focused on the suitability of each scheme to provide Unique Identifiers for Earth science data objects, to provide Unique Locators for the objects, to serve as Citable Locators, and to uniquely identify the scientific contents of

data objects if the data were reformatted. Of all the identifier schemes assessed, the one that most closely meets all of the requirements for an Unique Identifier is the UUID scheme. Any of the URL/URI/IRI-based identifier schemes assessed could be used for Unique Locators. Since there are currently no strong market leaders to help make the choice among them, the decision must be based on secondary criteria. While most publications now allow the use of URLs in citations, so that all of the URL/URI/IRI based identification schemes discussed in this paper could potentially be used as a Citable Locator, DOIs are the

Communicated by H.A. Babaie

Electronic supplementary material The online version of this article (doi:10.1007/s12145-011-0083-6) contains supplementary material, which is available to authorized users.

R. E. Duerr (✉)
National Snow and Ice Data Center,
University of Colorado at Boulder,
Boulder, CO 80309-0449, USA
e-mail: rduerr@nsidc.org

R. R. Downs
Center for International Earth Science Information Network
(CIESIN), Columbia University,
202 Geoscience, Lamont-Doherty Earth Obs.,
Palisades, NY 10964, USA
e-mail: rdowns@ciesin.columbia.edu

C. Tilmes
NASA Goddard Space Flight Center,
Greenbelt, MD 20771, USA
e-mail: Curt.Tilmes@nasa.gov

L. E. Bermudez
Open Geospatial Consortium (OGC),
483B Carlisle Drive,
Herndon, VA 20170, USA
e-mail: lbermudez@opengeospatial.org

P. Slaughter
Earth Research Institute,
University of California at Santa Barbara,
Santa Barbara, CA 93106-3060, USA
e-mail: peter@eri.ucsb.edu

B. Barkstrom
NASA/NOAA,
Asheville, NC 28804, USA
e-mail: brbarkstrom@gmail.com

W. C. Lenhardt
Oak Ridge National Laboratory,
Oak Ridge, TN, USA
e-mail: ledhardtc@ornl.gov

J. Glassy
R&D, Lupine Logic Inc.,
1121 East Broadway St, Suite 128,
Missoula, MT 59802, USA
e-mail: jglassy@lupinelogic.com

identification scheme currently adopted by most commercial publishers. None of the identifier schemes assessed here even minimally address identification of scientifically identical numerical data sets under reformatting.

Keywords Digital identifiers · Unique Identifiers · Permanent identifiers · Global unique persistent identifiers

Introduction

The problem of identity has vexed humanity throughout all of recorded history. A wide variety of methods; from assigned identifiers to taxonomic techniques and beyond; have historically been used to resolve the issue of whether this thing, whatever or whomever it may be, is what it purports to be. Yet none have ultimately proved to be flawless. Not surprisingly then, the issue of identity is just as much an issue in this digital era if not more than it has been. Given the mutability of digital objects it would be surprising indeed if it were not more of an issue. This presents a quandary for science given its foundations in the concept of repeatability. How can one repeat what cannot be identified? In the Earth sciences the problem is even more acute. Unlike other fields of research, the majority of observations in the Earth sciences are not repeatable. They are made at a distinct place and time whose circumstances cannot be exactly repeated, in contrast with laboratory experiments. Thus, the observations are unique and irreplaceable. One would think that this uniqueness would make identification easier; yet the realities of current scientific practice and technology means that it just isn't so.

Not surprisingly then, a number of identification schemes have been implemented by various communities; academic, commercial, and non-profit. Many of these schemes purport to be the answer to the question of identification, at least for that community (Garritty et al. 2009). But is this so? Or at least is this the case for the Earth Sciences? That is the question that was posed to NASA's Earth Science Data Systems Technology Infusion Working Group (hereafter called the TIWG) and the Federation of Earth Science Information Partners (ESIP) Data Stewardship and Preservation cluster (hereafter called the Stewardship cluster) shortly after it was formed in 2008. This paper addresses this question.

Consideration of this question led the authors to define four primary use cases. Answering it led them to restrict the assessment to only two classifications of data, as well as to generation of a number of criteria against which each identification scheme could be assessed. After discussing

You are catching up with the literature in your field and notice a paper describing an interesting feature in the MODIS Land Surface Temperature product over Patagonia in 2010. You think that is strange because you've been working with the same data, did not notice the feature the article discusses, and the results of your research would be impacted if the feature truly exists. Moreover, when you go back to look at the data again, you still don't see the feature described. Thinking that perhaps the paper was discussing some other version of the data, you contact the author to find out what data they were using and they provide the identifier for the exact data they used.

Fig. 1 Typical scenario demonstrating the need to able to identify current data

reasons for using identifiers, this paper considers nine different identifier schemes in terms of:

- Naming syntax,
- Technical value of the scheme,
- User value,
- Value to archives or data centers, and
- Use case support.

The section after this discussion contains the results of the assessments. The paper concludes with a brief discussion and summary of the results, and a summary of future work.

Reasons for using unique identifiers

There are several reasons why methods for identifying data are important. Figures 1 and 2 provide example scenarios for motivation.

While working on a project to understand the long-term trends in the mass balance of glaciers in Alaska's Kenai peninsula, you undertake a literature search and find a paper describing lidar measurements made in the area 40 years ago that would be helpful to have to compare to more recent measurements. The paper claims that the data were deposited at the World Data Center-A for Glaciology and cites the identifier used by the WDC for these data. Unfortunately, the World Data Center system was disbanded and reconstituted early in the 21st century and responsibility for the holdings of the WDC-A for Glaciology was migrated to a variety of other institutions and agencies several times over the intervening years. You'd like to obtain a copy of these lidar measurements.

Fig. 2 Typical scenario demonstrating the need to identify data over a long period of time

The first reason is to provide a user with simple capabilities for uniquely and unambiguously identifying the data of interest (Parsons et al. 2010). Digital data are not immutable. Scientifically interesting data may have been processed and reprocessed many times using different algorithms for a variety of scientifically meaningful reasons. For example, simply stating that AVHRR Brightness Temperature data were used in a study does not provide enough information for another scientist to use the identical data set, much less identical files, in replicating the original study. The fact that there were multiple AVHRR instruments on multiple spacecraft, each with unique instrument characteristics, complicates the issue even further. In fact, since older versions of the data may be replaced by later versions, the actual data themselves have often been deleted due to space or cost concerns, although the need for identification and reference remains. There may be value to maintaining identifiers that resolve to metadata including provenance and context information for a data product used in scientific research even if the data is no longer available, although it may be that the scientific goal of being able to replicate original results is then impossible.

The second reason for having a unique identifier is to help users find and access the data regardless of where they currently reside or how the responsibility for managing the data has changed over time. Considering the resource requirements for data management, it may be necessary to transfer data between archives over time to ensure their preservation. If data identifiers are only unique within each archive or data center, then the identifiers at some archives could be the same as the identifiers assigned by other archives. However, to maximize the utility of the data, they still need to be discoverable and accessible to users regardless of how responsibility for the data migrates over time.

The third reason for having unique identifiers is to facilitate management of data over time. As technology evolves, so do the data management systems based on those technologies. As a result the identifier assigned to a particular data set could also change. Over time, the accumulation of such changes can become quite complex as the identifiers reflect the operating systems and applications used in each era. Moreover, the titles assigned to a data set may also morph over time. For example, a time series may be extended or as data from an additional suite of ground stations may be added to a compilation. Such changes make it difficult to ensure that a data set referenced decades ago is the same data set as one referenced today.

The fourth reason for having unique identifiers is to facilitate data citation in publications. Data citation serves at least two purposes. First, as mentioned earlier, the ability to precisely identify the data used in original research is

fundamental to being able to reproduce research results and is therefore key to the scientific method. Second, as noted by the International Polar Year Data and Information Service (IPYDIS) “by encouraging proper citation of data sets, data providers and publishers receive appropriate credit for their efforts, the perception of data management as a discipline improves, and it is easier to track the use and impact of the data” (How to Cite a Dataset 2008).

Use cases

In the earth sciences four use cases are quite common. While others cases exist, the authors considered these to be most useful for differentiating between identifier schemes and will be used in discussing the results of the assessments. The use cases are:

- Unique Identifier: To uniquely and unambiguously identify a particular piece of data, no matter which copy a user has
- Unique Locator: To locate an authoritative copy of the data no matter where they are currently held
- Citable Locator: To identify cited data
- Scientifically Unique Identifier: To be able to tell that two data instances contain the same information even if the formats are different

Each of these use cases will be discussed in greater detail in the sections below.

Unique identifier

In order to ensure the continued viability of any digital object, there need to be multiple copies of that object, ideally scattered broadly around the world. This truism is even the centerpiece of common digital library systems such as LOCKSS (Maniatis et al. 2005). In the earth sciences, reuse of data is relatively common. In simple reuse, the user copies the original data resulting in potentially many more copies of an object. Ideally, an identifier for the original data would also be copied in order to ensure traceability to the original data source. Such an identifier would need to be location independent. In other words, every copy of the object no matter where it was located would be identified by and potentially would contain that identifier. However, it is easy for data users to rearrange or reformat data, so that unique identifiers that assume no rearrangement or reformatting may not be sufficient for all scenarios. We defer discussion of this complication to the last use case.

For individual files, data producers often use filenames for identifiers. However, because these names are easily

changed, this identifier approach appears less than ideal. Ideally, an identifier would be:

- Generated at the time the object is created;
- Placed within the object itself; and
- Referenced within related information about the object, such as an associated metadata file.

Such an identifier could be used in databases or other catalog systems as the unique key for that object. Such identifiers should be difficult to change once established, suggesting that the producer should embed them within the data object. Given the importance of reproducibility and verifiability, such identifiers could have additional trust value beyond that of a simple identifier so they could be used as a part of a verification process. The fundamental operational concept for such an identifier is that it is *created once and never modified thereafter*.

Practically speaking, this implies that such an identifier needs to be globally unique yet creating one should not require any naming authority beyond that of the original data producer. The point being that any data producer anywhere, at any time, including an investigator working remotely in the field, should be able to generate an identifier for the data they are producing at that time.

Unique locator

If data, either individual files or collections of files, are to be useful for science they need to be discoverable. Given an identifier used at one time, it must be possible to find the same data at any later time. Yet, history has shown that custodianship of scientific data changes over time. For example, NASA Earth Science mission data typically starts at an investigator-led production facility, moves to an active archive, and should be archived at a designated permanent repository (NASA 2006). Even permanent repositories have been known to move over time as the fortunes of their parent organizations rise and fall. Thus unlike the location independent identifiers ideally suited to the previous use case, identifiers used for this purpose must be location invariant, they point to an object; but are not necessarily located within or are a part of the object. Moreover, rather than pointing to just any or every copy of the object, they must point to the currently trusted source(s) for authentic copies of the data.

Sometimes data are reprocessed using later versions of algorithms or calibration that are intended to improve their quality. Commonly, older versions of the data are viewed as obsolete and are no longer recommended for future research. Since there is a cost to maintaining data, they may be deleted to make room in an archive for data that may be viewed as more valuable. Identifiers should still

reference the old data, especially if they were used in published works. To preserve the ‘persistent’ quality of the ‘locator’ aspect of these identifiers, they should still be resolvable even if the data have been deleted. This resolution usually takes the form of a ‘landing page’ pointing out the fact that the data have been deleted, and directing the requestor to the version of the data that replaced the original one. Ideally, the reference should include metadata, which is typically much smaller than the data, and thus less of a burden for an archive to maintain. This metadata should include provenance and context information for the missing data, so that a determined investigator might be able to reproduce the missing data if fortune allows it.

Like the previous use case, these location invariant identifiers need to be globally unique. However, unlike the previous use case, external naming authorities beyond the original producer are not only necessary but also useful for dealing with data whose custody may move from one organization to another at some point in the future. Having an external naming authority minimizes the possibility that the dissolution of an archive or its parent organization will make it impossible to continue using existing identifiers when data transitions to a new repository. Also, to avoid generating identifiers that may be abandoned in the future, these identifiers should not be generated until an archival authority makes the decision to make the data permanently available. Lastly, the operational concept for these identifiers is that *once they are created, they will require ongoing maintenance into the future*.

Citable locator

While a complete discussion of data citation is beyond the scope of this paper, calls for researchers to cite the data used in their work are increasing. For example, the American Geophysical Union (AGU) Council has recently asserted that data should “be credited and cited like the products of any other scientific activity” (AGU Council 2009). Consequently Earth science needs concise mechanisms to identify the data used in research, particularly identifiers that researchers can incorporate into a standard citation. Once incorporated into a publication, such an identifier should lead readers at any time in the future to the exact data used in the work that led to the publication.

Such identifiers are very similar to those in the previous use case. They should be location invariant, pointing to the currently accepted authoritative source(s) of the data and should be globally unique. Having an external name authority is an advantage in case the originating organization terminates operations, and ongoing maintenance can be expected once the identifier is created. However, such

identifiers need to be broadly accepted and used by journal publishers and would ideally be the same type of identifier as publishers use themselves to identify articles and publications. In addition, it may not be practical to cite each and every one of the potentially hundreds of thousands or millions of files used in a particular research project, although such precise identification may be needed for verification of results. However, for providing credit to the data producers, such identifiers could facilitate identification of entire data sets. Identification of specific subsets of a larger data set may need further research.

Scientifically unique identifier

The fourth use case, which requires the ability to tell that two data instances contain the same information even if their formats are different, is somewhat different than the other three use cases. Dealing with this case requires more detailed consideration of the mechanisms under which information technology can modify the bit-for-bit representation of data while retaining its scientific equality. For example, conversion of integer values to floating point ones does not usually change their equivalence. Likewise, rearranging the array ordering, from row-major to column-major, does not affect the ability of a researcher to recognize the data in the rearranged format as identical with the original—as long as he or she knows about the mapping. Whether data from a temperature record are represented as ASCII text in a csv file or as binary data or as any of a host of other data formats is irrelevant to the conclusions that can be drawn from an assessment of change over time of the data (Tilmes et al. 2010).

While Altman (2008) and others are contributing to the development of Universal Numeric Fingerprints (UNFs), this approach appears to rely on some format authority to provide a canonical format for Earth science data. In the probable absence of such an authority, it appears likely that numerical Earth science data “objects” are not uniquely identifiable and immutable. Rather, they appear likely to be members of “equivalence classes” that will require a rather more sophisticated treatment than the ones being considered for the other three use cases.

Ignoring this quandary for the moment, this use case is very similar to that of the Unique Identifier use case. If such an identifier were to exist, it must be location independent, would need to be generated when the object is initially created and regenerated or revalidated each time the object is copied, would move as the object is moved, should be placed within the object, but referenceable from within related objects such as metadata files or database records, ideally should have additional trust value, should not require a universal name authority and must be globally unique.

Types of data needing identifiers

This paper discusses data at only two levels of aggregation where identifiers are needed. These are

1. a data set as a whole, by which we mean a set of data described as a unit in a metadata record, such as a record in the Global Change Master Directory
2. an individual object within a data set, such as a file or a granule, the latter being the smallest entry in an archive’s inventory.

This paper does not consider identifiers for other aggregation levels, such as data collections from different instruments, different versions of data from the same instrument, data generated on demand, subsets of a data set that themselves consist of many files, or individual data base records or record fields.

Assessment criteria

Based on an analysis of the reasons for using identifiers (see section above) and a series of discussions within the TIWG and the Stewardship Cluster, a total of 14 criteria were generated against which to assess each of the identifier schemes. These criteria fall roughly into three categories:

1. *Technical value*, assessing the technology underlying the identifier scheme;
2. *User value*, assessing how the scheme is likely to make end user’s lives easier or more difficult; and
3. *Archive value*, assessing whether an identifier scheme is likely to make the job of data management easier or more complicated.

Table 1 provides a more detailed list of questions within these three criteria.

Assessments

Based on the assessment criteria in Table 1, the authors analyzed nine persistent identifier technologies and systems (in alphabetic order)

- Archival Resource Keys (ARKs)
- Digital Object Identifiers (DOIs)
- Extensible Resource Identifiers (XRIIs)
- Handle System (Handles)
- Life Science Unique Identifiers (LSIDs)
- Object Identifiers (OIDs)
- Persistent Uniform Resource Locators (PURLs)

Table 1 Assessment criteria for technology to assign persistent identifiers to earth science data

Category	Criterion
Technical value	<ol style="list-style-type: none"> 1. Is it recommended by any standards bodies? If so, which ones? 2. Are there any security issues with it? 3. How scalable is it to very large numbers of objects? 4. Is it interoperable with other identification schemes? 5. How compatible is it with the main internet naming schemes? 6. Does it require third party maintenance (e.g., a registry and a registry maintainer?) 7. How dependent is it on a naming authority and how do you know that naming authority will be in business decades or more into the future? 8. What is the expected longevity of the underlying technologies? How long will the scheme remain viable?
User value	<ol style="list-style-type: none"> 1. Will publishers allow it in a citation? 2. Does the identification scheme have any additional trust value? 3. Does the identifier have meaning? (Should identifiers be transparent or opaque?)
Archive value	<ol style="list-style-type: none"> 1. How maintainable is the identification scheme when data migrates from one archive to another (or even from one location in an archive to another)? 2. Are there any actual charges or costs involved (e.g., DOI's may cost only a few cents each; but, even a small archive may have dozen's of millions of files that should have identifiers)? 3. Does the identification scheme handle data that is not on the web? What about physical objects?

- Uniform Resource Identifiers/Names/Locators (URIs/URNs/URLs)
- Universally Unique Identifiers (UUIDs)

The results of each assessment are described below.

We do note one common aspect of the identifier schemes that use Web addresses: potential security issues related to the registry or naming authority organizations. While properly formatted Web addresses do not contain scripts and do not appear to be particularly vulnerable to attack, the authors did not attempt to assess IT threat scenarios against registry or naming authorities as part of our evaluation.

Archival resource key

Developed by the California Digital Library in 2003, an Archival Resource Key (ARK) is a Uniform Resource Locator (URL) specifically intended to serve as a long-term persistent identifier (Kunze 2003). The key premise underlying the ARK is that object persistence depends on specific commitments by the provider, which are included in the metadata of the identifier, and not on any particular object naming conventions. As a consequence, an ARK URL connects a user to three things: an “object, its metadata, and a statement” by its provider of their commitment to preservation of that object (Kunze and Rodgers 2008).

Naming syntax

An ARK contains five parts: [http://NMAH/]ark:/NAAN/Name[Qualifier]

Where, according to Kunze and Rodgers (2008):

- NMAH is an optional and mutable Name Mapping Authority Hostport (usually a hostname)
- The “ark:” label,
- The Name Assigning Authority Number (NAAN),
- The assigned Name, and
- An optional and possibly mutable Qualifier supported by the Name Mapping Authority

Kunze and Rodgers also state that a “NAAN and Name together form the immutable persistent identifier for the object independent of the URL hostname” and used in “a Web browser, the ARK leads the user to the named object” (2008). Appending a question mark to that same ARK returns a human and machine-readable metadata record about that object, while appending two question marks returns the provider’s commitment statement. A variety of tools are available for creating, binding, and resolving ARKs (Kunze and Rodgers 2008).

An example ARK: <http://ark.cdlib.org/ark:/13030/tf5p30086k>

The first part, “<http://ark.cdlib.org/>”, is not used for comparing ARKs for equivalence, and can be replaced at any time. It is in fact optional, so a synonym considered equivalent to the example object would be simply represented as “ark:/13030/tf5p30086k”.

Technical value

While the ARK specification was submitted to the Internet Engineering Task Force (IETF), the submission expired

without being accepted. There are no known security issues associated with ARKs and the ARK specification is highly scalable to large numbers of objects. The ARK standard is interoperable with other systems since each ARK is simply a URL and can be treated as such.

The California Digital Library (2010) currently sponsors the central authority to register and assign NAANs. Presently, a total of 27 Name Authorities exist including Google, the Internet Archive, the World Intellectual Property Organization, and the Digital Curation Centre. In addition, “the ARK specification describes a lookup algorithm for finding a new NMAH” (Kunze 2003) for an ARK if a user ever discovers one that isn’t available (Kunze and Rodgers 2008). The underlying technologies are simple and free. As a result, the scheme is expected to remain viable as long as organizations keep using it.

User value

URLs are recognized elements of a bibliographic citation (Chicago Manual of Style 2006; Publication Manual of the American Psychological Association 2001). As a result, ARKs can be cited like any other URL. On the other hand, ARKs have less acceptance by scientific journal publishers than do DOIs.

ARKs have significant additional trust value, as simply using an ARK is a strong indication that the organization recognizes the need for persistence. Every ARK includes an accessible reference to a nuanced policy statement by the archive steward regarding their commitment toward maintenance of the object.

Archive value

The ARK specification was purposely designed to facilitate migration of objects from one archive to another or to different locations within an archive. The main part of the identifier includes a reference to the naming authority and a name recognizable by that authority. The addition of the NMAH makes the identifier into a complete URL. If the NMAH becomes invalid, a centralized lookup can discover a new one. This makes it easy for all the contents under the responsibility of a given naming authority to move from place to place.

The name portion of an ARK is opaque and could be encoded using any other scheme. The content of the name is controlled by the naming authority and could be used to expose information, though the standard discourages this practice. The standard also includes mechanisms for expressing hierarchical nesting of objects and multiple expressions of an object (in multiple formats/languages/etc.)

There are currently no charges associated with ARKs themselves though organizations can contribute their names

to the N2T (Name-to-Thing Resolver 2007) resolution service for \$30/year. ARKs can be used to identify anything—not just objects available on the web.

Use case support

As an URL-based identification system ARKs are location invariant not location independent. They do require a name authority. As such they do not meet the full requirements for a Unique Identifier (see [Summary Discussion](#) section). However, like other URL-based identification schemes, ARKs would be acceptable Unique Locators. ARKs are not widely used within the publishing community, so are not optimal as a Citable Locator. ARKs in and of themselves do not make it possible to verify whether or not the contents of a data file or data set are unchanged under format transformations or content rearrangement, so are unusable as Scientifically Unique Identifiers.

Digital object identifiers

A Digital Object Identifier (DOI) is a unique identifier linked to a specific object, which must be a clearly defined piece of intellectual property; but which can either be tangible or intangible. The DOI system is one of the oldest digital identification systems in existence, having begun in 1998 with the creation of the International DOI Foundation (2010a).

Naming syntax

The DOI system consists of the DOI names, which are Uniform Resource Identifiers (URIs); the Handle System, used for DOI name resolution; and descriptive metadata based on the <indecs> framework (Rust and Bide 2000). The identifier, or DOI name, has two sections: a prefix and a suffix separated by a forward slash, “/”. The prefix itself has two parts separated by a “.”, the string “10” which indicates that this handle is a DOI, and a unique string assigned to the organization by a registry service.

DOI example from ORNL DAAC

DOI name: “10.3334/ORN LDAAC/840”

The value “10.3334” is the prefix assigned to the Environmental Science Division at Oak Ridge National Laboratory. The suffix, “ORN LDAAC/840”, is the locally unique identifier for the SAFARI 2000 data set distributed by the ORNL DAAC. The combination of the prefix plus the suffix ensures that the DOI is globally unique. In this case the second forward slash is part of

the local ID number.

Recommended Citation Including DOI:

Moody, E. G., M. D. King, S. Platnick, C. B. Schaaf, and F. Gao. 2006. SAFARI 2000 MODIS L3 Albedo and Land Cover Data, Southern Africa, Dry Season 2000. Data set. Available on-line [<http://daac.ornl.gov/>] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/840.

Resolvable URL: <http://dx.doi.org/10.3334/ORNLDAAC/840>

Technical value

As previously mentioned, prefixes are assigned to organizations by a registry service. Any one organization may have multiple prefixes. The combination of a prefix and a locally unique suffix was designed to ensure uniqueness and avoids the need for centralized allocation of DOI names. As such, the DOI scheme is theoretically infinitely scalable. In theory any character consistent with Unicode v2.0 may be used in a DOI. To date, the practice has been that the registry service portion of the prefix contains only numbers.

The DOI system is currently under consideration by the International Organization for Standardization as an ISO standard under Technical Committee 46, Subcommittee 9, Working Group 7. DOIs are also a registered Uniform Resource Identifier (URI) by the IETF (The International DOI Foundation 2010c) and the DOI syntax is an ANSI/NISO standard (Z39.84-2005) (The International DOI Foundation 2010b).

User value

DOIs have been widely adopted by publishers and are used to identify objects such as individual articles within a journal. When using a DOI in a citation, the prefix “doi:” is added to the DOI name to indicate that the identifier is a DOI. In addition, DataCite, an international organization recently formed to “support researchers by providing methods for them to locate, identify, and cite research datasets” also promotes the use of DOIs for digital science data (DataCite 2011).

Beyond their acceptance by the publishing community, DOIs currently are not known to provide any additional trust value, although DataCite is currently working on developing a specification for the metadata that should accompany a DOI name. Given membership by the California Digital Library in the DataCite consortium, it is not unreasonable to expect that considerations of trust may play a role in this or other DataCite efforts. Even though the DOI format is meant

to be an opaque identifier, existing identifiers can be incorporated into the suffix allowing for naming transparency.

Archive value

DOIs are interoperable with existing numbering schemes, regardless of whether recognized international numbering schemes, such as the ISBN, or a localized collection numbering scheme is used, as may be implemented at a given data center. This is accomplished simply by adding the DOI handle or prefix to the existing number, e.g. doi:10.3334/978-0316015844.

The DOI infrastructure consists of the identifier numbering syntax, a metadata schema for representing information about the object, a registry service, and a resolver service. The monetary costs vary and depend on the type of object being registered and the fee structure of the registrar. In the experience of the ORNL DAAC, implementation costs are minimal and consist mostly of staff time needed to assign and register a DOI prior to publishing a data set.

The DOI is designed to retain its relevance even if the object changes location or if the ownership of the object changes.

Use case support

As a URL-based identification system DOIs are location invariant not location independent and they do require a name authority. As such they do not meet the full requirements for a Unique Identifier (see [Summary Discussion](#) section). However, like the other URL-based identification schemes, DOIs would be acceptable Unique Locators. Moreover, given their wide-spread acceptance by the publishing community and support by DataCite, DOIs meet all of the requirements for a Citable Locator at least at the data set level, where the per item costs associated with DOIs are not likely to be cost prohibitive. DOIs in and of themselves do not make it possible to verify whether or not the contents of a data file or data set are unchanged under format transformations or content rearrangement, so are unusable as Scientifically Unique Identifiers.

Extensible resource identifier

The Extensible Resource Identifier (XRI) and the related Extensible Resource Descriptor (XRD) discovery format and XRI Data Interchange (XDI) semantic data sharing protocol specifications are being developed under the auspices of the Organization for the Advancement of Structured Internet Standards (OASIS).

Naming syntax

XRIs are formed of a series of segments, each of which may contain a permanent identifier or a transient reassignable identifier. The local context symbol “*” associated with a segment indicates that the segment value is not permanent; while “!” indicates that the “association between the identifier and the resource” it represents must never be changed (Reed 2010a).

Unlike most of the identification schemes assessed, XRI syntax allows for incorporation of semantic concepts through the use of global and local context symbols. Four “global context symbols have been defined” (Bartolomeo and Kovacikova 2009) by the OASIS XRI Technical Committee (2005):

- “=” designating a human as the ultimate name authority
- “@” designating an organization as the ultimate name authority
- “+” indicating a generic dictionary concept in the absence of an ultimate authority (e.g., any noun)
- “\$” where the ultimate authority is a standards organization, such as the OASIS XRI TC.

An example XRI (not bound to any protocol) follows:

@nsidc.org+dataset*newName!(doi:10.12345/OriginalName)

where:

- nsidc.org is the current authority for the object
- +dataset expresses the concept that NSIDC has data sets
- *newName is NSIDC’s current name for the data set
- !(doi:10.12345/OriginalName) is a permanent DOI that was assigned to the data when it was located at some other archive. The () notation indicates a cross-reference to another naming scheme.

The http binding to such a URL would become:

http://xri.net/@nsidc.org+dataset*newName!(doi:10.12345/OriginalName)

Technical value

Version 2.0 of the XRI specification narrowly failed a ballot to become a standard in 2008 (McRae 2008), primarily due to objections by the W3C Technical Advisory Group (Williams 2008). After working with the W3C TAG to understand their issues, the XRI TC is developing a third version of the XRI syntax specification (Bartolomeo et al. 2010). This version clarifies the relationship between URIs and XRIs by specifying that XRIs are a profile of a URI or Internationalized Resource Identifier (IRI) that provides additional structure and semantics. Version 2.0 of the XRI syntax specification is available (OASIS XRI

Technical Committee 2005) and specifications for a resolution service and bindings to HTTP, HTTPS, mailto, and info are actively being developed. Once completed, these new materials will be submitted for an OASIS standard vote (Reed 2010b).

XRIs were specifically developed to be independent of transport and access protocols, thereby allowing unambiguous assertions “that the same resource is being identified across different protocols, e.g., HTTP, HTTPS, FTP, SMTP, XMPP” and in different contexts (Reed 2010a). In addition, XRIs were designed to allow identifiers from multiple identifier schemes to be combined and shared; to allow delegation of naming authority for subsets of a namespace, to provide both human and machine readable identifiers, and to provide a mechanism for obtaining trusted responses on receipt of a XRI resolution request.

Since XRIs are being specifically developed to support resolution across multiple protocols, they could prove to be relatively impermeable to changes in technologies. Mechanisms for developing bindings for new protocols are well defined. As an example, Bartolomeo and Kovacikova describe the potential use of XRIs in combination with OMA’s XML Document Management specification to “enable interoperability of user profile data management between telecommunication and Internet services” (2009). Indeed, the largest extent XRI user community is a coalition of companies focused on Personal Data Exchange (PDE) technologies such as OpenId and Information Cards; while use in mobile bar code applications is another active area of work (Reed 2010b). However, given the incomplete state of the specifications and previous W3C objection to it, the future of this identifier scheme is not clear.

User value

XRIs have no more security issues than any other URI/URN-based scheme. The specification is too new to have been adopted by the publishing community; but it should be noted that http bound XRIs are fully qualified URLs, so could be used in a citation anywhere a URL can be used. The “XRI cross-reference syntax permits the inclusion of identifier metadata” (Reed 2010a) that could provide additional trust value, such as encrypted or hashed values that can be verified. Whether the identifiers are opaque or have meaning is up to the naming authority.

Archive value

Commercial name authorities exist for people and organizations—a small yearly fee is charged; but beyond that, communities are free to become their own authorities and evolve, as they deem necessary. Given the XRI cross-referencing capability, migrating data from one archive to

another should be straightforward. It does not matter whether the resource is renamed during the process, a proper resolution service would find it under any identifier. Lastly, it should be noted that any kind of resource could be described by an XRI.

Use case support

XRIs are location invariant not location independent and they do require a name authority. As such they do not meet the full requirements for a Unique Identifier (see [Summary Discussion](#) section). However, like other URI-based identification schemes, XRIs would be acceptable Unique Locators. While they do not have wide-spread acceptance of the publishing community, so are not optimal as a Citable Locators, they do have an advantage over DOIs for individual items within a data set in that there are no per item charges. XRIs in and of themselves do not make it possible to verify whether or not the contents of a data file or data set are unchanged under format transformations or content rearrangement, so are unusable as Scientifically Unique Identifiers.

The Handle System

Invented by the Corporation for National Research Initiatives (CNRI), the Handle System facilitates the assignment and management of unique global persistent identifiers to locate digital resources over time, in a manner that is independent of current or future storage locations (Corporation for National Research Initiatives 2009a, b). Using the Handle System to assign an independent permanent identifier to a digital resource and to maintain that association enables the current location of the resource to be obtained from its permanent identifier, known as a Handle (Kahn and Wilensky 2006). The Handle System is used by thousands of organizations to assign persistent identifiers, or Handles, to various kinds of digital resources, including scientific data and research-related information (Corporation for National Research Initiatives 2009b). The previously discussed DOI is a specific instantiation of the Handle protocol, with a particular set of social and business contracts.

Naming syntax

Consisting of two parts, a Handle contains a prefix, an array of characters that uniquely identifies the assigning organization, and a suffix, an array of characters that uniquely identifies a resource from all other resources that have been identified by the assigning organization.

Two types of URIs can represent a Handle:

- As a URN, a Handle is expressed in the form, hdl: prefix/suffix.
- Alternatively, as a URL, expressed in the form, <http://hdl.handle.net/prefix/suffix>, a Handle is resolvable in a web browser.

For example, <http://hdl.handle.net/4263537/5555> resolves to <http://www.handle.net/> the home page of the Handle System.

Technical value

The Handle System is specified in Request for Comments (RFC) 3560 (Sun et al. 2003a), 3651 (Sun et al. 2003b), and 3652 (Sun et al. 2003c), published by the IETF for review by the Internet community. The Handle System also has been designated as the standard required for uniquely identifying repositories and learning content distributed by the United States military (Department of Defense 2006).

The Handle System offers security, scalability, reliability, and interoperability. Security provisions include authentication of clients and servers to prevent unauthorized administrative access, capabilities to encrypt or disable access to confidential values, and options to request a digitally signed response (Sun et al. 2003a). The absence of design limits on the number of services within the Handle System, the number of sites within each service, or the number of handle servers within each site facilitates replication, reliability, and scalability (Corporation for National Research Initiatives 2007a). The Handle System is recognized as a fundamental infrastructure for enabling “interoperability in the Internet” (Denning and Kahn 2010, p. 36), also affording interoperability by enabling implementations using other systems, including systems for platform-independent web services (Smith 2007), domain name resolution (Paskin 2006), content typing (Blanchi and Petrone 2001), session initiation protocol mobility (Khouri et al. 2007), “rights management and discovery services” (Xiaofeng et al. 2010), and the DOI System (Chandrakar 2006).

When a Handle is specified in the form of a URL, the browser accesses the Global Handle Registry (GHR), maintained by CNRI, which points to the Local Handle Service (LHS) to access the current location of the resource. Any organization can establish itself as an LHS and serve as its own naming authority by registering with the GHR and installing a Handle Server to assign and maintain unique identifiers for resources. Sustainability of the Handle System is fostered by a business model that requires a one-time registration fee, which CNRI charges for assigning a prefix to an organization, and an annual maintenance fee for the GHS to point to the LHS (Corporation for National Research Initiatives 2007b). At the time of writing, each charge is nominally \$50.

User value

An organization that serves as a Handle naming authority can specify a scheme for suffix values as a simple number sequence or include additional identifying information to be assigned to resources (Corporation for National Research Initiatives 2007a; Lyons 2005). While the integrity of automatically assigned Handles can be established initially (Sun 2001), changes to the current locations of resources must be maintained by the naming authority to provide trustworthy persistent identification. Handles also can contribute to trust within a repository federation (Tansley 2006).

Like any other URL, the URL form of a Handle identifying a particular data set can be included within the references for an article prepared for publication. Given the proliferation of Handle servers and the subsequent plethora of open access resources that have been assigned Handle identifiers, their use in citations of various works in publications, including Earth science data and imagery, is relatively common. Handles also can be used within a data citation along with other identifiers (Altman and King 2007).

Archive value

The Handle System technology is openly available for free use worldwide, under a public license (Corporation for National Research Initiatives 2006). An installed Handle Server includes administrative tools that can be used either on site or remotely to manage the Handle Server and to maintain the Handles that it assigns. These tools can be used to change the location of a resource without occurring charges, if used by a naming authority responsible for the archive (Corporation for National Research Initiatives 2007a). A naming authority does not incur fees for assigning persistent identifiers. The Handle System also can be used to identify physical resources or information that is not accessible on the web by pointing to the internet address of a description that represents a physical resource, a contract, or a transaction (Kahn and Lyons 2006).

Institutional repository software is often implemented with Handle servers to automatically assign Handles to each collection, object, and data stream archived (Biswas and Paul 2010). In such installations, the Handle that is assigned to each object is encapsulated within the object as the value for the Identifier element of its Dublin Core metadata (Dublin Core Metadata Initiative 2008).

Use case support

Like other URI-based identification schemes, Handles are location invariant not location independent and they do require a name authority. As such they do not meet the full

requirements for a Unique Identifier (see [Summary Discussion](#) section). However, like the other URI-based identification schemes, Handles would be acceptable if not optimal Unique Locators and are being used to develop data publication services for Australia (Burton and Treloar 2009). While they do not have as much acceptance among commercial publishers as DOIs, they are accepted by the open access publishing community (Klerkx et al. 2010) and have an advantage over DOIs for individual items within a data set by avoiding per item charges. Handles in and of themselves do not make it possible to verify whether or not the contents of a data file or data set are unchanged under format transformations or content rearrangement, so are unusable as Scientifically Unique Identifiers.

Life science unique identifier

The Life Science Unique Identifier (LSID) effort began in 2003 by the Informatics Infrastructure Consortium (I3C) as a scheme to uniquely name life science entities of interest. Their goal was to “define a simple, common way to identify and access biologically significant data, where that data is stored in files, relational databases, in applications, or in internal or public data sources” (Life Science Identifier Resolution Project 2010).

Projects such as the following are actively using and exploring the use of LSIDs: BioMOBY, Generation Challenge Programme, Index Fungorum, Taxonomic Search Engine, and the Northern Temperate Lakes LTER Network (TDWG Globally Unique Identifiers: LSID 2010). The Life Science Identifier Resolution Project serves as a key reference for LSID development (2010).

Naming syntax

The LSID is expressed in the following format, as described by the Taxonomic Databases Working Group (2006):

URN:LSID:<Authority>:<Namespace>:<ObjectID>[:<Version>]

Salamone offers the following example of an LSID (Salamone 2002):

“urn:lsid:ncbi.nlm.nih.gov:GenBank:T48601:2.”.

Where the colon-delimited sub-parts of the LSID are:

1. urn:lsid—mandatory preface for the URN
2. ncbi.nlm.nih.gov—internet domain of the organization that assigned the LSID to the data
3. GenBank—the name of the data resource
4. T48601—name of a specific data element within the resource
5. 2—version number of the specific data element.

Technical value

The LSID specification has been proposed as a URN scheme to the I3C and has been adopted as a specification by the Object Management Group (Bafna et al. 2008). As a URN-based scheme, LSIDs carry the same security risks as other URN schemes in common use. LSIDs are naturally interoperable with other Internet approaches—the URN syntax and semantics are by definition compatible with other Internet naming schemes.

A companion protocol in formation, the LSID resolver protocol, allows caching at a variety of scales to facilitate resolution. As a result, the scheme is very scalable when deployed in distributed (or even singular) databases (Smith and Szekely 2005; LSID-Developer 2010). There are no hard requirements for third party maintenance, though actual implementation of LSID repositories (whether distributed or consolidated by a central naming authority) could rely on 3rd party maintenance.

An emerging controversy surrounding the evolution of the LSID (and LSID resolver) schemes exists (LSID 2010). At question, is whether HTTP URIs can perform an equivalent role, such that the LSID scheme violates the web architecture practice of re-using existing URI schemes. It is not clear at this time, what impact this controversy will have on long term use of this identification scheme.

User value

LSIDs embody limited intrinsic semantics, via the alphanumeric authority and namespace elements. Name assigning authorities are free to assign the final qualifying element as they choose, as long as the result is unique. As a result, transparency or opacity in naming is a choice of the name assigning authority. Additionally, it is common practice in some communities to use some form of cryptographic hash of the data, such as a UUID, as the ObjectID, providing additional trust value to the resulting identifier (Hyam 2009). Lastly, some journals have begun to include the LSID in articles, including Zookeys (2010) and Zootaxa (2010), which both publish articles that use the LSID to identify individual authors as well as data.

Archive value

The LSID scheme is no more or less maintainable than any other URN scheme. Migrating large collections of LSIDs from one database to another will require the usual diligence to ensure that no name-space collisions have occurred, but the basic properties of the scheme naturally attempt to avoid this as well as possible. There are no charges to issue or use LSIDs (Taxonomic Databases Working Group 2006). While the URN scheme itself is a

web-internet centric protocol, LSIDs may be used to uniquely identify any appropriate entity including physical entities when used in conjunction with a technology or protocol, such as bar codes and/or RFID tags, for labeling such entities.

Use case support

Like other URI-based identification schemes, LSIDs are location invariant not location independent and they do require a name authority. As such they do not meet the full requirements for a Unique Identifier (see [Summary Discussion](#) section). However, like the other URI-based identification schemes, LSIDs would be acceptable Unique Locators, though it is not clear that re-use of these identifiers within the Earth Science community is warranted while the on-going controversy over the existence and evolution of the LSID is unresolved. While they have some acceptance with publishers within the life sciences, they do not have the widespread acceptance within the Earth science publishing community, so are not optimal as a Citable Locator. Given the aforementioned controversy it is not clear whether or not LSIDs will have broader acceptance. LSIDs in and of themselves do not make it possible to verify whether or not the contents of a data file or data set are unchanged under format transformations or content rearrangement, so are unusable as Scientifically Unique Identifiers.

Object identifier

The Object Identifier (OID) is used to establish a name for an object in computing. This identification is heavily overloaded, since many communities use them for different purposes (Object Identifier 2010), including

- Naming Internet Nodes, typically as maintained by the Internet Assigned Numbers Authorities (IANA)
- For naming object types in X.509 certificates
- Within LDAP schemas for directory structures (Wahl et al. 1997)
- In the context of the Simple Network Management Protocol (SNMP), an OID identifies objects in a Management Information Base
- As identifiers in relational databases, such as Postgres and Oracle
- Identifying health care information in exchange standards, where the OIDs form globally unique identifiers

Naming syntax

“OIDs are composed of multiple occurrences of digits and the ‘.’ character. Lexical equivalence is achieved by exact

string match” (Mealling 2000). As examples, the numeric form of the Internet OID is 1.3.6.1; while IANA-assigned OIDs have a base OID of 1.3.6.1.4.1, and so on.

The appropriate form of an OID was heavily discussed during its formative years (Larmouth 2003; Lyons 2000; ASN 2003). Around 2000, the syntax for OIDs that may be used in communication protocols was formalized into Abstract Syntax Notation One (ASN.1) (Dubuisson 2001; Larmouth 1999). The ITU has published encoding rules for this language under the name “Recommendation X.200” (ITU 1994) and under ISO (ISO7498-3 1997). Compilers for it are also available, although the full language is rather complex (Dubuisson 2001).

Technical value

In simple form, OIDs form a tree that is readily extensible. Tree searches for leaves are typically $O(\log N)$ for binary structure and at worst $O(N)$ for the linear case, indicating reasonable scalability. As a hierarchical structure OIDs are potentially interoperable with any other hierarchical identifier scheme. In addition, a basic OID structure can be supported by a variety of technologies.

Because the underlying computer science is simple and convenient for describing hierarchical information structures, OIDs have been used informally, semi-formally, and formally. Alvestrand (1997) provides a quite concise, WWW-available summary of OIDs. Based on his description, it is easy for any individual or organization to define a hierarchical identifier schema that can be registered with one of the extant registration authorities (OID-Registry 2010). An important characteristic of such registration is that it provides delegation authority, so that the registration authority provides registration to a certain point in the naming hierarchy and then cedes further expansion to the registrant (ITU 2005). In other words, if a registrant obtains a base OID of 1.2.3.4, they then have full authority to generate subsidiary OIDs 1.2.3.4.# or to delegate authority for some subset of their OIDs to other authorities, for example, to delegate OIDs 1.2.3.4.5.# to another organization. While 3rd party OID registries are required, the organizations that act as registries (e.g., Library of Congress, IANA, etc.) appear likely to be as long-lived as the federal government and equivalent international bodies.

User value

One important application of OIDs is in the library cataloging approach adopted by the Library of Congress for Machine-Readable Cataloging records (MARC). These records have been used to record bibliographic entries for library catalogs (Furrie 2009; Network Development and

MARC Standards Office 1999; 2006). They are also related to the ISO standard Z39.50. The Z39.50 International Standard Maintenance Agency maintains an OID registry of Z39.50 Object Identifiers (Registry of Z39.50 Object Identifiers 2009).

Unlike URL-based identifier schemes and DOIs, publishers have not to date accepted OIDs as part of a citation. Agreement by professional societies and publishers would be needed in order to develop such a scheme. While in principle, it should be possible to develop OID-based schemes that encapsulate ownership and rights information that could provide additional trust value for the data, in practice this development has not been done.

The numeric format of the OID scheme is implicitly opaque. However, the common practice of assigning character strings or specific definitions to OID values or components would allow development of a transparent naming scheme.

Archive value

Once assigned a base OID, an organization or federation of organizations become responsible for determining how subsidiary names are assigned and managed.

In some cases, registration is provided for free. Alvestrand (1997) notes that the original intention was to hand out free registrations. Registries include

- IANA—which freely distributes OIDs within its “Private Enterprises” branch
- ANSI—which distributes OIDs within its “US Organizations” branch for \$1,000–\$2,500 depending on whether or not a string translated version of the ID is desired
- BSI—which distributes OIDs within its “UK Organizations” branch

It is not immediately obvious how to maintain an OID during data migration from one facility to another. OIDs can be used to identify non-Web data or physical objects.

Use case support

Unlike the other identification schemes assessed, it is not at all clear that OIDs are either location invariant or location independent and they do require a name authority. As such they do not meet the requirements for either a Unique Identifier or a Unique Locator. In addition, they have not gained recognition from the publishing community, so are not optimal as a Citable Locator. OIDs in and of themselves do not make it possible to verify whether or not the contents of a data file or data set are unchanged under format transformations or content rearrangement, so are unusable as Scientifically Unique Identifiers.

Persistent uniform resource locator

Persistent Uniform Resource Locator (PURL) technologies are at least a decade old (Lynch 1997). A PURL is both a URL and a URI, where the URL describes a persistent location that redirects to the current location of a web resource. Weibel (2007) provides an informal commentary on the development of PURLs. The Library of Congress (1997) provides a somewhat more formal description.

Naming syntax

PURLs have four parts:

- 1) A protocol—typically http
- 2) A resolver address
- 3) A domain—a hierarchical identifier analogous to a URL path name that allows PURLs to have different maintainers
- 4) A name—assigned by the user

An example PURL: <http://purl.oclc.org/NET/EMILLER/> where

- http is the protocol
- purl.oclc.org is the resolver address
- net is the domain and
- EMILLER is the web resource (Miller 2007).

Currently, this PURL redirects to <http://www.w3.org/People/EM/>, the home page of Eric Miller, the President of Zepheira, Inc. who “served as a Research Scientist at MIT’s Computer Science and Artificial Intelligence Laboratory”, was one of the co-developers of PURLs, was previously at OCLC and was “co-founder and Associate Director of the Dublin Core Metadata Initiative” (Zepheira 2011).

Technical value

Originally developed and hosted by the Online Computer Library Center (OCLC), the OCLC continues to provide name resolution and administration services for PURLs. According to the OCLC Information and Services for Publishers (2010):

“OCLC is a nonprofit, membership, computer library service and research organization dedicated to the public purposes of furthering access to the world’s information and reducing information costs. More than 72,000 libraries in 171 countries and territories around the world have used OCLC services to locate, acquire, catalog, lend and preserve library materials”.

The development of PURLs has been informed by developments in both MARC and Dublin Core (Powell

2004; Rust 1998), while “OCLC’s active participation in the IETF Uniform Resource Identifier working groups” drove the development of the PURLs (OCLC 2010).

In March 2009, in recognition that a single registration authority without multi-site replication may be a potential single point failure, Zepheira, Inc. and the National Center for Biomedical Ontology, banded together to develop a PURL Federation which will allow multiple resolution services to cooperate in PURL resolutions (PURL Federation 2010). An Alpha version of this extension to the PURL system is expected.

No security issues are known, beyond those endemic with any URL-based system. Given the exhibited durability of the system, its close ties to IETF and the W3C, and its current evolution to better support the Semantic Web it is likely that PURLs will remain viable for quite some time.

User value

The name portion of the PURL can be either opaque or meaningful as desired by the user, and is compatible with many other naming conventions. Since PURLs are simply URLs, they can be used in a citation in any place where a URL can be used. PURLs, in and of themselves, are not known to have any additional trust value. However, it is possible to choose a scheme for assigning the name portion of the PURL that does have additional trust value, for example a scheme based off a hash of the data itself.

Archive value

The PURL and PURLz systems are open source and freely available, and the PURL user community appears to be fairly active. The current OCLC-based system is free of charge. In 2007, OCLC and Zepheira announced that they would work together to re-architect OCLC’s PURL services “to more effectively support the management of a ‘Web of data’” (Online Computer Library Center 2007). In addition to enhancements to reflect current W3C standards, the updated version of PURLs allows for the extension to objects other than web resources that are important to the Semantic Web. In particular the updated PURL services can handle links to resources such as people, organizations, concepts, and scientific data.

Like any URL-based identifier scheme, persistence depends on maintenance and the commitment of the provider. Given proper attention to future flexibility considerations, migration of data from one archive to another, or from place to place within an archive is easily managed. In particular, PURL reliance on partial resolution allows a great deal of flexibility at all scales from moving individual files, entire data sets, or an entire archive.

Use case support

Like other URI-based identification schemes, PURLs are location invariant not location independent and they do require a name authority. As such they do not meet the full requirements for a Unique Identifier (see [Summary Discussion](#) section). However, like the other URI-based identification schemes, PURLs would be acceptable Unique Locators. While they do not have the wide-spread acceptance of the publishing community, so are not optimal as a Citable Locator for data sets as a whole, they do have an advantage over DOIs for individual items within a data set in that there are no per item charges. PURLs in and of themselves do not make it possible to verify whether or not the contents of a data file or data set are unchanged under format transformations or content rearrangement, so are unusable as Scientifically Unique Identifiers.

Uniform resource identifier/name/locator

A Uniform Resource Identifier (URI) is a naming and addressing technology that uses “a compact sequence of characters to identify” World Wide Web resources (Berners-Lee et al. 2005). IRIs (Duerst and Suignard 2005) complement URIs by allowing the use of “characters from the Universal Character Set” (Unicode/ISO10646). URIs encompasses both URLs (Berners-Lee et al. 1994) and URNs (Moats 1997). In general, URLs, URNs, and IRIs could all be used to uniquely identify data collections.

Naming syntax

The syntax of these kinds of identifiers is familiar from their common use.

Technical value

The Network Working Group at the IETF maintains the URI specifications. URI technologies are used in domain name systems and in systems that enable email and File Transfer Protocol (FTP).

URLs use domain names to locate resources. Domain names have a well-defined naming structure. The Domains can be categorized as generic top-level domains (.com, .org), country code domains (.us, .fr), and second level or lower domains. The Internet Corporation for Assigned Names and Numbers (ICANN) is in charge of managing the root Domain Name System (DNS). ICANN delegates the top domain names to the Network Information Center (NIC). Each country manages their country code level domains.

URNs do not imply availability of the resource being identified. The IANA is in charge of processing applica-

tions for URNs. Usually only the URN scheme is registered, not the actual identifiers. Since there is no universal registry for URNs, it is difficult to use them or resolve them.

The common belief about the difference between URNs and URLs is that a URL should be resolvable, while a URN does not have to be. This may be due in part to the fact that according to their designers, URNs “are intended to serve as persistent, location-independent resource identifiers” (Moats 1997). This also implies to many that a URN is more persistent, since it is not dependent on a web domain that can change or disappear over time.

User value

For many purposes, the use of URLs is preferred to URNs, since even though both provide identifier mechanisms, URLs provide easy navigation to the resource being identified, letting programs and users obtain information in a transparent way. In addition, the use of URNs does not guarantee persistence nor determine all ownership issues, although it can be argued that, given the relatively complex steps required to obtain and manage a URN, organizations that do so are likely to be more stable. At the same time, URLs clearly suffer from greater transience overall, and their use in situations that do not provide URI resolution will likewise cause frustration, as users enter a URL into their browser and get a “404: resource not found” message. So neither mechanism provides a guarantee of the existence or discoverability of an object.

While plain URLs are not necessarily persistent, they are accepted as a part of a citation by many journals. While the URL/URN/URI schemes are likely to be around for a very long time, given their ubiquitous use internationally, individual URL or URN values have no particular trust value. URLs in particular, are known for their impermanence. URI-based naming schemes can either be opaque or transparent depending on the needs of the designer.

Archive value

An organization that chooses to use URIs as its identifiers will need to maintain the web domain, manage the structure of the URIs and maintain the URL redirects (Cox et al. 2010) for the long-term. For example, the Open Geospatial Consortium currently uses URLs to uniquely identify standards, types of documents, coordinate reference systems, etc. The OGC Naming Authority Board is in charge of creating and approving each identifier and its resolutions. Identifiers could reference documents, images, electronic mailboxes, files, or other digital resources. In the Semantic Web (Berners-Lee et al. 2001) URIs not only describe

common web resources, but also concepts (e.g. person, place) or real things (e.g. John Smith, data set X).

While redirection can be used to maintain access to items that have moved from one archive to another, in practice there are many problems with this solution. For example, if an original URL contains information that identifies or brands an item as belonging to some physical organization, the odds of the next archive being willing to maintain that original URL becomes very small. In addition, while redirection can take the user to the correct object, a bookmark created from the original landing page will not persist.

Use case support

URNs were designed to be location independent. As such they would meet the requirements for Unique Identifiers. However, using URNs as Unique Identifiers for Earth Science data, would require designing a naming scheme and applying to IANA for approval, unless one of the existing approved URN identifier schemes were adopted. Since they would not be location invariant they would neither make an acceptable Unique Locator nor Citable Locator. On the other hand, URLs are not location independent, so do not fulfill the requirements of a Unique Identifier; but are location invariant so could conceivably act as a Unique Locator. Using them as such however, requires a long-term commitment by the archive and is problematic at best in the case where the data moves from one archive to another. While URLs are widely used within citations, their general impermanence makes them suboptimal as Citable Locators. Neither URNs nor URLs, in and of themselves, make it possible to verify whether or not the contents of a data file or data set are unchanged under format transformations or content rearrangement, so neither are unusable as Scientifically Unique Identifiers.

Universally unique identifier

A Universally Unique Identifier (UUID) is a number that is 16-bytes (128-bits), as specified by the Open Software Foundation's Distributed Computing Environment (DCE) standard (Kong 1995; The Open Group 1997). This standard allows distributed systems to create UUIDs without coordinating with a central authority while expecting that no other system will create and use the same UUID. Information from multiple remote systems labeled with UUIDs can then be combined on a server without name collisions (The Open Group 1997).

Naming syntax

A UUID contains 36 characters, of which 32 are hexadecimal digits that are arranged as 5 hyphenated groups, for example:

0a9ecf4f-ab79-4b6b-b52a-1c9d4e1bb12f

A well-known use of UUIDs is in Microsoft's Globally Unique Identifiers (GUIDs). Microsoft's "RPC run-time libraries use UUIDs to check for compatibility between clients and servers and to select among multiple implementations of an interface" (MSDN 2010). Other uses of UUIDs include Linux's ext2/ext3 file system, Mac OS X, and KDE, which all are based on implementations available from the e2fsprogs distribution (Ts'o 2010).

Technical value

Procedures for establishing UUIDs are specified in ISO/IEC 11578:1996 "Information technology—Open Systems Interconnection—Remote Procedure Call (RPC)" and in ITU-T Rec. X.667 | ISO/IEC 9834-8:2005. Also, the (Internet Engineering Task Force) IETF has published Proposed Standard RFC 4122 (Leach et al. 2005).

The first version of the UUID contained the Media Access Control (MAC) address and a timestamp of when the UUID was created and can therefore be used to trace back to the machine that created it. Announcing the existence of a valid IP address is considered to be a security risk. Later versions of the UUID standard do not contain the MAC address, relying on random number generators or a hashing scheme instead (Leach et al. 2005). Version information is included within 4 of the 128 bits available.

A common usage of UUIDs is as a unique label that can be associated with a less unique identifier, for example as a database primary key associated with a filename. In this limited context UUIDs are compatible with other identification schemes. Because UUIDs are unique and persistent, they make excellent Uniform Resource Names; however, some other scheme is required to form a URL, so UUIDs have limited use in discovering or accessing the item named (Leach et al. 2005).

User value

Due to the 16 byte length of the UUID, the number of possible permutations is 2^{128} . While it is not guaranteed that a UUID will be unique across all domains, it is highly unlikely that two identical UUIDs will be created. The UUID is designed to be opaque—information contained in them is not meant to be human readable. UUIDs have additional trust value when one of the hashing schemes is used when minting them. In these cases, users may re-hash the data to verify that no changes have been made since the UUID was minted. Unless incorporated as part of a URL, use of UUIDs within a citation is unknown.

Archive value

Unlike DOIs or other URL-based identification schemes, UUIDs do not need to be recreated or maintained when data

is migrated from one location to another. The underlying technology used to create UUIDs is simple and relatively easy to implement on most operating systems. The tools for creating them are included with many operating systems and computer languages and are freely downloadable. UUIDs could be used to label physical objects in addition to objects in the digital domain.

Use case support

UUIDs meet the all of the requirements for a Unique Identifier. However, since they are not location invariant they would not make either acceptable Unique Locators or Citable Locators. UUIDs in and of themselves do not make it possible to verify whether or not the contents of a data file or data set are unchanged under format transformations or content rearrangement, so are unusable as Scientifically Unique Identifiers.

Summary discussion

Table 2 shows each of the identifier schemes in terms of its suitability for each of the presented use cases, including use as a Unique Identifier, a Unique Locator, a Citable locator, or a Scientifically Unique Identifier. While the simplicity of such a table can be convenient, a discussion of the issues of each use case is necessary to address the complexity and potential use of these identifier schemes. In the following sections, we address the results of the assessment in terms of the use cases defined above for both data sets and individual data files or granules.

Unique identifier

Of all of the identifier schemes assessed, the one that most closely meets all of the requirements for a Unique Identifier is the UUID scheme. UUIDs are globally unique and do not require a naming authority. Specific versions of the generation algorithm can be used to encode additional trust information such as a message digest. Software to generate UUIDs is readily available, so it should be possible to create them as needed from virtually anywhere in the world. UUIDs are used as unique keys in catalogs and within metadata records and can be used within the objects themselves. While UUIDs would work well at both a data set and a data file level, UUIDs are particularly suited to identification of individual objects.

Beyond UUIDs, any of the URL or URI-based schemes could be used if the requirement that there be no central name authority is relaxed and/or if some scheme to guarantee the global uniqueness of the name is in place. Indeed, it is common in some disciplines to combine a Unique Identifier with a Unique Locator to create a single identifier that satisfies both use cases. For example, in the Life Sciences the ZooBank uses UUID's for the Object ID portion of their LSIDs (Pyle and Michel 2008).

Relaxing the requirement that there be no central name authority may seem like a reasonable option for data centers and large research groups that have the means to either become a name authority or to develop relationships with an external authority, especially since the data these group generate is not likely to be acquired and identified in the field, though this is a less viable option for individual investigators or small research teams. However, it is not

Table 2 Suitable identifiers for each use case where solid green indicates high suitability, vertical yellow stripes indicates good to fair suitability; and orange diagonal stripes indicates low suitability

Identifier Type	Unique Identifier		Unique Locator		Citable Locator		Scientifically Unique Identifier	
	Dataset	Item	Dataset	Item	Dataset	Item	Dataset	Item
ARK	Yellow stripes	Yellow stripes	Green	Green	Yellow stripes	Yellow stripes	Orange stripes	Orange stripes
DOI	Yellow stripes	Orange stripes	Green	Green	Green	Yellow stripes	Orange stripes	Orange stripes
XRI	Yellow stripes	Orange stripes	Green	Green	Yellow stripes	Yellow stripes	Orange stripes	Orange stripes
Handle	Yellow stripes	Orange stripes	Green	Green	Yellow stripes	Yellow stripes	Orange stripes	Orange stripes
LSID	Yellow stripes	Orange stripes	Yellow stripes	Yellow stripes	Yellow stripes	Yellow stripes	Orange stripes	Orange stripes
OID	Orange stripes	Orange stripes	Orange stripes	Orange stripes	Orange stripes	Orange stripes	Orange stripes	Orange stripes
PURL	Yellow stripes	Orange stripes	Green	Green	Yellow stripes	Yellow stripes	Orange stripes	Orange stripes
URL/URN/ URI	Yellow stripes	Orange stripes	Green	Green	Yellow stripes	Yellow stripes	Orange stripes	Orange stripes
UUID	Yellow stripes	Green	Orange stripes	Orange stripes	Orange stripes	Orange stripes	Orange stripes	Orange stripes

clear that the need for ongoing maintenance of the combined identifier as the data moves through its lifecycle from the research community into an archive or archives outweighs the benefits of this path, though clearly the level of maintenance required would be considerably less for collections of objects, than for individual objects.

It should also be noted that globally unique identifiers can be generated by investigators using any of the URI-based schemes, provided of course that the investigator has pre-reserved a namespace with a provider. Of course, the risk with this approach is the creation of large numbers of what appear to be Unique Locators (e.g., Handles, DOIs, PURLs, ARKs, etc.) that are not actually resolvable. This can occur for a variety of reasons, for example early in the research process data is not generally accessible though it may be shared within a research group or with collaborators at other institutes, or later in the data lifecycle when an investigator/archive determines that only certain versions or portions of the data should be published or archived. At this point it is not clear how large numbers of non-resolvable identifiers will impact the longevity of a Unique Locator technology, though the value of that technology can be expected to decrease in the eyes of the user community.

Unique locator

Any of the URL/URI/IRI-based identifier schemes assessed could be used here and since there are currently no strong market leaders to help make the choice among them, the decision must be based on secondary criteria. Of the suite of schemes available, ARKs perhaps provide the most additional trust value due to explicit statements that are included in the metadata about the preservation commitments of the provider, while XRIs perhaps provide the most protection against future technologic change since they were designed to be independent of transport and access protocols. The Handle system has perhaps the largest user base, which could be taken as a demonstration of its expected longevity, though the PURL claim to support by OCLC, IETF, and W3C is perhaps a potent counter argument. Of the suite, DOIs, at least at the individual file level, are perhaps the most problematic given that the current per DOI financial charges which, while small on a per item basis, quickly mount up when the number of objects is very large as it often is in the Earth sciences. Lastly, while local URL schemes can fit the bill in the short term, in the long term they provide the additional risk that in the event that data must move to a new repository, that new repository will be reluctant or unable to continue supporting the existing local URL scheme. In all cases, it is unwise to place organizational names within these

identifiers, since should data move to a new repository it is very unlikely that they would be willing to continue maintaining identifiers containing some other organization's name (Willet 2010).

Citable locator

While most publications now allow the use of URLs in citations, so that all of the URL/URI/IRI based identification schemes discussed in this paper could potentially be used as a Citable Locator, DOIs are the identification scheme currently adopted by most commercial publishers in the Earth sciences. Open access publishers, such as ARIADNE have adopted the Handle System to generate persistent identifiers, which also are UUID-compliant, for learning objects published (Klerkx et al. 2010). While DataCite does support other identifiers, DOIs are the data set identification scheme primarily supported by DataCite, an international consortium of libraries and information centers whose vision is “to support researchers by providing methods for them to locate, identify, and cite research datasets with confidence” (DataCite). When the intent of a citation is to provide credit to the data producer, so that a citation of a data collection as an entity is appropriate, the modest per DOI charges are not a stumbling block for well-funded data repositories at the data set level. Less-funded repositories could use Handles, ARKs, and PURLs, etc. When the costs of DOIs have been relaxed, “one can easily imagine supporting some collections using handles and some collections using DOIs” (Crosas 2011), as well as other technologies. If in the future data repositories were to develop mechanisms that allow researchers to associate a specific list of data objects used with a citable identifier, an association guaranteed to be maintained by the archive into the future, a modest charge perhaps in the form of a publication fee would be appropriate.

Scientifically unique identifier

While none of the identifier schemes assessed here even minimally address this use case, research into mechanisms to address the issue are underway in a number of fields, most notably social science where Altman (2008) and others are contributing to the development of Universal Numeric Fingerprints (UNFs). This work depends on finding an authority that can provide an acceptable “canonical” format for the many different kinds of Earth science data. In the probable event that no such authority emerges, it appears that Earth science data collections (whether files or larger collections) are not resolvable into distinct “objects” that can be uniquely identified. Rather, a particular collection is a member of an equivalence class, which requires a mapping to other members of the class

in order to demonstrate equivalence (Barkstrom 2011). Deeper discussion of this issue is beyond the scope of this paper.

Future work

The results of this study are largely based on a review of the literature and theoretical considerations of the needs for data identity in the Earth Sciences. While it would be convenient to recommend a single identification scheme or technology that addresses data identity issues for the Earth Sciences, the issues of data identification are too complex for the authors to make such a recommendation. Indeed, two of the use cases identified—the Unique Locator and Scientifically Unique Identifier—resulted in no clear winning identifier schemes. Mid-study the Stewardship Cluster was presented with the opportunity to gain experience with each of the identifier technologies assessed herein. The purpose of this test bed activity is to assess operational considerations with each technology. Hopefully, such considerations may clarify the choices, particularly with the Unique Locator use case.

In addition, over the course of this study several best practices were identified which should be gathered together and promulgated to the community through processes such as the NASA Standards Process Group (SPG) or the Global Earth Observing Systems of Systems (GEOSS) Best Practices registry. Similarly, several areas where best practices are needed were identified. It is likely that the Stewardship Cluster will address these concerns.

Finally, the TIWG has recently agreed to submit DOIs, as citable locators for data sets, and UUIDs, as unique identifiers for individual data items or granules, to the NASA SPG for endorsement as Earth Science Data System (ESDS) standards in the upcoming year.

Conclusion

The review of technologies available for assigning persistent identifiers to digital Earth science data, while not exhaustive, illustrates choices and issues to be considered when evaluating opportunities for establishing systems to assign global persistent identifiers to scientific data and research-related information planned for dissemination. The review of these persistent identifier technologies demonstrates capabilities that are being developed to enable the identification of data and aspects of the data for citation in the reference sections of publications.

The analysis revealed that the decision to select a specific technology for assigning persistent identifiers to Earth science data depends on the uses that are intended for

the identifiers by the scientific archive or data center that hosts and disseminates the data. Four primary use cases were identified as being of central interest in the Earth Sciences. All four use cases are applicable to data sets as a whole or for individual data items. Of the technologies evaluated, there does not appear to be a single persistent identifier technology that would be ideal for every use. Each identifier technology has unique qualities that could be applicable to specific identification needs. Considering such capabilities, a particular archive could decide to deploy multiple persistent identifier technologies to meet the needs of their stakeholders.

Ideally, systems used for archiving Earth science data should be designed to support multiple types of identifiers. In addition to the variations in uses of identifiers for data that have been described, persistent identifiers may be needed to identify other characteristics of data. For example, individuals, such as authors or other contributors may need to be uniquely identified. Similarly, organizations that contribute to the creation or publication of data may need to be described. Given such needs for unique persistent identifiers, additional research is needed on technologies to facilitate the assignment of globally unique persistent identifiers and on the uses of such technologies and the artifacts used to identify geospatial data and its characteristics. Considering the needs for identifying data and the variety of capabilities for identification that have been reviewed, unique identifiers should be assigned to Earth science data and used to cite the data as references in publications (see Online Resource 1 for a simple example of how this might work).

Acknowledgments The authors are grateful for the support received from the National Aeronautics and Space Administration (NASA), including support received for Robert Downs under contract NNG08HZ11C and the support for Ruth Duerr received under contract NNG08HZ07C and grants NNX08AN99A and NNX10AE07A. The authors are also grateful for the support received from the National Science Foundation under grant ARC 0946625. Lastly, the authors are grateful to the members of NASA's TIWG and the ESIP Stewardship Cluster who materially contributed to the results of the paper through many discussions during monthly teleconferences, list serve discussions and twice yearly meetings.

References

- Altman M (2008) A fingerprint method for scientific data verification. *Adv Comput Inf Sci Eng* 311–316. doi:10.1007/978-1-4020-8741-7_57
- Altman M, King G (2007) A proposed standard for the scholarly citation of quantitative data. *D-Lib Mag* 13(3/4). ISSN 1082–9873. <http://dlib.org/dlib/march07/altman/03altman.html>. Accessed 13 August 2010
- Alvestrand HT (1997) Object identifiers registry. <http://www.alvestrand.no/objectid/>. Accessed 14 August 2010
- Bafna S, Humphries J, Miranka DP (2008) Schema driven assignment and implementation of life science identifiers (LSIDs). *J Biomed Inform* 41(5):730–738. doi:10.1016/j.jbi.2008.05.014

- Barkstrom BR (2011) On the uniqueness of earth science data. In preparation
- Bartolomeo G, Kovacikova T (2009) User profile management in next generation networks. Telecommunications, 2009. AICT '09. Fifth Advanced International Conference on, 129–135. Venice: IEEE, May 24. doi:10.1109/AICT.2009.28. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5072418. Accessed 21 August 2010
- Bartolomeo G, Kovacikova T, Petersen F (2010) A novel approach to use data federation in next-generation networks. *Int J Commun Syst* 23:802–816. doi:10.1002/dac.1083
- Berners-Lee T, Masinter L, McCahill M (1994) Uniform Resource Locators (URL). Request for Comments: 1738. Internet Engineering Task Force (IETF). <http://www.ietf.org/rfc/rfc1738.txt>. Accessed 13 August 2010
- Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. *Sci Am* 184:34–43
- Berners-Lee T, Fielding R, Masinter L (2005) Uniform Resource Identifier (URI). Request for Comments: 3986. Internet Engineering Task Force (IETF). <http://www.ietf.org/rfc/rfc3986.txt>. Accessed 13 August 2010
- Biswas G, Paul D (2010) An evaluative study on the open source digital library softwares for institutional repository: special reference to Dspace and greenstone digital library. *Int J Libr Inf Sci* 2(1):1–010
- Blanchi C, Petrone J (2001) An architecture for digital object typing. Corporation for National Research Initiatives. <http://www.cnri.reston.va.us/software/repository/repo-whitepaper.pdf>. Accessed 13 August 2010
- Burton A, Treloar A (2009) Publish my data: a composition of services for ANDS and ARCS. Proceedings of the Fifth IEEE International Conference on e-Science. 164–170. doi:10.1109/e-Science.2009.31
- Chandrakar R (2006) Digital object identifier system: an overview. *The Electr Libr* 24(4):445–452. doi:10.1108/02640470610689151
- Corporation for National Research Initiatives (2006) Handle system public license agreement (Version 2). <http://hdl.handle.net/4263537/5030>. Accessed 14 August 2010
- Corporation for National Research Initiatives (2007a) HANDLE.NET (version 6.2) technical manual (Version 2). <http://hdl.handle.net/4263537/5031>. Accessed 13 August 2010
- Corporation for National Research Initiatives (2007b) Handle system service agreement (Version 3). <http://hdl.handle.net/4263537/5029>. Accessed 14 August 2010
- Corporation for National Research Initiatives (2009) The handle system. <http://www.handle.net/>. Accessed 13 August 2010
- Corporation for National Research Initiatives (2009b) Quick facts. <http://www.handle.net/factsheet.html>. Accessed 13 August 2010
- AGU Council (2009) The importance of long-term preservation and accessibility of geophysical data. http://www.agu.org/sci_pol/positions/geodata.shtml. Accessed 15 September 2010
- Cox S, Schade S, Bermudez LE, Simonis I (2010) “OGC identifier policy—a case for http URI”, Open Geospatial Consortium, Document No.: 10-124r1
- Crosas M (2011) The dataverse network®: an open-source application for sharing, discovering and preserving data. *D-Lib Magazine* 17 (1/2). doi:10.1045/january2011-crosas
- DataCite (2011) <http://datacite.org/>. Accessed 11 January 2011
- How to Cite a Dataset (2008) international polar year data and information service. <http://ipydis.org/data/citations.html>. Accessed 21 August 2010
- Denning PJ, Kahn RE (2010) The long quest for universal information access. *Commun ACM* 53(12):34–36. doi:10.1145/1859204.1859218
- Department of Defense. Instruction Number 1322.26 (DoDI 1322.26) (2006) <http://www.dtic.mil/whs/directives/corres/pdf/132226p.pdf>. Accessed 14 August 2010
- MSDN (Microsoft Developer Network) (2010) GUID structure. <http://msdn.microsoft.com/en-us/library/aa373931%28VS.85%29.aspx>. Accessed 15 September 2010
- Dublin Core Metadata Initiative (2008) DCMI metadata terms. <http://dublincore.org/documents/dcmi-terms/>. Accessed 14 August 2010
- Dubuisson O (2001) ASN.1: communication between heterogeneous systems, translated from the French by Phillippe Fouquart, Morgan Kaufman, San Francisco. ISBN: 0-12-633361-0. Also available online: <http://www.oss.com/asn1/dubuisson.html>. Accessed 14 August 2010
- Duerst M, Suignard M (2005) Internationalized Resource Identifiers (IRIs). Request for comments: 3987. Internet Engineering Task Force (IETF). <http://www.ietf.org/rfc/rfc3987.txt>. Accessed 13 August 2010
- PURL Federation (2010) Persistent URLs. <http://www.purlz.org/>. Accessed 14 August 2010.
- Furrie B (2009) Understanding MARC Bibliographic: machine-readable cataloging, 8th ed. Library of Congress. <http://www.loc.gov/marc/umb/>. Accessed 14 August 2010
- Garrity G, Thompson L, Ussery D, Paskin N, Baker D, Desmeth P, Schindel D, Ong P (2009) Executive summary: studies on monitoring and tracking genetic resources. *Standards in Genomic Sciences*, North America, 1, Jul 2009. Available at: <http://standardsingenomics.org/index.php/sign/article/view/signs.1491>. Date accessed: 17 Sep. 2010
- Hyam R (2009) A position on LSID's. <http://www.hyam.net/blog/archives/325>. Accessed 27 December 2010
- ISO7498-3 (1997) Data networks and open system communication OSI networking and system aspects—Naming, Addressing and Registration. International Standard ITU-T Rec. X.650 (1996) j ISO/IEC 7498-3:1997. <http://www.itu.int/rec/T-REC-X.650/en>. Accessed 14 August 2010
- ITU (1994) X.200: Information technology—open systems interconnection—basic reference model: the basic model. <http://www.itu.int/rec/T-REC-X.200/en>
- ITU (2005) Procedures for the operation of OSI registration authorities: general procedures and ASN.1 Object Identifier tree top arcs, ITU-T Rec. X.660 (2004) | ISO/IEC 9834-1:2005, <http://www.itu.int/ITU-T/studygroups/com17/oid/X.660-E.pdf>. Accessed 14 August 2010
- Kahn RE, Lyons PA (2006) Representing value as digital objects: a discussion of transferability and anonymity. *J Telecommun High Technol Law* 5(1):189–198. <http://www.colorado.edu/law/jthtl/Issues/Volume5Issue1.htm>. Accessed 14 August 2010
- Kahn R, Wilensky RA (2006) Framework for distributed digital object services. *Int J Digit Libr* 6(2):115–123. doi:10.1007/s00799-005-0128-x Previously published in May 1995. <http://hdl.handle.net/cnri.dlib/t95-01>. Accessed 14 August 2010
- Khoury JS, Jerez HN, Abdallah CT (2007) Efficient user controlled inter-domain SIP mobility authentication, registration, and call routing. Proceedings of the 4th Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, 2007. MobiQuitous 2007. doi:10.1109/MOBIQ.2007.4451064 http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4451064. Accessed 14 August 2010
- Klerkx J, Vandeputte B, Parra G, Santos JL, Van Assche F, Duval E (2010) How to share and reuse learning resources: the ARIAD-NE experience. Sustaining TEL: from innovation to learning and practice. *Lect Notes Comput Sci* 6383/2010:183–196. doi:10.1007/978-3-642-16020-2_13
- Kong MM (1995) DCE: an environment for secure client/server computing—Open Software Foundation's Distributed Computing Environment—Technology Information. Hewlett-Packard Journal
- Kunze JA (2003) Towards electronic persistence using ARK identifiers. <http://www.cdlib.org/inside/diglib/ark/arkcdl.pdf>. Accessed 3 August 2010

- Kunze JA, Rodgers RPC (2008) The ARK identifier scheme. Network Preservation Group. <http://www.cdlib.org/inside/diglib/ark/ark-spec.html>. Accessed 3 August 2010
- Larmouth J (1999) ASN.1 Complete. Morgan Kaufmann, London, p 472
- Larmouth J (2003) Registration authorities for OID components. <http://www.itu.int/itudoc/itu-t/com17/tutorial/84393.html>. Accessed 14 August 2010
- Leach P, Mealling M, Salz R (2005) A Universally Unique Identifier (UUID) URN namespace. The Internet Society. <http://www.ietf.org/rfc/rfc4122.txt>. Accessed 15 September 2010
- Library of Congress (1997) The relationship between URNs, Handles, and PURLs. <http://memory.loc.gov/ammem/award/docs/PURL-handle.html>. Accessed 14 August 2010
- Life Science Identifier Resolution Project (2010) GeekNet, Inc. SourceForge.net. <http://lsids.sourceforge.net/>. Accessed 12 August 2010
- LSID (2010) Wikipedia. <http://en.wikipedia.org/wiki/LSID>. Accessed 21 August 2010
- LSID-Developer (2010) GeekNet, Inc. SourceForge.net. http://sourceforge.net/mailarchive/forum.php?forum_name=lsid-developer. Accessed 21 August 2010
- Lynch C (1997) Identifiers and their role in networked information applications. <http://www.arl.org/bm-doc/identifier.pdf>. Accessed 14 August 2010
- Lyons WE (2000) Object identification and registration. http://isote.iso.org/livelihood/livelihood/fetch/2000/2489/Ittf_Home/MoU-MG/Moumg178.ppt. Accessed 14 August 2010
- Lyons S (2005) Persistent identification of electronic documents and the future of footnotes. *Law Libr J* 97(4):681–694. http://www.aallnet.org/products/pub_lj_v97n04/2005-42.pdf. Accessed 14 August 2010
- Maniatis P, Roussopoulos M, Guili TJ, Rosenthal DS, Baker M (2005) The LOCKSS peer-to-peer digital preservation system. *ACM Trans Comput Syst* 23(1):2–50. doi:10.1145/1047915.1047917
- McRae MP (2008) Failed OASIS standard ballot of XRI syntax v2.0. xri message. Organization for the Advancement of Structured Internet Standards (OASIS). <http://lists.oasis-open.org/archives/xri/200806/msg00001.html>. Accessed 21 August 2010
- Mealling M (2000) A URN namespace of object identifiers. Request for Comments: 3001. Internet Engineering Task Force (IETF). <http://www.ietf.org/rfc/rfc3001.txt>. Accessed 12 August 2010
- Miller E (2007) Eric Miller's home page. <http://www.w3.org/People/EM/>
- Moats R (1997) URN syntax. Request for comments: 2141. Internet Engineering Task Force (IETF). <http://www.ietf.org/rfc/rfc2141.txt>. Accessed 13 August 2010
- Name-to-Thing (N2T) Resolver (2007) <http://n2t.info/>. Accessed 13 August 2010
- NASA (2006) Requirements for archiving, distribution and user services in EOS Data and Information System (EOSDIS), 423-10-69 earth science data information system project. GSFC, Greenbelt
- Network Development and MARC Standards Office (1999) Library of Congress Control Number (LCCN)—restructuring to accommodate century change. Library of Congress. <http://www.loc.gov/marc/lccn.html>. Accessed 14 August 2010
- Network Development and MARC Standards Office (2006) Structure of the LC Control Number. Library of Congress. http://www.loc.gov/marc/lccn_structure.html. Accessed 14 August 2010
- OASIS XRI Technical Committee (2005) Extensible Resource Identifier (XRI) Syntax V2.0, Committee Specification, 14 November 2005. Organization for the Advancement of Structured Internet Standards (OASIS). <http://www.oasis-open.org/committees/download.php/15376>. Accessed 21 August 2010
- Object Identifier (2010) Wikipedia. http://en.wikipedia.org/wiki/Object_identifier. Accessed 14 August 2010
- OCLC (2010) PURLs: persistent uniform resource locators. <http://purl.oclc.org/>. Accessed 14 August 2010
- OID-Registry (2010) France Telecom. <http://www.oid-info.com/>. Accessed 14 August 2010
- Online Computer Library Center (2007) OCLC to work with Zephira to redesign OCLC's PURL service. <http://www.oclc.org/news/releases/200669.htm>. Accessed 12 August 2010
- Online Computer Library Center (2010) OCLS information and services for publishers. <http://publishers.oclc.org/en/default.htm>. Accessed 22 August 2010
- Parsons M, Duerr R, Minster JB (2010) Data citation and peer review. *EOS, Trans Am Geophys Union* 91(34):297–304
- Paskin N (2006) Naming and meaning of digital objects. In: Kia N, Atta B, Pierfrancesco B (eds) *Proceedings of the 2nd International Conference on Automated Production of Cross Media*
- Powell A (2004) Guidelines for assigning identifiers to metadata terms. Dublin Core Metadata Initiative. <http://www.ukoln.ac.uk/metadata/dcmi/term-identifier-guidelines/>. Accessed 14 August 2010
- Publication Manual of the American Psychological Association (2001), American Psychological Association, Washington, DC. <http://www.apastyle.org/>. Accessed 14 August 2010
- Pyle R, Michel E (2008) ZooBank: developing a nomenclatural tool for unifying 250 years of biological information. *Zootaxa*, 1950, 39–50. http://www.azoresbioportal.angra.uac.pt/files/publicacoes_Pyle%20et%20al_2009_Zootaxa.pdf. Accessed 16 March 2011
- Reed, D (2010a) Extensible Resource Identifier (XRI) Version 3.0, Working Draft 02. 18 January 2010. <http://www.oasis-open.org/committees/download.php/33876/xri-syntax-3.0-wd02.doc>. Accessed 11 January 2011
- Reed D (2010b) Personal communication with Ruth Duerr.
- Registry of Z39.50 Object Identifiers (2009) Z39.50 International standard maintenance agency. Library of Congress. <http://www.loc.gov/z3950/agency/defs/oids.html>. Accessed 14 August 2010
- Rust G (1998) Metadata: the right approach. *D-Lib Mag.*, July/August 1998. <http://www.dlib.org/dlib/july98/rust/07rust.html>. Accessed 14 August 2010
- Rust G, Bide M (2000) The<indecs>metadata framework: principles, model and data dictionary. WP1a-006-2.0, June. http://www.doi.org/topics/indecs/indecs_framework_2000.pdf. Accessed 27 December 2010
- Salamone S (2002) LSID: an informatics lifesaver. *Bio-IT World*. <http://www.bio-itworld.com/archive/011204/lsid.html>. Accessed 13 August 2010
- Smith AJ (2007) Developing handle system® web services at Cornell University. *D-Lib Mag.* 13(9/10). ISSN 1082–9873. <http://www.dlib.org/dlib/september07/smith/09smith.html>. Accessed 14 August 2010
- Smith D, Szekely B (2005) LSID best practices: a guide to deploying life science identifiers. <http://www.ibm.com/developerworks/opensource/library/os-lsdbp>. Accessed 21 August 2010
- Sun SX (2001) Establishing persistent identity using the handle system. *Proceedings of the Tenth International World Wide Web Conference, Hong Kong*. <http://www10.org/cdrom/posters/1136.pdf>. Accessed 14 August 2010
- Sun S, Lannom L, Brian B (2003a) Handle system overview. Request for comments: 3650. Internet Engineering Task Force (IETF). <http://hdl.handle.net/4263537/4069>. Accessed 14 August 2010
- Sun S, Reilly S, Lannom L (2003b) Handle system namespace and service definition. Request for comments: 3651. Internet Engineering Task Force (IETF). <http://hdl.handle.net/4263537/4068>. Accessed 14 August 2010
- Sun S, Reilly S, Lannom L, Petrone J (2003c) Handle system protocol (ver 2.1) specification. Request for comments: 3652. Internet Engineering Task Force (IETF). <http://hdl.handle.net/4263537/4086>. Accessed 14 August 2010
- Tansley R (2006) Building a distributed, standards-based repository federation: the China digital museum project. *D-Lib Mag.*, 12(7)

- 8). ISSN: 1082–9873 <http://www.dlib.org/dlib/july06/tansley/07tansley.html> Accessed 14 August 2010
- Taxonomic Databases Working Group (2006) Biodiversity information standards. http://www.tdwg.org/fileadmin/subgroups/guid/LSIDs_for_Biologists.pdf Accessed 21 August 2010
- TDWG Globally Unique Identifiers: LSID (2010) GBIF Wiki, Global Biodiversity Information Facility. <http://wiki.gbif.org/guidwiki/wikka.php?wakka=LSID> Accessed 21 August 2010
- The Chicago Manual of Style Online, Fifteenth Edition (2006) The University of Chicago Press, Chicago. <http://www.chicagomanualofstyle.org/home.html> Accessed <http://www.chicagomanualofstyle.org/home.html>
- The International DOI Foundation (2010a) The digital object identifier system. <http://www.doi.org/> Accessed 13 August 2010
- The International DOI Foundation (2010b) Overviews and standards. http://www.doi.org/about_the_doi.html#standards Accessed 13 August 2010
- The International DOI Foundation (2010c) TC 46 information and documentation. http://www.iso.org/iso/standards_development/technical_committees/list_of_iso_technical_committees/iso_technical_committee.htm?commid=48750 Accessed 13 August 2010
- The Open Group (1997) DCE 1.1: remote procedure call—Universal unique identifier <http://www.opengroup.org/onlinepubs/9629399/apdxa.htm> Accessed 15 September 2010
- Tilmes C, Yesha Y, Halem M (2010) Tracking provenance of earth science data. *Earth Sci Inform* 3(1–2):59–65. doi:10.1007/s12145-010-0046-3.
- Ts'o, T (2010) E2fsprogs: Ext2/3/4 filesystem utilities, sourceforge. Net. <http://e2fsprogs.sourceforge.net/> Accessed 11 January 2011.
- University of California Curation Center (UC3), California Digital Library (2010) Name Assigning Authority Number (NAAN) Registry. <http://www.cdlib.org/inside/diglib/ark/natab> Accessed 3 August 2010
- Wahl M, Coulbeck A, Howes T, Killis S (1997) Lightweight directory access protocol (v3): attribute syntax definitions. Request for Comments: 2252. Internet Engineering Task Force (IETF). <http://www.ietf.org/rfc/rfc2252.txt> Accessed 12 August 2010
- Weibel S (2007) PURLy gates and gift horses. Six Apart, Ltd., TypePad. <http://weibel-lines.typepad.com/weibelines/2007/11/purly-gates-and.html> Accessed 14 August 2010
- Willet P (2010) NOID: Nice Opaque Identifier (Minter and Name Resolver). <https://confluence.ucop.edu/display/Curation/NOID> Accessed 15 September 2010
- Williams, S (2008) TAG recommends against XRI from Williams, Stuart (HP Labs, Bristol) on 2008-05-21 (www-tag@w3.org from May 2008). W3C Public Mailing List Archives, World Wide Web Consortium (W3C). <http://lists.w3.org/Archives/Public/www-tag/2008May/0078> Accessed 21 August 2010
- Xiaofeng, G., Ying, L., and Sun, S. X. 2010. Federated content rights management for research and academic publications using the handle system. *D-Lib Magazine*, 16, 11/12, doi:10.1045/november2010-guo
- Zepheira (2011) <http://zepheira.com/about/people/eric-miller/> Accessed 21 January 2011
- ZooKeys (2010) Pensoft Publishers. <http://pensoftonline.net/zookeys/> Accessed 21 August 2010
- Zootaxa (2010) Magnolia Press. <http://www.mapress.com/zootaxa/> Accessed 21 August 2010