



# Persistent Identifiers for Earth Science Provenance

Curt Tilmes

*Curt.Tilmes@umbc.edu*

**eBiquity Research Group Presentation**

**February 25, 2009**

**UMBC** *eBiquity*

- Background
- Identification
- Persistence
- Actionable Identifiers
- (a few) Identifier issues
- Earth Science Data
- (some) Identifier schemes:
  - W3C: URI, URL, URN
  - UUID: Universally Unique Identifier
  - OID: Object Identifier
  - PURL: Persistent URL
  - DOI: Digital Object Identifier
  - XRI: Extensible Resource Identifier
  - ARK: Archival Resource Key
  - N2T: Name to Thing Resolver
- References

- ❑ Historically, published scientific research includes a description of the experiment that yielded the results – in sufficient detail to reproduce the experiment and get the same results.
- ❑ Reproducibility(among other things) -> Credibility -> Trust
- ❑ Modern research in earth science (and other fields) depends often involves sifting through mounds of data from a variety of sources (field sensors, satellite data, etc.) and applying various algorithms to reduce/transform/massage that data in various ways
- ❑ The data is likely the result of the work of hundreds of individuals over decades.
- ❑ Representing the provenance of such scientific results in a manner conducive to exploration, understanding and reproducibility is one of my interests.
- ❑ A key to this sort of representation is identifiers.

- ❑ All of the “artifacts” involved in the provenance of a scientific result
  - Data
  - Algorithms
  - Documentation
  - Sensors/Instruments/Instrument platforms
  - People (reputation)
  - Organizations (reputation)
  - Published scientific papers (add to credibility)
  - Computer systems
  - Abstract things like “a data transformation event” or “a validation experiment”
  - An ephemeral execution of a web service

- ❑ How do you identify a person?
- ❑ Consider Shakespeare's "*Romeo and Juliet*"
  - Is that a good "identifier"?
  - The intellectual content of the play
  - A published book with the play
  - A specific book (with a little jelly on page 32)
  - A performance of the play
  - A translation into another language
  - Can I cite an act, scene, line, page, paragraph?  
("microattribution")
  - ...
- ❑ The library folks have a very good handle on this for various content and media.

- ❑ “It is intended that the lifetime of a [persistent identifier] be permanent. That is, the [persistent identifier] will be **globally unique forever**, and may well be used as a reference to a resource well **beyond the lifetime of the resource it identifies or of any naming authority** involved in the assignment of its name.”

[http://www.doi.org/doi\\_presentations/overview\\_slides\\_4Dec2007/071205DOIOverview.ppt](http://www.doi.org/doi_presentations/overview_slides_4Dec2007/071205DOIOverview.ppt)

- ❑ My definition – I want the provenance web leading to a published component of the scientific literature to live as long as the publication is scientifically valid. (In fact, you can use a citation chain to determine when the identifier is no longer referenced in any way.)

- ❑ 'Actionable' Identifier = *Can I click on it?*
  - What happens if the resource is no longer around? We (NASA archive) delete old, obsolete data that takes up expensive space.
- ❑ Even if the data is gone, I'd still like to keep the identifier around...
- ❑ What happens if valuable data is moved from one “steward” to another? (We do this all the time...)
  - An entire archive taken over by another organization
  - A single dataset within the archive moved from one organization to another
  - What about data served from multiple locations?
  - What about data served in multiple formats?

- ❑ Data itself vs. a specific representation/format of that data
- ❑ Content invariance: Subject to correction? Subject to revision? What is the difference?
- ❑ Consider a reprocessing with identical inputs and algorithms, but in a slightly different computing environment..
- ❑ Does more than one identifier refer to the same resource?
- ❑ Can you compare two identifiers for equivalence?
- ❑ What happens when the resource itself moves?
- ❑ What happens when there is a new 'steward' for the resource?
- ❑ What happens if the resource physically resides in multiple places?
- ❑ What about “produce on demand”? Can be “real-time” or not..
- ❑ How are resources cited? discovered?
- ❑ Identifier standards, security, scalability, compatibility
- ❑ Dependent on central registries or authorities?
- ❑ Should identifiers be opaque or meaningful (include bits of metadata with semantics)? Structured with hierarchies? Can I predict a URI? ('OpenURL')

- ❑ Consider a published research paper that concludes some fact and says the data came from NASA instrument “MODIS”.
  - There are two MODIS instruments flying right now.
  - The set of standard products have 5 different reprocessing “collections”.
  - The data are too big to keep forever, usually all but the last two versions are deleted.
  - There are dozens of different algorithms that derive products from the captured data
  - A different calibration of the level 1 data can affect the level 2 and level 3 data

## ☐ Metadata

- We have a tremendous amount of metadata for the content of data (Date, Orbit number, quality flags, etc.)
- There is a distinct set of metadata for the content container (file size, file format, digital signature)

## ☐ Versions

- Every algorithm has strict configuration management with versions mapping to revisions (should have better documentation)
- What does “version” mean to data?
- Consider Algorithm X of version 1.2 is used to produce file A
- If we revise algorithm X and reprocess with version 1.3, the produced file A is different, we note in its metadata that it was produced with version 1.3
- Now what happens if we recalibrate the instrument that produced the data that was fed to algorithm X?

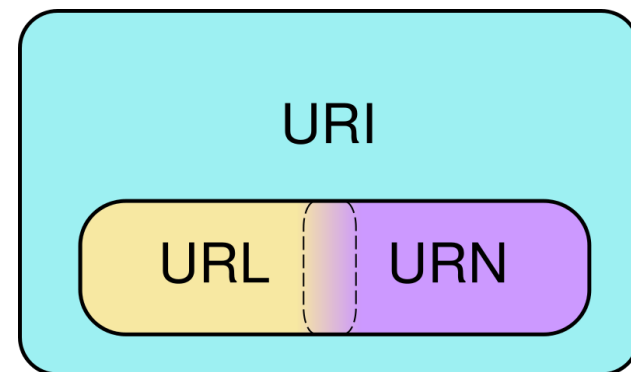
- ❑ URI – Uniform Resource Identifier
- ❑ URN – Name = What is it called?
- ❑ URL – Locator = Where can I find it?

**<scheme>:<scheme specific identifier>**

*http://example.org/something*

**urn:<namespace>:<namespace specific string>**

*urn:isbn:0451450523*



- ❑ A scheme for distributed systems to independently create unique identifiers without central coordination
- ❑ A 16-byte (128-bit) number (Enough to make 1 trillion UUIDs every nanosecond for over 10 billion years)
- ❑ Several different versions based on MAC address, time, hashing, random numbers.
- ❑ Canonical representation to make them easy to recognize:

*550e8400-e29b-41d4-a716-446655440000*

*urn:uuid:550e8400-e29b-41d4-a716-446655440000*

- ❑ Very Formal, ISO standard hierarchical naming scheme
- ❑ Think a truly global, universal directory tree similar to a unix directory tree. Any organization can register and get a “path” in the tree and populate it how they please.
- ❑ Several different notations in use:

```
{iso(1)member-body(2)f(250)type-org(1)ft(16)test(99)88}
1.2.250.1.16.99.88
oid://1/2/250/1/16/99/88
urn:oid:1.2.250.1.16.99.88
```

- ❑ Very simple indirect mapping that redirects from a PURL to a URL with standard HTTP redirect
- ❑ Includes “partial redirects” to relocate whole hierarchies
- ❑ Multiple PURLs could map to one URL (don't do that!)
- ❑ What about bookmarks?

**<scheme>://<PURL resolver>/<name>**

*http://purl.org/mypath/mydocs/mydoc*

- ❑ A framework for persistent identification
- ❑ A federation of “Registration Agencies” that control portions of the namespace
- ❑ RA pay fees based on volume to register DOIs

**<RA id>/<RA specific part>**

*10.1007/978-3-540-89965-5\_23*

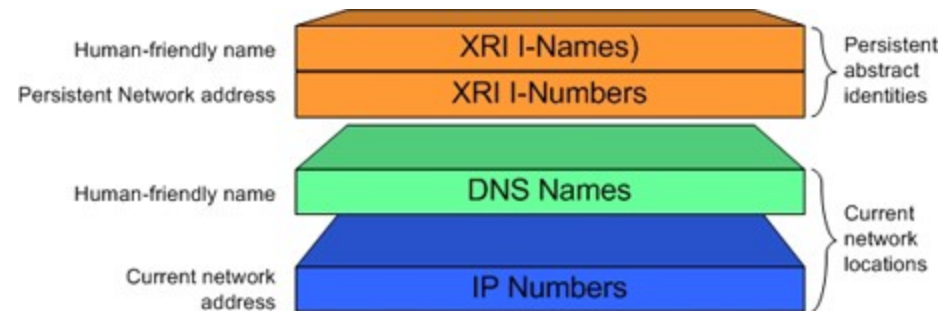
*http://dx.doi.org/10.1007/978-3-540-89965-5\_23*

*urn:info:doi:10.1007/978-3-540-89965-5\_23*

- ❑ From OASIS (Organization for the Advancement of Structured Information Standards)
- ❑ “Brokers” register with XDI.ORG and handle registration of identifiers (brokers are accredited and have “reputation”)
- ❑ I-numbers – machine friendly identifiers (like IP addresses) for any resource
- ❑ I-names – human friendly identifiers that resolve to an I-number
  - Various types =Person, @trademark, +anything

**xri://<authority>/<path>**

=*Curt.Tilmes*



<http://www.xdi.org/xri-and-xdi-explained.html>

- ❑ Hierarchical like URIs, but includes “cross-referencing”
- ❑ Fragments can be either persistent(!) or reassignable(\*)
- ❑ Formal methods for normalization and comparison
- ❑ Any URI could be a global authority, not just a “hostname:port”

*xri://@example.com/something*

*xri://(mailto:john.doe@example.com)/favorites*

- ❑ Apparently actively opposed by (portions of) W3C. Last standard proposal was voted down.
- ❑ *“We are not satisfied that XRIs provide functionality not readily available from http: URIs.”*

<http://www.equalsdrummond.name/?p=130>

- ❑ Scheme for Long-Term, Persistent Actionable Identifiers
- ❑ From California Digital Library, John Kunze
- ❑ Organizations must make a commitment to “long-term,” “persistent”, “actionable”
- ❑ PURLs, Handles, etc. add indirection, but not (necessarily) organizational commitment
- ❑ Goals:
  - Identifiers deliver you to objects (where feasible) (not a “404”)
  - Identifiers deliver you to object metadata
  - Identifiers deliver you to statements of commitment

- ❑ Name Mapping Authority Hostport (NMAH)= replaceable address that can be used to resolve the identifier (Kunze calls this the “booster rocket”)
- ❑ Name Assigning Authority Number (NAAN) = centrally registered list of organizations
- ❑ Append '?' for metadata and '??' for commitment policies
  - Human readable and structured for computers
- ❑ If the NMAH goes away, you can still find the object by checking central registry for new NMAH for the NAAN

**[ <NMAH> / ] ark : / <NAAN> / <Name>**

*ark:12345/myname*

*http://some.org/ark:12345/myname*

*http://other.org/ark:12345/myname*

- ❑ <http://n2t.info>
- ❑ Very simple resolver mirrored by consortium volunteers under one hostname

`http://n2t.info/<global prefix>/<local part>`

<code>http://n2t.info/NAA/...</code>	<i>(NAA = N2T NAA Number, eg, 12345)</i>
<code>http://n2t.info/ark:/NAA/...</code>	<i>(NAA = ARK NAA Number, eg, 12345)</i>
<code>http://n2t.info/urn:NAA:...</code>	<i>(NAA = URN Naming Authority, eg, nbn)</i>
<code>http://n2t.info/hdl:NAA/...</code>	<i>(NAA = Handle Naming Authority, eg, 12345)</i>
<code>http://n2t.info/doi:NAA/...</code>	<i>(NAA = DOI Naming Authority, eg, 10.12345)</i>
<code>http://n2t.info/purl:/NAA/...</code>	<i>(NAA = PURL Resolver Hostport, eg, purl.org)</i>

- ❑ "Cool URIs Don't Change." <http://www.w3.org/Provider/Style/URI>
- ❑ "Naming and Addressing: URIs, URLs, ..." <http://www.w3.org/Addressing/>
- ❑ "Object Identifier (OID)" <http://www.oid-info.com/>
- ❑ "The Digital Object Identifier (DOI) System." <http://www.doi.org/>
- ❑ "Persistent Uniform Resource Locator" <http://purl.org/>
- ❑ " A Universally Unique Identifier (UUID) URN Namespace"  
<http://www.ietf.org/rfc/rfc4122.txt>
- ❑ "XRI (Extensible Resource Identifier)"  
<http://www.xdi.org/xri-and-xdi-explained.html>
- ❑ "ARK (Archival Resource Key)"  
<http://www.cdlib.org/inside/diglib/ark/arkspec.html>
- ❑ "Name-to-Thing (N2T) Resolver" <http://n2t.info>

Thanks to NASA ESDSWG and Ruth Duerr of NSIDC