The Realities of Implementing Identifier Schemes for Data Objects **[1]**

Submitted by superadmin on Fri, 2012-06-29 16:35 Thursday, July 19, 2012 - 15:30 to 17:00
Event: Summer Meeting 2012 [2]
Session Type: Breakout [3]
Expertise Level: Intermediate [4]
Identifier: doi:10.7269/P3C8276K
Collaboration Area: Data Management Training [5]
Data Preservation [6]
Preservation and Stewardship [7]
Products and Services [8]
Abstract/Agenda:
Study of nine different identifier schemes by the Data Stewardship Committee resulted in recommendations for use of certain schemes for data objects. The recommendations were tempered by the caveat that implementation of the nine schemes could change the recommendations. For the second phase of the study. a test was set up to implement the nine identifier schemes by assigning identifiers to two different data sets and at least some of the components of the data sets. This presentation will present the findings based on that test for discussion by the various ESIP committees and clusters interested in furthering the discussion of identifiers for data objects, and for non-data objects.

Notes:
Identifiers session

Nancy and Greg

Implementing ID Schemes for Data objects.

Introduction/background –

Data stewardship and product services works

Data stewardship cluster was interested in what identifiers could be used for data objects and came out with 4 use cases

1) for unique identification 2) unique location, 3) citable location and 4) scientifically unique identification. These developed 8 schemes and resulted in a paper – *On the utility of identification schemes for digital earth science data …*
http://www.springerlink.com/content/52760gq3h200gw38/?MUD=MP [9]

A few samples of what questions were asked – how scalable is it, will publishers allow it, how maintainable is it when migrating data? Etc. They created a full table, which is on the data stewardship wiki –

http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Identifiers/Table [10]

They then thought about implementation of these schemes and the recommendations were feasible, in a test bed on a few different data sets. To see what the issues were and if this was an impact on the initial recommendations.

Two different data sets were used, glacier photo collection – an image collection; and a numerical

time series set from MODIS.

The full document can be viewed on the wiki –

http://wiki.esipfed.org/images/b/bf/OrigRegs.pdf [11] (does nto work, not right link).

Operational questions asked – full list is on the wiki, examples were given, relating to archival issues and discovery. They also have answers, but not all have been documented at this time.

Now on to DOI's – Greg.

Greg is going to talk about the different identifier types and the issues they saw along the way. Particularly they are going to focus on DOI's. Commonly used in journal publications. If you want to create a DOI how do you do it? There is an international IDF that distributes these. And they in turn hand off registration to a handful of agencies, crossref is one of them, and a new one, DataCite which is for data citations. This is handed off to allocators. CDL – California Digital Libraries is one of them. You sign a contract, with user responsibilities that the DOI points to something and you have to pay a subscription fee. They used to be charged per DOI but that created a barrier to some.

DOIs – they made data set level identifiers for their two test collections. They used EZID to create them, http://n2t.net/ezid/ [12] and these can be resolved to URLs or within the EZID system.

They registered these with some citation metadata as well, which is maintained in EZID and used for searching in other systems.

They also looked at assigning IDs at the granule level. They needed URL's or something that could point to where these granules would live. This worked well in the glacier photos but for the MODIS, URL needs to be specific and tied to one granule but this was not entirely true for the MODIS data. The granules did not have a stable web URL. DataCite however requires "landing pages" for URLs, something that describes the resource and where it might actually be, but this is currently not being enforced. Landing pages, and not to perhaps a file or other data itself.

Question – someone mentioned they have links to PDFs but asked if DataCite is asking for html pages being linked. Greg – it has not been fully defined, but Curt mentioned that this has been discussed quite a bit as well, so this should be noted. Greg – this is something that should be considered when doing granular level work.

Question – what if there are two sites that have the data published, distributed, are all cases using the same DOI? And what happens if the data is slight changed; does the DOI remain the same? Greg – that is a good discussion topic, and the paper they wrote, identified this as an issue. These identifiers do not answer this question. Commenter, this is for publications, so if I get a DOI, do I

give credit to where I found the DOI, or who created the data set? Stephen/Nancy – you are often given an example on how to cite it, and there is metadata included which gives more information about the data standard.

Another point about assigning DOIs to granule's is scalability. Do you assign unrelated (generated) names? Each granule is independent of each other. But this requires some maintenance of what identifier relates to what granules.

Another option is to start with a base identifier, and add a local name which identifies the granule. Curt – is the base identifier the identifier for the collection? Greg – it does not have to be that way but it can be. Cant rename the granules thought, but it does not require a database to manage the set.

Third option, base URL with a partial redirect. Only one identifier to manage but the granules must be managed as a group. This is powerful because management is difficult but now we have tied all of those granules together. This does not preclude doing the second approach at a later date. But the DOI does not currently support this method, while PURL's do, and so does Handles 7+.

Question - in either of the second two, can you have a granule that is part of more than one collection set? Greg – we are identifying the granule and that is independent of the collection.

Metadata – citation metadata (a limited set with controlled vocabulary) is required. Citable identifiers – you could have a DOI with no associated metadata, but the publishers, have been talking to DataCite and EZID to harvest the metadata.

Question – if you have an operational data set, and it looks like the granule is a data file. If I am doing research over a year, when I cite that data, do I cite one per day file that I use? Greg – that is a good question and ESIP has come up with recommendations for data citation guidelines. The recommendations is to cite a data cite with qualifications about the subset. Question – so you would have one for the data set and one for each day? Ruth – no, you would cite the collection, and you would detail that you used specific dates within that collection. Ruth – you would only have individual citations if it made sense to do so for your collection. SO for the photo collection there was rich metadata that distinguished them from each other. Whereas the other collection, MODIS, has 1.6 million granules, but it does not need unique identifiers for each item. You would cite the collection and explain what part of the collection you used.

Should you assign persistent identifiers to granules? It really depends on what is being done with it.

ARKs – archival resource keys. Developed and maintained by the CDL. These are similar to DOIs. They can be embedded in URLs and they have hierarchical parts, they are globally supported and support having associated metadata and finally the way to create them is through EZID, the same idea as DOIs.

Differences between ARKs and DOIs. Table on the slides, DOIs are more supported, more robust. ARKs are only maintained by CDL only, but NSCENT is expanding. Costs – DOIs are no longer charged but they still have costs involved and in CDL ARKs are cheaper. Curt – the costs is not related to the numbers? Nancy – this is dependent on the size of your organization. Curt – I thought it had to do with 10 a year vs. 1 millions. Nancy – No, that is not it. It depends on the size of the organization and their funding. So NASA would be more than a nonprofit. But they are just doing cost recovery and this is still reasonable costs.

ARKs are not tied to a landing page. DOIs have required metadata, ARKs do not have that but supports them as optional. DOIs are typically published things while CDL is not putting constraints on ARKs. Finally acceptance, publishers are more likely to accept DOIs but there is some interest in ARKs. This is greatest in terms of citations.

Curt – we had some discussion of a hybrid scheme – did you do any exploration of that idea? Nancy – no we did not.

PURLs – persistent URLs. Go to Purl.org, create a domain – this has to be unique, can't use something someone was already using. This was a hard process. The persistent identifier with a organization name in the identifier is a complicated process. If the ownership changes that can be awkward and difficult.

Created a few exact purls as well as some redirect purls for the glacier photos. The registration took seconds, they used a batch API to register the collection and the documentation for that was incorrect and it was a very difficult process. Finally, there was no method for association citation or metadata information with the purl. What happens if it is broken? Maybe you can figure out what it was pointing to if you had some metadata. This is actually kind of key.

UUIDs – not a lot to say about these. The real use case for these are unique identifiers embedded during the creation of the object. But you cannot use it as a locator, usually used in programing languages.

Question – is there any question to update metadata in DOIs and ARKs? Greg – yes, and there are APIs to do that.

OIDs – long dotted strings of numbers. Each node in the tree is the authority for a sub tree. A distributed authority system. There is no global infrastructure for these ideas. No authoritative. And during distribution of these, there is check to see who has ownership. It feels a little naked to be using OIDs. So how do you create one? If there is no system. What is the value for creating something in this system? How do you change ownership when it is built in to the number itself?

There is an OID repository which registers an UUID to an OID, this is human mediated.

As they tried to use the system – to create an OID for the glacier photo system – they got changes

made to their description and naming convention. For the MODIS they were told it can only be registered for an organization. So they did that for the NSIDC, so they could register a node and any sub node requests would go through them… but they never received any of the requests they sent to themselves for sub nodes! So they gave up.

XRIs – contentious. They were not really sure what is going on with them. The W3Cs rejected them and there are not any updates since 2008. The resolver for this system redirects to inames.net, supposed to be replacement for DNS, but as of now, major I names like google and facebook are open and free, so it must not be being used. There is no benefit over DOIs.

Handles – the handle system defined by RFCs, and are the same system as DOIs but DOIs start with 10. They are run on local servers; you have to register your local server and if any requests are made it gets sent to your local server. There are mirrors etc but it works. There are some firewall issues. The software comes with a java client, and it works great. The handle system supports metadata which is used for administration, but otherwise it is not being used in the Handle system. They are not actually storing it in the handle system when you do searches. There does not seem to be much value in this option.

LSID – life science identifiers, URN and a domain name of the owner of the resource. There is no global infrastructure, but it is resolved through DNS. Does not have a central registration system but still works. In testing these, they were intended to be layered on existing databases, and there were multiple installations and packages and validation programs. They are problematic do to the naming convention for use. There is also low adoption rate in the community.

Matt – just in case, it is up to the resolver to determine how to resolve it and most people were using a DNS, but you can use another different system and that made it a bit controversial. Does not think it worked out.

What did we learn from all this? The most suitable identifiers are DOIs, ARKs, and Handles. All three are unique locators, and go to a target URL. DOIs and ARKs are probably better for citations, given that they have metadata in some form so publishers can harvest that and Handles require more investment and costs. For Handles, you have to run a dedicated server and maintain it in perpetuity.

UUIDS best as unique identifiers

LSIDs possibly but unsuitable for locators.

PURLs no support and poor API.

Lease suitable, XRIs and OIDS.

Question – what about versioning, did you look at that? Nancy and Greg – we did not and perhaps we have been remiss on that. Ruth – there are some thoughts on that in the recommendations and guidelines. Curt – in the supplemental materials there is something about how to handle them. Ruth – it was nice that the testing confirmed the results of the paper.

Matt – who is ratifying the guidelines? Ruth – it started in this group when this was a cluster, then it went to the ESIP community and Geodata2011. Public data was solicited. In January, every ESIP member has a vote and it was voted on at the ESIP assembly meeting.

Matt – this is interesting, this is the first time I have heard of ESIP acting as a standards body – Ruth – we are not standards, and these are recommendations. Nancy – lots of vetting. Curt – the resolution also says it is a living document that this committee is required to maintain. It will be maintained and active. Bring up issues to the committee. Nancy – the same with submissions guidelines. Ruth – it would have to go to the entire assembly for vote.

Ruth – asked if this is being published to the wiki? Nancy – put some on there today, but they are planning on doing something more formal then that.

Next steps – finished summary spreadsheet with questions – it is not complete, and that will probably be on the wiki but it will also be included in the report that they are writing for data stewardship, and vetting it there and then ESIP as well. We might try and publish it in the same venue as the previous report.

Question – based on these findings and these recommendations – the way this community seems to be heading, do we have a sense of how this will change as we scale out? In costs and management of how many DOIs will be minted if we are talking about granule level DOIs? And are there similar works in other communities like archeology? Nancy – we have not looked at that, we scoped it based on the questions in the initial report. But it has come up, and we tried to address it as much as possible, but there are a lot of things to be done. And social science data sets are asking the same questions. That is something that we might want to continue to monitor in this data stewardship community.

Curt – Nate has some slides on NASAs implementation of these recommendations.

Nate – from ESDIS project, they have a wiki

Not sure if this is the right link:

http://earthdata.nasa.gov/wiki/main/index.php/Main_Page [13]

Each DOI should have a unique landing page, will be assigned to each new product or version, ESDIS will manage this, they have an agent number and are minting DOIs. Trying to develop one approach across ESDIS. As long as there are changes needed, they would be maintained by one location, but that is not currently done, it is done by hand but will be automated. Will use DOI suffix syntax – [mission]/[instrument]/data[m][n]

M is the processing level of the product. N is the date.

DOI for each new product, starting with 7 missions. Will embed the DOI in the product itself and its metadata. Will have a new landing page for each product. But also will create citation web pages too. Will add DOIs to the list of documents being archived. Shoulder concept – enables creation of non-interfering data domains. When changes at the data level, they refer to a particular data center and they will manage their "shoulder".

Ruth – there are various terms I have been hearing, NSIDC has 20 MODIS sets they are achieving, some are sea some are snow, and how many DOIs is that? 2 or is that 20? Nate – you would have a local management of that at NSIDC, if you decide that you want it to be 20, and then you would request 20. It is responsible for those products to determine what makes sense. Is it used at the granule level or is it always used as a collection?

Curt – the ESIP recommendation is per product per collection. So the individual references in a paper can refer to the product. Ruth – so right answer is 20.

Curt – ESIP is working with Nancy and company for the commons, for items that have the level of maturity that they need a DOI. We have a data stewardship planning session tomorrow morning, to discuss what we need to investigate until the next ESIP meeting. We might want to note for investigation at that meeting, and identifiers is just one of many. W3C providence group is moving to their last call on interoperability specifications. We need an Earth science extension to that, so we want to feel out the community to see how we want to approach that. Please try to make that meeting. We would really like input from other people about specific issues we could address as a committee.

Ruth – other topics, or change or moving forward, tomorrow's meeting would be the place to bring that up.

Session Leads:

Name: Nancy Hoebelheinrich [14]
Organization(s): Knowledge Motifs [15]
Email: nhoebel@kmotifs.com [16]

Notes takers:

Name: Sarah Ramdeen [17]
Organization(s): School of Information and Library Science UNC-CH [18]
Email: ramdeen@email.unc.edu [19]

Teaser: Study of nine different identifier schemes by the Data Stewardship Committee resulted in recommendations for use of certain schemes for data
Keywords: testbed [20]

Source URL: https://commons.esipfed.org/node/459

Links
[1] https://commons.esipfed.org/node/459
[2] https://commons.esipfed.org/event/summer-meeting-2012

[3] https://commons.esipfed.org/session-type/breakout
[4] https://commons.esipfed.org/taxonomy/term/261
[5] https://commons.esipfed.org/collaboration-area/data-management-training
[6] https://commons.esipfed.org/collaboration-area/data-preservation
[7] https://commons.esipfed.org/collaboration-area/preservation-and-stewardship
[8] https://commons.esipfed.org/collaboration-area/products-and-services
[9] http://www.springerlink.com/content/52760gq3h200gw38/?MUD=MP
[10] http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Identifiers/Table
[11] http://wiki.esipfed.org/images/b/bf/OrigRegs.pdf
[12] http://n2t.net/ezid/
[13] http://earthdata.nasa.gov/wiki/main/index.php/Main_Page
[14] https://commons.esipfed.org/node/394
[15] https://commons.esipfed.org/taxonomy/term/285
[16] mailto:nhoebel@kmotifs.com
[17] https://commons.esipfed.org/node/557
[18] https://commons.esipfed.org/taxonomy/term/373
[19] mailto:ramdeen@email.unc.edu
[20] https://commons.esipfed.org/taxonomy/term/289