Normalizing Resource Identifiers using Lexicons in the Global Change Information System

Linking Earth Science Identifiers, Concepts, and Communities

Brian Duggan[14]
bduggan@usgcrp.gov

Curt Tilmes[2]
curt.tilmes@nasa.gov

Steven Aulenbach[14]
saulenbach@usgcrp.gov

Robert E. Wolfe[12]
rewolfe@usgcrp.gov

Justin C. Goldstein[14]
jgoldstein@usgcrp.gov

Gerald Manipon[3]
geraldjohn.m.manipon@jpl.nasa.gov

[1]US Global Change Research Program
1717 Pennsylvania Ave NW
Washington, DC 20006

[2]NASA Goddard Space Flight Center
8800 Greenbelt Rd
Greenbelt, MD 20771

[3]NASA/Jet Propulsion Laboratory
4800 Oak Grove Dr
Pasadena, CA 91011

[4]University Corporation for Atmospheric Research
P.O. Box 3000
Boulder, CO 80307

ABSTRACT

Earth Science informatics involves collaboration between multiple groups of people with diverse specializations and goals, often using variations in terminology to refer to common resources. The uniformity of the resource identifiers often does not cross organizational boundaries. Because of this, permanent, widely used, unambiguous identifiers for resources are elusive. We examine real world cases of changing and inconsistent identifiers which inherently work against persistence and uniformity. We also present a solution which mediates factors in these situations; namely the creation of lexicons: mappings of sets of terms to URIs which are curated within the Global Change Information System (GCIS).

We discuss aspects of the GCIS which facilitate the use of lexicons: an information model which disambiguates resources, a RESTful API which provides metadata through content-negotiation, and a strategy for long term curation of URIs, including mechanisms for handling changes to URIs and variations in terms used by different communities while providing persistent URIs and preserving relationships between resources.

We provide working definitions of *term*s, *context*s, and *lexicon*s, and relate them to the practical challenges of disambiguation and curation. We also discuss the mechanisms employed and architecture of the GCIS, and how these choices facilitate representation of persistent identifiers and mappings of them to identifiers used colloquially within various earth science communities of practice.

Copyright is held by the author/owner(s).
WWW2015 Workshop Linked Data on the Web (LDOW2015)
This material was developed with federal support through the U.S. Global Change Research Program under National Science Foundation contract no. NSFDACS13C1421.

Keywords

Linked Data, URI, Co-reference

1. INTRODUCTION

1.1 Background

The U.S. Global Change Research Program (USGCRP) was established in 1989 by Presidential Initiative and mandated by the U.S. Congress in the Global Change Research Act (GCRA) of 1990 to "assist the Nation and the world to understand, assess, predict, and respond to human-induced and natural processes of global change."[1] The USGCRP has recently sponsored the creation of the Global Change Information System (GCIS) to better coordinate and integrate the use of federal information products on changes in the global environment and the implications of those changes for society.

In May, 2014, the USGCRP released the Third National Climate Assessment (NCA3). This 800 page document, authored by 300 people, each of which are affiliated with multiple organizations, has 30 chapters, 161 findings, 290 figures and 3,395 references. The references refer to publications, including government reports, peer-reviewed scientific journal articles, and books. The publications are often supported by datasets that could be based on observations, measurements, processed or derived data, or model projections. Model projection datasets are created by runs of models constrained by scenarios. Observations and measurements are taken using instruments on platforms. The information in the NCA3 provided a starting point for the contents of the GCIS. The GCIS information model includes representations of reports, chapters, findings, figures, people, organizations, references, publications, platforms, instruments, models and scenarios. The GCIS was used to support the

production of the report and the dissemination of the web version of the report.

1.2 Motivation

A key design goal of the GCIS was to disambiguate and identify distinct resources, and provide references to sources of information. Other goals included: supporting scientific traceability and reproducibility, facilitating the creation of reports such as the NCA3, providing a scalable backend for rich web versions of the NCA3 and other reports, providing an API for other types of applications, providing the ability to run structured queries about disparate types of earth science information, and facilitating the discovery and representation of connections between earth science information managed by independent organizations.

The GCIS has been implemented as a RESTful API, whose endpoints are URIs which are part of a knowledge base formalized by an ontology and distributed using a SPARQL endpoint to a triple store. The API supports content negotiation, and the HTML representations form a navigable web site.

1.3 Lexicons

This paper focuses on the concept of lexicons and the processes and techniques involved in the creation and maintenance of mappings from persistent URIs to pre-existing Earth Science identifiers. In particular, we discuss challenges and techniques for dealing with colloquial identifiers (*terms*) which are often specific to communities of practice. We also discuss our techniques for maintaining long term persistent identifiers, and working with changing or inconsistent terms.

2. RELATED WORK

The general problem of disambiguation of resources has been known for some time, dating back at least to Leibnitz's formulation of the *Identity of Indiscernibles* in 1686 [5].

In 2006, the WWW's Technical Architecture Group addressed issues surrounding identification and URIs by distinguishing between *resources* and *information resources*. An HTTP request for a resource can return a 303 ("See Other") response which directs a user to an information resource [10]. The former in general does not have a representation which can be transmitted over HTTP, whereas the latter does.

As noted in [3], a service which provides sufficiently descriptive information about resources can provide disambiguation services.

[8] describes some of the difficulties of using automated techniques to discover equivalent identifiers. Despite these difficulties, various sophisticated attempts have been made, such as [11] and [12]. Besides automatic classification, an alternative technique has been the creation of a declarative language for resolving identifier ambiguities [4].

Once there are unambiguous URIs, if these can be mapped between RDF datasets, the mapping forms a "linkset" [2] which can be distributed, harvested and used. The representation of linksets can be further refined with careful use of "owl:sameAs" [6] and other relationships.

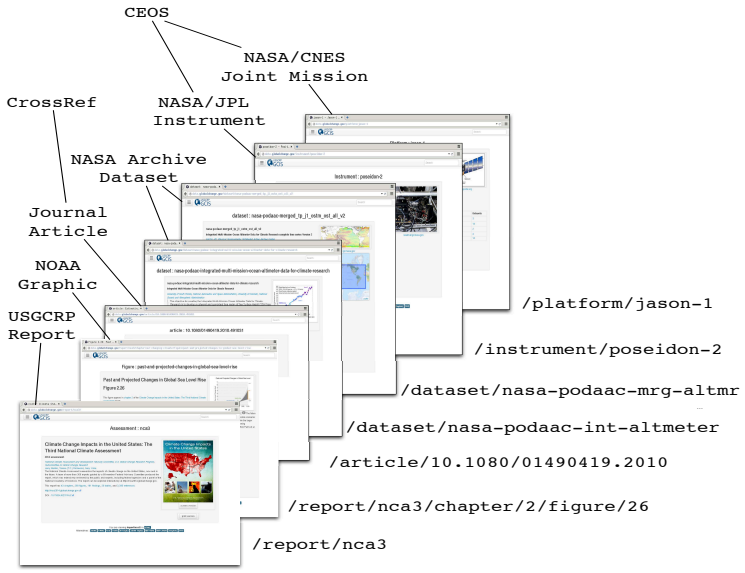

Figure 1: Organizations, Contexts, Identifiers: Past and Projected Changes to Global Sea Level Rise Figure in the Third National Climate Assessment

Various phrases have been used to describe this problem: the co-reference problem, the identity problem, disambiguation, and more. We chose the phrase "normalization" following its usage with relational databases and character encodings.

3. EXAMPLES

3.1 Traceability

Figure 2.26 of the Third National Climate Assessment[1] depicts Past and Projected Changes to Global Sea Level Rise (see figure 1). A sequence of resolvable URIs within GCIS traces this figure to a journal article which used a dataset which used another dataset, which was captured by an instrument on a platform. Each step in this sequence has a permanent URI within GCIS, but also refers to identifiers outside of GCIS; the journal article has a DOI managed by a publisher and resolvable using CrossRef, the first dataset has a URL managed by scientists, the second dataset has an identifier managed by a NASA Data Archive, the instrument and platform have identifiers created by the Committee on Earth Observing Satellites (CEOS). While all these organizations do provide machine-readable versions of their data, the identifiers are often curated independently. However, within each organization, there are terms which unambiguously identify a particular resource.

3.2 Identification

In some situations, there may be intentional uses of different names by different organizations. For instance, the "SAC-D/Aquarius" mission may also be called "Scientific Application Satellite-D"

> SAC-D/Aquarius is a cooperative international mission between CONAE (Comisión Nacional de

[1] `http://data.globalchange.gov/report/nca3/chapter/2/figure/26`

> Actividades Espaciales), Argentina, and NASA, USA. NASA uses the term [SAC-D/Aquarius] for the mission [..] At CONAE, which provides the spacecraft, the mission is referred to as Scientific Application Satellite-D [..] [9]

This is a clear case in which two different communities refer to the same resource differently. Such differences can propagate from narrative descriptions into identifiers found in serializations of information. When this happens, techniques for reconciling the identifiers become necessary.

3.3 Synchronization

Organizations in the domain of Earth Science distribute data, metadata, and information using a variety of serializations and interfaces, including:

ECHO 10	NASA Earth Observing System (EOS) Clearing House (ECHO)
ISO 19115	Geographic Information Metadata
FGDC	Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata
DIF	Directory Interchange Format, NASA's Global Change Master Directory (GCMD)
DCAT	W3C Data Catalog Vocabulary
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
CSV, JSON, YAML	Miscellaneous Serializations

Information conveyed varies by the standard and by the implementation. Despite these differences, the defining characteristics of resources generally remain consistent across representations. This enables a subset of information to be brought into the GCIS for the purpose of identification and disambiguation. In other words, when deciding what to harvest from other sources, a critical question is: does this piece of information help to distinguish this resource from other similar resources?

3.4 Communities

Science teams working on remote sensing missions produce scientific data and send them to data archives. A primary concern of archives is to maintain the fidelity of the science data as they are received. Because of this division of responsibilities, it's not clear which community would be better suited to take on the task of harmonizing identifiers across data producers. Instead differences in choices of identifiers may be passed on to end users of the data.

4. CONCEPTS

4.1 Terms as Identifiers

Issues surrounding characters, case folding, encoding, and strings are often omitted when resources are identified in narrative situations. This creates ambiguities in representations and postpones normalization issues. This motivates our definition of the word "term" which is based on the Universal Character Set (UCS).

Definition 1. We define a *term* to be a sequence of characters from the Universal Character Set (UCS) which is used as an identifier for a resource by a group of people.

The linked data glossary [7] defines a "term" using the notion of a controlled vocabulary, and also defines a controlled vocabulary using the notion of a term. We explicitly make use of Unicode characters in order to avoid circular definitions like this.

Within the GCIS, terms are encoded in UTF-8 and potentially normalized with Normalization Form C (NFC). We note that W3C recommendations for string matching are still evolving [13].

4.2 Communities of Practice

Within communities of practice, terms are created and used as part of the activities and communication between members of the community. Because of this, disambiguation of terms across communities is a secondary consideration. Within a community, a resource can be identified clearly when it is being referenced in a manner which is consistent with other similar resources. With this in mind, we group together terms using the type of resource.

Definition 2. We define a *context* to be a set of terms used to identify resources of the same type.

In example 3.1, "mission" and "instrument" are contexts.

We then define a lexicon by putting together the contexts.

Definition 3. We define a *lexicon* to be a set of contexts used by a particular community.

Returning to example 3.1, the CEOS lexicon uses "Mission" and "Instrument" contexts to group together terms.

4.3 GCIDs

As previously noted, entities in the GCIS are identified uniquely using a URI. The URI for a particular entity is called a GCIS Identifier, or GCID.

Definition 4. The *GCID* is the URI for an entity in the Global Change Information System.

Note that any GCID can be used in SPARQL queries, and can also be resolved using the GCIS as an endpoint and using content-negotiation.

One organization's "instrument" may be another one's "sensor". One organization's "platform" may be another one's "mission" or a third one's "source". We use lexicons to represent the way NASA's Physical Oceanography Active Data Archive Center (PODAAC), ECHO, GCMD and CEOS all refer to the same resource:

```
Lexicon | Context       | Term      | GCID (*)
-------------------------------------------------------
podaac  | Source        | JASON-1    | /platform/jason-1
ceos    | MissionId     | 286        | /platform/jason-1
gcmd    | prefLabel     | JASON-1    | /platform/jason-1
echo    | ShortName     | JASON-1    | /platform/jason-1
podaac  | Sensor        | POSEIDON-2 | /instrument/poseidon-2
ceos    | InstrumentId  | 182        | /instrument/poseidon-2

(*) under http://data.globalchange.gov
See also: http://data.globalchange.gov/lexicon
```

4.4 Linksets

When a term is an identifier which is part of another triple store, we can use "owl:sameAs" to connect the two identifiers and form a linkset. We treat dbpedia as a lexicon with a single context ("resource"). An application of this is writing a federated SPARQL query to compare crowdsourced information in dbpedia (or wikidata) to authoritative information from CEOS. This can belp improve the quality of the data in both places.

5. IMPLEMENTATION

5.1 Lexicon Interface

5.1.1 Creating, Updating

Creating a lexicon involves several steps:

1. Identifying a distinct set of terms.
2. Choosing a context.
3. Choosing an existing lexicon or adding a new lexicon.
4. Associating the terms with GCIDs.

An example of performing step 3 in an HTTP transaction follows; in this example we are associating "Aqua" a term used by CEOS, with the GCID "/platform/aqua" and the context "Mission":

```
PUT /lexicon/ceos/Mission/Aqua
Host: data.globalchange.gov
Content-Type: application/json

{ "gcid" : "/platform/aqua" }
```

An alternative interface is available which allows the terms and context to be sent in the payload, rather than the URI.

5.1.2 Querying

Looking up a term using a lexicon involves sending a GET request for an IRI containing the context and the term, and receiving a status code of 303 and a Location header, to indicate the corresponding GCIS URI, if one exists.

```
Request:
    GET /lexicon/ceos/Mission/Aqua HTTP/1.1
    Host: data.globalchange.gov
Response:
    303 See Other
    Location: /platform/aqua
```

5.2 System Architecture

The GCIS architecture incorporates elements of relational and semantic systems: cascading updates, referential integrity, strict type checking and other well-established features of relational databases are all valuable in maintaining the quality of the URIs and their relationships.

HTTP requests to the GCIS RESTful interface are handled by querying this relational database. The database contains simple explicit tables for resources that use natural identifiers as primary keys. JSON structures are formed using the names of the columns as names of the keys in the JSON objects. There is also a generic table which may be thought of as a parent table for many of the tables (i.e. using table inheritance).

Relationships between resources are stored in one of two ways: 1. Foreign keys between base tables. 2. Relationships between two entries in the generic table, via a mapping table, which may be annotated with a semantic relationship.

Triples are generated data from the tables to fill in text templates which output turtle. The turtle is parsed and used to populate a triple store. The triple store is rebuilt weekly; there are no incremental updates.

5.3 Terms

When new terms appear, they are captured in the GCIS. Scripts run periodically and pull information from various sources. The mechanism for assimilating new terms is to provide operators with notifications of unmatched terms; operators then manually associate the new terms with GCIS resources.

Example 1. A new satellite achieves orbit. The information in CEOS reflects this new information. It is pulled into GCIS, and a new entry is created. A new default GCIS identifier is created from a descriptive field (but it may be updated, as described below). The relational database is populated with information that reflects that data source. Since no term yet exists for this in the ceos lexicon, a new one is created which maps the term used by CEOS to the newly created identifier in the GCIS.

Example 2. An instrument on a spacecraft begins to produce new data. A science team processes the data and sends them to an archive. The archive makes the data available. In this case, a new term has been created, and it must be matched to a GCID manually. The new data is ingested, no match occurs. An operator notices this and manually associates the new term with an existing GCID.

Example 3. Two organizations use different terms for one satellite. Both organizations distribute data collected by an instrument on board the satellite. In this case, there are two lexicons, one for each organization, and each one has a context which has terms which map to the same GCID.

5.4 URIs

5.4.1 Creation

URIs are formed using primary keys in the relational database. The values of the columns comprising the primary key are

assembled along with the name of the table, into a URI which corresponds to a row of data in a table. The representation of a resource may involve joining to other related tables.

5.4.2 Persistence

When the value of a primary key column is changed, a cascading foreign key update will change the values in any related tables. Also, triggers will change the columns in the generic (parent) table. Because the templates are based on information in the database, these changes will automatically propagate to the semantic representation of the resource.

Also, for any changes, a note is written to an audit table which contains the old identifier and the new one.

When the GCIS receives a request for a resource that does not exist, it first checks the above audit log. If an entry exists for the requested identifier, this entry is used to construct a URL to which a redirect is then returned.

This mechanism allows for changes to identifiers in the GCIS while continuing to provide consistent endpoints in the API.

5.4.3 Preserving Mappings

Changes to the parent table described above also trigger changes to the lexicon tables. So the complete flow for an identifier change is:

```
API or web form
-> primary key of base table
-> primary key of generic table (cascading update)
-> gcid entry in list of terms (trigger)
-> turtle template (which uses database queries)
-> Triple store (by ingesting the rendered template)
-> SPARQL endpoint
```

5.4.4 Validation of Existing Information

Identifying changes to terms is important in order to effect changes like the ones above. This requires continuous validation. In order to perform this validation, scripts must periodically check to see if the terms are still valid. If there is a mapping from terms to URLs, this can be accomplished through a simple HEAD request. If not, more advanced techniques may be necessary.

6. CONCLUSIONS AND FUTURE WORK

Lexicons are a practical way of mapping identifiers within the earth science community to each other, when uniform identifiers for resources do not exist. The GCIS provides resolvable URIs which serve to fill the gap between organizations with varying terms for the same resource. Providing both a semantic and relational mechanism for storage, and both a semantic and RESTful API allows the GCIS to have the advantages of both architectures, namely referential integrity and backwards compatibility, as well as flexibility and adaptability. While we have seen some success in using cross comparisons of data to improve quality of individual data sources, more work can be done in this area. There is also work to be done in scaling up the number and type of data sources. Another future improvement is to provide useful user interfaces for scalable human disambiguation.

7. ACKNOWLEDGMENTS

Thanks to the various people and organizations who have contributed to the GCIS and the development of its information model, including Andrew Buddenberg and the Technical Support Unit at NOAA's National Climatic Data Center, Xiaogang Ma and the group at the RPI's Tetherless World Constellation, and the University Corporation for Atmospheric Research.

8. REFERENCES

[1] U.S. Public Law 101-606(11/16/90) 104 Stat. 3096-3104, 1990 Global Change Research Act of 1990

[2] Z. Akar, T. G. Halaç, O. Dikenelli, and E. E. Ekinci. Querying the web of interlinked datasets using VOID descriptions. 2012.

[3] D. Booth. URIs and the myth of resource identity. In *Identity, Reference, and the Web Workshop at the WWW Conference*, 2006.

[4] A. Dimou, M. V, E. S, P. Colpaert, R. Verborgh, E. Mannens, and R. V. D. Walle. RDF mapping language (rml) a generic language for integrated RDF mappings of heterogeneous data. In *Linked Data on the Web, WWW 2014, Seoul, South Korea, 2014*, 2014.

[5] H. Halpin. Identity, reference, and meaning on the web. In *Proceedings of the Workshop on Identity, Meaning and the Web (IMW06)*, 2006.

[6] H. Halpin and P. Hayes. When owl:sameas isn't the same: an analysis of identity links on the semantic web. In *Linked Data on the Web, WWW 2010*, 2010.

[7] B. Hyland, G. Atemezing, M. Pendleton, and B. Srivatstava. Linked data glossary. Working group note, W3C, June 2013. `http://www.w3.org/TR/2013/NOTE-ld-glossary-20130627/`.

[8] A. Jaffri, H. Glaser, and I. C. Millard. Managing URI synonymity to enable consistent reference on the semantic web. In *Workshop on Identity, Reference, and the Web (IRSW) at ESWC2008*, 2008.

[9] H. Kramer. SAC-D (Satélite de Aplicaciones Científicas-D)/Aquarius Mission. `https://directory.eoportal.org/web/eoportal/satellite-missions/s/sac-d`, 2015. [Online; accessed 12-March-2015].

[10] R. Lewis. Dereferencing HTTP URIs. Draft tag finding, W3C, May 2007. `http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14`.

[11] F. Maali, R. Cyganiak, and V. Peristeras. Re-using cool URIs: Entity reconciliation against LOD hubs. In *Linked Data on the Web, WWW 2011, Seoul, South Korea*, 2011.

[12] D. B. Nguyen, J. Hoffart, M. Theobald, and G. Weikum. Aida-light: High-throughput named-entity disambiguation. In *Linked Data on the Web, WWW 2014, Seoul, South Korea*, 2014.

[13] A. Phillips. Character model for the world wide web: String matching and searching. W3C working draft, W3C, July 2014. `http://www.w3.org/TR/charmod-norm/`.