

Normalizing Resource Identifiers using Lexicons in the Global Change Information System

Linking Earth Science Identifiers, Concepts, and Communities

Brian Duggan¹³, Curt Tilmes², Steven Aulenbach¹³,
Robert E. Wolfe¹², Justin C. Goldstein¹³, Gerald Manion²

¹US Global Change Research Program

²National Aeronautics and Space Administration

³University Corporation for Atmospheric Research

<http://data.globalchange.gov>



Outline

Introduction

- Global Change Information System (GCIS)

- Resource Identifiers

- Lexicons

Examples

- Traceability

- Identification

Concepts

- Terms, Contexts, Lexicons

Implementation

- Interface

- Architecture

- Identifier Changes

Conclusion

- Lessons Learned

- Challenges and Future Work

Introduction

Global Change Information System (GCIS)

Resource Identifiers

Lexicons

Examples

Traceability

Identification

Concepts

Terms, Contexts, Lexicons

Implementation

Interface

Architecture

Identifier Changes

Conclusion

Lessons Learned

Challenges and Future Work

Global Change Information System (GCIS)

The U.S. Global Change Research Program (USGCRP)

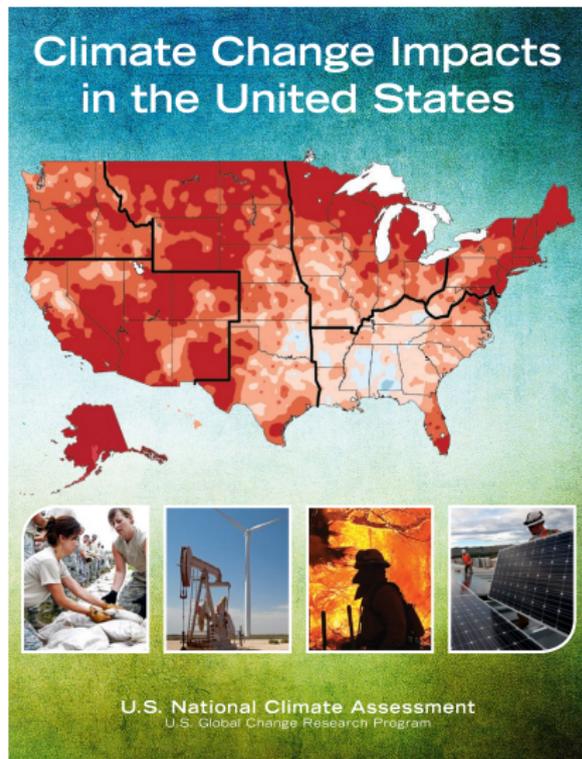
- ▶ U.S. Congress, 1990 : Global Change Research Act, establishes USGCRP
- ▶ to “assist the Nation and the world to understand, assess, predict, and respond to human-induced and natural processes of global change.” – Global Change Research Act
- ▶ confederation of 13 federal agencies in the U.S. Government
- ▶ overseen by White House Office of Science and Technology Policy
- ▶ Global Change Information System (GCIS) established 2012
- ▶ 2014 : released the third National Climate Assessment (NCA3)

Global Change Information System (GCIS)

The Third National Climate Assessment (NCA3).

“Highly influential scientific assessment“

- ▶ 829 pages
- ▶ 30 chapters
- ▶ 300+ authors
- ▶ 161 findings
- ▶ 284 figures
- ▶ 3,395 references
 - ▶ journal articles
 - ▶ books
 - ▶ reports
- ▶ datasets
- ▶ models
- ▶ platforms
- ▶ instruments



Global Change Information System (GCIS)

GCIS: an open-source web based resource for traceable, sound, global change data, information and products.

- ▶ Provides common identifiers across diverse systems.
- ▶ Supports report production.
- ▶ Backend API for dynamic NCA3 front end:
<http://nca2014.globalchange.gov>.
- ▶ Content negotiation for all URLs.
- ▶ HTML representations form follow-your-nose site.
- ▶ SPARQL endpoint:
<http://data.globalchange.gov/sparql>
- ▶ Semantic and relational data model.
- ▶ Identifies and disambiguates global change information.

Resource Identifiers

Terms

- ▶ RCP 8.5
- ▶ sresa2, SRES A2
- ▶ Terra, EOS AM-1, 80eca755-c564-4616-b910-a4c4387b7c54
- ▶ MODIS, 119
- ▶ NASA, 026:00
- ▶ 1.2, 8.3 (findings, figures)
- ▶ PODAAC-TPTMR-REP01
- ▶ Also: DOIs, ISSNs, ISBNs, ORCIDs, sometimes URIs

GCIS URIs (GCIDs)

`http://data.globalchange.gov`

- ▶ `/article/10.1080/15287390801997625`
- ▶ `/report/usfs-pnw-gtr-855`
- ▶ `/report/nca3/figure/global-temperature-and-co2`
- ▶ `/report/nca3/table/decisions-scales`
- ▶ `/report/nca3/finding/extreme-precipitation-increase`
- ▶ `/organization/nasa`
- ▶ `/person/0000-0001-6667-7047`
- ▶ `/dataset/nca3-cddv2-r1`
- ▶ `/platform/terra`
- ▶ `/instrument/modis`

Lexicons

Communities of practice use context-dependent terms as identifiers.

- ▶ Report collaborators
Authors, Science analysts, Editors, Graphic designers, Web developers, Project managers
- ▶ Data Managers
- ▶ Data Producers
- ▶ Modelers
- ▶ Scientists
- ▶ Policy Makers
- ▶ Committees, Federations
- ▶ Publishers
- ▶ Libraries

Introduction

Global Change Information System (GCIS)

Resource Identifiers

Lexicons

Examples

Traceability

Identification

Concepts

Terms, Contexts, Lexicons

Implementation

Interface

Architecture

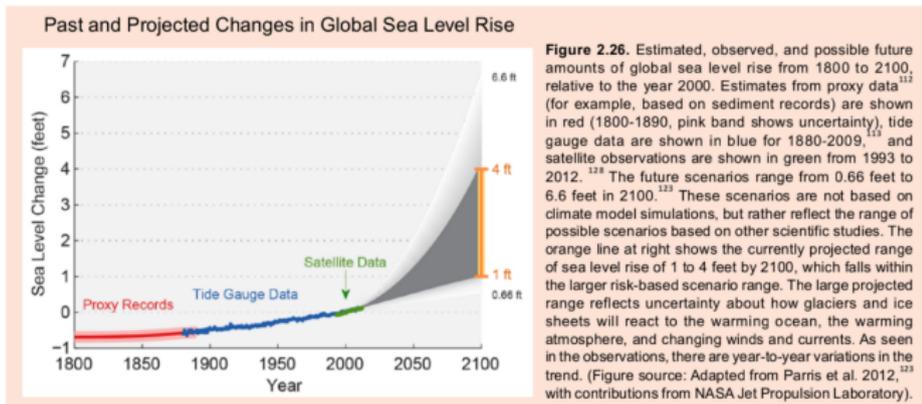
Identifier Changes

Conclusion

Lessons Learned

Challenges and Future Work

Third National Climate Assessment, Figure 2.26



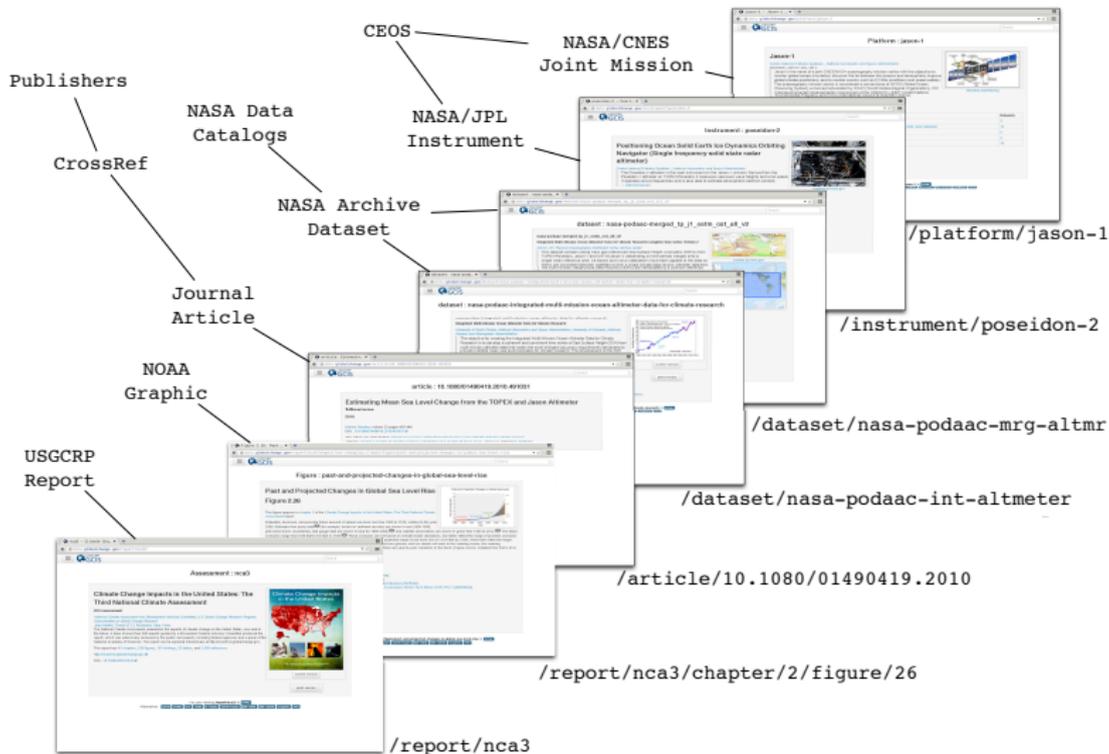
U.S. GLOBAL CHANGE RESEARCH PROGRAM

45

CLIMATE CHANGE IMPACTS IN THE UNITED STATES

<http://data.globalchange.gov/report/nca3/chapter/2/figure/26>

Traceability



Identification

<http://data.globalchange.gov/platform/jason-1>

Source NASA JPL Physical Oceanography Distributed Active Archive Center (PODAAC)

Mission Committee on Earth Observation Satellites (CEOS)

Platform Label NASA Global Change Master Directory (GCMD)

Platform Name NASA Earth Observing System Clearing House (ECHO)

PODAAC

```
-> GET http://podaac.jpl.nasa.gov/ws/search/dataset/?datasetId=PODAAC-USWCO-ALTO1  
<- ... <podaac:sourceShortName>JASON-1</podaac:sourceShortName>...
```

CEOS

```
-> GET http://database.eohandbook.com/database/missiontable.aspx  
<- ... Jason-1...  
<- ... 286...
```

GCMD

```
-> GET http://gcmdservices.gsfc.nasa.gov/static/kms/platforms/platforms.rdf  
<- <skos:Concept rdf:about="4ea59dad-ed94-453e-a991-62c790a1d101"  
<- ... <skos:prefLabel xml:lang="en">JASON-1</skos:prefLabel>
```

ECHO

```
-> GET https://api.echo.nasa.gov/catalog-rest/echo_catalog/datasets.echo10  
<- <Platform><ShortName>JASON-1</ShortName><LongName>Jason-1</LongName>...
```

Also, OAI-PMH, FGDC, DIF, ISO 19115, ECHO 10, CSV, JSON, ...

Introduction

Global Change Information System (GCIS)

Resource Identifiers

Lexicons

Examples

Traceability

Identification

Concepts

Terms, Contexts, Lexicons

Implementation

Interface

Architecture

Identifier Changes

Conclusion

Lessons Learned

Challenges and Future Work

Terms, Contexts, Lexicons

Term A sequence of characters from the Universal Character Set (UCS) which is used as an identifier for a resource by a group of people.

Context A set of terms used to identify resources of the same type.

Lexicon A set of contexts used by a community.

Lexicons map terms to GCIDs.

Terms are SKOS “lexical labels” used as identifiers.

Terms, Contexts, Lexicons

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>

Introduction

Global Change Information System (GCIS)

Resource Identifiers

Lexicons

Examples

Traceability

Identification

Concepts

Terms, Contexts, Lexicons

Implementation

Interface

Architecture

Identifier Changes

Conclusion

Lessons Learned

Challenges and Future Work

Interface

Creating terms

```
POST /lexicon/ceos
{ "context" : "MissionId",
  "term" : "286",
  "gcid" : "/platform/jason-1" }
```

```
# Alternative
PUT /lexicon/ceos/MissionId/286
{ "gcid" : "/platform/jason-1" }
```

```
# Lexicon lookup
GET /lexicon/ceos/MissionId/286
303 See Other
Location: /platform/jason-1
```

Interface

Creating, updating resources and URIs

POST /platform

```
{ "identifier" : "jason-1", ... }
```

POST /platform/jason-1

```
{ "identifier" : "jason-1-renamed", ... }
```

GET /platform/jason-1

303 See Other

Location: /platform/jason-1-renamed

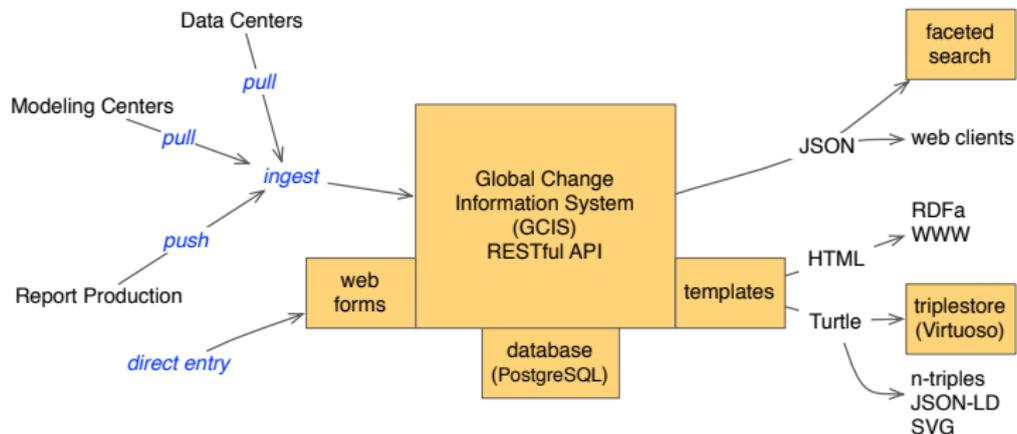
GET /lexicon/ceos/MissionId/286

303 See Other

Location: /platform/jason-1-renamed

Architecture

Information Flow



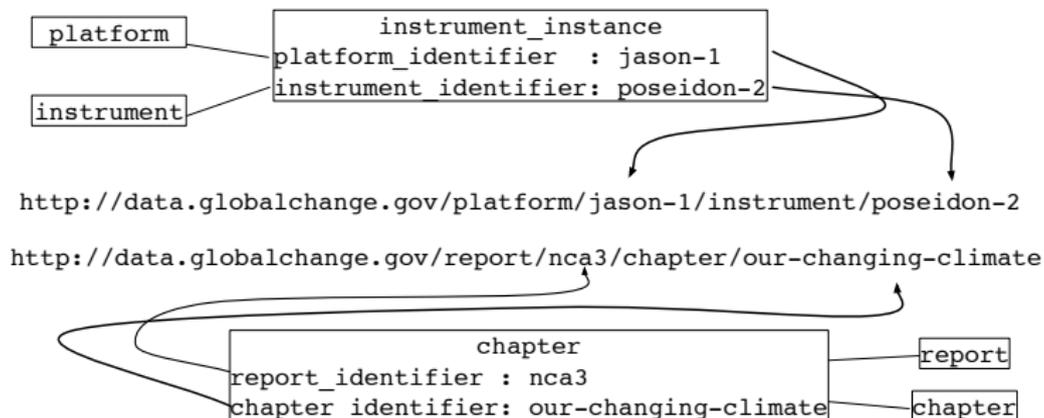
Identifier Changes

Relational representation

- ▶ PostgreSQL audit tables
- ▶ Check audit tables before a 404 response, maybe redirect.
- ▶ Foreign keys used when possible.
- ▶ Self-joinable common parent table with extra fields.
- ▶ Mapping table supports entity-activity-agent (PROV).
- ▶ Cascading updates and triggers.

Identifier Changes

Natural primary keys form unique URIs



Composite primary keys as foreign
keys with cascading updates

Identifier Changes

Change propagation

- ▶ API or web form
- database tables (cascades, triggers, audit)
- lexicon tables (triggers, audit)
- turtle template (uses database)
- Triple store (scrape)
- SPARQL endpoint

Identifier Changes

Turtle templates

platform.ttl.tut

```
<<%= current_resource %>>
dcterms:identifier "<%= $platform->identifier %>";
dcterms:title "<%= $platform->name %>""^^xsd:string;
dbpprop:launchDate "<%= $platform->start_date%>""^^xsd:dateTime;
dbpprop:deactivated "<%= $platform->end_date %>""^^xsd:dateTime;
% for my $instrument ($platform->instruments) {
  gcis:hasInstrument <<%= uri($instrument) %>>;
% }
a gcis:Platform .
%= include 'other_identifiers'
```

other_identifiers.ttl.tut

```
<<%= current_resource %>>
...
% for my $term (terms(current_resource)) {
  skos:altLabel "<%= $term %>";
  % if ($term->same_as) {
    owl:sameAs <<%= $term->same_as %>>;
  % }
  ...
% }
```

Identifier Changes

Turtle templates

```
<http://data.globalchange.gov/platform/jason-1>
  dcterms:identifier "jason-1";
  dcterms:title "Jason-1"^^xsd:string;
  dbpprop:launchDate "2001-12-09T00:00:00"^^xsd:dateTime;
  dbpprop:deactivated "2013-07-03T00:00:00"^^xsd:dateTime;
  gcis:hasInstrument <http://data.globalchange.gov/instrument/poseidon-2>
  gcis:hasInstrument <http://data.globalchange.gov/instrument/laser-retroreflector-array>;
  gcis:hasInstrument <http://data.globalchange.gov/instrument/doris-ng>;
  gcis:hasInstrument <http://data.globalchange.gov/instrument/jason-microwave-radiometer>;
  gcis:hasInstrument <http://data.globalchange.gov/instrument/blackjack>;
  a gcis:Platform .

<http://data.globalchange.gov/platform/jason-1>
  skos:altLabel "286";
  gcis:hasURL "http://database.eohandbook.com/database/missionsummary.aspx?missionID=286";

  skos:altLabel "Jason-1";
  gcis:hasURL "http://database.eohandbook.com/database/missionindex.aspx#J";

  skos:altLabel "Jason-1";
  gcis:hasURL "http://wikipedia.org/wiki/Jason-1";
  owl:sameAs <http://dbpedia.org/resource/Jason-1>;

  skos:altLabel "JASON-1";
  gcis:hasURL "http://podaac.jpl.nasa.gov/datasetlist?ids=Platform&values=JASON-1" .
```

Introduction

Global Change Information System (GCIS)

Resource Identifiers

Lexicons

Examples

Traceability

Identification

Concepts

Terms, Contexts, Lexicons

Implementation

Interface

Architecture

Identifier Changes

Conclusion

Lessons Learned

Challenges and Future Work

Lessons Learned

- ▶ Opaque identifiers incur technical debt.
- ▶ Ad hoc terms are often identifiers.
- ▶ Identifiers change without notice.
- ▶ Few APIs provide changesets.
- ▶ Federated queries need lexicons.

Challenges and Future Work

- ▶ Identification of aggregates (systems, series).
- ▶ Interfaces to scale up human disambiguation.
- ▶ Lexicons in the “long tail” of science.
- ▶ Optimizing audit tables for identifiers.

Thanks

Thanks: NOAA National Climatic Data Center, Tetherless World Constellation, Andrew Buddenberg, Hook Hua, Brian Wilson, Brent Newman, Xiaogang Ma

Brian Duggan

`bduggan@usgcrp.gov`

`http://github.com/usgcrp/gcis`

`http://data.globalchange.gov`