



EarthCube Council of Data Facilities Registry Working Group



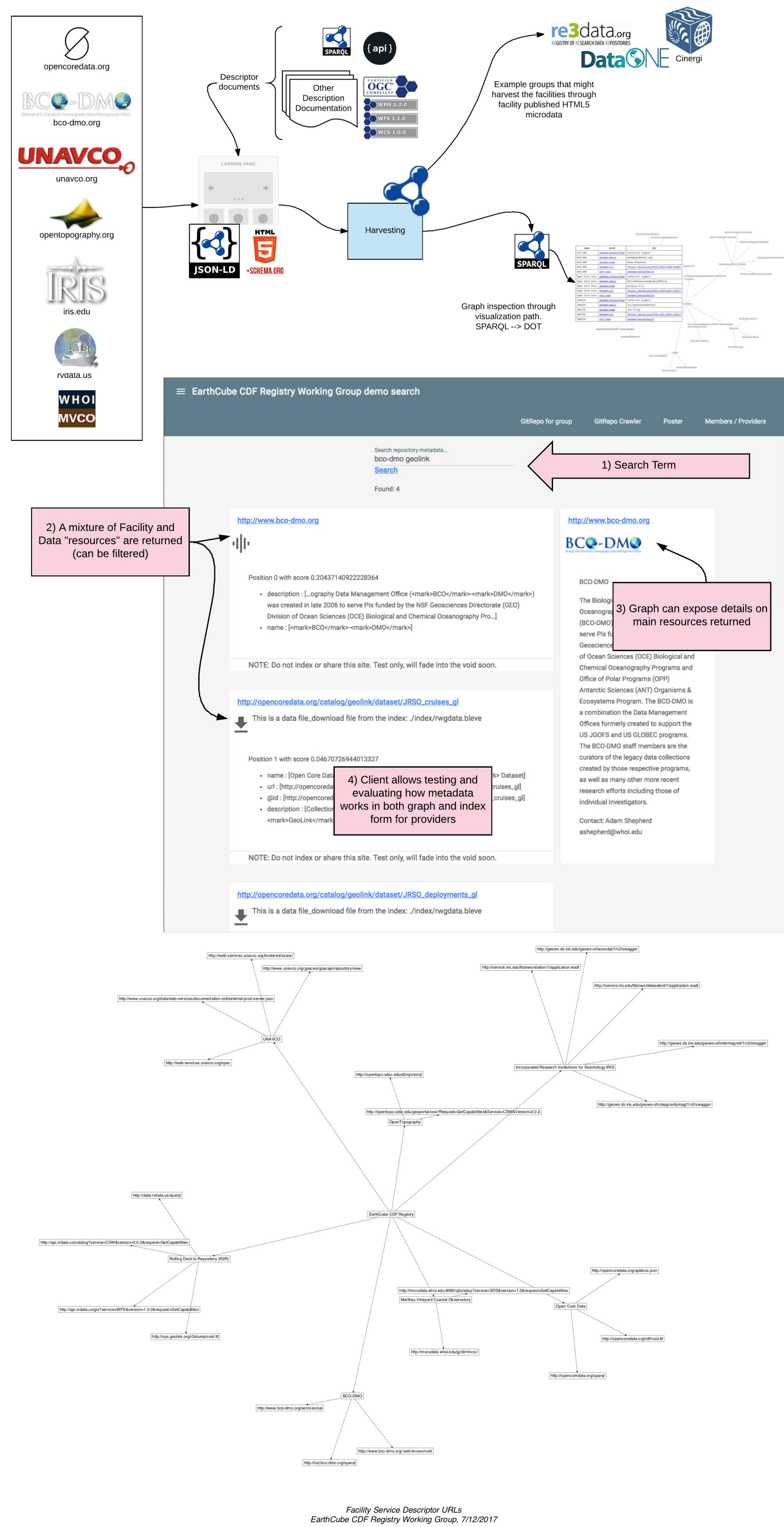
A report on goals and status

About

The EarthCube Council of Data Facilities (CDF) is a federation of existing and emerging geoscience data facilities that serves as a foundation for EarthCube and cyberinfrastructure for the geosciences. The CDF Facilities Registry Working Group formed to review alignment of existing approaches to research facility description and discovery. The involved parties include the EarthCube CDF, Coalition for Publishing Data in the Earth and Space Sciences (COPDESS) and the Registry of Research Data Repositories (re3data).

Publish & Harvest Test

A set of repositories are participating in a pilot/prototype publish & harvest demonstration.



Support of FAIR Data Principles

- 1) **Findable** - machine readable repository information can be harvested and supports aggregation sites like re3data
- 2) **Accessible** - repository metadata and data sets are open and accessible
- 3) **Interoperable** - standard representation of repository metadata adopted across the community supports interoperability
- 4) **Re-usable** - Leveraging re3data.org and schema.org allows for repository metadata schema to be used by others

Tim Ahern
Director of Data Services
IRIS DMC
tim@iris.washington.edu

Adam Shepherd
Software Developer
BCO-DMO
ashepherd@whoi.edu

Robert Arko
Tech. Director
UNOLS/R2R
arko@ideo.columbia.edu

Shelley Stall
Assistant Director
American Geophysical Union & COPDESS
sstall@agu.org

Douglas Fils
Data Manager
Ocean Leadership
dfils@oceanleadership.org

Mike Stults
Software Engineer
IRIS DMC
mike@iris.washington.edu

Danie Kinkade
Data Manager
BCO-DMO
dkinkade@whoi.edu

Michael Witt
Associate Professor
Purdue University & re3data
mwitt@purdue.edu

Community Input

A copy of this poster and all code is at:
<https://github.com/fils/CDFRegistryWG>

- Earthcube.org/group/council-data-facilities
- re3data.org
- COPDESS.org
- ESIPfed.org



The Council of Data Facilities Charter states that “[Geoscience data facilities] **provide direct societal benefits** through increased coordination, collaboration, and innovation in the acquisition, curation, preservation, and dissemination of geoscience data, tools, models, software, and services.” Additionally, **data management and preservation of scientific data is best conducted by domain repositories** that provide these curation and preservation services for datasets specific to targeted domains.

Today, NSF-funded **domain repositories have no common way to share** with the larger Earth and space science (ESS) community information about each repository and their data holdings. The CDF Registry Working Group is **developing guidelines for what information would be valuable to share and a machine-readable method to publish that information.**

The CDF Registry Working Group is working with the Coalition of Publishing Data in the Earth and Space Sciences (**COPDESS.org**) and the Registry of Research Data Repositories (**re3data.org**) to develop repository metadata that will be valuable to CDF’s community and build on the metadata schema already adopted by the international scientific repository community represented by re3data.

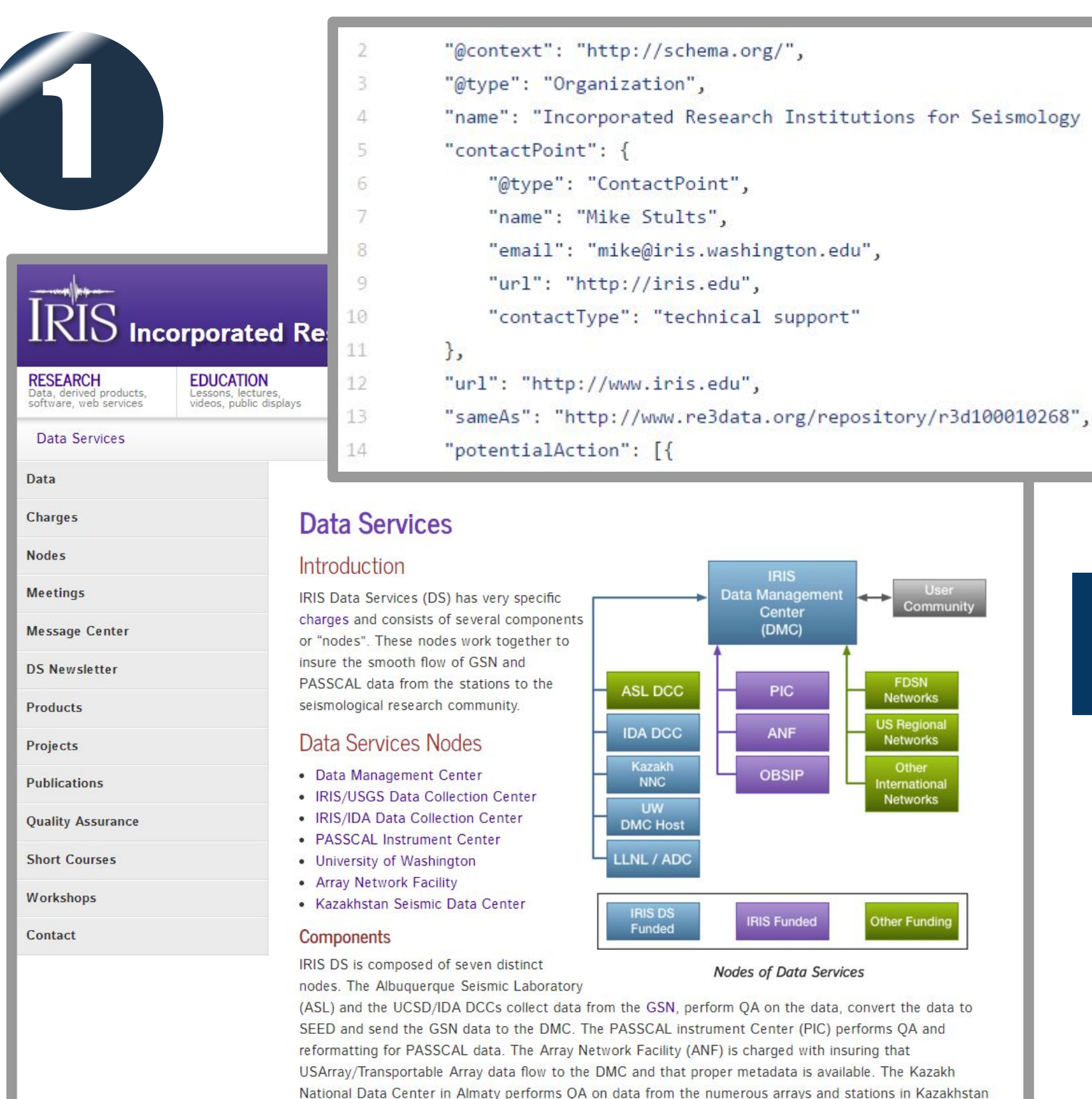
Eight CDF members are supporting the pilot portion to test out the draft guidelines for publishing repository metadata. By using schema.org for the base representation of the machine readable information (JSON-LD), the solution includes **built in governance that is sustainable** into the future.

Objectives:

1. Formalize a set of **repository parameters-of-interest** to CDF members.
2. Review the **alignment** of those parameters with re3data and COPDESS.
3. Develop strategies for CDF members to **express/expose** this information.
4. Develop a means to **encode this schema in a machine readable** format.
5. Demonstrate the use of schema.org for **publishing and accessing** this metadata.
6. **Leverage re3data** as a reference implementation for collecting and exposing this metadata.

Benefits:

1. Repositories have **control over their metadata** and can update at any time.
2. These new standard guidelines for **publishing repository metadata can be adopted across all ESS repositories** and other scientific domains to support repository discovery and access.
3. CDF will **recommend these standards** to their membership and **work towards adoption** by all NSF-funded ESS repositories as complying with the standard.



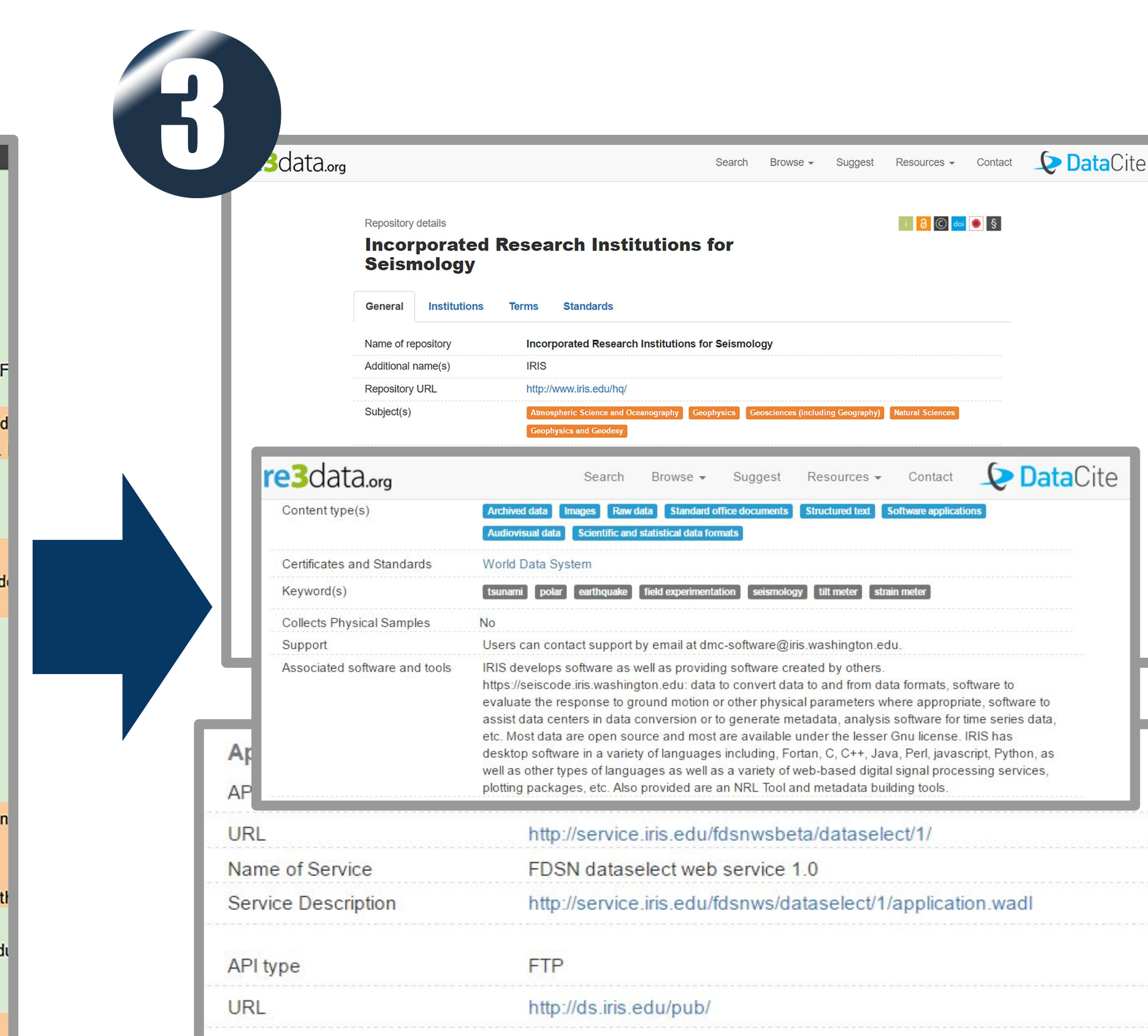
Facility exposes metadata about itself using JSON-LD embedded in its landing web page

Next Steps:

- Complete pilot effort to test guidelines.
- Coordinate with schema.org as a potential external vocabulary for repository metadata for on-going management and governance.
- Provide guidelines to CDF members and other ESS NSF repositories to embed metadata information as JSON-LD (schema.org).
- Monitor and encourage adoption by CDF members and ESS NSF repositories

1	2	3
Facility name	re3data mapping	example value
1. Facility name	repositoryName	IRIS Data Management Center
2. URL	repositoryURL	http://ids.iris.edu
3. Identifier	repositoryIdentifierType	DOI
4. Mission statement	repositoryIdentifierValue	http://doi.org/10.17616/R3X697
5. Sponsoring agency	missionStatementURL	http://www.iris.edu/about_us/mission-statement
6. Domain	institutionName	National Science Foundation
7. Keyword	responsibilityType	funding
8. Data collected	subject	Seismology
9. Data formats	keyword	"Earthquakes", "Geophysics", "Tsunamis"
10. Services	asType	Other
11. Registered By	apiURL	http://service.iris.edu/idsnewsbeta/datasetselect/1/
12. Restrictions on data access	apiDocumentation	http://service.iris.edu/idsnewsbeta/datasetselect/1/
13. Data collection and archival	dataAccessRestriction	registration
14. Associated software and tools	dataLicenseName	CC
15. Collects physical samples	dataLicenseURL	https://creativecommons.org/licenses/by/2.0/
16. Support	policyName	Data Acceptance Policy for Seismological Data
17. Community engagement/control	policyURL	http://www.iris.edu/files/programs/data_service
18. Quality assurance/control	qualityManagement	Quality Guidelines for Data Submission
19. Cloud resources utilized	qualityPolicy	http://www.iris.edu/files/programs/data_service

Metadata to be harvested by re3data: the green elements above map directly to the re3data schema, orange elements will be extended using schema.org



Mock-up of extended facility information represented in the re3data web interface after it has been harvested