

NSF EarthCube's GeoCODES

Geoscience
Cyberinfrastructure for
Open Discovery in the
Earth Sciences

Mike Bobak

I ILLINOIS

NCSA | National Center for
Supercomputing Applications



EarthCube

Transforming Geosciences Research

NSF EarthCube 2012-22

- Cyberinfrastructure to integrate Geoscience community efforts
 - Integrate data -> info -> knowledge management efforts
 - Sharable tools, methods, .. via community involvement
- Funded projects include a wide variety of products
 - Tools, data, services, apps and other community resources for geoscience researchers
 - Such as: Argovis 2.0, Pangeo, Planet Microbe, Sparrow, Macrostrat, StraboSpot, ICEBERG, ASSET, QGreenland, Heliportal, AMGeO, ML4POSE, InGeO, X-DOMES, CHORDS, Seaview, Ocean Protein Portal, Advancing netCDF-CF, Linked Earth, Paleobiology Database, eODP, GeoDeepDive, Data Discovery Studio, Digital Rocks Portal, BALTO, Throughput
- The EarthCube Office (earthcube.org/ECO) is a resource to all of these
- ECO manages the earthcube.org/geocodes project/resources

What is GeoCODES?

GeoCODES is:

- An NSF Earthcube program effort to better enable cross-domain discovery of and access to geoscience data & research tools.
- Geoscience Cyberinfrastructure for Open Discovery in the Earth Sciences

GeoCODES is made up of three components:

- Metadata standards
 - An evolving standard for exposing data called science on schema.org
- A set of tools
 - to index relevant data from partners within the Council of Data Facilities (CDF) who have adopted *science on schema*
 - plus a prototype portal to query that data & now resources from:
- A Resource Registry
 - Used to register and discover relevant tools
 - Now searchable and linkable via metadata

(FAIR) Metadata standards

go-fair.org

FAIR Principles Implementation Networks News Events Resources About GO FAIR

Data Together: Joint commitment by GO FAIR, CODATA, RDA & WDS

The four major international data organisations – GO FAIR, CODATA, RDA & WDS – commit to working together to optimise the global research data ecosystem and to identify the opportunities and needs that will trigger federated infrastructures to service the new reality of data-driven science. [\[more\]](#)

DATA TOGETHER

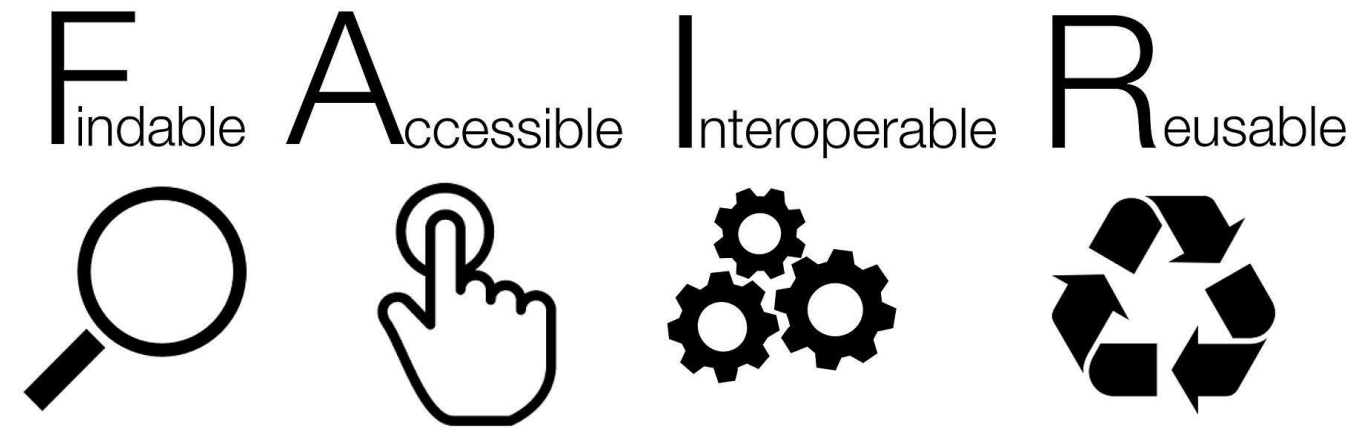
What is in it for you?

GO FAIR Today

Implementation Networks

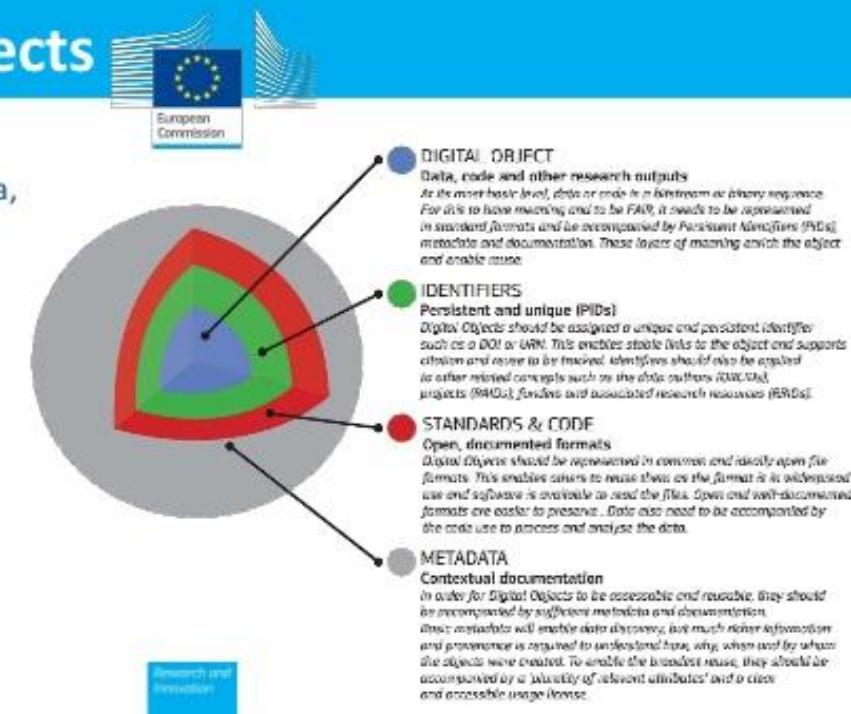
GO CHANGE **GO TRAIN** **GO BUILD**

FAIR Principles



FAIR Digital Objects

- Digital objects can include data, software, and other research resources
- Universal use of PIDs
- Use of common formats
- Data accompanied by code
- Rich metadata
- Clear licensing



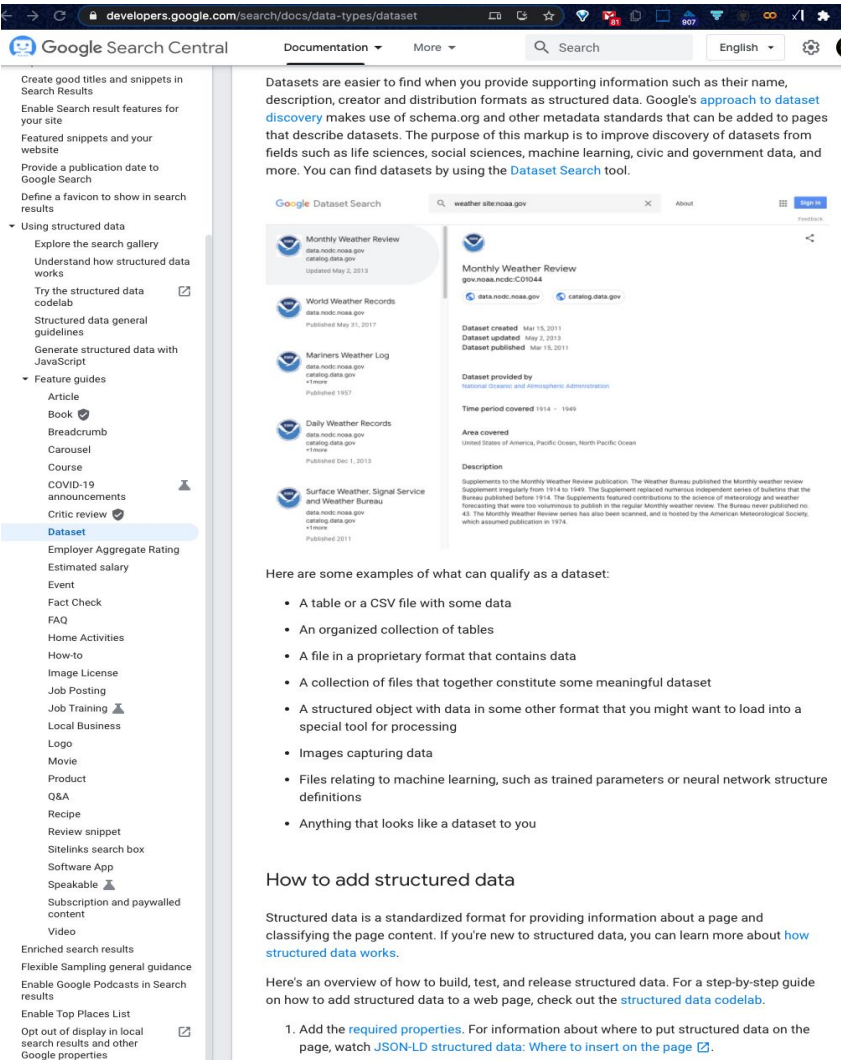
more at: developers.google.com/search/docs/data-types/dataset

Evolving standard science on schema.org

- Used by all search engines, rich snippets
- Metadata harvest part of standard crawl
- Available to dataset-search if [schema:Dataset](https://schema.org/Dataset)
- Try it datasetsearch.research.google.com

Tools to index and a portal to query that data

- We do a focused Geoscience crawl of CDF repos
- We have search sites with extended functionality
- For finding data, & now Resource Registry software..



The screenshot displays the Google Dataset Search results page for the query 'weather.noaa.gov'. The interface includes a sidebar on the left with navigation options like 'Using structured data', 'Feature guides', and 'Dataset'. The main content area shows a list of search results, with the top result being 'Monthly Weather Review' from 'data.noaa.gov'. This result is expanded to show detailed metadata, including the dataset's creation and update dates, the provider (National Oceanic and Atmospheric Administration), the time period covered (1914-1949), and the area covered (United States of America, Pacific Ocean, North Pacific Ocean). A description of the dataset is also provided. Below the search results, there is a section titled 'Here are some examples of what can qualify as a dataset:' which lists various types of data collections. Finally, a section titled 'How to add structured data' provides instructions on how to add structured data to a web page, including a link to a 'structured data codex' and a numbered list of steps.

Google Dataset Search

weather.noaa.gov

Monthly Weather Review
data.noaa.gov
Updated May 2, 2013

World Weather Records
data.noaa.gov
Published May 21, 2017

Mariners Weather Log
data.noaa.gov
Published 1957

Daily Weather Records
data.noaa.gov
Published Dec 1, 2013

Surface Weather, Signal Service and Weather Bureau
data.noaa.gov
Published 2011

Monthly Weather Review
gov.noaa.noaa.gov
Catalog data.gov

Dataset created: Mar 15, 2011
Dataset updated: May 2, 2013
Dataset published: Mar 15, 2011

Dataset provided by:
National Oceanic and Atmospheric Administration

Time period covered: 1914 - 1949

Area covered:
United States of America, Pacific Ocean, North Pacific Ocean

Description:
Supplements to the Monthly Weather Review publication. The Weather Bureau published the Monthly weather review Supplement irregularly from 1914 to 1949. The Supplement contained numerous independent series of tables that the Bureau published before 1914. The Supplements featured contributions to the science of meteorology and weather forecasting that were not considered to publish in the regular Monthly weather review. The Bureau never published no. 43. The Monthly Weather Review series has also been scanned, and is hosted by the American Meteorological Society, which assumed publication in 1914.

Here are some examples of what can qualify as a dataset:

- A table or a CSV file with some data
- An organized collection of tables
- A file in a proprietary format that contains data
- A collection of files that together constitute some meaningful dataset
- A structured object with data in some other format that you might want to load into a special tool for processing
- Images capturing data
- Files relating to machine learning, such as trained parameters or neural network structure definitions
- Anything that looks like a dataset to you

How to add structured data

Structured data is a standardized format for providing information about a page and classifying the page content. If you're new to structured data, you can learn more about [how structured data works](#).

Here's an overview of how to build, test, and release structured data. For a step-by-step guide on how to add structured data to a web page, check out the [structured data codex](#).

1. Add the [required properties](#). For information about where to put structured data on the page, watch [JSON-LD structured data: Where to insert on the page](#).

Evolving standard science on schema.org

- Start with `schema.org:Dataset`
- W3C's Data Catalog Vocabulary (DCAT)
- Some structured tables via W3C CSVW
- Science extensions to schema.org
 - github.com/ESIPFed/science-on-schema.org
- Can start with minimal `json-ld` `<script>`
 - en.wikipedia.org/wiki/Linked_data in `json`

Tools to index and a portal to query that data

- We do a focused Geoscience crawl of CDF repos
- We make further use of more metadata
- Eg. linking data to tools.. from Resource Registry

[/science-on-schema.org/examples/dataset/minimal.jsonld](https://science-on-schema.org/examples/dataset/minimal.jsonld)
`<script type="application/ld+json">`

```
{
  "@context": {
    "@vocab": "https://schema.org/"
  },
  "@type": "Dataset",
  "name": "Removal of organic carbon by natural bacterioplankton communities
as a function of pCO2 from laboratory experiments between 2012 and 2016",
  "description": "This dataset includes results of laboratory experiments which measured
dissolved organic carbon (DOC) usage by natural bacteria in seawater at different pCO2 levels. ....",
  "url": "https://www.example-data-repository.org/dataset/472032",
  "version": "2013-11-21",
  "keywords": ["ocean acidification", "OA", "Dissolved
Organic Carbon", "DOC", "bacterioplankton respiration",
"pCO2", "carbon dioxide", "elevated pCO2", "oceans"],
  "license": "CC-BY-4.0"
} </script>
```

Access to GeoCODES data/tool search

- Present public site:
 - geocodes.earthcube.org
- POC for using clowder
 - earthcube.clowderframework.org
 - mbobak-ofc.ncsa.illinois.edu/search.htm
- Newer demo with faceted-search:
 - alpha.earthcube.org
 - mbobak-ofc.ncsa.illinois.edu/landing.html

geocodes.earthcube.org /about.html on the crawl and serving



Organization

The sources for geodex comes mostly from collaboration with the EarthCube Council of Data Facilities (CDF).



Providers

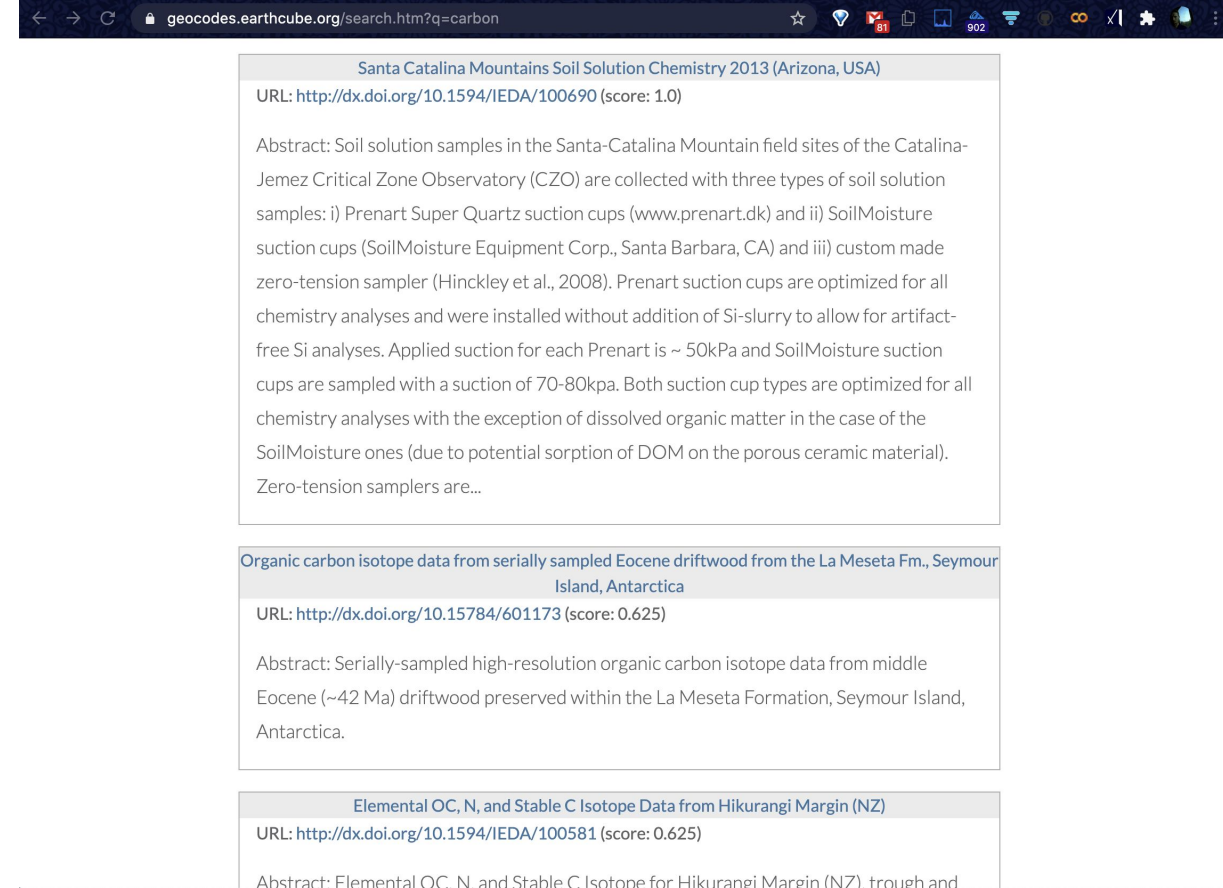
CDF members who express their resources via structured data on the web approaches can be indexed.



Indexing

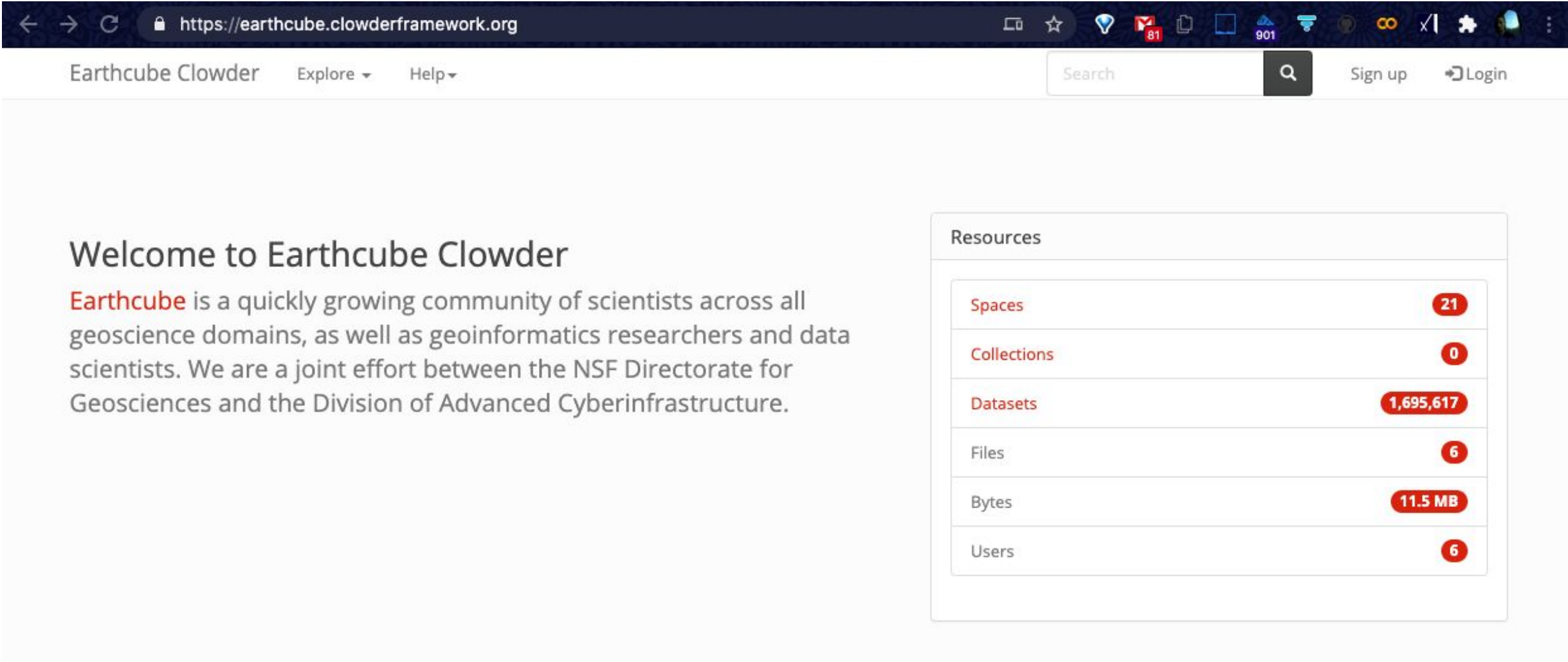
Geodex uses the gleaner program (gleaner.io) to build the index and (GROW) as a server. See the about section for more.

Results of search: Title, Abstract, DOI URI to data page at repo



Other versions include a links to full metadata incl. in (a reskinning of clowder to this)

ClowderFramework.org data-facility/repository per *space*, plus one for Resources/tools..





High-Resolution Topography Data and Tools

0 0 1



Neotoma Paleoecology Database and Community is an online hub for data, research, education, and discussion about paleoenvironments. Anyone with an Internet connection can access Neotoma.

11955 0 1



OpenSky is the home for
NCAR/UCAR research and historical
materials as well as other collections



Magnetics Information Consortium
(MagIC)
Promoting Information technology
Infrastructures for the International
paleomagnetic, geomagnetic and
rock magnetic community.

4136 0 1

opencoredata

Open Core Data is an implementation of the RDA Digital Object Cloud. Open Core Data contains digital objects from the continental and ocena drilling research projects funded by the National Science Foundation. These objects are described using the structured data on the web patterns pro...

18171 0 1



unavco

Transforming understanding of Earth systems and hazards using geodesy.

5086 0 1



ssdb.iodp

The Site Survey Data Bank (SSDB) is a repository for site survey data submitted in support of International Ocean Discovery Program (IODP) proposals and expeditions. SSDB serves different roles for different sets of users

5344 0 1



hydroshare

HydroShare is CUAHSI's online collaboration environment for sharing data, models, and code.

4185 0 1

Clowder organization

- One *space* per data-facility
- *Datasets* hold metadata
- Also a Resources space:

Allows for

- dataset & tool search
- metadata/annotation
- linking out to get the data
- & sometimes (assoc) tool/s

The screenshot shows the Earthcube Clowder web interface. The browser address bar displays the URL: `earthcube.clowderframework.org/spaces/5f87c52ee4b0a4d76fb2c3ce`. The page header includes the "Earthcube Clowder" logo, navigation links for "Explore" and "Help", a search bar, and "Sign up" and "Login" buttons. The main content area is titled "resource_registry" and contains a paragraph explaining the EarthCube Resource Registry (ECRR) project. Below this, there is a "Public Data" tab and a section titled "Datasets" with the subtitle "Viewing most recent datasets". A red button labeled "View All Datasets" is located to the right of the "Datasets" section. The "Datasets" section displays a grid of dataset cards. Each card includes a title, a description, and a set of icons representing various metrics (e.g., 0 datasets, 0 collections, 1 member, 4 views, 0 downloads). The visible dataset cards are: "Seismic Analysis Code (SAC) format", "Access of Oceanic Protein Datasets", "UK Linked Open Data Register", "GeoTIFF 1.0 format", "URI Template specification", and "Earth Cube Resource Registry Ontology". On the right side of the page, there are sections for "Statistics" (showing 1 member, 0 collections, and 274 datasets), "External Links" (with a link to `https://earthcube.org/resource_registry`), and "Access" (with a "PUBLIC" button).

Clowder search results

& a result's metadata(tab) tree listing

earthcube.clowderframework.org/search?query=carbon


Earthcube Clowder Explore Help Search Sign up Login

Search

carbon


Search Syntax Help
Metadata Search

Results




SensorML urn:sunburst:sensor:SAMI-CO2
Wed Nov 04 19:50:22 GMT 2020

* Measures the partial pressure of carbon dioxide pCO2 in water from 200-600 µatm (ranges above 600 are available by request) * Uses a highly precise and stable colorimetric reagent method * Provide researchers with valuable in-situ time series data * Deployable to depths up to 600 meters * Can be deployed in the ocean or in freshwater * Long-term depolyments - can run for more than a year taking hourly measurements * Can support up to 3 external instruments such as PAR, dissolved oxygen, chlorophyll fluorometer, or CTD * Can support inductive modems or external loggers if required. * Biofouling Package available for deployments in productive environments <https://xdomes.tamucc.edu/srr/sensorML/urn-sunburst-sensor-SAMI-CO2.html>




Soil chemical properties, periodic
Tue Nov 17 15:54:46 GMT 2020

Carbon and nitrogen concentrations from the top 30 cm of the profile. Data are reported by horizon (organic vs. mineral) within a soil core. <https://data.neonscience.org/data-products/DP1.10078.001>



Root chemical properties
Tue Nov 17 15:54:46 GMT 2020

Carbon and nitrogen concentrations in root biomass, either from periodic collections of surface soil (0-30 cm) or from one-time soil Megapit sampling in increments to 2 m depth. <https://data.neonscience.org/data-products/DP1.10102.001>



Sediment chemical properties
Tue Nov 17 15:54:46 GMT 2020

earthcube.clowderframework.org/datasets/5fa305fee4b097cab4a0021b

Earthcube Clowder Explore Help

Files Metadata Extractions Visualizations Comments (0)

Metadata

Extracted by <http://clowder.ncsa.illinois.edu/extractors/deprecatedapi> on Nov 4, 2020

@type: Dataset

isAccessibleForFree: true

alternateName: urn:sunburst:sensor:SAMI-CO2

description: * Measures the partial pressure of carbon dioxide pCO2 in water from 200-600 µatm (ranges above 600 are available by request) * Uses a highly precise and stable colorimetric reagent method * Provide researchers with valuable in-situ time series data * Deployable to depths up to 600 meters * Can be deployed in the ocean or in freshwater * Long-term depolyments - can run for more than a year taking hourly measurements * Can support up to 3 external instruments such as PAR, dissolved oxygen, chlorophyll fluorometer, or CTD * Can support inductive modems or external loggers if required. * Biofouling Package available for deployments in productive environments

includedInDataCatalog:

url: <https://xdomes.tamucc.edu/srr/>

@id: <https://xdomes.tamucc.edu/srr/>

keywords: oceanography,CO2

license: <https://creativecommons.org/licenses/by/4.0/>

name: SensorML urn:sunburst:sensor:SAMI-CO2

url: <https://xdomes.tamucc.edu/srr/sensorML/urn-sunburst-sensor-SAMI-CO2.html>

version: 2020-04-17 17:00:00

provider:

@type: Organization

legalName: Regional Ocean Acidification: Northwestern Gulf of Mexico

name: OAR Northwestern Gulf of Mexico

url: http://hulab.tamucc.edu/OAP/OAP_index.htm

@id: data.gcoos.org

publisher:

@type: Organization

New landing page demo, informed by a NCSA UX study



Results filtered by faceted metadata:

- Data, Tools, or both
- Science Domain
- Place (geolocation)
- Publisher
- Date (of publication)

The screenshot displays the EarthCube Faceted Search interface. The browser address bar shows the URL: `alpha.geocodes.earthcube.org/facetedsearch_live.html?q=carbon&searchTypes=all`. The search bar contains the term "carbon". The page title is "Faceted Search of EarthCube resources".

On the left side, there are four faceted metadata filters:

- Resource Type**: A dropdown menu showing "data" with a count of (279).
- Science Domain**: A dropdown menu showing "Interdisciplinary Earth Data Alliance (IEDA)" with a count of (186).
- Place**: A dropdown menu showing "U.S. Antarctic Program (USAP) Data Center" with a count of (93).
- Publisher/Repo**: A dropdown menu showing "Date" with a count of (3).

On the right side, there are two search results displayed. Each result includes a title, a description, an abstract, and keywords. The first result is titled "3D multichannel seismic field data from the San Luis Pass region off Galveston, Texas, acquired by the R/V Brooks-McCall in 2013 (BM1310)" and is from the "Interdisciplinary Earth Data Alliance (IEDA)". The second result is titled "Abrupt Change in Atmospheric CO2 During the Last Ice Age" and is from the "U.S. Antarctic Program (USAP) Data Center".

At the top right of the search results, there is a "Sort by:" dropdown menu, a "Clear all filters" button, and a "279 Results" indicator.

Future work:

- Linking data with tools ..
- Automatic launching of tools with data
- Search on map
- Getting these benefits in clowder via:
 - triple store sync with clowder
 - embedding science on schema

Transect data of coral species and other substrate types collected in the field using line transects in Palau and Yap in 2017 and in the Federated States of Micronesia in 2018

[Website](#) [Cite](#) [Metadata](#)

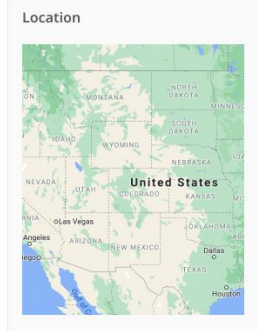
Type: Data

Abstract: As part of the reef-composition survey of Palau (7°30' N, 134°30' E) and Yap (9°32' N, 138°7' E), 10-meter long, 2 to 5-meter depth transects were conducted. Coral species along the transect were recorded along with substrate types and other organisms present. Surveys in Palau were conducted from June 2nd to June 24th, 2017, and from June 25th to July 6th, 2017 in Yap. In Pohnpei (6.2°N, 158.2°E) and Kosrae (5.3°N, 162.9°E) FSM, six 10-meter transects were used to measure the benthic composition for every centimeter, at each site of 48 sites. Corals were recorded to species level, except massive Porites and encrusting Montipora, which were recorded in the field as growth forms. All other organisms along each transect were identified to the highest possible taxonomic resolution.

Creator: Robert van Woesik

Publisher: Florida Institute of Technology

Date: 2020-09-08



Downloads

[Download TIFF](#)

[Download Shapefile](#)

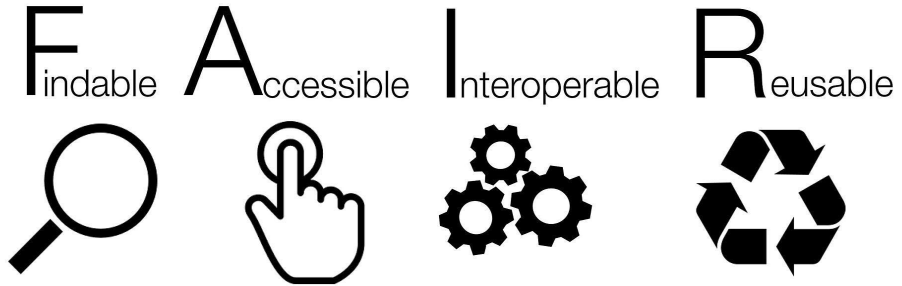
Related Data

- ▲ Coral densities and extension rates from scientific literature collected in the field or in laboratories
- ▲ Sea urchin size, density, and species from transects surveyed in Palau and Yap in 2017 and in the Feder...
- ▲ Parrotfish species, density counts, and fish length from field-video surveys in Palau and Yap in 2017...
- ▲ Transect data of coral species and other substrate types collected in the field using line transects in...
- ▲ Bacterial cell counts and Dissolved Organic Carbon (DOC) measurements from R/V Atlantis AT32, AT34...

Compatible Tool

- ▲ NetCDF classic format (netCDF)
- ▲ TopBraid Composer Free Edition
- ▲ LinkedEarth
- ▲ McIDAS grid file format (McIDASGrid)
- ▲ Application for Extracting and Exploring Analysis Ready Samples (AppEARS)

Faster time to science
via metadata use
to get more



resources

GeoCODES

- Involved with standard metadata adoption
- Crawl & Search made easier with more functionality
- Now incl Resource-Registry search & linking w/data

Try it at:

<https://geocodes.earthcube.org>

<https://earthcube.clowderframework.org>

<https://alpha.earthcube.org>

Can take questions later: @Mike Bobak

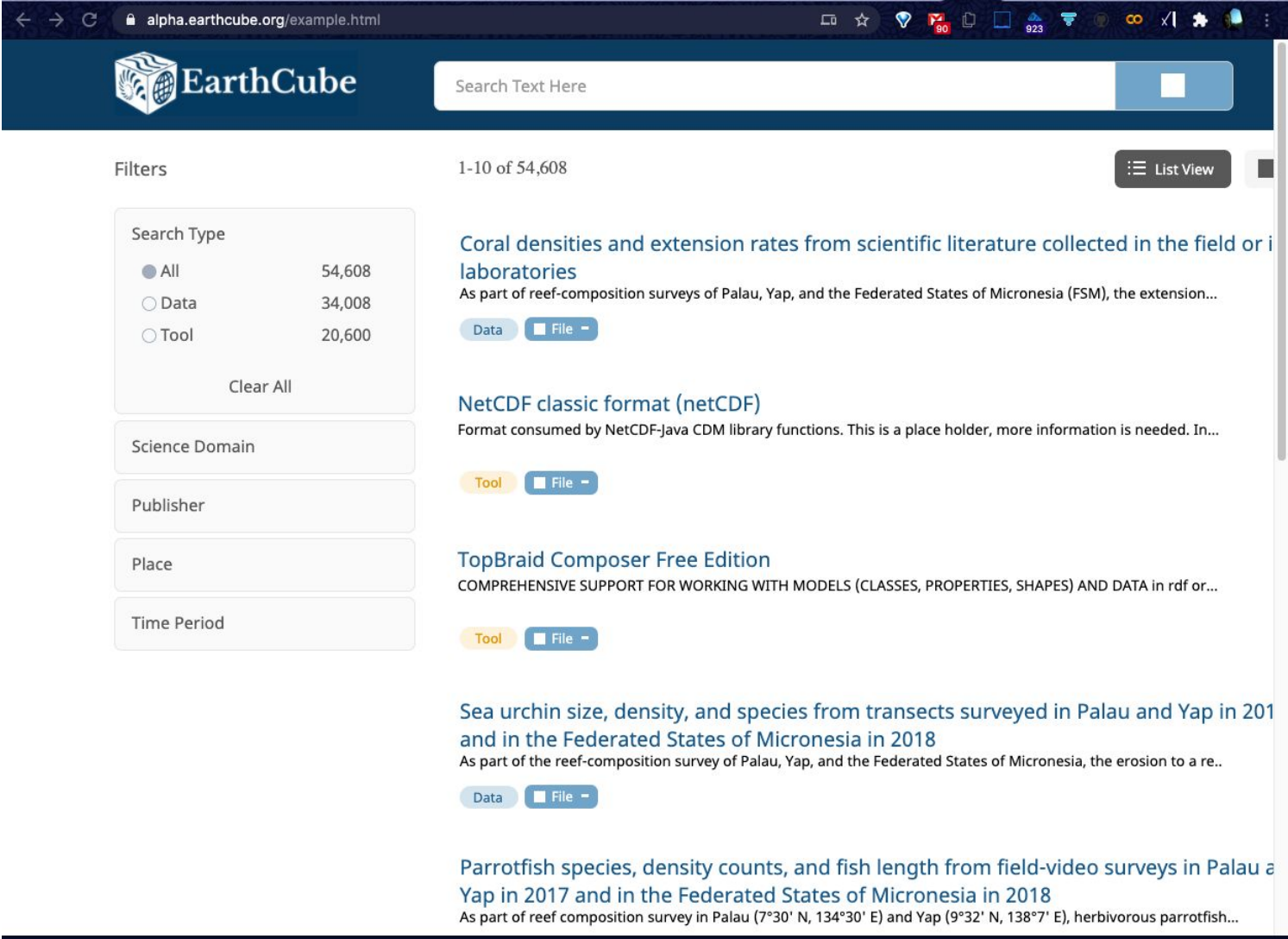
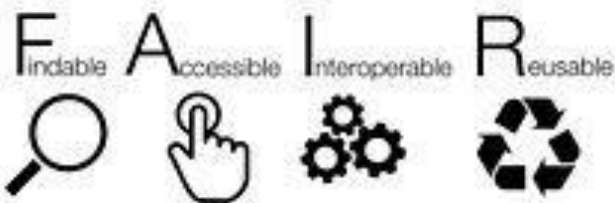
extra slides

There will be a way to get metadata ‘*details*’ & upcoming *data tool linkages* as well.

Also opening of the result’s *data & associated tools* in a *NoteBook* is also possible.

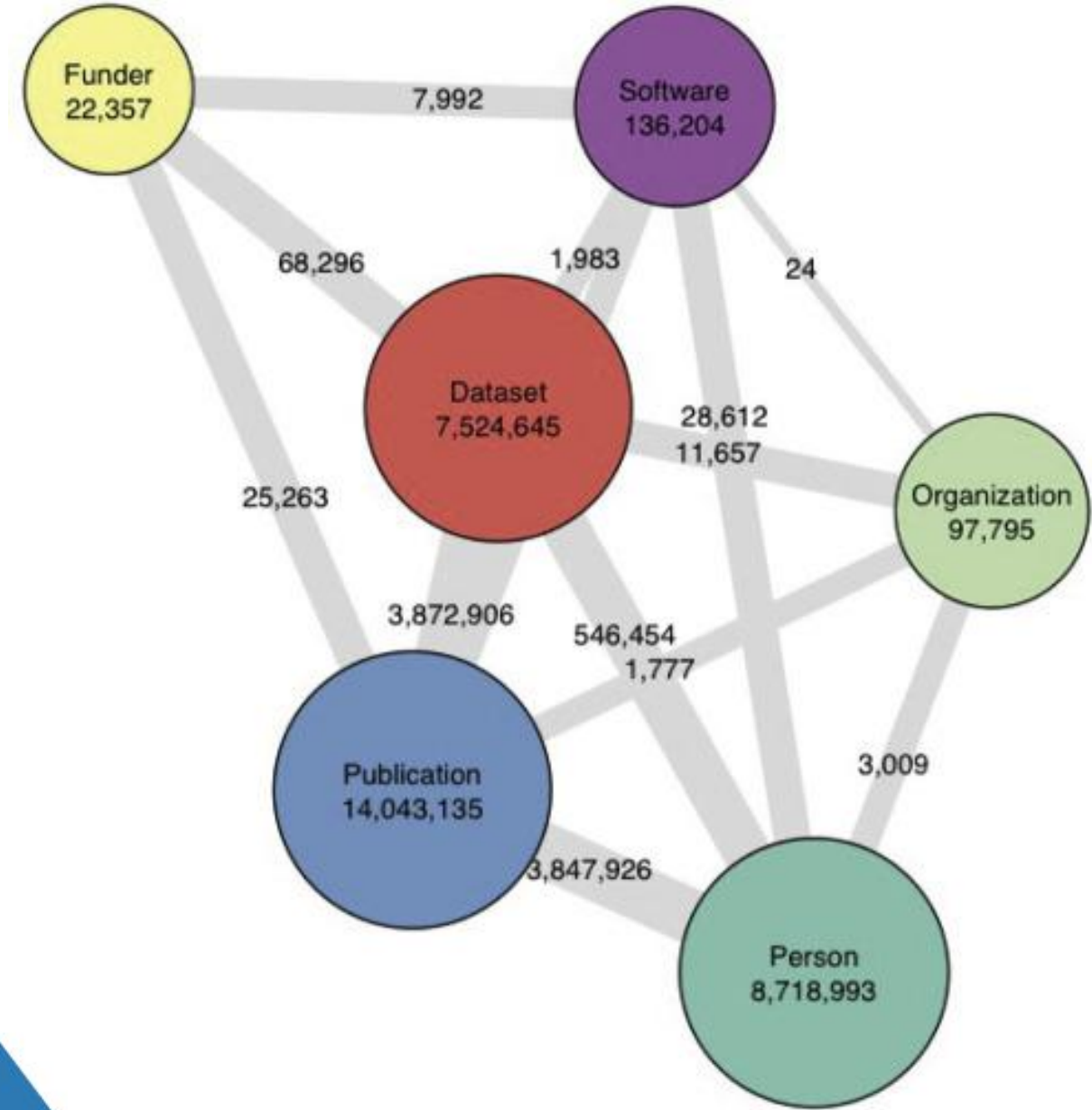
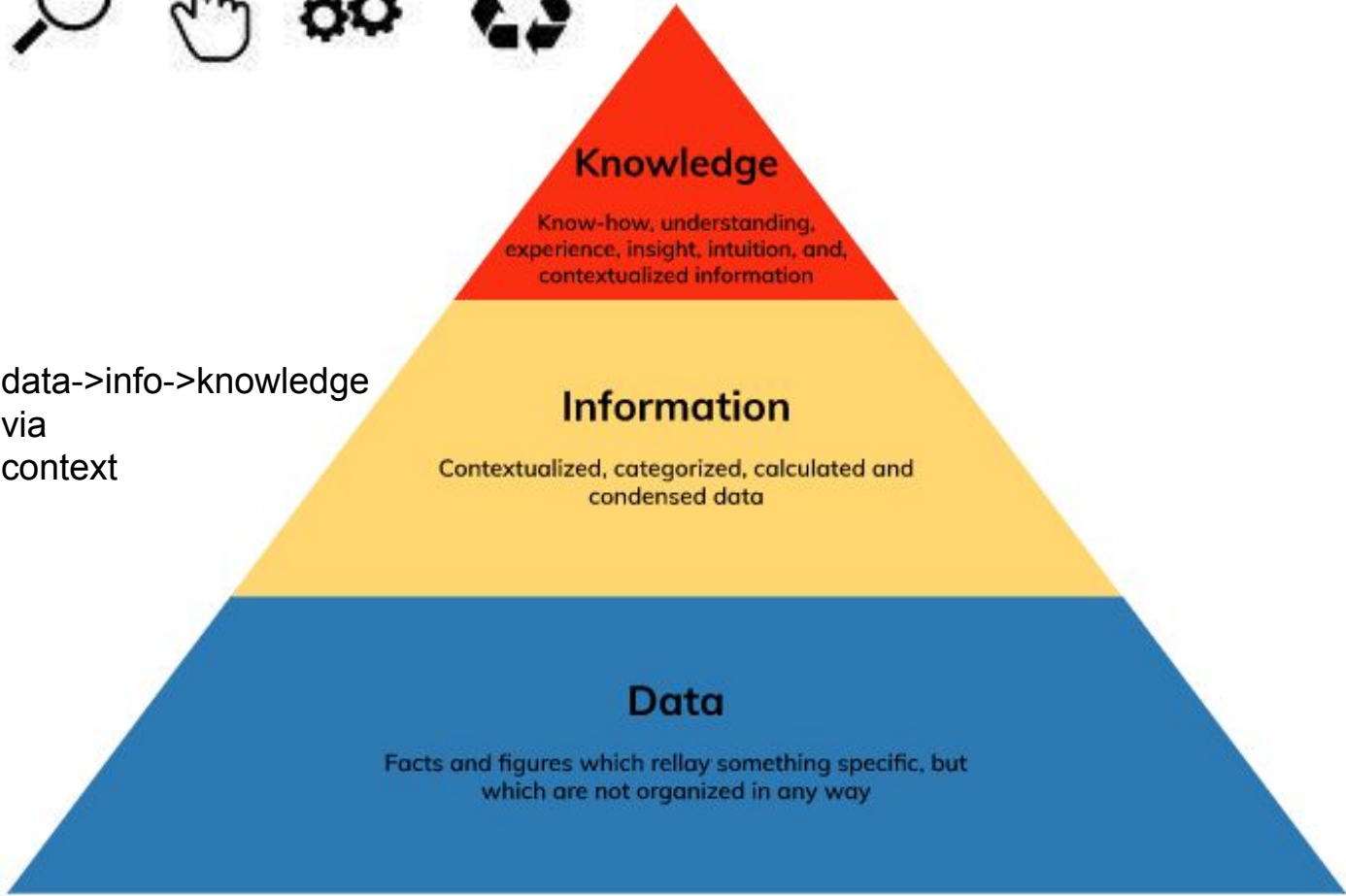
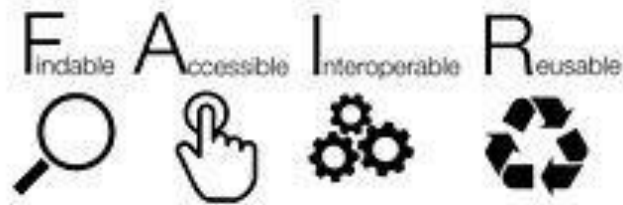
Throughput is an EC project that might help us bring in some more of these linkages

Linked-Data is what makes these resources:

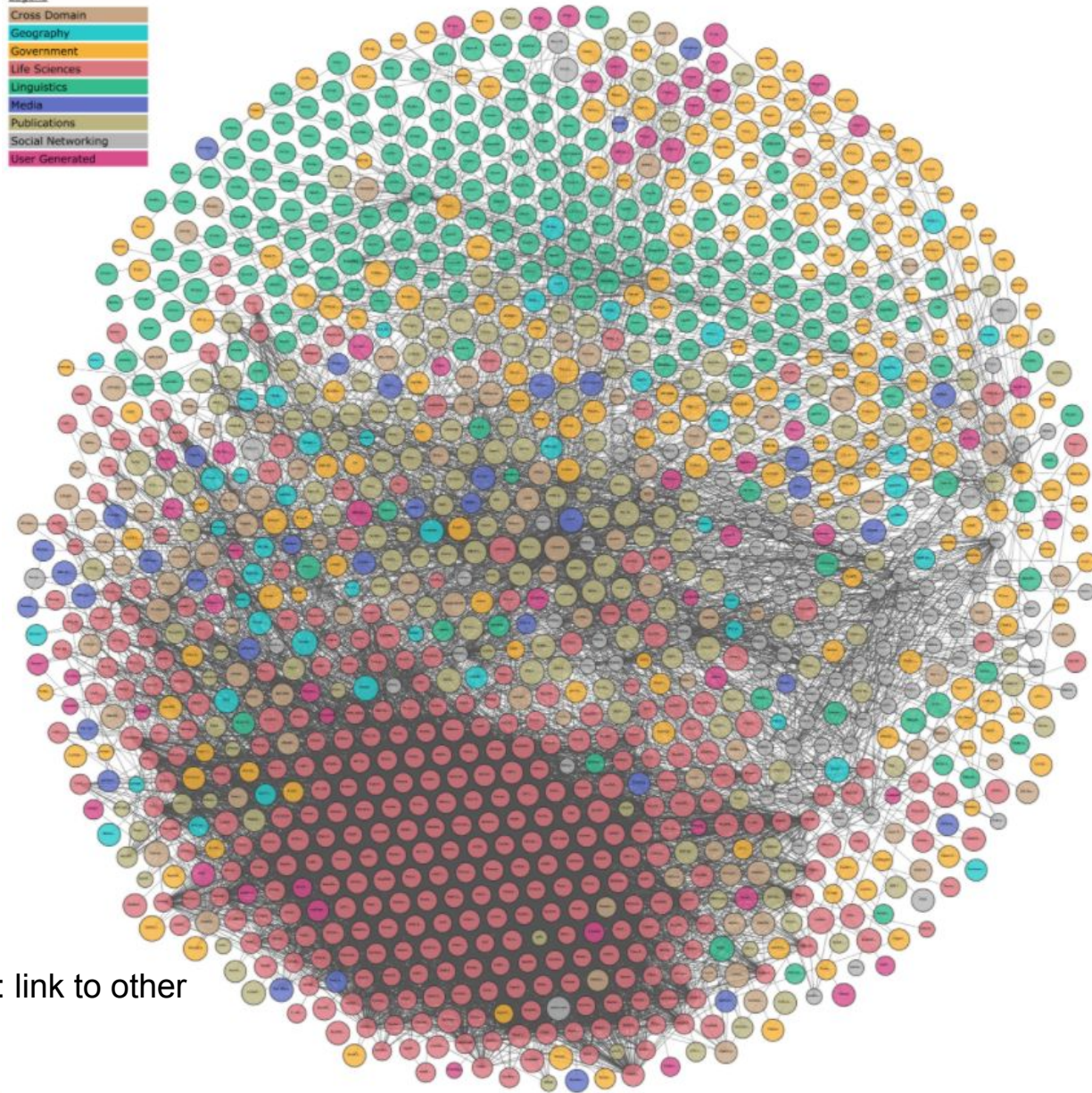
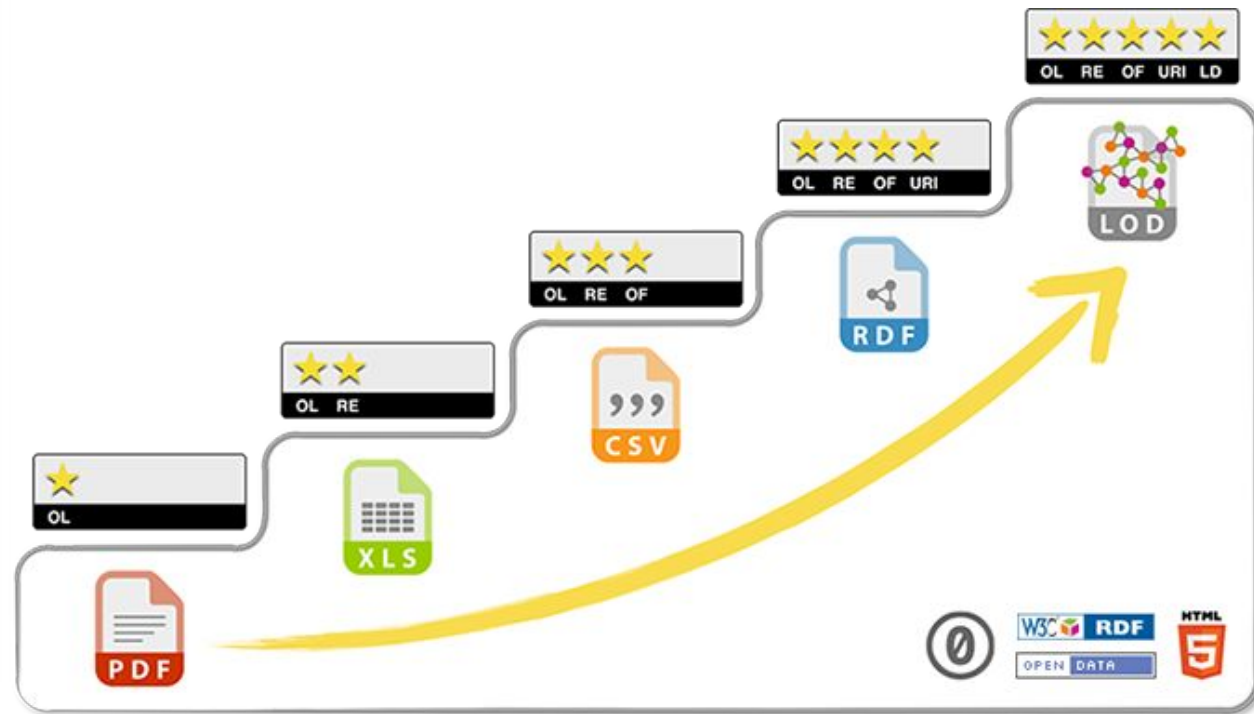


Here is the *static UX results page*, with all the filter-widgets collapsed, and icons to highlight what is *data* and what is a *tool*.

Throughput is an EC project that might help us bring in some more of these linkages
Linked-Data is what makes these resources



5stardata.info/en last star is linking to the
LinkedOpenData cloud lod-cloud.net



Available as: 1: open online, 2: structured, 3: non-proprietary, 4: ref via URIs, 5: link to other formats

FAIR Principles

[Home](#) · [FAIR Principles](#)

> FAIR Principles

- > **F1: (Meta) data are assigned globally unique and persistent identifiers**
- > **F2: Data are described with rich metadata**
- > **F3: Metadata clearly and explicitly include the identifier of the data they describe**
- > **F4: (Meta)data are registered or indexed in a searchable resource**
- > **A1: (Meta)data are retrievable by their identifier using a standardised communication protocol**
- > **A1.1: The protocol is open, free and universally implementable**

In 2016, the '[FAIR Guiding Principles for scientific data management and stewardship](#)' were published in *Scientific Data*. The authors intended to provide guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

A practical "how to" guidance to go FAIR can be found in the [Three-point FAIRification Framework](#).

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the [FAIRification process](#).

- F1. (Meta)data are assigned a globally unique and persistent identifier**
- F2. Data are described with rich metadata (defined by R1 below)**
- F3. Metadata clearly and explicitly include the identifier of the data they describe**
- F4. (Meta)data are registered or indexed in a searchable resource**

Implementable

- > **A1.2: The protocol allows for an authentication and authorisation where necessary**
- > **A2: Metadata should be accessible even when the data is no longer available**
- > **I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation**
- > **I2: (Meta)data use vocabularies that follow the FAIR principles**
- > **I3: (Meta)data include qualified references to other (meta)data**
- > **R1: (Meta)data are richly described with a plurality of accurate and relevant attributes**
- > **R1.1: (Meta)data are released with a clear and accessible data usage license**
- > **R1.2: (Meta)data are associated with detailed provenance**
- > **R1.3: (Meta)data meet domain-relevant community standards**
- > **How to GO FAIR**
- > **FAIRification Process**

Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure. For instance, principle F4 defines that both metadata and data are registered or indexed in a searchable resource (the infrastructure component).

GeoCODES

- Involved with standard metadata adoption
- Crawl & Search made easier with more functionality
- Now incl Resource-Registry search and linking

Try it at:

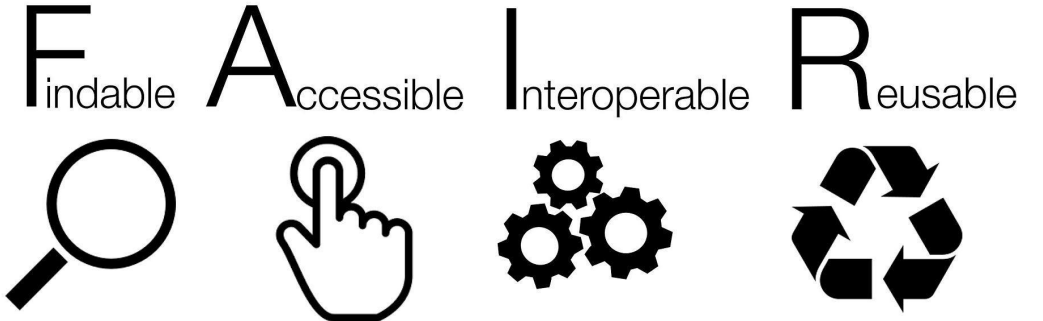
<https://geocodes.earthcube.org>

<https://earthcube.clowderframework.org>

<https://alpha.geocodes.earthcube.org>

Can take questions later: @Mike Bobak

*Faster time to science
via metadata use
to get more*



resources