Documenting Spatial Data

Reina Chano Murray

Geospatial Data Curator & Applications Administrator
Johns Hopkins University

Sept 28, 2022

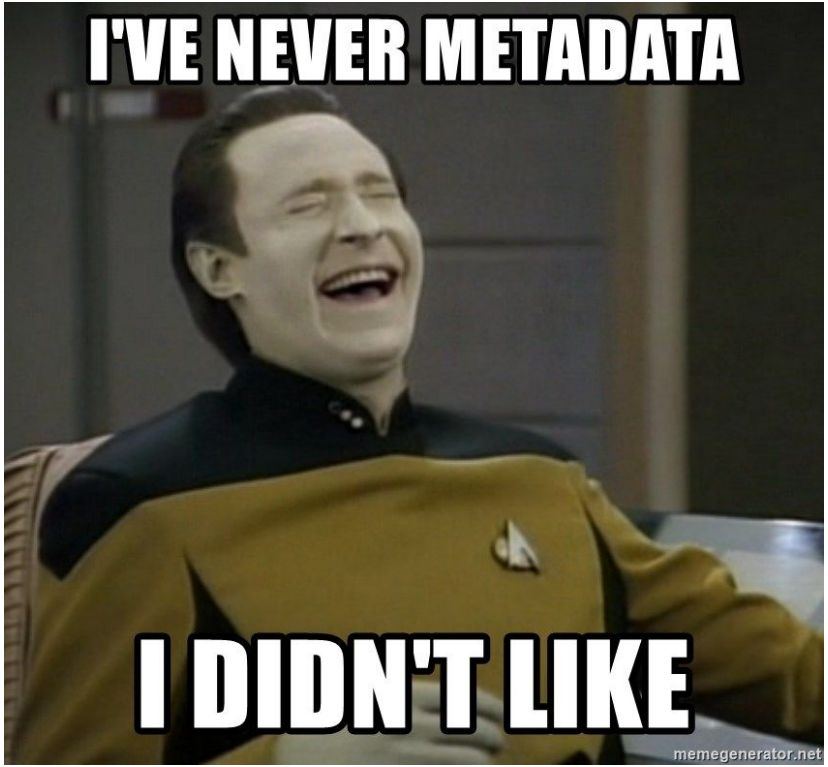
What We'll Cover

Metadata

- What is metadata?
- Why is it important?
- What are metadata standards?

Geospatial Metadata

- Geospatial metadata standards
- Geospatial metadata schemas
- Common elements to include
- Different levels of data
- Other recommendations for documentation (READMEs!)



Metadata

What is “metadata”?

*Metadata is **data** about **data**. It describes information like who collected it, when, for what purpose, and the level of quality… You can think of metadata as **little messengers to the future**.*

- Sarah Wakamiya, 2018

https://www.nps.gov/articles/what-in-the-world-is-metadata.htm

What is “metadata”?

- Provides context to data
  - making it easier to retrieve, use or manage the data in question

- Has some structure:

  *“the metadata is collected so that it can fulfill a useful purpose, and sorted into known categories. It is this notion of structure that turns raw information into **actionable metadata**.”*

  Jenn Riley, NISO, Understanding Metadata: What is Metadata, and What is it For?: A Primer, 2017)

What is “metadata”?

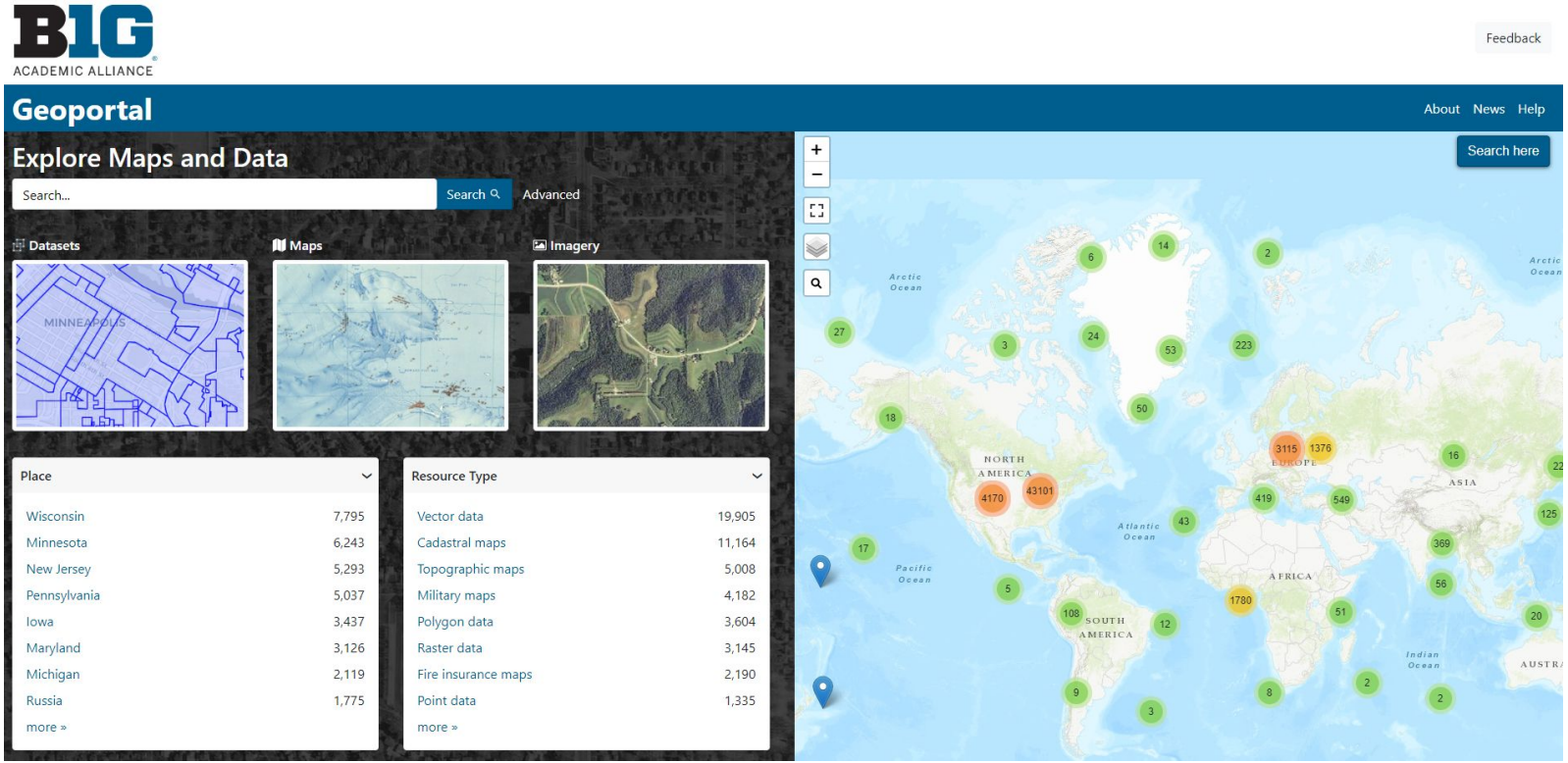
Descriptive metadata	For finding or understanding a resource
Administrative metadata - Technical metadata - Preservation metadata - Rights metadata	 - For decoding and rendering files - Long-term management of files - Intellectual property rights attached to content
Structural metadata	Relationships of parts of resources to one another
Markup languages	Integrates metadata and flags for other structural or semantic features within content

Jenn Riley, NISO, Understanding Metadata: What is Metadata, and What is it For?: A Primer, 2017, page 6)

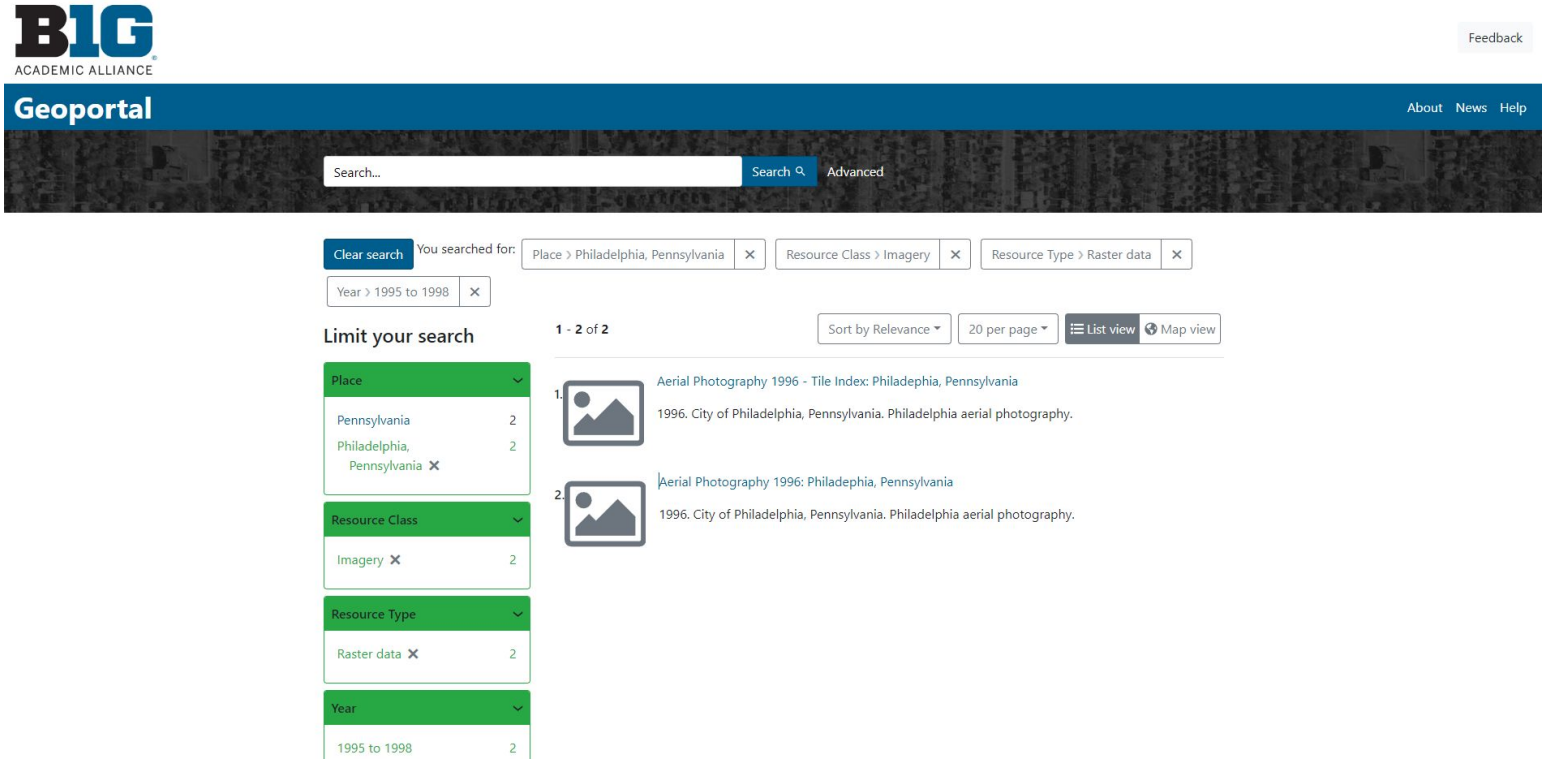
Why is metadata important?

Metadata makes your data more discoverable and understandable

What do we mean by “discoverable”?



What do we mean by “discoverable”?



Why is metadata important?

Good metadata helps others

trust,

validate,

reuse

and

build upon

your data.

Why is metadata important?

Tom Chatfield, a data management expert with the Bureau of Land Management bravely shared one real example from his own agency. A particular BLM field office developed an overall resource management plan for its region which included a map of oil and gas potential for its area. Naturally, this is an issue that registers strong opinions both for and against oil and gas recovery on public lands. When reviewed in the advent of litigation, BLM discovered that ***no Data Quality information existed for the map****. No data sources could be identified, nor was there any discussion of the analysis methodology by which the oil and gas potential information was developed for the map. No geologic or geographic data were identified.*

This was potentially disastrous in terms of defending the resource management plan. BLM was forced to develop a brand new map and closely document the provenance of the source data and methodology used to determine the potential ratings on the map. The ***cost exceeded several thousands of dollars in staff time and production costs****, not to mention the* ***intangible costs of loss of trust in the agency by partners and the public****.*

Zolly, L.S., Henkel, H.S., Hutchison, V.B., Langseth, M.L., Thibodeaux, C.J., 2015, USGS data management training modules—metadata for research data [Slide 10]: U.S. Geological Survey, https://doi.org/10.5066/F7RJ4GGJ.

What are metadata standards?

- Provides agreement on the meaning and semantics of the data
- Think of metadata standards as a set of rules, or an agreement, that specify the minimum amount of information that should be documented about a dataset (and how)
- standardization makes it easier for both **humans** and **computers** to find and understand your data
- Metadata standards include **schemas**
    - schemas provide the overall structure for the metadata - provide a set of elements that should be used to describe a dataset

How many standards are there?

- So. Many.
- Check out Jenn Riley's Metadata Map: http://jennriley.com/metadatamap/
- Other sources
  - https://rdamsc.bath.ac.uk/
  - https://fairsharing.org/
  - https://datastandards.directory/
  - https://www.iso.org/home.html

Commonly Required Core Elements

- **Title** – The name of the resource. Titles should include place names and dates when available.
- **Creator** – The person(s) or organization(s) that created the resource.
- **Date** – The publication date for the resource.
- **Subject** – Theme or topic keywords of the resource chosen from a controlled vocabulary such as Library of Congress Cartographic Genre and/or ISO Topic Category terms.
- **Description** – A description of the resource that may include additional relevant information, like data creation methods, data sources, and special licenses. If the description is minimal or lacking, it can be improved by concatenating available metadata fields, such as title, date, format, and place.
- **Rights** – Values such as "Public" or "Restricted" if applicable, as well as any licensing information.

Geospatial Metadata

Common Geospatial Metadata Standards

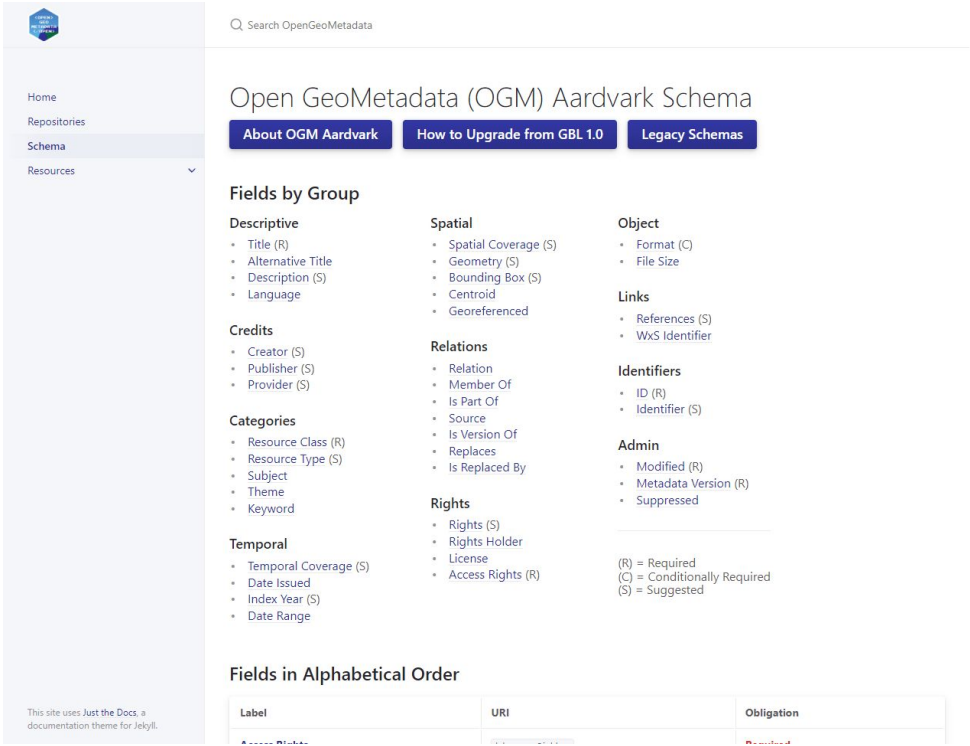
ISO 191xx	"In 1999 the International Organization for Standardization (ISO) Technical Committee (TC) 211 Geographic Information / Geomatics was tasked to harmonize the FGDC Content Standard for Digital Geospatial Metadata (CSDGM) with other formal and defacto standards that support the documentation of geospatial data and services. The resultant ISO 19115: Geographic information - Metadata standard was finalized in 2003 and endorsed by the FGDC in 2010. A series of additional ISO 191** standards have been developed to update, extend, and supplement the 19115 standard."
FGDC CSDGM	"Content Standard for Digital Geospatial Metadata (CSDGM), Vers. 2 (FGDC-STD-001-1998) is the current version of this FGDC authored and endorsed standard. Executive Order 12906, directed that Federal agencies document geospatial resources using this standard. ***Since the publication of EO 12906, the FGDC has endorsed several ISO Geospatial Metadata Standards that are now encouraged for use."***
ArcGIS Metadata	"The ArcGIS metadata format contains elements that can store all content in all metadata standards that are supported by ArcGIS Desktop. It includes all Federal Geographic Data Committee (FGDC) *Content Standard for Digital Geospatial Metadata (CSDGM)* metadata content, all ISO 19139 metadata content, all North American Profile metadata content, and all INSPIRE metadata content."
ESRI-ISO	An older metadata standard from ArcGIS Desktop that can be upgraded to the ArcGIS metadata format
INSPIRE	The European Union standard for spatial data sets and services

Common Elements to Include

- Bounding box
- Geographic location
  - (place keywords -- such as https://www.geonames.org/)
- Spatial representation / geometry type
- Projection / coordinate reference system
- Provenance (how the data was made)
  - Geoprocessing workflow: geoprocessing tools used to create data
  - System/Software environment: software and software version used to create data
- Geospatial metadata schema used
- Maintenance schedule (if data will be updated)

NOTE: some of this get automatically taken care of by your GIS; others require manual entry

Example of a Geospatial Metadata Schema

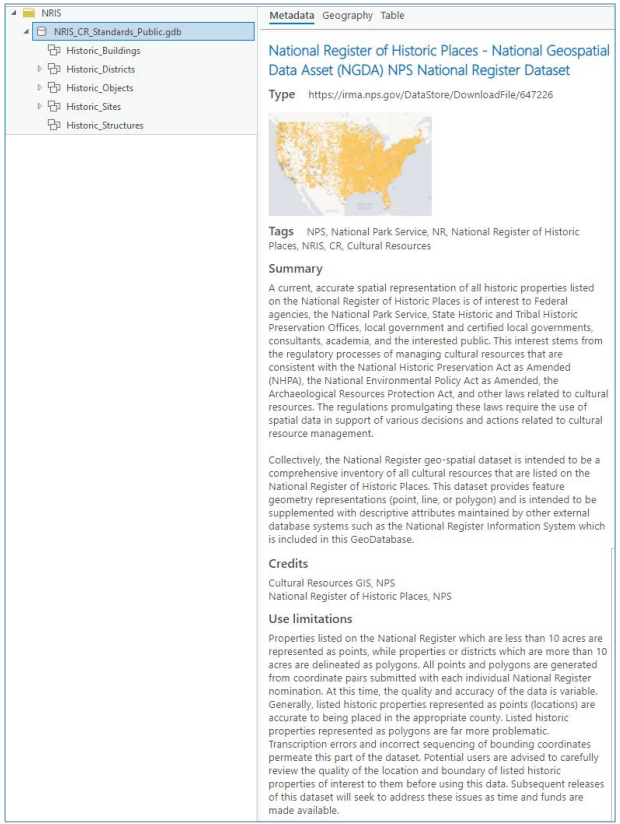


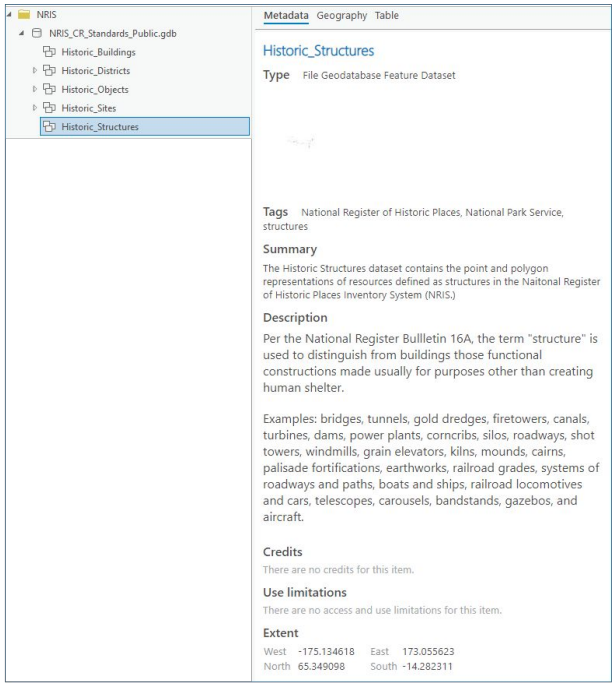
https://opengeometadata.org/docs/ogm-aardvark

Different Levels of Data to Document

- File-level
- Geodatabase/collection-level
- Presentation-level

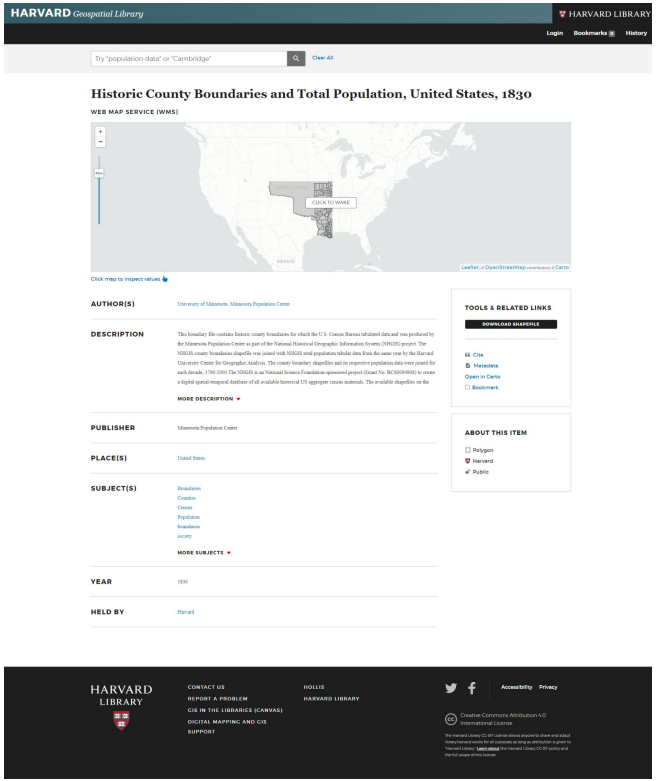
Collection/Geodatabase-level + File-level

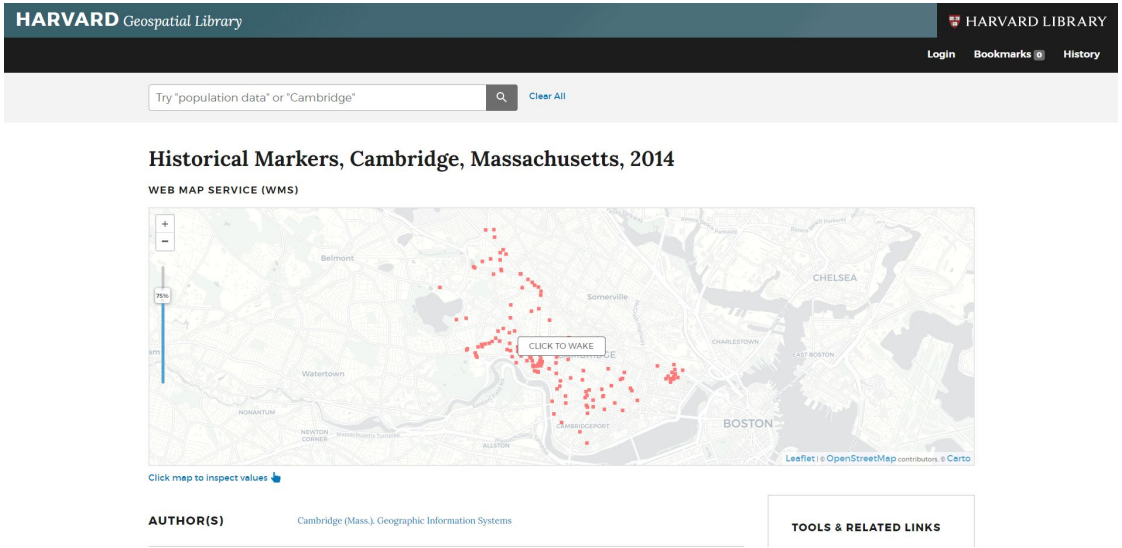




Stutts M. 2014. National Register of Historic Places. National Register properties are located throughout the United States and their associated territories around the globe. https://irma.nps.gov/DataStore/Reference/Profile/2210280

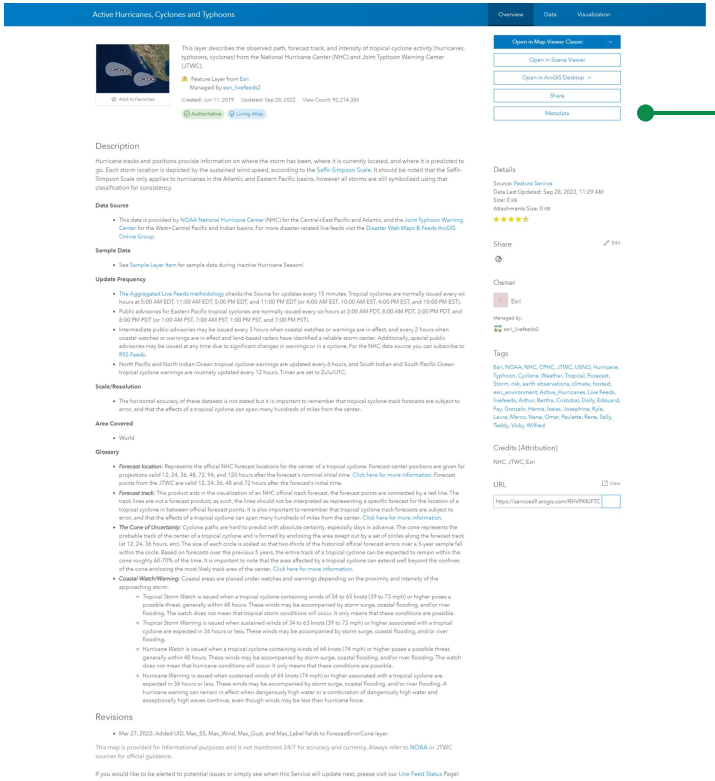
Presentation-level





https://hgl.harvard.edu/catalog/harvard-nhgis-pop1830

Presentation-level



Active Hurricanes, Cyclones and Typhoons

ISO-19139 Metadata

Metadata Information:

Metadata language:
Metadata character set: utf8

Last update: 2022-08-16

Metadata contact - pointOfContact:

Organization's name: Esri

Scope of the data described by the metadata: dataset
Scope name: dataset

Name of the metadata standard used: ISO 19139 Geographic Information - Metadata - Implementation Specification
Version of the metadata standard: 2007

Metadata identifier: 248e7b5827a34b248647afb012c58787

Resource Identification Information:

Citation:

Title: Active Hurricanes, Cyclones and Typhoons

Reference date - publication: 2022-08-16

Themes or categories of the resource: climatologyMeteorologyAtmosphere, environment

Place keywords:

Keywords: United States of America

Abstract:

Hurricane tracks and positions provide information on where the storm has been, where it is currently located, and where it is predicted to go. Each storm

Forecast points from the JTWC are valid 12, 24, 36, 48 and 72 hours after the forecast's initial time.Forecast track: This product aids in the visualizatio

The cone represents the probable track of the center of a tropical cyclone and is formed by enclosing the area swept out by a set of circles along the fore

Purpose:

This layer describes the observed path, forecast track, and intensity of tropical cyclone activity (hurricanes, typhoons, cyclones) from the National Hurr

Dataset language: eng
Dataset character set: utf8

Status: completed
Resource constraints:

Constraints:

Limitations of use: This work is licensed under the Esri Master License Agreement.View Summary | View Terms of Use

Resource constraints:

Legal constraints:

Access constraints: otherRestrictions.

Limitations of use: Other Constraints

Resource constraints:

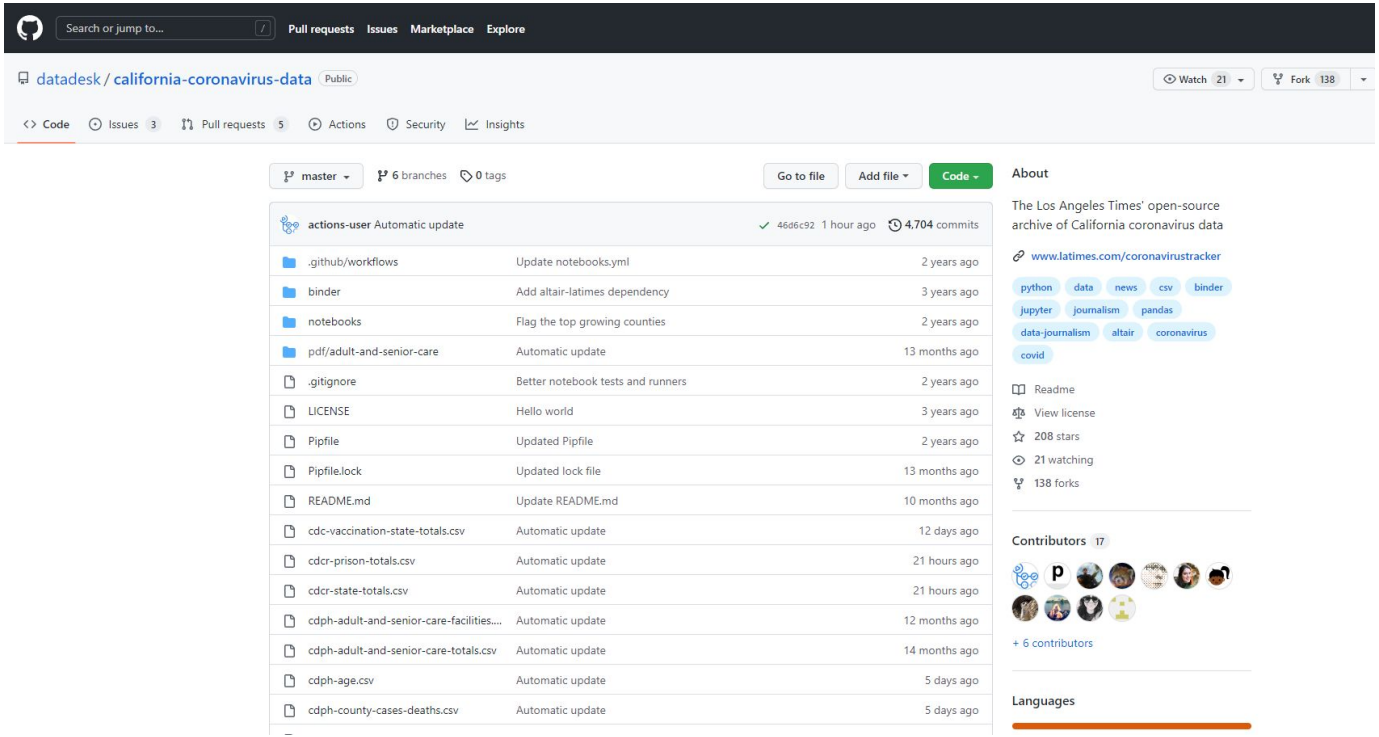
Legal constraints:

Use constraints:

Limitations of use: Other Constraints

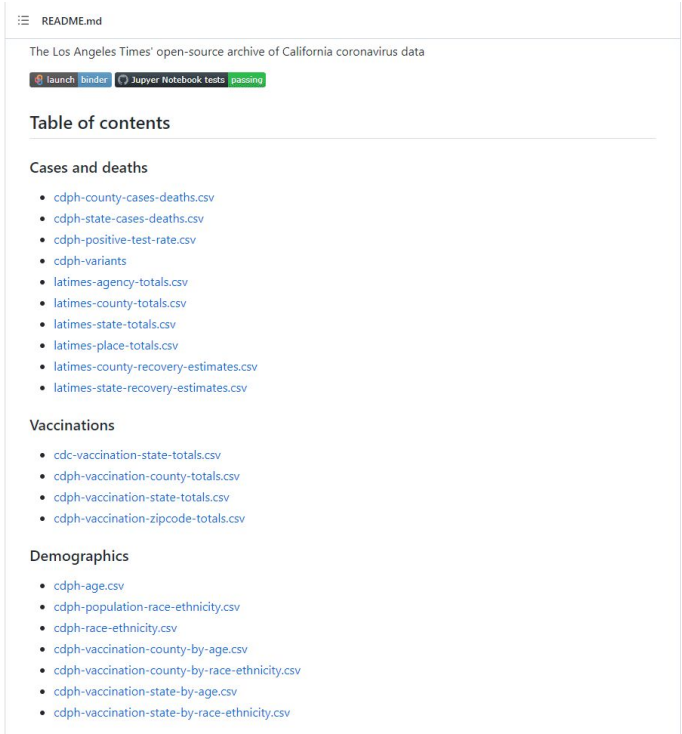
https://www.arcgis.com/home/item.html?id=248e7b5827a34b248647afb012c58787

Presentation-level



https://github.com/datadesk/california-coronavirus-data

Presentation-level



README.md

The Los Angeles Times' open-source archive of California coronavirus data

launch binder | Jupyer Notebook tests passing

Table of contents

Cases and deaths

- cdph-county-cases-deaths.csv
- cdph-state-cases-deaths.csv
- cdph-positive-test-rate.csv
- cdph-variants
- latimes-agency-totals.csv
- latimes-county-totals.csv
- latimes-state-totals.csv
- latimes-place-totals.csv
- latimes-county-recovery-estimates.csv
- latimes-state-recovery-estimates.csv

Vaccinations

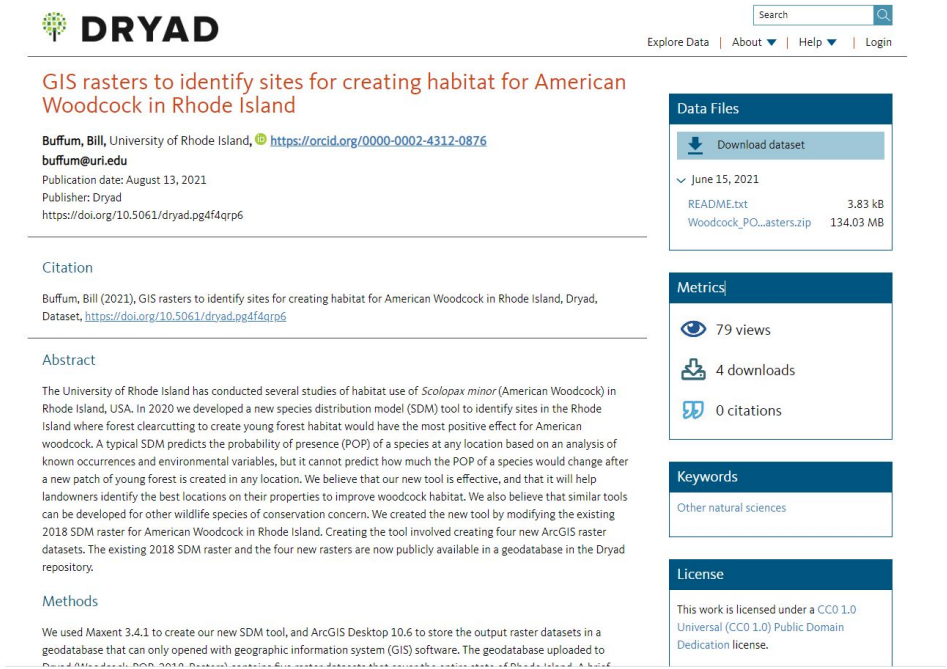
- cdc-vaccination-state-totals.csv
- cdph-vaccination-county-totals.csv
- cdph-vaccination-state-totals.csv
- cdph-vaccination-zipcode-totals.csv

Demographics

- cdph-age.csv
- cdph-population-race-ethnicity.csv
- cdph-race-ethnicity.csv
- cdph-vaccination-county-by-age.csv
- cdph-vaccination-county-by-race-ethnicity.csv
- cdph-vaccination-state-by-age.csv
- cdph-vaccination-state-by-race-ethnicity.csv

https://github.com/datadesk/california-coronavirus-data

Presentation-level



DRYAD

Search

Explore Data | About | Help | Login

GIS rasters to identify sites for creating habitat for American Woodcock in Rhode Island

Buffum, Bill, University of Rhode Island, https://orcid.org/0000-0002-4312-0876
buffum@uri.edu
Publication date: August 13, 2021
Publisher: Dryad
https://doi.org/10.5061/dryad.pg4f4qrp6

Citation

Buffum, Bill (2021), GIS rasters to identify sites for creating habitat for American Woodcock in Rhode Island, Dryad, Dataset, https://doi.org/10.5061/dryad.pg4f4qrp6

Abstract

The University of Rhode Island has conducted several studies of habitat use of *Scolopax minor* (American Woodcock) in Rhode Island, USA. In 2020 we developed a new species distribution model (SDM) tool to identify sites in the Rhode Island where forest clearcutting to create young forest habitat would have the most positive effect for American woodcock. A typical SDM predicts the probability of presence (POP) of a species at any location based on an analysis of known occurrences and environmental variables, but it cannot predict how much the POP of a species would change after a new patch of young forest is created in any location. We believe that our new tool is effective, and that it will help landowners identify the best locations on their properties to improve woodcock habitat. We also believe that similar tools can be developed for other wildlife species of conservation concern. We created the new tool by modifying the existing 2018 SDM raster for American Woodcock in Rhode Island. Creating the tool involved creating four new ArcGIS raster datasets. The existing 2018 SDM raster and the four new rasters are now publicly available in a geodatabase in the Dryad repository.

Methods

We used Maxent 3.4.1 to create our new SDM tool, and ArcGIS Desktop 10.6 to store the output raster datasets in a geodatabase that can only opened with geographic information system (GIS) software. The geodatabase uploaded to

Data Files

Download dataset

June 15, 2021

README.txt	3.83 kB
Woodcock_PO...asters.zip	134.03 MB

Metrics

79 views

4 downloads

0 citations

Keywords

Other natural sciences

License

This work is licensed under a CC0 1.0 Universal (CC0 1.0) Public Domain Dedication license.

Methods

We used Maxent 3.4.1 to create our new SDM tool, and ArcGIS Desktop 10.6 to store the output raster datasets in a geodatabase that can only opened with geographic information system (GIS) software. The geodatabase uploaded to Dryad (Woodcock_POP_2018_Rasters) contains five raster datasets that cover the entire state of Rhode Island. A brief description of the five rasters is provided below. We provide a detailed description of how we created our new SDM tool, and how the tool can be used, in the following journal article: **Bill Buffum, Roger Masse and Scott R. McWilliams (2021) Novel Use of Species Distribution Modeling to Identify High Priority Sites for American Woodcock Habitat Management. Northeastern Naturalist 28: p. 233-247.**

Woodcock_POP_2018: This raster (10m pixels) shows the 2018 probability of presence for woodcock in RI. It is based on Maxent modelling of three years woodcock location data, obtained by catching 68 woodcock in RI in the spring, attaching transmitters, and tracking their movements until August. The Maxent model was projected onto 2018 spatial data.

Woodcock Upland POP 2018: This raster (50m pixels) shows the 2018 probability of presence for woodcock in RI. It is similar to Woodcock POP 2018, but it has 50m pixels and only includes upland forest areas which are suitable for clearcuts to create young forest habitat.

Woodcock POP After Forest Mgt 2018: This raster (50m pixels) was the primary product of our new SDM tool. It shows the predicted probability of presence for woodcock in RI if young forest habitat is created at any upland forest location in RI.

Increase in Woodcock POP After Forest Mgt 2018: This raster (50m pixels) shows the predicted increase in probability of presence for woodcock in RI if young forest habitat is created at any site. It was produced by using raster calculator to subtract the POP values of Woodcock Upland POP 2018 from Woodcock POP After Forest Mgt 2018

Top Priority Sites for Forest Mgt: This raster (50m pixels) shows the top priority (top 20% and top 40%) sites for creating young forest habitat for woodcock in Rhode Island. The top 20% sites include area where both the predicted POP increase after forest management and the predicted POP after forest management are in the top 20% class. The top 40% sites include areas where both the predicted POP increase after forest management and the predicted POP after forest management are in the top 40% range.

Usage Notes

The ArcGIS raster datasets can only be viewed by using geographic information system (GIS) software such as ArcGIS Desktop or ArcGIS Pro (Environmental Systems Research Institute, Redlands, CA). The geodatabase, a compressed version of which is uploaded to Dryad, contains a total of 619 files, but when it is opened with GIS software, the five raster datasets will be visible and easily opened.

https://datadryad.org/stash/dataset/doi:10.5061%2Fdryad.pg4f4qrp6
https://orcid.org/0000-0002-4312-0876

Tips for Documentation

- READMEs are your best friend
- Document your data along the way - saves you time at the end
- Use descriptive file names
- If you’re using geospatial desktop software or web GIS, create your metadata in the platform/software you start in (inheritance)
    - Convert metadata into XML

README Files - What is it?

A README file is a text file containing key information about your data which gives the reader a general understanding of the purpose and history of your data set, how it is organized, and how it can be used. You can think of a README file as a **manual for your data**.

You can use it to capture components of your data that are not adequately captured in the metadata contained with your geoprocessing application.

README Files - Why create one?

While some of the information contained within your README file may overlap with the content you entered in the metadata within your geoprocessing tool, it is still a good idea to create a separate file that lives outside of your geospatial file.

This is helpful for not only if you share your data with others, but is also beneficial to yourself if you need to revisit your data in the future.

Furthermore, some geospatial formats (ie, GeoJSON) or other formats you might save your file as (csv) will not store metadata like a shapefile.

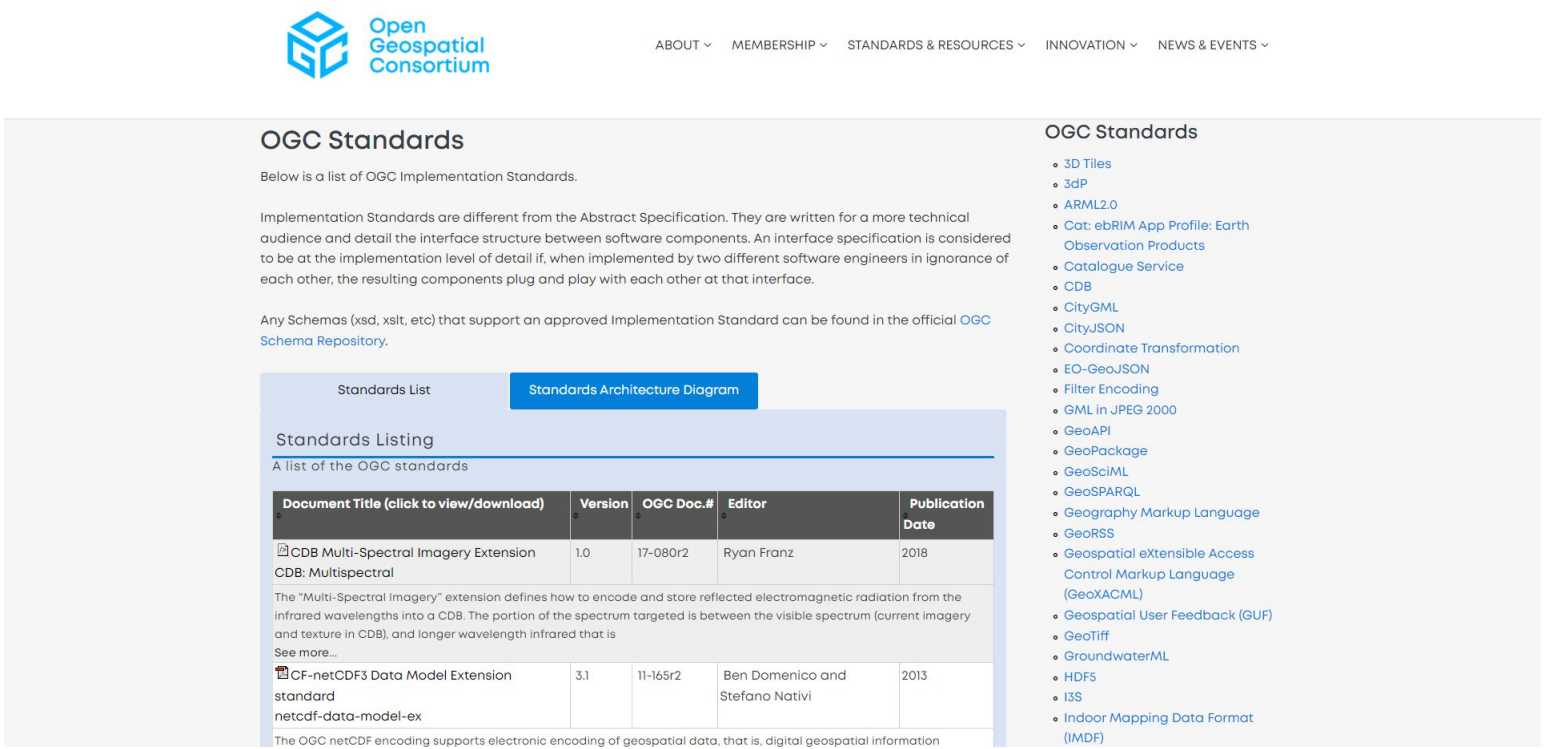
Cover all your bases.

README Files - What to include?

- High-level/core information such as purpose of the data set, where the files can be found
- Geospatial-specific elements such as coordinate ref. system,
- Workflow/software environment such as software version, data manipulations which occurred outside of your geoprocessing tool, data version history
- Details about your files such as the purpose of key files and your file naming scheme

See https://data.research.cornell.edu/content/readme

OGC *Implementation* Standards



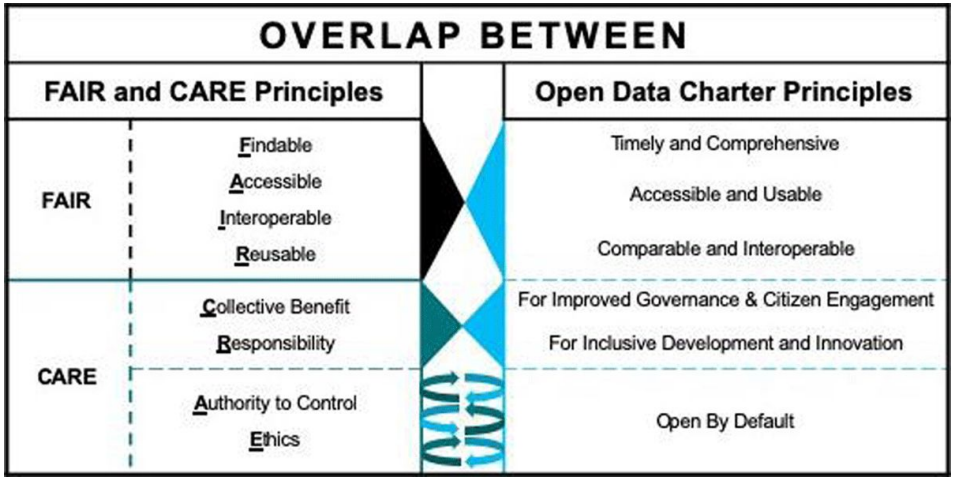
https://www.ogc.org/docs/is

Data Principles to Know

FAIR
https://www.go-fair.org/fair-principles/

CARE
https://www.gida-global.org/care

Open Data Charter
https://opendatacharter.net/principles/



Questions?

Reina Chano Murray

https://github.com/reinacmurray

reinamurray@jhu.edu