

Open Knowledge Network Workshop: Geosciences Report Out

Gary Berg-Cross
 Ruth Duerr
 Daniel Garijo
 Yolanda Gil
 Tsengdar Lee
 Andrew Moore
 Shashi Shekar
 Jia Zhang

4-5 October 2017

USC Information Sciences Institute Yolanda Gil gil@isi.edu 1

Geosciences Queries: Users

- Scientists
 - Sometimes spend months or years locating datasets by hand
- Citizens
- Businesses
- Policy makers
- Resource managers (eg water)
- City planning
- First responders

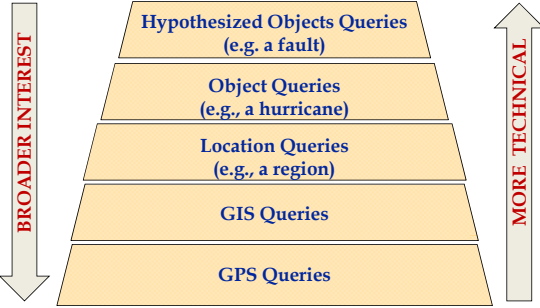
USC Information Sciences Institute Yolanda Gil gil@isi.edu 2

Geoscience Queries: Characteristics

- Data is grounded in space/time
 - Gridded at different granularities, interpolation possible
 - Subsetting and aggregation
- Data is often stored in efficient grid-based formats
 - E.g., NetCDF, Shape files
- Target objects may be challenging to define
 - Objects may have unclear boundaries
 - E.g., the Chesapeake Bay
 - Object boundaries may vary over time
 - E.g., a storm, a hurricane
 - Objects may be unobservable
 - Eg the San Andreas Fault
- Uncertainty: in observations, time, locations, models
- Provenance is very important

USC Information Sciences Institute Yolanda Gil gil@isi.edu 3

Geosciences Queries: Layers



USC Information Sciences Institute Yolanda Gil gil@isi.edu 4

Short-Term Steps: (I) Understand the Nature of Geo Queries

Regular Citizens	Scientists
<ul style="list-style-type: none"> ■ Example: What is the trend in air quality in the last month in Bethesda? ■ Approach: Get 100K relevant queries from major search engine and analyze trends and patterns 	<ul style="list-style-type: none"> ■ Example: Find fisheries data for the last 50 years in the upper Hudson river ■ Approach: Analyze EarthCube reports from 50 vision workshops in many geoscience areas (corals, geomorphology, etc.)

USC Information Sciences Institute Yolanda Gil gil@isi.edu 5

Short-Term Steps: (II) Making Existing Geo Data Searchable

- Question: What would be the minimum effort to make existing datasets indexed by search engines?
 - Is indexing through schema.org sufficient?
 - Do we need to promote space and time standards?
- Problem: Shortcomings of schema.org
 - Space/time representation is too simplistic
 - No clear immediate benefits to data providers
 - Complexity of use
- Approach:
 - Suggest specific improvements to schema.org

USC Information Sciences Institute Yolanda Gil gil@isi.edu 6

Other Short-Term Steps

- Connecting other kinds of data through space/time
 - Financial data, health data, etc.
- Linking data and OKN to encyclopedias and reference works
- Ratings from consumers
- Capture intent of queries
 - Customization and presentation to types of users
 - Granularity, uncertainty, etc.
- Data quality

Mid-Term Steps:

(I) The Social Network of Data (think Facebook)

Connecting data with people and organizations

- Landing page for key datasets
- People could
 - Comment on data
 - Rate data
- Link data to providers
- Link users of same datasets
 - Providers can see it!

Connecting data with software

- Link to visualization software that is useful for a dataset
- Link to pre-processing software appropriate for data type/format
- Link to models can be used with a dataset

Mid-Term Steps:

(II) NL Query Interfaces for Space/Time Data

- Problem: lots of data already available through APIs but only accessible if one knows endpoint and language
 - NASA, CUAHSI, BCO-DMO, LTER,...
- Approach:
 - Make them accessible through search engines
 - Natural language queries focused on data on time and space should retrieve data from existing repositories
 - Build on the analysis of queries proposed earlier

Other Mid-Term Steps

- Data fusion tools
 - When many datasets/sources are available
- Easy mapping across vocabularies/ontologies
 - Immediate need for SWEET <> ENVO
- "What-if scenario" queries
 - What would happen if there is a drought for 5 years
 - What happens to my house if water levels reach 1m in river pass
 - Are my insurance premiums likely to raise
- Real-time data
 - Eg rainfall during a tornado
- Very large amount of "dark data"
 - Collected by individual investigators, shelved in hard drives