The OKN KONQUER project: toward knowledge-based querying of semantically-enhanced biomedical and geoscience data sources

Lucila Ohno-Machado (PI), Peter Rose, Ilya Zaslavsky (UC San Diego),
Hua Xu, Kirk Roberts (UT Health Science Center - Houston),
Joseph Hamman (National Center for Atmospheric Research),
George Alter (Inter-university Consortium for Political and Social Research)

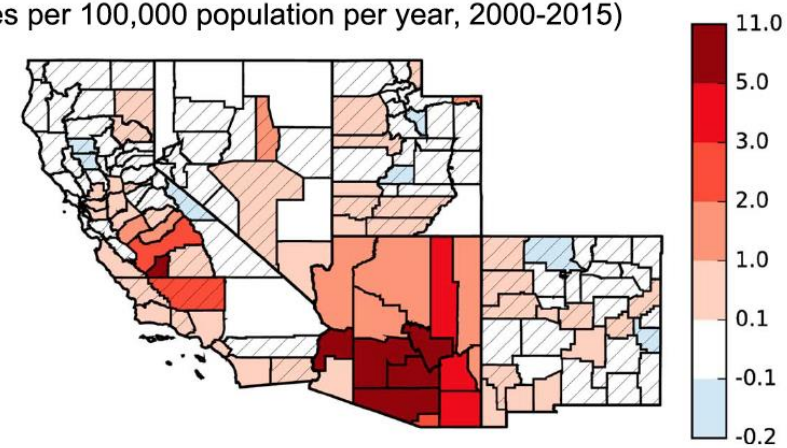
Valley Fever

Hotspots

- San Joaquin Valley CA
- Southcentral AZ

149,000 case reports in southwestern US (2000-2015)

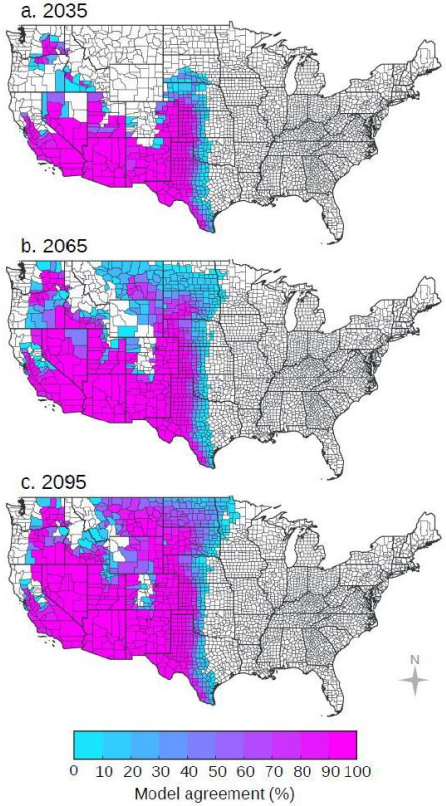
c. Annual valley fever incidence trends
(cases per 100,000 population per year, 2000-2015)



ME Gorris, et al. (2018) GeoHealth, 2

Climate and environmental drivers

Predicted spread of Valley Fever

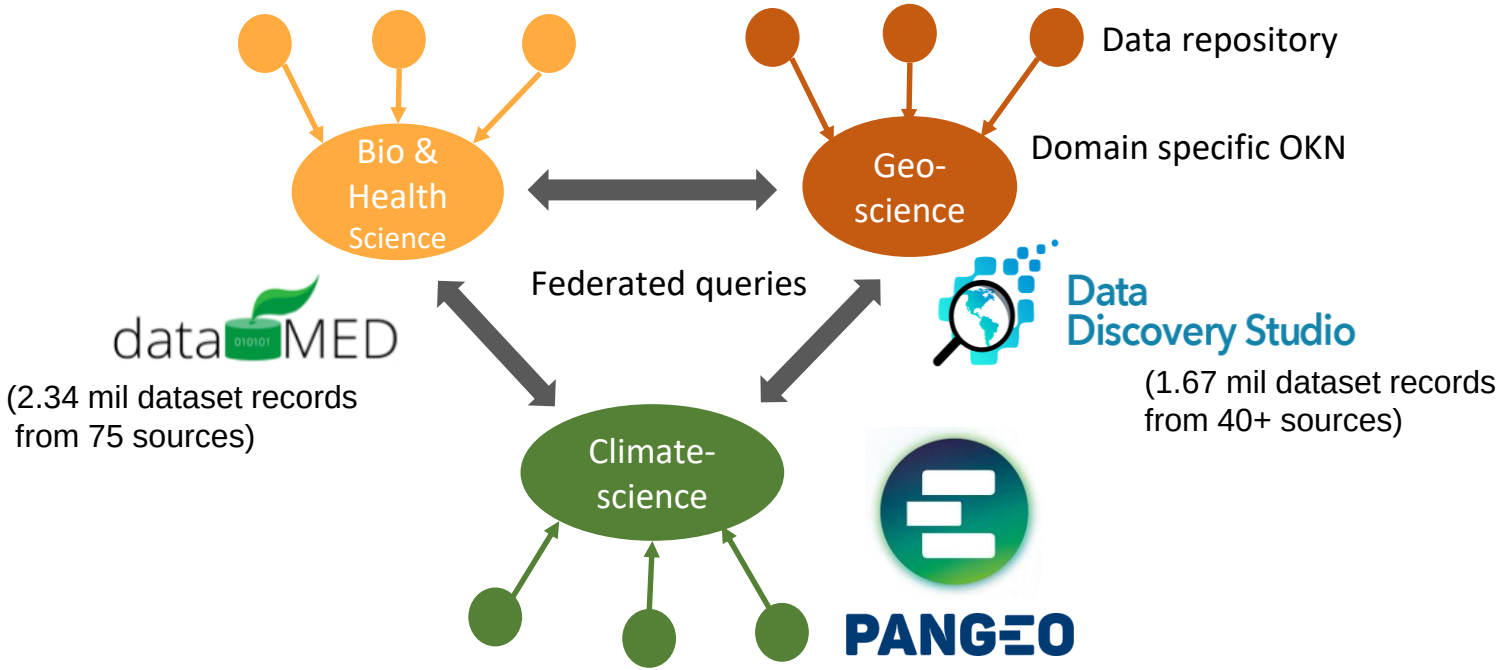


ME Gorris, et al. (2019) GeoHealth, 3

Our vision is to deliver a search engine, KONQUER

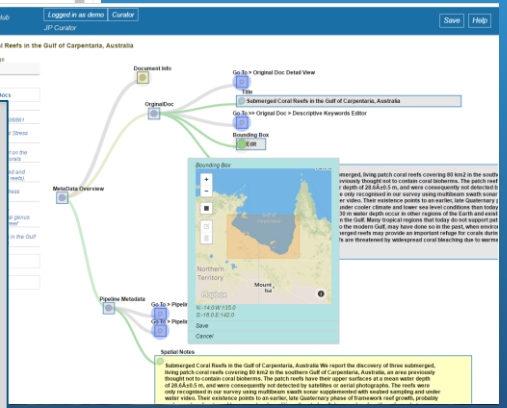
For researchers to obtain and integrate relevant data sets from multiple scientific domains

KONQUER: Knowledge Open Network and Queries for Research



Use case: Finding data sets to answer, "Was the number of cases of Valley Fever increased as a result of precipitation levels in California's Central Valley in 2016?"

EarthCube Data Discovery Studio



- Standard metadata formats and APIs
- Automated semantic processing and indexing pipeline
- Manual metadata editing and validation
- Provenance tracing
- Faceted search + full text, map, time
- Collection management
- Jupyter notebooks from any resource or collection
- Schema.org export for Google indexing

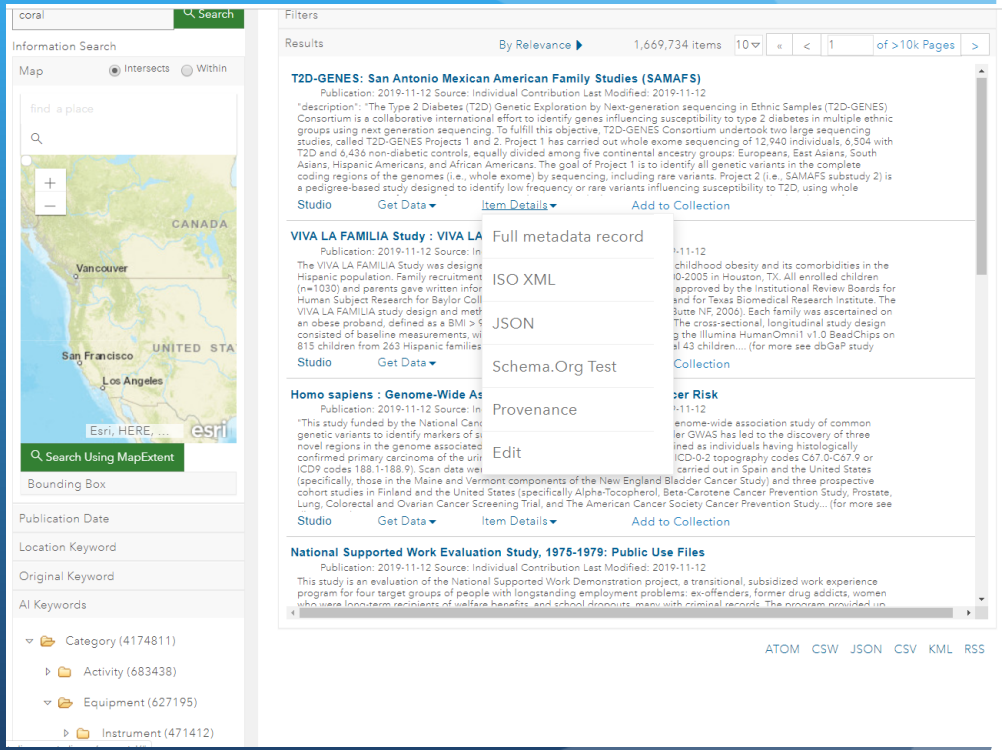
1.67 million records

From 40+ repositories and EC contributions

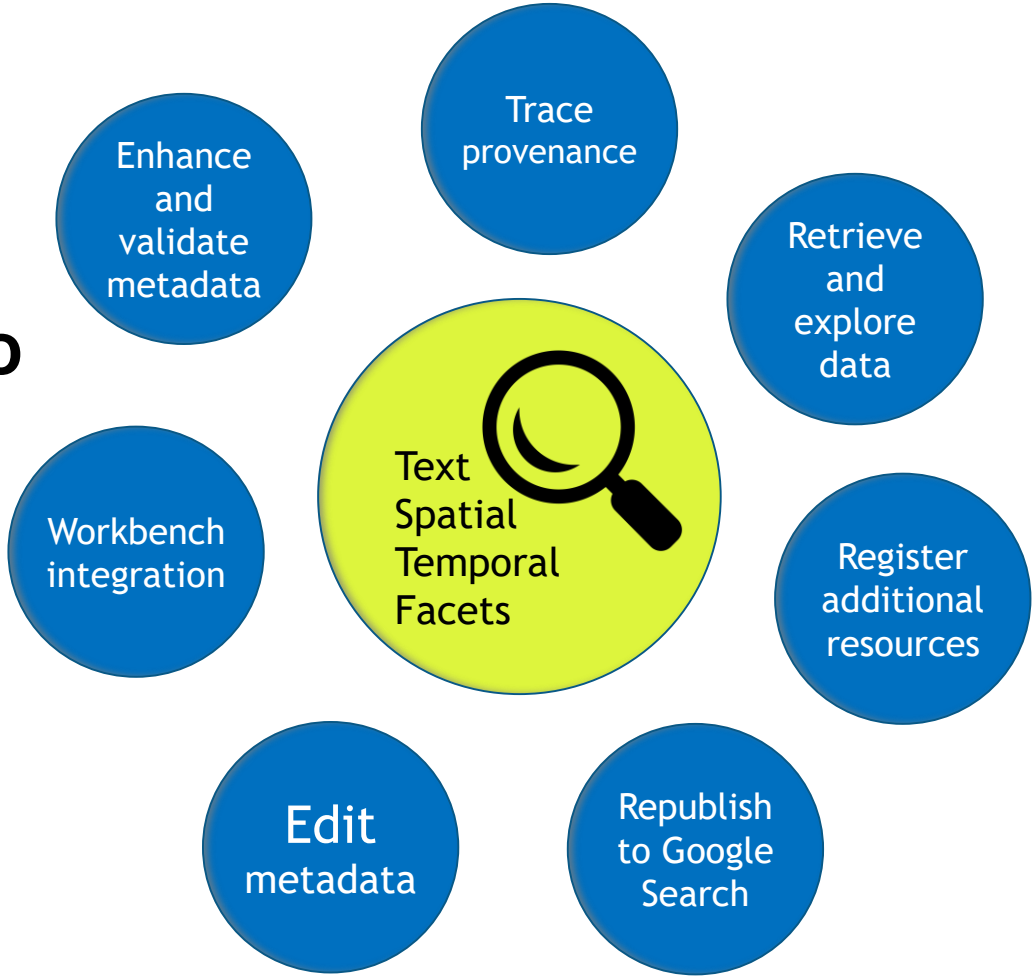
Metadata automatically enhanced through CINERGI

Content Enhancement Components

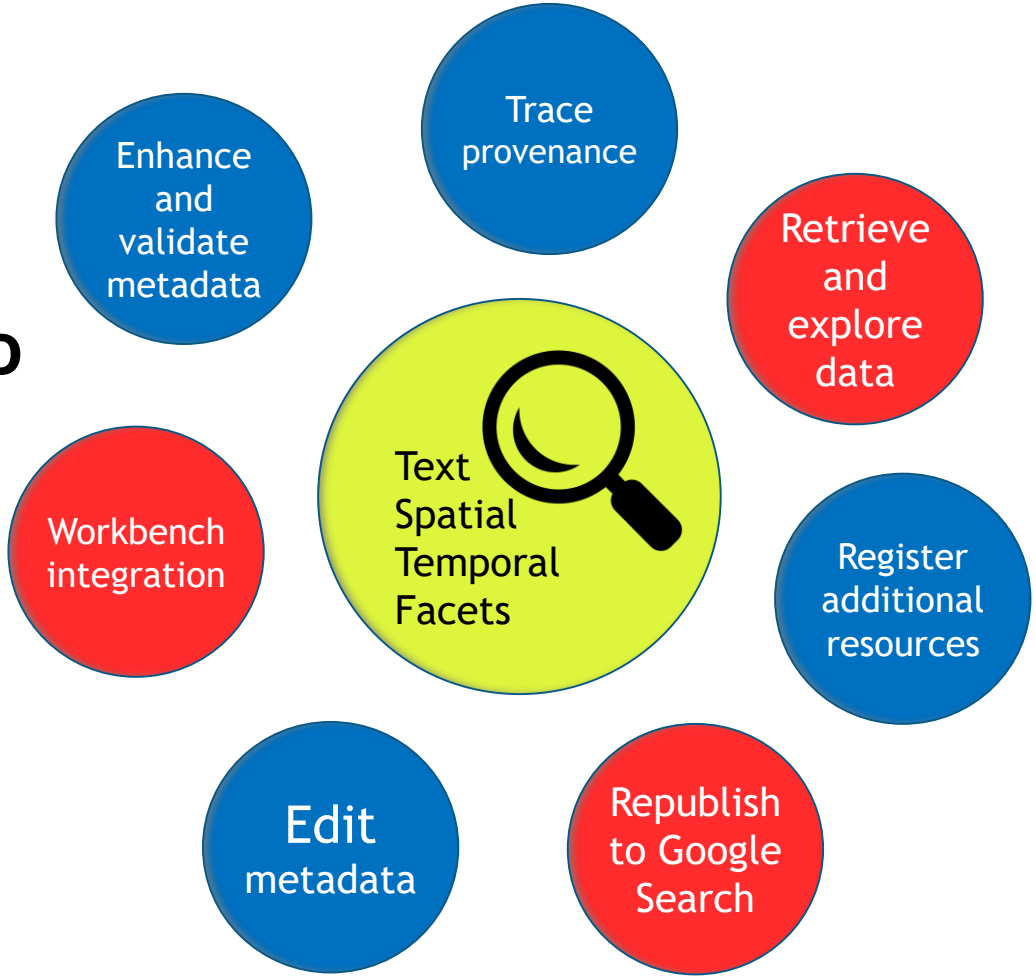
- Common enhancer API
- Provenance recording: W3C PROV and Neo4J
- Spatial enhancer (bounding boxes)
- Keyword enhancer
  - Materials; Processes; Equipment; Methods; Features; Activities; Science Domains; Geologic age; Organizations; Resource types
- Organization Enhancer
  - Associate with Virtual Authority Identifiers
- Collection Enhancer
  - Add keywords to a metadata collection
- Schema validation



Beyond Search:
Data Discovery Studio



Beyond Search:
Data Discovery Studio



From Data Discovery to Research Workflows in EarthCube

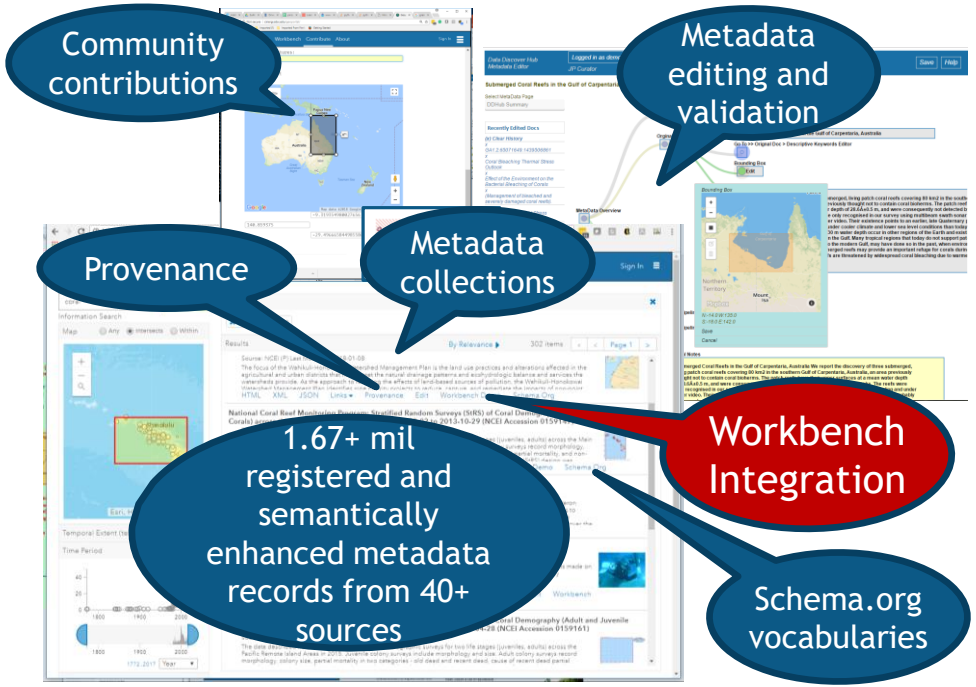
Data Discovery

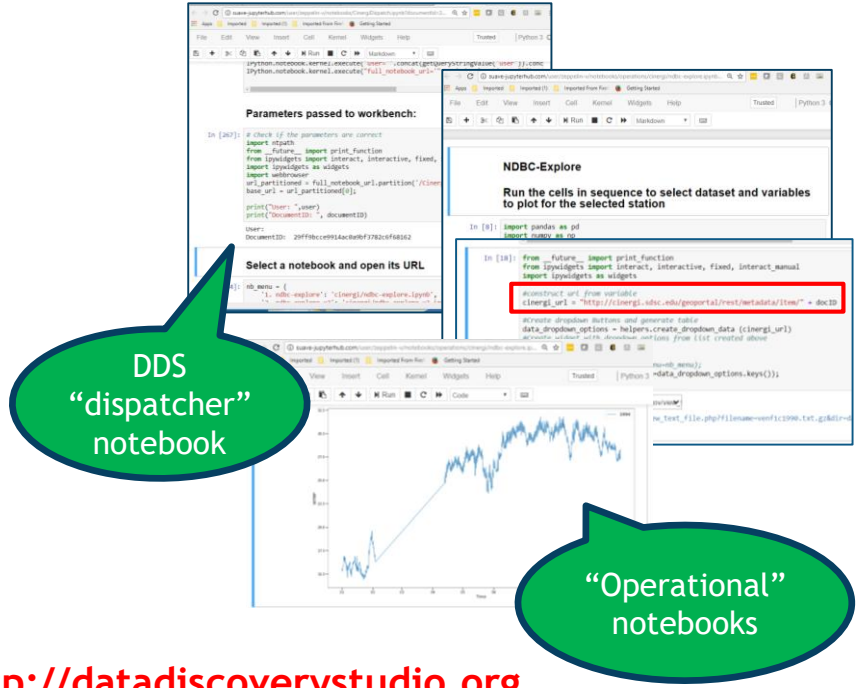
Document IDs →

← *Usage metadata*

Data Retrieval and Analysis

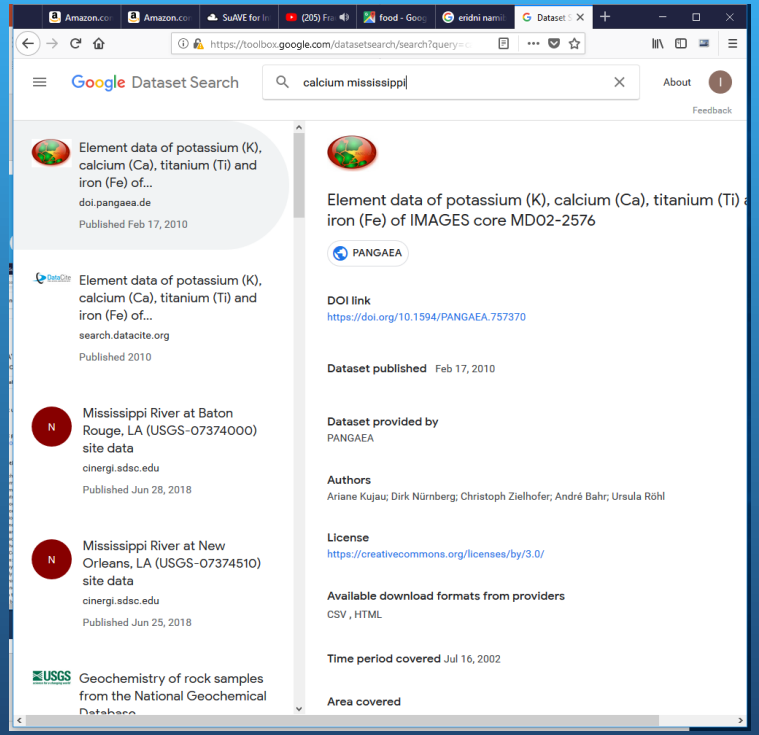
jupyterhub





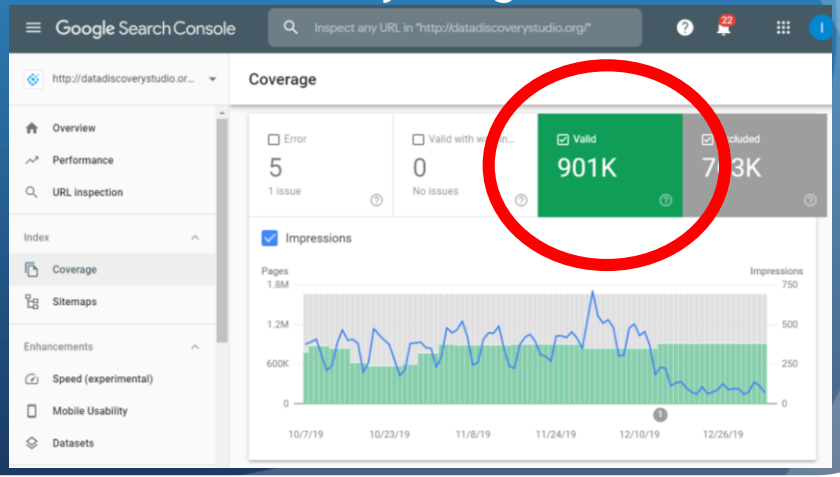
http://datadiscoverystudio.org

Connection with Google Dataset Search

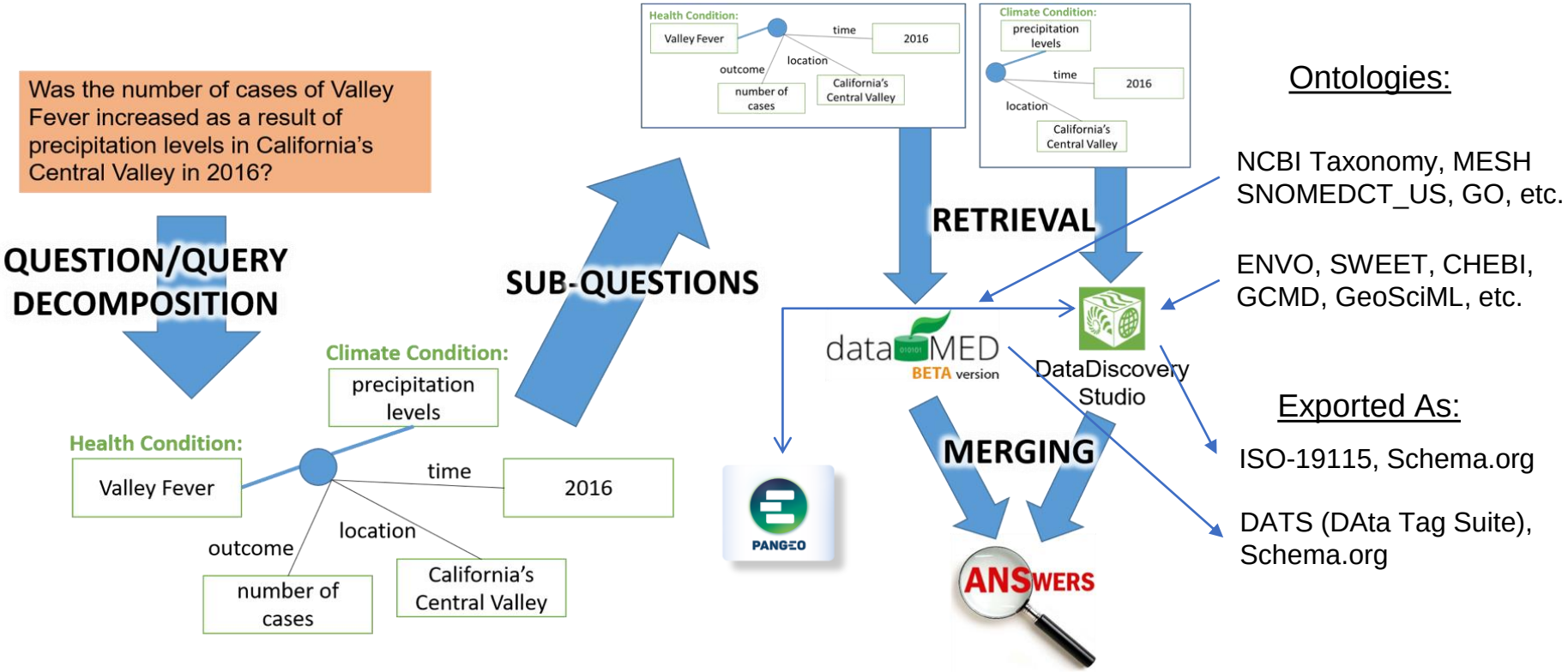


https://toolbox.google.com/datasetsearch

- Schema.org markup is created on the fly for all records
  - If you don't yet publish your datasets in schema.org – you may do so through the Data Discovery Studio
- Resources indexed by Google dataset search



Workflow


Was the number of cases of Valley Fever increased as a result of precipitation levels in California's Central Valley in 2016?
QUESTION/QUERY DECOMPOSITION
SUB-QUESTIONS
Climate Condition:
precipitation levels
Health Condition:
Valley Fever
time
2016
outcome
location
number of cases
California's Central Valley
Health Condition:
Valley Fever
time
2016
outcome
location
number of cases
California's Central Valley
Climate Condition:
precipitation levels
time
2016
location
California's Central Valley
RETRIEVAL
dataMED
BETA version
DataDiscovery Studio
PANGEO
MERGING
ANSWERS
Ontologies:
NCBI Taxonomy, MESH SNOMEDCT_US, GO, etc.
ENVO, SWEET, CHEBI, GCMD, GeoSciML, etc.
Exported As:
ISO-19115, Schema.org
DATS (DAta Tag Suite), Schema.org

Key questions and ongoing hurdles

- **How to build a knowledge network <u>across domains</u>, and keep it updated?**
- **How to build an OKN over large <u>collections of datasets</u>?**
- **How to <u>federate knowledge queries</u> across such collections?**

<u>Ongoing work and hurdles:</u>

- Spatial indexing DataMed records using DDStudio's Spatial Enhancer/NLP
  - Patient locations typically not available (privacy and logistics reasons)
  - Need to resolve mismatches in spatial IDs and geographies for integration
- Co-registering datasets between DDStudio and Pangeo
  - Interoperability between catalogs and search systems, STAC entries for selected DDStudio records
  - Convergence on the use of Jupyter hub infrastructure
- Prototyping federated queries
  - Other ways to link data besides spatially?