

# 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)



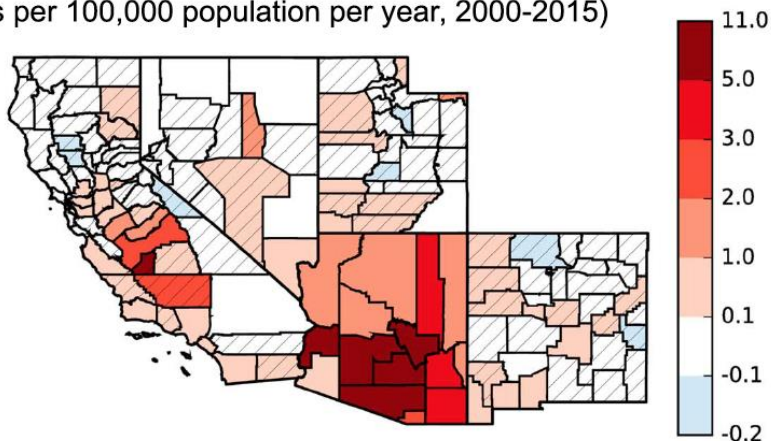
# Valley Fever

## 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)

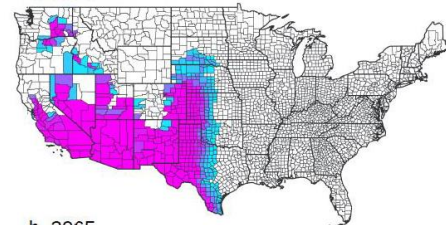


## Climate and environmental drivers

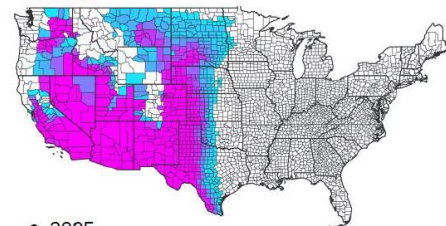


## Predicted spread of Valley Fever

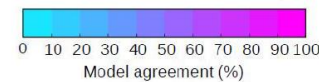
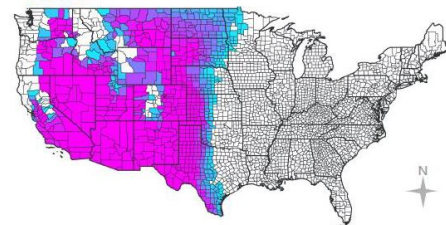
a. 2035



b. 2065



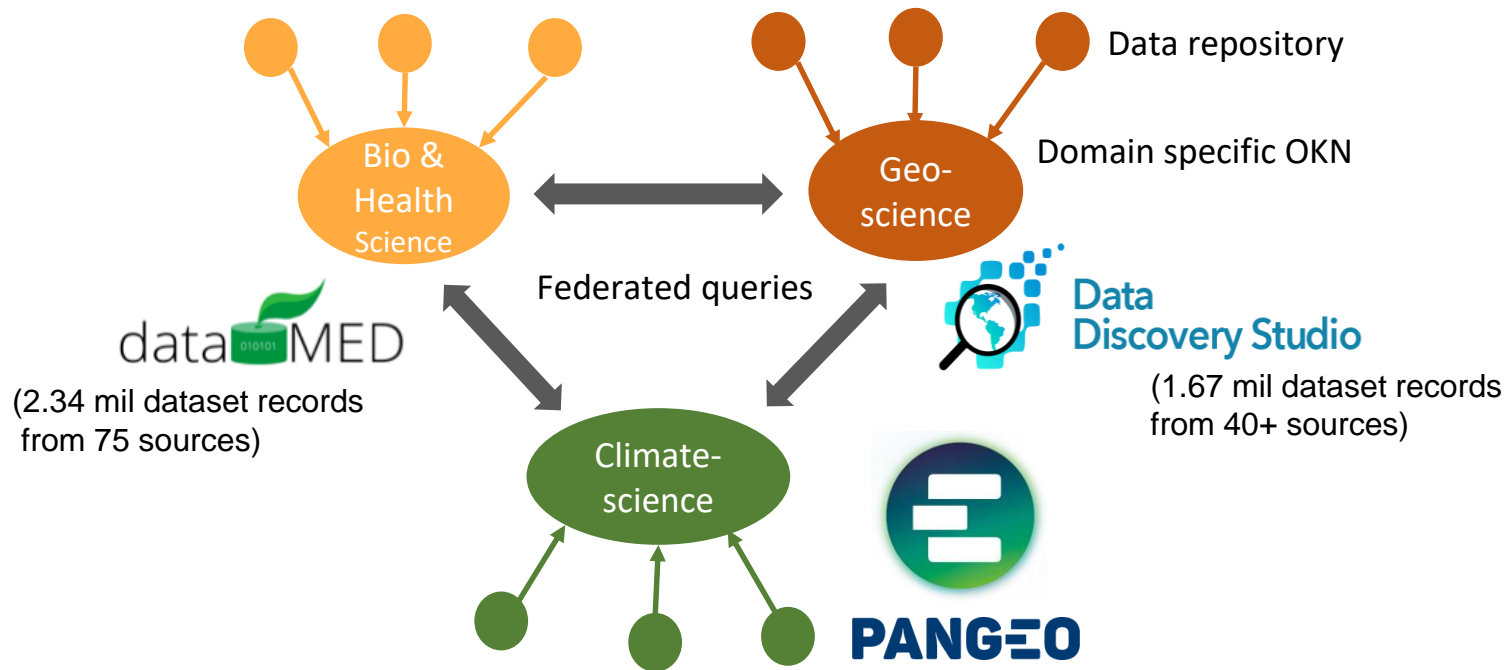
c. 2095



# 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



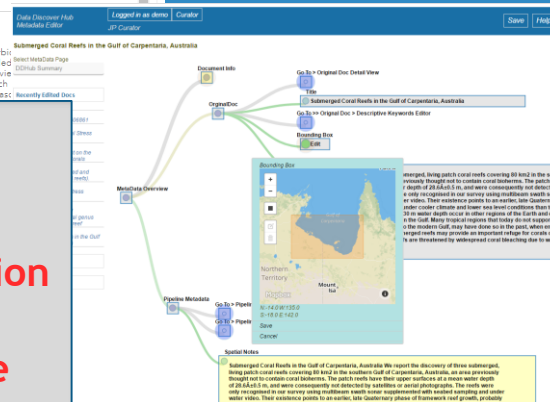
*Use case:* Finding data sets to answer, “Was the number of cases of Valley Fever increased as a result of precipitation levels in California’s Central Valley in 2016?”

# EarthCube Data Discovery Studio

The screenshot shows the EarthCube Data Discovery Studio interface. At the top, there are navigation tabs: Search, Collection, Contribute, About, and Help. Below the navigation bar is a search bar with a 'Search' button. To the left of the search results is a map showing the location of the study area. The search results section displays the title 'T2D-GENES: San Antonio Mexican American Family Studies (SAMAFS)' and a brief description. Below the description, there are links for 'Get Data', 'Item Details', and 'Add to Collection'. The interface also shows a 'Filters' section and a 'Results' section with a count of 1,669,734 items.

1.67 million records

- 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



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

The screenshot displays a web application interface with a search bar at the top left containing the text 'coral'. Below the search bar is a map of the United States with markers for Vancouver, San Francisco, and Los Angeles. To the right of the map is a search results panel. The panel has a 'Filters' section at the top, followed by a 'Results' section showing 1,659,734 items. The results are listed in a table with columns for 'Studio', 'Get Data', 'Item Details', and 'Add to Collection'. The first result is 'T2D-GENES: San Antonio Mexican American Family Studies (SAMAFS)', followed by 'VIVA LA FAMILIA Study: VIVA LA FAMILIA', 'Homo sapiens: Genome-Wide Association Study', and 'National Supported Work Evaluation Study, 1975-1979: Public Use Files'. The interface also includes a 'Browsing Box' on the left side of the results panel, showing a hierarchy of categories and keywords.

coral Search

Information Search

Map Intersects Within

find a place

Search Using MapExtent

Browsing Box

Publication Date

Location Keyword

Original Keyword

All Keywords

Category (4174811)

Activity (683438)

Equipment (627195)

Instrument (471412)

Filters

Results By Relevance 1,659,734 items 10 of >10k Pages

**T2D-GENES: San Antonio Mexican American Family Studies (SAMAFS)**

Publication: 2019-11-12 Source: Individual Contribution Last Modified: 2019-11-12

"description": "The Type 2 Diabetes (T2D) Genetic Exploration by Next-generation sequencing in Ethnic Samples (T2D-GENES) Consortium is a collaborative international effort to identify genes influencing susceptibility to type 2 diabetes in multiple ethnic groups using next generation sequencing. To fulfill this objective, T2D-GENES Consortium undertook two large sequencing studies, called T2D-GENES Projects 1 and 2. Project 1 has carried out whole exome sequencing of 12,940 individuals, 6,504 with T2D and 6,436 non-diabetic controls, equally divided among five continental ancestry groups: Europeans, East Asians, South Asians, Hispanic Americans, and African Americans. The goal of Project 1 is to identify all genetic variants in the complete coding regions of the genomes (i.e. whole exome) by sequencing, including rare variants. Project 2 (i.e. SAMAFS substudy 2) is a pedigree-based study designed to identify low frequency or rare variants influencing susceptibility to T2D, using whole

Studio Get Data Item Details Add to Collection

**VIVA LA FAMILIA Study: VIVA LA FAMILIA**

Publication: 2019-11-12 Source: Individual Contribution Last Modified: 2019-11-12

The VIVA LA FAMILIA Study was designed to identify genetic variants influencing susceptibility to type 2 diabetes in multiple ethnic groups using next generation sequencing. To fulfill this objective, T2D-GENES Consortium undertook two large sequencing studies, called T2D-GENES Projects 1 and 2. Project 1 has carried out whole exome sequencing of 12,940 individuals, 6,504 with T2D and 6,436 non-diabetic controls, equally divided among five continental ancestry groups: Europeans, East Asians, South Asians, Hispanic Americans, and African Americans. The goal of Project 1 is to identify all genetic variants in the complete coding regions of the genomes (i.e. whole exome) by sequencing, including rare variants. Project 2 (i.e. SAMAFS substudy 2) is a pedigree-based study designed to identify low frequency or rare variants influencing susceptibility to T2D, using whole

Studio Get Data Item Details Add to Collection

**Homo sapiens: Genome-Wide Association Study**

Publication: 2019-11-12 Source: Individual Contribution Last Modified: 2019-11-12

"This study funded by the National Cancer Institute aims to identify markers of susceptibility to cancer in the genome associated with common primary carcinomas of the United States (ICD9 codes 140-208.9). Scan data will be made available to the research community through the National Cancer Institute's Genomic Data Commons (GDC) portal.

Studio Get Data Item Details Add to Collection

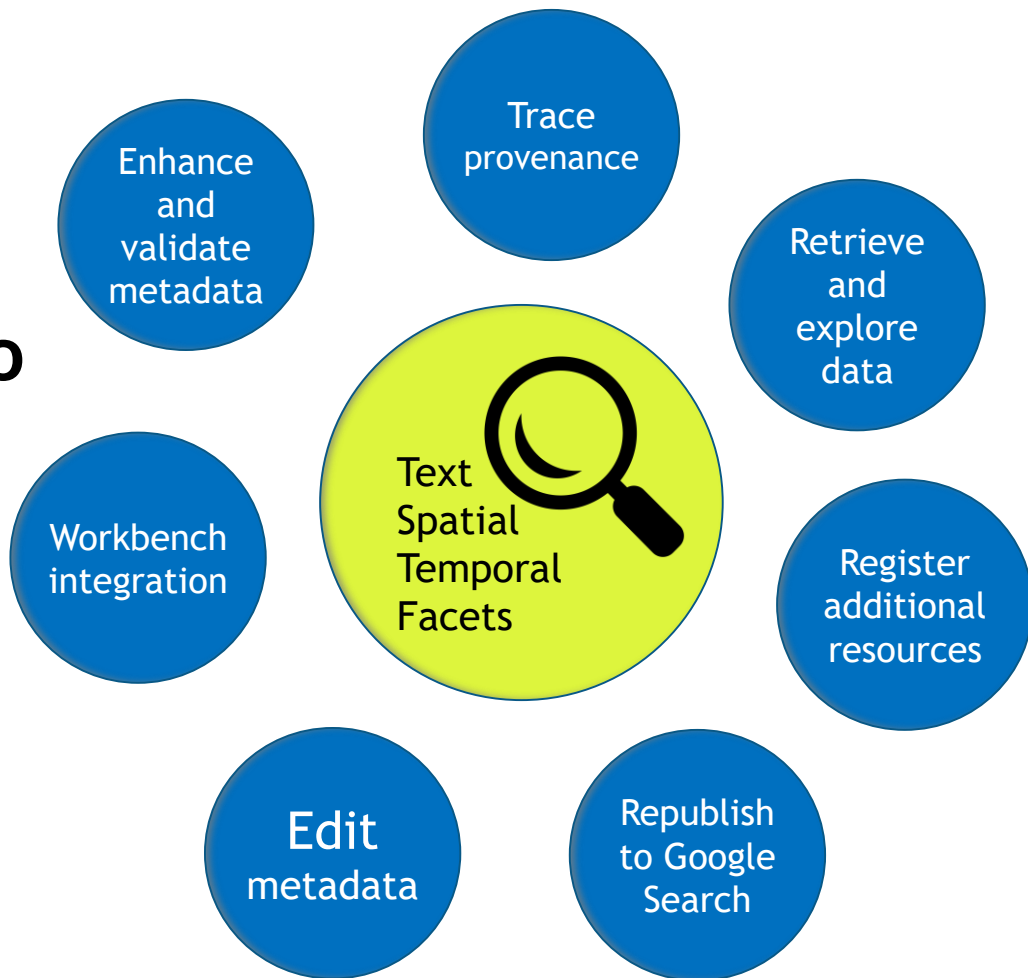
**National Supported Work Evaluation Study, 1975-1979: Public Use Files**

Publication: 2019-11-12 Source: Individual Contribution Last Modified: 2019-11-12

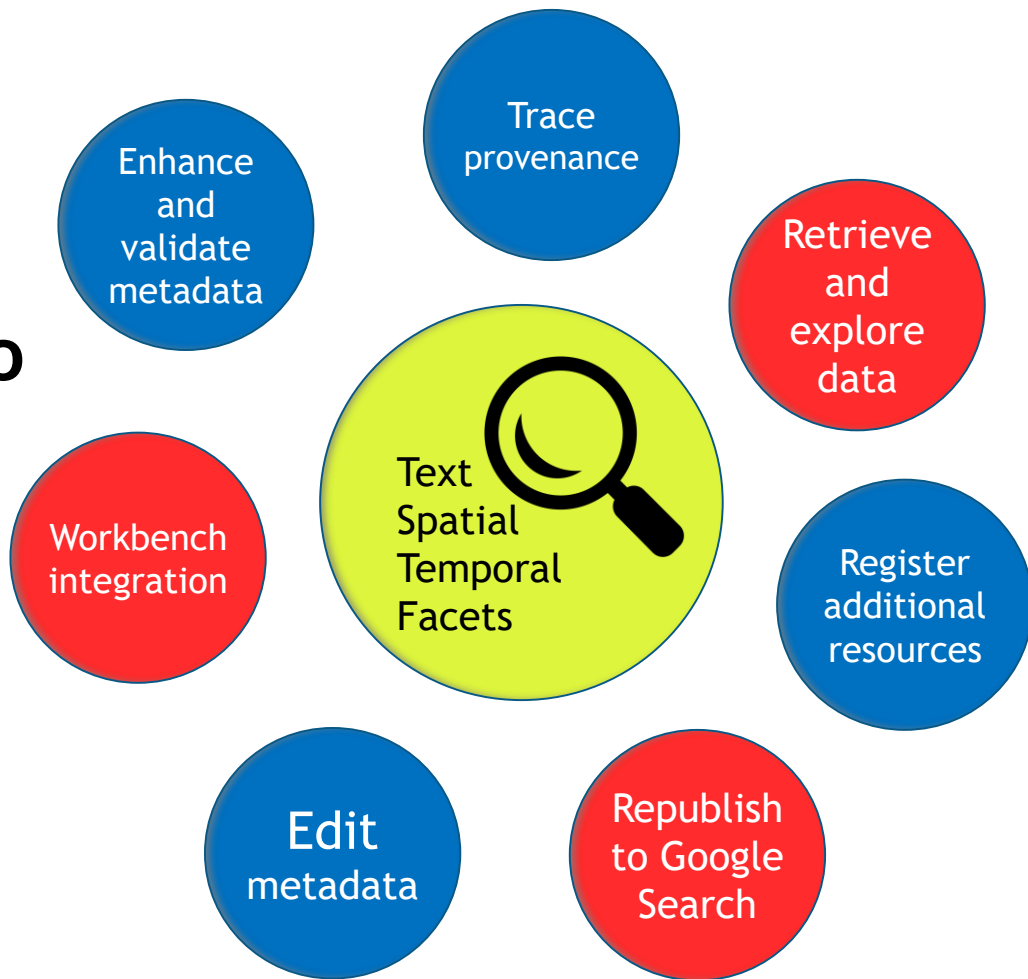
This study is an evaluation of the National Supported Work Demonstration project, a transitional, subsidized work experience program for four target groups of people with longstanding employment problems: ex-offenders, former drug addicts, women who were long-term recipients of welfare benefits, and school dropouts. Many with criminal records. The program provided in

ATOM CSW JSON CSV KML RSS

# 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



Community  
contributions

Metadata  
editing and  
validation

Provenance

Metadata  
collections

1.67+ mil  
registered and  
semantically  
enhanced metadata  
records from 40+  
sources

Workbench  
Integration

Schema.org  
vocabularies

DDS  
“dispatcher”  
notebook

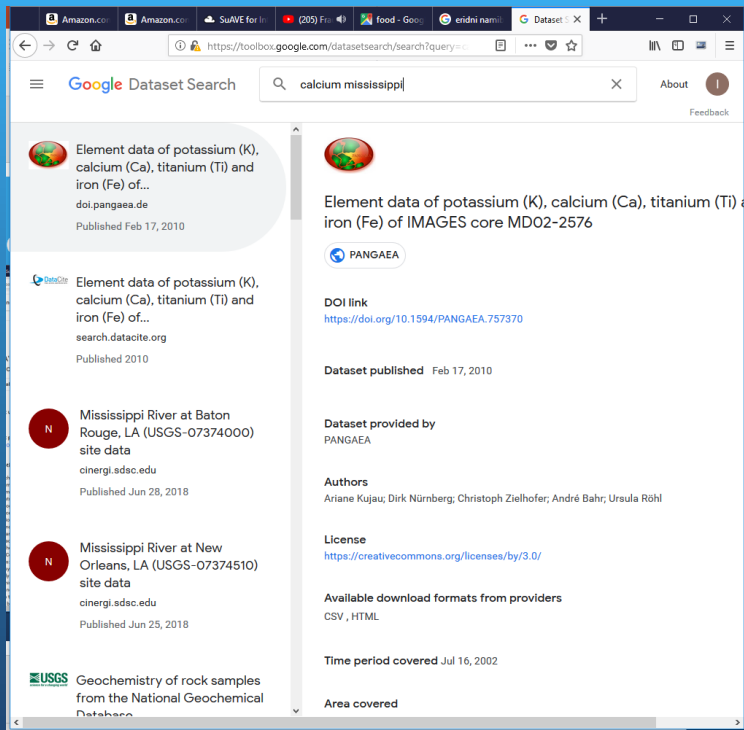
“Operational”  
notebooks

<http://datadiscoverystudio.org>

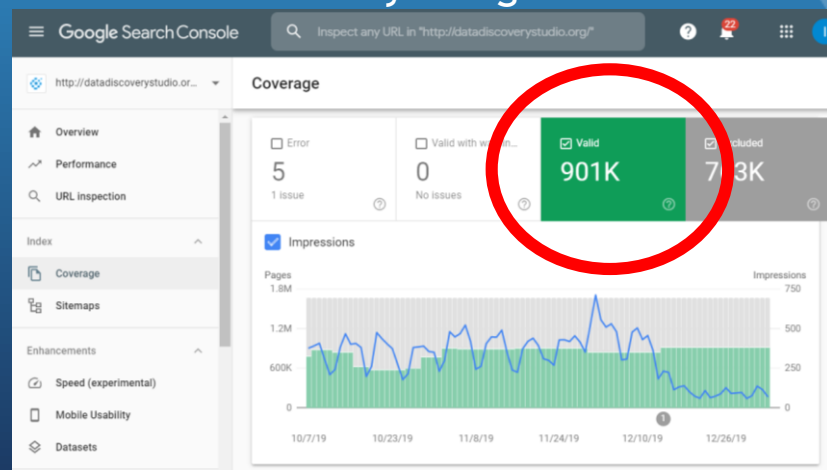


# Connection with Google Dataset Search

- 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

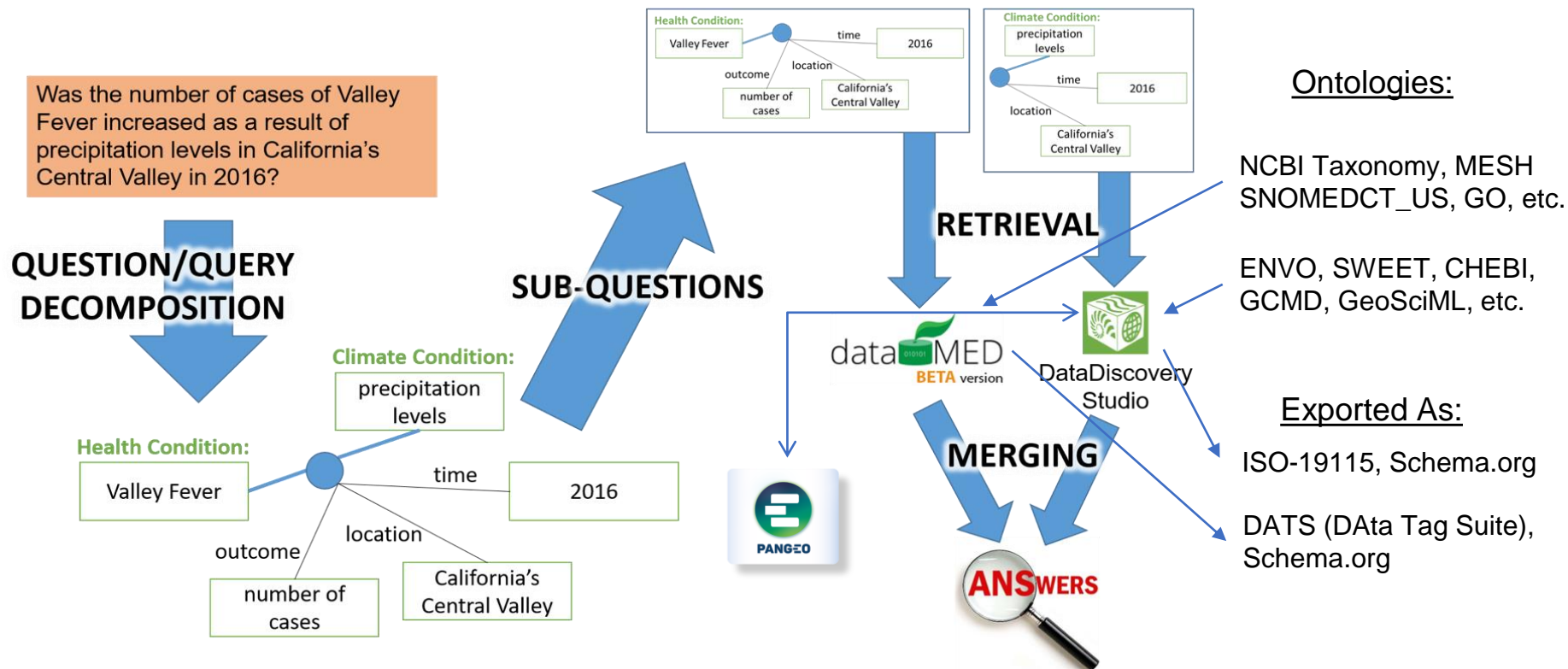


The screenshot shows the Google Dataset Search interface with the search query 'calcium mississippi'. The results list several datasets, including 'Element data of potassium (K), calcium (Ca), titanium (Ti) and iron (Fe) of...' from pangaea.de and 'Mississippi River at Baton Rouge, LA (USGS-07374000) site data' from cinergi.adsc.edu. The interface includes a search bar, navigation icons, and a list of dataset cards with details like DOI links and publication dates.



<https://toolbox.google.com/datasetsearch>

# Workflow



# Key questions and ongoing hurdles

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

## Ongoing work and hurdles:

- 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?