



# Designing for Serendipity: Research Data Curation in Topic Spaces

Sara Lafia  
Ph.D. Candidate in Geography

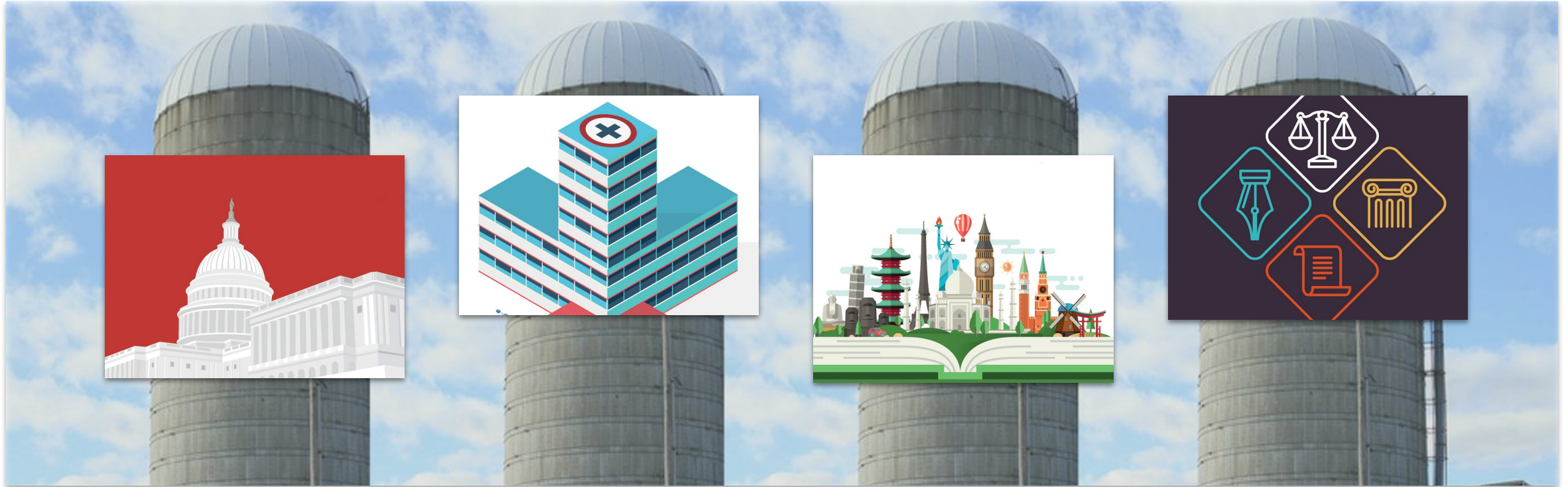
Committee: Werner Kuhn (chair),  
James Frew, Kelly Caylor, Daniel Montello

July 13, 2020

**spatial@ucsb**  
CENTER FOR SPATIAL STUDIES

**UC SANTA BARBARA**

# How can researchers find **related data** without needing to know disciplinary terms?



Political Science

Health

Urban Planning

Law



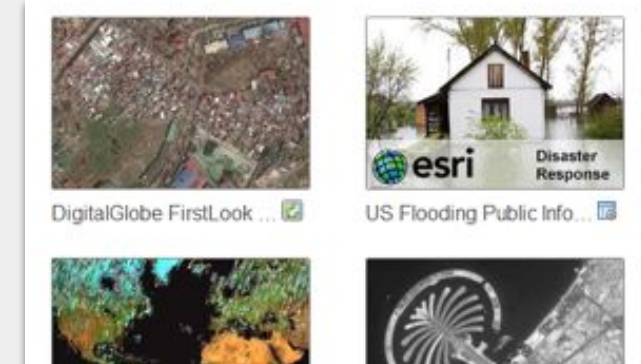
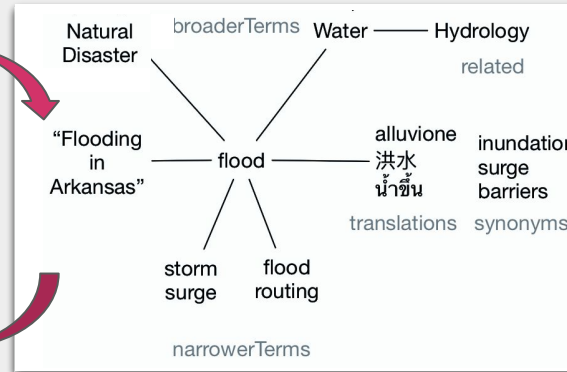
# Bookshelves curate books by **topics**, supporting search and discovery.



- **Research data:** *documents* (Buckland, 1997) and *metadata* (Mayernik, 2016) used or generated by researchers
- **Curation:** *organization* of data to maximize meaningful access (Fear, 2013) and to *support* bibliographic objectives (Svenonius, 2000)

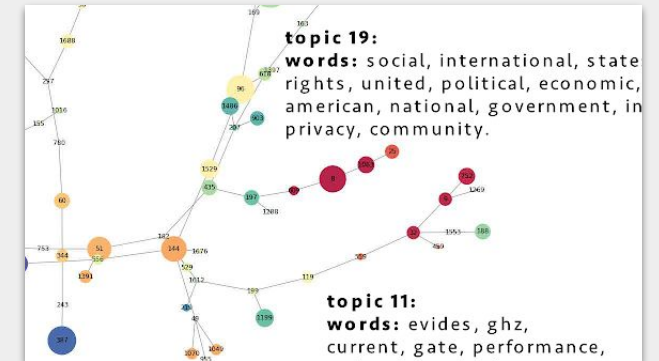
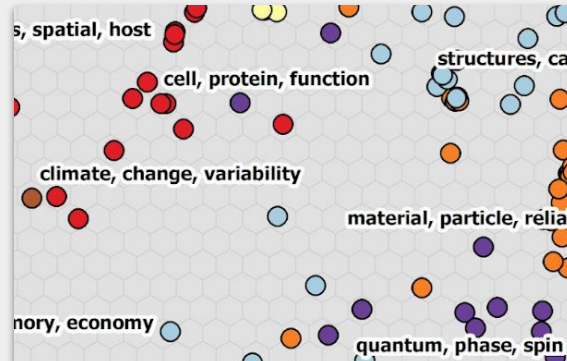
## Verbalization (Study 1)

How can we map topics of interest, expressed in **users' terms**, onto the language of metadata?



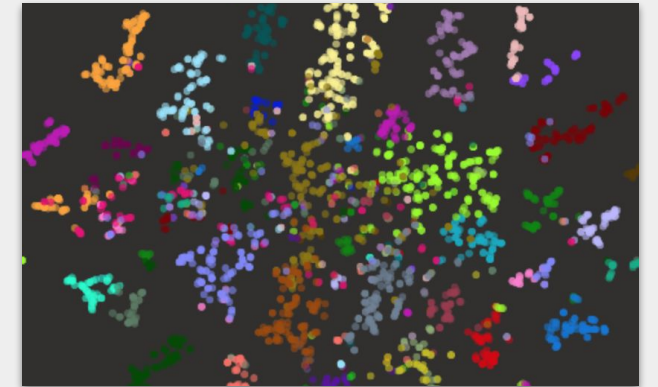
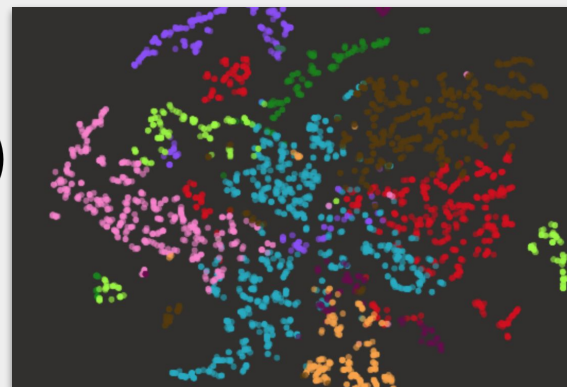
## Spatialization (Study 2)

How can we elicit and spatially represent the topics of research data to convey their **similarity**?



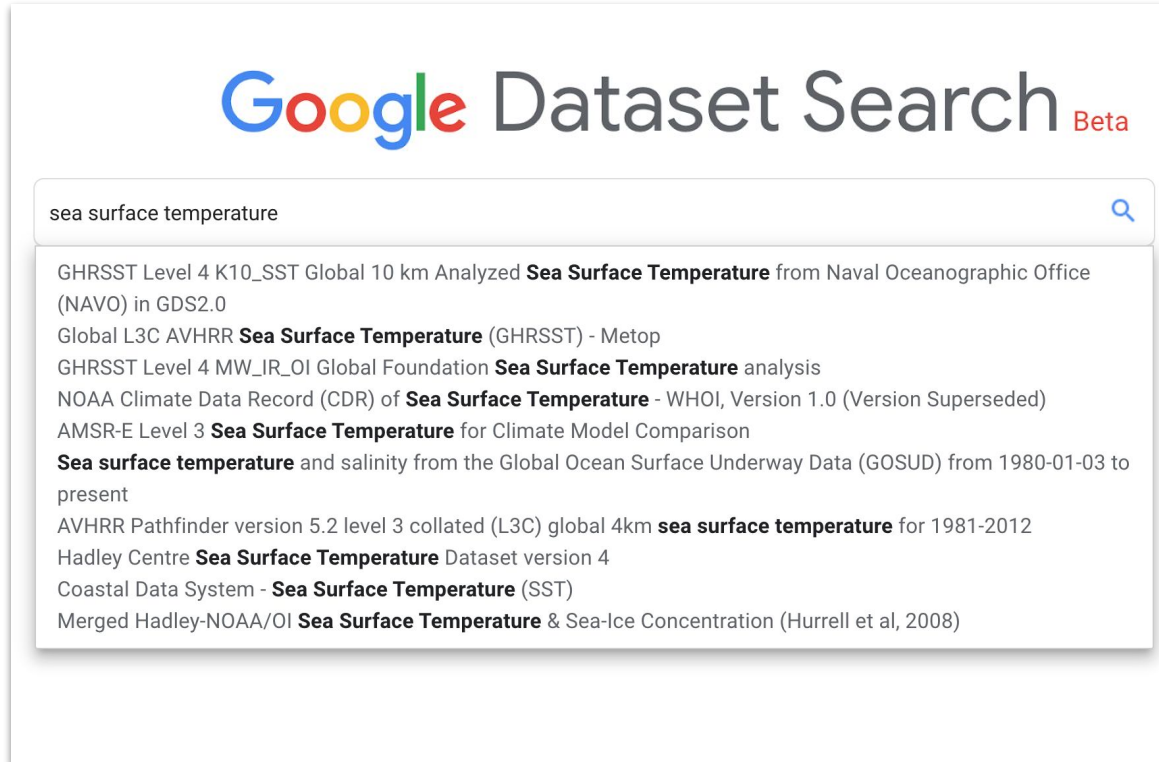
## Generalization (Study 3)

How can we represent the topics of a multidisciplinary body of research at multiple **levels of detail**?





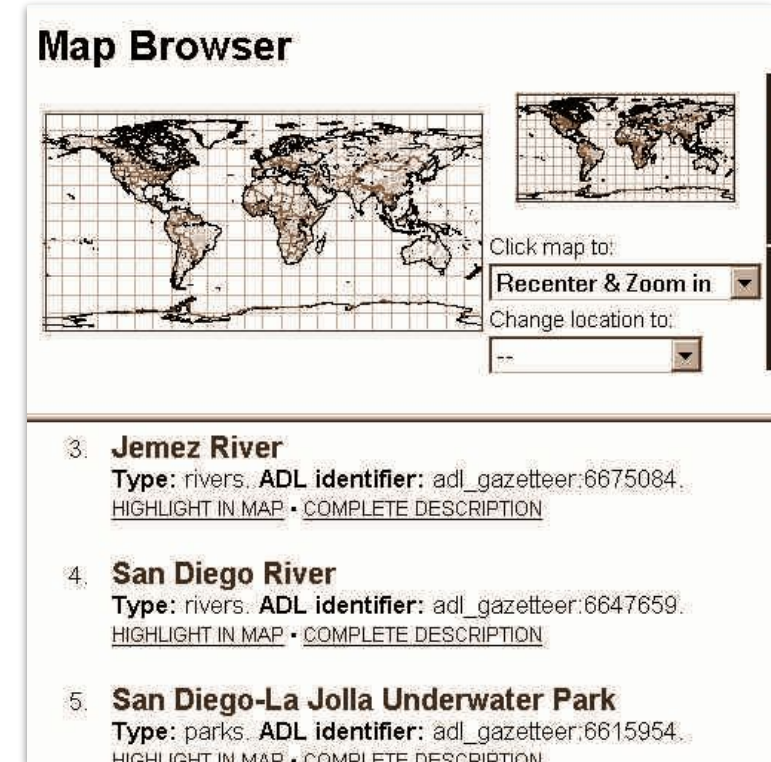
# Seeking research data



The screenshot shows the Google Dataset Search interface. At the top, the text "Google Dataset Search Beta" is displayed. Below it, a search bar contains the text "sea surface temperature". A list of search results is shown below the search bar, including:

- GHR SST Level 4 K10\_SST Global 10 km Analyzed **Sea Surface Temperature** from Naval Oceanographic Office (NAVO) in GDS2.0
- Global L3C AVHRR **Sea Surface Temperature** (GHR SST) - Metop
- GHR SST Level 4 MW\_IR\_OI Global Foundation **Sea Surface Temperature** analysis
- NOAA Climate Data Record (CDR) of **Sea Surface Temperature** - WHOI, Version 1.0 (Version Superseded)
- AMSR-E Level 3 **Sea Surface Temperature** for Climate Model Comparison
- Sea surface temperature** and salinity from the Global Ocean Surface Underway Data (GOSUD) from 1980-01-03 to present
- AVHRR Pathfinder version 5.2 level 3 collated (L3C) global 4km **sea surface temperature** for 1981-2012
- Hadley Centre **Sea Surface Temperature** Dataset version 4
- Coastal Data System - **Sea Surface Temperature** (SST)
- Merged Hadley-NOAA/OI **Sea Surface Temperature** & Sea-Ice Concentration (Hurrell et al, 2008)

**Information lookup** with keywords  
(Hearst, 2011; Ithaca S + R Faculty Survey, 2016)

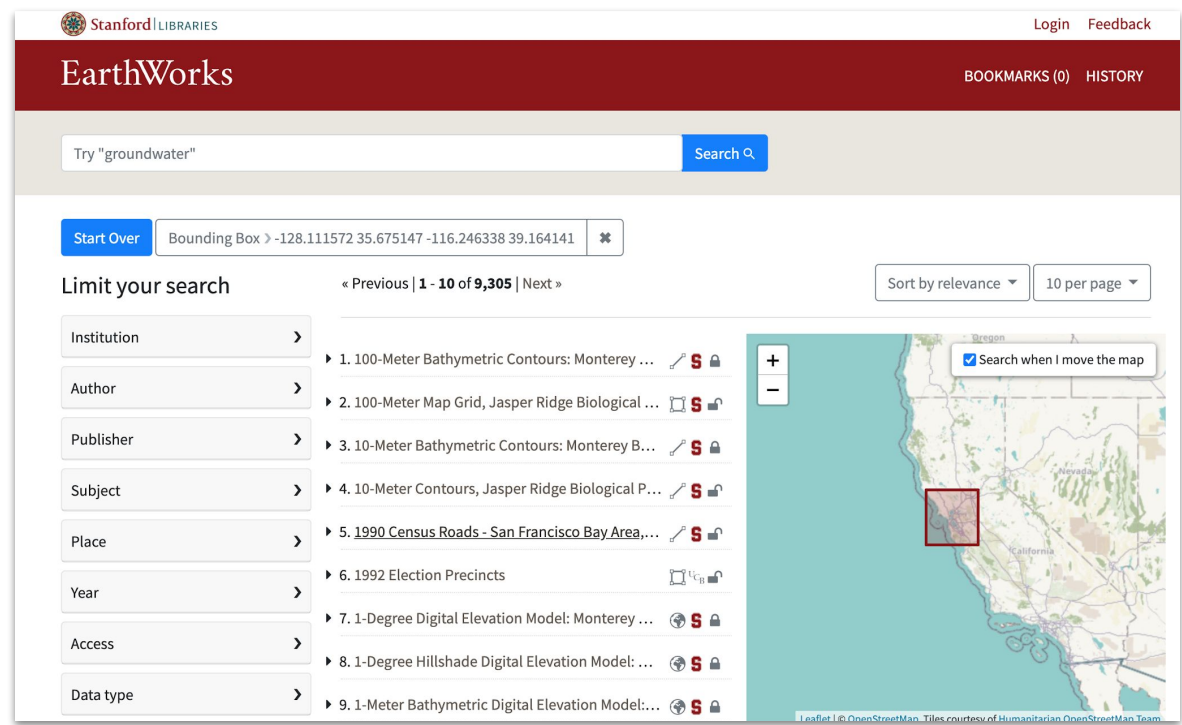


The screenshot shows a "Map Browser" interface. At the top, the text "Map Browser" is displayed. Below it, there is a world map with a grid overlay. To the right of the map, there are two smaller maps showing different views of the world. Below the maps, there are two input fields: "Click map to:" and "Change location to:". Below these fields, there are two buttons: "Recenter & Zoom in" and "Change location to:". Below the buttons, there is a list of search results:

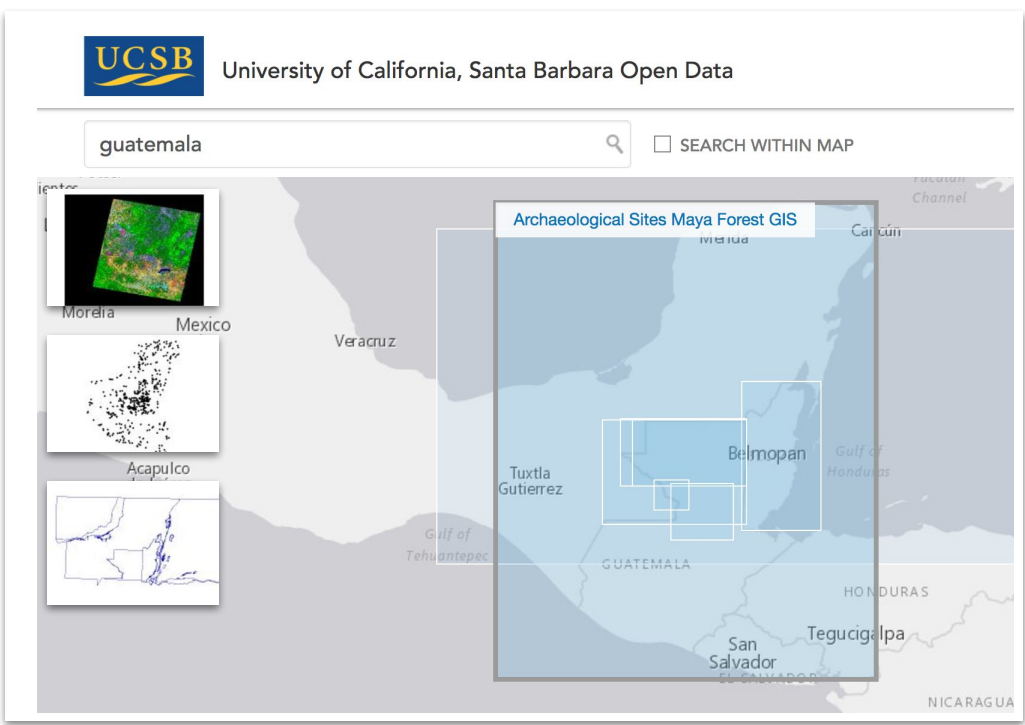
- 3. **Jemez River**  
Type: rivers. ADL identifier: adl\_gazetteer:6675084.  
[HIGHLIGHT IN MAP](#) • [COMPLETE DESCRIPTION](#)
- 4. **San Diego River**  
Type: rivers. ADL identifier: adl\_gazetteer:6647659.  
[HIGHLIGHT IN MAP](#) • [COMPLETE DESCRIPTION](#)
- 5. **San Diego-La Jolla Underwater Park**  
Type: parks. ADL identifier: adl\_gazetteer:6615954.  
[HIGHLIGHT IN MAP](#) • [COMPLETE DESCRIPTION](#)

**Exploratory search** in a geographic map  
(Smith and Frew, 1995)

# Organizing research data



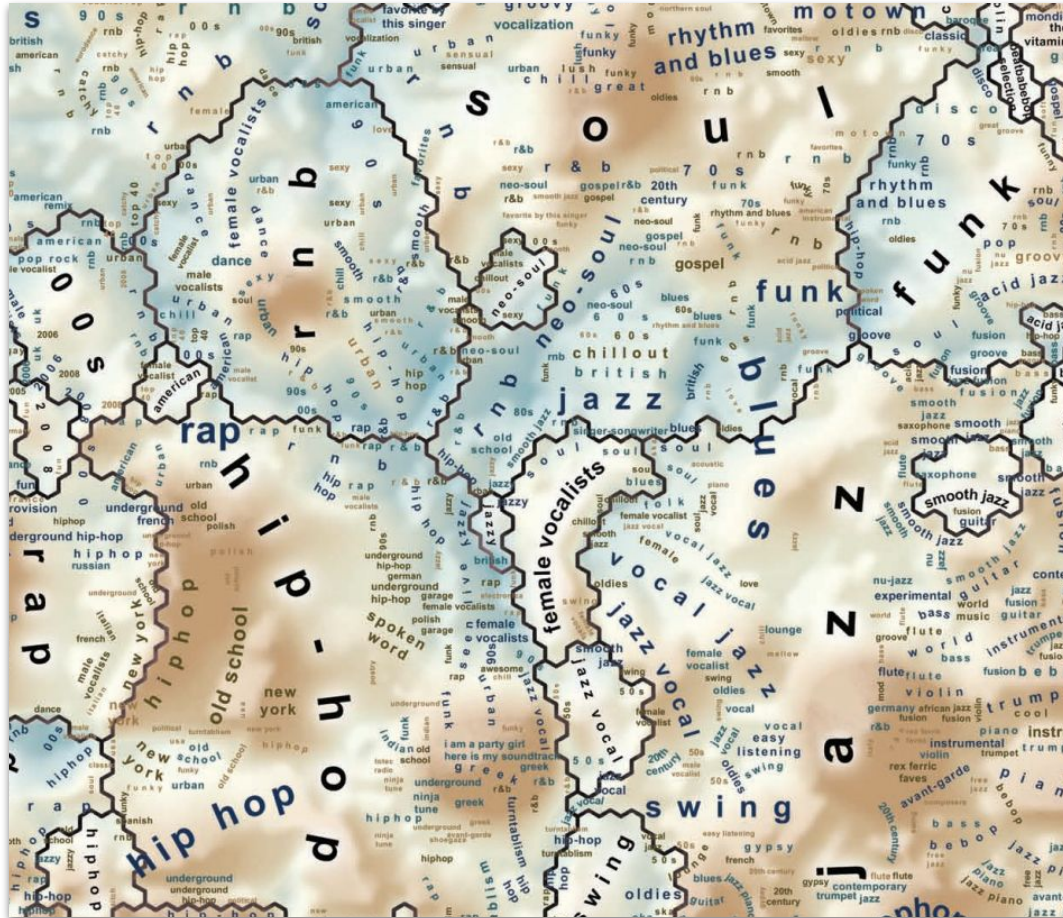
**Geographic** data: location, time, theme  
(Sinton, 1978; Durante and Hardy, 2015)



Multidisciplinary **research data**  
(Lafia et al., 2016)



# From geographic organization to **topic spaces**



## Landscape of musical genres from Last.fm (Biberstine et al., 2010)

- **Spatialization:** mapping physical space to abstract domains through spatial *metaphors* (Kuhn, 1996)
- **Distance-similarity metaphor:** *nearby* data items are semantically *similar* (Montello et al., 2003)

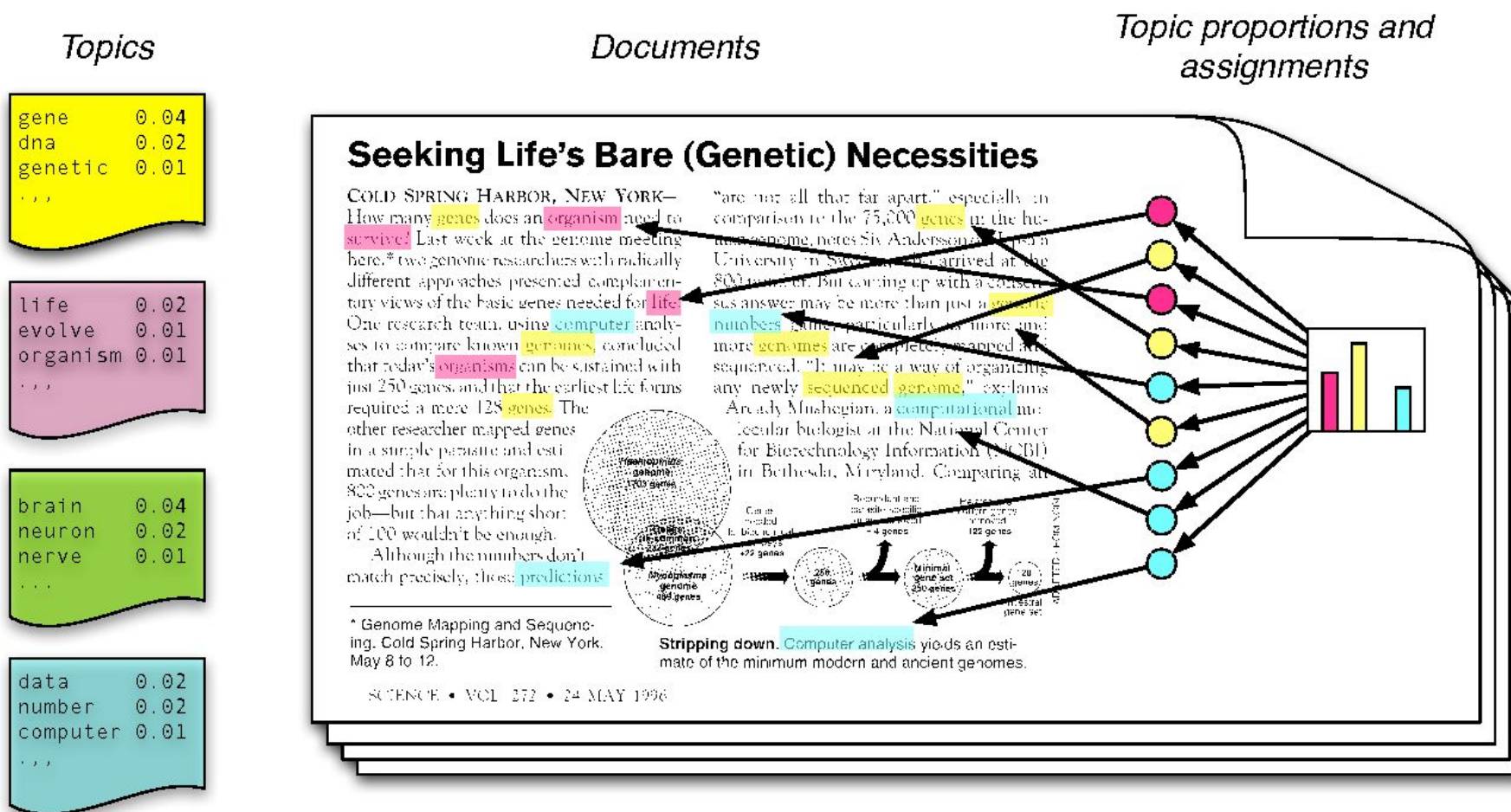
# Making topics explicit: **semantic annotation**



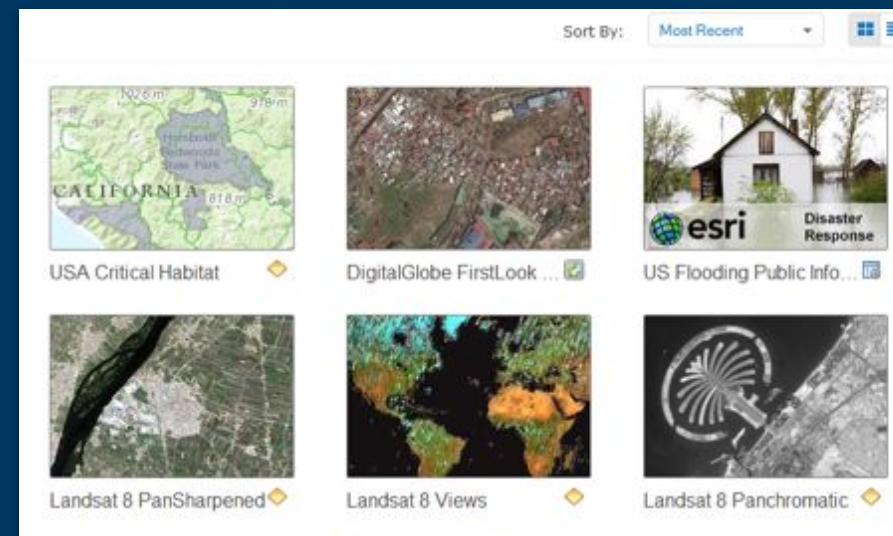
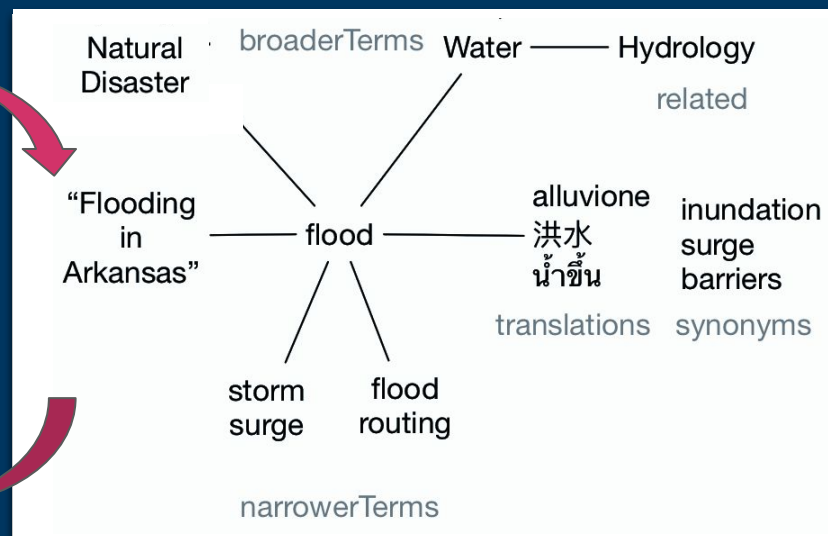
**Hierarchical** classification versus other spatial conceptualizations  
(Gärdenfors, 2000)



# Making topics explicit: **topic modeling**



Documents as **mixtures** of topics (Latent Dirichlet Allocation)  
(Blei, 2012)



*Lafia et al. (2018)*

# Verbalization

## Improving Discovery of Open Civic Data (Study 1)



# Open data initiatives allow public data **access** but do not guarantee **discoverability**.

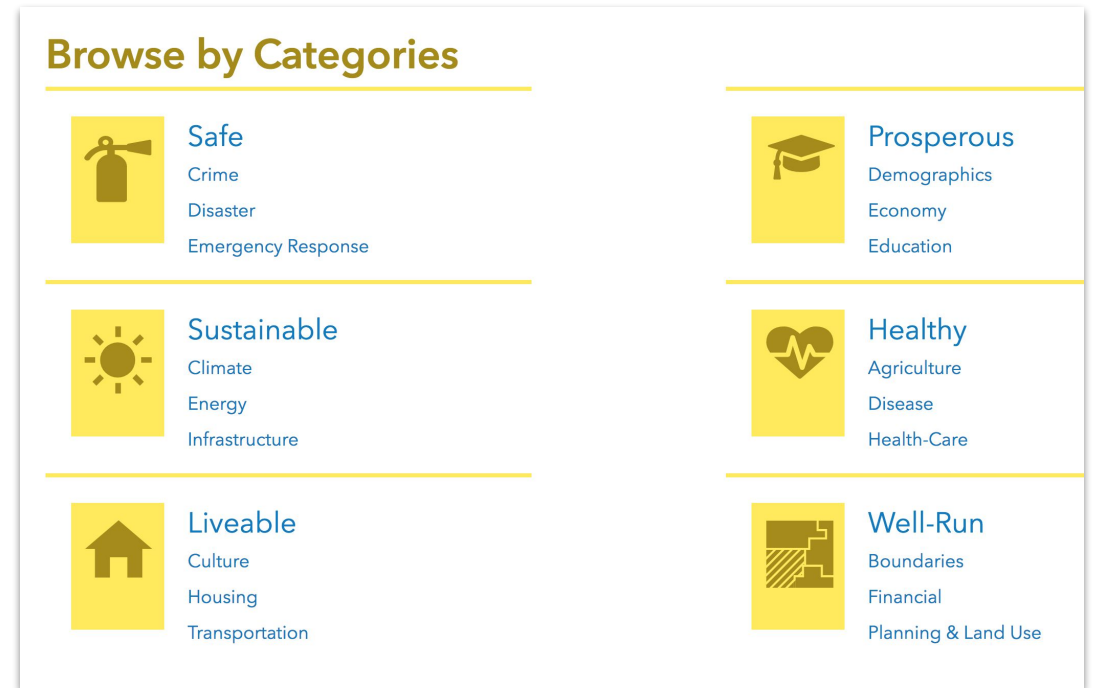
The screenshot shows the Michigan's Open Data Portal. The header is blue with the 'data.MICHIGAN.gov' logo on the left and navigation links (Michigan.gov, Open Michigan, Developers, Resources, Sign In) on the right. The main banner features a photo of the Michigan State Capitol dome with the text 'Michigan's Open Data Portal' and a description: 'View and analyze data provided by a variety of Michigan state agencies in one easy to view catalog. Create charts and graphs, filter and summarize data, and develop maps with the open data.' Below the banner, a featured dataset is titled 'MDOT Fatalities and Serious Injuries MI Public Roads' with a sub-label 'Infrastructure'. The description states: 'Annual and rolling averages of traffic crashes, fatalities and serious injuries on Michigan public roads, as well as data-driven predictions for each category.' It includes a 'Less' link, a 'Tags' section with 'non-motorized, crashes, trans, mdot, bicycle, and 7 more', and an 'Updated' date of 'October 1, 2019' with 'Views' of '50'. An 'API Docs' link is at the bottom right.

The screenshot shows the 'Crashes in DC' dataset page on the Open Data DC portal. The header is dark blue with the 'Open Data DC' logo and navigation links (App Gallery, Data Stories, Developer Starter Kit, Data Policy, Feedback, Handbook). The main banner features a map of DC with the text 'Open Data DC'. Below the banner, the dataset is titled 'Crashes in DC' with the source 'City of Washington, DC | DCGISopendata'. The description reads: 'Crashes on the street segment line network of the District of Columbia maintained by the District Department of Transportation (DDOT). In addition to locations, a related table consisting of crash...'. The dataset details include 'Type: Feature Layer', 'Last Updated: December 10, 2019', 'Rows: 232,975', and 'Tags: accidents, bike, crash, Crashes, fatality, injuries, pede...'. The 'Data' icon is visible in the top left of the dataset card.

Source: <https://hub.arcgis.com/>

# How can we map topics of interest, expressed in users' terms, onto the language of metadata?

1. **Select a base vocabulary of geospatial categories**
2. Extend vocabulary with concept hierarchies
3. Tag metadata with terms from concept hierarchies
4. Evaluate portal implementations



**Vocabulary problem** in human system communication (Furnas et al., 1987)



# How can we map topics of interest, expressed in users' terms, onto the language of metadata?

1. Select a base vocabulary of geospatial categories
- 2. Extend vocabulary with concept hierarchies**
3. Tag metadata with terms from concept hierarchies
4. Evaluate portal implementations



ASK A LIBRARIAN

The Library of Congress > [Linked Data Service](#)

## Public safety

### URI(s)

- > <http://id.loc.gov/authorities/subjects/sh2008002399>
- > <info:lc/authorities/sh2008002399>
- > <http://id.loc.gov/authorities/sh2008002399#concept>
- > [Safety, Public](#)

### Broader Terms

- > [Human services](#)

### Narrower Terms

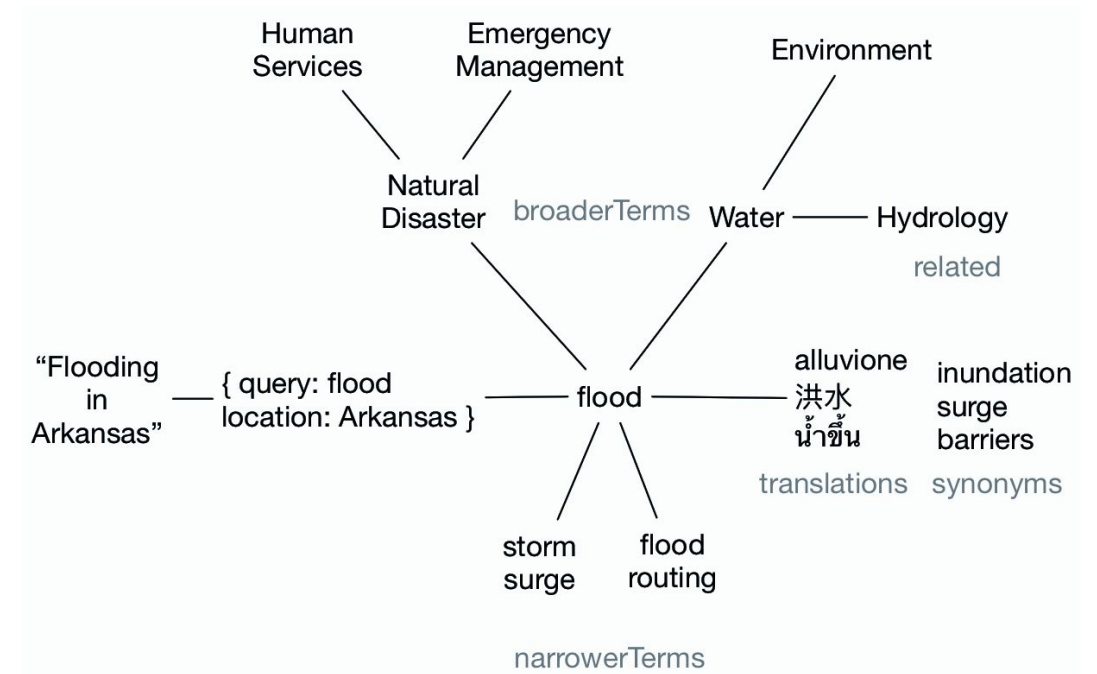
- > [Civil defense](#)
- > [Crime prevention](#)
- > [Emergency management](#)
- > [Fire extinction](#)
- > [Fire prevention](#)
- > [Offenses against public safety](#)
- > [Police](#)
- > [Traffic safety](#)

### Exact Matching Concepts from Other Schemes

- > [public safety](#)

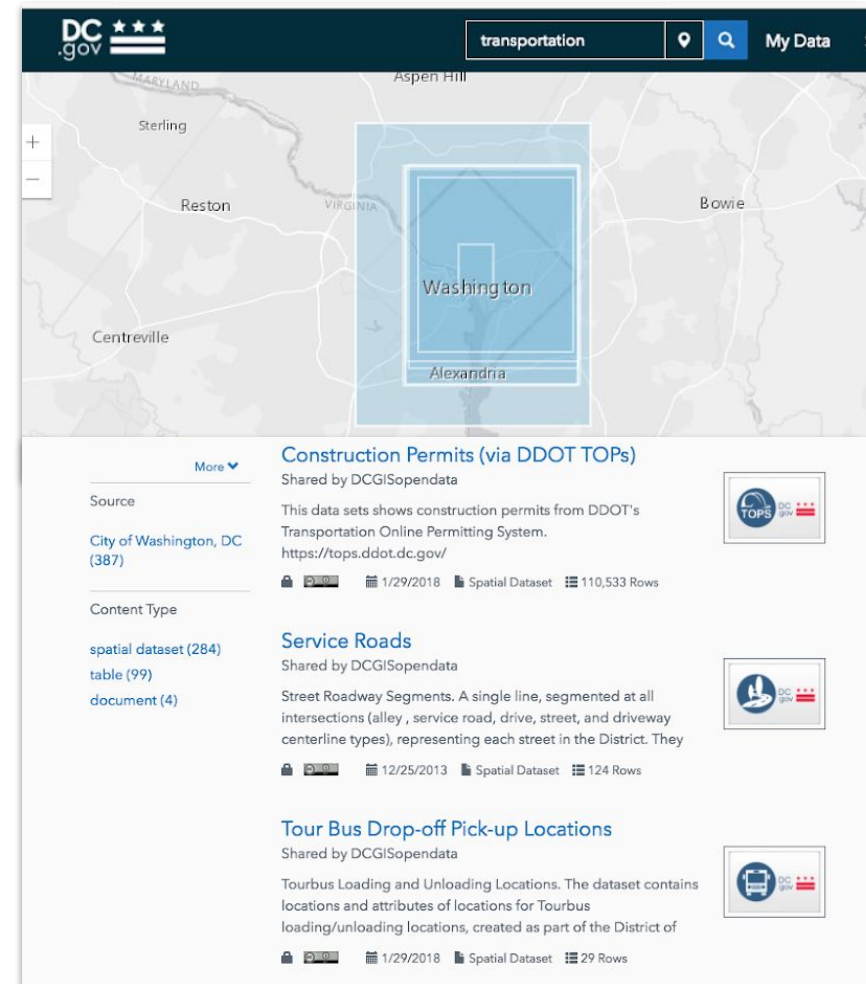
# How can we map topics of interest, expressed in users' terms, onto the language of metadata?

1. Select a base vocabulary of geospatial categories
2. Extend vocabulary with concept hierarchies
- 3. Tag metadata with terms from concept hierarchies**
4. Evaluate portal implementations



# How can we map topics of interest, expressed in users' terms, onto the language of metadata?

1. Select a base vocabulary of geospatial categories
2. Extend vocabulary with concept hierarchies
3. Tag metadata with terms from concept hierarchies
4. **Evaluate portal implementations**





# Curation protocol for semantic annotation

Q

bicycle accident

✕

MDOT Fatalities and Serious Injuries MI Public Roads

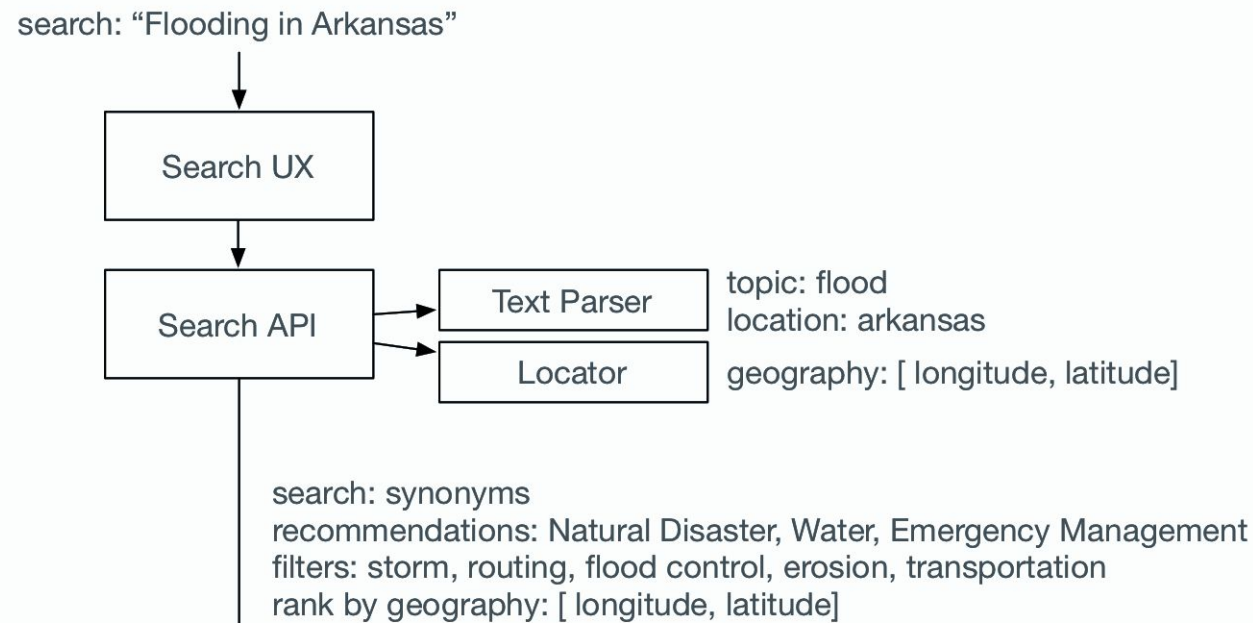
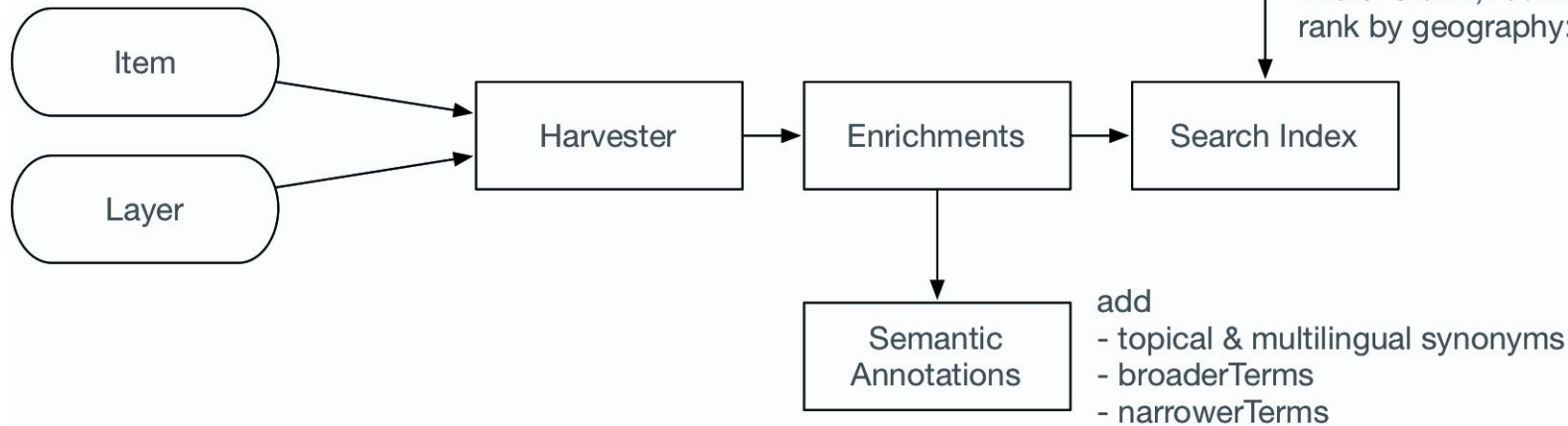
Infrastructure

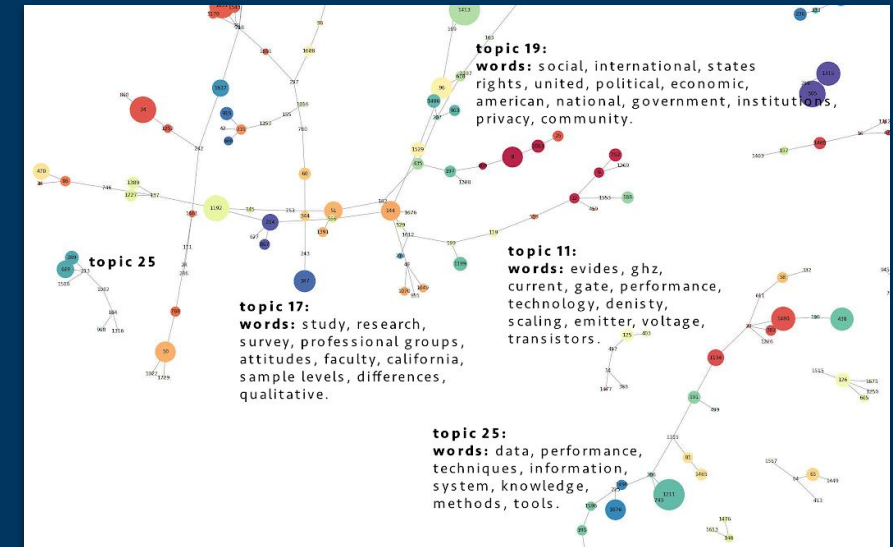
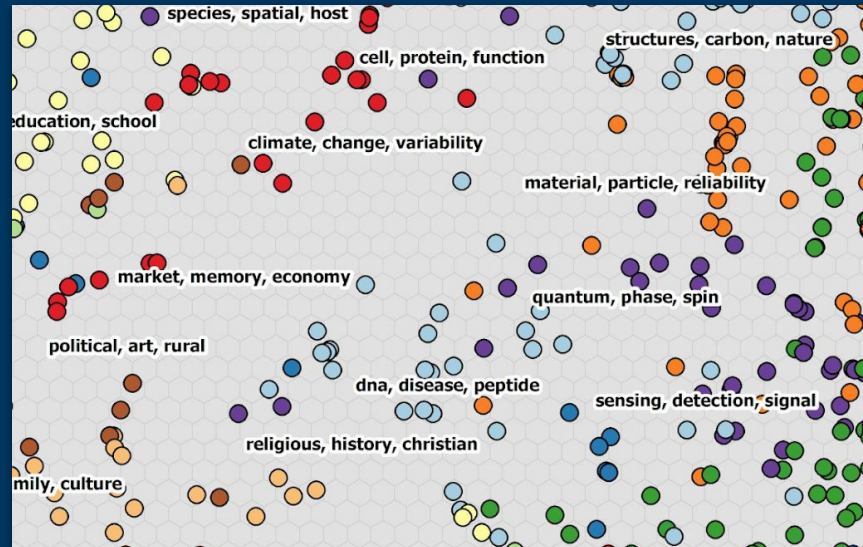
Annual and rolling averages of traffic crashes, fatalities and serious injuries on Michigan public roads, as well as data-driven predictions for each category.

Crashes in DC

City of Washington, DC | DCGISopendata

Crashes on the street segment line network of the District of Columbia maintained by the District Department of Transportation (DDOT). In addition to locations, a related table consisting of crash...





Lafia et al. (2019)

*Verbalization* → *Spatialization*

Enabling the Discovery of Thematically Related Research Objects with Systematic Spatializations (Study 2)

# Related academic research is often described with **different** terms across disciplines.

ADRL

Alexandria Digital Research Library

About • FAQ • Usage Guidelines •

All Fields

Search...

Search

Limit your search

Format

Collection

Contributor

Genre

Date

Academic Department

Rights

UCSB electronic theses and dissertations

In partnership with the Graduate Division, the UC Santa Barbara Library is making available theses and dissertations produced by UCSB students. Currently included in ADRL are theses and dissertations that were originally filed electronically, starting in 2011. In future phases of ADRL, all theses and dissertations created by UCSB students may be digitized and made available.

Genres: Dissertations, Academic

Format: Text, Collection

ARK: ark:/48907/f3348hkh

Local Identifier: etds

Items in this Collection

« Previous | 1 - 10 of 2,277 | Next »

Search Collection

Go

Source: <https://www.alexandria.ucsb.edu/collections>

A Temporal Approach to Defining Place Types based on User-Contributed Geosocial Content

Author: McKenzie, Grant Donald

Degree Grantor: University of California, Santa Barbara. Geography

Degree Supervisor: Krzysztof Janowicz and Martin Raubal

Place of Publication: [Santa Barbara, Calif.]

Publisher: University of California, Santa Barbara

Creation Date: 2015

Issued Date: 2015

Topics: Information Science, Geodesy, and Geography

Keywords: Semantic Signatures  
Geosocial Networking  
Point of Interest  
User-generated Content  
Place  
Temporal Signatures

Representations of an Urban Neighborhood : Residents' Cognitive Boundaries of Koreatown, Los Angeles

Author: Bae, Crystal Ji-Hye

Degree Grantor: University of California, Santa Barbara. Geography

Degree Supervisor: Daniel R. Montello

Place of Publication: [Santa Barbara, Calif.]

Publisher: University of California, Santa Barbara

Creation Date: 2015

Issued Date: 2015

Topics: Asian American studies, Social psychology, Urban planning, and Geography

Keywords: Boundaries  
Mental maps  
Koreatown  
Urban neighborhoods  
Cognitive regions  
Los angeles

Overview

Background

Verbalization

Spatialization

Generalization

Conclusions

18



# How can we elicit and spatially represent the topics of research data to convey their similarity?

1. **Collect metadata for research documents**
2. Model topics of document titles and abstracts
3. Generate spatializations in field and network spaces
4. Demonstrate similarity relations based on distance

<i>Metadata element</i>	<i>Requirement</i>
<b>Title</b>	<b>50 words or less</b>
Year of publication	2011 – 2016
Degree grantor	Academic department
Degree supervisor	Academic advisor
<b>Detailed abstract</b> ...	<b>Problem statement, description of methods and procedures used to gather data, summary of findings; no word limit</b>

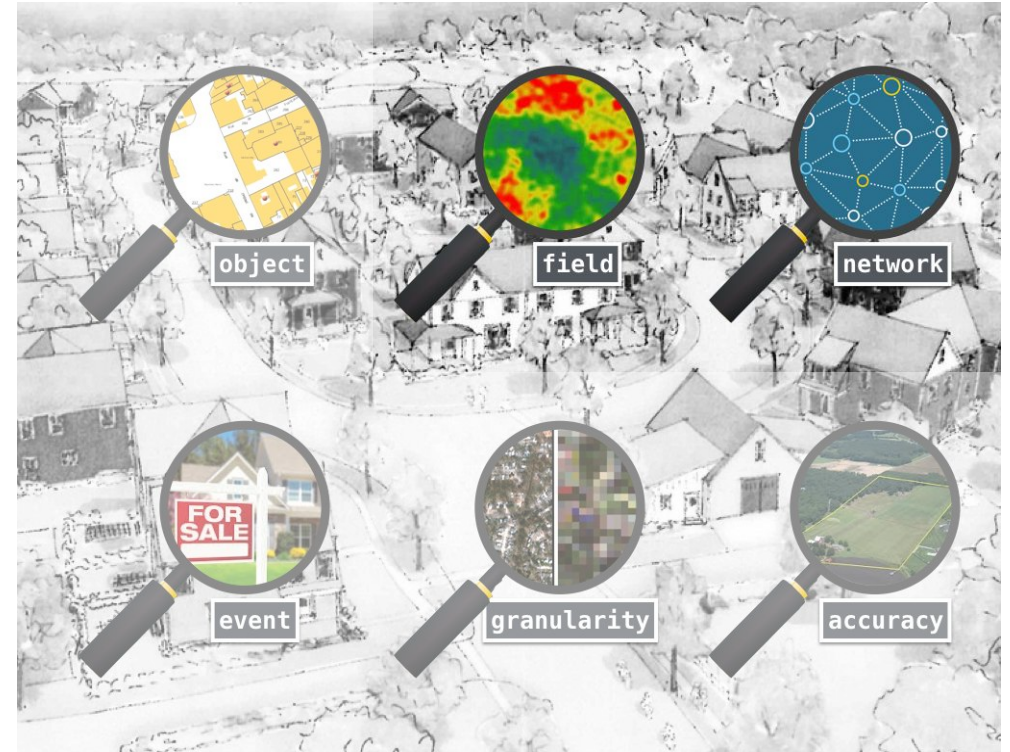
# How can we elicit and spatially represent the topics of research data to convey their similarity?

1. Collect metadata for research documents
- 2. Model topics of document titles and abstracts**
3. Generate spatializations in field and network spaces
4. Demonstrate similarity relations based on distance

<i>Title</i>	<i>topic 0</i> ( <i>'species',</i> <i>'spatial',</i> <i>'host'...</i> )	<b><i>topic 1</i></b> ( <b><i>'urban',</i></b> <b><i>'region',</i></b> <b><i>'local'...</i></b> )	<i>topic 2 ...</i> ( <i>'species',</i> <i>'population',</i> <i>'coastal...'</i> ) ...
Direct and Indirect Contributions of Photodegradation to Litter Decomposition in a California Grassland	0.47	<b>1.54E-04</b>	9.48E-05
Representations of an Urban Neighborhood: Residents' Cognitive Boundaries of Koreatown, Los Angeles	1.00E-04	<b>0.47</b>	1.14E-04
Household and Community Organization at Nimatlala, an Island Chumash Village on Limuw (Santa Cruz Island), California...	9.98E-05	<b>0.19</b>	0.33

# How can we elicit and spatially represent the topics of research data to convey their similarity?

1. Collect metadata for research documents
2. Model topics of document titles and abstracts
3. **Generate spatializations in field and network spaces**
4. Demonstrate similarity relations based on distance

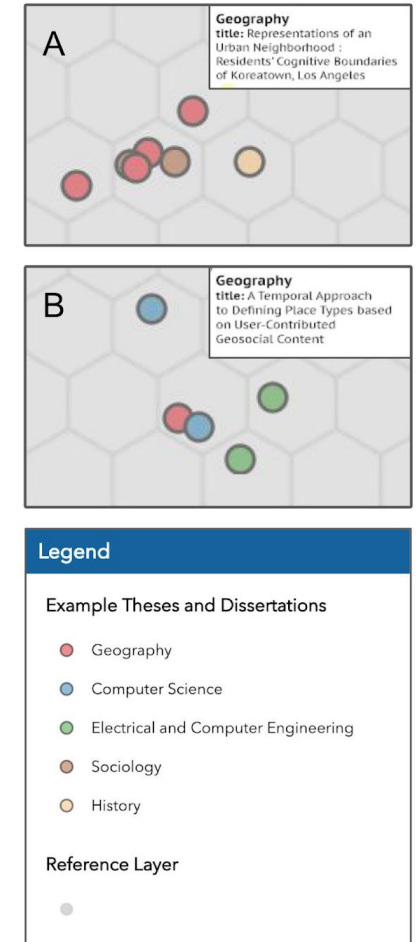
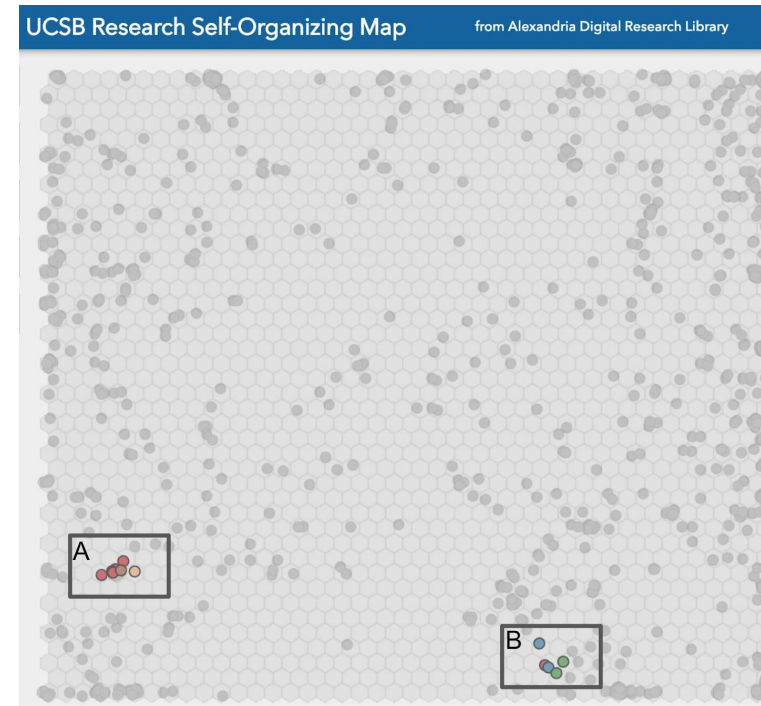


**Core concepts** of spatial information  
(Kuhn, 2012)

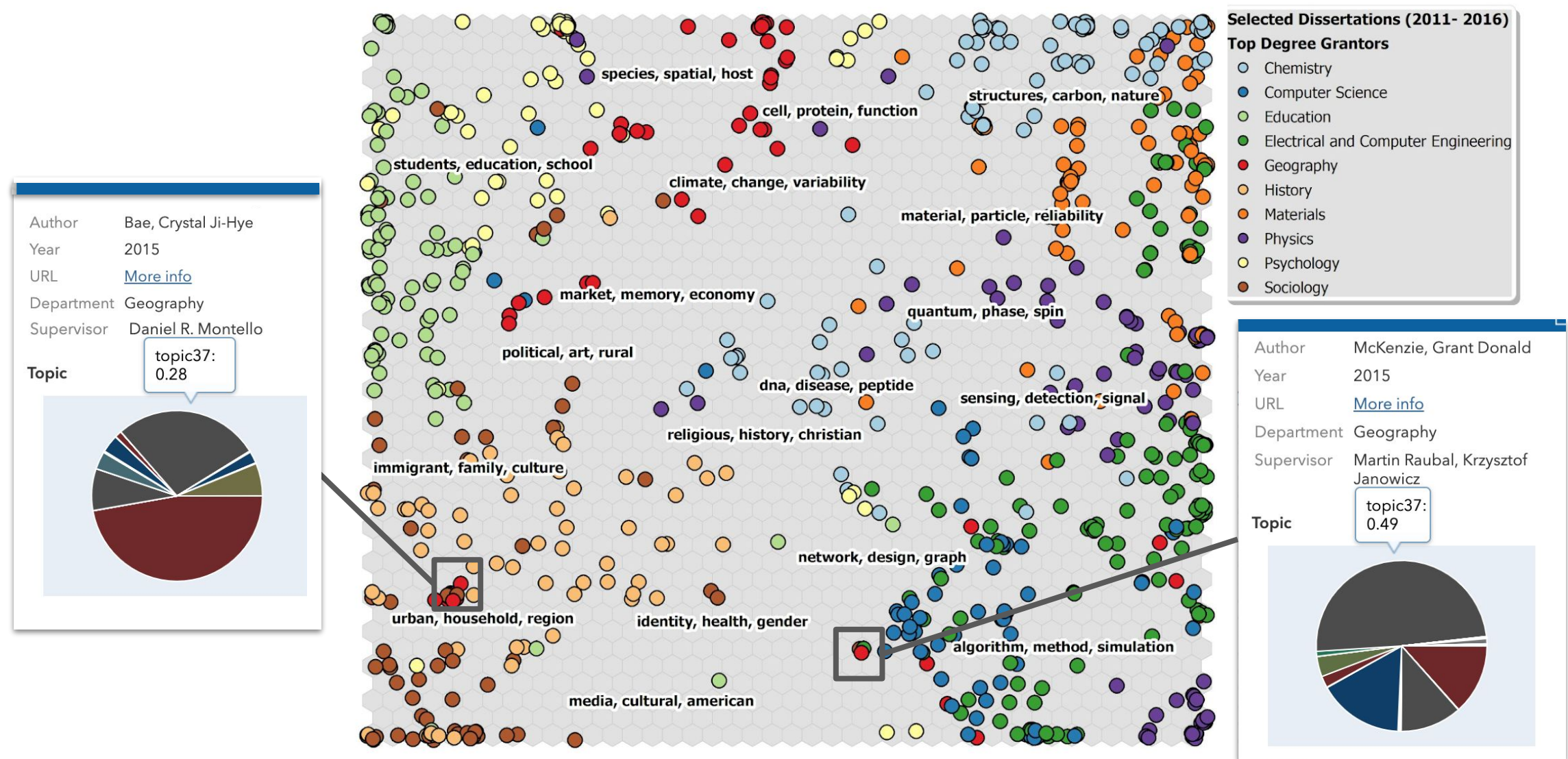


# How can we elicit and spatially represent the topics of research data to convey their similarity?

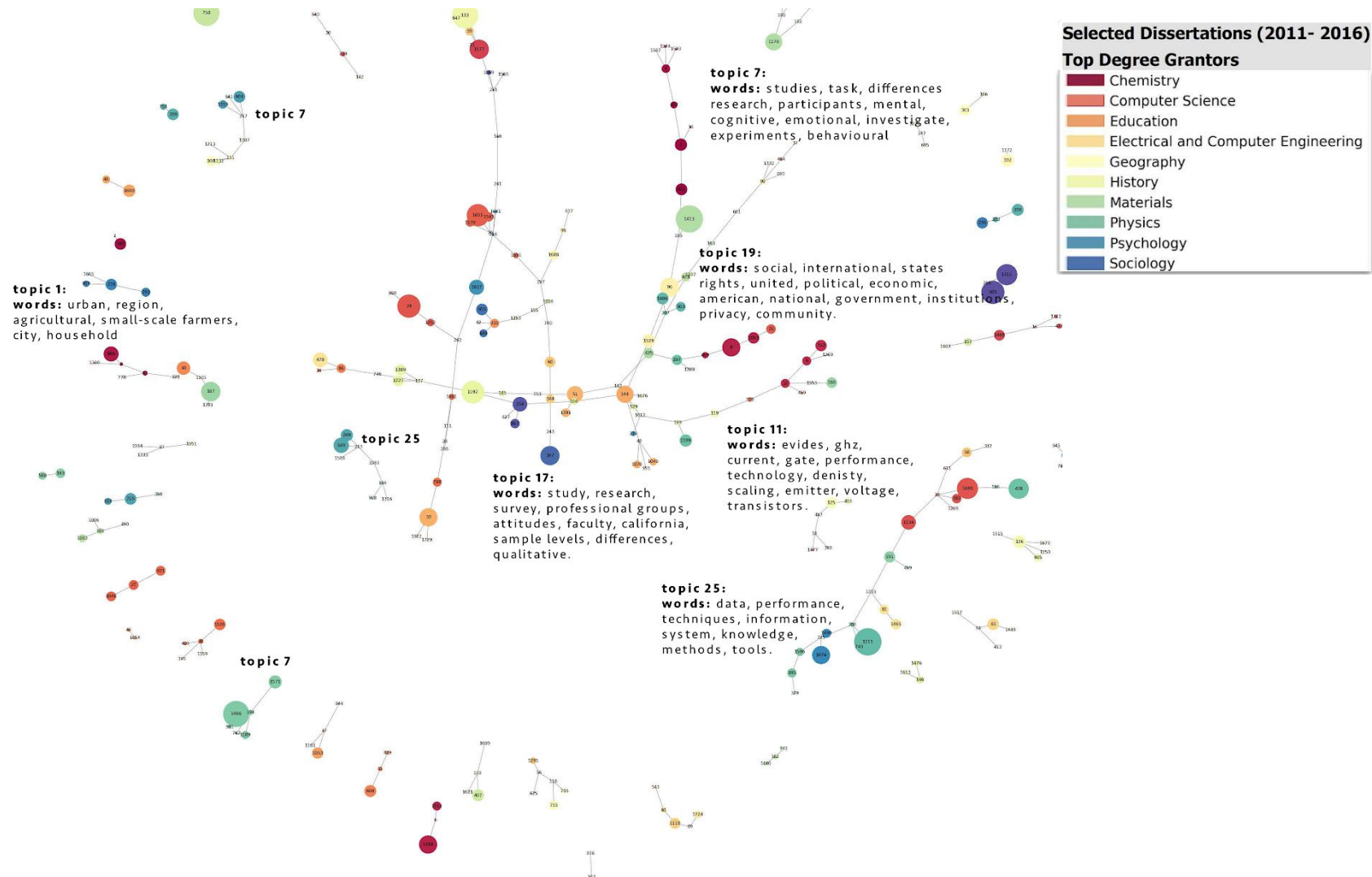
1. Collect metadata for research documents
2. Model topics of document titles and abstracts
3. Generate spatializations in field and network spaces
4. **Demonstrate similarity relations based on distance**



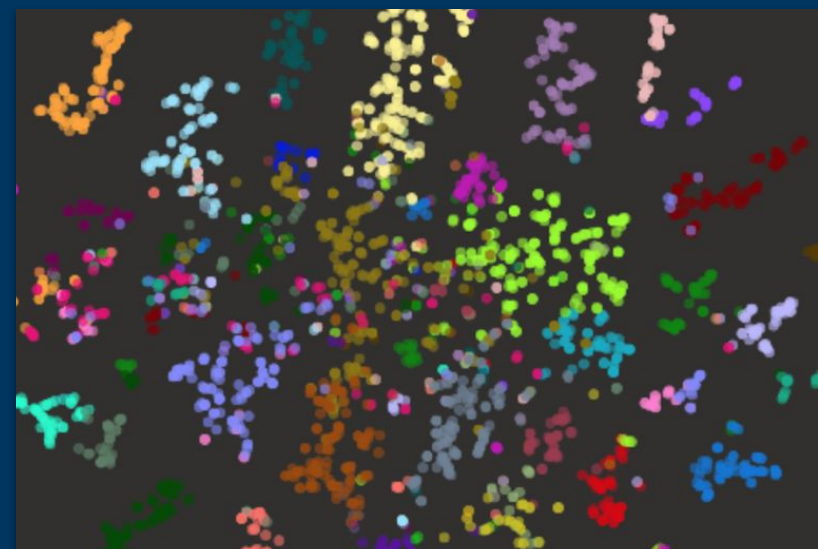
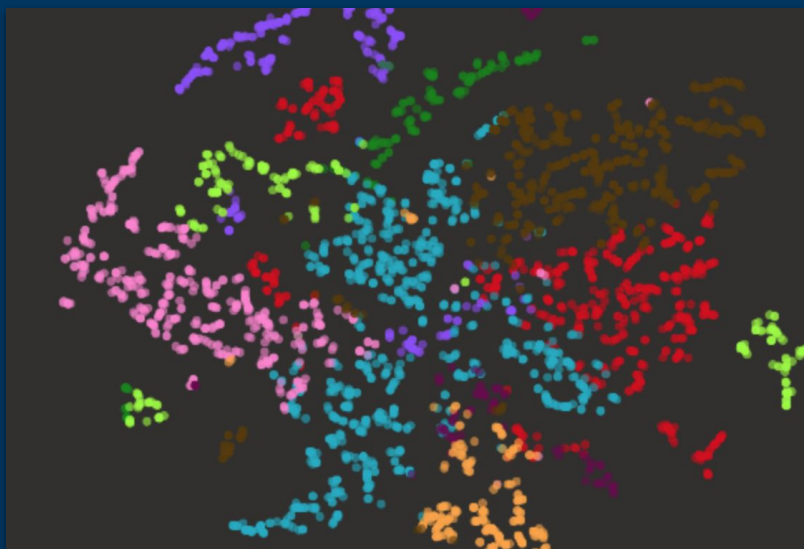
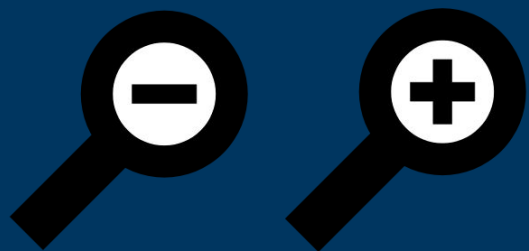
# Field of research topics



# Network of research topics







*Lafia et al. (2020)*

*Spatialization* → *Generalization*

**Mapping Research Topics at Multiple Levels of Detail**  
(Study 3)

# Research productivity is difficult to **quantify** and **compare** across disciplines.

## Plumes and Blooms



Each year, winter rains wash sand, mud and other terrestrial debris into the Santa Barbara Channel. Then, during the spring and summer, phytoplankton populations increase dramatically and ultimately provide the primary energy source for the entire marine food web. These alternating patterns of... more

Tags: Earth Systems Science

## Cheadle Center for Biodiversity & Ecological Restoration

The mission of the Vernon and Mary Cheadle Center for Biodiversity and Ecological Restoration (CCBER) at UCSB is to support:



Education  
Management  
Management, Restoration and Conservation

Tags: Environmental Information Management

## Bermuda Bio Optics Project

The Bermuda Bio-Optics Project (BBOP) is a long term study of the factors contributing to the regulation of the underwater light field in the open ocean and the resulting biogeochemical impact. These studies are done, on average, once a month in conjunction with the Bermuda-Atlantic Time Series... more



Tags: Earth Systems Science

## Snow Hydrology Research Group

The Snow Hydrology Research Group is part of the Donald Bren School of Environmental Science and Management at the University of California, Santa Barbara. It is also a member of the ESIP Federation (Earth Science Information Partners). The primary research focus of this group is NASA's REASoN (... more



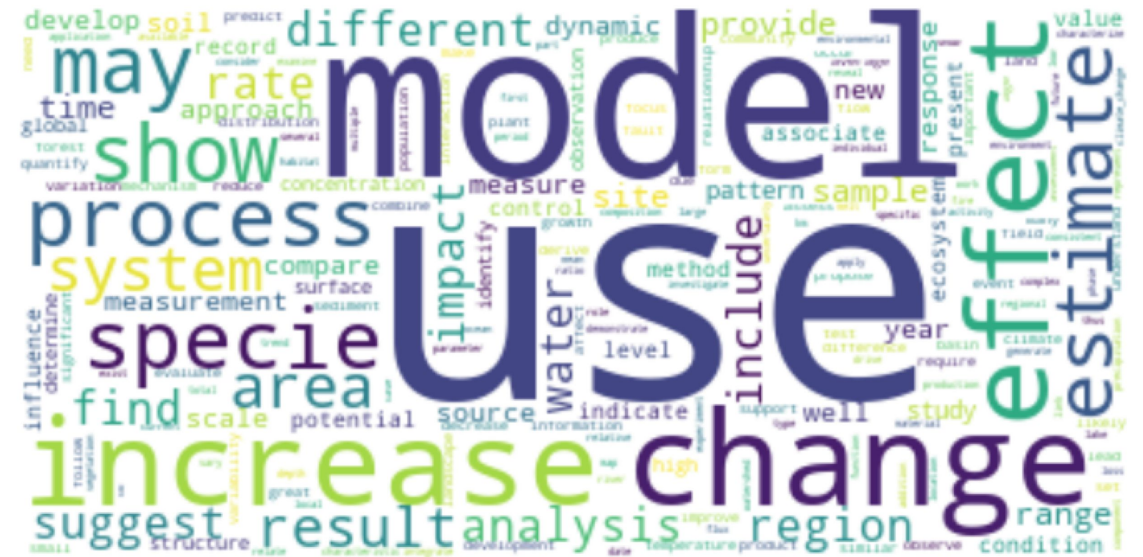
Tags: Earth Evolution, Earth Systems Science, Environmental Information Management, Human Impacts



Source: <https://www.eri.ucsb.edu/>

# How can we represent topics of a multidisciplinary body of research at multiple levels of detail?

1. **Analyze and process document metadata**
2. Select number of topics to model based on coherence
3. Spatialize topics at a coarse and a detailed level
4. Deploy a map dashboard and interpret results

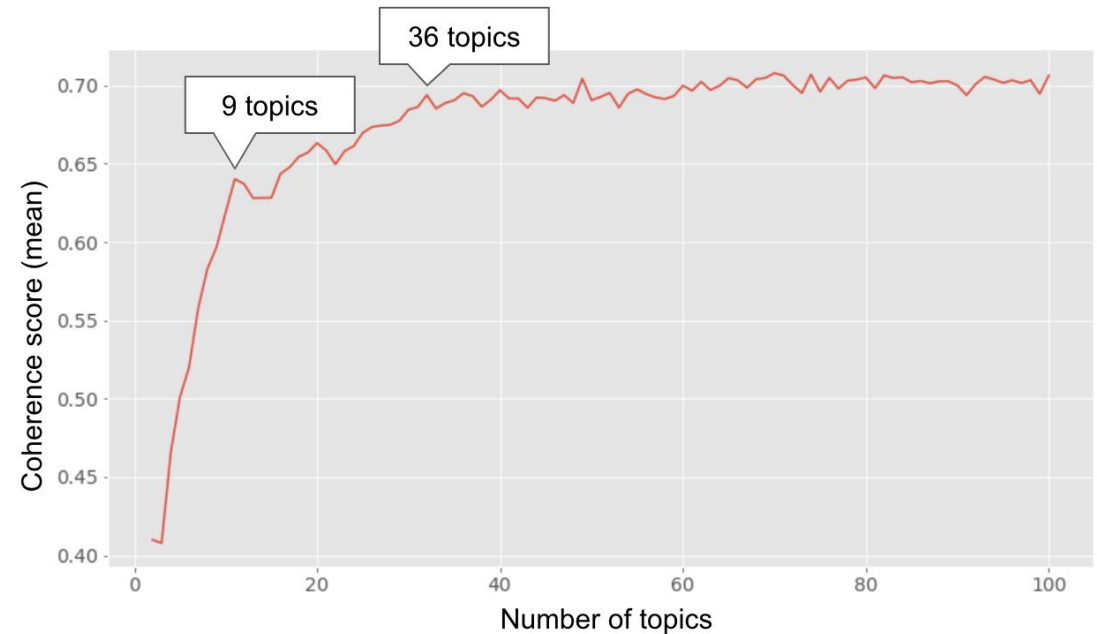


**Frequent terms** in ERI's 3,770 research documents (2009 – 2019)



# How can we represent topics of a multidisciplinary body of research at multiple levels of detail?

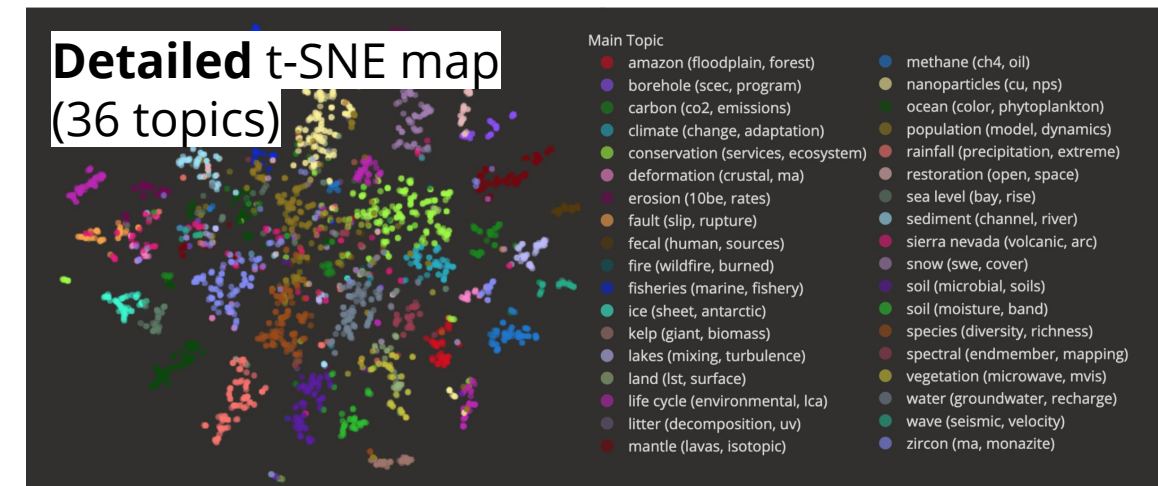
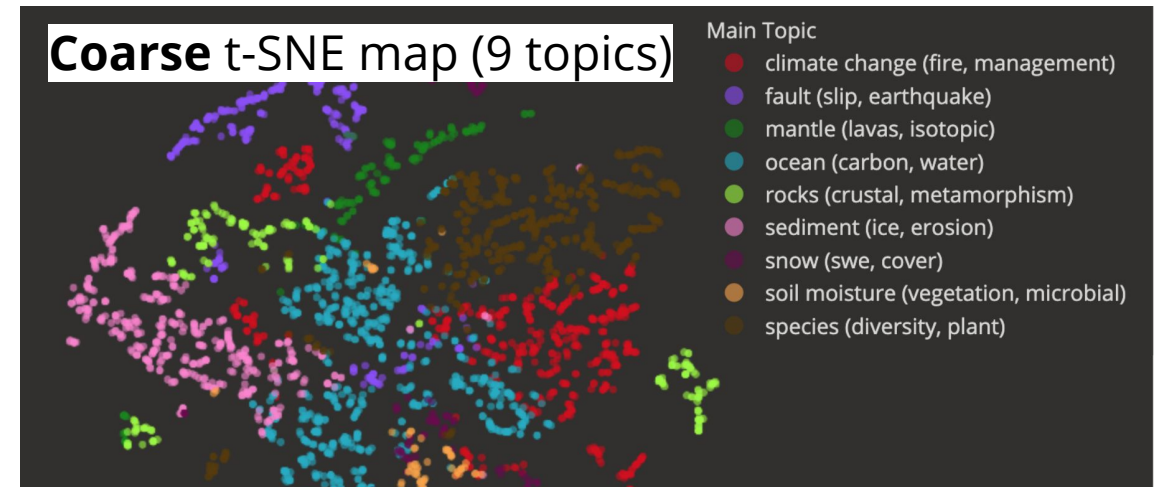
1. Analyze and process document metadata
- 2. Select number of topics to model based on coherence**
3. Spatialize topics at a coarse and a detailed level
4. Deploy a map dashboard and interpret results



**Coherence** scores for NMF topic models with 2 – 100 topics

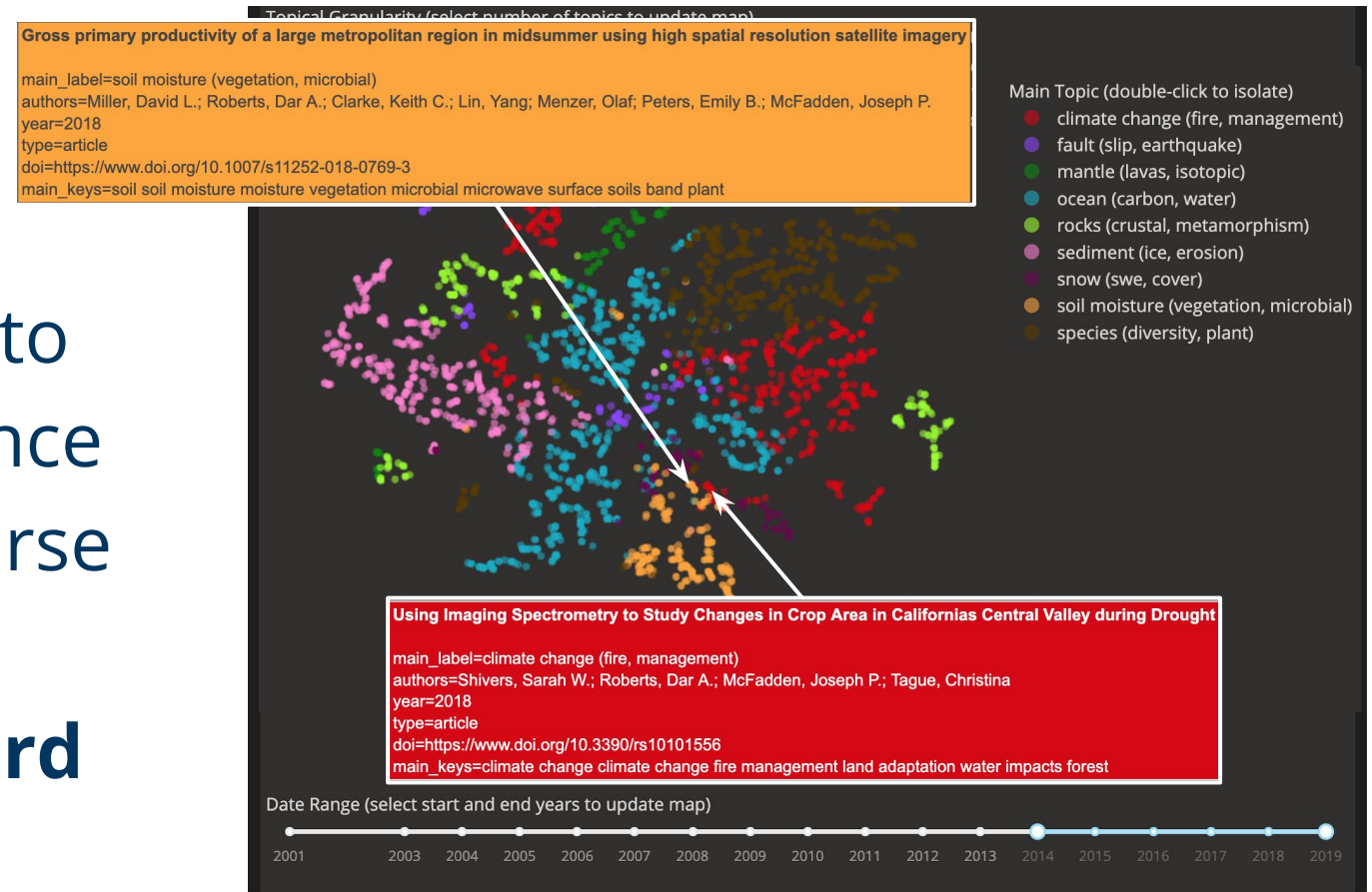
# How can we represent topics of a multidisciplinary body of research at multiple levels of detail?

1. Analyze and process document metadata
2. Select number of topics to model based on coherence
3. **Spatialize topics at a coarse and a detailed level**
4. Deploy a map dashboard and interpret results



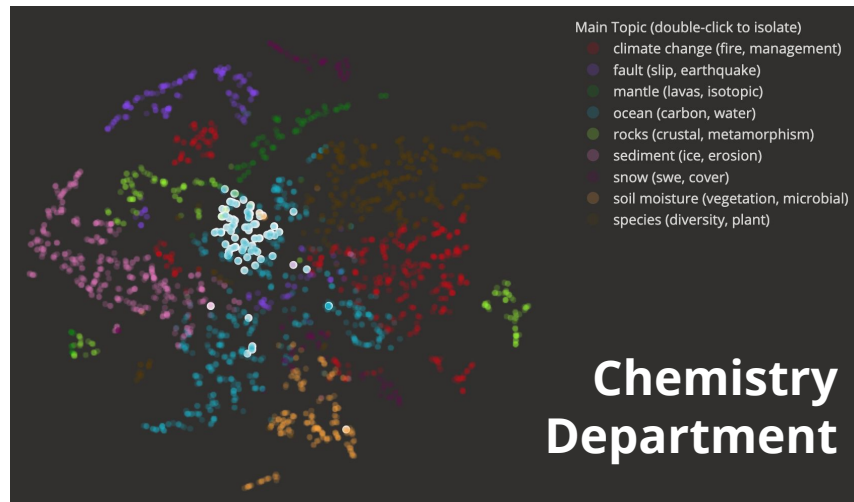
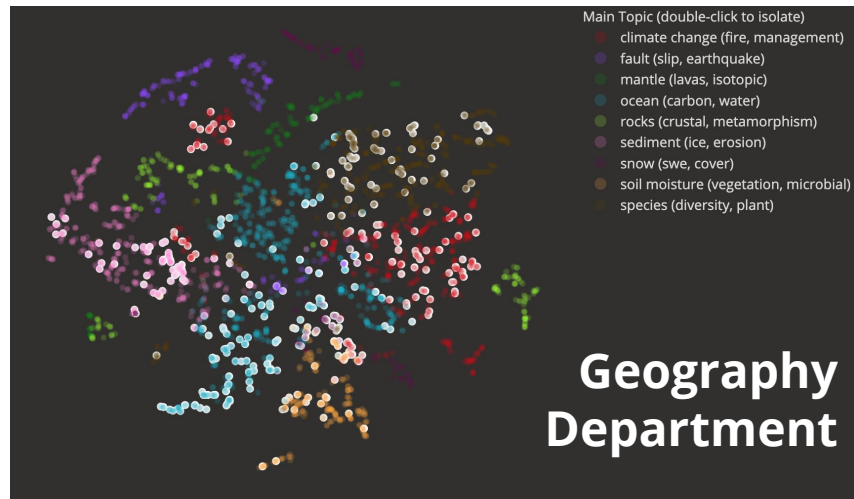
# How can we represent topics of a multidisciplinary body of research at multiple levels of detail?

1. Analyze and process document metadata
2. Select number of topics to model based on coherence
3. Spatialize topics at a coarse and a detailed level
4. **Deploy a map dashboard and interpret results**



Source: <https://eri-research-dashboard.herokuapp.com/>

# Can these maps support **high-level** views of research at a multidisciplinary institute?

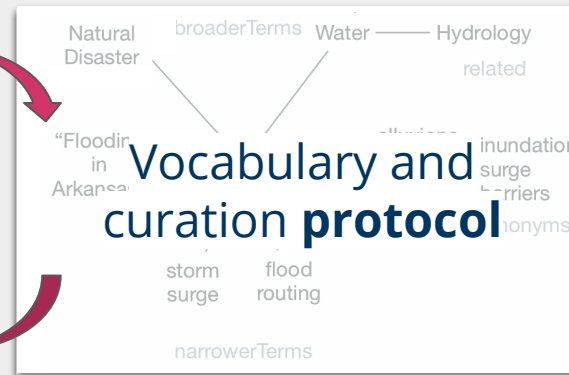


- **Review questions:** support for standard “research accomplishment” questions (e.g. trends, specialities)
- **Researcher survey:** ERI’s research, their research, detection of events (e.g. center funding, faculty hires)



### **Verbalization** (Study 1)

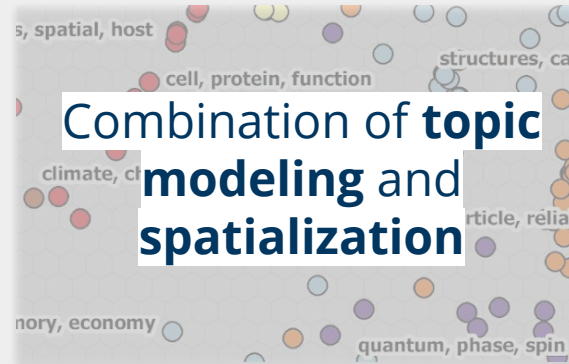
User terms mapped to system terms with a hierarchical vocabulary



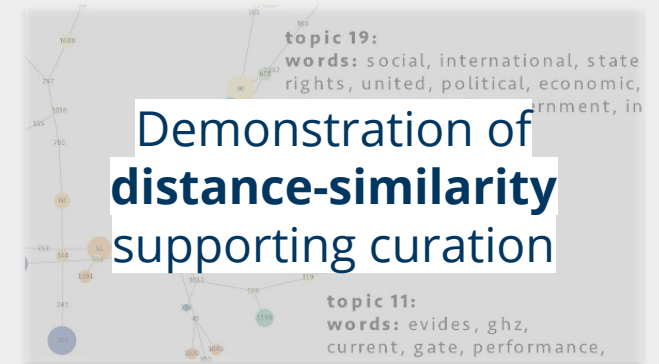
Expansion of search terms across **levels of government**

### **Spatialization** (Study 2)

Research topics elicited from metadata configured as both a field and a network



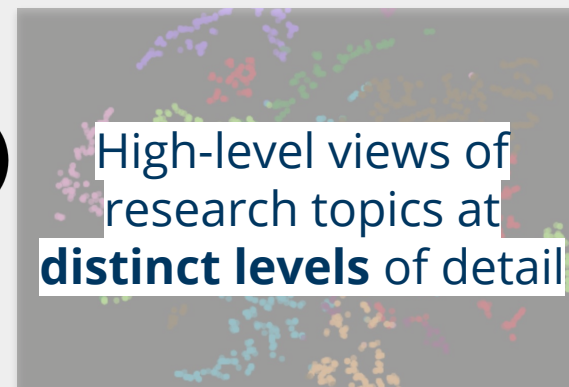
Combination of **topic modeling** and **spatialization**



Demonstration of **distance-similarity** supporting curation

### **Generalization** (Study 3)

Research topics elicited from metadata configured in temporally-sequenced maps



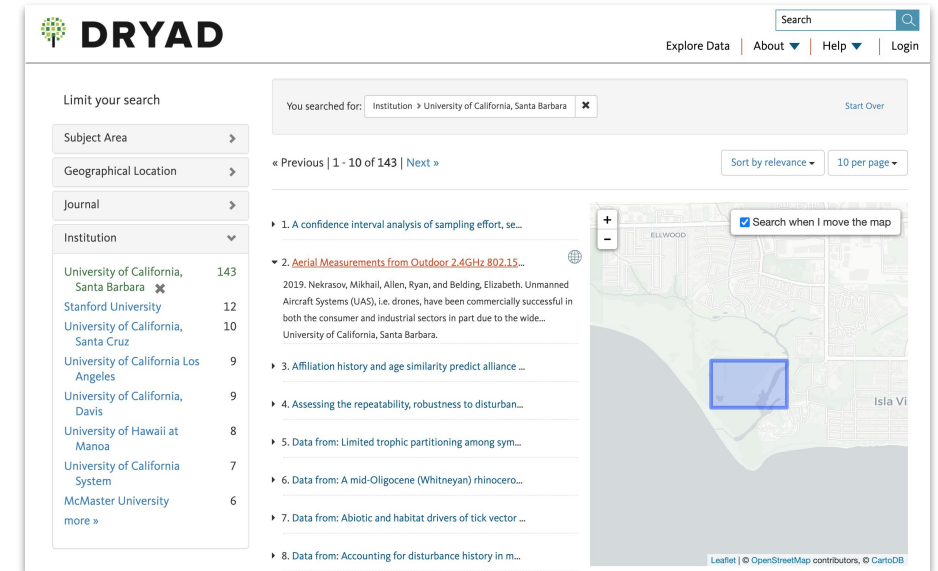
High-level views of research topics at **distinct levels** of detail



**Spatial** support for the institutional review process

# Limitations

- **Evaluation baseline:** innovating previously unseen solutions
- **Feedback mechanisms:** potential for cross-study tasks and insights
- **Research data proxies:** adoption of data curation policies



UCSB joins **Dryad** Data Repository  
(143 items contributed so far)

# Open Questions

- How can core concepts of spatial information further support the **spatial curation** of research?
- Which **curatorial actions** impact data discovery and reuse?
- How can **recommendation** and **question-answering** support data discovery and reuse?



**Core concepts** of spatial information  
(Kuhn, 2012)

# References

Biberstine, J., Duhon, R. J., Börner, K., Hardy, E., & Skupin, A. (2010). A semantic landscape of the Last.fm music folksonomy: Using a self-organizing map. *Cyberinfrastructure for Network Science Center*.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

Buckland, M. K. (1997). What is a “document”? *Journal of the American Society for Information Science*, 48(9), 804-809.

Durante, K., & Hardy, D. (2015). Discovery, management, and preservation of geospatial data using hydra. *Journal of Map & Geography Libraries*, 11(2), 123-154.

Fear, K. M. (2013). Measuring and anticipating the impact of data reuse (Doctoral dissertation).

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11), 964-971.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT Press.

Hearst, M. A. (1999). User interfaces and visualization. In *Modern Information Retrieval*, 257-323.

Kuhn, W. (1996, August). Handling data spatially: Spatializing user interfaces. In *Proceedings of the 7th International Symposium on Spatial Data Handling*, Delft, The Netherlands (pp. 12-16).

Kuhn, W. (2012). Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, 26(12), 2267-2276.

Lafia, S., Jablonski, J., Kuhn, W., Cooley, S., & Medrano, F. A. (2016). Spatial discovery and the research library. *Transactions in GIS*, 20(3), 399-412.

Lafia, S., Turner, A., & Kuhn, W. (2018). Improving discovery of open civic data. In 10th International Conference on Geographic Information Science (GIScience 2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Lafia, S., Last, C., & Kuhn, W. (2019). Enabling the discovery of thematically related research objects with systematic spatializations. In *14th International Conference on Spatial Information Theory (COSIT 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Lafia, S., Kuhn, K., & Caylor, K. (2020). Mapping research topics at multiple levels of detail. Manuscript submitted for publication.

Mayernik, M. S. (2016). Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, 67(4), 973-993.

Montello, D. R., Fabrikant, S. I., Ruocco, M., & Middleton, R. S. (2003, September). Testing the first law of cognitive geography on point-display spatializations. In *International Conference on Spatial Information Theory* (pp. 316-331). Springer, Berlin, Heidelberg.

Sinton, D. (1978). The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study. *Harvard papers on geographic information systems*.

Smith, T. R., & Frew, J. (1995). Alexandria digital library. *Communications of the ACM*, 38(4), 61-62.

Svenonius, E. (2000). *The intellectual foundation of information organization*. MIT press.





## Special thanks to:

Andrew Turner, Pranav Kulkarni,  
Daniel Fenton, and Alexander Harris  
(Esri R&D), Christina Last (University  
of Bristol), André Bruggmann  
(University of Zurich)

Sara Lafia  
Ph.D. Candidate in Geography

Committee: Werner Kuhn (chair),  
James Frew, Kelly Caylor, Daniel Montello