



Designing for Serendipity: Research Data Curation in Topic Spaces

Sara Lafia
Ph.D. Candidate in Geography

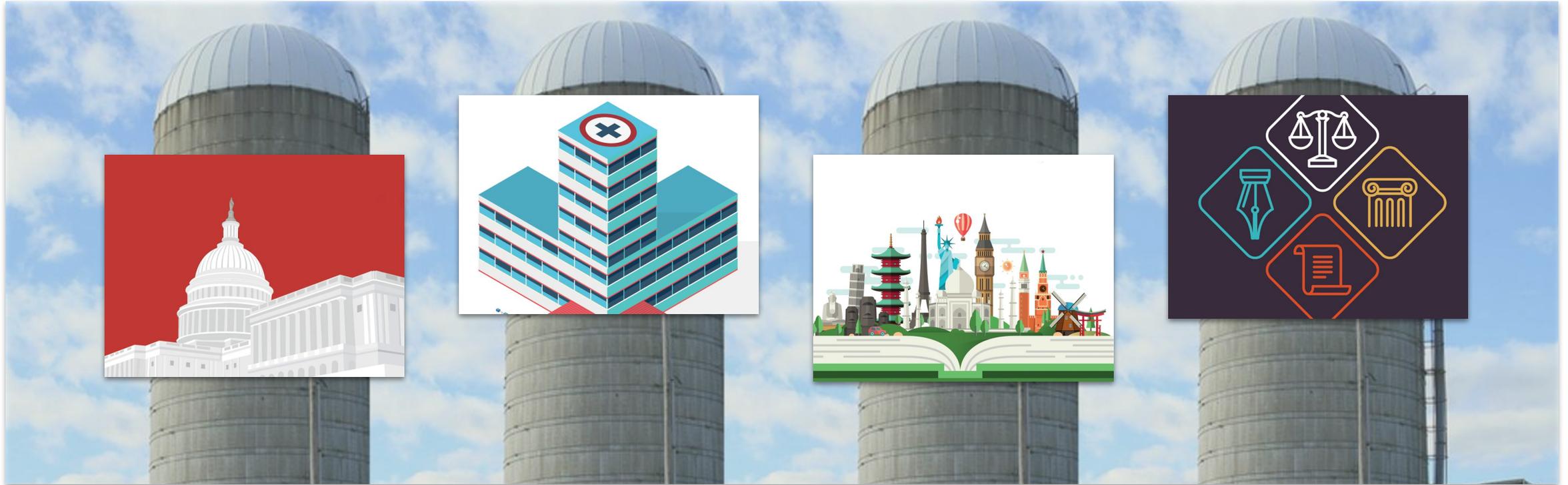
Committee: Werner Kuhn (chair),
James Frew, Kelly Caylor, Daniel Montello

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

Overview	Background	Verbalization	Spatialization	Generalization	Conclusions
-----------------	------------	---------------	----------------	----------------	-------------

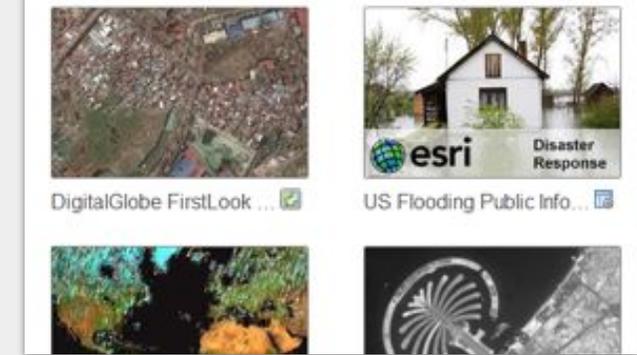
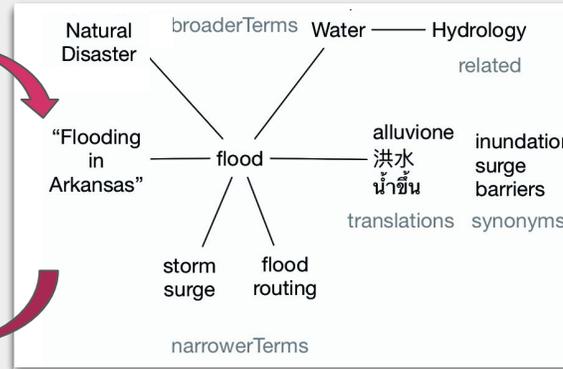
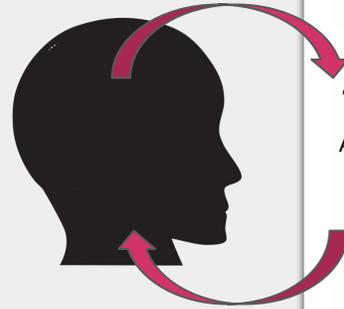
Bookshelves curate books by **topics**, supporting search and discovery.



- **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)

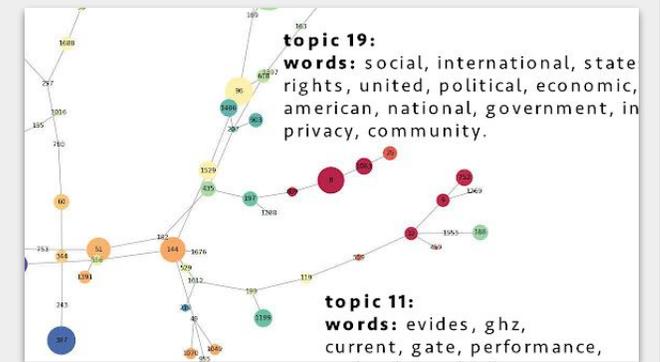
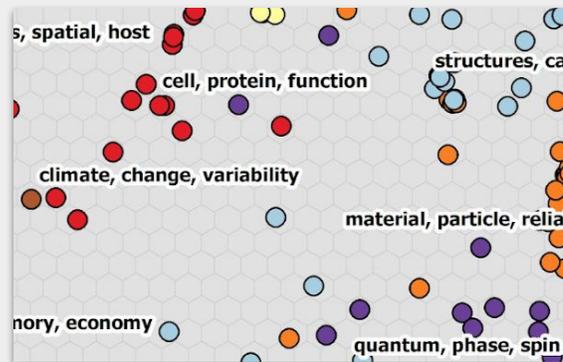
Verbalization (Study 1)

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



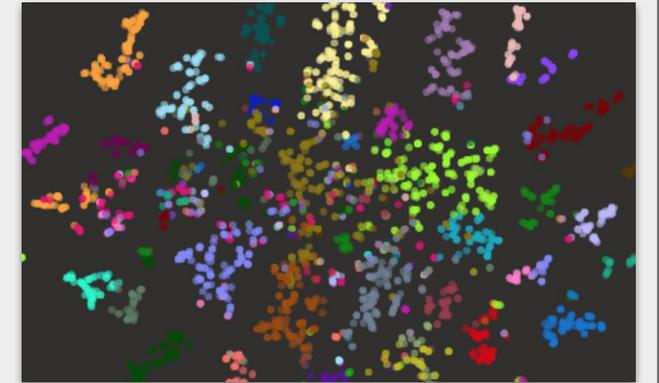
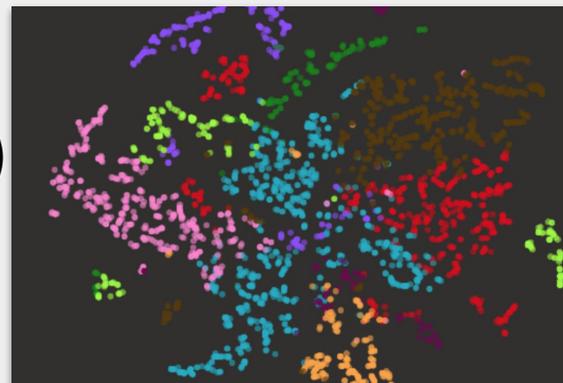
Spatialization (Study 2)

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

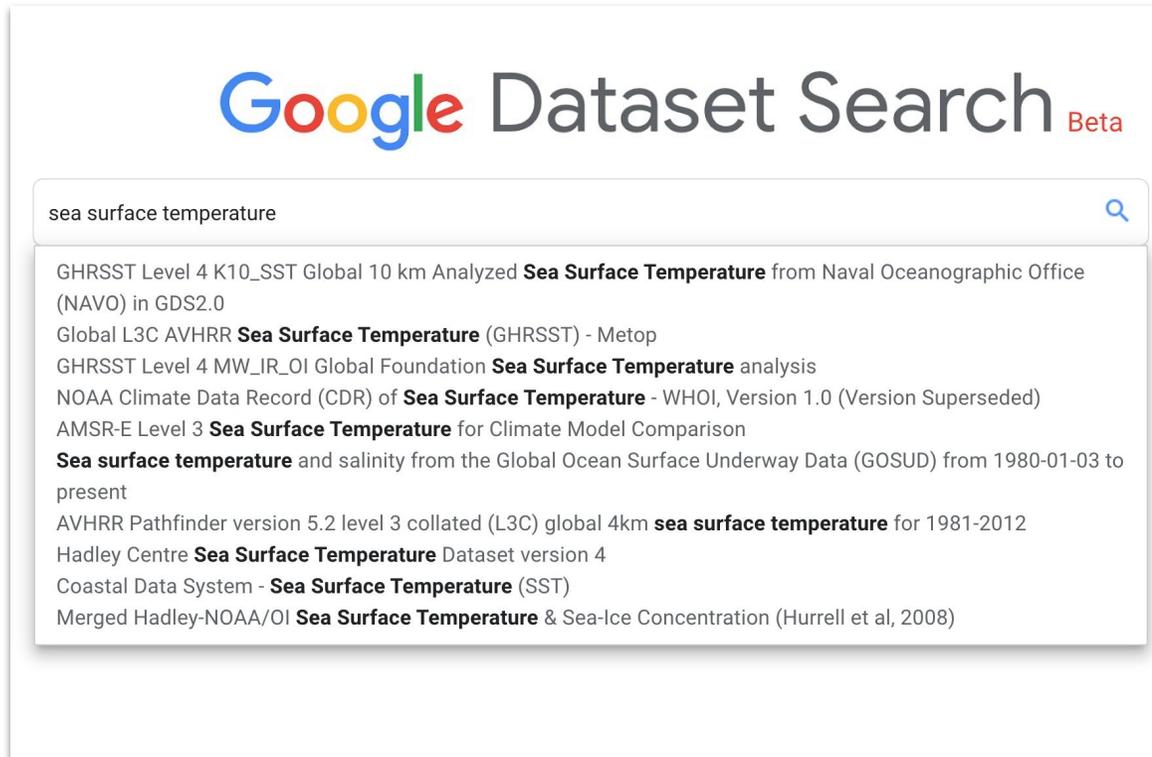


Generalization (Study 3)

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



Seeking research data

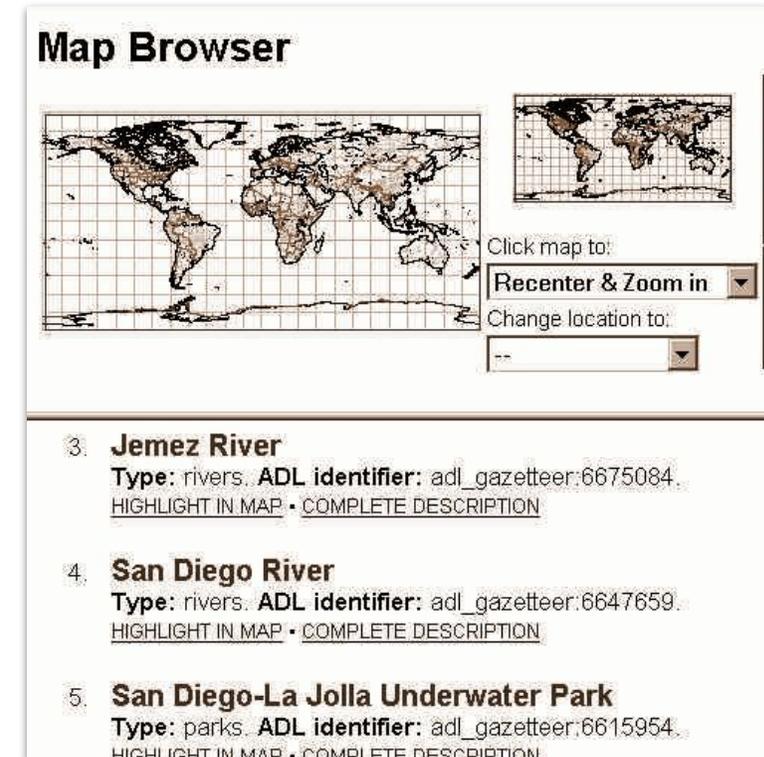


Google Dataset Search **Beta**

sea surface temperature

- GHRSSST 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** (GHRSSST) - Metop
- GHRSSST 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)



Map Browser

Click map to:
Recenter & Zoom in
Change location to:

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

The screenshot shows the Stanford EarthWorks website interface. At the top, there is a search bar with the text "Try 'groundwater'" and a "Search" button. Below the search bar, there are navigation options: "Start Over", "Bounding Box" with coordinates "-128.111572 35.675147 -116.246338 39.164141", and a "Search" button. A "Limit your search" section is visible on the left, with filters for Institution, Author, Publisher, Subject, Place, Year, Access, and Data type. The main content area displays a list of search results, including "1. 100-Meter Bathymetric Contours: Monterey ...", "2. 100-Meter Map Grid, Jasper Ridge Biological ...", "3. 10-Meter Bathymetric Contours: Monterey B...", "4. 10-Meter Contours, Jasper Ridge Biological P...", "5. 1990 Census Roads - San Francisco Bay Area, ...", "6. 1992 Election Precincts", "7. 1-Degree Digital Elevation Model: Monterey ...", "8. 1-Degree Hillshade Digital Elevation Model: ...", and "9. 1-Meter Bathymetric Digital Elevation Model: ...". A map of California is shown on the right, with a red box highlighting the San Francisco Bay Area. The map includes a "Search when I move the map" checkbox and a "Sort by relevance" dropdown menu.

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

The screenshot shows the UCSB Open Data website interface. At the top, there is a search bar with the text "guatemala" and a "SEARCH WITHIN MAP" button. Below the search bar, there is a map of Guatemala with several search results displayed as thumbnails. The thumbnails include a satellite view of a region, a map of Mexico with a red box highlighting the state of Morelia, and a map of Acapulco. The main map shows the geographical outline of Guatemala and its neighboring countries: Mexico to the north, Honduras to the east, and El Salvador to the south. The map also shows the Gulf of Tehuantepec and the Gulf of Honduras. The search results are titled "Archaeological Sites Maya Forest GIS".

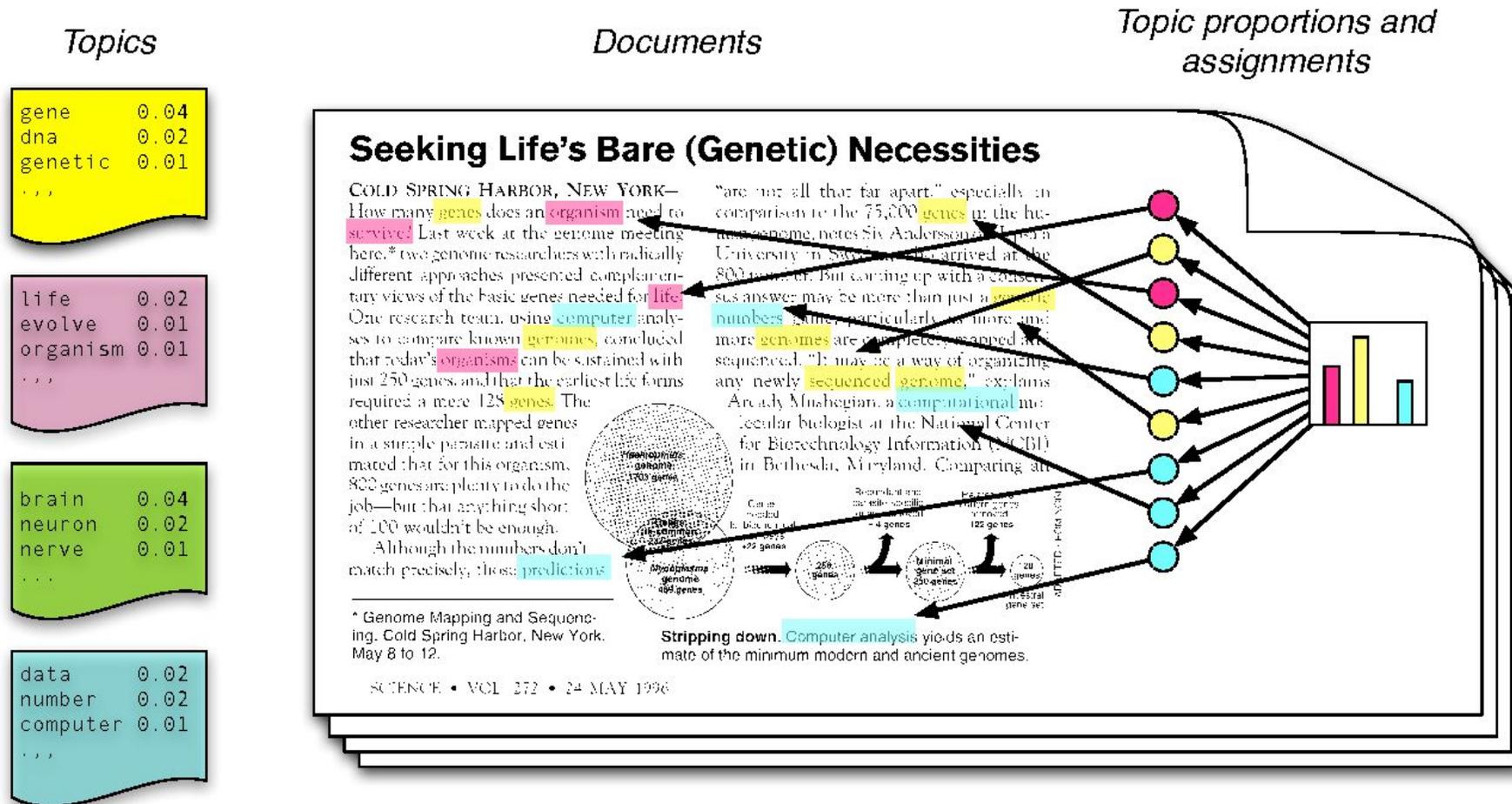
Multidisciplinary **research data**
(Lafia et al., 2016)

Making topics explicit: **semantic annotation**

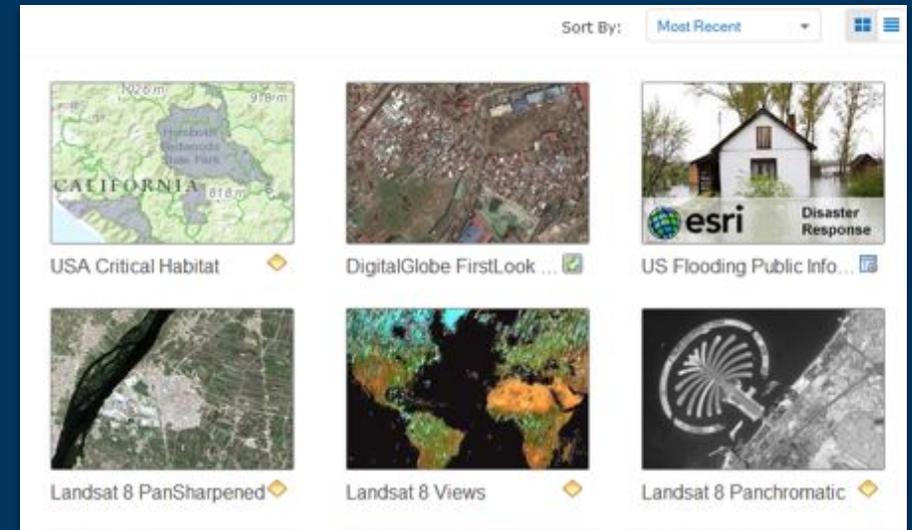
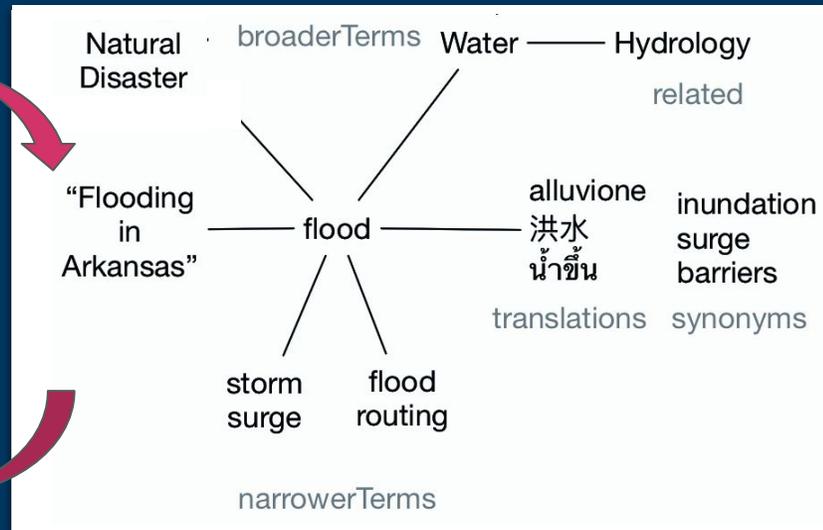
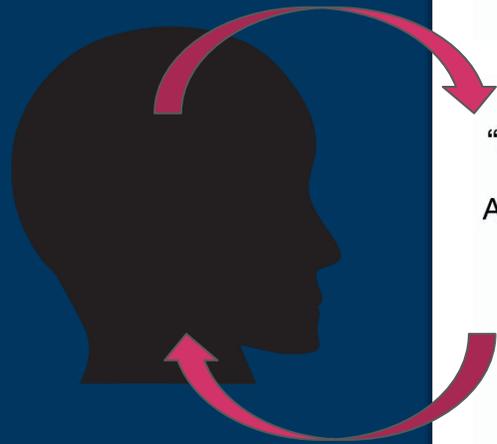


Hierarchical classification versus other spatial conceptualizations
(Gärdenfors, 2000)

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**.

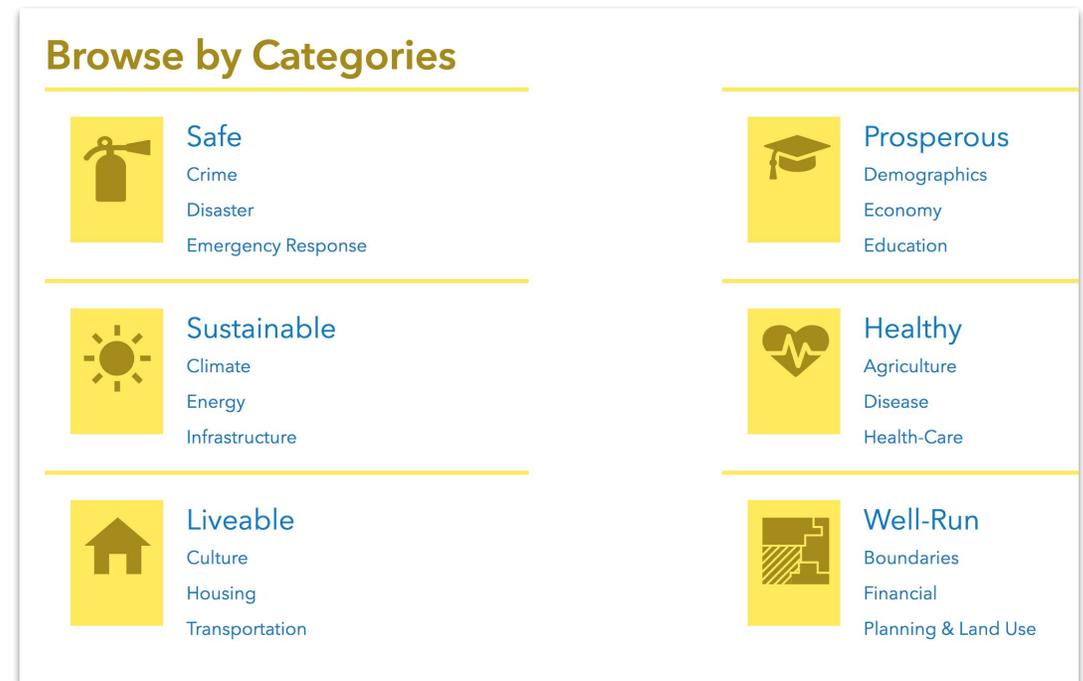
The screenshot shows the Michigan's Open Data Portal. The header includes the logo 'data.MICHIGAN.gov' and navigation links for 'Michigan.gov', 'Open Michigan', 'Developers', and 'Resources'. The main heading is 'Michigan's Open Data Portal' with a subtext: 'View and analyze data provided by a variety 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.' The featured dataset is 'MDOT Fatalities and Serious Injuries MI Public Roads' under the 'Infrastructure' category. 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, 'Updated October 1, 2019', 'Views 50', and 'Tags non-motorized, crashes, trans, mdot, bicycle, and 7 more'. There is also a link to 'API Docs'.

The screenshot shows the Open Data DC website. The header includes the logo 'Open Data DC' and navigation links for 'App Gallery', 'Data Stories', 'Developer Starter Kit', 'Data Policy', 'Feedback', and 'Handbook'. The main heading is 'Open Data DC' with a background image of a city street grid. The featured dataset is 'Crashes in DC' by 'City of Washington, DC | DCGISopendata'. The description states: '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...'. It includes 'Type: Feature Layer', 'Last Updated: December 10, 2019', 'Rows: 232,975', and 'Tags: accidents, bike, crash, Crashes, fatality, injuries, pede...'.

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

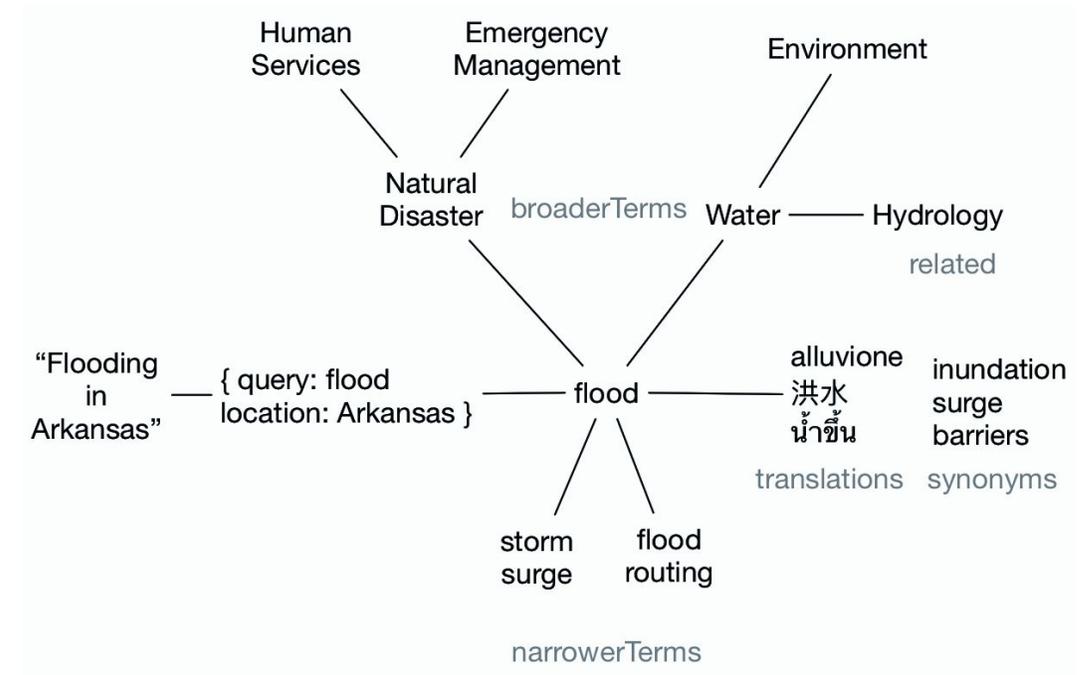
- > [Civil defense](#)
- > [Crime prevention](#)
- > [Emergency management](#)
- > [Fire extinction](#)
- > [Fire prevention](#)
- > [Offenses against public safety](#)
- > [Police](#)
- > [Traffic safety](#)

Exact Matching Concepts from Other Schemes

- > [public safety](#)

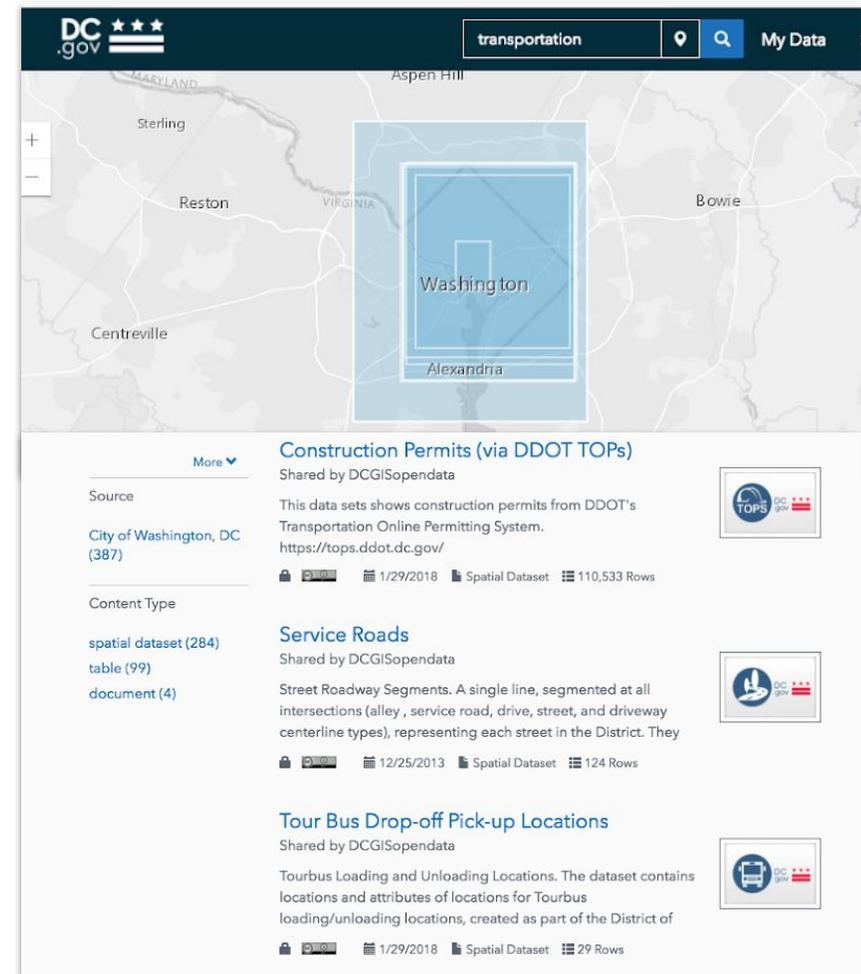
How can we map topics of interest, expressed in users' terms, onto the language of metadata?

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

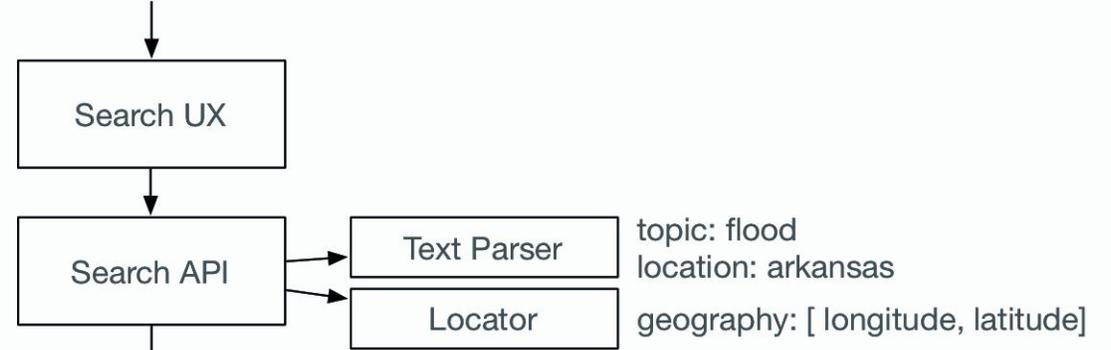
MDOT Fatalities and Serious Injuries MI Public Roads Infrastructure

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

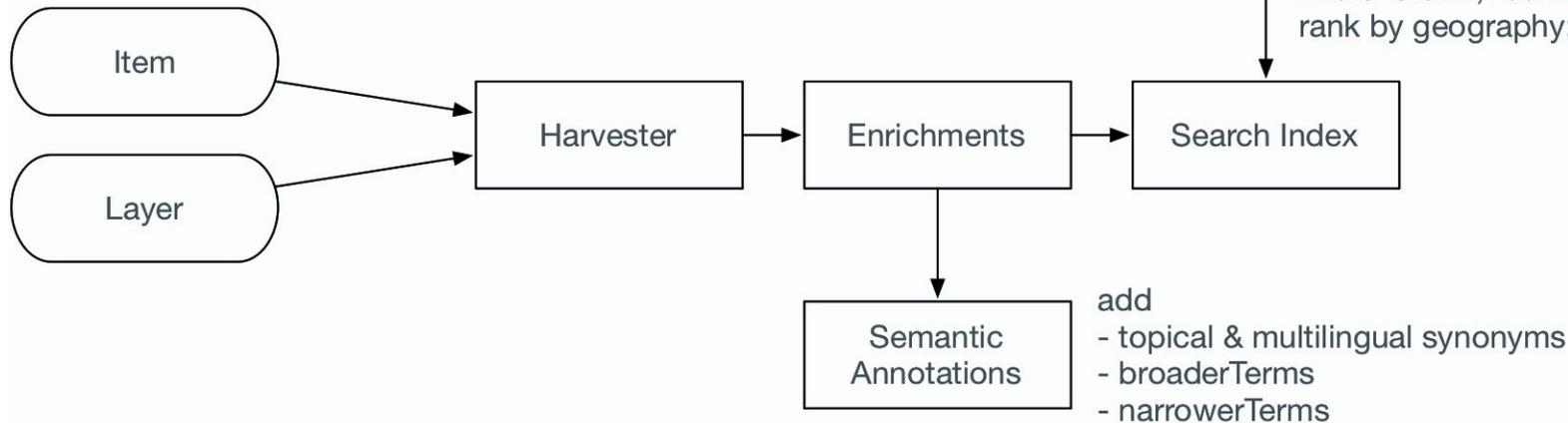
Crashes in DC
City of Washington, DC | DCGISopendata

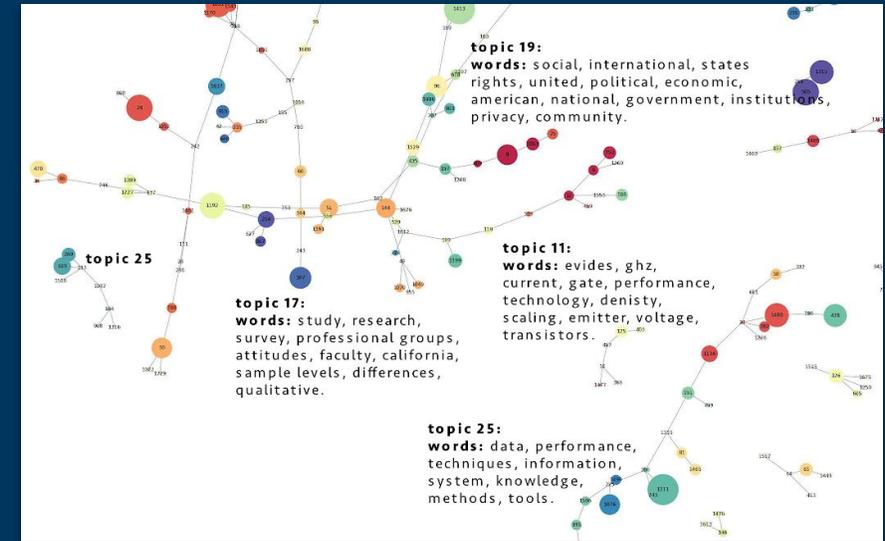
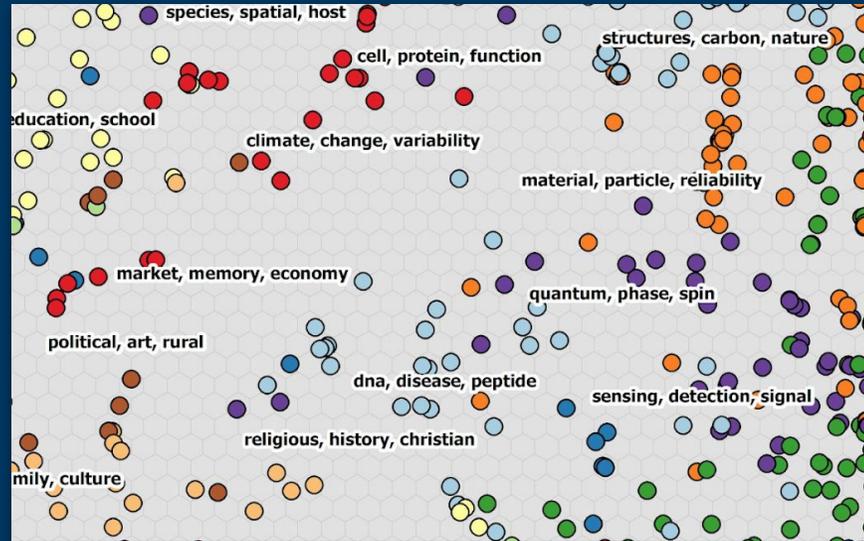
Crashes on the street segment line network of the District of Columbia maintained by the District Department of Transportation (DDOT). In addition to locations, a related table consisting of crash...

search: "Flooding in Arkansas"



search: synonyms
 recommendations: Natural Disaster, Water, Emergency Management
 filters: storm, routing, flood control, erosion, transportation
 rank by geography: [longitude, latitude]





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.

The screenshot shows the ADRL website interface. At the top left is the ADRL logo. The main header contains the text 'Alexandria Digital Research Library' and navigation links for 'About', 'FAQ', and 'Usage Guidelines'. Below the header is a search bar with a dropdown menu set to 'All Fields' and a 'Search' button. The main content area is divided into two columns. The left column, titled 'Limit your search', contains several filter buttons: 'Format', 'Collection', 'Contributor', 'Genre', 'Date', 'Academic Department', and 'Rights'. The right column features the title 'UCSB electronic theses and dissertations' followed by a descriptive paragraph: '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.' Below this text are metadata fields: 'Genres: Dissertations, Academic', 'Format: Text, Collection', 'ARK: ark:/48907/f3348hkz', and 'Local Identifier: etds'. At the bottom of this section is a 'Search Collection' input field with a 'Go' button. A footer for this section reads 'Items in this Collection' and '« Previous | 1 - 10 of 2,277 | Next »'.

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

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

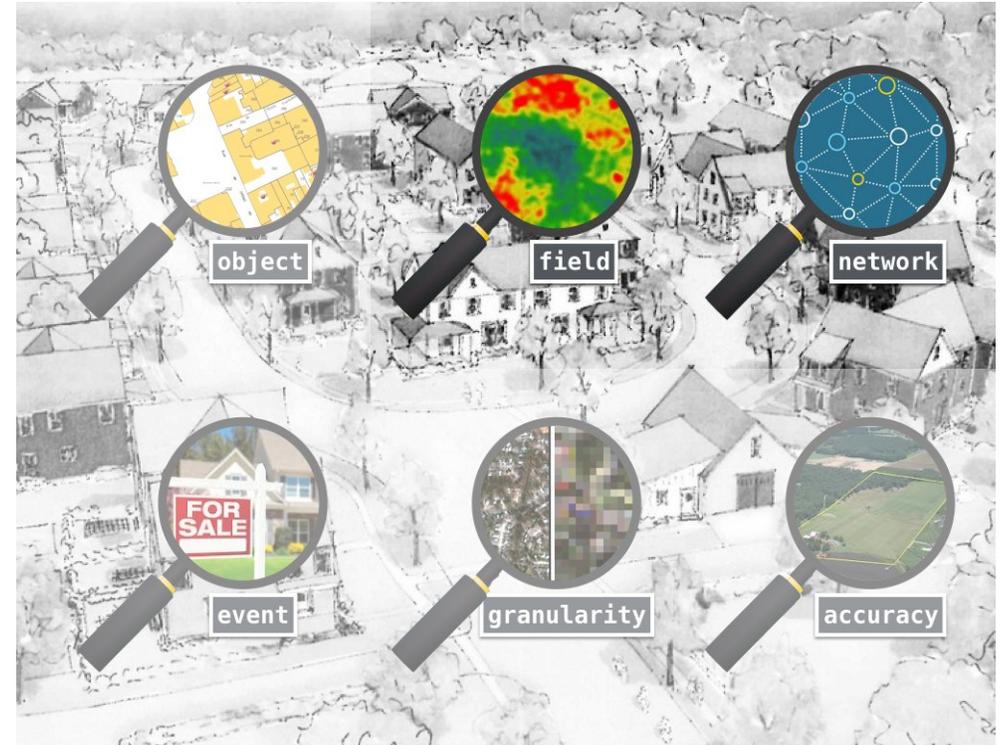
How can we elicit and spatially represent the topics of research data to convey their similarity?

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

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

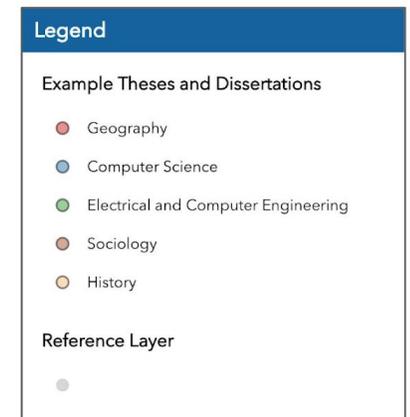
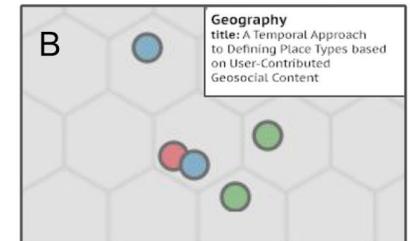
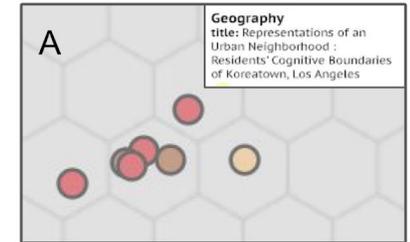
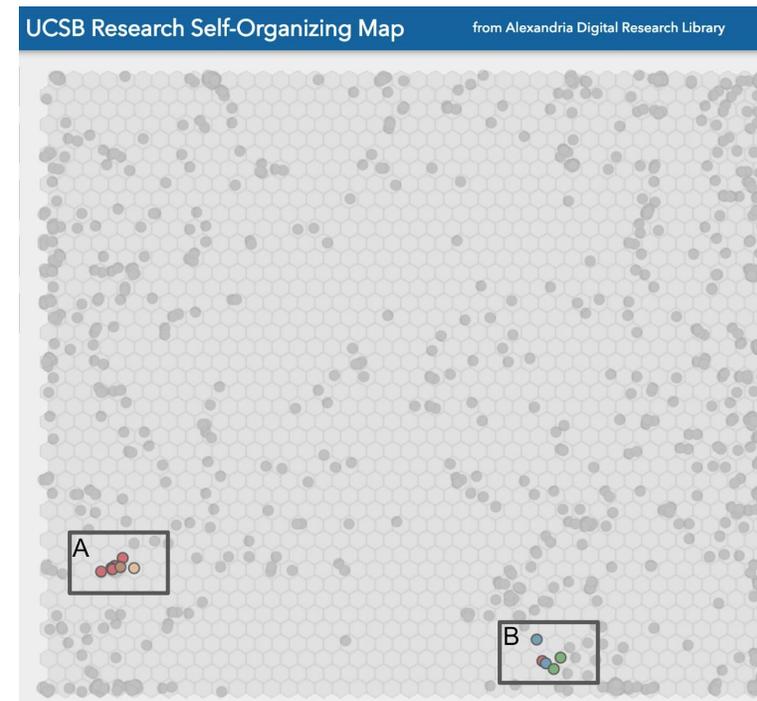
1. Collect metadata for research documents
2. Model topics of document titles and abstracts
- 3. Generate spatializations in field and network spaces**
4. Demonstrate similarity relations based on distance



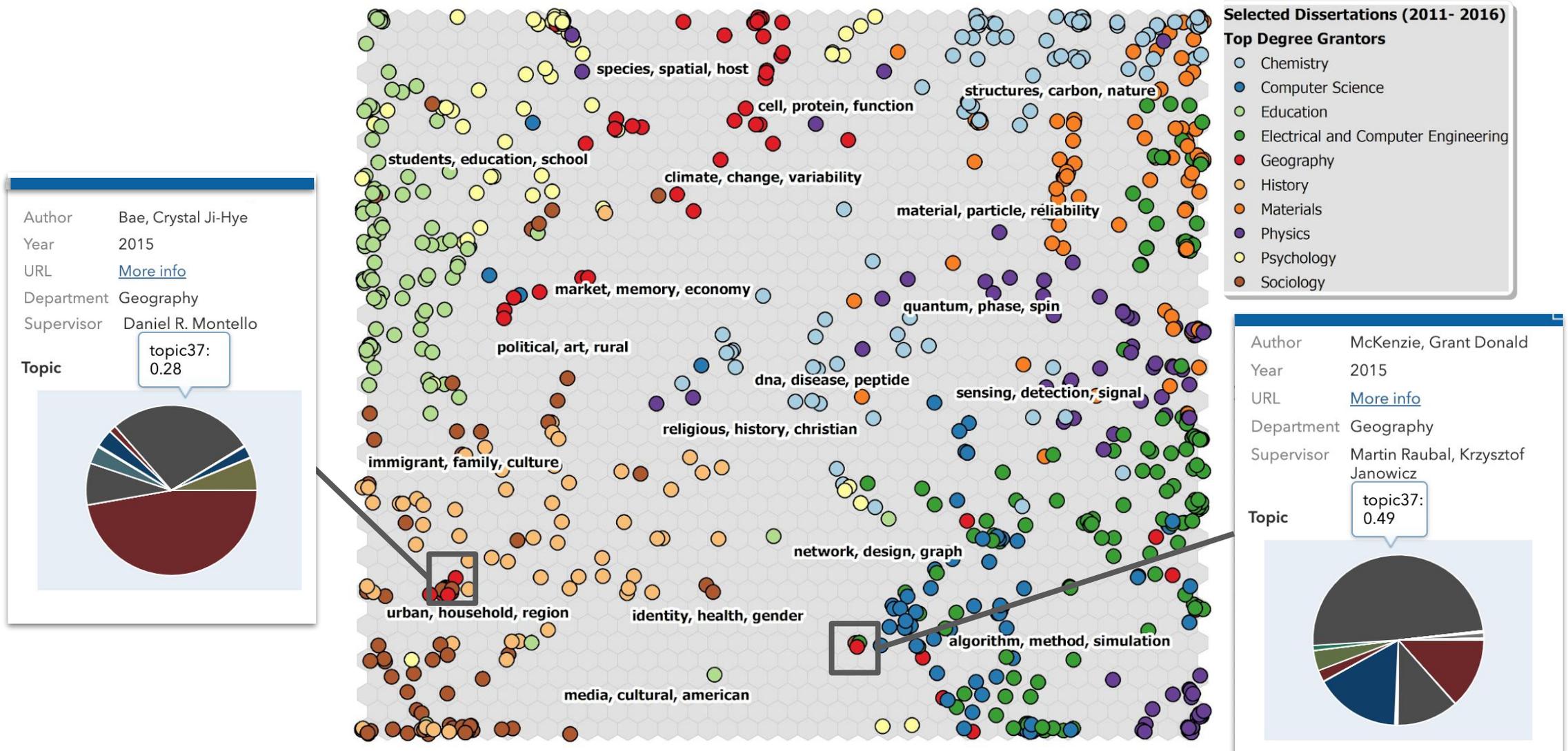
Core concepts of spatial information
(Kuhn, 2012)

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

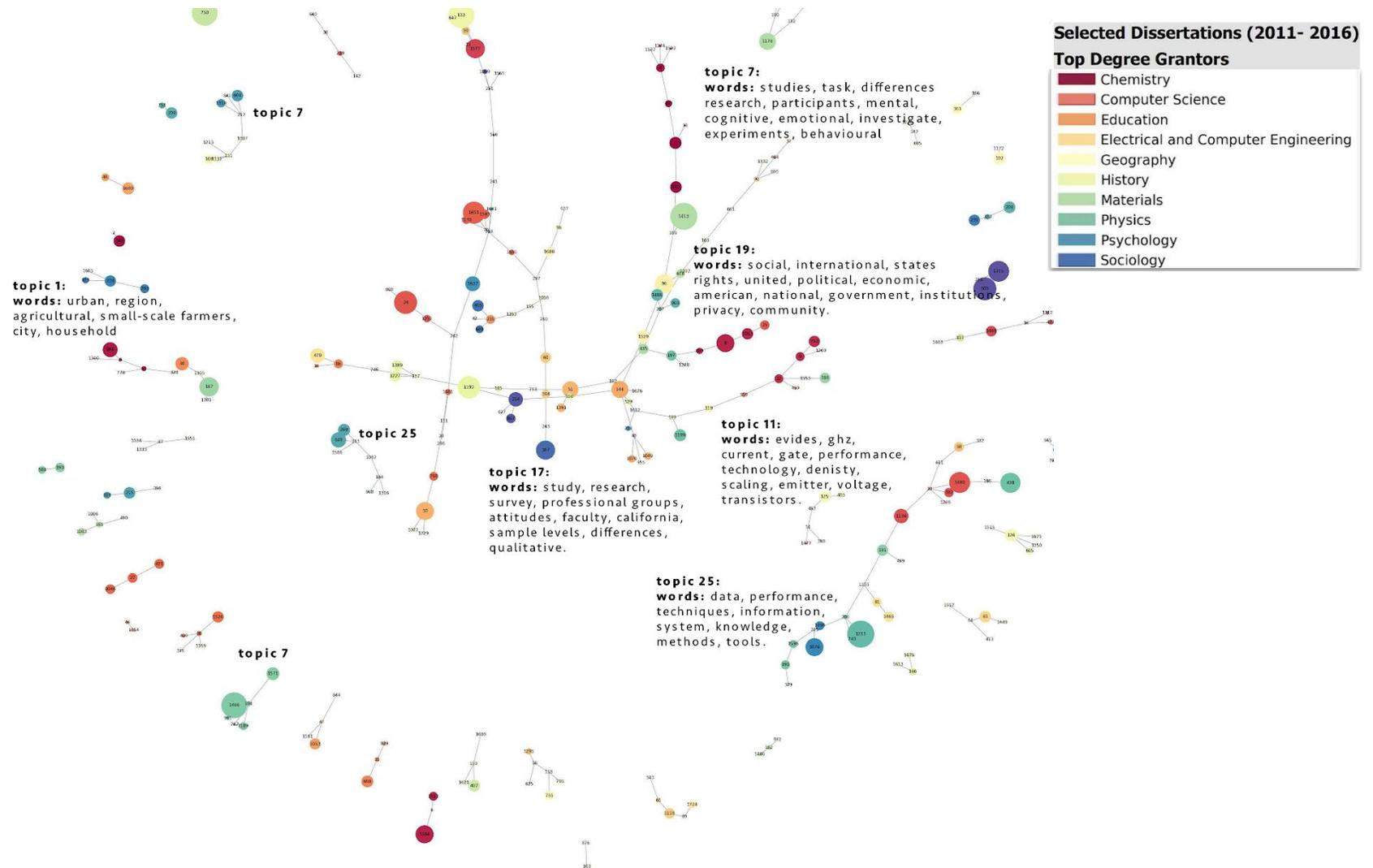
1. Collect metadata for research documents
2. Model topics of document titles and abstracts
3. Generate spatializations in field and network spaces
4. **Demonstrate similarity relations based on distance**

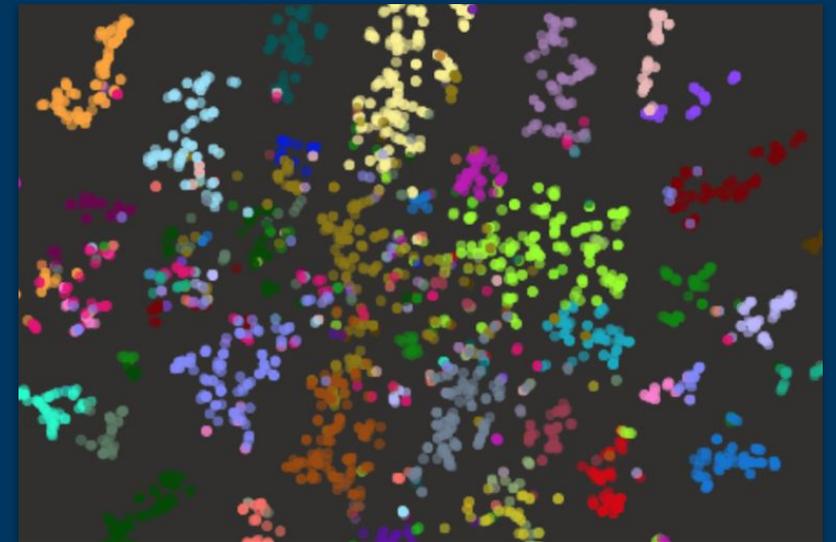
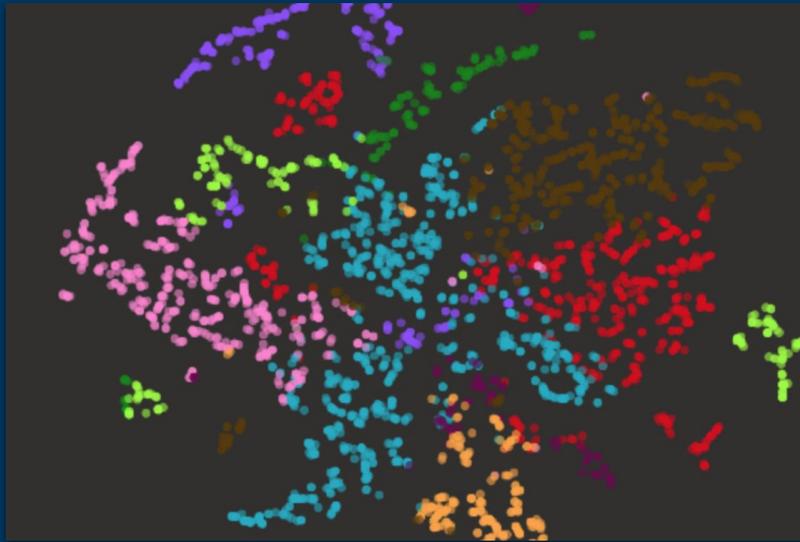
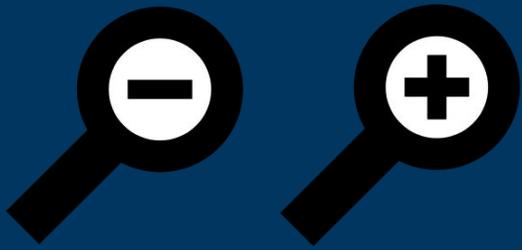


Field of research topics



Network of research topics



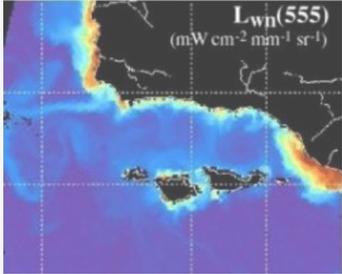
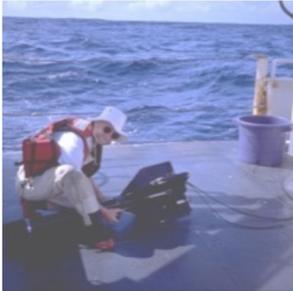


Lafia et al. (2020)

Spatialization → *Generalization*

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

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

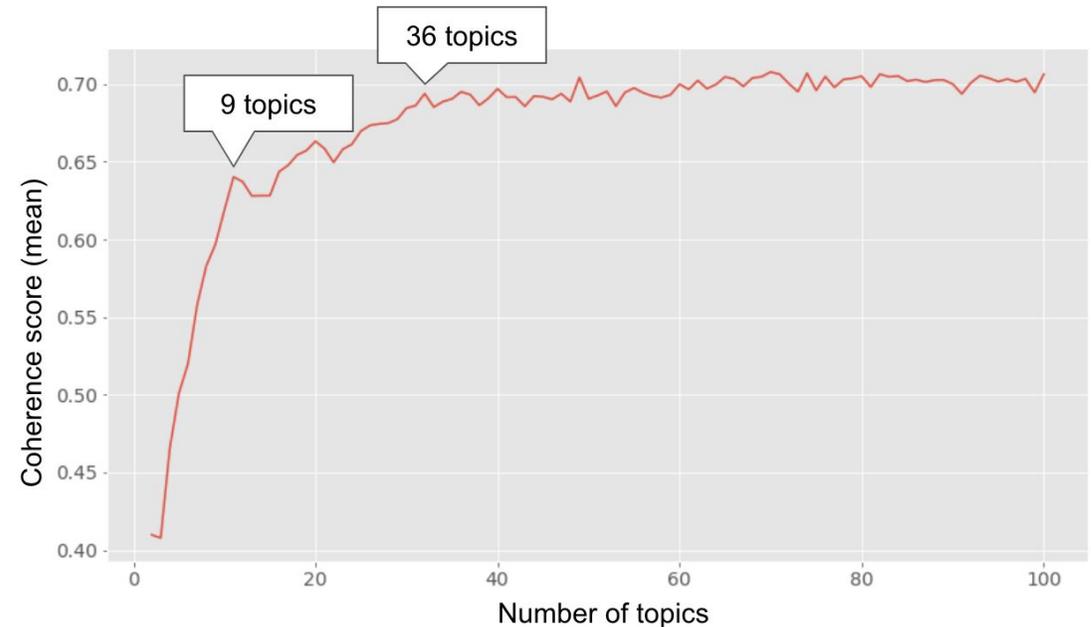
<h3>Plumes and Blooms</h3>  <p>Each year, winter rains wash sand, mud and other terrestrial debris into the Santa Barbara Channel. Then, during the spring and summer, phytoplankton populations increase dramatically and ultimately provide the primary energy source for the entire marine food web. These alternating patterns of... more</p> <p>Tags: Earth Systems Science</p>	<h3>Cheadle Center for Biodiversity & Ecological Restoration</h3> <p>The mission of the Vernon and Mary Cheadle Center for Biodiversity and Ecological Restoration (CCBER) at UCSB is to support:</p>  <p>Education Management Management, Restoration and Conservation</p> <p>Tags: Environmental Information Management</p>
<h3>Bermuda Bio Optics Project</h3> <p>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</p> <p>Tags: Earth Systems Science</p> 	<h3>Snow Hydrology Research Group</h3> <p>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</p> <p>Tags: Earth Evolution, Earth Systems Science, Environmental Information Management, Human Impacts</p> 



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

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

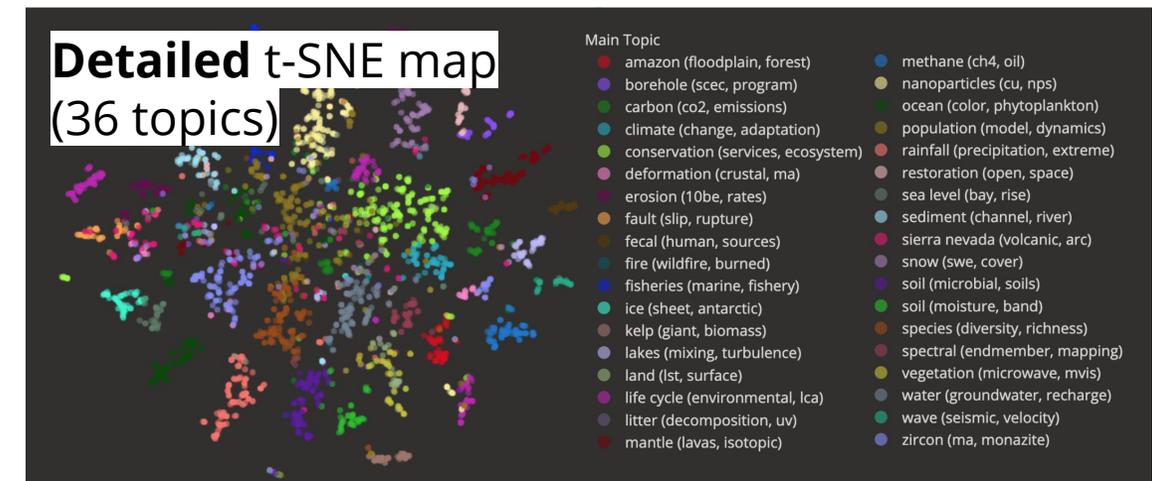
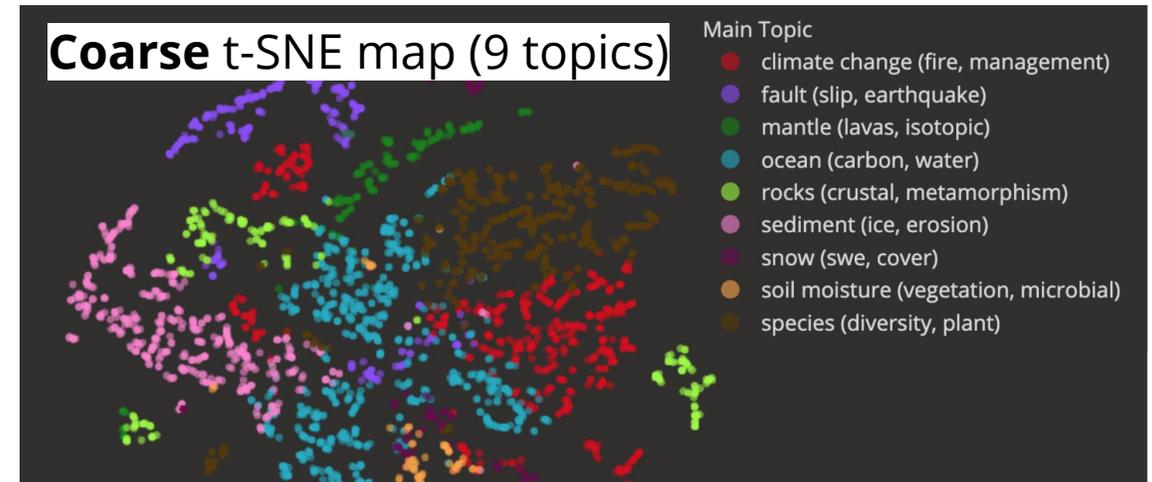
1. Analyze and process document metadata
- 2. Select number of topics to model based on coherence**
3. Spatialize topics at a coarse and a detailed level
4. Deploy a map dashboard and interpret results



Coherence scores for NMF topic models with 2 – 100 topics

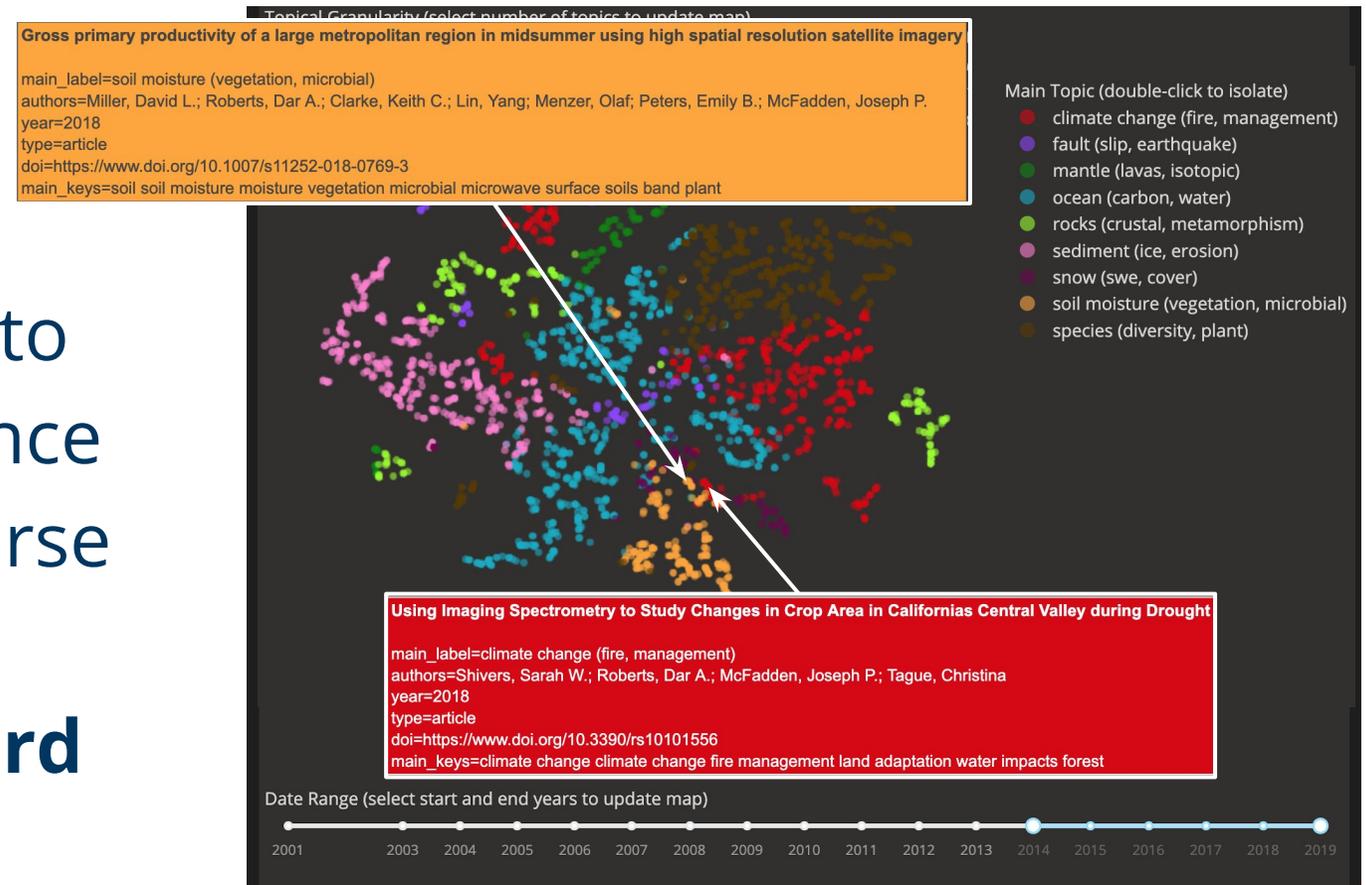
How can we represent topics of a multidisciplinary body of research at multiple levels of detail?

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



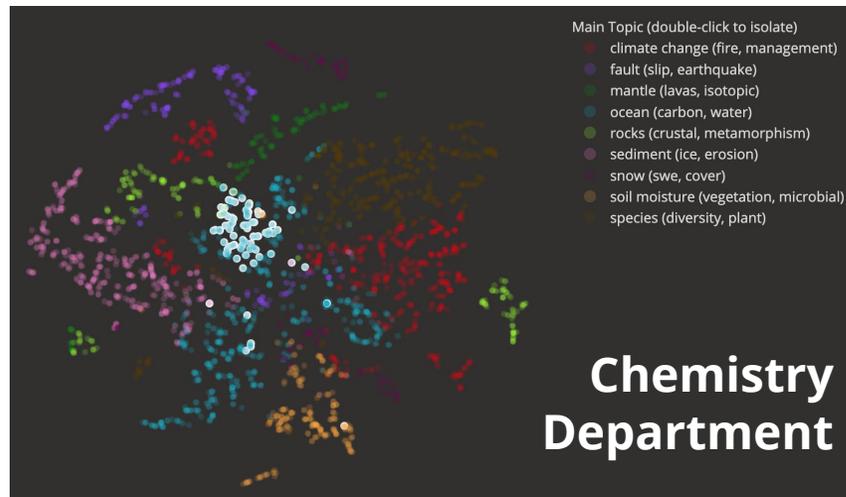
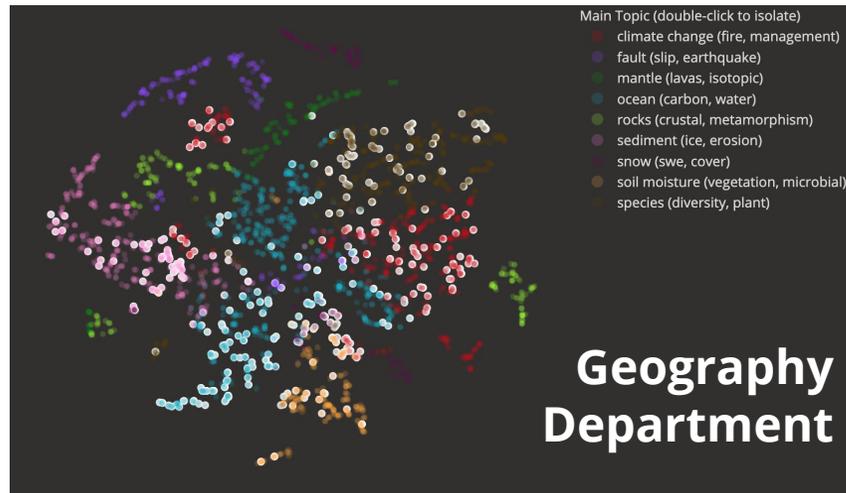
How can we represent topics of a multidisciplinary body of research at multiple levels of detail?

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/>

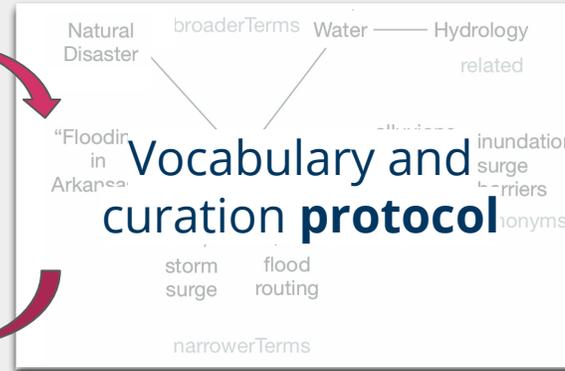
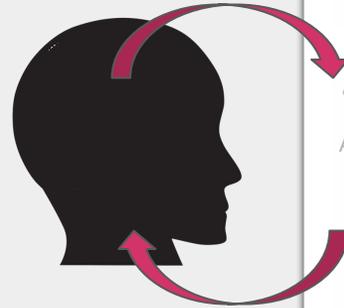
Can these maps support **high-level** views of research at a multidisciplinary institute?



- **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)

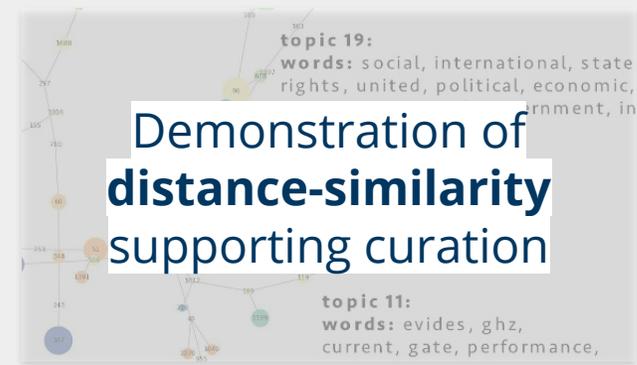
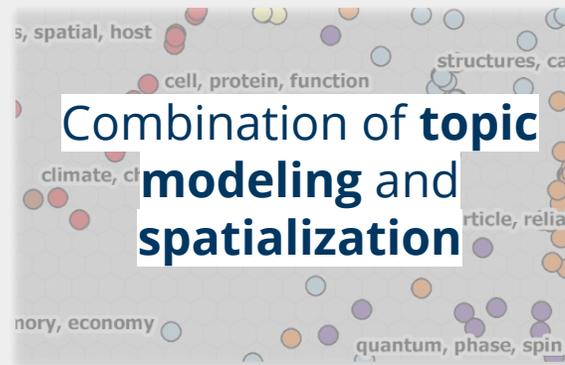
Verbalization (Study 1)

User terms mapped to system terms with a hierarchical vocabulary



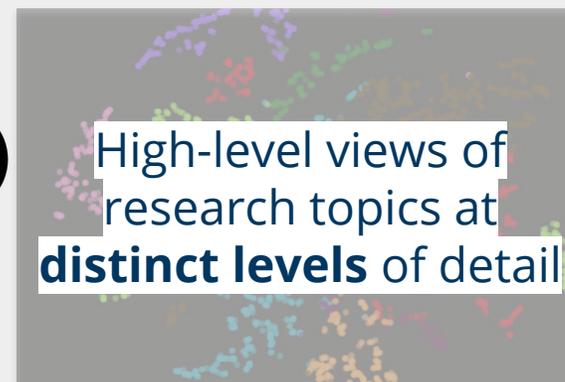
Spatialization (Study 2)

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



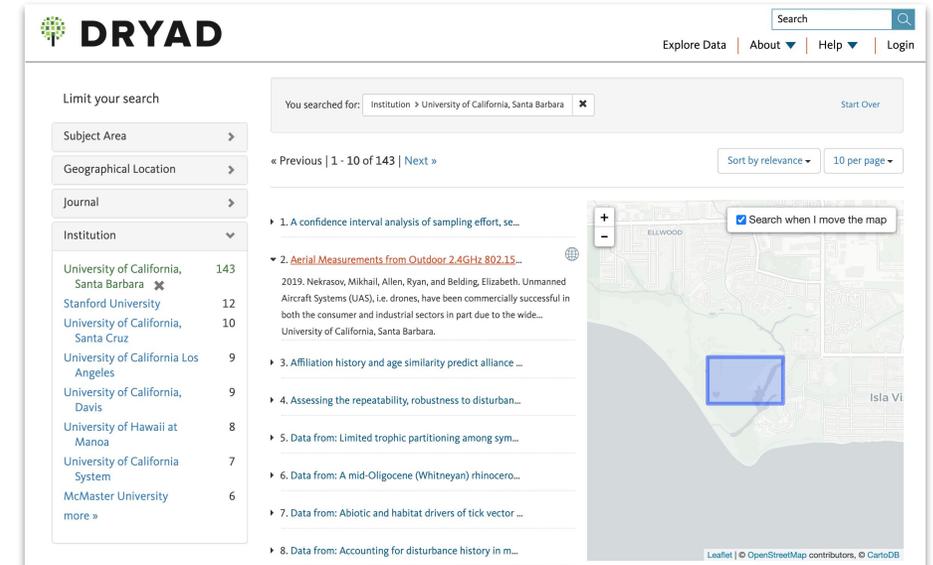
Generalization (Study 3)

Research topics elicited from metadata configured in temporally-sequenced maps



Limitations

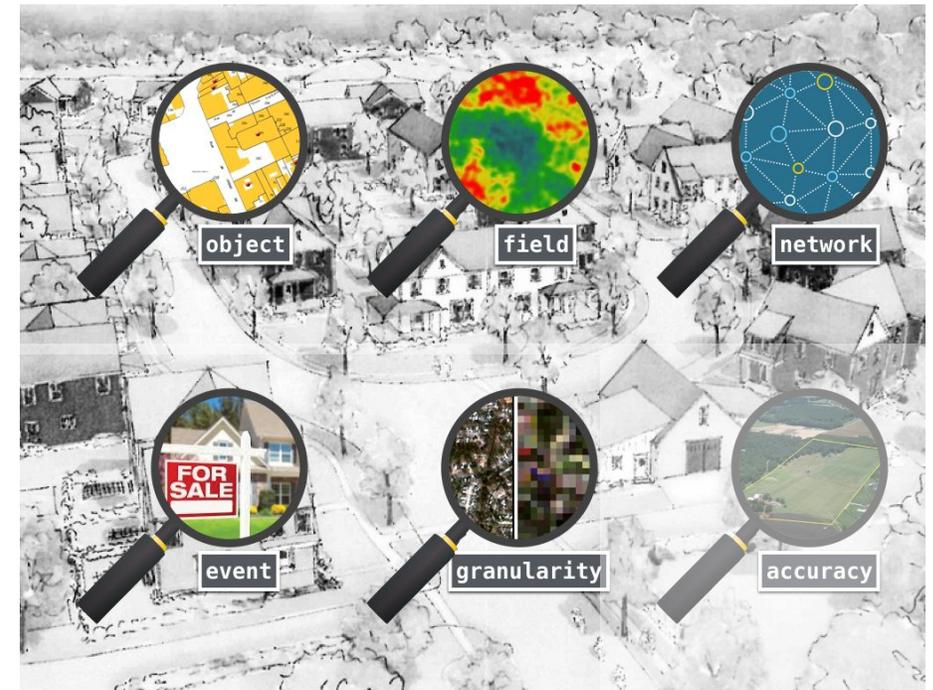
- **Evaluation baseline:** innovating previously unseen solutions
- **Feedback mechanisms:** potential for cross-study tasks and insights
- **Research data proxies:** adoption of data curation policies



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