

UC Santa Barbara

UC Santa Barbara Electronic Theses and Dissertations

Title

Designing for Serendipity: Research Data Curation in Topic Spaces

Permalink

<https://escholarship.org/uc/item/5647q82f>

Author

Lafia, Sara Katherine

Publication Date

2020

Peer reviewed|Thesis/dissertation

University of California
Santa Barbara

Designing for Serendipity: Research Data Curation in Topic Spaces

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Geography

by

Sara Lafia

Committee in charge:

Professor Werner Kuhn, Chair
Professor James Frew
Professor Kelly Caylor
Professor Daniel Montello

September 2020

The Dissertation of Sara Lafia is approved.

Professor James Frew

Professor Kelly Caylor

Professor Daniel Montello

Professor Werner Kuhn, Committee Chair

July 2020

Designing for Serendipity: Research Data Curation in Topic Spaces

Copyright © 2020

by

Sara Lafia

Curriculum Vitæ

Sara Lafia

Education

2020	Ph.D. in Geography, University of California, Santa Barbara.
2017	M.A. in Geography, University of California, Santa Barbara.
2014	B.S. in Urban Planning & B.S. in Geography, California State Polytechnic University, Pomona. Valedictorian.

Fellowships and Scholarships

2019	Mellichamp PhD Research Fellowship: UCSB 21st Century Global Dynamics
2018	Robert Raskin Scholarship: Earth Science Information Partners
2017 - 2018	Jack & Laura Dangermond Graduate Fellowship: UCSB Department of Geography
2012 - 2014	Greater Research Opportunities Fellowship: U.S. Environmental Protection Agency
2012 - 2013	Prete Environmental Education Fellowship: Lyle Center for Regenerative Studies

Awards

2019	Excellence in Teaching Award: UCSB Department of Geography
2019	Womxn/Hacks Hackathon Winner: Best Use of Google Cloud Platform): UCSB
2018	SAP + Esri Spatial Hackathon Finalist: Geo machine learning): Esri DevSummit, Palm Springs, CA

Travel Grants

2014 - 2020	Jack & Laura Dangermond Travel Grants: UCSB Department of Geography
2019	FORCE11 Scholarly Communication Institute Scholarship: UCLA 2019
2017	Graduate Student Synthesis Workshop Invitation: National Socio-Environmental Synthesis Center
2016	DataONE Users Group Meeting Travel Grant: National Center for Ecological Analysis and Synthesis

2014

Google Geo for Higher Education Summit Invitation: Googleplex,
Mountain View, CA

Publications

Refereed Academic Journal Articles

Lafia, S., & Kuhn, W. (2018). Spatial Discovery of Linked Research Datasets and Documents at a Spatially Enabled Research Library. *Journal of Map & Geography Libraries*, 14(1), 21-39.

Coggins, B. L., **Lafia, S.**, & Torghabeh, B. V. (2018). Dramatic Change in North Korea: Instability and Human Flight Propensity. *North Korean Review*, 14(1), 49-70.

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.

Refereed Conference Proceeding Papers

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., 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.

Allen, C., Hervey, T., **Lafia, S.**, Phillips, D. W., Vahedi, B., & Kuhn, W. (2016). Exploring the Notion of Spatial Lenses. In International Conference on Geographic Information Science (pp. 259-274). Springer, Cham.

Short Papers and Workshop Proceeding Papers

Lafia, S., Xiao, J., Kuhn, W., & Hervey, T. (2019). Talk of the Town: Discovering Government Data via Conversational Voice Assistants. In 14th International Conference on Spatial Information Theory (COSIT 2019). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Golubovic, N., Krintz, C., Wolski, R., **Lafia, S.**, Hervey, T., & Kuhn, W. (2016). Extracting Spatial Information from Social Media in Support of Agricultural Management Decisions. In Proceedings of the 10th Workshop on Geographic Information Retrieval (p. 4). ACM.

Lafia, S., & Staehli, L. (2016). From Research Objects to Research Networks: Combining Spatial and Semantic Search. In SDW@ GIScience (pp. 35-40).

News Articles and Reports

Lafia, S., & Kuhn, W. (2019). Enabling the Spatial Discovery of Research Data in Libraries. In US National Report. (US National Committee for the International Cartographic Association).

Seltmann, K., **Lafia, S.,** Paul, D., James, S., Bloom, D., Rios, N., ... & Davis, E. (2018). Georeferencing for Research Use (GRU): An Integrated Geospatial Training Paradigm for Biocollections Researchers and Data Providers. *Research Ideas and Outcomes*, 4, e32449.

Lafia, S. & Kuhn, W. (2017). Opening Access to University Scholarship through Spatial Discovery. *ESRI, ArcUser* (pp. 66 - 67).

Presentations

- | | |
|------|---|
| 2019 | Enabling the Discovery of Thematically Related Research Objects. COSIT: Regensburg, Germany |
| 2019 | Talk of the Town: Discovering Government Data. COSIT: Regensburg, Germany |
| 2019 | Spatially Supported Thematic Discovery of Research Data. COSIT: Regensburg, Germany |
| 2019 | Re-creating and Exploring Research Neighborhoods. FORCE11 Summer Institute: UCLA |
| 2018 | Spatial Data Discovery at UCSB: Future Directions. Library Open Access Week: UCSB |
| 2018 | Improving Discovery of Open Civic Data. GIScience: Melbourne, Australia |
| 2018 | Discovering Spatial Data for Our Changing Earth. ESIP: Tucson, Arizona |
| 2017 | Discovering and Sharing Campus Scholarship. Center for Spatial Studies Lightning Talks: UCSB |
| 2017 | Linking Geographic Footprints to Topic Spaces. Graduate Slam Competition: UCSB |
| 2016 | From Research Objects to Research Networks. WAML: Stanford University |
| 2016 | Spatial Discovery and the University Library. TGIS Session at ESRI User Conference: San Diego, CA |
| 2016 | People and Places: Resource Discovery in a Library Context. AAG: San Francisco, CA |
| 2016 | Degrees of Separation: Measuring Musicians and Places. Center for Spatial Studies: UCSB |

2017	Research Trajectories: Topic Models of UCSB Graduate Research. MAT EoYS: Santa Barbara, CA
2015	Computing with Core Concepts of Spatial Information. COSIT: Santa Fe, New Mexico

Research Experience

2016 - present	Graduate Student Researcher: UCSB Center for Spatial Studies
2019	Research Assistant for WhatEvery1Says (WE1S) with Dr. Alan Liu: UCSB Department of English
2019	Research Assistant for global-e with Dr. Michael Curtin: UCSB Department of Film and Media Studies
2017	Engineering Intern: Esri Research & Development, Arlington, VA
2017	Research Assistant for Dr. Bridget Coggins: UCSB Department of Political Science
2016	Informatics Intern: NCEAS, Santa Barbara, CA
2013 - 2014	NASA DEVELOP Student Contractor: Jet Propulsion Laboratory, Pasadena, CA
2012	NSF-REU Summer Research Assistant: University of Nevada, Reno, NV

Teaching and Mentoring Experience

2015; 2017; 2019	Research Mentor for Undergraduate Students: Center for Spatial Studies, UCSB
2017 - 2018	Teaching Associate (Instructor of Record): Department of Geography, UCSB
2016	Workshop Instructor with Dr. Katja Seltsmann: iDigBio Georeferencing Workshop, NCEAS
2015	Freshman Discovery Seminar Co-Instructor with Dr. Chandra Krintz: UCSB
2014 - 2018	Teaching Assistant: Department of Geography, UCSB

Academic Service

2017; 2019	Faculty Search Committee Graduate Representative: Environmental Data Science & GIS, UCSB
2015; 2017; 2019	Meeting Co-organizer: Spatial Discovery Expert Meetings, Center for Spatial Studies

2018 - 2019	Graduate Student Association Representative: Associated Student Government, UCSB
2018	Workshop Co-organizer: Core Computations on Spatial Information at GIScience in Melbourne, Australia
2017 - 2018	Graduate Student Representative: Associated Students Campus Bicycle Committee, UCSB
2018; 2019	Manuscript Reviewer: Journal of Spatial Information Science; International Journal of Digital Earth
2017	GIS Help Desk Consultant: Center for Spatial Studies, UCSB

Certifications

2020	Certificate in College and University Teaching (CCUT)
2020	Information Technology & Society Ph.D. Emphasis (CITS)
2017	Summer Teaching Institute for Associates (STIA)

Software

- Spatial: ArcGIS, ArcGIS Server, QGIS, GDAL
- Other: Protégé, Processing, Git

Languages

- Java
- Python
- HTML5, CSS, Javascript

Affiliations

- Association of American Geographers (AAG)
- Association for Computing Machinery (ACM)
- Association for Information Science and Technology (ASIS&T)
- Cartography and Geographic Information Society (CaGIS)

Abstract

Designing for Serendipity: Research Data Curation in Topic Spaces

by

Sara Lafia

Researchers seeking relevant data across disciplines confront the challenge of navigating technical descriptions. How can curation support the serendipitous discovery of related research data? Everyday spaces like bookshelves are designed to support browsing and exploration by placing similar resources closer together. Space and time are foundational ordering relations for knowledge organization. I ask how this ordering, which is well-established in the geographic context, can be translated to locate and organize research data in abstract topic spaces. This dissertation develops methods for making the latent topics of research metadata explicit. These methods produce spatial configurations where related research topics are co-located in neighborhoods. This has the potential to support serendipitous discovery by offering researchers ways to discover related data. I test this notion in three studies that develop topic spaces for research data curation. The first part of this dissertation in Chapter 2 focuses on supporting research data discovery with a common terminology. I develop a crosscutting base vocabulary of geospatial topics to help users discover related government data in a ubiquitous open civic data platform. Semantic annotation expands search terms by mapping users' vernacular onto the language of metadata. In the second part of this dissertation, I shift away from addressing terminological search to supporting spatial curation by developing topic spaces. In Chapter 3, I develop two kinds of topic spaces for curating research theses and dissertations: landscapes and networks. I use topic modeling to determine the latent semantic similarity of research metadata and then produce topic spaces from these using spatialization

techniques. In Chapter 4, I spatialize an institute’s multidisciplinary body of research, producing topic maps at two distinct levels of detail. Emerging spatial patterns, like centrality and proximity, support high-level narratives about cross-disciplinary research activities that complement the quantitative metrics currently cited in reviews of institutional research. Together, these three studies demonstrate strategies for developing topic spaces in which diverse, yet related, multidisciplinary research data are curated. Future research will extend these methods by tracing the impact of specific curatorial actions contributing to research data discovery and reuse.

Contents

Curriculum Vitae	iv
Abstract	ix
1 Introduction	1
1.1 Synopsis	1
1.2 Background and Motivation	5
1.3 Research Questions and Methods	13
1.4 Permissions and Attributions	17
2 Improving Discovery of Open Civic Data	20
2.1 Introduction	21
2.2 Background	23
2.3 Challenges and Approaches	25
2.4 Methods	29
2.5 Results	38
2.6 Discussion and Outlook	42
3 Enabling the Discovery of Thematically Related Research Objects	45
3.1 Introduction	46
3.2 Enabling Research Discovery Across Disciplines	49
3.3 Conceptual Approach: Making the Choices of Spaces Systematic	50
3.4 Technical Approach: Implementing Field and Network Spatializations	55
3.5 Application: Discovering Thematically Related Research	60
3.6 Conclusion	65
4 Mapping Research Topics at Multiple Levels of Detail	68
4.1 Introduction and Motivation	69
4.2 Background	72
4.3 Methods	77
4.4 Results	86
4.5 Evaluation	92

4.6	Conclusions	100
5	Conclusions and Future Work	103
5.1	Summary and Discussion	103
5.2	Contributions	108
5.3	Limitations	109
5.4	Open Questions	110
	Bibliography	113

Chapter 1

Introduction

1.1 Synopsis

This dissertation is comprised of three research articles that support data curation. The first part of this dissertation develops a high-level, common terminology to support data description. The second part of this dissertation takes a spatial rather than verbal approach, developing configurations of research topics that can be generalized and presented at distinct levels of detail. Together, these approaches strive to support the serendipitous discovery of research data.

In this dissertation, “research data” are broadly interpreted to include research documents [1], objects [2], and metadata [3] used or generated by researchers. This includes spatial government datasets, graduate theses, and descriptions of funded research projects. This dissertation contributes a series of methods that make the latent topics of research data explicit and configure them into curated topic spaces. The bookshelf, which organizes books into curated neighborhoods of topics, serves as a design analogy for this approach. Curation is a process of long term care that involves organizing data to maximize meaningful and enduring access to them [4]. Thus, data curation directly influ-

ences search, including browsing, exploration, and serendipitous discovery [5]. I ask how curation can support the discovery of related research data without requiring knowledge of disciplinary terms. This question relates to the misalignment between human concepts and system terms, called the vocabulary problem in human system communication [6], which presents major challenges to identifying related research topics in multidisciplinary settings like open data portals, university research libraries, and research institutes.

In the first part of this dissertation (Chapter 1), I introduce foundational concepts that motivate and inform this work. I discuss information seeking behavior, specifically distinguishing exploration from lookup paradigms. I also review research data curation practices and initiatives from government and research library perspectives. Finally, I review existing information visualization and spatialization applications and discuss how they address challenges related to “accessing task-appropriate scientific data” [7].

The first research article in this dissertation (Chapter 2) focuses on improving the discovery of open civic data. Open data initiatives enable the public to access spatial data that are vital for governance, but these initiatives do not guarantee data discoverability. Data descriptions vary greatly and can be highly technical, making it difficult for people to find and use data. This study addresses the mismatch between users’ colloquial search terms and descriptions of civic data in an open geoportal. I construct a crosscutting base vocabulary that describes core geospatial topics of open civic data and extend it by including broader, narrower, and synonymous terms. I put the extended vocabulary into production as a curation protocol, semantically annotating existing metadata of geographic civic data available in several open data portal instances from Esri’s ArcGIS Hub. This results in a mapping between users’ colloquial search terms and metadata descriptions. Data that were initially difficult to find with a given set of keywords become discoverable with expanded keywords, allowing users to search for general topics of concern and discover relevant, authoritative data across levels of government.

The second research article (Chapter 3) addresses the challenge of identifying related research data in the multidisciplinary setting of a university library repository. It remains difficult to identify research addressing similar topics, including subjects and methods, from different disciplines since commonalities in their descriptions are latent. Instead of mapping user terms to system terms with a hierarchical vocabulary, I move toward a geometric approach. This is informed by Gärdenfors’ conceptual spaces [8], which model similarity relations in knowledge representations with distances. I develop two kinds of abstract topic spaces from research theses and dissertations hosted by the UCSB Library: a field or landscape, implemented as a self-organizing map [9], and a network. These spatializations exploit similarity relations determined with topic modeling to co-locate and arrange related topics into neighborhoods. To implement the spatializations, I extend the geographic capabilities of ArcGIS Online by mapping topics to 2-dimensional field and network spaces that convey the similarity of theses and dissertations by distances. This results in representations of research data in each space that follow the first law of cognitive geography [10] in which similar research topics are found closer together. This spatial approach to locating and organizing the topics of research data has the potential to increase awareness of related research activities across academic disciplines.

The third research article (Chapter 4) extends and evaluates the spatialization approach established in the previous study to support an institutional review of UCSB’s Earth Research Institute (ERI). The spatializations produced in this study explicitly incorporate the notion of granularity, resulting in structured, multi-level maps of research topics that enable reviewers to “read” a body of research composed of publications and projects “at a distance” [11]. The topic modeling and spatialization approaches developed in this study improve upon design decisions made in the previous study; topic modeling with non-negative matrix factorization [12], rather than Latent Dirichlet allocation [13], produces more coherent inputs to spatialization. Furthermore, mapping the

research documents with t-distributed stochastic neighbor embedding [14], rather than self-organizing maps [9], supports the systematic development of reproducible, multi-level topic maps. Cross-disciplinary topics emerge at each level of detail, revealing the latent thematic structure of ERI's body of research. Additionally, time is incorporated into the maps, supporting the detection of events like the acquisition of funding or addition of researcher expertise. Researchers affiliated with the institute evaluate the topic clusters, or neighborhoods, of related research documents presented in maps at two distinct levels of detail. The maps of research topics are found to strengthen reviews by exposing thematic expertise, relationships between researchers, topical distributions and clusters of work, and the evolution of these aspects over time. The relationships among the institute's researchers and departments highlighted in the topic maps support the development of qualitative narratives and complement existing institutional review metrics.

Figure 1.1 illustrates the three research article chapters in this dissertation.

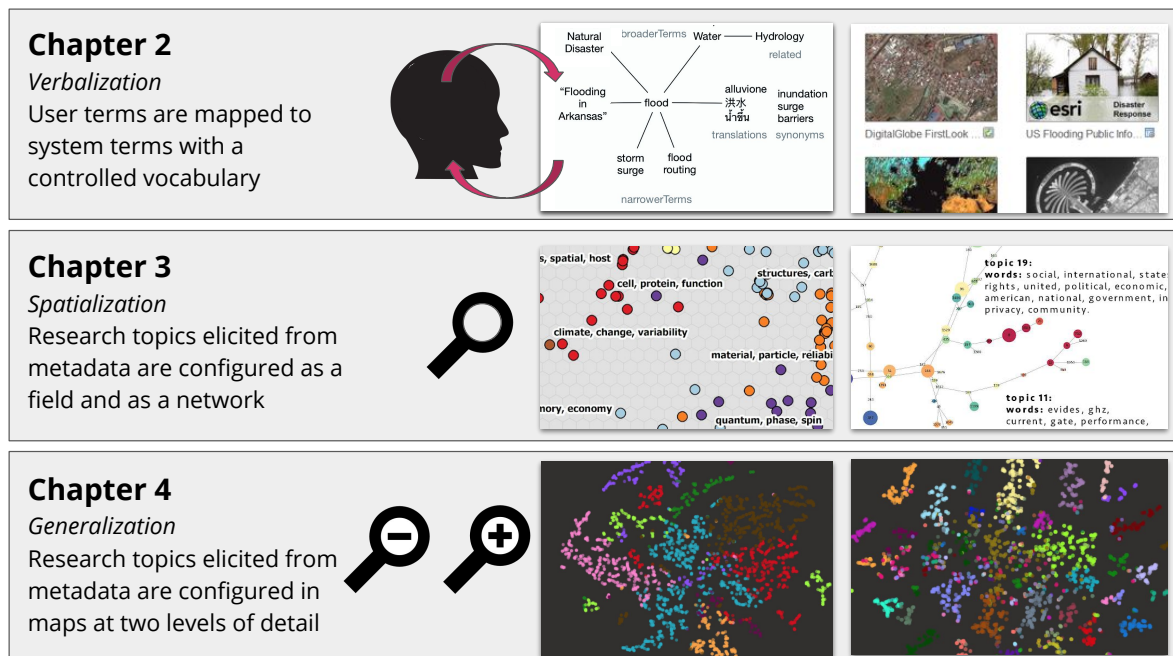


Figure 1.1: Overview of methods in dissertation chapters

The final part of this dissertation (Chapter 5) summarizes key findings. It shows how the first and second research articles provide contrasting approaches to addressing the vocabulary problem [6]. In Chapter 2, term expansion guides the mappings from the user to a system, supporting *verbalization*. Chapter 3 instead takes a *spatialization* approach, eliciting topics from research metadata and presenting them to the user in two kinds of spatial configurations: a landscape, or field, and a network. Chapter 3 also develops a method for research data curation that combines topic modeling and spatialization. This method is extended and evaluated in Chapter 4, which applies the notion of “distant reading” from the Digital Humanities to support the *generalization* and mapping of research topics at two levels of detail, revealing distinct patterns of research activity. Together, these research articles demonstrate spatial strategies for leveraging the latent topics in research data to support curation and discovery.

1.2 Background and Motivation

Various libraries, repositories, and archives organize their resources geographically to serve targeted user groups. Projects, like the Alexandria Digital Library, have conducted user studies, for example surveying the needs and perceptions of earth scientists, information specialists, and educators; a combination of online surveys and ethnographic studies informed system improvements [15]. Related approaches, such as the GeoBlacklight initiative develop personas, which consider the distinct research needs of users including professors, Ph.D. students, librarians, and GIS instructors [16]. Resources that are *geographically* discoverable reference the spatial domain consisting of “the Earth’s surface and near-surface” [17]. Digital library projects, including the Alexandria Digital Library¹

¹<http://legacy.alexandria.ucsb.edu/>

and GeoBlacklight², along with data repositories, like DataONE³ and Dryad⁴, have enabled geographic browsing and exploration by delineating the footprints of resources; a common strategy is to assign resources spatiotemporal bounding boxes exposed in geographic maps [18]. This allows researchers seeking resources about particular parts of the world to discover geographically co-located resources across providers.

I recently collaborated with the UCSB Library to experiment with enabling geographic search for research data, including species counts and archaeological surveys, hosted across institutional repositories and lab servers [19]. I used ArcGIS Online to georeference and expose traditional spatial data, like shapefiles and imagery, as well as non-spatial data, like publications and reports, by assigning explicit geographic footprints to the data and exposing them through an open data portal. Their metadata were also transformed into linked data using existing geospatial vocabularies like GeoLink⁵, resulting in a protocol for describing spatial research data and linking them to non-spatial data [20]. These approaches improved awareness of ongoing research across campus by geographically co-locating data across contributors.

While the challenge of curating research data by geographic location has been largely addressed, this idea has not yet been extended to support research data curation by topic “location”. This dissertation aims to inform the design of maps that use research topics to curate research data for discovery. Rather than making research data *geographically* discoverable through georeferencing however, I focus on making research data *spatially* discoverable, referring to any spatiotemporal frame at any spatial resolution, and including non-Cartesian spaces [17].

A familiar example of thematic co-location of resources is the bookshelf, where similar

²<https://github.com/geoblacklight>

³<https://www.dataone.org/>

⁴<https://datadryad.org/stash>

⁵<http://schema.geolink.org/1.0/base/main.html>

books are grouped, for example by Dewey Decimal Classification. This is an example of a design choice that supports browsing and exploration by organizing thematically related content to make patterns and relationships apparent. Powerful, spatial search is not yet widely supported for research data [21]. This is a missed opportunity, particularly in a multidisciplinary context where it is often difficult to identify related research across academic disciplines. This is due in part to the fact that research data are often published in disciplinary silos; for example, ecological data are frequently published in the Knowledge Network for Biocomplexity⁶, while terrain data are often shared through OpenTopography⁷. Researchers who work across these siloes must know the many places where they can look for research data and how they should specify their search terms when they look for research data in each context.

Data providers, including government portals and library repositories, stand to benefit from making the latent similarity of diverse research topics apparent to users seeking data for research. Organizational affordances that are well-known from geography can be implemented in abstract spaces to impose a useful spatial structure on research data. This dissertation does this by spatially curating research data across multidisciplinary contexts, conveying the relatedness of topics through spatial structures, like topic clusters forming neighborhoods or sharing connectivity. The spatial organization of research topics enables users to explore their similarity, or dissimilarity, by their proximity.

Spatialization, which transforms high-dimensional data into lower-dimensional, geometric representations on the basis of similarity [10] underpins thematic organization in this dissertation. Spatialization is used to re-create conceptual neighborhoods, as bookshelves do. While bookshelves provide one example of thematic neighborhoods that support discovery, they are not easily reconfigured. They also require users to browse in

⁶<https://knb.ecoinformatics.org/>

⁷<https://www.opentopography.org/>

a particular category, even though most information items could be classified in many possible ways, limiting physical items to one physical location. Spatialized interfaces can be far more flexible than library bookshelves by indexing items in multiple ways, offering multi-dimensional possibilities to support views of research data [22].

Like georeferencing, spatialization provides a foundation for curating research data; rather than assigning geographic locations to research data, they can be located in an abstract space. Tobler’s first law of geography posits that in geographic space, everything is related to everything else, but nearby things are more related than distant things [23]. Previous work with the UCSB Library showed that organizing data by geographic location enabled exploration and discovery, as spatial relations like proximity, overlap, and neighborhood help users discover data about similar parts of the world [19]. Spatialization extends this idea with the distance-similarity metaphor, in which nearby elements are considered similar; this has been applied to organize large, thematically diverse, and multidimensional collections [10]. Thus, spatialization offers an approach for co-locating related research topics in abstract spaces.

The following sections describe the foundation for this dissertation, which aims to support curation by spatially organizing the topics of research data. This discussion of current practices and limitations in research data curation from the perspectives of institutions, users, and systems also situates the contributions of this dissertation.

1.2.1 Research Data Curation

*How are **institutions** curating research data? What limitations do they face?*

Academic libraries increasingly manage campus research data repositories. The domain-neutral identity of academic libraries makes them an ideal “meeting place” for researchers and research data from many disciplines. Libraries are well-positioned to support research

data curation, including the creation of formal metadata; curation models, like the UK Digital Curation Centre’s Lifecycle model, ensure data discovery, access, and use in the present and in perpetuity [18].

Subject librarians and researchers often work together to create high-quality metadata. The identification of a single set of metadata elements to ensure research data discovery and reuse remains an unanswered question; metadata elements tend to vary across communities of practice. Despite this variation, metadata themselves are increasingly viewed as valuable data and “first-class citizens” for analysis [2]. A push to improve the description of research data, for example by annotating research metadata and workflows with ontologies, has in some cases resulted in semantically rich research objects; these are aggregations of resources linked to other data by digital associations [2]. Semantically enabled research objects can be thought of as extended documents; just as indexing objects gives them a documentary status [1], making them persistently available on the web makes it possible to cite objects [24].

Some government data portals [25], smart city applications [26], and publisher clearinghouses [27] actualize this vision by publishing ontologies alongside data descriptions that aid in interpretation and reuse; however, semantic annotation still is not widely adopted. Data descriptions, including abstracts and tags, are frequently produced without controlled vocabularies, contributing to the vocabulary problem in human-system communication [6]. A remaining core challenge is subject indexing, which involves adequately describing the *aboutness* of data for discovery and reuse [28]. Along with “location” and “time”, “theme” is one of three key dimensions of spatial information [29] and is the attribute that reflects *aboutness*. In my research, I do not seek to define the *aboutness* for research data, but rather, I leverage existing data descriptions, either by semantically annotating (enriching) or topic modeling (decomposing) them.

1.2.2 Exploration Paradigms

*How do **users** seek research data? What challenges do they face?*

To understand how users look for information, it is important to consider the variety of tasks that they want to perform as well as the design of interfaces that enable lookup [22]. Lookup tasks can be distinguished into *information lookup* and *exploratory search* tasks for learning and investigating; in particular, investigation is a high recall task, requiring users to find “a large proportion of relevant information available”, which describes many academic research scenarios [22]. Browsing differs in the ways that users interact with information structures and interfaces. Methods for organizing search results include *faceted categories* corresponding to a relevant dimension or feature type and *clustering* results according to some measure of similarity [22]. Such methods organize results into meaningful groupings that help users understand and act.

While researchers increasingly turn to convenient search engines offering keyword-based omniboxes [30], this style of searching doesn’t help users understand what data exist or what search terms are available to refine or improve their queries. These “bibliographic objectives” are well-understood in library systems, which are designed with the following scenarios in mind: 1) *finding* assumes a search for a known document; 2) *collocating* assumes a search for similar types of information; 3) *choice* assumes making a selection among a number of similar documents; 4) *acquisition* assumes obtaining access to an entity; and 5) *navigation* assumes that a user can explore entities through generalization, association, aggregation, equivalence, association, and hierarchy [31].

The diversity of these bibliographic objectives show that searching for scholarly information like research data often necessitates exploration and browsing, which are not currently supported by generic search engines. Alternatively, the academic digital library can play a unique role in supporting users searching for research data. Digital libraries

should be places where users can: 1) *discover*, by combining online and human sources; 2) *retrieve*, by posing initially vague questions and iteratively reformulating queries against SDIs and federated web catalogs; 3) *interpret*, by summarizing, clustering, ranking, visualizing, and performing statistical analysis on results; 4) *manage*, by drawing from both shared and local data; and 5) *share*, by communicating across domains, in collaboration with colleagues [32]. I seek to enable these user-centric bibliographic objectives by curating topic spaces that improve the discoverability of research data.

1.2.3 Information Visualization and Spatialization

*How do **systems** support data discovery? What are their limitations?*

Well-designed systems are responsive to users' preferred categorizations; in library systems, Dewey Decimal Classification has been broadly adopted due to its success in supporting the bibliographic objectives discussed in Section 1.2.2. Books displayed on a library shelf form thematically cohesive neighborhoods of material that can be useful to explore when conducting research. By contrast, in geographic information systems (GIS), users often struggle to find appropriate tools to support research tasks; toolboxes tend to be organized in non-intuitive hierarchies, which can impede search and discovery [33]. In general, hierarchical classification schemes often do not align with intuitive semantic conceptualizations of relationships [8]. Rather, people are better able to understand organization visually, where relations like similarity are conveyed through distance [34]. For example, viewers of a spatialized display of documents from a large digital library intuitively understood relationships among them, like similarity, based on their proximity [35]. Such displays, or *spatializations*, are intended to exploit people's familiarity with space in everyday life to produce intuitive information spaces.

Vision is a powerful modality of interaction; information visualizations externalize and

amplify cognition [36]. Spatializations are a special class of visualizations that can be treated as if they represent spatial information [34]. Spatialization reduces the cognitive load of system users by abstracting great amounts of detail into visually perceptible spatial patterns [35]. It does this by mapping physical space to abstract domains through spatial metaphors [37] and by transforming high-dimensional attribute data into lower-dimensional, geometric representations on the basis of similarity [10]. As a technique, it has been applied to visually convey topics latent in knowledge domains. Examples of this include the visualization of topics from conference abstracts [38] and the visualization of co-citation networks from large databases of publications [39]. Spatializations have also been empirically evaluated, supporting the idea that users perceive groups of related items shown in point [40], network [41], and region displays [42] to be similar. These studies also revealed that the distance-similarity metaphor works differently in each kind of space. In point displays, users generally perceived points that were closer together to be more similar unless they formed clusters or lines [40]. In networks, similarity judgements followed network links rather than other spatial distance measures [41]. Finally, regions greatly modified distance-similarity judgements, as region membership superseded other spatial distance measures [42].

More recently, cartographically-informed spatializations have been shown to improve data discovery in data-rich environments [43]. When interpreted as cartographic maps, spatializations can reveal spatial relations like proximity, overlap, neighborhood, and presence or absence [44]. These relations convey the similarity or dissimilarity of information and highlight outliers. A relevant example of a cartographically-sound spatialization applied in a library context is an “information landscape” of a digital library’s holdings [45]. The spatialized “terrain” allows users to: 1) explore and navigate the landscape, 2) gain an overview of the collection, and 3) discover new relationships based on the relative positions of content [45]. This system supports browsing, offering users

an important alternative to “known-item searches”. Spatialization has the potential to improve the discoverability of data curated by academic libraries and other related institutions by organizing and presenting similar resources. The intuitive ideas that underlie spatialization, like the distance-similarity metaphor, are expected to make it easier for users to browse for, and discover, related research data in spatialized displays.

1.3 Research Questions and Methods

How can researchers be made aware of related data without needing to know disciplinary terms? This dissertation addresses this question by operationalizing the distance-similarity metaphor to co-locate related topics from research data. The first study in this dissertation explores approaches for expanding topics in data descriptions:

- **RQ1.** How can topics of interest, expressed in users’ ways of talking, be mapped onto the language of metadata? (Chapter 2)

Users of civic data often search for information using colloquial terminology, while organizations creating metadata primarily rely on technical vocabularies that reflect scientific or organizational hierarchies. As a consequence, the thematic categories, terms, geo-references, or attribute schemas of official civic metadata often do not align with users’ search terms. For example, citizens searching for data about bicycle collisions in a given area are unlikely to use the same search terms that organizations, like department of transportation agencies, use to describe relevant data. Users may also search with broad terms, such as traffic safety, and will then not discover data tagged with narrower official terms, such as vehicular crash. Augmenting official metadata with semantic annotations has been found to expand the discoverability and relevance recommendations of data, supporting more inclusive access to them. Thus, a preliminary step toward improving

users' ability to find thematically relevant data is to expand their search terms.

I address this question in the context of integrated networks of open geospatial data, focusing on increasing the discoverability of civic data available through a novel open access collaboration platform: Esri's ArcGIS Hub. To expand users' search terms, I develop a method for automated semantic annotation, which aggregates similar thematic and geographic information. Provider-contributed data descriptions are harvested. From these, I develop a crosscutting base vocabulary of core geospatial topics to map provider to consumer descriptions. The base vocabulary supports the description of geospatial topics and reuses broader, narrower, related, and translated concepts from existing ontologies including Library of Congress Subject Headings, Princeton WordNet, and the USGS Thesaurus. This yields a method for semantically annotating data with core geospatial topics for consumers that is integrated into Esri's ArcGIS Hub, a platform already in use by many government agencies. This semantic annotation method is applied to several prototypical open data portals, resulting in improved data discovery for users (relevance of results) and improved coordination of work for organizations.

This demonstrates a method for mapping user queries to authoritative metadata. Existing ontologies are used to expand the tags associated with civic data, making the topics present in the provider descriptions of civic data semantically richer. However, this solution uses term-matching to expand the topics of geospatial data, implying similarity through related, broader, and narrower concepts rather than by making the relationships among the topics explicit to the user. The next study in this dissertation builds on the idea of term expansion and recommendation, asking how the topics of similar research data can be made spatially and visually explicit for the user:

- **RQ2.** How can the topics of research data be systematically elicited and spatially represented to convey their similarity? (Chapter 3)

I address this in the context of an academic library repository: the Alexandria Digital Research Library (ADRL), which curates non-spatial research data and which does not offer spatial search capabilities. This contrasts with the previous context, where civic data were discoverable through a map-based interface and where thematic, keyword-based search complemented geographic search. In the previous study, the research data were both geographic and non-spatial datasets siloed by data provider terminology across levels of government. In this case, the research data are theses and dissertations siloed by the terminologies used in academic disciplines. Despite these differences, both studies address the “vocabulary problem in human-system communication” [6] as they address disconnects between data users and providers stemming from disciplinary terminologies.

The strategies developed in this study are distinctly spatial rather than verbal. Rather than building hierarchies of terms to expand metadata, I develop spatializations to curate research topics. Core concepts of spatial information [33] provide a set of spatial configurations by which non-spatial data can be organized. The set of concepts (location, field, object, network, event, granularity, accuracy) identify and specify what spatial information is about, “at a level above data models, but independent of particular application domains” [33]. Metadata from the collection of research documents are harvested. I use natural language processing and topic modeling to elicit topics from documents, which are modeled using two distinct content concepts: field and network. The resulting spatial views of research become interchangeable lenses supporting distinct views of research data and distinct questions about them; depending on the similarity relation of interest, certain lenses are better-suited to particular questions. For example, a field (or landscape) view of topics supports questions about distribution and frequency, while a network view of topics supports questions about connectivity and centrality. Each view relies on representations of distance (e.g. proximity, degree) to communicate thematic similarity. Novel views of the same data can be thought of as multiple maps of the same

domain, answering different questions and providing different sets of techniques for exploring them [46]. This results in two spatialized interfaces that enable users to search for related research data across academic disciplines.

This approach shifts away from mapping the language of user queries. Instead, topics are elicited from research data descriptions and are mapped in spatial configurations based on topic similarity. Core concepts of spatial information guide the spatial configuration of non-spatial research data and inform the questions that each view supports. This enables users to take novel spatial views of research topics in a research library collection. The final study in this dissertation explores the ways in which core concepts of spatial information can be used as lenses to produce spatializations, specifically focusing on the systematic selection of levels of detail for both topic models and spatializations:

- **RQ3.** How can the topics of a multidisciplinary body of research be spatially elicited and represented at multiple levels of detail? (Chapter 4)

I address this question in the context of an institutional review of the Earth Research Institute (ERI) at UC Santa Barbara. Quantitative evaluation metrics, like publication counts and amounts of funding, have traditionally been used to summarize ERI's complex, multidisciplinary body of research. There is an expressed need to complement these quantitative evaluation metrics with high-level, qualitative narratives supported by maps of research topics. This need also suggests a clear audience and role for these maps, which makes it relatively easier to evaluate their performance in the context of an institutional review. This is contrasted with the previous study, in which topic modeling and spatialization are used to develop map prototypes that are not formally evaluated.

I improve upon the topic modeling and spatialization methods of the previous study to develop maps of research topics at two distinct levels of detail. These coarse and detailed levels draw upon “distant reading” methods from the Digital Humanities that support

the development of qualitative narratives, comprehension, and discovery of knowledge from vast bodies of text [11]. The maps also use core concepts of spatial information [33] to embed documents as research *objects* in a *field* of continuous topic values. Rather than using these core concepts of spatial information as distinct lenses for viewing research data however, they are used in combination to produce maps of research topics at multiple levels of *granularity*. This work addresses the unresolved problem of choosing adequate levels of detail for topic models and spatializations.

The maps are evaluated by leading researchers affiliated with the institute in preparation for a formal review. Cross-disciplinary topics of research publications and projects emerge at each level of detail, revealing the latent thematic structure of the institute’s multidisciplinary research. These multi-level maps of research topics seem to complement quantitative evaluation metrics that typically inform institutional reviews by exposing thematic expertise, relationships between researchers, topical distributions and clusters of work, and the evolution of these aspects over time.

1.4 Permissions and Attributions

This dissertation has been supported by a private grant⁸ awarded to the UC Santa Barbara Library and the Center for Spatial Studies. The goal of this grant has been to study the challenges and strategies that libraries and researchers face when trying to discover research data on diverse platforms and in a variety of environments. This grant has supported the development of methods that enable researchers to identify research data located across disparate web-accessible databases, repositories, and catalogs.

The content of Chapter 2 is the result of a collaboration with Andrew Turner (Esri R&D) and Werner Kuhn (UC Santa Barbara). This work was part of a research in-

⁸<http://spatial.ucsb.edu/research/spatial-discovery>

ternship at Esri’s Research and Development Office. It was previously published in the 10th International Conference on Geographic Information Science (GIScience 2018) [47] under a Creative Commons Attribution (CC-BY) license and is available here: <https://drops.dagstuhl.de/opus/volltexte/2018/9337/>. The acceptance rate reported by conference organizers for full articles at the GIScience conference in 2018 was 37 percent (17/46). As the lead author on this article, I was responsible for the conceptualization, methodology, and writing. Andrew Turner and Werner Kuhn contributed to conceptualization, writing, editing, supervision, and project administration. Pranav Kulkarni, Daniel Fenton, and Alexander Harris of Esri Research and Development are gratefully acknowledged for their contributions to data curation and protocol deployment.

The content of Chapter 3 is the result of a collaboration with Christina Last (University of Bristol) and Werner Kuhn (UC Santa Barbara). This work was previously published in the 14th International Conference on Spatial Information Theory (COSIT 2019) [48] under a Creative Commons Attribution (CC-BY) license and is available here: <https://drops.dagstuhl.de/opus/volltexte/2019/11110/>. The acceptance rate reported by conference organizers for full articles at COSIT 2019 was 27 percent (8/30). As the lead author on this article, I was responsible for the conceptualization, methodology, data curation, writing, and visualization. Christina Last contributed to the methodology for producing network visualizations and wrote the first version of the associated sections during her time as a visiting scholar at UCSB’s Center for Spatial Studies. Werner Kuhn contributed to conceptualization, writing, editing, supervision, and project administration. André Bruggmann and Sara Fabrikant of the University of Zurich are gratefully acknowledged for their contributions to project conceptualization and methodology during André’s time as a visiting scholar at UCSB’s Center for Spatial Studies

The content of Chapter 4 is the result of a collaboration with Kelly Caylor (UC Santa Barbara Earth Research Institute) and Werner Kuhn (UC Santa Barbara). The work was

submitted to a special issue on paradigm shifts in information science in the Journal of the Association for Information Science and Technology (JASIST) on July 3, 2020. The research article has been released on the EarthArXiv preprint server for early feedback and is available here: <https://doi.org/10.31223/osf.io/523ex>. If accepted for publication, it will be published under a Creative Commons Attribution (CC-BY) license. As the lead author on this article, I was responsible for the conceptualization, methodology, data curation, writing, and visualization. Werner Kuhn and Kelly Caylor contributed to conceptualization, writing, editing, supervision, and project administration. The members of the Earth Research Institute's Advisory Board are gratefully acknowledged for their feedback and support.

Chapter 2

Improving Discovery of Open Civic Data

Abstract: We describe a method and system design for improved data discovery in an integrated network of open geospatial data that supports collaborative policy development between governments and local constituents. Metadata about civic data (such as thematic categories, user-generated tags, geo-references, or attribute schemata) primarily rely on technical vocabularies that reflect scientific or organizational hierarchies. By contrast, public consumers of data often search for information using colloquial terminology that does not align with official metadata vocabularies. For example, citizens searching for data about bicycle collisions in an area are unlikely to use the search terms with which organizations like Departments of Transportation describe relevant data. Users may also search with broad terms, such as “traffic safety”, and will then not discover data tagged with narrower official terms, such as “vehicular crash”. This mismatch raises the question of how to bridge the users’ ways of talking and searching with the language of technical metadata. In similar situations, it has been beneficial to augment official metadata with semantic annotations that expand the discoverability and relevance recommendations of

data, supporting more inclusive access. Adopting this strategy, we develop a method for automated semantic annotation, which aggregates similar thematic and geographic information. A novelty of our approach is the development and application of a crosscutting base vocabulary that supports the description of geospatial themes. The resulting annotation method is integrated into a novel open access collaboration platform (Esri's ArcGIS Hub) that supports public dissemination of civic data and is in use by thousands of government agencies. Our semantic annotation method improves data discovery for users across organizational repositories and has the potential to facilitate the coordination of community and organizational work, improving the transparency and efficacy of government policies.

2.1 Introduction

In recent decades, great strides have been made to encourage data creators and providers to make the findings of their research or results of their activities publicly accessible. Researchers receiving grant funding now face mandates to preserve and expose data resulting from their research [3]. Parallels can be drawn between the mounting movement surrounding open access in academia and similar movements well underway in the civic arena surrounding shared municipal data; all levels of government, from Federal agencies to city governments, have started exposing data [49]. Open data, also known as open Public Sector Information, contribute to citizens' rights to public access of government information. Open data policies at various levels of government have stimulated and guided the publication of both spatial and non-spatial government data [50]. The resulting creative downstream use of civic datasets is staggering, ranging from mobilization of grassroots citizen initiatives to uptake by private application developers [26]. By making civic data about a range of topics, from departmental budgets to bicycle

collisions, consumable through APIs, governments such as the City of Los Angeles¹ have become better connected to their citizenry.

However, simply making data accessible online does not guarantee their discoverability [51]. The likelihood of discovering thematically relevant geospatial data is still quite low; this is due to two key geospatial issues. The first issue is that data produced by co-located and adjacent governments are often described differently. Thus, discovering spatial data about bicycle collisions provided by neighboring governments, such as Arlington and Fairfax, VA along with state data, for example, is not trivial. This is because data, such as bicycle collision statistics, are described in a heterogeneous way by neighboring municipalities and by various levels of government. A second issue is that civic data are not described using terms that public consumers use. Governments may collect and provide traffic collision statistics, while consumers may want to assess community safety for cyclists.

It is unrealistic to imagine all providers of civic data conforming to a single metadata standard or providing suites of additional colloquial keywords to resolve these issues. In fact, the multiplicity of inward-looking open data policies at various levels of government make this untenable [50]. Instead, we ask how semantic mappings can bridge the gap between terms used in peoples' daily lives and terms from technical governmental metadata, thus improving the recall and precision of open civic data. Our approach bridges data provider and data user terms by developing a crosscutting base vocabulary that expands core geospatial topics and can be used to better describe civic data. We demonstrate the value of our approach by applying the vocabulary to automatically annotate data on a novel open access platform. The contributions of this work are as follows:

- A **system** for harvesting provider-contributed data descriptions

¹<http://geohub.lacity.org/>

- A **base vocabulary** of core geospatial themes mapping provider to consumer descriptions
- A **protocol** for semantically annotating data with core geospatial themes for consumers

The remainder of this paper is organized as follows. Section 2.2 provides background on the studied open data platform. Section 2.3 surveys challenges of and approaches focused on improving data discovery. Section 2.4 discusses the method developed to enrich tags during metadata harvesting. Section 2.5 describes the resulting implementation. Section 2.6 discusses the results of the work and presents a research outlook.

2.2 Background

In order to validate the method design and evaluate results, this work integrates semantic annotation into an open access collaboration platform, Esri’s ArcGIS Hub². This platform exposes organizational data via ArcGIS, which is a geospatial data management, visualization, and analytics system used by governments, industry, academia, and other organizations to support planning and operations. ArcGIS integrates desktop software with cloud-hosted tools and data services for distributed information access that can be shared privately or with the public. Using ArcGIS Online, members of organizations and the public can create, edit, and share maps and other data. This global system organizes a content-rich catalog of information across a breadth of scientific themes and operational domains.

ArcGIS Hub is a new open access platform that supports and organizes civic engagement and direct collaboration between governments and their constituents. ArcGIS

²<http://hub.arcgis.com/>

Hub extends the ArcGIS Online system with new capabilities for open data sharing, configurable metadata catalogs, integration with regional and national metadata registries such as Data.gov³, and analysis tools for the public to visualize and share perspectives on data relationships.

Governments and other enterprises can use ArcGIS Hub to create custom websites for open data sharing that allow the public to easily search, access and download data. ArcGIS Hub's primary audience are the general public: people and groups outside of the organizations sharing the data. While ArcGIS Hub integrates with proprietary software, it also serves as a standalone platform that enables anonymous, public access to datasets from any other platform or data provider; it is not necessary to have any authentication credentials in order to discover or use open access data shared through ArcGIS Hub.

As of early 2018, over 100,000 datasets had been made available through ArcGIS Hub by more than 5,000 governments, academic institutions, and other organizations. These datasets are discoverable by search term, specified by user keyword, and by area of interest, which can be specified by map interface. The current state of search in ArcGIS Hub is based on keyword matching, which matches user queries against dataset titles, descriptions, and tags. A limitation of this type of search however, is that it fails to capture broader or related contexts of the query, only returning content that has a title, description, or tags containing the input term. For example, a search for "bicycle" would not return related content, such as "pedestrian", or broader content, such as "transport".

Civic data providers are primarily focused on making their data available and secondarily focused on making their data discoverable to public consumers, often only providing descriptions or tags when required and often using domain-specific terms. This creates semantic and schematic barriers to data discovery, resulting in a gulf between terms that users and terms that providers use to describe and search for the same data.

³<https://www.data.gov/>

Resulting challenges to discoverability and current approaches to address them are the focus of the next section.

2.3 Challenges and Approaches

Data shared through public repositories satisfy basic accessibility requirements, but are often siloed and difficult to discover. A recent report from the Open Research Data Task Force [52] found that the two main challenges to using open data are: 1) finding data to use and 2) (re)using them. While this is especially true of academic data scattered across diverse domain repositories, it is also true of civic data. The current silos for civic data are not simply organizational, but semantic and schematic, rooted in the technical vocabularies used to categorize and structure data [18]. The main challenges to reusing civic data are the domain-specific terms used to describe data and their attribute schemata [52].

Innovations from the arenas of academia, government, and industry demonstrate contrasting, yet complementary, approaches to addressing discoverability challenges [3]; advances in each arena also inform this work. Recent innovations in discoverability have resulted from the implementation of linked data technologies, which allow for data to be self-describing [53]. The uptake of linked data technologies has resulted in an ever-expanding graph of shared knowledge⁴, replete with reusable ontologies from many domains. Linked data technologies address key semantic and schematic challenges, aiding in many arenas such as in the discovery of scientific data for reuse and discovery across integrated civic data streams [2, 25].

⁴<http://lod-cloud.net/>

2.3.1 Semantic Challenges

The first challenge to civic data discovery is semantic. Semantic heterogeneity is understood to result from differing mental models of phenomena as well as from differences in naming conventions; naming heterogeneity can be overcome with term mappings using thesauri, but cognitive heterogeneity is understood to be a more difficult problem to solve in the absence of a minimum set of common definitions [54]. Our work focuses on overcoming heterogeneous naming of semantically similar content, resulting from divergent metadata standards.

The rigor and quality of data classification and tagging schemes can vary greatly by data provider. In the case of highly curated data, such as Federal data layers shared through Esri’s Living Atlas of the World,⁵ tags for each dataset have high agreement and control, grouping the data into one of several predefined themes: demographics, transportation, landscape, oceans. . . Similarly, data conforming to the ISO 19115 metadata standard⁶ adhere to a highly controlled vocabulary describing what the contents are about by keyword: agriculture, biota, economy, health. . . However, as of early 2018, only 66,000 (about 8 percent) of the 760,000 items in the ArcGIS Hub catalog had formal metadata.

Metadata files in ArcGIS Hub are also not indexed for search; instead, keyword search in Esri’s ArcGIS Hub is based on search by regular expression against the titles, descriptions, and tags of content. Organizations contributing data supply their own tags and descriptions, which results in varying levels of quality. Relatively few tags are based on a controlled vocabulary and descriptions of data have varying levels of completeness. This results in a situation where search for “bicycle collisions” returns results for Washington D.C., where data have been assigned the tags of “transportation”

⁵<https://livingatlas.arcgis.com/en>

⁶<https://www2.usgs.gov/science/about/thesaurus-full.php>

and “collision”, but not for the neighboring city of Alexandria, VA where the data have been tagged with “transit” and “accident”.

2.3.2 Schematic Challenges

The second challenge to civic data discovery is schematic. Schematic heterogeneity is understood to result from variations of conceptual schemata within or across disciplines; it can be overcome by schema integration [54].

Governmental organizations such as law enforcement agencies that report traffic accidents, including bicycle collisions, adhere to such integrated specifications, in this case the Model Minimum Uniform Crash Criteria (MMUCC)⁷ developed by the National Highway Traffic Safety Administration (NHTSA). This data model provides a reporting schema; local agencies can adapt it as needed, but it defines a minimum set of uniform fields that can be identified across municipal crash datasets. These criteria specify attribute names (i.e. “County Name”), definitions, and expected data types (i.e. “GLC Code”). Another well-adopted data model developed with interoperability in mind is the Local Government Information Model⁸. Similarly, it defines feature datasets (i.e. “Facilities Streets”), feature classes (i.e. “street lane width”), and attribute fields (i.e. “lane width, type: small integer”).

Where common data models are used, it is possible to easily reuse, and even combine, datasets. However, the majority of data discoverable through Esri’s ArcGIS Hub do not conform to any common data models. Attribute fields are defined ad-hoc and are also not indexed for search unless specified separately as tags.

⁷<https://www.nhtsa.gov/mmucc>

⁸<http://solutions.arcgis.com/local-government/help/local-government-information-model/>

2.3.3 Linked Data Approaches

The need for improved access to civic data parallels that for academic data. Just as research groups, or even academic domains, publish and reuse data according to different standards across various repositories, governmental agencies and municipalities also adhere to a variety of standards with varying levels of quality. The rise of Internet of Things (IoT) technology, which is enabling the evolution of “smart cities”, has also created new sets of challenges related to the volume, velocity, and variety of civic data streams. The challenges that have made heterogeneous civic data difficult to integrate and harmonize in the past have been successfully met by semantic annotation of data streams, which enables their alignment [55].

Rather than semantically annotating civic data after the fact, some governments have adopted linked open data principles as a standard for data sharing; “smart cities” such as London⁹ and Dublin¹⁰ have launched campaigns to expose operational city service data streams in an open, consumable format [26]. Esri Ireland for example now serves national geospatial information as linked data, consumable through an API [56]. In a linked data framework, it is not only easier for both humans and machines to consume civic data, but it is also easier to combine data from multiple sources, for example across levels of government.

One reason for this is that semantically annotated data can be dereferenced, resolving issues of uncertainty concerning attribute values or terms. For example, the United Nations Sustainable Development Goals ontology resolves terminological ambiguity while tracking progress toward shared goals on a multinational scale [25]. The outcomes of such successful linked data approaches motivated us to develop a similar method for semantically annotating civic data in order to improve user search. This method is the

⁹<http://connected-data.london/>

¹⁰<http://smartdublin.ie/>

focus of the next section.

2.4 Methods

In order to improve the discoverability of civic data, we have developed and implemented a base vocabulary and a semantic annotation system. Semantic annotation augments official metadata with relevant tags supplied by a vocabulary, thus expanding the relevance recommendations of data. The method taken to develop and implement an automated semantic annotation system is summarized in the following steps:

1. Formalize base vocabulary for core geospatial themes
2. Extend vocabulary by reusing existing concept hierarchies
3. Augment existing metadata with extended tag hierarchies
4. Evaluate system performance for search

2.4.1 Formalizing the Base Vocabulary

A key contribution of this work is the development and formalization of a compact base vocabulary that maps prototypical themes of government departments to aspects of users' lives. This vocabulary addresses two geospatial problems: 1) it makes data shared by governments that are co-located or adjacent discoverable; and 2) it makes descriptions of the phenomena that data are about semantically relevant to public users. The base vocabulary categories shown in Figure 2.1 were developed in collaboration with civic stakeholders, municipal staff, research organizations, and Esri's Local Government Team¹¹. The vocabulary holistically organizes data and tools, allowing them to be referenced.

¹¹<http://www.esri.com/software/arcgis/arcgis-for-local-government>

Browse by Categories



Safe

Crime
Disaster
Emergency Response



Prosperous

Demographics
Economy
Education



Sustainable

Climate
Energy
Infrastructure



Healthy

Agriculture
Disease
Health-Care



Liveable

Culture
Housing
Transportation



Well-Run

Boundaries
Financial
Planning & Land Use

Figure 2.1: ArcGIS Hub categories reflect existing themes assigned to datasets manually as tags

While these categories reflect typical organizational structures of civic government, they also capture core geospatial themes that communities want to track and measure. These categories are currently used as search facets for data in ArcGIS Hub. While they may structurally reflect issues that communities prioritize, they may not reflect the terms that community members may use when searching for this data. They also may not reflect the terms that a given organization uses to describe its data.

In order to formalize ArcGIS Hub Categories, we began by building a thesaurus of concepts modeled in Protégé¹², an open source ontology editing software. We opted for a pragmatic adoption of the Simple Knowledge Organization System (SKOS) to model these concepts for a number of reasons: SKOS supports flexible modeling of hierarchical relationships; it is used widely across numerous domains; and it is often used in term expansion activities¹³. For these reasons, we were able to reuse authoritative and deref-

¹²<https://protege.stanford.edu/>

¹³<https://www.w3.org/TR/skos-ucr/>

erenceable concepts already published to the Semantic Web by organizations also using SKOS.

Some data available through ArcGIS Hub, such as layers exposed through Esri's Living Atlas of the World, have already been classified and tagged with ArcGIS Hub Categories. These include broader categories like "healthy" and narrower categories like "disease".

However, user-specified terms are not reflected in the ArcGIS Hub Categories. Analysis of the ArcGIS Hub query log revealed that users of Esri's ArcGIS Hub tend to search for data using terms that relate to their own colloquial conceptualizations of theme and geography. In a sample of 470,796 queries performed in 2015, only 12,257 (or 2.6 percent) used any form of the predefined categorical Hub keywords, (i.e. "healthy", "transportation", ...). This means that the majority of themes present in user searches likely take another form. This could mean that users are searching with synonyms of these keywords (i.e. "well-being"), or narrower concepts (i.e. "bicycle"), which would not yield results. Similarly, in the same sample of queries, only 64,353 (or 27.3 percent) use geographic references, like coordinates, addresses, place types, or zip codes in their searches. Similarly, geographic concepts that reflect place hierarchy (i.e. "Ronald Reagan National Airport is in Arlington County, VA") or proximity (i.e. "Reagan Airport is next to East Potomac Park") are not reflected in results.

2.4.2 Extending the Base Vocabulary

We imported existing concepts matching the Hub category tags from Library of Congress Subject Headings (LCSH)¹⁴, Princeton WordNet 3.1¹⁵, and the USGS Thesaurus¹⁶. Reusing these three vocabularies to describe civic data is novel, as they have

¹⁴<http://id.loc.gov>

¹⁵<http://wordnet-rdf.princeton.edu>

¹⁶<https://www2.usgs.gov/science/about/thesaurus-full.php>

been developed and traditionally used to describe library resources and scientific data. These vocabularies provide sufficient terminological coverage for extending the Hub categories shown in Figure 2.2.

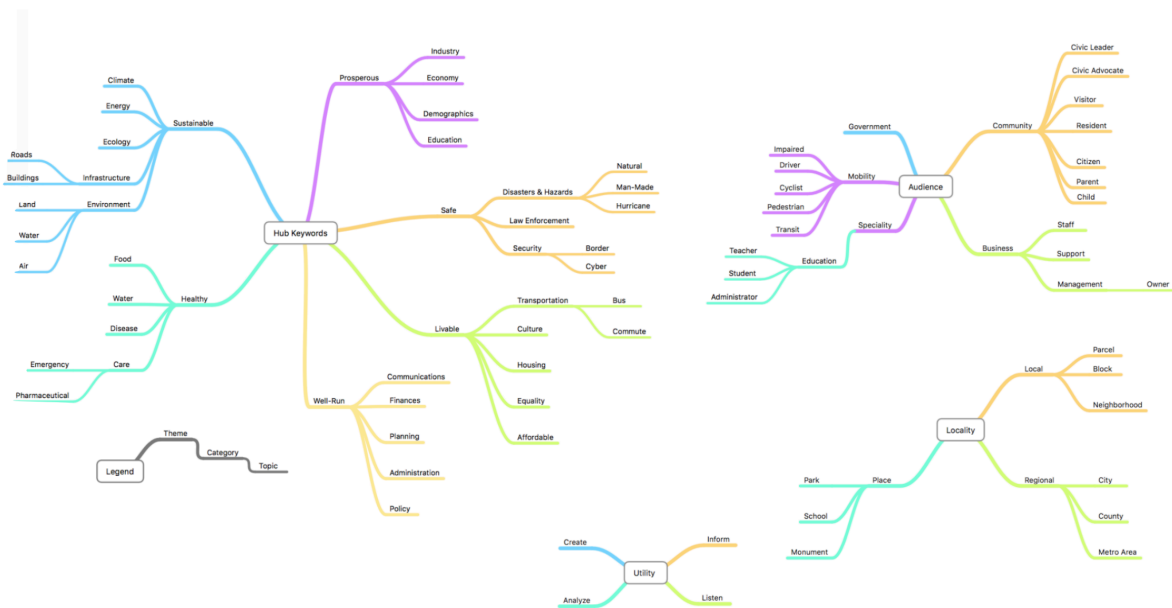


Figure 2.2: Extension of ArcGIS Hub terms to related categories in existing vocabularies

LCSH are a controlled and well-defined set of terms used for resource classification. In addition to providing a stable identifier, LCSH concepts also adhere to a SKOS scheme and provide broader, narrower, and related concepts for each term. For example, “agriculture” in LCSH has useful variants “farming” and “husbandry”, narrower terms like “agronomy”, and related terms like “food supply” and “land use, rural”. LCSH is designed to be used as a thesaurus; its subject headings provide bibliographic access to related subject matter.

Similarly, WordNet terms are also available in in a SKOS scheme and are consumable as RDF, a linked data model. WordNet is a lexical database that combines the capabilities of a dictionary and a thesaurus for the English language. Concepts matching Esri Hub categories were retrieved from WordNet synsets, which are sets of synonyms with

translations. For example, the synset for “agriculture” in WordNet includes “husbandry” and “farming” along with multilingual translations for each. Designed to support cognitive science applications, WordNet is suitable for information retrieval, text classification, and translation tasks [57].

A final source of Hub concept extension comes from the United States Geological Survey (USGS) Thesaurus, which is currently under development. As such, it provides identifiers without dereferencing; despite this, it is a rich source of authoritative scientific definitions and related terms in a SKOS scheme. For instance, it provides examples of the term agriculture used in the topics of “farming” and “horticulture”. The USGS Thesaurus is designed to aid public interpretation of science web resources and topics.

Other sources were experimented with but ultimately were not implemented. Schema.org¹⁷ was considered for thematic and geographic expansion, but was rejected as its top-level concepts are too broad, while narrowing too quickly. Geonames and DBPedia were also investigated, but have not yet been implemented; concepts from these sources may be included in the near future, as both are rich sources of colloquial place-types and themes found in users’ daily lives. It will be possible to extend the base vocabulary following the method developed in this work as other candidate vocabularies are considered.

To further expand ArcGIS Hub terms, we undertook additional mappings from existing categories to community standards, including INSPIRE¹⁸, FGDC¹⁹, and ISO 19115 data specifications. INSPIRE provides 34 spatial data themes, which specify common data models and code lists. INSPIRE themes aim to support the creation of a European Union spatial data infrastructure. These themes include “hydrography”, “transport networks”, and “protected sites”. Similarly, the National Geospatial Data Asset (NGDA)

¹⁷<http://schema.org/docs/schemas.html>

¹⁸<https://inspire.ec.europa.eu/data-specifications/2892>

¹⁹<https://www.fgdc.gov/what-we-do/manage-federal-geospatial-resources/a-16-portfolio-management/themes>

provides a set of 16 themes with appointed lead agencies and the aim of supporting data interoperability. These themes include “climate and weather”, “land use-land cover”, and “soils”. Finally, ISO 19115 provides a set of 19 themes, including terms like “biota”, “health”, and “oceans”. Each of these community standards function as a controlled vocabulary for describing spatial data resources in their respective metadata contexts; their terms overlap to varying extents.

Pragmatically, we were interested in areas of term overlap, as mapping these standardized community terms to the expanded set of ArcGIS Hub terms establishes semantic links between thematically related resources. Various agencies conform to these standards when describing their data. Federal agencies, such as the U.S. Geological Survey, use NGDA themes to describe resources shared through ArcGIS Hub. The FGDC for example maintains a keyword thesaurus with these terms and points to it as a best-practices resource for publishing datasets to open data clearinghouses; it states that “the more robust your theme keyword list, the more likely it can be located by others (and yourself)”. While this is true in principle, describing data with controlled keywords alone will not make data readily discoverable for public consumers of data who often search for data using colloquial terminology.

In order to augment official metadata, the controlled vocabularies for INSPIRE, NGDA, and ISO 19115 were incorporated into the expanded ArcGIS Hub terms. We designated mappings between related terms from each controlled vocabulary in Protégé using the SKOS predicate *related*. Thus, a term like “transportation” has: *related* terms from INSPIRE (“Transport networks”), NGDA (“Transportation”), ISO 19115 (“Transportation”); *broadier* and *narrower* terms from LCSH and USGS Thesaurus (“public transit”); and *synonyms* and *translations* from WordNet (“ES - transporte”). Each of these tags becomes a triple statement pointing to externally defined resources.

2.4.3 Augmenting Existing Metadata

We exported the base vocabulary from Protégé as triple statements in Terse RDF Triple Language (Turtle)²⁰ syntax and imported them into a Fuseki²¹ triplestore, set up as a public endpoint. The vocabulary is stored as a graph that can be queried using SPARQL syntax, which allows for queries across multiple endpoints. Figure 2.3 shows an example of a query template in Fuseki returning query results in JSON to be integrated as auxiliary metadata.

The figure consists of two side-by-side screenshots. The left screenshot shows the 'SPARQL query' interface in a web application. It includes a header with 'query', 'upload files', 'edit', and 'info' buttons. Below the header, there's a section for 'EXAMPLE QUERIES' with buttons for 'Selection of triples' and 'Selection of classes'. A 'PREFIXES' section shows buttons for 'rdf', 'rdfs', 'owl', 'xsd', and a plus icon. The 'SPARQL ENDPOINT' is set to 'http://34.229.180.217:8080/fuseki/category/query' and the 'CONTENT TYPE' is set to 'JSON'. The main area contains a SPARQL query template:

```

2 PREFIX hub: <http://www.esri-hub.com/vocab/>
3 PREFIX owl: <http://www.w3.org/2002/07/owl#>
4 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
5 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
6 PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
7 PREFIX ontology: <http://wordnet-rdf.princeton.edu/ontology#>
8 PREFIX loc: <http://id.loc.gov/authorities/subjects/>
9
10 SELECT ?label ?synonym
11 WHERE {
12   ?term skos:prefLabel ?label .
13   ?term skos:altLabel ?synonym .
14   FILTER contains( lcase(?label), "agriculture")
15 }

```

The right screenshot shows the 'Term Expander' interface. It has a 'Term' input field with 'agriculture' and a 'Fetch dataset from Elasticsearch' button. Below the input, the expanded terms are displayed in a JSON format:

```

{
  "synonyms": [
    "Farming",
    "Husbandry",
    "agribusiness",
    "agricultura",
    "agroindustria",
    "factory farm"
  ],
  "translation": [
    "الزراعة",
    "زراعة",
    "فلاحة",
    "agricultura",
    "landbrug",
    "laborantza",
    "nekazaritza",
    "فلاحت",
    "maanviljely",
    "maanviljelys",
    "maatalous",
    "agriculture",
    "cucuk tanam",
    "pengebunan",
    "penternakan",
    ...
  ]
}

```

Figure 2.3: SPARQL query template (left) and expanded terms (right) for term “agriculture”

ArcGIS Hub includes a search index of aggregated dataset records from all data providers. When organizations like governments indicate their data is public, ArcGIS Hub compiles multiple metadata sources into a custom search index to support multiple content search and discovery services.

²⁰<https://www.w3.org/TeamSubmission/turtle/>

²¹<https://jena.apache.org/documentation/>

The search index process, shown in Figure 2.4, includes three phases: harvesting, validation, and enrichment. During harvesting of a dataset, Hub collects metadata from the ArcGIS Online item information, associated formal metadata, the feature service and feature layer definition, and data attribute aggregate statistics. Validation includes heuristics to measure metadata completeness, support for secure connections with HTTPS, and query responsiveness, which determines if the data are actually accessible. During the enrichment phase, a dataset is decomposed into relevant keywords which are then sent to the semantic query service to retrieve new semantic tags that are then attached to the dataset metadata.

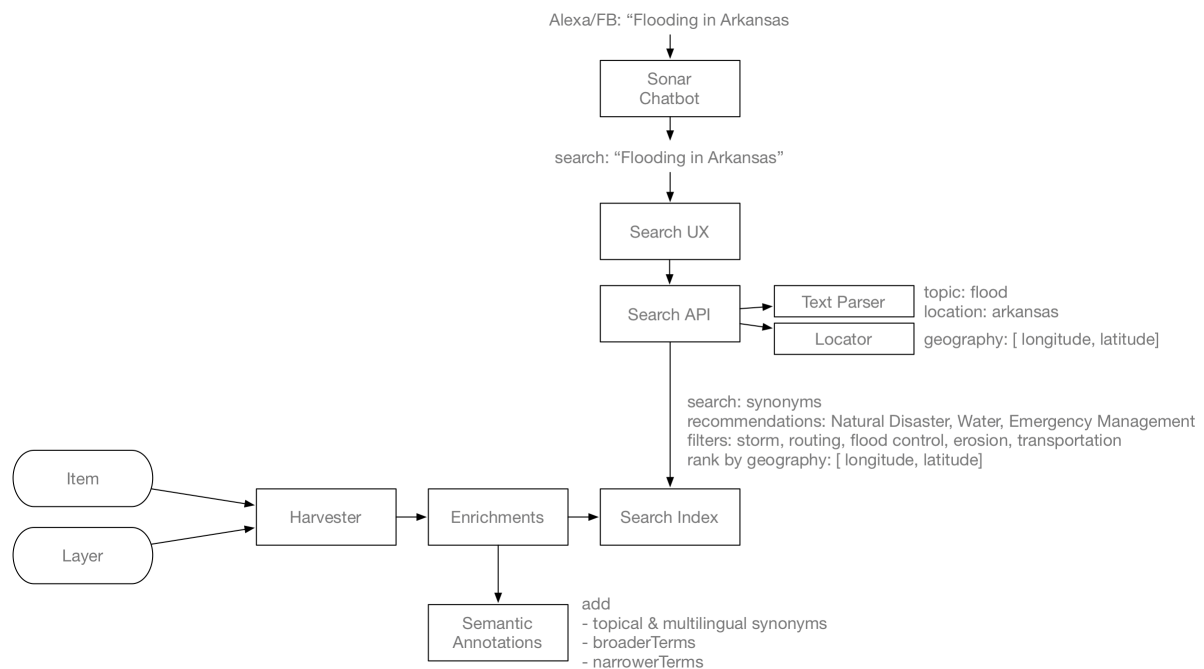


Figure 2.4: Semantic annotations added to metadata, supporting search through query expansion

For example, Flood Zone data from Evansville, Indiana are tagged “Evansville, Vanderburgh County, Flood Zones, IN, environmental”. Using each of the terms from each of the tags results in a superset of synonyms, translations, broader terms, and narrower terms, shown in Figure 2.5. These terms are each added to the dataset record in the

search index using an internal semantic annotation service. The semantic annotation service is an internal API that hosts the base vocabulary as a queryable API using the Apache Jena Fuseki server. This server supports defined requests to build a set of tags that expand the dataset metadata for broad, narrow, translated, and similar terms.

At query time, these additional terms can be used to match user queries such as “human health”, or “impact assessment” that may not have another similar word match in the dataset metadata collection but will now have results based on matching these new, additional semantic tags. The semantic tags also include translations such as “air pasang” (Indonesian) or “nousuvesi” (Finnish). Beyond similar terms, there are broader terms such as “Natural disasters” and “Water” and narrower terms such as “Flood damage prevention” and “Forest influences” that can be used to recommend new search terms to the user for refining their search results.

2.4.4 Evaluating System Performance for Search

In practice, search for data is now semantically aided; related content, such as synonymous terms, can be retrieved when inferred as thematically related. For example, a search for traffic accidents can now return other content related to a broader concept of transportation’ as pedestrian fatalities. While only a small fraction of data (about 8 percent) in ArcGIS Hub initially included formal metadata, semantic annotations added related metadata in the form of related terms, supporting data discoverability and integration.

In order to evaluate the contributions of our approach, we consider that semantically enabled search wasn’t previously possible: this informs our baseline criteria. Search efficacy is measured accordingly using several methods: conversion rates through usage analytics tracking, usability testing, and relevance judgment evaluation.

Usage analytics tracking measures all user interactions with the ArcGIS Hub web application. This includes search inputs, filter interactions and result selection. We define several conversion funnels corresponding to expected user outcomes, which include downloading the data, creating an information product such as a web map or a Story Map, or bookmarking a view of the data for later use. These conversions indicate that a good search result was returned. We can then compare conversion results with and without semantic annotation.

Usability testing includes defining a workflow that human test subjects perform while being monitored by researchers. Listening to stream of consciousness verbal evaluations and observing interface interactions denotes perceptions of different search modalities and outcomes. This testing may be performed in-house or in collaboration with stakeholders.

Lastly, relevance judgment evaluation asks a similar set of users to evaluate the quality of search results as: perfect, relevant, partially relevant, or irrelevant. The scores for each result are tallied and compared with the optimal result and rank ordering to define the quality of the search relevance, due to semantic annotations or without semantic annotations. The results of these evaluation measures are forthcoming at this time of writing.

2.5 Results

Governments, academic institutions and other organizations publish open data to encourage the creative reuse of information for new purposes. ArcGIS Hub allows these organizations to create websites that enable search and discovery of their authoritative data, as well as recommend data shared through other groups. The Bureau of Transportation's Geospatial Statistics site is shown as one such example in Figure 2.6. Visitors can perform simple searches through their web browser or mobile device, or request in-

formation through new digital media chatbots on Facebook and Amazon Alexa.

Extending dataset metadata with semantic annotations expands the discoverability of information through colloquial and multilingual search associations. Figure 2.4 illustrated how search queries use the semantic search index to parse and retrieve relevant datasets.

To use the semantic search API, Hub implements a REST HTTP API for structured queries from web browsers, mobile apps, and custom embeds; it uses a JSON-Schema self-documenting hypermedia API and includes search index attribute filters and facets. An API search query is first split into relevant parameters for keywords, time, location, and provider. The keywords are compared with the semantic annotation tags for similarity matches; the time, location and provider are used as filters. The result includes a relevance-ranked list of datasets as well as aggregate facets of topics, data types, and providers for further filtering.

The semantic annotations augment the search relevance matches by comparing search keywords with terms that may not have existed in the original metadata document, but describe the dataset with alternative labels that match these queries. Figure 2.5 shows an example of expanded semantic tags that are compared for relevance ranking, including multilingual terms, as well as the broader and narrower aggregate terms that can be used in search interface facets.

2.5.1 Building Data Networks

Semantic annotation supports additional use cases beyond metadata querying. ArcGIS Hub includes a global catalog of data from governments of various administrative levels: local council and departmental, metropolitan, provincial, regional, national, and multinational organizations. Each government follows a varying set of metadata and keyword standards that may not overlap with other governments, even if the organizations

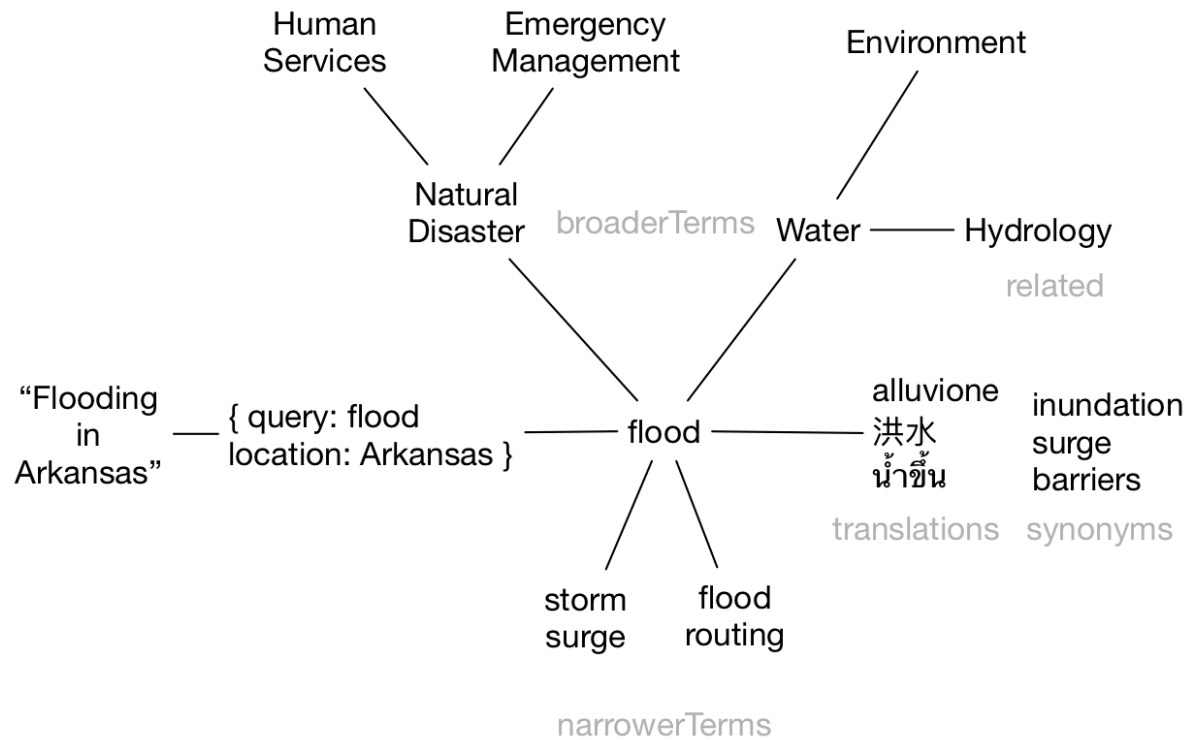


Figure 2.5: Search queries are parsed and compared with semantic annotations to expand matches and provide additional facets

are geographically adjacent or coincident. This can make integration of data across municipal boundaries problematic, resulting in lost productivity or detriments to operations and safety.

Semantic annotations support data integration by organizing datasets into common thematic groupings, which increase the discovery and utilization of similar datasets across municipal data providers. By way of example, consider several civic datasets provided by neighboring municipalities such as road networks, public schools, moving violations (e.g. vehicle speeding citations), and reported crashes between vehicles, bicycles or people. Additionally, there are regional and national datasets provided by agencies that also include transit networks (bus stops and train stations): FARS (Fatality Analysis Reporting System).

In order to track progress toward thematic community initiatives, such as “Vision Zero”, discovery of relevant data must be possible across all levels of government. Vision Zero is a strategy to eliminate all traffic fatalities and severe injuries, while increasing safe, healthy, equitable mobility for all. Potential Federal data sources for tracking a “Vision Zero” Initiative are shown in Figure 2.6. However, without semantic annotation, there is uncertainty as to whether a search for traffic data will return relevant results across other Hub sites at a state, county, or municipal level.

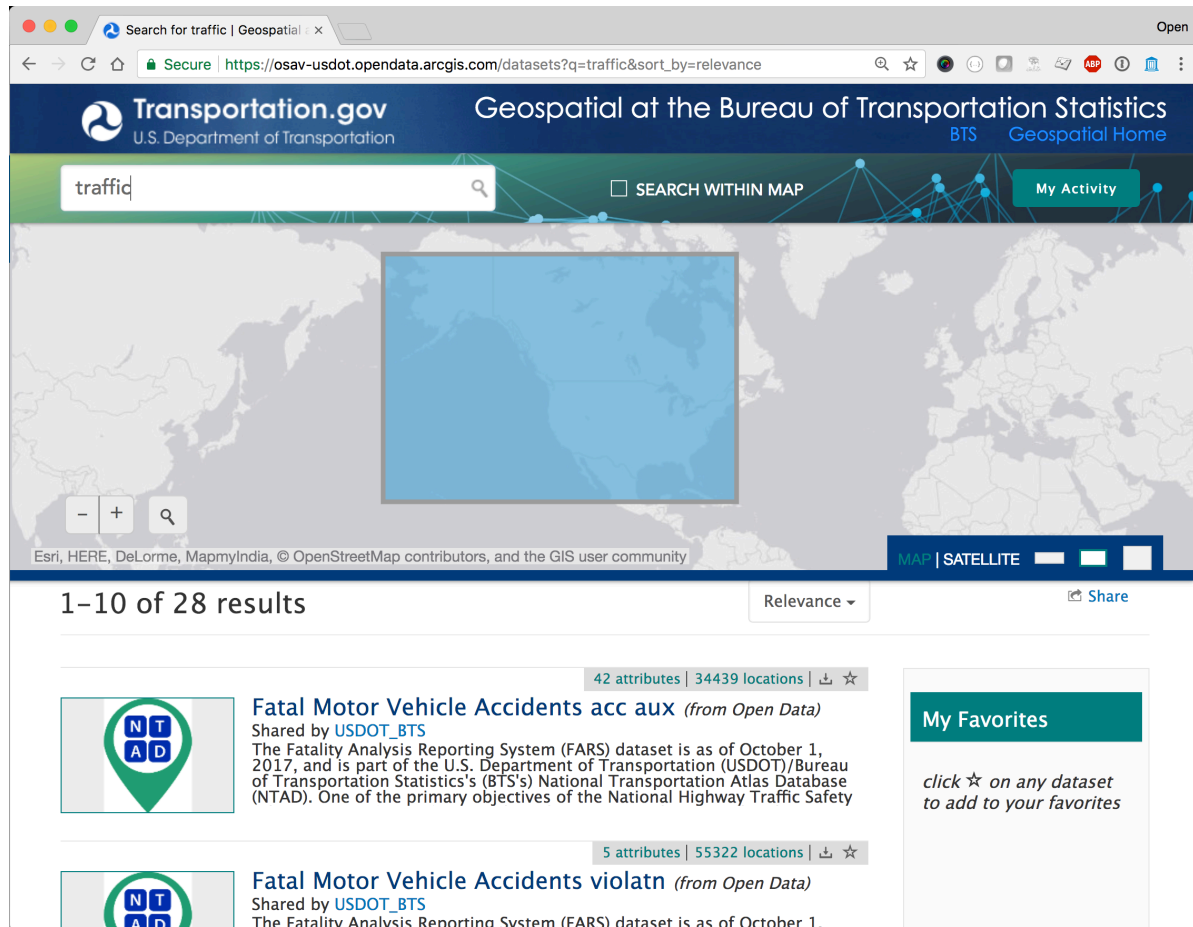


Figure 2.6: U.S. Department of Transportation traffic related datasets sorted by relevance

ArcGIS Hub builds the search index that includes each of the four example local municipal datasets from each municipality. This includes the original metadata and

the additional semantic annotations on the datasets that associate them with related thematic groupings. Searching just the category term has mixed, or missing, results from some provider catalogs. Figure 2.7 compares search results across the GIS catalogs of the District of Columbia, State of Maryland, and County of Arlington, exposed through ArcGIS Hub.

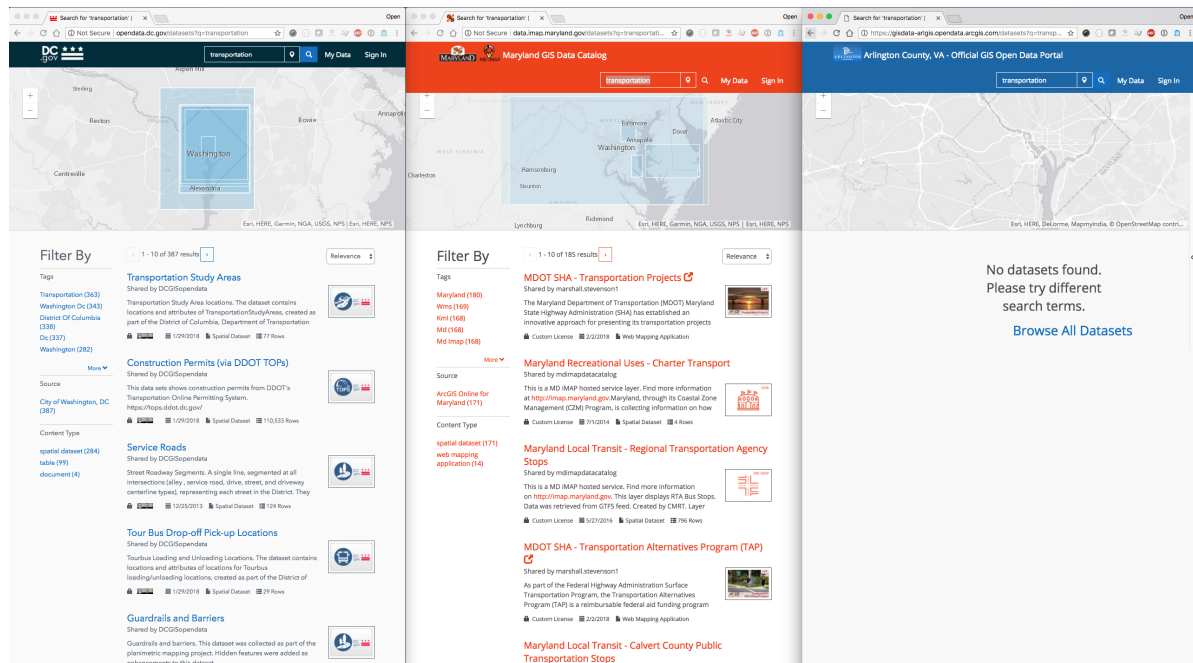


Figure 2.7: Comparing searches for “transportation” before adding semantic annotations

By comparison, when colloquial terms are used, there are similar results from all local providers. Figure 2.8 compares search results across the same GIS catalogs for related terms.

2.6 Discussion and Outlook

The work presented in this paper improves data discovery through the application of semantic annotations to civic data, which facilitate transparency and coordination

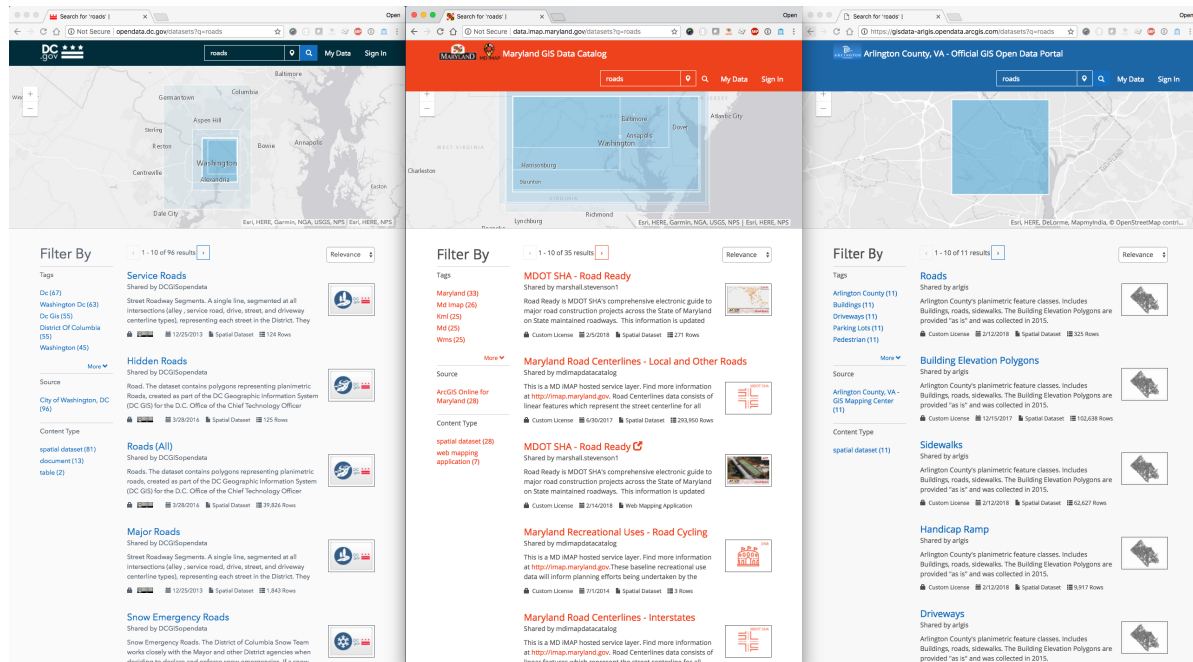


Figure 2.8: Comparing searches for related term “roads” after adding semantic annotations

of work; semantic search enables the exploration and discovery of relationships among organizations’ data that were previously unknown.

Several areas of research are continuing from this work. We plan to expand and refine the base vocabulary to better support bi-directional term expansion. This will allow users to discover new datasets by improving traversal of the base vocabulary’s relations, like broad and narrow terms, for both thematic and geographic concepts. We anticipate that alignment with new ontologies, such as the U.N.’s Sustainable Development Goals Ontology, and application of our methods in related domains, such as academic libraries, will continue to improve data discovery across organizational repositories.

On a larger scale, the lessons learned from our research can be applied to new domains and extended along the following dimensions.

The Sustainable Development Goals (SDGs) are the results of an ambitious global initiative to improve the health and well-being of people and communities. They consist

of 17 goals, 169 targets and 232 data indicators that will measure and monitor progress towards the SDG. These targets and indicators include a semantic graph that relate to socioeconomic terms, municipal planning, and other related governance sectors. Work is ongoing with several national mapping agencies and the United Nations to integrate their semantic graphs with the base vocabulary presented in this paper.

We are also applying the methods developed in this paper to data discovery in the context of digital research libraries. While libraries have long been the traditional brokers of knowledge, today's queries are largely mediated by commercial digital search engines [31]. Yet, libraries are taking on new roles, facilitating discovery, and often co-production, of knowledge [19]. Semantically annotated data can be more easily discovered and retrieved via queries that traverse knowledge graphs, regardless of the endpoints where they are hosted. Academic libraries are poised to serve as a semantically-neutral meeting ground where domain data can be aggregated and made spatially and thematically discoverable, similar to ArcGIS Hub.

Chapter 3

Enabling the Discovery of Thematically Related Research Objects

Abstract: It is challenging for scholars to discover thematically related research in a multidisciplinary setting, such as that of a university library. In this work, we use spatialization techniques to convey the relatedness of research themes without requiring scholars to have specific knowledge of disciplinary search terminology. We approach this task conceptually by revisiting existing spatialization techniques and reframing them in terms of core concepts of spatial information, highlighting their different capacities. To apply our design, we spatialize masters and doctoral theses (two kinds of research objects available through a university library repository) using topic modeling to assign a relatively small number of research topics to the objects. We discuss and implement two distinct spaces for exploration: a field view of research *topics* and a network view of research *objects*. We find that each space enables distinct visual perceptions and questions about the relatedness of research themes. A field view enables questions about

the distribution of research objects in the topic space, while a network view enables questions about connections between research objects or about their centrality. Our work contributes to spatialization theory a systematic choice of spaces informed by core concepts of spatial information. Its application to the design of library discovery tools offers two distinct and intuitive ways to gain insights into the thematic relatedness of research objects, regardless of the disciplinary terms used to describe them.

3.1 Introduction

In recent decades, the curation of scholarship and its access mechanisms have shifted from physical to virtual spaces. In the 1990s, physical card catalogs were migrated to online databases, trading collocation for scalability [58]. Similarly, library shelves with thematically collocated material are today largely accessed through virtual spaces, such as digital repositories organized by faceted categories [22]. This shift has increased the potential for exchange of scholarly information on the Web through semantically rich *research objects* [2]. While online library services may provide scholars with access to millions of research objects, they do not necessarily improve the ability of scholars to serendipitously discover related objects. Such a capacity was naturally built into the physical spaces of book shelves, albeit in a limited form.

Spatialization can recreate specially designed two-dimensional thematic spaces, such as neighborhoods and networks of themes. These spaces support exploration, browsing, and navigating and can be exploited in future search and discovery services, complementing standard known-item searches [45].

Exploratory search is already supported by library services, like GeoBlacklight¹ and DASH², which index research objects *geographically* and enable discovery and access

¹<https://geoblacklight.org/>

²<https://datadryad.org/search>

through map interfaces. Such services curate and expose research objects based on their geographic footprints, derived from the named places that they are about (if any). They enable the integration of research perspectives by geographic locations, revealing spatial patterns, such as clusters or gaps [44]. They are especially useful in a university setting where research objects from different disciplines may refer to the same places [19]. However, geographic space only captures geographic notions of location and relatedness. Location, time, and theme are dimensions that can be used to organize observations [29] including research objects. Since in many cases, the temporal organization of research objects is comparatively straightforward (e.g. indexing research objects by their date of publication or the period they are about and displaying them using a time slider), we take on the bigger challenge of representing the relatedness of research themes.

We address this challenge by literally mapping it to the existing solution for discovery by geographic location. In other words, we ask how exploratory search for research objects can be improved by maps of thematic spaces in which related research themes are placed closer together. Conceptually and technically, we adapt our previous work to expose research objects by their geographic footprints [19] to enable discovery in specially designed two-dimensional thematic spaces, which we implement using spatialization techniques. Spatializations exploit people’s familiarity with spaces in everyday life to produce intuitive visual information spaces that convey similarity through distance [37]. Spatializations, like self-organizing maps informed by cartographic principles, have been applied to efficiently visualize knowledge domains, such as the subdisciplines of geography [59]. Various types of spatializations, including point maps [40], network maps [41], and regions [42] have been proposed and empirically evaluated, demonstrating that viewers correctly interpret nearby items in abstract space as similar. Analogous to the first law of geography [23], this first law of cognitive geography states that viewers believe that closer things tend to be more similar [40].

Yet, spatialization remains underexploited, particularly in libraries, which have to deal with vast and context-dependent thematic search spaces. We see this as an opportunity to experiment with spatialization in a multidisciplinary university library repository of research objects. What further distinguishes our approach is that the spatial views we develop are designed based on core concepts of spatial information³; in this theory, a base concept (location), a set of content concepts (field, object, network, event), and a set of quality concepts (granularity, accuracy, provenance) capture what spatial information is about. This theory positions spatial information at a level above data models, independent of particular application domains [33]. We use these concepts to design two kinds of spatializations: fields of research topics and networks of research objects. A field of research topics reveals their distribution, while a network of research objects reveals their connectivity and centrality. We implement these two spatial views by selecting the spatialization techniques of a self-organizing map [9] and of a planar network. To obtain the necessary visual interfaces for these abstract spaces, we extend the capabilities of the same web GIS platform (ArcGIS Online) that we previously used to display and discover research objects geographically. We show how the spaces that we design are configurable and enable intuitive exploration and discovery of related research objects across disciplines.

The remainder of this paper is organized as follows. In Section 3.2, we present a motivating scenario to illustrate the challenge of discovering related research. Section 3.3 explains our conceptual approach to systematize the design of search spaces through the core concepts of spatial information. In Section 3.4, we implement spatializations of research objects from a university repository of masters and doctoral theses. In Section 3.5, we apply the spatializations, demonstrating the types of questions that they enable with examples from the previous search scenario. Finally, in Section 3.6, we envision discovery

³<https://www.researchgate.net/project/Core-Concepts-of-Spatial-Information>

in spatializations informed by other core concepts of spatial information.

3.2 Enabling Research Discovery Across Disciplines

Discovering thematically related research in a multidisciplinary setting is both important and challenging. This is a consequence of the siloing of scientific perspectives on the world into different disciplines and the heterogeneous terminologies used within them [60]. Specifically, scholars may find it challenging to identify collaborators and methods outside of their discipline. This is problematic, given that scientific studies and applications of geographic information are increasingly transdisciplinary [33]; they may, for example, combine knowledge from sociology and psychology, or borrow methods from computer science and engineering.

As a motivating scenario, consider two published Geography theses: *Representations of an Urban Neighborhood: Residents' Cognitive Boundaries of Koreatown, Los Angeles* [61]; and *A Temporal Approach to Defining Place Types based on User-Contributed Geosocial Content* [62]. How could the authors of these theses have gone about finding collaborators studying related topics or using related methods? Even for trained interdisciplinary researchers, disciplinary terminologies make it hard to discover related research, resulting in missed sources, insights, and opportunities for collaboration. How can researchers be made aware of thematically related research without needing to know its disciplinary terms?

A common approach to reduce mismatches in keyword-based search is to use ontologies to expand the set of search terms [22]. Such network-based approaches are often based on Linked Open Data and in the case of web journals, enable the discovery of networked data about authors, reviewers, and editors [63]. However, this approach loses the more intuitive similarity relations in the construction of terminological hierarchies

[8], whose relations (e.g. broader, narrower) may not always be meaningful to the user. Thus, we propose to complement the terminological approach with an innovative spatial approach affording similarity judgments on research themes. Just as designs for successful everyday spaces, like neighborhoods and street networks, follow spatial patterns [64] and support important cognitive strategies, so can the designs for visual spaces that enable serendipitous discovery. These spatial patterns and strategies are well-understood in the geographic case (consider navigation or perspective-taking) and spatialization carries them over to abstract thematic spaces. The organizational affordances of space, well-known from geographic as well as desktop spaces, can be built into abstract spaces, creating useful and intuitive spatial structures for research themes.

3.3 Conceptual Approach: Making the Choices of Spaces Systematic

The core concepts of spatial information [33] offer a systematic approach to defining spatial structures by providing a typology of geographic (and other) spaces to guide the organization and interpretation of spatially referenced data. Thus, we recast spatialization as a conceptual choice of a lens through which to view data (i.e. viewing research objects as a field or network). The core concepts of spatial information provide lenses that enable distinct views on spatialized relationships, such as similarity. To go beyond purely cartographic design [44], we make our choices of spaces more systematic by basing our spatializations on those two core concepts that have a solid mathematical formalization: *fields*, formalized by continuous functions from location to theme, and *networks*, formalized by graph theory.

3.3.1 Choices of Spaces and their Entailments

We first review previous work to create *field* and *network* spatializations, highlighting their underlying spatial theories that inform and evolve our approach. Our thesis is that, if treated systematically and formally, there are distinct choices of spatial concepts that carry perceptual powers; these enable specific types of questions and associated insights.

Landscapes and Fields: We begin with an example from Wise’s [65] pioneering intelligence work, where a spatialized display of news documents shows viewers intuitive similarity relationships based on their proximity in the display. Documents are treated as objects, with *k-means* and *complete linkage hierarchical clustering* used to project documents to a two-dimensional plane. This results in a spatialization, where the position of every news document is surrounded by a neighborhood of topics. A surface is then fit over the display, representing a terrain with peaks of high frequency terms drawn from the corpus.

While this work introduces the metaphor of a landscape or terrain to information visualization, it conflates the field of topic vectors with one of topic frequencies, essentially performing a local map algebra operation. The two field views (topic neighborhoods and topic frequencies) can be separated and an additional view of documents as objects can be added; each affords different types of reasoning (on similarity, frequency, and clustering). In our work, we will show this idea for the case of research objects. While we omit frequencies, which are not supported by adequate amounts of data, we further develop the object view into a network view that shows specific connections between documents.

Another example of an information landscape is Fabrikant’s [45] spatialization of a digital library’s holdings. Like Wise’s approach, multidimensional scaling is used as a projection method to create a surface of keywords. However, Fabrikant’s work extends the landscape metaphor by explicitly referencing three spatial concepts: 1) distance

(similarity), 2) scale (level of detail), and 3) arrangement (dispersion and concentration), based on Golledge’s primitives of spatial knowledge [66]. These concepts are used to systematically inform what users can do in the landscape: looking (overview), navigating (to discover items of interest), changing level of detail, selecting individual documents, and discovering relationships between documents (detail on demand). While this example moves toward conceptual formalization, it does not yet support multiple views based on different spatial concepts.

Networks and Graphs: “Maps of science” visualize research networks, ranging from co-citation networks to expertise profiles [39]. Brner et al. visualize a network of millions of university research articles embedded in an abstract spherical space. The network is rendered in a pseudo-Mercator projection, based on the idea that a Riemannian perspective, which uses a sphere as the layout surface, offers continuous linkages. However, it is unclear what additional costs or benefits this choice imparts, as some network properties (like centrality) may be more challenging for viewers to ascertain in such a view.

The extraction of spatial and temporal information from digital text archives can inform more systematic spatializations [67]. Bruggmann and Fabrikant embed a network of toponyms (nodes) and their relationships (edges) in a geographic map to illustrate their connectivity and hierarchy. The inclusion of time in their analysis enables interesting questions about how certain places have risen or fallen in prominence over some period; this is encoded by node size (frequency of mention) along with edges (co-references with another place). The resulting networks are clear and effective, highlighting important relationships, like centrality, through systematic choices of node roles, edge roles, weighting, and embedding.

3.3.2 Locating Research Objects in Topic Space

Our conceptual design addresses university theses, which do not have any inherent way of locating them. We therefore model them as research objects in an n -dimensional vector space of topics. To locate them, we perform topic modeling on their titles and abstracts. Although the full text is available for most theses, we consider them to be adequately described at the metadata level; our approach gains efficiency and practicality, as only commonly available metadata are required for spatialization. Topic modeling assigns each thesis a vector of keywords (standing in for their topics) locatable in a two-dimensional topic map. We chose to assign topics to research objects, as this supports useful exploratory data analyses [13].

Field-based model: Rather than using the topic model to compute on the similarities of theses, we spatialize it into a topic map that supports visual pattern detection and similarity inferences. Our first spatialization is based on the field concept, with topics as the field attribute. Fields enable questions about the value of an attribute at any position in a given spatial and temporal domain. Field-based models underlie, but do not imply the use of, a landscape metaphor. They involve explicit choices of a spatio-temporal framework and a type of attribute (scalar, vector, spinor, or tensor).

We create a self-organizing map (SOM) using the vectors of words that result from the topic model. The SOM creates a field with a two-dimensional abstract spatial framework and a vector attribute. It represents topic locations as hexagonal cells into which point objects (representing the theses) fall. This can be seen as an example of a relative Leibniz space, generated based on objects, rather than a pre-established absolute Newtonian space [68]. The SOM satisfies the criteria for field-based models as follows:

- In its **spatio-temporal framework**, time is held constant (covering the entire period of available theses), location is controlled by the topic map, and theme is

measured.

- The measured attribute **value** is an n -dimensional topic vector of words associated with the topic, ordered by their probability of occurring in theses on the topic.
- Furthermore, the topic field is **continuous**, in that a small move in position in any of six directions results in a small change in attribute value.

Network-based model: Our second choice of spatialization is based on the network concept. Networks provide views of objects that are not supported by a field view, such as questions about direct connections between objects and their centrality in the network [33]. Graphs formalize network models and give them inferential power and versatility.

Network models in general require the following explicit choices [69]: what plays the role of a node?, what plays the role of an edge?, how are edges labeled or weighted?, do they have direction?, and is there an embedding of the nodes, edges, or both in another space? Like the field-based model, the planar network that we produce also exemplifies a relative Leibniz space. Our network spatialization of theses rests on the following choices:

- The theses (research objects) are conceptualized as **nodes**.
- The **edges** are defined based on a **binary** topical relation between theses; if two research objects have at least one of five “top topics” in common, they share an edge.
- The edges are **weighted** by the value of the topic attribute (0–1).
- The edges are **non-directed**, as topic sharing is symmetrical.
- The nodes are **embedded** in a planar space, also based on value of the topic attribute.

3.4 Technical Approach: Implementing Field and Network Spatializations

We spatialized masters and doctoral theses accessible through the Alexandria Digital Research Library (ADRL), a repository⁴ curated by the UC Santa Barbara Library. It is named for the original Alexandria Digital Library (ADL), a project in which users could access multimedia library objects through a map interface [70]. Experimental work on ADL also resulted in a prototype information landscape of library objects based on frequent keywords [45]. Despite the lineage that ADRL shares with the original ADL geo-library project, it does not offer any spatial search capabilities, neither in geographic nor in thematic space; this design limitation presents an opportunity to develop spatial views that enable the discovery of research objects. We use established topic mapping and spatialization techniques [43] to:

- harvest the metadata of research theses from the ADRL repository,
- compute and assign topics to the theses using topic modeling, and
- spatialize the topics, producing a self-organizing map (SOM) and a network.

3.4.1 Metadata Harvesting

For our experiment, we chose research theses published by graduates of UC Santa Barbara between 2011 and 2016 that represent all 53 academic departments granting graduate degrees. The theses are accessible through a public-facing search interface, which provides keyword-based search and facets. The metadata are not accessible through an API, so we obtained permission from the UCSB Library to harvest them for the 1,731

⁴<https://alexandria.ucsb.edu/collections/f3348hkz>

research theses using a combination of *WGET*⁵ and the Python libraries *Crummy* and *Beautiful Soup* 4⁶. The metadata follow the Portland Common Data Model⁷ and are comprised of: a unique identifier; a title (of 50 words or less); a year of publication; an author; a degree grantor; a degree supervisor; a language; and a detailed abstract (no word limit) containing a problem statement, a description of methods and procedures used to gather data, and a summary of findings. Researcher contributed (uncontrolled) keywords were only available for research theses added after 2017, so we did not include keywords in our topic model.

3.4.2 Topic Modelling

We produced a topic model using *MALLET*⁸, an open-source package developed for text-based machine learning applications. We applied Latent Dirichlet Allocation (LDA) to model the topics present in the combined text of the title and abstract of each thesis [13]. LDA is used to determine the thematic relatedness of theses by attributing the presence of each word in the combined title and abstract text to a topic characterized by a word vector. LDA is suitable, as it has been applied to many similar dimensionality reduction and classification problems [13]. LDA largely succeeds in capturing the notion of relatedness (relative to the set of inputs) despite the fact that different terms are used within those inputs (e.g., variability and change are likely to be grouped into a single topic). Thus, it is a pragmatic solution for dealing with complex notions of topics and their relatedness.

We removed the standard English stop words using a list from the MALLET package. We then experimented with between 30 to 100 topics, roughly corresponding to the num-

⁵<https://www.gnu.org/software/wget/>

⁶<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

⁷<https://pcdm.org/2016/04/18/models>

⁸<http://mallet.cs.umass.edu/>

ber of academic departments at UCSB, which indicates a rather coarse topic granularity and targets the cross-disciplinary scope of our inquiry. We found that 71 topics provided the lowest log-likelihood value, a criterion that optimizes for the tightest possible lower bound [13]. We then assigned topic probabilities to the research objects, coded from 0 to 70. We chose to leave the topics unlabelled; they are characterized only by their word vectors. The assignment of topics provides the basis for relatedness in the following steps.

3.4.3 Field Spatialization

We adapted a method developed by Bruggmann to spatialize the output of a topic model [67] by using a self-organizing map (SOM) toolbox⁹ for ArcGIS 9 written by Lacayo-Emery. This toolbox implements the SOM algorithm [9] in existing cartographic software, leveraging its clustering and dimensionality reduction to produce a 2-dimensional map that is readily visualized.

We set the following parameters: the x / y dimension of the SOM was 42 x 42 (1,764 hexagons); the length of training was 50,000 / 500,000 runs; and the initial neighborhood radius was 42 / 6. We used the probability distribution matrix that resulted from topic modeling to produce our SOM template in ArcGIS Desktop. For cartographic readability, we only display theses from the most productive departments (those with over 50 theses). This resulted in a SOM showing 775/1,731 theses from 10 departments. Figure 3.1 shows the SOM, which is also published to ArcGIS Online as an interactive web application¹⁰.

3.4.4 Network Spatialization

We applied a hierarchical clustering method adapted from Leicht et al. [71], which is a compromise between the single-linkage clustering method (in which a single edge

⁹<http://code.google.com/p/somanalyst>

¹⁰<http://arcg.is/0vyezH>

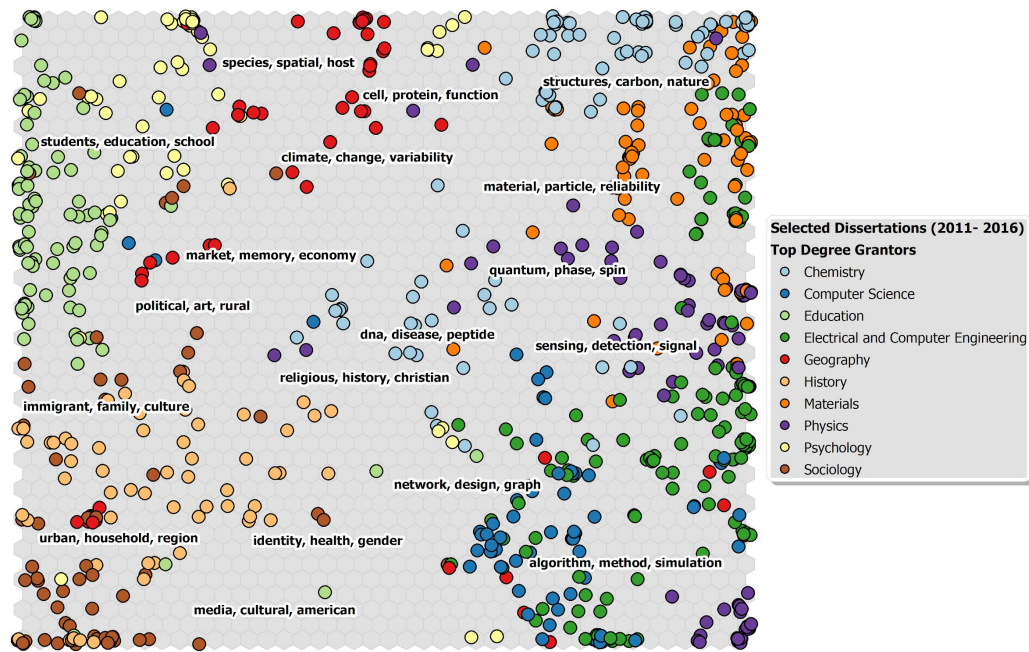


Figure 3.1: Theses (color coded by academic department) located in the SOM (gray tessellated topic field of themes, shown with representative terms from word vectors)

is defined based on the most related pairs of nodes) and average-linkage clustering (in which an edge is defined based on the average relatedness of all pairs of nodes). We used the *tidyverse*¹¹ package in R to construct the edge list, assigning theses the role of nodes and shared topics the role of edges; for cartographic readability, we restrict shared topics to 5. Specifically, each thesis is characterized by the same topics and associated word vectors used to produce the SOM. For example, if *Thesis A* is characterized by Topics 2, 11, 22, 34, and 60 and *Thesis B* is characterized by Topics 4, 11, 27, 33, and 51 they share one edge based on shared Topic 11.

We scale node size relative to the amount that two nodes share a corresponding topic; thus, a larger node corresponds strongly with its shared topic and a smaller node does not. For example, if *Thesis A* is characterized 70% by Topic 11, its node size will be 0.7 (out of a maximum size of 1). We also embed nodes in a planar space (distinct

¹¹<https://www.tidyverse.org/>

from that of the SOM) that reflects how strongly each node corresponds to its “top-topic”; the position of each node reflects the value (0–1) of the top topic vector. To enable comparisons between the SOM and the network, we randomly sampled without replacement 775 nodes, embedded in a planar space, and connected them with edges standing in for a “top-five” topic. Figure 3.2 shows the network constructed with the *networkx*¹² library, which is also published in a reproducible Jupyter Notebook¹³ and deployed using Binder¹⁴.

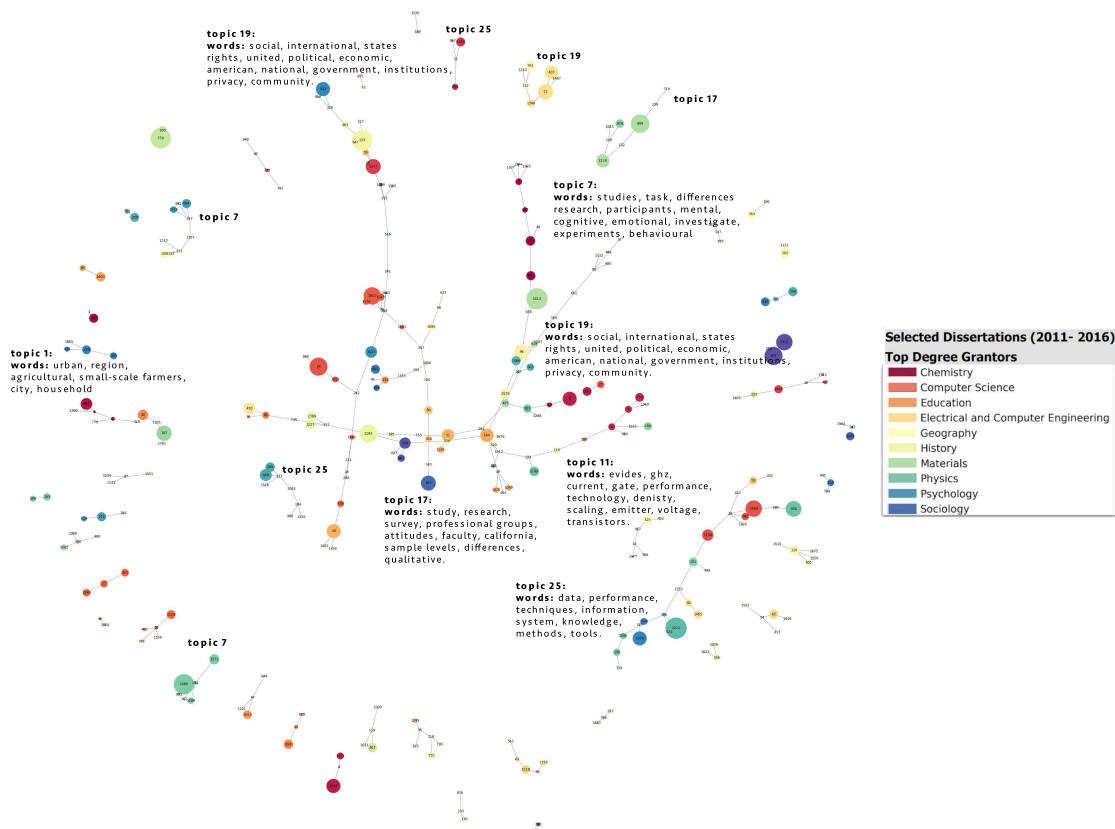


Figure 3.2: Theses (color coded by academic department) connected by their shared five top topics (shown with representative terms from word vectors)

¹²<https://networkx.github.io/>

¹³https://github.com/saralafia/adrl/tree/master/3_network

¹⁴<https://mybinder.org/v2/gh/saralafia/adrl/master>

3.5 Application: Discovering Thematically Related Research

The spatializations that we produce enable scholars to discover thematically related research objects, unlike the current ADRL, which does not offer any such capabilities. We apply the field and network concepts of spatial information to the motivating scenario offered in Section 3.2, referencing specific research objects related to the theses from the scenario. Patterns of relatedness are interpreted using Golledge’s spatial primitives of *distance*, *arrangement*, and *scale* [66], which have informed previous conceptual formalizations [45].

3.5.1 Questions Enabled by a Field of Research Topics

Both the field, in the form of a self-organizing map (SOM), and the research objects used to produce it enable the discovery of related research topics. Fields enable questions about value (i.e., research topic) at a given location. A continuous field function satisfies Tobler’s First Law of Geography [23], so that nearby topics in the SOM are similar. For pairs of objects, similarity can therefore be assessed by *distance*. Researchers interested in a particular area of research can see related theses by examining those closest to that area of interest in the SOM. Closely related research objects tend to fall within the area’s neighborhood (i.e. a single hexagonal topic location or an aggregate of several such cells).

In the case of Bae’s research from our scenario, the SOM displays six research objects from Geography, History, and Sociology within a neighborhood. Neighborhoods can be defined based on various distance thresholds. In addition to shared topics, relatedness may also reflect shared methods and techniques, as these are typically captured in abstracts as well. McKenzie’s research, for example, is in a neighborhood of research

objects from Computer Science and Electrical and Computer Engineering. While the subject matter of some research is different (e.g., photography or drugs), the theses share methods (e.g. spatial, data, search and learning, place, knowledge). Figure 3.3 illustrates these related research objects from the scenario, located in the SOM.

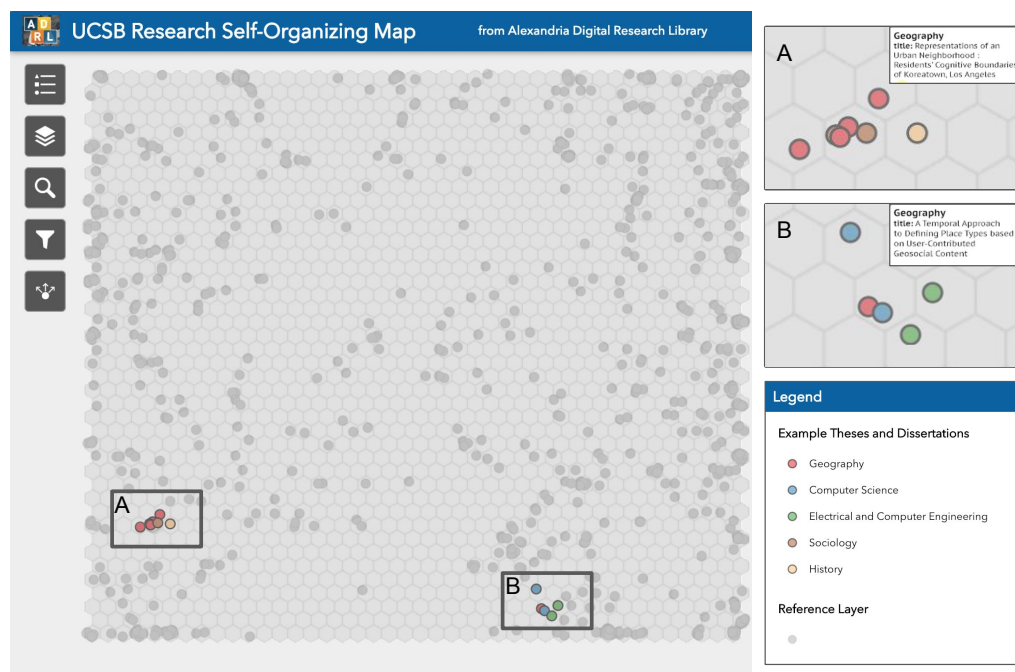


Figure 3.3: Selected theses (color coded by academic department) located in the SOM and surrounding: (A) Bae’s geography thesis; and (B) McKenzie’s geography thesis

Beyond similarity of themes or methods, *arrangement*, such as the dispersion or concentration of research themes in a topic space, are also demonstrated in the field view. Theses that address the urban, region, local topic are clustered and centered in the SOM, indicating that this topic pertains to many theses; conversely, topics (and their associated research objects) at the periphery of the SOM are less related to other research topics (e.g. dna, disease, peptide) and pertain to fewer theses. Compared with concentrated theses from other departments (like Materials, shown previously in Figure 3.1) the Geography department theses are dispersed; although Bae and McKenzies’ theses share topics (urban, region, local... and “models, based, system...”), they are still distant from each

other.

The field view with the thesis objects placed in it also reveals the presence and absence of research areas among existing theses. Portions of the field that do not contain any theses show research areas that may not be addressed in the repository, possibly suggesting interesting themes not yet studied and signaling opportunities for research at the boundary between disciplines. It should be noted that such gaps can also result from distortions in distance; cartogram techniques, which we have not yet applied to our field view, can be used to account for this by warping the SOM basemap [72]. Nonetheless, gaps between History and Geography surrounding Bae's research for example might suggest opportunities for integration of subject matter and techniques in this area (e.g., in the spatial humanities).

Scale in the field view is determined by topic modeling (number of input topics) and the parameters of the SOM (spatial resolution of the cells that locate topics). The size of the cells in relation to the whole field, and the dimensions of the field influence the position of topics and research objects. In our SOM, only one other thesis shares a top topic with McKenzie's research; this would likely change if the resolution of the cells changed, resulting in different topic groupings. Prevalent themes of research objects are visible at multiple levels. At the repository level shown in Figure 3.1, a prevalent topic appears to be about spatial, visual, search and relates to research across many departments, including Psychology, Geography, and Computer Science. Prevalent topics of departments can also be seen from the color coding of theses by academic department (rather than by academic advisor or year of publication, which would be other possible choices).

3.5.2 Questions Enabled by a Network of Research Objects

Questions about the similarity, distribution, and prevalence of research topics in a repository are handled in the SOM view; however, questions about explicitly modeled relationships between the objects are not. Networks deal with these questions by encoding the relationships in their edges: for instance, are the theses of Bae and McKenzie topically related, and if so how? Figure 3.4 illustrates how networks convey connectivity, showing topical correspondence between departments and topical diversity within departments.

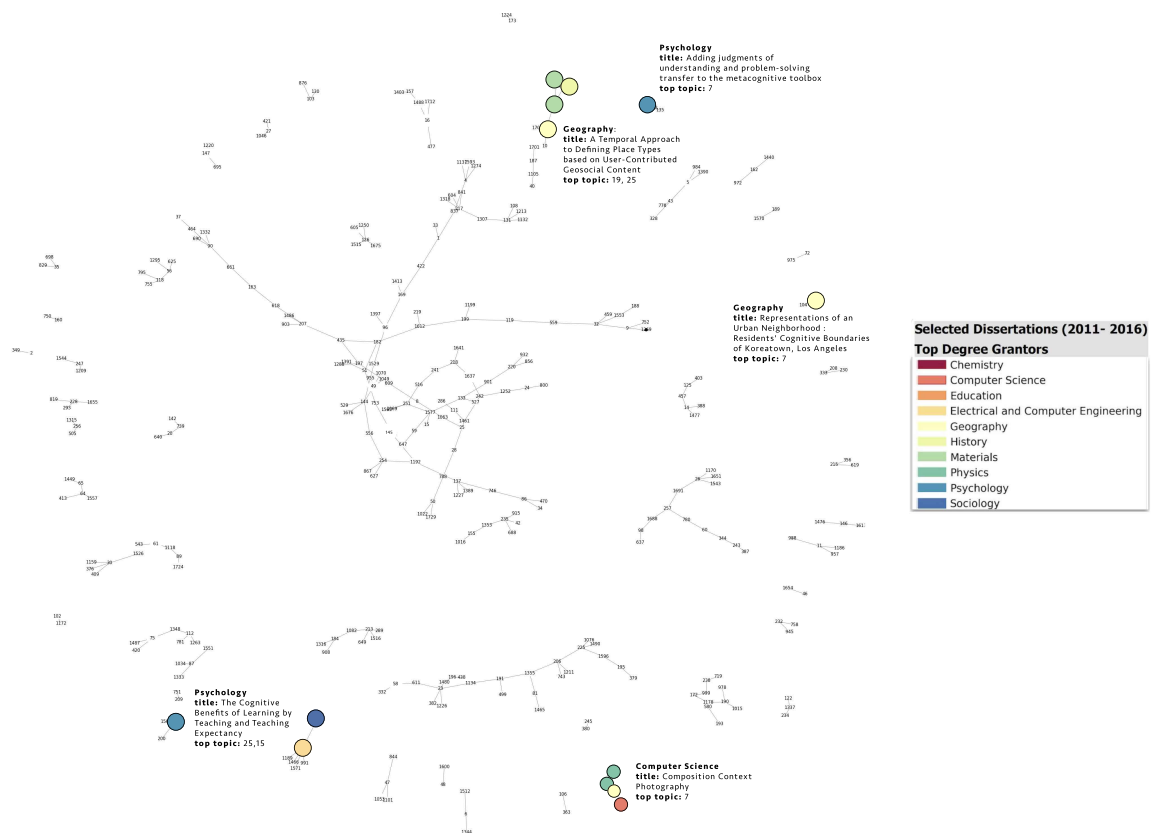


Figure 3.4: Selected theses (color coded by academic department and labelled by title) connected to other theses if the pair shares any five “top topics”

A network view answers questions about the specific relation encoded by network *distance*. Bae’s research is most thematically related to other research objects one edge

apart. A comparison between the network and the SOM shows additional similar theses, such as one from Marine Sciences, which is also characterized by the “urban, region, local...” topic.

In comparison to the SOM, where McKenzie’s thesis is located next to a Computer Science thesis, there is a larger distance between them.

In the network, McKenzie’s thesis is closer to History and Materials theses, characterized by Topics 19 (“international, social, political...”) and 25 (“data, performance, techniques...”). The Computer Science thesis shares a stronger topical relation with Geography and Physics theses, which are characterised by Topic 7 (“image, multiple, technology...”).

Arrangement is related to node embedding; the most central topics in the network visualization are shared by the most research objects. More specifically, topics that intersect the central region of the network are less specific than topics that describe multiple research nodes. Niche topics are pushed toward the edge of the network; thus, theses that are heavily characterized by these topics cohere to them strongly. As shown in Figure 3.2, research nodes occupying a central location in the network are characterised by the generic terms “study, research, survey...” and by “studies, tasks, differences...”. Conversely, theses such as those represented by the specific terms “work, material, particle...” share the fewest edges and therefore, are least central. Such theses deal with technical themes shared only by a few departments (in this case, those of Materials, Chemistry, and Electronic Engineering).

The *scale* of the network view shows a hierarchy with three levels: individual research objects, academic departments, and the repository as a whole. The nodes and the edge relations in a network can be defined in many ways. A node could represent a particular researcher and its attributes could be a list of theses published or supervised by the academic. Instead of representing a shared topic, edges could stand for a shared advisor,

creating a network of academic families or schools’. While the choice to restrict edges to five top topics was pragmatic, it also illustrates the flexibility of the design approach; any kind of binary relations between research objects can be visualized.

3.6 Conclusion

3.6.1 Summary

In order to enable discovery in a multidisciplinary setting, we develop two systematic spatializations that allow users to identify thematically similar research objects. These spatializations provide a helpful alternative to known-item search by facilitating exploration; they do not require users to have prior disciplinary knowledge. To produce them, we conceptually reframe existing spatialization techniques using core concepts of spatial information. From this reframing, we produce two applications: a self-organizing map of research topics (a *field* view) and a network of connected research objects (a *network* view). In both spatializations, the relatedness of research objects can be ascertained by their *distance*; nearby topics (in neighborhoods) or objects (separated by an edge) are more related. The *arrangement* of topics and objects in each spatialization also indicates their overall relatedness; central research topics or objects tend to be more shared, while those on the periphery are niche. Finally, *scale* in both spatializations is determined during pre-processing (e.g. number of topics in the model) and spatialization (e.g. cell size; node or edge assignment). While made systematic, these choices are parameters that can be reconfigured during subsequent analysis.

3.6.2 Outlook

Spatializations in library services enable thematic search for research objects and complement our previous implementation of geographic search for them. Spatializing research themes extends the power of spatial search from geographically-referenced information into topic spaces, formalized in this work by core concepts of spatial information: *fields* and *networks*.

Information displays that index research by theme, location, and time [29] enable scholars to ask novel questions. The relatedness of research, indicated by proximity either in geographic location (e.g. Central American archaeology and entomology research) or thematic location (e.g. archaeological excavations of diverse ancient cultures) shows the potential for interplay between thematic and geographic views that our work enables. Furthermore, we envision allowing users to explore the spatializations in combination, gaining distinct yet complementary views of the same repository. While exploring a self-organizing map (SOM), a user can gain an overview of topics in the repository and from this, identify a specific area of interest. The subset of research objects falling into that area of the SOM can then be explored in the network, enabling further interrogation of connections. We are interested in assessing the design of our spatializations using standard usability tests, where test subjects are given questions to answer with each spatialized theme.

Temporal visualization beyond time sliders may also play a role in enabling research discovery. The meaningful representation of time-varying information [73] in disciplines like the digital humanities is notoriously fuzzy, inconsistent, and spatially variable [74]. We envision using temporal information inhering in research theses (e.g. publication date; events or periods studied) to be modeled by *events* and support reasoning on periods (time intervals).

Time, made explicit and linked to spatializations, could show how research topics evolve in geographic and thematic spaces. However, events do not yet seem to provide a useful metaphor for spatializations, as they are notoriously difficult to visualize in static maps.

Visualizing the quality (as opposed to the content) core concepts of spatial information, which include *granularity*, *accuracy*, and *provenance* [33], suggests many directions for future spatialization work. *Granularity*, or level of detail, relates both to geographic scale and to the coarsened or refined topics shown in spatializations. At present, granularity provides a clear and important intuition, as it relates directly to visualization (e.g. detail on demand). *Accuracy* relates to validity, possibly determined through comparison of multiple spatializations against domain ontologies. Finally, *provenance* may provide a way to explore the lineage of ideas (e.g. discovering related research through co-citation networks).

The long-term goals for this work are to increase awareness of relevant previous or ongoing research by applying spatial thinking to the discovery of thematically related work. Integrating research by spatialized topic, rather than siloing it by discipline, is likely to enable increased collaboration across academic disciplines. Much like browsing stacks of books in a physical library, exploring a spatialized library repository can transform a common research task into a learning opportunity or a serendipitous discovery.

Chapter 4

Mapping Research Topics at Multiple Levels of Detail

Abstract: Institutional reviews typically rely on scientometrics, like the h-index and impact factors of their participants, to assess research productivity. Productivity is not the only review criterion however, and scientometrics can be difficult to generate and compare in multidisciplinary settings. Distant reading methods from the Digital Humanities can complement the current quantitative evaluation paradigm; these methods support qualitative narratives, comprehension, and discovery of knowledge by arranging vast bodies of text into graphs, maps, and trees. To test this idea, we apply distant reading methods to a multidisciplinary body of research authored by 240 researchers from the Earth Research Institute (ERI) at UC Santa Barbara over the past decade. We model cross-disciplinary topics of research publications and projects emerging at multiple levels of detail. From these, we design maps that reveal the latent thematic structure of multidisciplinary research. ERI’s researchers use and evaluate these maps of research topics in the context of an institutional review to read ERI’s body of research at a distance, i.e. at multiple levels of detail. We find that our approach strengthens

the institutional review process by exposing thematic expertise, relationships between researchers, topical distributions and clusters of work, and the evolution of these aspects over time.

4.1 Introduction and Motivation

Universities and funding agencies request that organized, multidisciplinary research units summarize and report on their research, collaboration, and growth as part of periodic institutional reviews. Such reviews conduct meta-research on grant proposals, publications, and other products to provide a high-level view of research impact and productivity [75]. Yet, it is unclear how best to summarize and present multidisciplinary bodies of research in ways that generate useful insights and support effective reviews.

Bibliometrics and scientometrics supporting the quantitative study of published documentation and academic disciplines [76] have become cornerstones of institutional research assessments. Research administrators and funding agencies often use metrics, like the Hirsch Index (h-index) and the journal impact factor (JIF), to assess the impact and performance of departments or individual researchers and to monitor their collaborators or competitors [77]. Such metrics are trusted due in part to their perceived scientific legitimacy and because they offer indicators, which if appropriately selected and applied, can yield data to support performance monitoring and the selection of research priorities [78].

There are limitations to evaluative metrics, however. For one, metrics based on citation analysis are often not comparable across academic disciplines [79]. Incentivized metrics like citation counts, impact factors, and amounts of funding have been found to be vulnerable to manipulation [80]. Academics subject to performance evaluations have also criticized institutions for relying on such metrics at the expense of narratives [81].

Given these concerns, there is a growing desire to develop principles for multidisciplinary research evaluation that more adequately capture impact and quality. One strategy has been to complement evaluative metrics with high-level characterizations and narratives [82]. Another has been to develop knowledge domain visualizations, or maps of science, that chart the structure of vast realms of scientific literature by showing the development of research areas, their interconnections, and evolution [46].

Mapping is indispensable in many monitoring and planning contexts; without maps of the physical territory, it would be challenging to plan and manage the development of cities, landscapes, and infrastructure. Cadastral maps, for example, document ownership and other rights to the land; they also inform and communicate numerous planning interventions including strategic land use decisions, economic investment, and mitigation measures [83].

In general, space and time are fundamental ordering relations for knowledge representation [84]. The spatial turn observed in the social sciences and humanities has exploited the idea of spatial organization to facilitate cross-disciplinary exchange, allowing many lines of thought to converge [85]. In cognitive science, it has been claimed that conceptual spaces in which nearby concepts are similar, underlie human thinking and learning [8]. These powers of spatial representation underpin the idea of *spatialization*, which maps abstract domains to map spaces in which nearby elements are similar [37].

In this article, we examine the utility and benefits of spatialization for mapping the latent thematic structure of a multidisciplinary body of research. We further explore its potential to complement current quantitative approaches by aggregating research topics at multiple levels of thematic detail. To develop and test our ideas, we situate our research in the context of an organized research unit at UC Santa Barbara (UCSB): the Earth Research Institute (ERI)¹. ERI's stated mission is to support research and education

¹<https://www.eri.ucsb.edu/>

in the sciences of the solid, fluid, and living Earth. Core areas of research within the institute consist of Natural Hazards, Human Impacts, Earth System Science, and Earth Evolution. ERI supports faculty and researchers spanning 24 academic departments who are supported by 145 different funding agencies, which encompass the full breadth of Earth and Environmental Sciences. To date, ERI has taken an ad-hoc approach to characterizing their research based in part on anecdotal observations. For example, ERI expertise has recently broadened from earth science and crustal studies to include conservation and biodiversity topics.

To formally capture this kind of institutional knowledge about ERI’s evolving research, we propose a systematic, reproducible, and data-driven approach for eliciting cross-cutting research topics. We analyze research documents (publications and funded projects) by 240 researchers affiliated with ERI between 2009 and 2019 to produce research maps at multiple levels of detail. We then evaluate whether they are useful to researchers within the institution.

In the remainder of this article, we first situate our work in relation to existing approaches for abstracting and mapping information. We then describe our method to produce maps of a body of research at selected levels of detail. Finally, we report how leading ERI researchers evaluate the potential for our maps to support an institutional review². We find that our approach complements the review process by exposing and relating thematic expertise, highlighting relationships between projects or teams of authors, analyzing topical distributions and clusters of work, and tracking the evolution of all these aspects.

²A slight delay in the actual institutional review (resulting entirely from COVID-19) precluded feedback from external reviewers in time for our research project.

4.2 Background

Topic modeling and spatialization jointly support exploration and discovery at multiple levels of detail by performing dimensionality reduction. This is analogous to cartographic generalization for maps: computational and cognitive issues of complexity are addressed by deliberately reducing the level of detail in the representation [43]. Topic modeling offers a way to identify research topics latent in articles and projects that are not bounded by traditional silos, like academic departments and their terminologies. We use topic modeling to choose levels of detail and identify major topics shared by research documents. Spatialization is a means of abstracting large amounts of data to give a visual impression of their underlying structure, in particular similarity between concepts. Spatialization allows us to visualize these topics in such a way that nearby topics are more similar than distant ones. These methods jointly and sequentially support our goal of summarizing large amounts of thematic data and connections taking place at multiple levels of detail.

4.2.1 Topic Modeling

Topic models are statistical machine learning techniques that can uncover hidden structures in collections of documents; topic models group documents in which similar terms co-occur, forming sets of semantically related documents [86]. Latent Dirichlet Allocation (LDA) is a ubiquitous type of topic model in which documents are described by multiple topics [13]. Topic models have been applied to classify and summarize large collections of documents, as well as solve similarity judgement problems [13]. Topics themselves can also be of interest; for example, the National Institutes of Health and National Science Foundation have developed topic-based search interfaces to explore trends in related research projects [86].

4.2.2 Spatialization

Spatializations offer high-level views of contents through the familiar visual modality of maps [35]. The first law of cognitive geography, or distance-similarity metaphor, references the first law of geography, which states that “everything is related to everything else, but nearby things are more related than distant things” [23]. The distance-similarity metaphor treats distance in abstract spaces as metaphorically equivalent to dissimilarity [10]. Spatialization has been applied to organize large, thematically diverse, and multidimensional collections; for example, Figure 4.1 shows political survey data from Switzerland spatialized as a political landscape, resulting in a map of continuous ideological values with nearby cities sharing preferences [87].

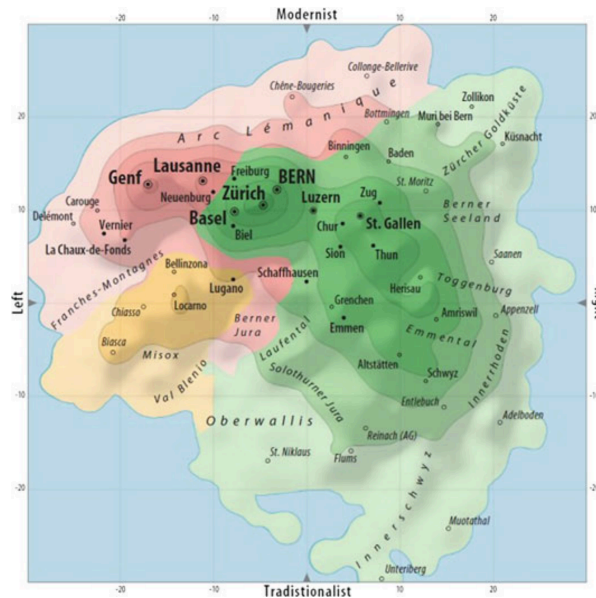


Figure 4.1: Spatialization of Swiss political ideology as a landscape (*Source:* Hermann and Leuthold, 2003; reprinted with author permission)

4.2.3 Organizing Spatializations by Core Concepts of Spatial Information

To further systematize spatializations, we apply the theory of core concepts of spatial information [33]. The concepts provide a high-level vocabulary with which to ask and answer questions about phenomena in space and time (Table 4.1). They capture distinct ways of thinking about space and computing with spatial information; thus, they are applicable to geographic as well as other spaces. They provide a set of interchangeable lenses through which research data can be spatialized and viewed [88].

We produce a *field* of continuous topic values from the texts of research documents with a topic value at each position. This can be thought of as a landscape or surface of topic values. Research documents conceptualized as *objects* can then be embedded in this topical landscape. *Networks* of research documents may be linked, for example by shared topics or authors. *Events* may be introduced to answer questions about what has happened, is happening, or may happen to researchers and departments, for example.

In our previous work, we found that mapping research documents as objects embedded in fields revealed their distribution and coverage, while linking them in a network revealed their topical connectivity and centrality in the collection, among other insights [48]. In this article, we concentrate on fields and objects, but further demonstrate the relevance of core concepts of spatial information by enabling the detection of events through a timeline.

4.2.4 Level of Detail

Previous studies have shown that levels of detail in spatialized displays, such as hierarchical regions, shape viewers' interpretation of the similarity of elements like news articles [10]. The problem of choosing adequate levels of detail for topic models and

Concept	Definition	Intuition	Question	Examples
Location	a description of where things are by spatial relations	things are located relative to each other or in reference frames	where is it?	in the center of town; at a latitude and longitude
Field	an attribute with values everywhere in a region and at all times during a period	fields continuously map positions and times to attributes	what's the value at a given position and time?	today's air temperatures at 8am everywhere in the state
Object	an individual in space and time, with properties and relationships	objects have identity	what's the value of an object property at a given time?	a building with address and owner
Network	a set of objects with links between pairs	networks capture connectivity	what connects two objects at a given time?	a bus network in a city
Event	an occurrence at some time, involving participants (fields, objects, and/or networks)	events change participants	what's the value of an event property at a participant location?	a tornado, an epidemic, a house sale, a road closure

Table 4.1: Summary of core concepts of spatial information

spatializations is unresolved.

The selection of an appropriate number of topics is a central challenge in topic modeling; selecting too few leads to overly broad topics while selecting too many leads to redundancy [89]. Best practices recommend a combination of human evaluation strategies and topic coherence measures [86].

Spatialization relies on generalization methods for merging individual features into groups. This is analogous to cartographic generalization, which performs hierarchical

clustering based on feature similarity and results in changing geometric representations and labels for the features at each level of detail [43]. The strict membership requirements imposed by hierarchical clustering may not be the best way to represent relationships between levels of detail in abstract spaces, however. Alternative tree-like structures, like semi-lattices, could more adequately represent concept organization, for example as sets of partially overlapping information [90].

4.2.5 Distant Reading as an Abstraction Paradigm

“Distant reading” [11] provides methods for deliberately abstracting and visualizing unstructured text; in order to analyze hundreds of novels for example, it is necessary to render fewer elements in order to offer a sharper sense of high-level themes and their interconnection [91]. Distant reading uses graphs, maps, and trees to spatially configure units, like genres and novels, and reveal latent structures in their source material [11]. These methods are part of a suite of macroscopic research devices, including knowledge domain visualizations [46] and spatializations, that enable the study of patterns at multiple levels of detail over time.

Graphs, maps, and trees offer a typology of spatial and temporal views that support the exploration of large bodies of literary texts. Graphs are quantitative diagrams that produce temporal signatures revealing patterns and cycles like the rise and fall of literary genres. Maps are spatial diagrams that order components to reveal structures in the space of the narrative universe. Trees are morphological diagrams that emphasize the passage of time, revealing divergence and drift of literary styles. We believe that these models can also guide abstraction over scientific texts, which in our case are large numbers of abstracts from publications and grant proposals. Our application appears to be less ambitious, but deals with the same type of raw data: natural language texts following

some conventions of style.

While texts are the real objects of literature, they may not be the right objects of knowledge for literary history [11]. By analogy, in the institutional review of a research unit, available data about publications and grants are the real objects of science, but are not the right focus for understanding it. Currently, institutional reviews focus on productivity metrics, which do not address this problem. Abstraction is necessary to support an understanding of broader patterns and trends latent in research documents, such as the emergence and evolution of shared topics.

4.3 Methods

We develop methods for producing maps that support the distant reading of research activities at distinct levels of detail. These maps show research topics and their evolution over time. To produce the maps, we first retrieve research documents and model the topics of their titles and abstracts. We then feed the output of topic models at two distinct levels of detail into spatialization algorithms to produce maps of the research topics. The data and code for our experiments with topic modeling and spatialization, which inform our approach, are publicly available in our project’s Github repository³.

4.3.1 Data Sources

We analyze publications and funded projects from ERI’s 240 researchers active from 2009 - 2019. These data from the past decade cover research activities and financial aspects including: publications issued by the unit, funded project proposals, grants, and contracts. We gather publication metadata with the Dimensions API, which is freely available for non-commercial use. ERI maintains records of active researchers and

³<https://github.com/saralafia/ERI-maps>

their funded projects. We only consider publications and funded projects with titles and abstracts, leaving 3,770 research documents for analysis (3,108 publication abstracts with titles and 662 funded project abstracts with titles).

4.3.2 Text Pre-processing

We prepare the research document titles and abstracts for topic modeling by removing records with identical identifiers (DOIs), removing HTML tags, and reformatting ASCII extended characters. Figure 4.2 shows that the research documents are relatively concise with an average length of 1,678 characters; they range in length from 128 to 7,083 characters.

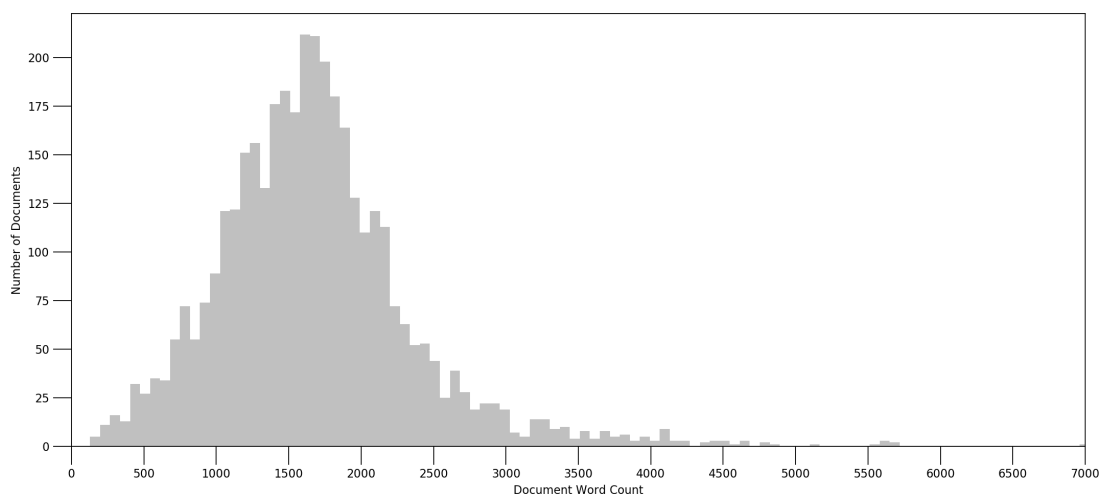


Figure 4.2: Distribution of research documents by length (word count)

Next, we reformat the research documents following a standard natural language processing pipeline [86]. We first determine distinct document terms using term frequency-inverse document frequency, or tf-idf [92]. This reflects the relative importance of a term to a document in a corpus and is often used as a weighting factor in information re-

trivial applications; this measure seeks to balance specific terms that show up frequently in relatively few documents (e.g. polymerase) with those that show up frequently across many documents (e.g. sample). Frequent terms, scaled by how often they appear in the documents, are shown in Figure 4.3. We find that many of these frequent terms describe research methods (e.g. estimate) rather than subject matter (e.g. snow).

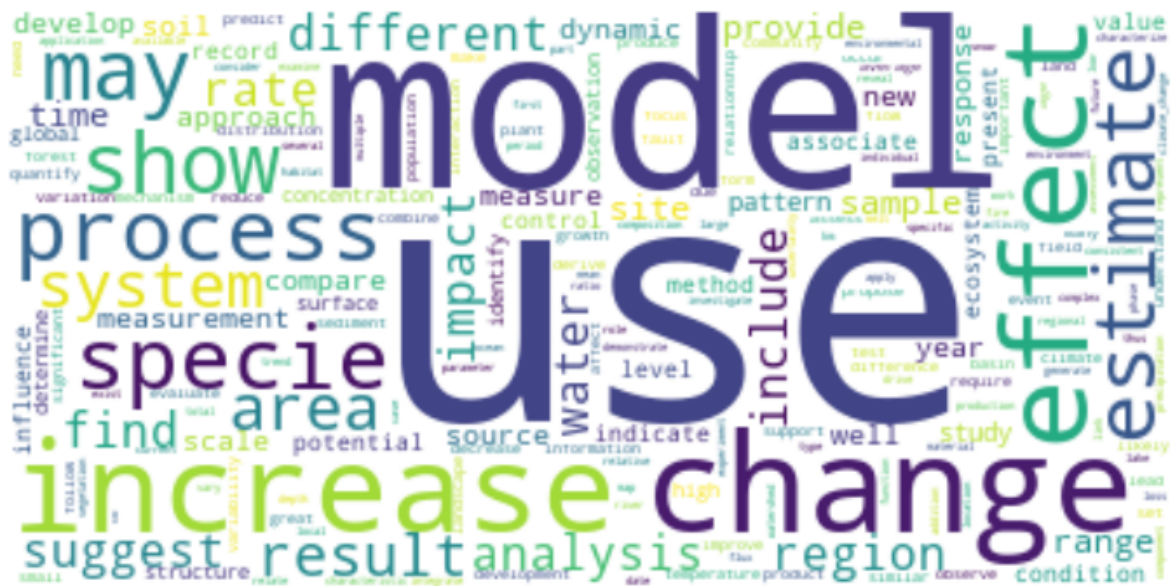


Figure 4.3: Word cloud of frequent terms from the processed research documents

We remove common articles from the documents, as well as frequent and generic terms identified with tf-idf that qualify as stopwords. These terms are ‘data’, ‘study’, ‘project’, ‘research’, ‘collaborative’, ‘include’, ‘result’, ‘increase’, ‘high’, ‘low’, ‘large’, ‘include’, ‘based’. We then tokenize the terms, convert them to lowercase forms, and construct n-gram models (bigrams, trigrams) to preserve contiguous sequences of words (e.g. ‘climate change’). Finally, we lemmatize terms to resolve them to their base forms (e.g. ‘changing’ to change). These steps produce a dictionary of word identifiers and word frequencies along with a document corpus to use in topic modeling.

4.3.3 Topic Modeling

We consider two main kinds of topic modeling approaches: probabilistic (Latent Dirichlet allocation, hierarchical LDA) and matrix factorization (non-negative matrix factorization). In probabilistic approaches, each document is treated as a mixture of a small number of topics; words and documents get a probability score for each topic. The LDA algorithm is a generative probabilistic model that represents documents as mixtures of topics composed of words with certain probabilities [13]. It assumes that similar words occur in similar contexts and aims to discover latent topics underlying the documents. Given our interest in modeling topics at multiple levels of detail, we also consider extensions of LDA proposed for learning topic hierarchies like hierarchical LDA, or hLDA. It estimates the structure of a topic hierarchy and partitions documents nonparametrically [93].

In matrix factorization approaches, dimensionality reduction methods from linear algebra are used to decompose a document-term matrix into a smaller set of matrices, which can be interpreted as a topic model [12]. Non-negative matrix factorization (NMF) is a dimensionality reduction technique for decomposing samples, which in the case of topic modeling are documents. Like in LDA and hLDA, documents are represented as term vectors, which can be combined into a document term matrix. However, documents are represented as combinations of topics rather than likelihoods. In NMF, term weighting can also be used to boost distinctive terms by term frequency inverse document frequency, or tf-idf [92]. LDA has been positioned as an alternative to tf-idf, offering improved insights into inter- or intra- document statistical structure [13].

To determine a suitable number of topics to model, we consider the cognitive implications of Miller’s Law, which proposes that the average person can hold approximately 72 “chunks” of information in working memory (e.g. 7 digits, 6 letters, 5 words), limiting

the simultaneous perception and processing of information by humans [94]. Miller’s Law, applied to our topic models, suggests a coarse level of detail (72 topics) that reviewers should be able to consider at once. For a suitable number of topics at a more detailed level, we reapply Miller’s Law to each chunk of the coarse level, resulting in bounds of (5 x 5) and (9 x 9), or a range of 25 - 81 chunks.

We use these heuristics to guide the selection of models in combination with coherence as an interpretability measure. Coherence is a measure of the extent to which top terms representing a topic are semantically related, relative to how they relate to other terms in the corpus [95]. Coherence is considered to be more human interpretable for evaluating topic model quality than other measures, including perplexity and log-likelihood [89]. It is based on the fundamental idea in classification that the members of a class should be more similar to each other than to members of other classes. We generate topic models with high coherence scores using each technique, which allows us to select and compare topic models at different levels of detail [95].

Our topic modeling experiments are detailed in reproducible Jupyter notebooks in our project repository⁴. We first produce LDA models, selecting those with high coherence scores. We find topics related to methods like “modeling”, “analysis”, and “task” that are not indicative of ERI’s research subjects. This may be an artifact of higher frequencies of methods-related terms like modeling, which were prominent in the word cloud (Figure 4.3). In LDA, such frequent terms are not weighted or normalized as they are in other topic modeling approaches like NMF.

Next, we explore hierarchical LDA, or hLDA. We find that the nonparametric partitioning of the corpus produces broad, generic topics at each level. hLDA also overemphasizes research outliers by distinguishing them into separate and comparatively small categories. It seemed to be a promising technique, given its ability to model topics at

⁴<https://github.com/saralafia/ERI-maps>

multiple levels of detail within a single model, but we decided against using it due to its unsatisfactory interpretation of ERI's broad field of multidisciplinary research. hLDA may prove to be useful for modeling topics in a more constrained field of research, however.

Several factors suggest that non-negative matrix factorization (NMF) is a more suitable topic modeling approach for our purposes than LDA or its variants. We find that NMF produces topic models with higher coherence scores than our LDA models by about 17 percent on average. This may be because NMF is better suited to model smaller or sparser datasets, like titles and abstracts, rather than full text [96]. We also find that NMF produces topics that are more indicative of subject matter, rather than methods. This may be due to term weighting with tf-idf, unlike LDA, which operates on raw term frequency [89].

To find coherent topic models, we generate NMF models across a range of topic numbers (2 - 100) and calculate their coherence scores. Figure 4.4 shows a rapid increase in model coherence between 2 - 10 topics, a more gradual increase in coherence between 20 - 40 topics, and a continual, but modest increase in coherence up to 100 topics. While the model with 100 topics has the highest coherence score, it does not obey our usability constraints; its large number of topics exceeds the number of chunks of information that reviewers would be able to keep in mind. Instead, we select two distinct levels of topic models: a coarse level with 9 topics (falling in the 5-9 topic range) and a more detailed level of 36 topics (falling in the 25-81 topic range), which are optimally coherent.

We generate the NMF models using Scikit-learn decomposition⁵ as follows. Non-negative matrix factorization decomposes a document-term matrix into a smaller set of matrices, which can be interpreted as a topic model (Lee and Seung, 1999). We normalize the document-term matrix with tf-idf and decompose it into a product of two matrices:

⁵<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html>

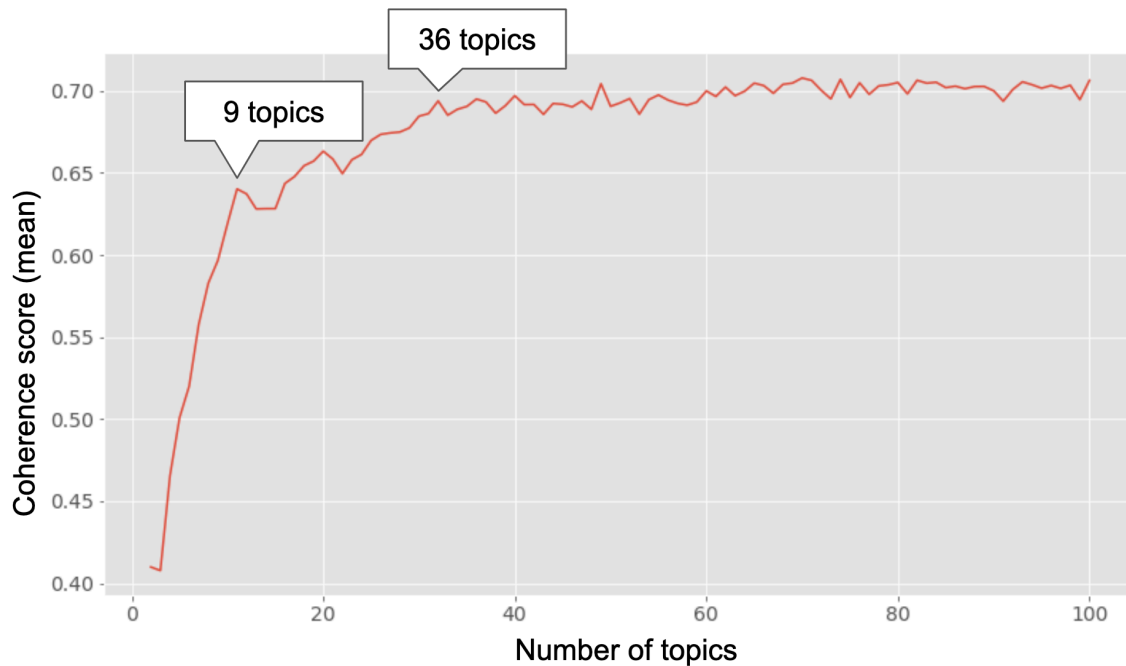


Figure 4.4: Coherence scores for NMF topic models with 2 - 100 topics

a topic-term matrix and a document-topic matrix. Figure 4.5⁶ shows how terms and topics can be combined to approximately reconstruct documents. The topic-term matrix shows the topics associated with particular terms; the highest ranked terms are used to obtain a descriptor for each topic based on the 10 top terms for each. The document-topic matrix describes each document by a mixture of topics and shows how strongly it is associated with a topic or set of topics. The document-topic matrix forms the input to the subsequent spatialization while the topic-term matrix is used to reference topics and terms from their descriptors.

⁶Adapted from Wikimedia (CC BY-SA) <https://upload.wikimedia.org/wikipedia/commons/7/7c/NMF.svg>

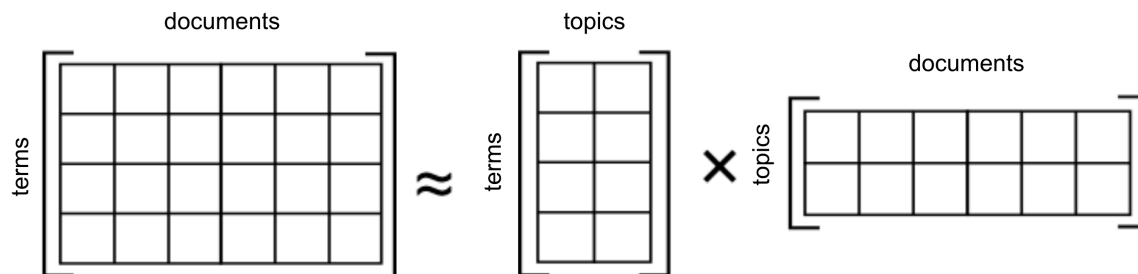


Figure 4.5: Conceptual overview of NMF decomposition of the document-term matrix

4.3.4 Spatialization

In our approach, the topic models serve as inputs to spatialization. Latent topics and their proximity can be represented in a number of ways, making the view of space underlying a spatialization a key choice [33]. Space can be seen 1) as a field, or a continuous surface, with a topic value at each position; 2) as a collection of objects, which are located in an abstract space; or as 3) a network, which links pairs of objects to show a relation of interest like connectivity, hierarchy, degrees of separation, or flow.

To present the high dimensional topic mixtures of research documents in two-dimensional spatial configurations, we consider graphs, maps, and trees as distant reading methods [11]. The methods to produce Moretti’s graphs, such as histograms and signatures, are well-understood but we decide not to pursue them because we do not have enough longitudinal data to support significant insights from graphs; our research documents span only 10 years. We develop a non-temporal variation on Moretti’s trees as part of our evaluation (Section 4.5.3) to study the relationship between topics at two levels of detail; instead of showing topic evolution, we use these variations on trees to show topic clustering. Our spatializations produce maps, which handle time and events in the way maps traditionally do through a sequence of annual snapshots.

To produce maps, we conceptualize research documents as objects located in a contin-

uous two-dimensional topic space. The topic mixture of each research document assigns it to a neighborhood of related research topics. To embed the research documents in a continuous topic space, we considered two dimensionality reduction techniques: t-distributed stochastic neighbor embedding, or t-SNE [14] and uniform manifold approximation and projection for dimension reduction, or UMAP [97]. Both t-SNE and UMAP are suitable for our purposes, as they model high-dimensional objects as points in a low dimensional space, clustering similar objects and spacing apart dissimilar ones.

We considered each of these techniques, evaluating how well they convey the similarity of research documents; these experiments are detailed in reproducible Jupyter notebooks in our project repository⁷. The inputs to the spatializations are the document-topic matrices resulting from the coarse (9) and detailed (36) NMF topic models. We first map research documents with t-SNE using manifold learning in Scikit-learn⁸. The t-SNE algorithm transforms the high dimensional document-topic matrix into a low dimensional coordinate representation. Each document is assigned a position based on its topic mixture, resulting in the placement of topically similar documents near each other and dissimilar documents farther apart. The UMAP process for assigning locations to research documents is nearly identical to that of t-SNE; a key difference is the assumption that documents are uniformly distributed on a complex surface, resulting in a distinct spatial configuration. We produced these with UMAP learn⁹. The axes in both t-SNE and UMAP are left unlabeled, as they describe complex curved paths in the original high-dimensional space and do not have human-interpretable meaning [14, 97].

We interactively explored the maps to interpret the effects of the map parameters, which balance local, pairwise similarity with global, inter-cluster similarity [98]. The first parameter influencing the size, distance, and shape of clusters is perplexity, which controls

⁷<https://github.com/saralafia/ERI-maps>

⁸<https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE>

⁹<https://umap-learn.readthedocs.io/en/latest/>

the number of nearest neighbors. Perplexity describes how well a probability distribution predicts a sample. In our maps, low perplexity values produce clearly delineated clusters while high ones allow for more global connectivity and less clearly delineated clusters. We select a perplexity of 7; typical values fall between 5 - 50 [14]. The second parameter is early exaggeration, which determines the compactness of clusters. This is an optimization method to create empty space in the map between clusters so they can achieve better global, in addition to local, organization [14]. We select an early exaggeration of 5; the default value in the Scikit-learn implementation is 12.

Both mapping approaches appear to produce similar results in terms of the spatial layout of documents and clusters with a few notable differences. Figure 4.6 shows that t-SNE produces local clusters of similar objects that are visually distinct, while UMAP allows for more outliers and preserves compact clusters; for instance, all red documents clustered and labeled with fault (seismic motion) are concentrated in UMAP, while they are split into three distinct clusters in t-SNE. The effects of uniform spacing are also visible in UMAP; the red and blue clusters are disjoint in UMAP but are partial neighbors in t-SNE. The arrangement of individual documents and clusters of documents in t-SNE conveys topical similarity well. Based on these observations, we deem t-SNE to be a more compelling technique for our purposes.

4.4 Results

Our methods address the question of how to systematically elicit and represent the major topics of a complex, multidisciplinary body of research at multiple levels of detail that show their similarities and evolution over time. They result in maps of research documents located in a continuous topic space, which exhibit topical proximity in neighborhoods and capture multiple levels of detail over periods of time. We explore whether

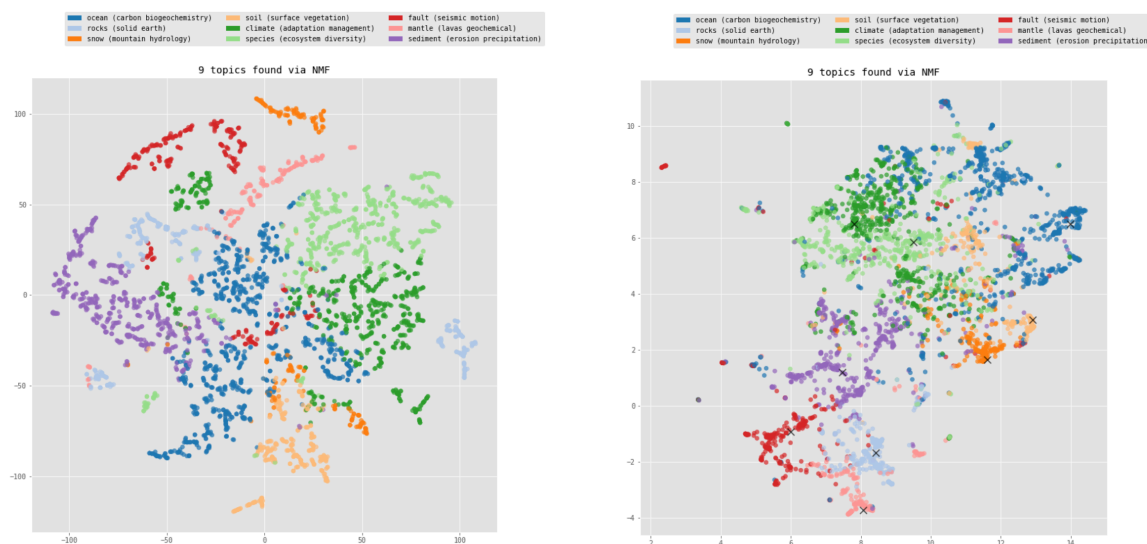


Figure 4.6: ERI research documents clustered by 9 topics with t-SNE (left) and UMAP (right); each color corresponds to the document’s main topic, labeled with three term descriptors

and how these maps of research topics support the qualitative assessment of a multidisciplinary body of research.

4.4.1 Reading Maps of Research Documents

The maps produced with t-SNE show research documents with similar topics forming neighborhoods at two distinct levels of detail. Documents are assigned to topic clusters, which are labeled with the first three terms from their topic descriptor. Topic modeling does not produce labels for the resulting topics, so this is a pragmatic choice that allows us to reference and interpret the topic clusters. The categorical colormap¹⁰ offers perceptually distinct categories for visualizing the relatively large number of topics in the detailed topic model.

In the coarse map with 9 topics shown in Figure 4.7, we observe patterns related to the centrality, size, contiguity, and proximity of clusters. Documents assigned to the

¹⁰<https://colorcet.holoviz.org/>

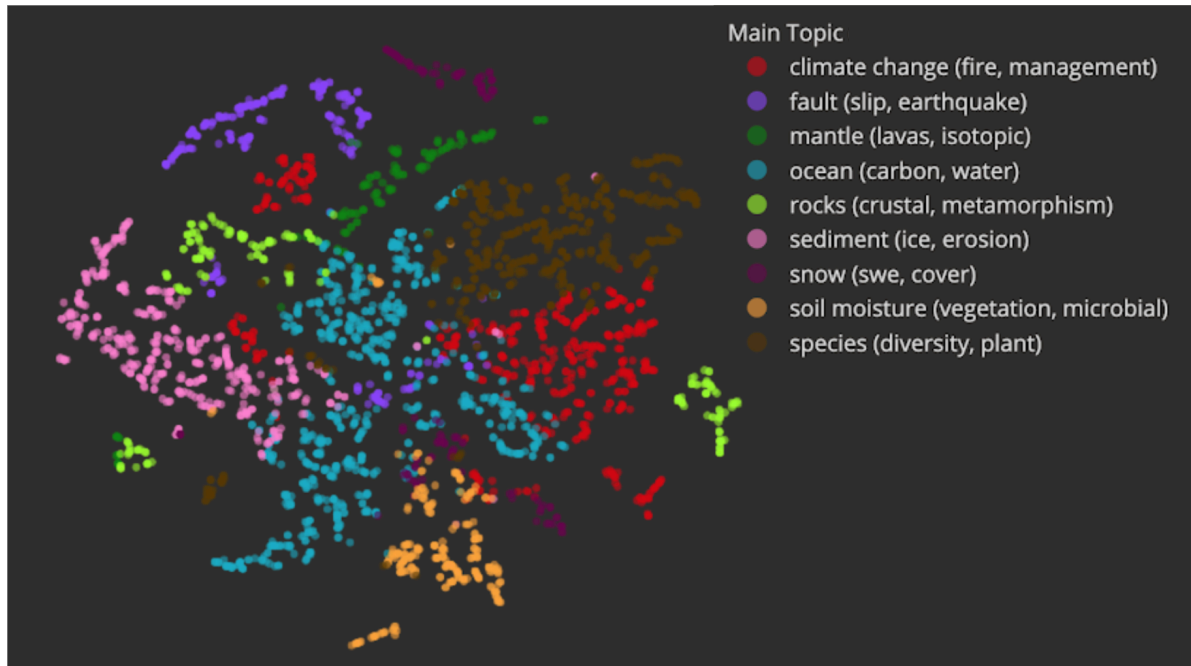


Figure 4.7: Coarse (9 topic) map of research documents (2009 - 2019)

large ocean cluster are in the center of the map while smaller clusters like snow, are on the periphery. This suggests that the documents described by the ocean topic are similar to more documents in the corpus than those assigned to the snow cluster, which may be more niche.

The cluster labeled rocks is small and discrete compared with the species cluster, suggesting that more of ERI's research is ecological rather than geological in nature; however, these disciplinary identities are not mutually exclusive. Documents can be characterized by more than one research topic in the map.

Documents in the soil moisture cluster are uniformly located in a similar region of the map, while others like those in the climate change cluster, are dispersed and non-contiguous. This suggests a lack of internal conformity within this cluster. Lower document dispersion in the soil moisture cluster suggests topical homogeneity, while higher dispersion in the climate change cluster suggests more heterogeneous documents.

The adjacency of the sediment cluster with the rocks, climate change, and ocean clusters suggests that its documents straddle, and sometimes bridge, these research areas, particularly those on the clusters' edges. Clusters located farther apart are also dissimilar. The snow and the soil moisture clusters are found on opposite sides of the map; however, other documents described by these topics are neighboring at the bottom of the map, converging around an edge of the climate change cluster. Indeed, the documents found there bridge these areas; they address snowmelt, surface temperature in forests, biomass accumulation, streamflow changes, and other related ideas.

While the coarse map presents a distant overview of ERI's research topics, the detailed map shown in Figure 4.8 reveals intricate patterns. The center population cluster borders other research areas including the species, ocean, and fisheries clusters. Another multi-topic cluster found at the bottom map periphery gathers similar public policy research from different topics, like mitigating climate change impacts on fisheries and earth system science in Canada.

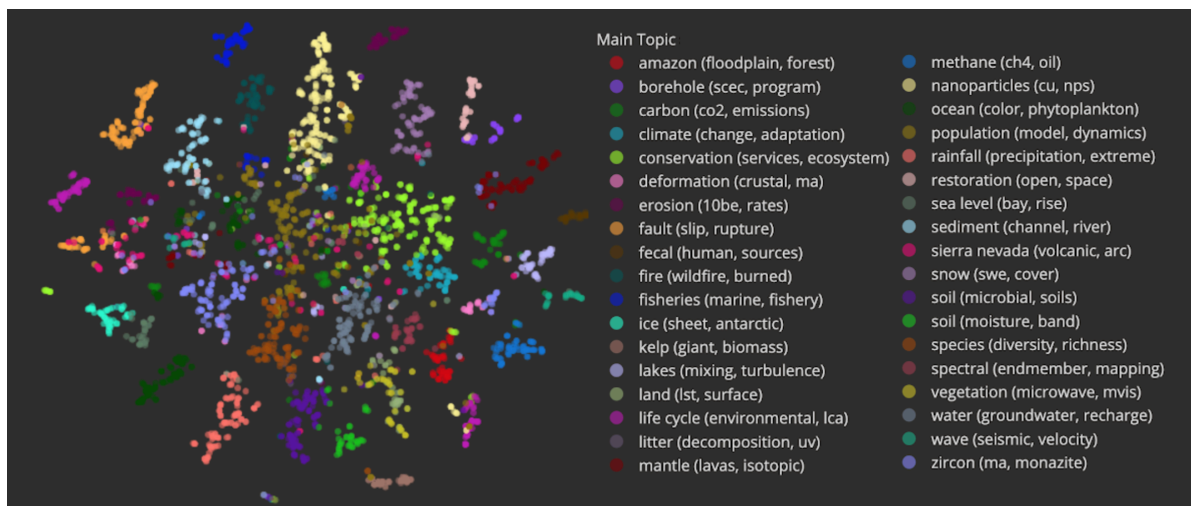


Figure 4.8: Detailed (36 topic) map of research documents (2009 - 2019)

The detailed map is made up of relatively even distributions of topic clusters. One exception is the fecal cluster on the right edge of the map, which is small and separated;

its nearest neighbor is the lakes cluster below it. A larger nanoparticles cluster at the top of the map is associated with ERI’s productive Center for Environmental Implications of Nanotechnology.

Central clusters tend to be less uniform than those at the edges. The water, conservation, and methane topic clusters are interspersed with documents addressing marine isotopes, stream mapping at a battlefield conservation site, and stream nitrate concentrations in mountainous watersheds. This is contrasted with the homogeneous clusters found at the edges, such as the ice cluster on the left edge dominated by documents addressing glaciers.

In the detailed map, we see that there are distinct, yet adjacent, areas of research involving similar researchers and shared ideas, such as integrating wildfire risk with the study of agricultural encroachment. The conservation and fire clusters are adjacent in the detailed map; in the coarse map, these documents fall under the climate change topic. In the detailed map, most fire research documents border the sediment and fisheries clusters, suggesting that documents about wildfire recovery and river restoration share similarities.

We have presented maps at two selected levels of detail - coarse (9 topics) and detailed (36 topics) - as described in our methods (Section 4.3). Yet, our method can reproduce and generate maps of research topics at any level of detail. The maps are systematically produced with the goal of improving upon the ad-hoc definition and interpretation of research thrusts in the institutional review process. “Reading these data-driven maps generates qualitative insights; they represent topics extracted from the text of research documents that require viewers to perceive patterns in them. They also possess emergent qualities, revealing more than the sum of their parts [11]; they show patterns in ERI’s research that were previously difficult or impossible to see when inspecting single documents, publication and project lists, or the work of individual researchers.

4.4.2 Deploying a Map Dashboard

To distribute and evaluate our maps, we deploy a public-facing dashboard¹¹ using Plotly, Dash for Python, and Heroku. The dashboard's About panel describes the map and allows users to select a level of detail, topics to map, and a year range. Figure 4.9 shows the Search panel, which allows users to filter data by ERI researcher or by academic department, returning metadata for a selected document, including its DOI when available. We make time explicit by showing a map snapshot for each year, which can be filtered by a range of years. This provides a backdrop for the interpretation of events, such as the acquisition of major grants or the hiring of new faculty in growing research areas. We provide evidence supporting these interpretations in our evaluation (Section 4.5).

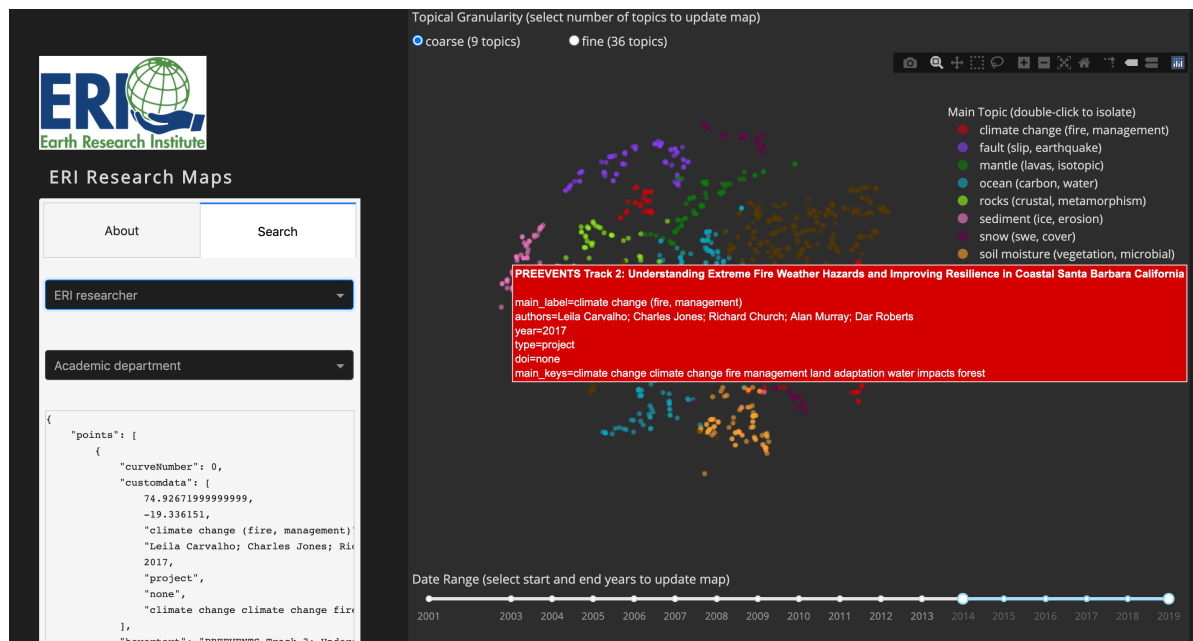


Figure 4.9: Search panel of the research map dashboard showing interactive features

¹¹<https://eri-research-dashboard.herokuapp.com/>

4.5 Evaluation

Do the maps we developed support “distant reading of research documents in the context of an institutional review? To answer this question more systematically, we evaluate the maps in three main ways. First, we use the maps to interpret and answer standard questions asked in the institutional review process. Second, we evaluate the maps in action, considering how they are used by the researchers whose work is being assessed [99]. We survey leading ERI researchers who determine if and how they think the maps support reading at a distance. Finally, we assess the relationship between the two maps by clustering the detailed topics and comparing the resulting clusters with the coarse topics.

4.5.1 Institutional Review Questions

How do maps of research topics support questions commonly posed to reviewers? Here, we consider the six institutional review questions about research accomplishment that UCSB’s organized research units must regularly address¹². They are currently answered using quantitative evidence, for example, numbers of publications by field of research and amounts of funding per researcher. While these benchmark questions are particular to UCSB, the concerns they address seem fairly representative of similar contexts elsewhere:

- **Research Quality and Significance:** Describe the quality and significance of research accomplished and in progress.
- **Trends and Research Specialties:** Comment on significant trends within the disciplines represented in the unit and relate these to current research specialties

¹²<https://www.research.ucsb.edu/organized-research-unit-oru-administration>

in your organized research unit (ORU).

- **Benefits to Campus and Departments:** Comment on how the ORU benefits the campus in general and academic departments in particular.
- **Participant Productivity, Influence, and Prominence:** Comment on the continuing productivity and influence of unit participants, locally as well as nationally. Comment on evidence of prominence in the fields represented in the ORU.
- **Collaborations and Interdisciplinarity:** Comment on the unit's collaborative/interdisciplinary work, its quality, and its impact on ORU research efforts and the campus.
- **Extramural Funding:** Describe the possible sources and availability of extramural funds to support the unit's-research. Are your participants sufficiently active in the pursuit of extramural funds in light of funding possibilities? How does the extent of annual extramural research funding compare with similar units nationwide?

We have claimed that maps of research topics can complement current evaluation metrics by supporting qualitative narratives. Here, we show how each of these questions can be addressed with maps of research topics:

Research Quality and Significance. Journal impact factors are a typical quantitative metric. Our maps complement this by generating a broader picture of cross-disciplinary topics from research publications. They highlight researchers' and departments' main topics and topical reach (diffuse or tightly clustered). Researchers in the Bren School of Environmental Science and Management are represented across all major topics, while those affiliated with Biology concentrate mainly in the species and oceans topic clusters.

Trends and Research Specialties. Funding agency priorities (e.g. NSF’s “10 big ideas”) and publisher classification schemes (e.g. fields of research) are typical sources of evidence. Our maps define research topics emerging from publications and projects that are not constrained by external classification schemes or historic disciplinary boundaries. The detailed map captures the topical diversity of research across affiliations while the coarse map emphasizes earth and environmental science topics unifying ERI’s researchers.

Benefits to Campus and Departments. Evidence includes faculty recruitment, research computing infrastructure, and educational outreach programs. Temporal sequencing in our maps can be used to assess the impact of events, like the inception of educational programs (e.g. the Kids in Nature Program in 2012) or influential funding (e.g. a 2017 NSF award to upgrade campus computing resources). While causality cannot be determined, it is interesting to note growth in certain topic areas following these events (e.g. a rise in ecological restoration projects following the start of educational programming and community outreach). These insights provide concrete and solid support over anecdotal discussions in institutional reviews.

Participant Productivity, Influence, and Prominence. The professional accolades of individual participants, such as awards, are often reported as evidence. Our maps provide a more objective picture of the topics that each researcher addresses by showing the topical distribution of each researcher’s documents. For example, geographer David Siegel’s work is concentrated mainly in the ocean and species topic clusters while geographer Dar Roberts’ work is more broadly dispersed across species, climate, ocean, snow, sediment, and soil moisture. While both accomplished researchers work extensively with remotely sensed imagery, differentiating their areas of expertise supports institutional management and reporting.

Collaborations and Interdisciplinarity. The affiliations of collaborators on funded projects are typically offered as evidence of interdisciplinarity. Our maps currently an-

notate each project by a single researcher and do not emphasize projects that have collaborators from multiple departments. This functionality could be added if ERI's leadership were interested to see who drives collaborations, not just what common topics they address.

Extramural Funding. This is currently based on award amounts. Our maps do not incorporate this kind of information, as these indicators already serve their purpose. The projects currently shown in the map have all been funded, but it could be valuable to also show the topics of unfunded projects, for example revealing changes to topics prioritized over time by funding agencies.

4.5.2 Researcher Survey

How do ERI's leading researchers interpret their own role in ERI's evolving research? We seek to understand researchers' interpretations of topics and relationships shown in the maps. To gather feedback, we administered an online survey to researchers on ERI's advisory board. This survey also served as a rehearsal and internal review for the imminent 5 year review in which the primary map users will be external reviewers in leadership positions at similar institutes. The survey was kept intentionally short and contained the following items:

- ERI Topics: Take a minute to explore the first map, both at the coarse and fine granularity. How well do you think these topics represent ERI's research overall?
- PI Topics: Next, find yourself in the Search panel. Your publications or projects will be highlighted. How well do you think this map represents your research?
- Topic Evolution: Finally, try filtering the research documents using the time slider. Do you notice any trends and do these coincide with any events in ERI's history that you can recall?

- Other: Do you have any other comments or ideas for improving this tool?

We received responses from 5/13 members of the ERI advisory board in the 4 days that were allocated to the survey. The main ideas that emerged from the responses can be separated into observations made from the maps and comments about map design. These responses provide suggestive evidence, which is summarized as follows:

ERI Topics. A majority (3/5) of respondents felt that the coarse map adequately described ERI’s research while the remainder had some objections. One noted that the coarse map lacks several important categories (e.g, biogeochemistry, inland waters, carbon cycle) but that the detailed map represents the range of research. Another felt that the topics reduced all of ERI’s research to physical entities that made it seem like a geology department. These concerns may relate to the design decision to label and color the documents by main topics; the labels include the first term from the topic descriptor with the second and third included in parentheses. Since topic modeling does not produce labels for the resulting topics, any succinct labeling in support of readability and verbalization skews the presentation. This feedback suggests that alternative approaches to labeling the topics could help because the objections raised were related to category names rather than the clustering of documents.

Researcher Topics. Respondents (3/5) felt that they understood the positions of their documents relative to ERI’s research landscape. Several mentioned that their assignments aligned with their identities as researchers; one noted I was largely in the species topic group and I do identify as a species-based researcher. Another felt that their work was categorized imperfectly at best as they work mainly on carbon cycling but had been associated with soils. These observations raise interesting challenges for visualizing perceived differences between researchers’ self-assigned specialties and positions assigned

to their work based on a relatively short period of time.

Topic Evolution. One researcher stated that trends in the map pointed to the founding of the UC Center for Environmental Implications of Nanotechnology at UCSB in 2013. Another noted that the map appears to start out along the edges then fills in the middle maybe it is selective hiring of people to bridge gaps? These interpretations speak to the utility of the spatialization approach; researchers are able to associate patterns in the map with probable events in which interdisciplinary research topics emerge, bridging traditional clusters. Changes in topical coverage following a faculty hire or large funding awards were observable to the respondents when they used the maps in combination with the time slider. Their observations demonstrate the kinds of insights that we envisioned the temporally sequenced maps might offer.

Other. Most of the comments about map functionality address click interactions, background color, alphabetization of lists, and other details that are readily changed. Suggestions for additional functionality included ways to browse lists of related documents based on shared topics, to visualize closely linked topics, and to search based on grants and papers. We expect to incorporate respondents' suggestions in preparation for the upcoming institutional review. We take the leading researchers' responses as a qualified endorsement of the generalization and visual presentation of work done at their institute.

4.5.3 Relationship Between Levels of Research Topics

Does the relationship between the detailed and coarse topics form a hierarchical tree or a lattice? This consideration is related to the granularity of the representation; many real-world phenomena, like administrative areas - cities in states within countries - can be represented by a tree but may function more like a lattice - city metropolitan regions

spanning state boundaries [90]. To study the relationship between topics in the detailed and coarse maps, we generate a dendrogram, which is a tree configuration to cluster the topics. It is different from the trees in distant reading [11], which capture the evolution of topics over time. Rather, the tree shown in Figure 4.10 coarsens the detailed topic model by clustering the 36 detailed topics into 9 clusters based on their similarity.

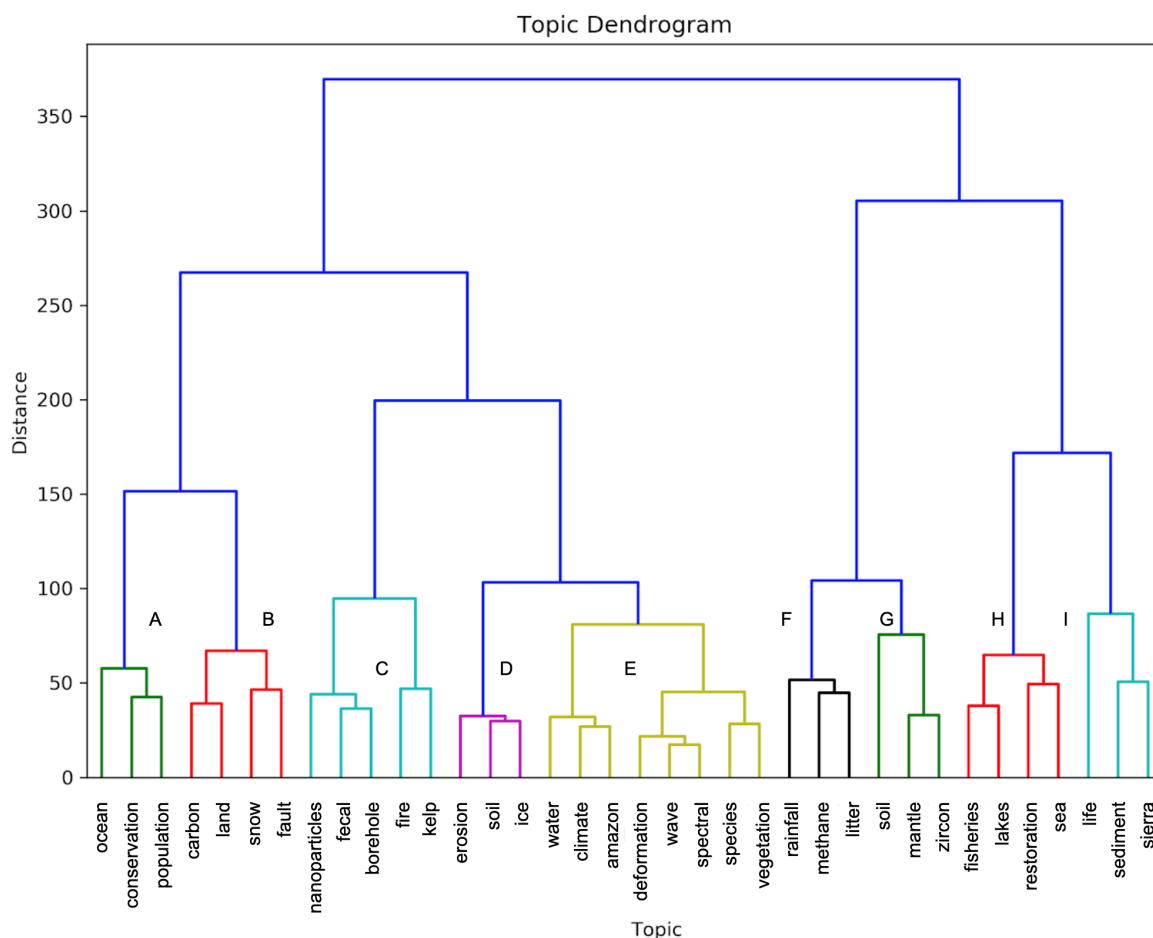


Figure 4.10: Dendrogram of 36 clustered topics labeled by the first term in their descriptors

We use SciPy¹³ to cluster the topics and render them in a dendrogram. We experimented with different linkage methods for clustering the topics and selected Ward's minimum variance [100]. It is agglomerative, beginning with each topic in a separate clus-

¹³<https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html>

ter; at each step, it merges the two most similar clusters, resulting in compact, even-sized clusters [101]. The vertical axis of the dendrogram represents the distance, or dissimilarity, between topics and the horizontal axis shows the labeled topics. Each of the nine clusters are assigned a color; for example, cluster A in the dendrogram is colored green and consists of three topics: conservation' and population, which are the most similar, and ocean'. The smallest clusters (A, D, F, G, and I) consist of three topics each and seem to be the most coherent, as to be expected. The largest cluster (E) consists of 8 topics and appears to be the most general, combining water', climate', vegetation', and species'.

To study the relationship between the topic models, we compare the resulting 9 topic clusters to the 9 coarse topics (Table 4.2). The clusters column records if the topics in a cluster share at least one term in common with a coarse topic. Terms associated with the clustered topics are based on the union of each member topic's ten most representative terms.

Topics 2 and 8 share terms with the topics in clusters 5 and 7. The other coarse topics seem to only partially correspond with the clusters. Topic 9 in the coarse model has the terms sediment, ice, and erosion. In the detailed model, these terms are split into distinct topics: sediment is a term in Topic 28, ice is a term in Topic 10, and erosion is a term in Topic 15. Each detailed topic is associated with a limited number of coarse topics, suggesting that the detailed topics are not hierarchical refinements of coarse ones, but that moving between levels (i.e. more or less distant) actually reconfigures the terms into a lattice of related topics.

Topic	Keywords	Clusters
1	ocean carbon water co2 phytoplankton global surface organic color emissions	A, B, D, E, F, G
2	ma rocks crustal metamorphism monazite crust zircon metamorphic deformation exhumation	E, G
3	snow swe snow cover cover water snow water modis snowmelt model water equivalent	A, B, E
4	soil soil moisture moisture vegetation microbial microwave surface soils band plant	B, D, E, G
5	climate change climate change fire management land adaptation impacts water forest	A, B, C, D, E, H
6	species diversity plant richness native biomass communities biodiversity effects ecosystem	A, C, E, G
7	fault slip earthquake rupture seismic motion ground faults earthquakes ground motion	B, C, E, I
8	mantle lavas isotopic crust 3he melt samoan geochemical 4he 3he 4he	E, G
9	sediment ice erosion rates 10be river sea glacial erosion rates precipitation	D, E, F, H, I

Table 4.2: Topics, descriptors with ten top terms, and corresponding topic clusters

4.6 Conclusions

The evaluation of the research maps demonstrates their potential to complement current approaches to institutional reviews. Our data-driven approach uses topic modeling and spatialization to derive high level topics from ERI’s body of research and make them discoverable by researchers and departments.

In the next phase of this research, we envision a dynamic system for generating maps of research topics based on a continually updating corpus of research documents, which would require supervised classification of unseen documents based on our current data. We are also interested in highlighting relationships among topics, departments, researchers, or other units by presenting the research documents as a network. While we successfully developed field and network views of research documents in our previous work [48], we did not exploit the potential of networks in this research for several rea-

sons. For one, networks based on quantitative information (e.g. co-citation frequency) are already established scientometric techniques [46] and we are interested in developing complementary approaches to systematically model research topics. As ERI is a multi-disciplinary institution, we also did not want to draw imbalanced comparisons between the contributions of individual researchers from different disciplines; to avoid this, we would prefer to produce network views of other significant units like academic departments. Finally, our approach embeds research documents as nodes; however, linking embedded objects would not allow us to see emerging patterns like documents that are highly connected and central.

A limitation of our approach is that while spatial information has three components (space, time, theme) we primarily take advantage of the thematic dimensions of data. While we incorporate temporal views in our maps, allowing for document subsetting by timespan and event detection, we could also consider making time the primary dimension. Previous work developing semantic signatures shows that time and space offer two ways to order knowledge [84]. Views ordered primarily by time could be thought of as temporalizations, rather than the spatializations we develop, and could take the form of graphs from distant reading, where the evolution of specific topics could be tracked.

We are also interested in extending the evaluation of our work to coincide with ERI's external review. This would help us understand how external reviewers who do not have a personal connection to ERI's research interpret and evaluate its research topics. To determine the applicability and maturity of our approach for adoption in a broader context, we would also be interested in expanding our survey to include researchers or leaders affiliated with similar organized research units. This would allow us to build consensus around strategies for adopting maps of research as robust spatial decision support tools.

At the outset of this article, we argued for maps of the research territory to provide

decision support. The maps we have produced give an impression of the underlying thematic structure of the research in the form of research neighborhoods that are meaningful within, and possibly across, institutions. Just as land use maps are used to manage resources and forecast growth in a regional planning context, maps of research can be used to do the same in an institutional setting. We envision maps of research topics being used internally (as part of the organized research unit's self-assessment) and externally (as a communication tool describing research trends and developments) which may be of interest to external reviewers, other research units, and the public.

Chapter 5

Conclusions and Future Work

5.1 Summary and Discussion

This dissertation develops approaches for curating research data in topic spaces. This is of increasing relevance for several reasons. There is a growing demand on institutions to promote data curation strategies that support discovery and reuse. The declaration of FAIR principles to guide scientific data stewardship [7], and recent advent of TRUST principles for digital repositories [102], speak to a growing cross-disciplinary scientific consensus around data curation questions. These principles are intended to guide curation practices by promoting access to research data and trust in data repositories.

In light of these institutional shifts, the methods developed in this dissertation offer strategies for increasing access to research data and improving their discoverability, particularly in multidisciplinary settings. Chapter 2 develops a crosscutting vocabulary of core geospatial topics to facilitate term expansion, thus hoping to improve the discoverability of open civic data. This semantic approach addresses the “vocabulary problem” [6] by cutting across institutional barriers and levels of government. Rather than developing a vocabulary and taking a term expansion approach however, Chapter 3 demonstrates

spatial methods to establish the similarity of research documents and present them according to a distance-similarity metaphor [10]. The spatial method developed in Chapter 3 cuts across institutional barriers posed by the language of academic disciplines. Chapter 4 demonstrates the relevance of spatial methods for curating research topics in a multidisciplinary, institutional setting by applying topic modeling and spatialization to support an institutional review. Multiple levels of detail in the spatializations support “distant reading” [11] of research documents where relevant patterns emerge in coarse or detailed maps. Abstracting and arranging the topics of research data is a first step toward supporting their serendipitous discovery and reuse.

This dissertation is also relevant for understanding the information behavior of particular user communities [103] with the goal of overcoming barriers to data discovery and reuse. This can inform institutional perspectives on data curation to support users’ search strategies and needs. In Chapter 2, the strategies developed to support data discovery map users’ vernacular to metadata with the goal of enabling the public to find authoritative, government data. This strategy is differentiated in Chapter 3 to support the discovery of related theses and dissertations by users like graduate students; in this case, the users and providers are similar. Finally, the strategies developed in Chapter 4 are evaluated by a formally organized group of earth scientists on the advisory board of a research institute; however, these mapping approaches have the potential to communicate high-level views of research trends to more general academic and public audiences.

Finally, this dissertation takes a system design perspective to address open questions about data discovery. The design of each study builds upon the decisions made in previous studies. Chapter 2 takes a hierarchical approach to expand technical terms but hides from view the similarity relations shared by data. Following this study, this dissertation takes a deliberate spatial approach to support the interpretation of relationships like similarity among information items [40]. In Chapter 3, similar theses and dissertations

are conveyed visually through spatializations rather than through hierarchies with the intentional design choice of co-locating similar research data [8]. In Chapter 4 similar design choices are made for the system that is developed with the added interaction of specifying the level of detail in the topic map.

The design choices are informed by core concepts of spatial information [33] which offer lenses for viewing and arranging research data. In Chapter 2, the research data are conceptualized as *objects* linked by relations established in a controlled vocabulary including broader, narrower, or related concepts. These relations form a hierarchical *network* of concepts, which are used to annotate the research data. In Chapter 3, two systematic views of the underlying research data, as a *field* and as a *network*, are produced to support distinct questions. In Chapter 4, the research data are *objects* embedded in a *field* of topics at two distinct levels of detail, determined by the *granularity* of the topic models and spatializations. The views of documents, which can be updated by time interval, also support the detection of *events* involving the research documents as participating *objects*. In each case, *location* is the base concept, guiding the interpretation of research data similarity. In Chapter 4, the location of research topics produces meaningful research *neighborhoods*, a concept that is strongly associated with the *field* notion given Tobler’s first law of geography [23] and the distance-similarity metaphor [10], which hold relevance for curation by co-locating related topics. These studies demonstrate the relevance of core concepts of spatial information beyond a geographic context, showing their utility for organizing research data in abstract information spaces. While quality concepts of data *accuracy*, *provenance*, and *completeness* are not addressed in this dissertation, they suggest additional design facets to consider in future studies.

Throughout this dissertation, evaluation has shifted from a focus on system design to incorporate user interpretations of the system; this is in keeping with related design work in this area. For instance, the initial focus of the spatialization research of Bruggmann

and Fabrikant [67] is on system development; they subsequently focus on evaluating the efficacy of their spatialized displays through user studies [104] with digital humanities scholars involving focus group research, a cognitive walkthrough, and a think aloud study. The arc of evaluation in this dissertation follows a similar pattern, focusing first on system development in Chapter 2, moving into prototyping in Chapter 3, and concluding with expert evaluation of system designs using surveys in Chapter 4.

This dissertation strives to address data usability challenges with methods from spatial information science. Chapter 2 addresses search problems by semantically annotating data. This approach focuses on “verbalization”, enhancing the terminology of users by expanding data tags. For example, Section 2.4 evaluates the system search performance by monitoring conversion rates, which divide the number of data downloads by the total number of site visitors. This measure indicates search success, implying that users are able to search using their terminology and find suitable data to download. The project partner, Esri, has since conducted more testing and improved evaluation mechanisms. An industry standard for measuring the success of features, such as semantically aided search, requires large amounts of A/B testing to compare versions of data portals against each other. Interspersing the A/B tests with experience sampling (for example, interrupting users with small surveys about their experience) can offer additional insights into user experiences. These usability testing strategies in portals with and without semantic annotation have the potential for high ecological validity; findings may be generalized to broader contexts. Given the involvement of A/B testing however, a more common measure of success is feature adoption. If users find a feature to be compelling, it is likely to be retained and improved. Semantically aided search has been adopted and is associated with increased conversion rates. Esri continues to develop its capabilities to expand user search terms by adopting a wider linked data framework¹ for semantic annotation.

¹<http://conceptnet.io/>

Chapter 3 moves from abstracting concepts through “verbalization” to “spatialization”. Since the tools developed in Section 3.5 are prototypes, their potential is demonstrated through answers to prototypical research questions. Instead of evaluating the spatialization approach through measures like conversion rates, Chapter 3 demonstrates the application of the spatializations to support research data discovery. This evaluation strategy develops a case study where two dissertations from geography researchers are located and interpreted in the spatialized displays. Questions enabled by a *field* of research topics, which highlights their distribution, are distinct from questions enabled by a *network* of research documents, which highlight their connectivity. This chapter demonstrates the viability of combining topic modeling and spatialization. Developing additional exploratory tasks for users to perform in the spatializations would support further usability evaluation; for example, asking users to locate items, identify relationships between attributes, or distinguish between spatial locations would offer evidence for how the spatializations align with users’ mental models [105]. Despite not undertaking this kind of user assessment, it was still possible to anticipate limitations and make adjustments to the spatialization methods in the following study; for example, there are some known limitations of self-organizing maps that made them unsuitable for conveying a distance-similarity metaphor [43]. Self-organizing maps do not have a consistent proximity metric; this can be addressed by warping the self-organizing map using a cartogram method [72]. They also appear to cluster items on their edges although they are rotationally invariant; this can be addressed by projecting the map onto a surface, like a torus, but such adjustments may be challenging for users to interpret. These issues limit the utility of self-organizing maps as communication devices.

For these reasons, the spatialization method is refined and evaluated in Chapter 4. In Section 4.5 the maps of research are evaluated in three main ways: they are applied to support typical questions asked in an institutional review; they are evaluated by leading

researchers affiliated with the research institute; and the relationship between the coarse and detailed maps are shown to take the form of a lattice rather than a strict hierarchy, confirming the idea that hierarchical representations of research topics are too strict. Instead, this approach produces adequate topic models at each level without imposing this rigid constraint. Furthermore, the surveys test the impact of the spatializations by assessing stakeholders. They show that the researchers who are represented in the space recognize themselves in it and can explain their relative locations, suggesting that the maps depict a reality that the stakeholder believes to be true [87].

5.2 Contributions

This dissertation advances our understanding of spatial curation’s potential to support research data discovery. It presents a system for improving civic data discovery through term expansion and two applications that combine topic modeling and spatialization to support data discovery in the multidisciplinary contexts of a university research library and an organized research institute. The research articles in this dissertation have focused on overcoming data discovery challenges by contributing methods and practical applications that advance geographic information science as follows:

Semantic annotation protocol for open geospatial data. Chapter 2 has contributed to the development of a system for harvesting provider-contributed data descriptions. It does so by developing a base vocabulary of core geospatial topics that map provider to consumer descriptions. This vocabulary is used to semantically annotate data with core geospatial topics to improve civic data discovery. This results in improved access to geospatial data across data providers and levels of government, which use distinct terminologies to describe their data products. Rather than creating new terms, this approach leverages existing authoritative vocabularies, including the Library of Congress

Subject Headings and the USGS Thesaurus, to achieve term expansion.

Systematic design choices for spatializations. Spatialization theory has established best practices by integrating cartographic principles to produce “cognitively plausible information visualizations” [43]. Chapter 3 in this dissertation further contributes to spatialization theory a systematic choice of spaces informed by core concepts of spatial information [33]. The application of this spatialization approach to research data curation offers distinct and intuitive ways to gain insights into the relatedness of research data, overcoming the disciplinary terms used to describe them.

Spatial decision support for institutional reviews. Chapter 4 in this dissertation contributes a method for producing maps of research topics at multiple levels of granularity. The maps support qualitative narratives, complementing existing quantitative approaches to institutional reviews. More broadly, this approach translates the notion of spatial decision support [83] from a *geographic* context (using maps of the built environment to inform planning decisions), to a *spatial* context (using topic maps of the research environment to inform institutional planning objectives).

5.3 Limitations

Establishing an evaluation baseline. Design science is concerned with “how things ought to be, with devising artifacts to attain goals” rather than “how things are”, which is a concern in the natural sciences [106]. Addressing usability problems in spatial information science through design means innovating previously unseen solutions. The methods developed in this dissertation are meant to allow users to do things that they could not do before. While Chapter 2 offers a baseline for expanded A/B testing, it was more challenging to establish a baseline in the studies presented in Chapters 3 and 4. Assessing user experiences in a controlled version of a search site before and after adding

semantic annotations to metadata for example would facilitate comparison.

Expanding testing and feedback mechanisms. Testing and collection of feedback in each study was fairly limited. Ideas for eliciting additional feedback would be applicable across the studies; for example, assigning users similar tasks across systems (a data portal, a data repository, and a data dashboard) and paradigms (information lookup and exploratory search) would reveal differences in perceptions of the interfaces. Developing a finer understanding of specific cognitive mechanisms or strategies contributing to user search success could be a productive direction for improving this research.

Use of proxies for research data. At the outset of this dissertation, the goal was to experiment directly with research data. However, institutional data curation policies and repository solutions have only recently started coming into effect. For example, UC Santa Barbara very recently joined Dryad², with the effect of making large collections of researcher-deposited data available for experimentation. At the time of writing however, no such agreements existed; in the absence of institutional agreements to support data curation, experiments reported in this dissertation were either conducted externally (in partnership with Esri) or reference other kinds of research products (such as theses, research articles, or grant proposals) as proxies for original research data. Despite this, the methods developed in this dissertation are relevant for metadata in general, suggesting their applicability to secondary research data.

5.4 Open Questions

How can core concepts of spatial information further support the spatial curation of knowledge? This dissertation has demonstrated the broad applicability of core concepts of spatial information [33] to organize and interpret information spaces

²<https://datadryad.org/search>

beyond the geographic context. These concepts have informed the spatialization of research topics through the content concepts of *field*, *object*, and *event*. Experiments with the content concept of *networks* in Chapter 3 did not prove to be very effective, mainly because the nodes (research documents) were embedded. It remains to be seen how alternative conceptualizations of research networks might support additional questions about the connectivity of research topics, especially in the case of institutional reviews. Furthermore, the quality concept of *granularity* makes the level of detail in topic models and topic spaces tangible by supporting “distant reading” of a body of research [11] in the institutional review process. In the long-term, core concepts of spatial information could facilitate the spatial curation of scientific knowledge with the potential to serve not only academic research units and the organizations they report to, but also broader audiences, such as libraries and the general public.

What kinds of curatorial actions have the most impact in terms of data discovery and reuse? While this dissertation has contributed methods and tools to improve data discovery, there is much more to be studied specifically from an archival standpoint³. For example, does reconciling variable names against a controlled vocabulary increase the likelihood that users will find, understand, and reuse the data? Data sharing policies have diverse motivations including research transparency, replication, and data reuse. These concerns also have institutional and policy implications, as the maintenance and provision of data is costly. Developing an understanding of the impacts of these kinds of specific curatorial actions has great potential to improve the discovery and reuse of research data across communities of practice, particularly with regard to the secondary impact of data reuse and the disciplinary diversity of data citations [4].

How can recommendation and question-answering systems support data discovery and reuse? Experimentation with emerging technologies, such as recom-

³https://www.nsf.gov/awardsearch/showAward?AWD_ID=1930645

mendation systems for research data, is a promising avenue. There is current interest in studying and developing infrastructure to support research data reuse⁴. This dissertation has focused on developing similarity measures for research data based on topic modeling that are presented visually in maps. In practice however, fitness for use is also influenced by other factors including perceived relevance to a research question and priorities within the application area of interest [18]. These considerations also have implications for research on semantic technologies for geographic question answering⁵. Foundational research is still needed on the form and concepts of spatial questions and answers, for example to support geo-analytical workflow composition. Incorporating quality dimensions such as *accuracy*, *provenance*, and *completeness* into data recommendation systems suggests a promising line of related research continuing from this dissertation.

⁴https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505730

⁵<http://spatial.ucsb.edu/giscience-geoQA-2021>

Bibliography

- [1] M. K. Buckland, *What is a document?*, *Journal of the American Society for Information Science* **48** (1997), no. 9 804–809.
- [2] S. Bechhofer, D. De Roure, M. Gamble, C. Goble, and I. Buchan, *Research objects: Towards exchange and reuse of digital knowledge*, *Nature Precedings* (2010).
- [3] M. S. Mayernik, *Research data and metadata curation as institutional issues*, *Journal of the Association for Information Science and Technology* **67** (2016), no. 4 973–993.
- [4] K. M. Fear, *Measuring and Anticipating the Impact of Data Reuse*. University of Michigan, 2013.
- [5] P. Morville and J. Callender, *Search Patterns: Design for Discovery*. O’Reilly Media, Inc., 2010.
- [6] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, *The vocabulary problem in human-system communication*, *Communications of the ACM* **30** (1987), no. 11 964–971.
- [7] M. D. Wilkinson, S.-A. Sansone, E. Schultes, P. Doorn, L. O. B. da Silva Santos, and M. Dumontier, *A design framework and exemplar metrics for fairness*, *Scientific Data* **5** (2018).
- [8] P. Gärdenfors, *Conceptual Spaces: The Geometry of Thought*, pp. 151–201. MIT Press, 2004.
- [9] T. Kohonen, *Self-Organizing Maps*. Springer, 1995.
- [10] S. I. Fabrikant, M. Ruocco, R. Middleton, D. R. Montello, and C. Jørgensen, *The first law of cognitive geography: distance and similarity in semantic space*, *Proceedings of GIScience 2002* (2002) 31–33.
- [11] F. Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso, 2005.

- [12] D. Lee and H. S. Seung, *Learning the parts of objects by non-negative matrix factorization*, *Nature* **401** (1999) 788–791.
- [13] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent dirichlet allocation*, *Journal of Machine Learning Research* **3** (2003), no. Jan 993–1022.
- [14] L. v. d. Maaten and G. Hinton, *Visualizing data using t-SNE*, *Journal of Machine Learning Research* **9** (2008), no. Nov 2579–2605.
- [15] L. L. Hill, L. Carver, M. Larsgaard, R. Dolin, T. R. Smith, J. Frew, and M.-A. Rae, *Alexandria digital library: user evaluation studies and system design*, *Journal of the American Society for Information Science* **51** (2000), no. 3 246–259.
- [16] D. Hardy and K. Durante, *A metadata schema for geospatial resource discovery use cases*, *Code4Lib Journal* (2014), no. 25.
- [17] M. F. Goodchild, *A geographer looks at spatial information theory*, in *International Conference on Spatial Information Theory*, pp. 1–13, Springer, 2001.
- [18] W. Bishop and T. H. Grubestic, *Geographic Information*. Springer, 2016.
- [19] S. Lafia, J. Jablonski, W. Kuhn, S. Cooley, and F. A. Medrano, *Spatial discovery and the research library*, *Transactions in GIS* **20** (2016), no. 3 399–412.
- [20] S. Lafia and W. Kuhn, *Spatial discovery of linked research datasets and documents at a spatially enabled research library*, *Journal of Map & Geography Libraries* **14** (2018), no. 1 21–39.
- [21] A. Ballatore, W. Kuhn, M. Hegarty, and E. Parsons, *Spatial approaches to information search*, *Spatial Cognition & Computation* **16** (2016), no. 4 245–254.
- [22] M. Hearst, *User interfaces and visualization*, *Modern Information Retrieval* (1999) 257–323.
- [23] W. R. Tobler, *A computer movie simulating urban growth in the Detroit region*, *Economic Geography* **46** (1970), no. sup1 234–240.
- [24] B. Lawrence, C. Jones, B. Matthews, S. Pepler, and S. Callaghan, *Citation and peer review of data: Moving towards formal data publication*, *International Journal of Digital Curation* **6** (2011), no. 2 4–37.
- [25] B. Smith and M. Jensen, *The UNEP ontologies and the OBO foundry.*, in *ICBO/BioCreative*, 2016.
- [26] R. Kitchin, *The real-time city? Big data and smart urbanism*, *GeoJournal* **79** (2014), no. 1 1–14.

- [27] G. Mai, K. Janowicz, Y. Hu, and G. McKenzie, *A linked data driven visual interface for the multi-perspective exploration of data across repositories.*, in *VOILA@ ISWC*, pp. 93–101, 2016.
- [28] B. Hjørland, *Subject (of documents)*, *Knowledge Organization* **44** (2017), no. 1 55–64.
- [29] D. Sinton, *The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study*, *Harvard Papers on Geographic Information Systems* (1978).
- [30] C. Wolff, A. B. Rod, and R. C. Schonfeld, *Ithaka S+ R US Faculty Survey 2015*, .
- [31] E. Svenonius, *The Intellectual Foundation of Information Organization*. MIT Press, 2000.
- [32] A. Paepcke, *Digital libraries: Searching is not enough; what we learned on-site*, *Corporation for National Research Initiatives* (1996).
- [33] W. Kuhn, *Core concepts of spatial information for transdisciplinary research*, *International Journal of Geographical Information Science* **26** (2012), no. 12 2267–2276.
- [34] S. I. Fabrikant, *Spatialization*, *International Encyclopedia of Geography: People, the Earth, Environment and Technology: People, the Earth, Environment and Technology* (2016) 1–7.
- [35] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, *Visualizing the non-visual: Spatial analysis and interaction with information from text documents*, in *Proceedings of Visualization 1995 Conference*, pp. 51–58, IEEE, 1995.
- [36] M. Card, *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999.
- [37] W. Kuhn, *Handling data spatially: Spatializing user interfaces*, in *Proceedings of the 7th International Symposium on Spatial Data Handling, Delft, The Netherlands*, pp. 12–16, 1996.
- [38] M. F. Goodchild and D. G. Janelle, *Specialization in the structure and organization of geography*, *Annals of the Association of American Geographers* **78** (1988), no. 1 1–28.
- [39] K. Börner, R. Klavans, M. Patek, A. M. Zoss, J. R. Biberstine, R. P. Light, V. Larivière, and K. W. Boyack, *Design and update of a classification system: The UCSD map of science*, *PLOS One* **7** (2012), no. 7 e39464.

- [40] D. R. Montello, S. I. Fabrikant, M. Ruocco, and R. S. Middleton, *Testing the first law of cognitive geography on point-display spatializations*, in *International Conference on Spatial Information Theory*, pp. 316–331, Springer, 2003.
- [41] S. I. Fabrikant, D. R. Montello, M. Ruocco, and R. S. Middleton, *The distance-similarity metaphor in network-display spatializations*, *Cartography and Geographic Information Science* **31** (2004), no. 4 237–252.
- [42] S. I. Fabrikant, D. R. Montello, and D. M. Mark, *The distance-similarity metaphor in region-display spatializations*, *IEEE Computer Graphics and Applications* **26** (2006), no. 4 34–44.
- [43] A. Skupin and S. I. Fabrikant, *Spatialization methods: a cartographic research agenda for non-geographic information visualization*, *Cartography and Geographic Information Science* **30** (2003), no. 2 99–119.
- [44] A. M. MacEachren, *How Maps Work: Representation, Visualization, and Design*. Guilford Press, 2004.
- [45] S. I. Fabrikant, *Spatialized browsing in large data archives*, *Transactions in GIS* **4** (2000), no. 1 65–78.
- [46] K. Börner, C. Chen, and K. W. Boyack, *Visualizing knowledge domains*, *Annual Review of Information Science and Technology* **37** (2003), no. 1 179–255.
- [47] S. Lafia, A. Turner, and W. Kuhn, *Improving discovery of open civic data*, in *10th International Conference on Geographic Information Science (GIScience 2018)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.
- [48] S. Lafia, C. Last, and W. Kuhn, *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, 2019.
- [49] A. Turner, “Desire paths to open data.” <http://highearthorbit.com/articles/desire-paths-to-open-data>, 2014.
- [50] A. Zuiderwijk and M. Janssen, *Open data policies, their implementation and impact: A framework for comparison*, *Government Information Quarterly* **31** (2014), no. 1 17–29.
- [51] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, *et. al.*, *Why linked data is not enough for scientists*, *Future Generation Computer Systems* **29** (2013), no. 2 599–611.

- [52] O. R. D. Taskforce, *Research data infrastructures in the UK : Landscape report*, tech. rep., Universities UK, 2017.
- [53] T. Berners-Lee, J. Hendler, and O. Lassila, *The semantic web*, *Scientific American* **284** (2001), no. 5 34–43.
- [54] Y. Bishr, *Overcoming the semantic and other barriers to gis interoperability*, *International Journal of Geographical Information Science* **12** (1998), no. 4 299–314.
- [55] S. Bischof, A. Karapantelakis, C.-S. Nechifor, A. P. Sheth, A. Mileo, and P. Barnaghi, *Semantic modelling of smart city data*, in *The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis)*, 2014.
- [56] C. Debruyne, É. Clinton, L. McNerney, A. Nautiyal, and D. O’Sullivan, *Serving Ireland’s geospatial information as linked data.*, in *International Semantic Web Conference (Posters & Demos)*, 2016.
- [57] G. A. Miller, *WordNet: a lexical database for English*, *Communications of the ACM* **38** (1995), no. 11 39–41.
- [58] N. Baker, *Discards*, *The New Yorker* (Apr, 1994) 6486.
- [59] A. Skupin, *The world of geography: Visualizing a knowledge domain with cartographic means*, *Proceedings of the National Academy of Sciences* **101** (2004), no. suppl 1 5274–5278.
- [60] D. Soergel, *The rise of ontologies or the reinvention of classification*, *Journal of the American Society for Information Science* **50** (1999), no. 12 1119–1120.
- [61] C. J.-H. Bae, *Representations of an Urban Neighborhood: Residents’ Cognitive Boundaries of Koreatown, Los Angeles*. University of California, Santa Barbara, 2015.
- [62] G. D. McKenzie, *A Temporal Approach to Defining Place Types Based on User-Contributed Geosocial Content*. University of California, Santa Barbara, 2015.
- [63] Y. Hu, K. Janowicz, G. McKenzie, K. Sengupta, and P. Hitzler, *A linked-data-driven and semantically-enabled journal portal for scientometrics*, in *International Semantic Web Conference*, pp. 114–129, Springer, 2013.
- [64] C. Alexander, *A Pattern Language: Towns, Buildings, Construction*. Oxford University Press, 1977.
- [65] J. A. Wise, *The ecological approach to text visualization*, *Journal of the American Society for Information Science* **50** (1999), no. 13 1224–1233.

- [66] R. G. Golledge, *Primitives of spatial knowledge*, in *Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems*, pp. 29–44. Springer, 1995.
- [67] A. Bruggmann and S. I. Fabrikant, *How to visualize the geography of Swiss history*, in *Proceedings of the AGILE'2014 International Conference on Geographic Information Science*, AGILE Digital Editions, 2014.
- [68] J. Nunes, *Geographic space as a set of concrete geographical entities*, in *Cognitive and Linguistic Aspects of Geographic Space*, pp. 9–33. Springer, 1991.
- [69] M. Newman, *Networks*. Oxford University Press, 2010.
- [70] T. R. Smith and J. Frew, *Alexandria digital library*, *Communications of the ACM* **38** (1995), no. 4 61–62.
- [71] E. A. Leicht, G. Clarkson, K. Shedden, and M. Newman, *Large-scale structure of time evolving citation networks*, *The European Physical Journal B* **59** (2007), no. 1 75–83.
- [72] A. Bruggmann, M. M. Salvini, and S. Fabrikant, *Cartograms of self-organizing maps to explore user-generated content*, in *26th International Cartographic Conference*, pp. 25–30, 2013.
- [73] R. T. Snodgrass, *Temporal databases*, in *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, pp. 22–64. Springer, 1992.
- [74] A. Rabinowitz, R. Shaw, S. Buchanan, P. Golden, and E. Kansa, *Making sense of the ways we make sense of the past: The PeriodO project*, *Bulletin of the Institute of Classical Studies* **59** (2016), no. 2 42–55.
- [75] J. P. Ioannidis, *Meta-Research: Why Research on Research Matters*, *PLoS Biology* **16** (2018), no. 3 e2005468.
- [76] D. Bawden and L. Robinson, *Informetrics*, pp. 165–181. Facet Publishing, 2015.
- [77] A. Jappe, D. Pithan, and T. Heinze, *Does bibliometric research confer legitimacy to research assessment practice? A sociological study of reputational control, 1972-2016*, *PloS one* **13** (2018), no. 6 e0199031.
- [78] P. Vinkler, *Indicators are the essence of scientometrics and bibliometrics*, *Scientometrics* **85** (2010), no. 3 861–866.
- [79] F. Radicchi, S. Fortunato, and C. Castellano, *Universality of citation distributions: toward an objective measure of scientific impact*, *Proceedings of the National Academy of Sciences* **105** (2008), no. 45 17268–17272.

- [80] D. Moher, F. Naudet, I. A. Cristea, F. Miedema, J. P. Ioannidis, and S. N. Goodman, *New principles for assessing scientists*, *Issues in Science and Technology* **35** (2018), no. 1 20–23.
- [81] P. McGuire, *How higher education’s data obsession leads us astray*, *The Chronicle of Higher Education* (Oct, 2019).
- [82] A. Polonioli, *In search of better science: On the epistemic costs of systematic reviews and the need for a pluralistic stance to literature search*, *Scientometrics* **122** (2020), no. 2 1267–1274.
- [83] G. Andrienko, N. Andrienko, P. Jankowski, D. Keim, M.-J. Kraak, A. MacEachren, and S. Wrobel, *Geovisual analytics for spatial decision support: Setting the research agenda*, *International Journal of Geographical Information Science* **21** (2007), no. 8 839–857.
- [84] K. Janowicz, *The role of space and time for knowledge organization on the semantic web*, *Semantic Web* **1** (2010), no. 1, 2 25–32.
- [85] S. Arias and B. Warf, *Introduction: The Reinsertion of Space into the Social Sciences and Humanities*, pp. 1–11. Routledge, London, 2009.
- [86] J. Boyd-Graber, D. Mimno, and D. Newman, *Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements*, pp. 225–255. CRC Press Boca Raton, FL, 2014.
- [87] M. Hermann and H. Leuthold, *Atlas der politischen Landschaften: Ein weltanschauliches Porträt der Schweiz*. Vdf Hochschulverlag an der ETH, 2003.
- [88] C. Allen, T. Hervey, S. Lafia, D. W. Phillips, B. Vahedi, and W. Kuhn, *Exploring the notion of spatial lenses*, in *The Annual International Conference on Geographic Information Science*, pp. 259–274, Springer, 2016.
- [89] D. Greene, D. O’Callaghan, and P. Cunningham, *How many topics? Stability analysis for topic models*, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 498–513, Springer, 2014.
- [90] S. C. Hirtle, *Representational structures for cognitive space: Trees, ordered trees and semi-lattices*, in *International Conference on Spatial Information Theory*, pp. 327–340, Springer, 1995.
- [91] T. Underwood, *Distant Reading and Recent Intellectual History*, *Debates in the Digital Humanities 2016* (2016) 530–533.
- [92] G. Salton, A. Wong, and C. Yang, *A vector space model for automatic indexing*, *Communications of the ACM* **18** (1975), no. 6 391–398.

- [93] T. L. Griffiths, M. I. Jordan, J. B. Tenenbaum, and D. M. Blei, *Hierarchical topic models and the nested Chinese restaurant process*, in *Advances in Neural Information Processing Systems*, pp. 17–24, 2004.
- [94] G. A. Miller, *The magical number seven, plus or minus two: Some limits on our capacity for processing information*, *Psychological Review* **63** (1956), no. 2 81.
- [95] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, *Optimizing semantic coherence in topic models*, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 262–272, Association for Computational Linguistics, 2011.
- [96] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, *A practical algorithm for topic modeling with provable guarantees*, in *International Conference on Machine Learning*, pp. 280–288, 2013.
- [97] L. McInnes, J. Healy, and J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, *arXiv preprint arXiv:1802.03426* (2018).
- [98] M. Wattenberg, F. Viégas, and I. Johnson, *How to use t-SNE effectively*, *Distill* **1** (2016), no. 10 e2.
- [99] P. Solomon, *Discovering information in context*, *Annual Review of Information Science and Technology* **36** (2002), no. 1 229–264.
- [100] J. H. Ward Jr, *Hierarchical grouping to optimize an objective function*, *Journal of the American Statistical Association* **58** (1963), no. 301 236–244.
- [101] M. S. Aldenderfer and R. K. Blashfield, *Cluster Analysis*, pp. 62–73. Sage Publications London, 1984.
- [102] D. Lin, J. Crabtree, I. Dillo, R. R. Downs, R. Edmunds, D. Giaretta, M. De Giusti, H. L’Hours, W. Hugo, R. Jenkyns, *et. al.*, *The TRUST principles for digital repositories*, *Scientific Data* **7** (2020), no. 1 1–5.
- [103] K. E. Fisher, S. Erdelez, and L. E. McKechnie, *Theories of Information Behavior*. Information Today, Inc., 2005.
- [104] A. Bruggmann and S. I. Fabrikant, *How does GIScience support spatio-temporal information search in the humanities?*, *Spatial Cognition & Computation* **16** (2016), no. 4 255–271.
- [105] E. L. Koua, A. MacEachren, and M.-J. Kraak, *Evaluating the usability of visualization methods in an exploratory geovisualization environment*, *International Journal of Geographical Information Science* **20** (2006), no. 4 425–448.

- [106] H. A. Simon, *The science of design: Creating the artificial*, *Design Issues* (1988) 67–82.