Search Facets and Ranking in Geospatial Dataset Search

Thomas Hervey[1]
Department of Geography, University of California, Santa Barbara, CA, USA
Center for Spatial Studies, University of California, Santa Barbara, CA, USA
http://www.spatial.ucsb.edu
thomasahervey@ucsb.edu

Sara Lafia
Department of Geography, University of California, Santa Barbara, CA, USA
Center for Spatial Studies, University of California, Santa Barbara, CA, USA
http://www.spatial.ucsb.edu
slafia@ucsb.edu

Werner Kuhn
Department of Geography, University of California, Santa Barbara, CA, USA
Center for Spatial Studies, University of California, Santa Barbara, CA, USA
http://www.spatial.ucsb.edu
werner@ucsb.edu

Abstract

This study surveys the state of search on open geospatial data portals. We seek to understand 1) what users are able to control when searching for geospatial data, 2) how these portals process and interpret a user's query, and 3) if and how user query reformulations alter search results. We find that most users initiate a search using a text input and several pre-created facets (such as a filter for tags or format). Some portals supply a map-view of data or topic explorers. To process and interpret queries, most portals use a vertical full-text search engine like Apache Solr to query data from a content-management system like CKAN. When processing queries, most portals initially filter results and then rank the remaining results using a common keyword frequency relevance metric (e.g., TF-IDF). Some portals use query expansion. We identify and discuss several recurring usability constraints across portals. For example, users are typically only given text lists to interact with search results. Furthermore, ranking is rarely extended beyond syntactic comparison of keyword similarity. We discuss several avenues for improving search for geospatial data including alternative interfaces and query processing pipelines.

2012 ACM Subject Classification Information systems → Environment-specific retrieval; Human-centered computing → Interactive systems and tools; Information systems → Retrieval effectiveness

Keywords and phrases search, portal, discovery, GIR, facet, relevance, ranking

Digital Object Identifier 10.4230/LIPIcs.GIScience.2021.I.5

1 Introduction

It is hard to overemphasize the value of open geospatial portals. In less than a decade, a new generation of Digital Earth [10] has unfolded as thousands of municipalities and other data stewards have created online open geospatial portals, turning voluminous isolated geospatial data into provisioned public resources. Every day, these data are used by citizens to learn about their community services and researchers to study their environments [26]. Some open geospatial portals, herein referred to as *portals*, are small, serving specific datasets from niche domains like soil science (e.g., TERENO), while others are broad, serving as aggregation platforms for datasets across many levels of government (e.g., Data.gov).

[1] corresponding author

While portals have been widely adopted, there is a need for more research to evaluate the state of spatially-enabled search across portals that curate scientific, library, and governmental data. Furthermore, there is a need to bring these communities of practice into conversation so that best practices from each can be shared.

In this paper, we review the state of search on open geospatial portals. We focus on how the design of search has developed across communities of practice for the curation of geospatial research data, civic data, and library data. To the best of our knowledge, there has not been a comprehensive examination of portal search functionality across user communities. This is problematic because these portals collect and disseminate geospatial data that is vital to governance and research. The geographic information science community needs to know what search functionality already exists to better inform future developments of geospatial search and geographic information retrieval (GIR).

This work specifically examines what a user can control while searching and how a portal processes and interprets user queries. We focus on search *facets* and *ranking functions*. As described by [16], we use the term facet to broadly mean a search control that allows a user to further specify a query. Facets let users filter out search results, add criteria for including specific search results, sort results, and sometimes navigate results in a specific way. A ranking function orders results by their relevance to a query. For example, a function could use TF-IDF (term frequency-inverse document frequency), a common numerical statistic comparing keyword frequencies between a query and a set of potential search results. By investigating these two components—facets and ranking functions—we can identify gaps in functionality and search effectiveness. Therefore, our research question is *what is the state of faceted search and ranking functions in open geospatial portals?*

To answer this question, we survey several dozen open portals including Data.gov, DataONE, UCSB's Alexandria Digital Library[2], and ArcGIS Hub [3], and examine the front-end and back-end functionality of their search engines. We first survey search facets and record user search controls, interaction modalities, presentation of results, and navigation. Second, we attempt to record how a portal processes and interprets a user query. Specifically, we record how corresponding data are deemed relevant and ranked. This step proves to be a formidable challenge since many portals do not publicly document how they process and reason on queries. Third, we conduct a qualitative sensitivity analysis of search. We construct several search scenarios, execute corresponding queries, and compare the search results both between portals and after reformulating queries.

In Section 2 we discuss previous work to survey portals and measure their functionality. We then describe our procedure for surveying portals in Section 3. Section 4 is a discussion of our results, including a summary of user search facets, ranking functions, alternative query results, and insights on noteworthy portals. We conclude with a discussion of shortcomings in the functionality of current portals and possible avenues for improving portal effectiveness.

2 Background

Portals curate, organize, and disseminate data for public consumption, often through public web applications[36]. In our work, we study portals on the web[4] that leverage an ecosystem of web services for storing, querying, rendering, mapping, and executing functions (like

[2] https://www.alexandria.ucsb.edu
[3] https://hub.arcgis.com/search
[4] https://portal.ogc.org/files/?artifact_id=6669

geocoding) against geospatial data [23]. In many cases, these portals are an extension of enterprise geographic information systems (e.g., ArcGIS Online) or are integral components of governmental Spatial Data Infrastructures (SDIs) (e.g., INSPIRE).

Many portals strive to uphold FAIR and open data[5] principles: effective and efficient data findability, accessibility, interoperability, and reusability [34, 36]. FAIR principles mark a significant turn for data curation, as they make it possible to evaluate data quality (e.g., fitness for use) across application domains [8]. These principles also influence how effectively a user can search for data. Despite the kind of curatorial "best practices" that FAIR principles suggest, user experiences within portals still vary quite a bit. Simple search interactions across different portals show that search differs noticeably. For example, a search experience for scientific research data is very different from a search for library holdings; there is still relatively little agreement on a similar set of "best practices" for spatial search [3]. Each of these communities has developed distinct metadata standards (e.g., Ecological Metadata Language[6] in the bioinformatics community), adopted different data curation platforms (e.g., DataONE[7]), and anticipated different user needs (e.g., search by taxon facets are in the bioinformatics community). Furthermore, it remains unclear how the spatial dimension of search should interact with other search dimensions, like time and theme, to ensure that a user can search for data effectively.

Search on a portal generally proceeds as follows: through a user interface (UI) a user executes a query, the portal processes and interprets the query (often using a service to georeference a query), and then the system returns relevant resources from a back-end content management system (CMS) or a database management system (DBMS) that houses indexed geospatial data and metadata. The user can then explore the results and either download a resource or reformulate their query to change results. On the back-end of a portal, a processing algorithm is used to rank results. This algorithm, which the ranking function is a part of, influences both UI design and how the system interprets a query. For example, a processing pipeline may follow precreated and hand-tuned rules for filtering query results (based on corresponding UI facets a user adjusts on a portal's interface). There may be other processing steps that a user cannot control (such as text normalization).

Modern portals are the primary outlet for search and discovery of geospatial data. Before portals, geospatial data were collected and curated separately by individual organizations. Both governments and GIS companies realized that they needed a better way to make geospatial data discoverable, usable, and interoperable [32, 13]. As web services for sharing data rose, the GIS industry became interested in using them as a medium for bringing both GIS and geospatial data to a larger audience. These services were the first way to get data to the public, but data quality was low and interoperability proved challenging to achieve. The digital earth and SDI movements allowed governments and organizations to centralize and build a technical infrastructure for managing geospatial data [24]. One such example is the U.S. Federal Geographic Data (FGDC) National Spatial Data Infrastructure[8].

The desire to leverage citizen-generated data moved SDIs in a new direction [14, 35]. Portals complemented SDIs, allowing for publication and aggregation of disparate geospatial data [9, 27, 35]. This made portals a popular and accessible form of a distributed GIS [30] yet publishing geospatial content remained technically difficult. Portal architecture began to

[5] We use the definitions for "open knowledge" and "open data" provided by the Open Knowledge Foundation at `https://okfn.org/opendata/`

[6] `https://eml.ecoinformatics.org/`

[7] `https://www.dataone.org/`

[8] `https://www.fgdc.gov/nsdi/nsdi.html`

evolved to adopt search capabilities. Today, portals disseminate and allow users to explore geospatial data. Some are "mashups" built on services such as Google Search or Google Maps for geocoding and OpenStreetMap for visualizing search results [12, 9].

Recent surveys of government-run open data portals across Australia [31] and the U.S. [33] noted that a large portion of portal growth is driven by governments who seek transparency and want to engage citizens in government initiatives. In Australia, for example, dozens of small and large government portals are successful, because they continually publish datasets, refine and clarify open data policies [7], and increase visibility through citizen engagement events like government-sponsored hackathons. Although these surveys are informative, their authors suggest that their survey methods are preliminary. Portal adoption across government, research, and libraries has been rapid in the last few years, so general measures for portal functionality, quality, and effectiveness are still in their infancy. Viewport-based GIR systems have been proposed to support comparison based on the semantic similarity of their features; however, such systems do not yet support realistic information needs [4].

Most current geospatial search challenges and opportunities are described in [29] including novel opportunities like personalized search and interpreting local intent [1], intelligent ranking algorithms based on machine-learned feature combinations [18], and challenges like cataloging cross-disciplinary geospatial search needs or bolstering theory-driven geoparsing methods. Some insights into how portals process and interpret user queries are available from the perspective of portal developers. Search and discovery scenarios in the library community are well illustrated by GeoBlacklight developers [15] and the Alexandria Digital Research Library Project [11]. Advances from the research data community are illustrated by the adoption of standards, such as FAIR data principles [34] and by the examination of challenges in developing domain repositories [25]. Lastly, the adoption of civic data portals across multiple levels of government in the U.S. and E.U. [36] and ArcGIS Online as an open data platform [20] illustrate how user data needs are anticipated and handled.

3 Survey Methods

In this work, we seek to understand three things about portals: 1) what facets a user can control that affect search results, 2) how a portal processes and interprets a user's query (for ranking results), and 3) if and how reformulating a query changes search results. By answering these questions, we gain an understanding of the current capabilities and limitations when a user searches on a portal.

We specifically surveyed 1) search facets, 2) interaction modalities (e.g., maps and text list views), 3) adherence to FAIR principles, and 4) ranking functions (e.g., BM-25, TF-IDF). Note that our methodology uses an individual search and judgement process run solely by the authors. For example, relevance judgements for search results and the adherence to FAIR principles were qualitative judgements. However, the interpretation of FAIR principles remains largely subjective and open to interpretation; to address this, FAIR metrics [34] including rubrics for tools, datasets, and repositories[9] are currently under development. The metrics that are being developed for repositories focus mainly on licensing, protocols, and resource description. Following this, we gave a portal a satisfactory rating for adherence to FAIR principles if its datasets followed at least three of the four main principles–findable, accessible, interoperable, resuable. We gave a portal a higher rating if its datasets followed all four principles, metadata were well documented, and the portal supported to its users,

[9] `https://fairshake.cloud/rubric/`

such as through blog posts on how to use and manipulate dataset metadata. Due to time constraints, we were not able to gather subjects and pool a larger set of judgements. However, in future work we plan to conduct A/B testing between a control and modified search system during which we will gather more judgements from test subjects.

To start, we created a list of sample portals to survey. We hand selected our sample from across three main communities that curate open geospatial portals: 1) civic data portals, 2) scientific research portals, and 3) library portals. These were drawn from four online sources that curate a list of portals, GIS data sources, and GIS learning resources (e.g., Awesome-GIS[10], Awesome-Geospatial[11]). We briefly visited and tested all of the portals on these curated lists and narrowed our list to 35 sample portals. In the remainder of this paper, we will discuss the results from nine unique and diverse portals.

A portal was considered for our list if it: 1) hosts 50 or more open geospatial datasets, 2) has datasets published within the past six months, and 3) provides a way for users to search for datasets. We wanted to achieve broad diversity in our sample. Therefore, we ended up surveying 35 portals that differ in purpose, topic, geographic coverage, or curating body. Two examples of purpose are citizen engagement and academic data reuse. Examples of curating bodies include governments and municipalities, libraries, non-government organizations, and academic institutions. Portals in our list needed to serve georeferenced datasets in formats like .geojson, .shp, .TIFF, or .netCDF (so that can be used in traditional GIS or spatial analysis tasks). For this reason, we intentionally did not survey gazetteers and point of interest (POI) search tools like the World Historic Gazetteer[12], Frankenplace[13], or Yelp[14].

For each portal in our list, we initially took a "follow-your-nose" approach to surveying. This means that when we arrived at the root of a site, we would read the home page and begin exploring by clicking on prominent links. We then reviewed any available documentation for users and developers. Documentation is also useful for understanding how curators articulate the purpose and suggested usage of a portal. When available, we also read open data policies, search and interface user guides, and technology and metadata descriptions.

We then tested a portal's search interface. We first navigated to the root search page (which sometimes was on the portal's home page). Once there, we recorded all the options that a user could control, which included the mode of interaction (e.g., map-based, list-based), navigation (e.g., number of pages, page hierarchy), and search facets (e.g., text search box, filters, map controls, sorting).

Next, we documented portal ranking functions to the extent possible. This was difficult because many systems use proprietary and/or closed-source search engines that do not disclose their ranking functions. Some portals have an application programming interface (API), to bypass the UI and access dataset metadata directly. In some cases, API documentation gave insights into how a query is parameterized and how results are ranked. In other cases, we were able to read documentation from portals that use open-source CMSs, such as CKAN, and a few portals gave us access to internal documentation on their ranking functions.

Our last step was to see how effective and sensitive search is on these portals. The purpose of this step was to 1) try and bolster our understanding of how non-disclosed ranking functions work, and 2) test how sensitive ranking functions are to changes in a user's query.

[10] `https://github.com/sshuair/awesome-gis`
[11] `https://github.com/sacridini/Awesome-Geospatial`
[12] `http://whgazetteer.org/`
[13] `http://www.frankenplace.com/`
[14] `https://www.yelp.com/`

To do this we, selected nine portals from our list on which to execute queries (listed in Table 1). Once we were familiar with their search pages and what specific datasets were available, we developed several search scenarios.

Table 1 Descriptive characteristics of a subset of surveyed portals. For each portal, we recorded the number of public datasets/cataloged items, the temporal range of datasets (starting from either the application time or the creation time), and the coverage (geographic and community focus).

Portal	Datasets	Time	Coverage
DataONE	820k +	1800 - present	global (environmental science)
Data.gov	250k +	mid-1800s - present	U.S. (none, authoritative)
ArcGIS Hub	178k +	1700s - present	global (none, semi-authoritative)
USGS	100k +	2000 - present	U.S. (none, authoritative)
ADRL	33k +	1860 - 2018	California, misc. (library data)
Tereno	1000 +	1995 - present	Germany (environmental science)
INSPIRE	6500 +	1900 - present	Western Europe (none, authoritative)
NASA	6600 +	1587 - present	global (Earth observations)
Heritage Gateway	60+ sources	prehistoric - present	England (structures and landmarks)

These scenarios were modified from three personas of application end users that are described in the GeoBlacklight concept design[15]; for a better understanding of our search scenarios, we refer readers to their descriptions [15]. These personas include a professor of History, a Ph.D. candidate in Environmental Science, and an undergraduate sophomore studying urban planning. Each persona has a motivation, scenario, and expectations of a portal. Although they are exclusively academic, they vary enough to realistically resemble search scenarios from other personas. Based on our interpretation of the persona descriptions, we created a specific search task and respective query to simulate that persona initiating a search. An example of a search scenario is as follows. History professor Brian Diaz needs data about historical and modern churches in Scotland. He does not have a lot of time and he likes using text search, but would also be happy narrowing results using a map. He searches two portals, Heritage Gateway and the INSPIRE Geoportal. He executes and refines the text of several queries without any additional facet adjustments. Example refinements include “churches”, “modern churches”, and “modern churches edinburgh”. When he cannot find relevant results, he tries modifying his queries (using reformulation techniques outlined in Table 2).

After the scenarios were created, we executed an initial query for each, recording the resulting datasets and the number of datasets we considered to be relevant to our search needs. We then iteratively reformulated the query 12 times, and re-recorded the results and portion of the results that we considered to be relevant. We repeated this process for each scenario on each portal with three different initial queries. Table 2 shows the types of query reformulations we used, which were gathered from [19, 21, 22].

4 Results

Some portals are small and have few datasets. For example, TERENO serves soil and geochemical datasets from a few environmental research observatories in Germany. Some portals are curated by small municipalities such as Mono County, California, U.S. Others are

[15] https://geoblacklight.org/documents/GeoBlacklight%20Concept%20Design%20v0.3.3.pdf

Table 2 Types of query reformulations executed during search scenarios. Reformulation types adapted from definitions found in [19, 22, 21].

Query	Reformulation	Type	Purpose
"modern churches"	"churches"	generalization	broaden
"churches"	"modern churches"	specialization	narrow
"modern churches"	"modern temples"	word substitution	change meaning
"modern churches"	"churches modern"	repeat	reformat
"modern churches"	"catholic cemeteries"	new	change meaning
"edinburgh"	"glasglow"	geo-modification	intent modification
"edinburg"	"edinburgh"	geo-correction	correct spelling
"edinburgh"	"edinburgh tx"	geo-disambiguation	placename disambiguation
"churches"	"churches edinburgh"	place insertion	narrow geographically
"churches edinburgh"	"churches"	place deletion	broaden
"food Scotland"	"food Europe"	granularity change	broaden or narrow

large with sophisticated search tools and have many datasets. For example, ArcGIS Hub serves many datasets of widely varying topics and global coverage. We strived for diversity and wanted to ensure that we were not just sampling popular portals. Furthermore, we believed that smaller portals may have more specific user controls since they likely wouldn't have to manage a large amount of diverse datasets. Several smaller portals included those run by The Nature Conservancy[16], Lithuania's federal government[17], and Cyprus's Department of Land and Survey[18].

As mentioned in Section 3, we sampled 35 portals from the lists and will now discuss nine in particular. These nine were chosen because they capture the diversity of all portals sampled. Table 1 includes descriptive characteristics of these nine portals including number of public datasets/cataloged items, the temporal range of datasets (starting from either the application time or the creation time), the geographic coverage, and the focus community/theme. Table 3 describes search facets and our understanding of the ranking functions on these portals. We show that the nine portals mostly show results in list form. first and all include lists. Some show results in map form first. We believe that all portals could improve in their employment of the FAIR principles. DataONE employed FAIR principles the best because they document the ways in which they promote FAIR use such as through webinars. The following subsections describe search facets, ranking functions, and results from query reformulations in more detail.

4.1 Front-End: User Search Controls

On almost all sample portals surveyed, users have the same core set of facets when searching. First, users are given an omnibox for entering free text. A user enters a text query using keywords or natural language. Second, users are given at least two pre-configured facets to refine their text search. Facets are typically located in a sidebar and are check box, radio button, or range slider toggles. A common selection facet is *tag*, which lets users select a descriptor tag that has been associated with a dataset. Approximately half of the portals surveyed have an advanced search feature with an extended interface. Typically, this lets

[16] http://maps.tnc.org/gis_data.html
[17] https://www.geoportal.lt/geoportal/web/en/
[18] https://eservices.dls.moi.gov.cy/#/national/geoportalmapviewer

Table 3 Search characteristics of a subset of surveyed portals. For each portal, we recorded the ease of navigation, the interaction modalities (e.g., list-based, map-based, visualization-based), the types of search facets (e.g., filters, result sorts), ranking functions, and degree of employment of FAIR principles.

Portal	Facets	Modality	FAIR princples	Ranking Function
DataONE	*filters:* [data attribute; annotation; data files; member node; creator; year; identifier; taxon; location], *sort by:* [most recent; identifier; title; author]	map + list	very good	- BM-25 - query expansion
Data.gov	*filters:* [topics; topic categories; dataset type; tags; format; organization type; organizations; publishers; bureaus; location], *sort by:* [relevance; time/date; popular; date added]	list	good	- TF-IDF
ArcGIS Hub	*filters:* [capabilities; source; content; type; tags], *sort by:* [relevance; most recent; trending; name]	list	good	- BM-25 - Query expansion
USGS	*filters:* [map controls; file format; extent; topic sub-category]	map	satisfactory	- TF-IDF
ADRL	*filters:* [search by all fields, title, subject, or accession number; format; collection; contributor; topic; place; genre; date; academic department; library location; rights], *sort by:* [relevance, year created, creator]	list	satisfactory	- TF-IDF
TERENO	*filters:* [“what?” by topic, keywords, sensor type, parameter; “where?” by metadata fields, catalog, regions, map extent; “when?” by date range]	map + list	satisfactory	- TF-IDF
INSPIRE	*text search:* Select country then search by dataset title, *filters:* [country; spatial scope; theme]	list	satisfactory	- TF-IDF
NASA	*map options:* [region, time, hand-drawn region], *filters:* [features; keywords; platforms; instruments; organizations; projects; processing levels; granule data format], *sort by:* [relevance; usage; end date]	list + map	satisfactory	- TF-IDF
Heritage Gateway	*filters:* [“where?” by search geocoder; “what?” by thesaurus of building, object, or evidence type; “who?” by associated person, architect; “when?” by date range, period; “resource?” by parent organization]	list or map	satisfactory	- TF-IDF

users specify additional filters based on less popular metadata. Third, once a user executes a query and the results are presented as a list, users are given a option for sorting the results. For example, the user can sort by relevance (discussed in Section 4.2), by the date that datasets were created/modified, or alphabetically by dataset title. Note that we did not extensively survey individual results or additional result pages, only the first page result list after a query was executed.

UI complexity and navigation varied substantially. Approximately half of the portals have an omnibox for text search or icons for pre-created search topics on their home page. The other half of portals lead users to search through a button or link with a label like *“find data”*. Approximately one fourth of the portals surveyed initially present results as a map or a map with a list. Users can then navigate the results geographically and further refine by clicking on a specific result, or a region that was labeled with the number of results located in that region.

4.2 Back-End: Query Processing and Interpretation

As previously mentioned, it is difficult to determine exactly how a portal processes and interprets a query (and determine which potential search results are relevant) without knowing their ranking algorithm. Fortunately, many portals are built using an open source CMS that use open-source search engines. Many portals in our survey, especially government portals,

used one of three CMSs for all or part of their back-end processing: CKAN, Socrata, or ArcGIS Hub (or ArcGIS OpenData). Previous surveys [31] suggest that investing in open data portals is typically expensive and labor intensive, so it's reasonable to assume that such systems are appealing for hosting and/or serving data. Other geographic data CMSs, such as GeoBlacklight or Samvera, are only popular within specific communities.

Facets are almost always used to forumate a query before execution. However, typically portal UIs include facets on search result pages so that users can refine their queries. Calculating relevance is at the heart of a ranking function. Typically, a relevance score for a potential search result is a composite value calculated by combining one or more weighted criteria. Most portals appear to use the TF-IDF algorithm for scoring potential search results. This ranking algorithm works by comparing keyword frequency between a query and one or more text attributes (such as title, abstract, and tags) of indexed datasets. Several portals, such as ArcGIS Hub, use hand-tuned boosting in their ranking functions to increase the importance of certain criteria (like keyword frequency in a dataset's tags). At least a quarter of the portals surveyed use a CMS (such as CKAN) that leverages Apache Lucene or Apache Solr as their search engine with no noticeable customization beyond the default search ranking function (i.e., using the bag-of-words model and TF-IDF). For example, searching on Data.gov, on the back-end, a user's query is most likely tokenized, sanitized, normalized, and converted compare with potential search results (which also represented as bags of words). ArcGIS Hub uses Elasticsearch[19] as a search engine and the BM-25 ranking algorithm. For the portals where we were able to see the private ranking function including ArcGIS Hub and Heritage Gateway, keyword frequency match is the most important weight for raking potential search results. We found that recency and popularity (e.g., download frequency) are occasionally used in calculating the score (upward of 25 percent of a score). DataONE and ArcGIS Hub, use query expansion to include results with relevant taxa and synonyms. For example, searching "robbery" on an ArcGIS Hub site will give similar but more granular thematic results such as "crime."

Open geospatial portals are unique from other open data portals due to their handling of space and space's relation to theme and time. This is a difficult task and the subfield of GIR is dedicated to effectively serving relevant geospatial information. A technique at the heart of GIR is georeferencing a textual query. This means disambiguating and resolving geographic references, often toponyms, properly interpreting spatial relations, and inferring the geographic intent of the query. In the most basic application, this means interpreting <*theme*><*spatial relation*><*location*> such as "churches in Poznań Poland" [28, 29]. We did not find any portal that interprets a query this way, although we believe that they do exist. At best, portals attempt to properly interpret the intent of thematic terms through techniques like query expansion. A spatial relation such as "near" is usually ignored or assumed to mean containment. Surprisingly, location appears to be ignored during query processing or specified separately in the UI the omnibox input. For example, several portals (DataONE, Heritage Gateway) let users specify a location by typing in toponyms in a separate text box, which is then geocoded and matched with pre-created regions.

4.3 Effects on Results from Query Reformulations

As stated in Section 3, we ran three search scenarios on our nine focus portals. Overall, the results were mostly what we expected. Most portals surveyed did not indicate a ranking function that leverages techniques beyond the bag-of-words model and TF-IDF

[19] `https://www.elastic.co/`

statistic. We determined this because repeat, reformulations had no effect but generalizations, specializations and word substitutions did. DataONE is an exception because the query "coral disease" yielded more results and more relevant results than "disease coral". In most cases, generalization resulted in more results and and about half the time more relevant results. All portals sampled were sensitive to generalization, specialization, and word substitution reformulations. Once again, this makes sense because most portals are keyword sensitive, so including or removing keywords tended to make a large difference.

In all of the portals surveyed, making geo-modifications, geo-corrections, or geo-disambiguations during a query reformulation did not have noticeable effects on query results. For example, a geo-disambiguation from "edinburgh" to "edinburgh tx" usually reduced the number of results likely because more keywords were included, not because a geography was disambiguated and constrained. A geo-correction from "edinburg" to "edinburgh" increased results. Changing spatial granularity didn't appear to be any different from word substitution except in a few portals like ArcGIS Hub. With access to documentation on Hub's query processing, we know that Hub leverages a process for comparing locations with different spatial granularities. For example, results for "scotland" were more frequent than results for "edinburgh" and not simply because of increased keyword frequency.

4.4 Noteworthy Examples

Search functionality on several portals was unique enough to merit distinction. These portals used novel techniques to make searching easier or more specific. The first portal is Heritage Gateway, which is a historic building and landmark portal for England. On their portal, users interact with search almost entirely through a map. Once users execute a query, results are displayed as symbolized points on a map. Users can then click on individual features to read more about them in a pop-up window or download them. The portal's advanced search lets users refine a text query with *where*, *what*, *when*, and *who* filter criteria. Figure 1 shows a portion of the XML schema for query processing with specific ways users can set criteria (e.g., specifying *where* using a reference system like *gridref*, *osgridref*, *latitude*, *longitude*, etc).

A second notable portal is DataONE. DataONE's UI is shown in in Figure 2. The search interface is primarily map-based with a sidebar for entering text queries and numerous filters. As users pan with the map, search results and facets automatically update to reflect only what datasets are visible. DataONE uniquely disseminates data via RSS, GeoRSS, and other data casting feeds. In 2020, DataONE may start a collaboration with RDMLA[20] for guiding data management and curation best practices that has the potential to support curation efforts in other scientific repositories and libraries.

4.5 Suggested Improvements

As the results show, most portals treat search similarly. On the front end, most portals implement an omnibox for text input and facets to balance user control with effort. On the back end, most portals use a bag-of-words model to represent queries and potential search results, and match them based on keyword frequency. Generic portals also appear to be similarly designed. However, we argue that the uniqueness of geospatial content merits more sophisticated search facets and ranking functions than those that are currently used.

We recommend that most portals transition to primarily map or other visualization modalities (like topic explorers), instead of lists and text boxes. For example, interfaces could orient around self-organizing topic maps [20] or geographic maps. Most interfaces that we

[20] `https://rdmla.github.io/`

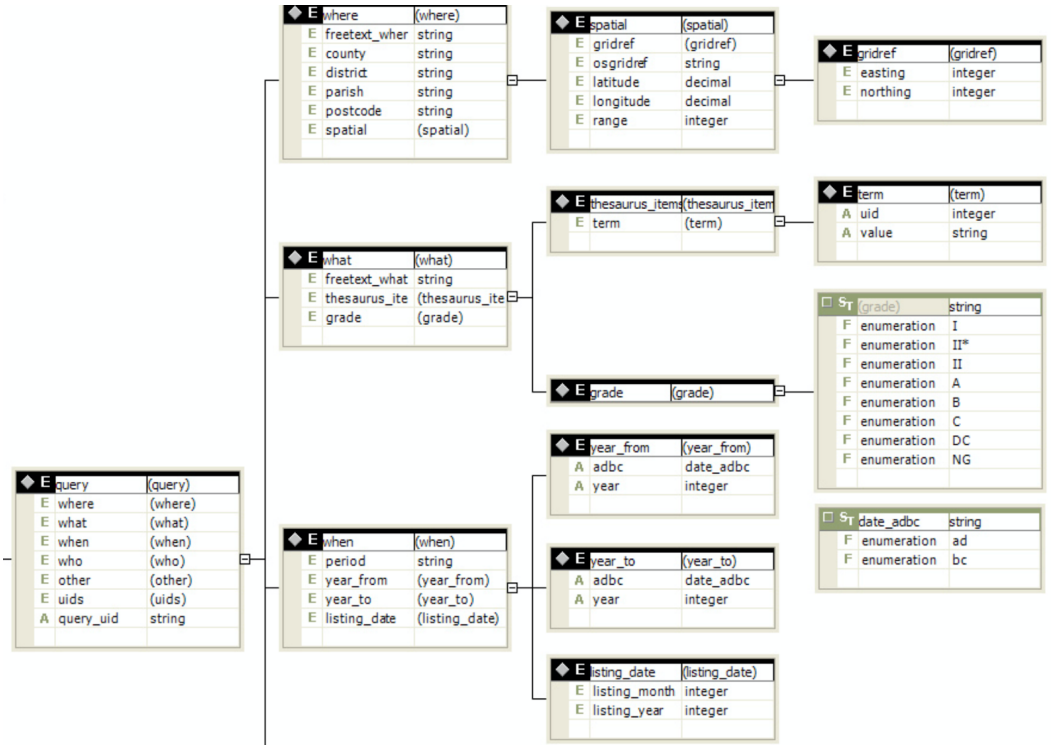

Figure 1 A portion of an XML schema for user query criteria on the Heritage Gateway portal. A text query can be refined by any number of top level criteria on the left including where, what, when, who, other, uids, or query-uid. Each criteria can be further specified such as year-from under the when criteria which is a custom date, or grade which is one of nine predetermined building grades such as I,II,III, DC, or NG.

surveyed currently offer minimal help tools. Therefore, offering users comprehensive guides or wizards for navigating search interfaces could simplify the search process. One idea would be to display a dataset in a map-based interface and then, using a wizard, ask users what aspects they would like to change in order to find other datasets. Regarding search facets, we believe that most portals include too many search filters instead of dedicating resources to improving natural language query interpretation. Effectively balancing functionality and usability is difficult, and unfortunately most web search interfaces fail at this [2].

In terms of developing explicitly spatial ranking and relevance metrics, we saw few examples of portals that did this. One suggestion is to build upon the multidimensional ranking scheme proposed by Sharma and Beard [5] that use space, time, and theme as dimensions. The spatial component of a result would be weighed based on topological relations; the temporal component would be weighed based on Allen intervals; and multiple thematic components would be selected by user in the form of "glyphs". Any score boosting for a dimension should be based on the portal's needs. This solution brings the three key dimensions of spatial information [6] to bear in developing ranking and relevance metrics for spatial data.

There is a clear need for finer grained spatial and thematic processing and interpretation. Few portals compare individual data values to a text query. Those that do only do so when a user specifies advanced search features. However, these features are typically relegated to comparing dataset metadata to a query, not individual data values within a dataset. In other

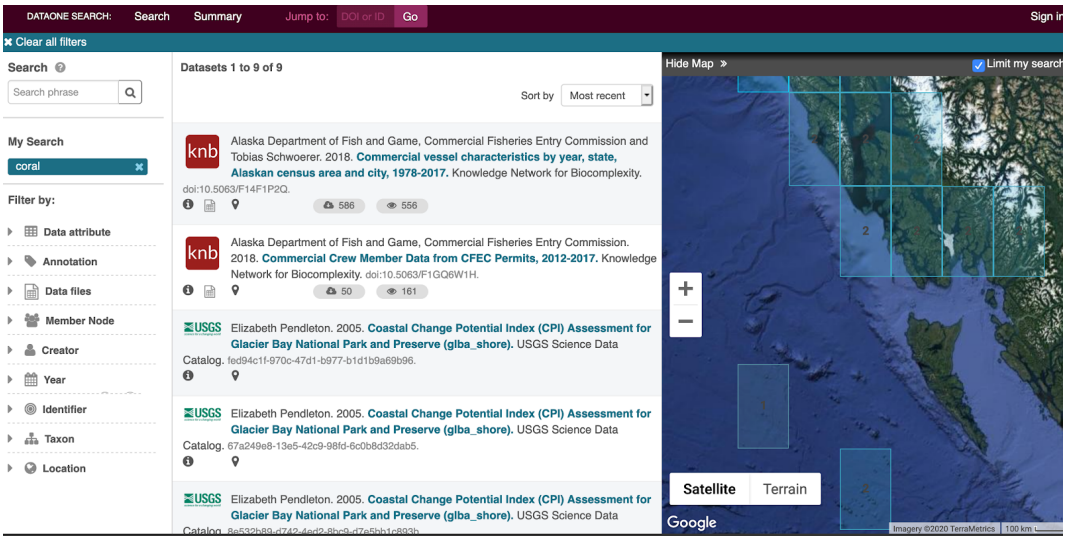

Figure 2 The search interface on DataONE. After executing a text query, users can filter results using criteria including data attribute, year, and taxon. The list of results changes when filters are applied or when a user pans the map and/or selects a gridded region of interest.

words, systems should extract numeric values and ordinal values (like "most" or "nearest") from a query and compare them with potential search results using at a minimum hand-tuned rules. These improvements parallel a Semantic Web goal of returning specific data points for a query, not just datasets [17].

5 Conclusion

In this work we have taken a critical look at the current state of search on open geospatial portals. We surveyed the front end of systems and focused on search facets, a type of control users have while searching. We then surveyed how the back ends of systems process and interpret queries, and how they rank relevant results. To corroborate our understanding about the back-end of these systems and test how effective searching is, we executed several search scenarios. In these scenarios we iteratively reformulated queries against nine specific portals. We found that most portals leverage an omnibox for raw text search and filters to refine them. We also found that most portals use a syntactic-based keyword frequency model for representing queries and potential search results (found in most basic search architectures). As expected, after most query reformulations, changes in results were simple and aligned with what we would expect from this model. We then described distinctive characteristics of nine unique portals and further detailed two notable portals and why they stood out as models for geospatial search.

Open geospatial data portals, which are growing in popularity as resources for accessing geospatial data, have an opportunity to be forefront models of advanced GIR and geospatial computing. However, based on the current state of search facets and ranking, there are several substantial improvements needed to make portals easier to use, easier to navigate, and adhere better to FAIR principles. Optimally, in addition to search, portals would more effectively enable serendipitous discovery.

There are several notable limitations to this work. First, we were not able to quantitatively assess the effectiveness of each portal surveyed. In future work, we plan to create effectiveness criteria based on explicit relevance feedback. Also, since many portals use proprietary search

engines, we were not able to explicitly see how their ranking functions work. However, the intention of this work is to survey, frame, and motivate a quantitative analysis of user search behavior. In future work, we plan to use query logs from the ArcGIS Hub platform to model search behavior. Through those efforts, we hope to see if and how search success and abandonment patterns relate to the limitations of the portals surveyed herein.

References

1. Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. Spatial variation in search engine queries. In *Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08*, pages 357–366, 2008. doi:10.1145/1367497.1367546.
2. Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.
3. Andrea Ballatore, Werner Kuhn, Mary Hegarty, and Ed Parsons. Special issue introduction: Spatial approaches to information search. *Spatial Cognition and Computation*, 16(4):245–254, 2016. doi:10.1080/13875868.2016.1243693.
4. Andrea Ballatore, David C Wilson, and Michela Bertolotto. A holistic semantic similarity measure for viewports in interactive maps. In *International Symposium on Web and Wireless Geographical Information Systems*, pages 151–166. Springer, 2012.
5. Kate Beard and Vyjayanti Sharma. Multidimensional ranking for data in digital spatial libraries. *International Journal on Digital Libraries*, 1(2):153–160, 1997. doi:10.1007/s007990050011.
6. Brian J.L. Berry. Approaches to Regional Analysis: A Synthesis. *Annals of the Association of American Geographers*, 54(1):2–11, 1964. doi:10.1111/j.1467-8306.1964.tb00469.x.
7. John Carlo Bertot, Ursula Gorham, Paul T. Jaeger, Lindsay C. Sarin, and Heeyoon Choi. Big data, open government and e-government: Issues, policies and recommendations. *Information Polity*, 19(1-2):5–16, 2014. doi:10.3233/IP-140328.
8. Bradley Wade Bishop and Carolyn Hank. Measuring fair principles to inform fitness for use. *International Journal of Digital Curation*, 13(1):35–46, 2018. doi:10.2218/ijdc.v13i1.630.
9. Christopher Bone, Alan Ager, Ken Bunzel, and Lauren Tierney. A geospatial search engine for discovering multi-format geospatial data across the web. *International Journal of Digital Earth*, 9(1):47–62, 2016. doi:10.1080/17538947.2014.966164.
10. Max Craglia, Michael F Goodchild, Alessandro Annoni, Gilberto Camara, Michael Gould, Werner Kuhn, David Mark, Ian Masser, David Maguire, Steve Liang, and Ed Parsons. A position paper from the Vespucci Initiative for the Advancement of Geographic Information Science. *International Journal of Spatial Data Infrastructures Research*, 3:146–167, 2008. doi:10.2902/1725-0463.2008.03.art9.
11. James Frew, Michael Freeston, Nathan Freitas, Linda Hill, Greg Janee, Kevin Lovette, Robert Nideffer, Terence Smith, and Qi Zheng. The Alexandria digital library architecture. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1513:61–73, 1998. doi:10.1007/pl00021470.
12. Tamar Ganor. An Integrated Spatial Search Engine for Maps and Aerial Photographs on a Google Maps API Platform. *Journal of Map and Geography Libraries*, 13(2):175–197, 2017. doi:10.1080/15420353.2016.1277574.
13. Christian Philipp Geiger and Jörn Von Lucke. Open Government and (Linked) (Open) (Government) (Data). *JeDEM - eJournal of eDemocracy and Open Government*, 4(2):265–278, 2012. doi:10.29379/jedem.v4i2.143.
14. Michael F. Goodchild, Pinde Fu, and Paul Rich. Sharing geographic information: An assessment of the geospatial one-stop. *Annals of the Association of American Geographers*, 97(2):250–266, 2007. doi:10.1111/j.1467-8306.2007.00534.x.
15. Darren Hardy and Kim Durante. A Metadata Schema for Geospatial Resource Discovery Use Cases. *Code4Lib Journal*, 25:1–1, 2014. URL: http://journal.code4lib.org/articles/9710.
16. Marti Hearst. User interfaces for search. *Modern Information Retrieval*, pages 21–55, 2011.

17 Krzysztof Janowicz, Frank van Harmelen, James A Hendler, and Pascal Hitzler. Why the Data Train Needs Semantic Rails. *AI Magazine*, 36(May):5–14, 2015. doi:10.1609/aimag.v36i1.2560.

18 Yongyao Jiang, Yun Li, Chaowei Yang, Fei Hu, Edward M. Armstrong, Thomas Huang, David Moroni, Lewis J. McGibbney, and Christopher J. Finch. Towards intelligent geospatial data discovery: a machine learning framework for search ranking. *International Journal of Digital Earth*, 11(9):956–971, 2018. doi:10.1080/17538947.2017.1371255.

19 Rosie Jones, Wei Vivian Zhang, Benjamin Rey, Pradhuman Jhala, and Eugene Stipp. Geographic intention and modification in web search. *International Journal of Geographical Information Science*, 22(3):229–246, 2008. doi:10.1080/13658810701626186.

20 Sara Lafia, Andrew Turner, and Werner Kuhn. Improving discovery of open civic data. *Leibniz International Proceedings in Informatics, LIPIcs*, 114(9):1–9, 2018. doi:10.4230/LIPIcs.GIScience.2018.9.

21 Tessa Lau and Eric Horvitz. Patterns of search: analyzing and modeling web query refinement. In *UM99 user modeling*, pages 119–128. Springer, 1999.

22 Chang Liu, Jacek Gwizdka, Jingjing Liu, Tao Xu, and Nicholas J. Belkin. Analysis and evaluation of query reformulations in different task types. *Proceedings of the ASIST Annual Meeting*, 47, 2010. doi:10.1002/meet.14504701214.

23 David J. Maguire and Paul A. Longley. The emergence of geoportals and their role in spatial data infrastructures. *Computers, Environment and Urban Systems*, 29(1 SPEC.ISS.):3–14, 2005. doi:10.1016/j.compenvurbsys.2004.05.012.

24 Ian Masser. *GIS worlds: creating spatial data infrastructures*, volume 338. Esri Press Redlands, CA, 2005.

25 Matthew S Mayernik. Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, 67(4):973–993, 2016. doi:10.1002/asi.23425.

26 Karen Okamoto. What is being done with open government data? An exploratory analysis of public uses of New York City open data. *Webology*, 13(1):1–12, 2016.

27 Ricardo Oliveira and Rafael Moreno. Harvesting, integrating and distributing large open geospatial datasets using free and open-source software. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 41(July):939–940, 2016. doi:10.5194/isprsarchives-XLI-B7-939-2016.

28 José M. Perea-Ortega, Miguel A. García-Cumbreras, and L. Alfonso Ureña-López. Evaluating different query reformulation techniques for the geographical information retrieval task considering geospatial entities as textual terms, 2012.

29 Ross S. Purves, Paul Clough, Christopher B. Jones, Mark H. Hall, and Vanessa Murdock. Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations and Trends® in Information Retrieval*, 12(2-3):164–318, 2018. doi:10.1561/1500000034.

30 Michael G. Tait. Implementing geoportals: Applications of distributed GIS. *Computers, Environment and Urban Systems*, 29(1 SPEC.ISS.):33–47, 2005. doi:10.1016/j.compenvurbsys.2004.05.011.

31 Akemi Takeoka and Christopher G Reddick. A longitudinal cross-sector analysis of open data portal service capability : The case of Australian local governments. *Government information quarterly*, 34:231–243, 2017. doi:10.1016/j.giq.2017.02.004.

32 W Tang and J Selwood. Spatial portals: Adding value to spatial data infrastructures. In *ISPRS Workshop on Service and Application of Spatial Data Infrastructure*, pages 14–16, 2005.

33 Jeffrey Thorsby, Genie N.L. Stowers, Kristen Wolslegel, and Ellie Tumbuan. Understanding the content and features of open data portals in American cities. *Government Information Quarterly*, 34(1):53–61, 2017. doi:10.1016/j.giq.2016.07.001.

34 Mark D. Wilkinson, Susanna Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino Da Silva Santos, and Michel Dumontier. Comment: A design framework and exemplar metrics for FAIRness. *Scientific Data*, 5:1–4, 2018. doi:10.1038/sdata.2018.118.

35 Phil Yang, John Evans, Marge Cola, Steve Marley, Nadine Alameh, and Myra Bambacus. The emerging concepts and applications of the spatial web portal. *Photogrammetric Engineering and Remote Sensing*, 73(6):691–698, 2007. doi:10.14358/PERS.73.6.691.

36 Anneke Zuiderwijk and Marijn Janssen. Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, 31(1):17–29, 2014. doi:10.1016/j.giq.2013.04.003.