Borst, Timo; Limani, Fidan



# Patterns for Searching Data on the Web Across Different Research Communities

LIBER Quarterly

*Suggested Citation:* Borst, Timo; Limani, Fidan (2020) : Patterns for Searching Data on the Web Across Different Research Communities, LIBER Quarterly, ISSN 2213-056X, LIBER, The Hague, Vol. 30, Iss. 1, pp. 1–21,
https://doi.org/10.18352/lq.10317

LIBER QUARTERLY
The Journal of the Association of European Research Libraries
Vol. 30 (2020) 1–21 | e-ISSN: 2213-056X

# Patterns for Searching Data on the Web Across Different Research Communities

**Timo Borst**

Leibniz Information Center for Economics, Kiel, Germany
t.borst@zbw.eu, orcid.org/0000-0002-2481-029X

**Fidan Limani**

Leibniz Information Center for Economics, Kiel, Germany
f.limani@zbw.eu, orcid.org/0000-0002-5835-2784

## Abstract

Being a concept quite familiar in the domain of information retrieval, data search in a web based environment has recently gained attention. With researchers and academic institutions increasingly publishing their data on the public web, traditional research workflows with respect to data search are subject to empirical analysis, user studies, re-engineering and service development. We investigate these workflows more in detail and introduce three patterns of web-based data search intended to serve both as a general reference and as a starting point for discipline specific adoptions. We give some real-world examples in terms of existing web applications and GUI components, thereby suggesting a combination of both generic and community specific approaches towards solutions for data search. We further analyze these patterns by means of empirical evidences we found in some research communities, before giving a summary and outlook on future work.

**Keywords:** data search; data discovery; data analysis; data exploration


Uopen Journals | http://liberquarterly.eu/ | DOI: 10.18352/lq.10317

## 1. Introduction

Search for data—or just 'data search' has been a key activity of any information agent, be it human or non-human. From their very beginnings in the 60s, modern information systems were designed to support information needs and information behaviour related to data (e.g., by crafting decision support systems in relation to human-computer-interaction, cf. Hirschheim & Klein, 2012), and myriads of database technologies and databases including data dictionaries, interfaces and query languages were primarily focused just on that purpose. So how could it still be a matter of ongoing or even increasing research and development? We see mainly two reasons for this: First, by the term 'data search' we apparently mean something more than traditional database retrieval: it is not only about searching and retrieving results from databases according to their local schemes and query interfaces, but primarily about searching across many and distributed, potentially growing data sources just in the way of a web search that is based on federated metadata and/or web pages describing the data. The second reason is a shift in the paradigm towards managing and publishing those data: Instead of keeping it on their local hard drives or access-restricted institutional servers, researchers and their funded projects are requested more and more to expose and publish their data on the web according to the FAIR principles (Wilkinson et al., 2016), which are promoting common web search. In certain contrast to the common access-restricted solutions and infrastructures, data search is mainly interested in publicly available (meta-)data.

This implies many considerations on an operational level, which are hardly being solved or tackled in terms of established conceptual frameworks, technical solutions or widely accepted information practices. For example, what resembles or distinguishes data search from web or document search in terms of related concepts such as similarity or relevance of search results? What is the surrogate from data to be exposed, transferred and indexed? What is the role of descriptive metadata in search processes, and how can they be generated or curated? How do textual queries relate to data which is essentially numerical? How can data search in one domain be transferred to other research disciplines? How are search results displayed with respect to data? How do we search and retrieve data at all? How can search results on datasets be ranked at best considering the fact that there might be other relevance factors than in traditional bibliographic resources (Lewandowski, 2015).

This paper cannot treat all these fundamental questions in detail, but elaborates *three basic patterns of data search* to be supported by web-based infrastructures: data must be first discovered, then explored, and finally analyzed. We start with a section on related work, and then continue by fixing these patterns more in detail to set up a conceptual framework intended for discussing, designing and finally implementing services supporting data search. We illustrate how these patterns could be supported by the example of already existing applications for web-based data search. Finally, we report about evidences we collected during a research project while investigating search behaviour and requirements from different research communities.

## 2. Related Work

With the formation of infrastructures for managing research data and related conceptual and programmatic approaches such as the FAIR principles, scientific data has recently become a first-class citizen of information objects. Likewise, research and development on web-based data search, including conceptual approaches, empirical studies and application development, have become new topics in information and computer science (Kern & Mathiak, 2015). Several works have addressed similarities and differences between document and data search (Kern & Mathiak, 2015; Stempfhuber & Zapilko, 2009), emphasizing the need for new approaches and solutions with respect to the latter. As information objects, data are different from documents, for they are primarily numerical and non-textual (Takeuchi, Sugiura, Akahoshi, & Zettsu, 2017). Kern and Mathiak (2015) observed that users put much more effort in detailed querying of data, expressing locations, ranges and operators at an average word count of nine. Especially in natural, experimental and observational sciences, data is inherently related to spatial and temporal indicators. While the discovery of documents may be supported by 'proxies' such as an abstract or a title representing a document's (protected or inaccessible) content, it is not evident what the appropriate abstract or title for a dataset could be (Gregory, Groth, Cousijn, Scharnhorst, & Wyatt, 2019a). Keyword or topic search for documents is typically conducted to get a list of potentially relevant documents dealing with a concept that I am likely to know more about after having studied those documents, while applied to data it may only be meaningful as a concept being functionalized via additional context information such as time, location or population, as expressed in verbose queries such as 'average income tax of US citizens in 1970'. All this does

not imply that metadata resp. metadata search is obstructive in the case of data: it is definitely a first important step towards discovery of data through scientific discourse—or, as Borgmann et al. (2016) put it: 'Without metadata, datasets may devolve into spreadsheets of unlabeled rows and columns or into indecipherable strings of numbers.'

To overcome the unsettling vagueness of document search and to make data search more precise, thus data better findable, several approaches have been undertaken. Within the domain of social sciences, Stempfhuber and Zapilko (2009) suggested to integrate document and data retrieval within one subject portal. Their approach is based on the assumption that document queries can be mapped to data queries by means of ontologies. While this approach looks feasible in the context of a subject portal integrating both literature and data from a particular discipline, recent user evaluations have shown that querying data, at least within a data repository, is fairly different both from a conceptual and operational point of view (Groth, Koesten, Mayr, de Rijke, & Simperl, 2018). Finally, (meta)data search services such as Dataverse,[1] Datacite,[2] Zenodo,[3] Elsevier DataSearch[4] or Google Dataset Search[5] have pushed datasets to become first-class citizens of research output, and not just appendices of publications.

With respect to publishing and providing scientific data as Linked Open Data and RDF datasets, Kunze and Auer (2013) have sketched a way to query datasets on the basis of their VOID descriptions similar to metadata records describing literature content. Moreover, they suggested referring to single data by means of its URI patterns, hence a dataset containing such a pattern to be considered as 'relevant'. But taking the binary decision if a URI is contained in a RDF dataset might be too weak to adopt traditional information retrieval concepts such as relevance, precision and recall, since it does not refer to both the category of information needs and their fulfillment, which might be a gradual decision within the overall judgement of being relevant or not. For instance, a dataset could be more specific about weather (as a scientific concept and model), when it does not just contain or mention this keyword or a geolocation in combination with measured data, but also subconcepts or variables holding data on related concepts such as air pressure, temperature or humidity (cf. Figure 3). However, this approach addresses an important aspect of data search, namely checking data on the background of research questions directing to concrete data. For instance, if one wants to know about the average temperature in London in 2003, I might find

relevant datasets containing just that piece of information. In their approach 'Data Near Here', Megler and Maier (2015) developed a web application and interface for querying observatory data from oceanography, which is derived from community-specific data and practices. Because of the well-structured data and their intimate knowledge of this community, they could design and implement an exploratory search for the data matching the scientific use cases they are mostly familiar with.

Data search has undoubtedly become a popular topic: e.g. the SIGIR Workshop on DATA:SEARCH'18 identified requirements and challenges with respect to the concept and implementation of data search (Groth et al., 2018). Three recent studies suggest a more empirical, user-oriented approach towards data search and discovery (Chapman et al., 2019; Khalsa, Cotroneo, & Wu, 2018; Wu, Psomopoulos, Khalsa, & de Waard, 2019). Promising implementations and practices for data search can be mostly observed in research communities for the simple reason that these environments evolved long before any web-based search solutions and general public search engines were introduced. In order not to reinvent the wheel, pre-existing or established local community solutions for data search should always be considered as a reference and benchmark.

There has been quite a tradition of modelling information seeking behaviour, with popular models from, e.g. Kuhlthau (1991), Ellis and Haughan (1997). These models were focused on literature search, referring to information practices where internet or literature database search was only one approach and channel among others. In particular, the approach from Kuhlthau (1991) addressed the topic of psychological or emotional states of researchers or students, which are not the subject of this paper. Instead, we rather adopt their idea of identifying basic search patterns, with some significant differences we experienced while investigating research communities in the realm of web-based data search.

## 3. Conceptual Framework

We now introduce the basic concepts or modes associated with the more general and complex notion of 'data search'. This notion may involve at least three different aspects, which in our opinion are easily mixed together

while analysing, discussing or designing applications for data search. The distinction is intended to structure discussions and conceptions concerning data search.

Data search is basically threefold and comprises the aspects of *discovering*, *exploring* and *analysing data*. In a typical environment for scientific information infrastructure, discovery of data is mainly supported by system agents such as (meta)data providers and metadata aggregators or service providers (among them general web search engines). Discovery services support the detection of datasets by matching search queries to descriptive metadata and delivering a list of registered or published, mostly protected, datasets from different sources. Their focus is on the relation between datasets, not on their actual content in terms of data structures or values. They support users in detecting datasets labelled with 'weather' or 'finance', while these general concepts might differ a lot with respect to a dataset's content in terms of sub concepts, variable names, values or data formats. Typical services for making datasets discoverable are registering, identifying, citing, harvesting (both exporting and importing metadata), and recommending datasets resp. their metadata, and there have emerged a wide range of impressive infrastructure services supporting those activities (Datacite, Dataverse, Elsevier, DataSearch, Zenodo).

As complementary activity, exploring data literally means investigating a dataset to find out more about its content or structure similar to searching within a document. For instance, a researcher might be interested in finance datasets that contain data on prices and income. For these subconcepts to be matched against a search query, they might be part of explicit variable names or labels, or part of a data dictionary or related documentation (e.g. questionnaires, the results of which have been aggregated into a table). A typical service supporting the exploration of datasets is documenting in terms of specifying, structuring, clustering and annotating data (Newson, 2019)—tasks that have hardly been tackled on a generic level because of the diversity of documentation practices across disciplines and time periods.

Lastly, the concept of analysing data refers to the most familiar and 'traditional' user behaviour, since it aims at filtering or selecting data according to values specific to that data, as done for decades in typical database operations. For instance, as a researcher I might be interested in all locations within a certain area that have temperatures lower than zero degree in a certain period of time. For the purpose of information needs like these to

be fulfilled, the interface must provide an appropriate query language that matches the data, whereas visualising the results may help to analyse and 'understand' the data better. While discovery and exploration require the indexing of metadata and the provision of data scheme information in terms of e.g. codebooks, analysing effectively needs accessing and loading that data into some environment that permits querying and processing in particular large volumes of data.

It is important to note that our three patterns of data search (see Table 1) are strongly analytical and, from an operational point of view, heavily related to each other when it comes to searching data in a real-life situation. As Gregory, Groth, Cousijn, Scharnhorst, and Wyatt, (2019b) have pointed out, data discovery in particular might be viewed as a socio-technical practice that happens within a certain social environment with diverse peers and proxies, such as colleagues, literature retrieval, conference presentations, or interdisciplinary networking and training events. Hence, a general framework for data search referring to all these constituents might be conceived as a moving target that is constantly under development and change, especially with respect to different community practices and social settings. On the other hand, researchers from any discipline nowadays deal with a more or less web-based environment and infrastructure that strives to provide data in a web-friendly and, consequently, more standardised way. Finally, it might be disputed that 'analysing' data (including related operations such as processing data) is conceived as an integral part and activity of data search. One might argue that when a dataset is selected for further research and analysis, search has come to an end. On the other hand, analysing data is strongly related to exploring it—with the difference that it is more about operating, editing and (re-)publishing the data.

In the following, we will explain our three search patterns in terms of application design and real-world examples.

## 4. Discovering, Exploring and Analysing Data

### 4.1. Discovering Data

Although data is primarily numerical, users accustomed to general web search look for data via text-based queries containing keywords that denote

*Table 1: Patterns of data search.*

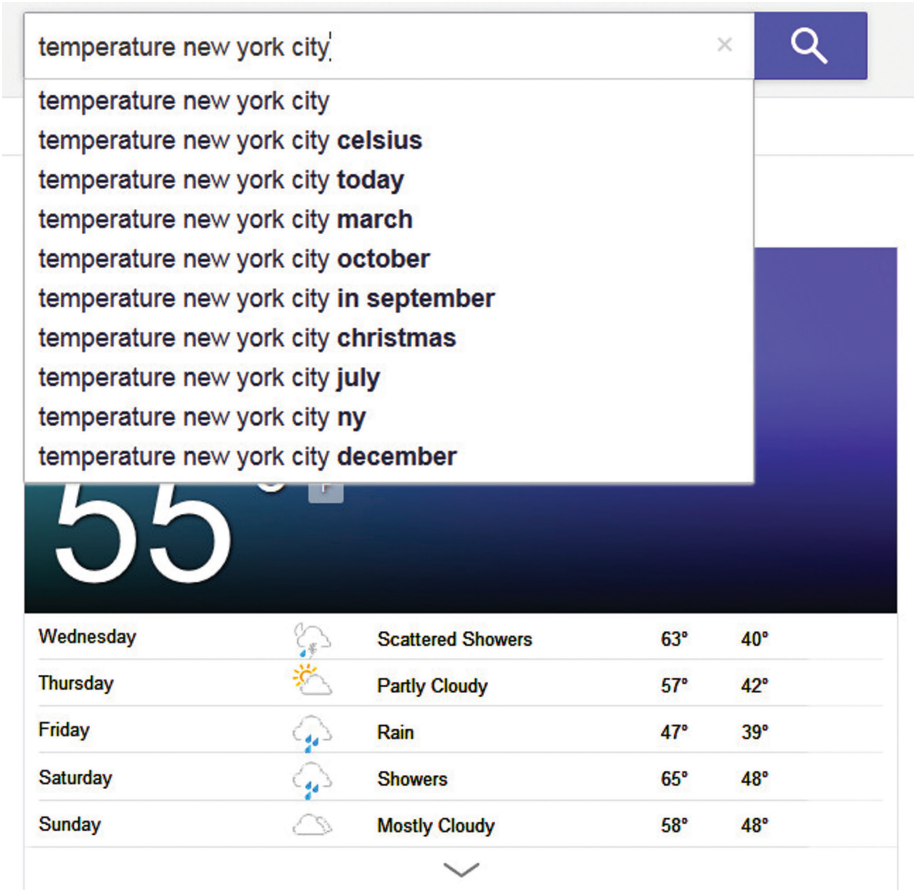
| Pattern | | Discover | | Explore | | Analyse |
|---|---|---|---|---|---|---|
| Description | | Serendipitous discovery of datasets on a general level by means of descriptive metadata (such as subjects or research method) | | Exploring content of a dataset (sometimes exposed as preview), which is summarised and accessible by means of dataset features | | Operating on a dataset including manipulating and reorganising data. Generating (and eventually publishing) derived datasets |
| Input/output | | Unstructured free text or predefined input/ list of (relevant) datasets | | Parameters with data values + ranges/list of datasets matching the input value | | Data values and parameters in combination with functions operating the (manipulated) dataset |
| Access to metadata and/or data? | | Metadata only | | Metadata + dataset features or summary in terms of a codebook or data dictionary | | Extensive access to any data and related provenance information including raw data a dataset is derived from |
| Access via? | | Search interface (apart from other proxies, such as peers or literature references) | | Search index incl. codebook information, interface for advanced search | | Local download of data/hosted data/API |
| Authentication required? | | No | | No | | Yes (depending on data provider's policy) |

the subject of interest. Search engines and discovery systems for datasets are normally run by repository systems or catalogue software that are strongly based on textual metadata residing in databases, files or HTML pages. Hence, users try very hard at querying data even more than they do in case of document search for the simple reason that they operate on the data much more than on a piece of literature, for example. Topic search would be a first step towards serendipitous discovery of datasets in particular from data providers and communities a user is not familiar with. By 'discovery' we mean a subject- or topic-oriented search on the level of descriptive metadata that results in a list of datasets a researcher might take into consideration for further exploration and/or analysis. We presume that a user first searches by a familiar concept, such as 'salinity', 'BSRN' or 'seagrass' (examples taken from indexed records from PANGAEA (n.d.) and Sea Around Us (n.d.)). Afterwards, the result is specified and filtered according to other metadata, such as origin, form, spatial and temporal indicators, measures, availability, trust or related literature (cf. metadata-based facets introduced in ICPSR (2019)).

Large data providers or metadata aggregators, such as Zenodo, Elsevier Data Search, or Google Dataset Search, support discovery and topic search just in the way they do document search: the query terms are matched with metadata, such as title or description of a dataset (Burgess & Noy, 2018). Querying a large search index such as Zenodo's (including research output other than datasets) delivers a list of search results that might be hard to filter, for descriptive metadata is not that reliable as it is in the case of literature search/discovery: datasets are often titled and described cryptically, while the data itself is often protected and inaccessible. Ironically, popular search engines such as Yahoo! or Google bypassed the issues with document indexing by identifying data-related user queries and rendering them differently from document results. Instead of providing a list of documents matching a query expression, Yahoo.com, for e.g., returns a table with figures and measures (see Figure 1).

Similar to other search engines, Yahoo.com extends the query by auto suggesting most used temporal indicators such as 'today' or 'March'. Indeed, this approach might work in an environment of a general web search engine, with considerable queries also from non-scientific user groups. But more academic search engines and discovery systems might still fall short of collecting and aggregating their usage data in that massive way, as Burgess and

*Fig. 1: Search suggestions for 'temperature New York City'.*



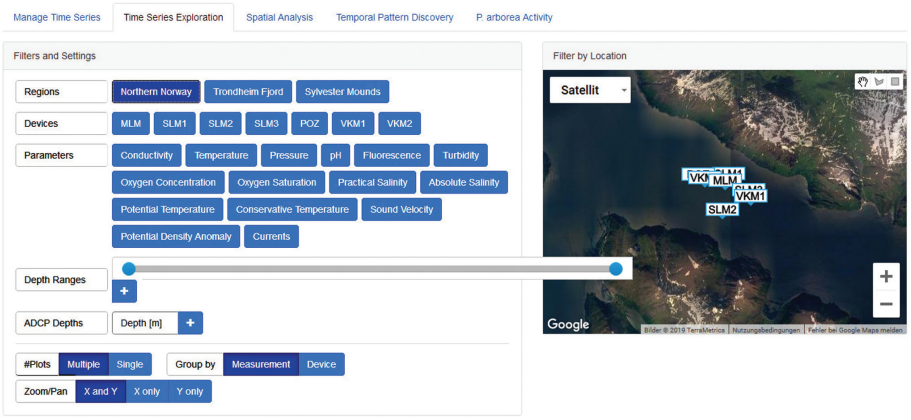

Noy (2018) pointed out. As a result, they could support querying of datasets across different research communities and user groups in two ways: either by supporting topic search for generic concepts, including mapping to more community-specific language, or by providing dedicated interfaces supporting spatial and temporal discovery of datasets.

One central topic of data discovery is the balance between generic and disciplinary search terms. In our analysis of information behaviour across different communities, we identified the gap between the generic topics that

users start the discovery process with, and the community-specific (sub)topics they typically rely mostly on to complete this process (final filtering and selection of results). Thus, when dealing with heterogeneous research communities, we need to consider both terminology 'breadth' and 'depth': besides the generic (descriptive metadata) terms, communities also want to be able to search by more discipline-specific terms. For instance, in social sciences there are dataset features especially useful for researchers to express their information needs. To researchers from this domain, metadata that describe the disciplinary aspects of a dataset, such as questionnaire design, topics covered, survey sample selection, etc., are instrumental in using a dataset. Every dataset that the data provider collects is described by a set of variables, and each variable contains, among others, information about the concept it refers to, and the topic that subsumes it. For example, a user prefers searching with a label 'auspan', which represents the concept of 'year move abroad', and is part of the topic of 'emigration'. All this is common terminology for researchers from this domain and, in addition to the generic terms used, it represents a key capability to localise datasets of interest.

Enabling users to search across disciplines is a basic goal for aggregators and search service providers. One possible approach is to refer to subject classifications for categorising datasets according to scientific (sub) disciplines that are suggested and maintained by the communities. Another approach would be fostering the discovery of datasets according to their spatial and temporal properties. Despite their heterogeneity, inconsistency and incompleteness across different research communities and their corresponding repositories, around 73% of metadata records, e.g., from the PANGAEA repository hold information on spatial and/or temporal information as two basic characteristics of data (Takeuchi et al., 2017). According to DataCite's metadata schema 4.3, geolocation has become a recommended metadata element with the properties for specifying both the longitude/latitude and the label of a geolocation. Some data repositories even operate specific APIs for querying geolocation related data, or provide GUI elements for locating a map region by drawing rectangles. These services might be supplemented with the option to name a region or geo-political entity in the sense of geocoding, such as 'European Community', 'Northern Norway' (in the example above) or 'City of London'. Hence labels, in particular for geo-political entities, may generally help to identify datasets across domains, but they are also crucial for referring to data within the scientific discourse (Figure 2).

*Fig. 2: GUI with labels for selecting a map area (example from http://oceantea.uni-kiel.de/).*




A final approach – that of semantic search – is to leverage structured descriptions of the terms that communities use to label their datasets. The aim is to go beyond a search query keyword, and handle cases such as synonyms, narrower/broader/related/associated, or otherwise relevant concepts to that term. One such example is the annotation of datasets from the ecology domain based on an ontology that improves the search experience, added as a feature to the Metacat metadata system (Berkley, Bowers, Jones, Madlin, & Schildhauer, 2009). Including domain knowledge in any case improves the overall search experience. Thus, this approach could also be combined with other approaches, where more semantic structure is required or could further assist the search process. One such combination could merge spatial and temporal properties with semantics as a means to improve the keyword-based search approach. A case of employing a geo-spatial taxonomy to structure the (geo-spatial) metadata terms of a geo-referenced dataset shows improved results over the keyword-based approach (Apache Lucene search engine implementation) (Li, Goodchild, & Raskin, 2014).

### 4.2. Exploring and Analysing Data

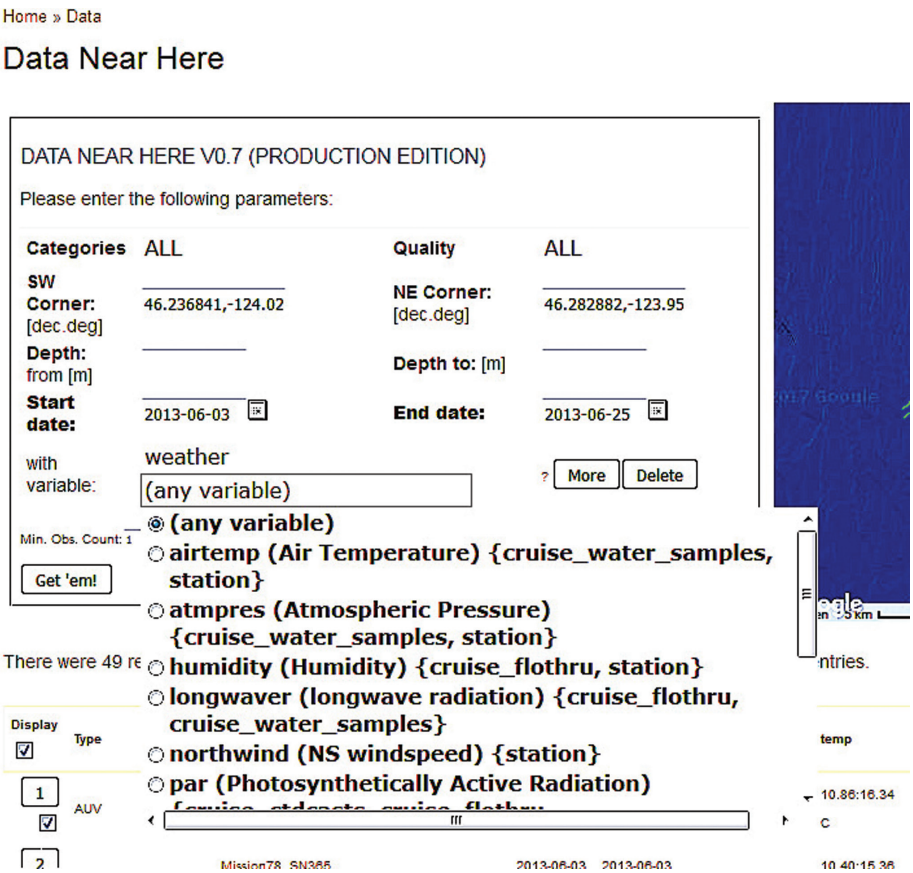
While discovery of data operates on the level of descriptive metadata of different, potentially related, datasets, exploration happens within a certain dataset that a researcher finally has selected to consider. In the pre-web era,

exploring data required domain-specific environment and knowledge, such as an interface for parametrising a query, or knowledge about the structure of the data. In particular within scientific disciplines, such as marine sciences, hydrology, climatology or oceanography, we found some elaborated applications for exploring data that has been transferred to the web. For instance, (Megler & Maier, 2015) demonstrate the requirements, processing steps and evaluations on user interactions and benefits with respect to a database application and interface suitable for research purposes. Taking into account the specific requirements from researchers from oceanography as user stories, they developed a database and interface for querying labels for concepts (e.g., 'weather') and associated variables (e.g., 'humidity' or 'temperature') (Figure 3).

While this rather elaborated approach requires a full-fledged database application with a search index accessible through the web, there is another option to publish datasets in a more light-weight format—a subset or preview -, in particular if the data is not supposed to be accessible outside a community's infrastructure. For this purpose, a dataset's characteristics could be summarised by means of a data dictionary, data scheme or codebook, informing the user about documented or generated characteristics, such as the label, data type or range of values of a variable.

Although some data centres and repositories already provide a web-compliant version of (a subset of) their data, including interfaces for exploring the content, this is still not at the very core of data driven research practices requiring specific infrastructure and software in terms of access, run-time, scalability and performance. In particular, large datasets with, e.g., petabytes of data are simply not eligible for being transferred from a data repository to, e.g., a user's desktop. While discovery and exploration of (meta-) data can be conducted off-site and with limited access to a subset of data, analysing datasets requires full access to both published and raw data, the latter sometimes even more required (Halevy et al., 2016). Data analytics and processing includes operations such as wrangling, curating and transforming the data, often from original material. And if analytics is conducted by means of visualisations – that become even more important as data increases in size – it needs to be plotted by special programming functions and libraries that are sometimes still beyond common techniques for web presentation.

*Fig. 3: Interface for data exploration provided by DNH project.*




## 5. Search Modes: Empirical Evidence from GeRDI

The GeRDI project (GeRDI, n.d.) provides a research infrastructure that targets long tail data across research communities and disciplines, and with it come different research data management practices and requirements. During the requirements gathering process, we mainly focused on services that an infrastructure like GeRDI should provide. However, one service in particular – search – enabled us to further explore the patterns manifested in the user stories. In this context, we identified different patterns that users from these communities prefer to see implemented in GeRDI, although there wasn't a common

understanding of or expectation from these patterns among its communities. The definitions of search modes presented in this paper help us (including the users) understand and classify the search-related requirements in GeRDI.

In the context of search patterns, generally, communities in GeRDI seem to clearly distinguish between the first two modes of search, i.e., discovery and exploration; in their view, analysis often seems quite close, if not part of exploration. Metadata- vs. data-level operations seems to be a pragmatic rule that helps these communities distinguishing between search patterns. To them, operating solely on metadata during search fits the scope of discovery mode, whereas operating on data for search fits the scope of exploration mode. Finally, the cases where they want to reuse or otherwise include the raw data of dataset in their workflows fit the scope of analysis mode.

The list below shows the GeRDI pilot communities based on which we based our empirical part. In some cases, there are multiple communities associated with a single (sub) domain, such is the case with the EREE communities, but we kept only the main communities for brevity:

- Alpine Environmental Data Analysis Center (AlpEnDAC)
- Microscopy and Bioinformatics (CBG)
- Digital Humanities (CRANE)
- Hydrology and River Basin Management (HFM)
- National Center for Tumor Diseases (NCT)
- Socio-Economic Panel (SOEP—DIW)
- UN International Strategy for Disaster Reduction and Environmental Computing (UNISDR)
- Digital geo-linguistics (Verba Alpina)
- Environmental, Resource and Ecological Economics (EREE)

### 5.1. Discovery in GeRDI

Regardless of the metadata granularity required to support it (GeRDI services operate on metadata), all communities in GeRDI need a service that provides

a basic mechanism for search. In terms of the search mode definitions, we encounter discovery with all GeRDI communities, and the role it plays differs depending on the community and/or domain. This mode is especially important for the scientific domains that work with resources such as text or data that is commonly represented via textual descriptions; Verb Alpina, CRANE, or SOEP are such examples. Communities from other disciplines, such as PALOZ (paleoceanography) or EREE, also rely on discovery mode, but they see it as part of a two-step search process. Since they have portals for more discipline-specific search, the discovery service in GeRDI provides them with an initial search functionality for users to assess whether to proceed towards the specialised portals that support discovery at a deeper, disciplinary level, and/or provide additional modes of search, such as exploration and analysis. To these communities, in addition to the search functionality, GeRDI also brings additional visibility to their dataset collections.

Having in mind the different requirements that communities in GeRDI have, discovery is not supported for all of them at the same level. While common descriptive metadata suffice to support discovery in one community, such as search based on geo-referencing (or location) or subject terms, there are communities that rely on disciplinary metadata elements during discovery search. For example, a community from the genetics domain wants to rely on a 'species' metadata element, whereas a community from social sciences relies on metadata such as 'variables' and 'concepts' of those variables to search corresponding data collections. Species, variables, and concepts are all metadata specific to these communities.

As part of this search pattern, we also explored the potential application of semantic search. The requirements-gathering process identified few knowledge organisation systems that research communities use, such as thesauri, vocabularies, and subject headings, but the domain heterogeneity prohibited us from adopting them at the metadata level across domains. In any case, we wanted to move towards more semantic-like services, and we chose metadata normalisation to lay the foundations for this. Namely, we analysed the harvested metadata of some of the common attributes, such as language, geolocation, and subject terms, and we wanted to assess the extent to which we can control their value range (thus their (value) diversity). Take the language attribute as an example: we were able to identify rules that map terms denoting the same language to a single (central) term, thus enable users to search in any of the language varieties present in the index, regardless of how it

was described by the contributing community. This normalisation approach is able to reduce 70+ language terms to 30. A semantic structure to describe languages, such as ontology, for example, accepted across communities, would be the way to proceed towards another "semantic upgrade". Other metadata elements in our collection did not exhibit much commonalities or (semantic) structure to rely on for a similar normalisation effort. For example, the subject terms present a high degree of variety, which is very challenging to describe or structure with KOSs across different domains. In this way, any attempts to search for resources based on subject across communities, while of great interest, currently proves difficult to reach.

Regardless of the level of search supported, discovery mode remains a common, required search scenario across the different pilot communities in GeRDI. In any case, in order to support this search mode, all the referred metadata elements need to be part of GeRDI schema, indexed and searchable.

### 5.2. Exploration and Analysis in GeRDI

As expressed during the requirements gathering process in GeRDI, whenever numerical data are considered, such as in the natural science domains (hydrology, alpine research, climate research, microscopy and bio-informatics, etc.), exploration and analysis search modes are required for a more complete search experience for the users. For example, users would want to retrieve datasets that fall in a certain range of values – be it temperature, location, or any other type of range – for a metadata field. Moreover, often, a preview of data is deemed important to get a first impression on unknown data sets. In this case, users prefer (via a service implementing the search) more access to and knowledge about datasets before engaging with one, which fits the typical case of an exploration mode. Another requirement from this category pertains to visualisation of time series to see if a dataset is relevant to the user. In this case, the relevance check is based on the value aggregations to provide a range of values for a (set of) metadata element(s) the user is interested in, as well as (community-specific) quality level assessment outcomes.

Despite their search requirements for discovery, exploration, and analysis, GeRDI communities do not always provide enough metadata to support them. Based on the metadata available for harvest, the only service that can be offered to all communities is that of discovery. As a result, due to lack

of metadata and supporting features, the search modes of exploration and analysis are not yet represented in the GeRDI search service. The search implementation in GeRDI followed an incremental development approach, starting with the discovery mode as a base service, and laying the grounds to evolve towards the exploration and analysis ones in the future releases.

## 6. Summary and Outlook

With respect to the evolving landscape of a research data infrastructure being increasingly exposed and made accessible through the web, concepts and best practices for providing and searching data on the web become an urgent topic. While the web fosters globally connected and trans-disciplinary research just by exposing, indexing and providing research output, approaches to data search are still built on traditional practices such as text-based search of metadata representing a dataset. As we have pointed out, we still regard some of those metadata as crucial for discovery and exploration of datasets. But at the same time we suggest those concepts to be mapped to operating data elements such as geolocation and data labels. Basically, we identified three modes of data search that might serve as a generic framework for identifying and categorising data search practices across different communities, as we did in the GeRDI project.

Concerning future work, we are just about aggregating, identifying and mapping harvested subject terms from different research communities and repositories. Moreover, in the context of the ongoing GeRDI project, we are developing a technical infrastructure for distributed research data management including services for particularly discovering and analysing datasets from different communities. While the discovery of datasets includes normalisation and filtering of their metadata, their analysis requires an environment for processing the data that is to be assisted by services for bookmarking and storing the datasets.

## Acknowledgments

The authors gratefully acknowledge financial support from the GeRDI project, funded by German Research Foundation (DFG), grants no. BO818/16-1 and HA2038/6-1.

## References

Berkley, C., Bowers, S., Jones, M. B., Madin, J. S., & Schildhauer, M. (2009). Improving data discovery for metadata repositories through semantic search. In L. Barolli, F. Xhafa, & H. Hui-Huang (Eds.), *CISIS 2009, International Conference on Complex, Intelligent and Software Intensive Systems* (pp. 1152–1159). New York: IEEE Explore Digital Library. https://doi.org/10.1109/CISIS.2009.122.

Borgman, C. L., Golshan, M. S., Sands, A. E., Wallis, J. C., Cummings, R. L., Darch, P. T., & Randles, B. M. (2016). Data management in the long tail: Science, software, and service. *International Journal of Digital Curation*, *11*, 128–149. https://doi.org/10.2218/ijdc.v11i1.428.

Burgess, M., & Noy, N. (2018). *Building Google dataset search and fostering an open data ecosystem.* [Blog]. Retrieved February 14, 2020, from https://ai.googleblog.com/2018/09/building-google-dataset-search-and.html.

Chapman, A., Simperl, E., Koesten, L. Konstantinidis, G., Ibáñez-Gonzalez, L., Kacprzak, E., & Groth, P. (2019). *Dataset search: A survey*. Retrieved November 13, 2019, from https://arxiv.org/abs/1901.00735.

Ellis, D., & Haughan, M. (1997). Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of Documentation*, *53*(4), 384–403. https://doi.org/10.1108/EUM0000000007204.

GeRDI. *Generic Research Data Infrastructure*. (n.d.). Retrieved November 13, 2019, from https://www.gerdi-project.eu/.

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019a). Searching data: A review of observational data retrieval practices in selected disciplines. *Journal of the Association for Information Science and Technology*, *70*(5), 2019, 419–432. https://doi.org/10.1002/asi.24165.

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019b). Understanding Data Search as a Socio-technical Practice. https://arxiv.org/abs/1801.04971v3.

Groth, P., Koesten, L., Mayr, P., de Rijke, M., & Simperl, E. (2018). DATA:SEARCH'18 – Searching data on the web. In L. Dietz, L. Koesten, & S. Verberne (Eds.), *SIGIR2018 Workshops: ProfS, KG4IR, and DATA:SEARCH* (pp. 65–73). Retrieved Febriary 11, 2020, from http://ceur-ws.org/Vol-2127/preface-datasearch.pdf.

Halevy, A., Korn, F., Noy, N. F., Olston, C., Polyzotis, N., Roy, S., & Whang, S. E. (2016). Goods: Organizing Google's datasets. In *Proceedings of the 2016 International Conference on Management of Data* (pp. 795–806). New York: ACM. https://doi.org/10.1145/2882903.2903730.

Hirschheim, R., & Klein, H. K. (2012). A glorious and not-so-short history of the information systems field. *Journal of the Association for Information Systems, 13*(4), 188–235. http://doi.org/10.17705/1jais.00294.

ICPSR. *Find & analyze data*. (2019). Retrieved November 13, 2019, from https://www.icpsr.umich.edu/icpsrweb/ICPSR/search/studies.

Kern, D., & Mathiak, B. (2015). Are there any differences in data set retrieval compared to well-known literature retrieval? In S. Kapidakis, C. Mazurek, & M. Werla (Eds.), *Research and advanced technology for digital libraries* (pp. 197–208). Cham, CH: Springer International Publishing. https://doi.org/10.1007/978-3-319-24592-8_15.

Khalsa, S., Cotroneo, P., & Wu, M. (2018). A survey of current practices in data search services. https://doi.org/10.17632/7j43z6n22z.1.

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, *42*(5), 361–371. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-%23.

Kunze, S. R., & Auer, S. (2013). Dataset retrieval. *Proceedings of the 2013 IEEE Seventh International Conference on Semantic Computing* (pp. 1–8). New York: IEEE Computer Society. https://doi.org/10.1109/ICSC.2013.12.

Lewandowski, D. (2015). *Ranking library materials*. Retrieved November 13, 2019, from https://arxiv.org/abs/1511.05806.

Li, W., Goodchild, M. F., & Raskin, R. (2014). Towards geospatial semantic search: Exploiting latent semantic relations in geospatial data. *International Journal of Digital Earth*, *7*(1), 17–37. https://doi.org/10.1080/17538947.2012.674561.

Megler, V. M., & Maier, D. (2015). Demonstrating 'Data near here': Scientific data search. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, (pp. 1075–1080). New York: ACM. https://doi.org/10.1145/2723372.2735360.

Newson, K. (2019). *Introducing the data curation tool.* Retrieved February 11, 2020, from https://spotdocs.scholarsportal.info/display/DAT/2019/09/26/Introducing+the+Data+Curation+Tool.

PANGAEA. *Data publisher for earth & environmental science*. (n.d.). Retrieved November 13, 2019, from https://pangaea.de/.

Sea Around Us. *Fisheries, ecosystems and biodiversity*. (n.d.). Retrieved November 13, 2019, from http://www.seaaroundus.org/.

Stempfhuber, M., & Zapilko, B. (2009). Integrated retrieval of research data and publications in digital libraries. In S. Mornati & T. Hedlund (Eds.), *Rethinking electronic publishing: Innovations in communication paradigms. Proceedings of the13th International Conference on Electronic Publishing (ELPUB 2009)* (pp. 613–620). Roma: Nuovo cultura https://dblp.org/rec/conf/elpub/StempfhuberZ09.

Takeuchi, S., Sugiura, K., Akahoshi, Y., & Zettsu, K. (2017). Spatio-temporal pseudo relevance feedback for scientific data retrieval. *IEEJ Transactions on Electrical and Electronic Engineering*, *12*(1), 124–131. https://doi.org/10.1002/tee.22352.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018, n.p. https://doi.org/10.1038/sdata.2016.18.

Wu, M., Psomopoulos, F., Khalsa, S. J., & de Waard, A. (2019). Data discovery paradigms: User requirements and recommendations for data repositories. *Data Science Journal*, *18*(1), 3. https://doi.org/10.5334/dsj-2019-003.

## Notes

[1] https://dataverse.org/home.

[2] https://search.datacite.org/.

[3] https://zenodo.org/.

[4] https://datasearch.elsevier.com.

[5] https://toolbox.google.com/datasetsearch.