HERE IS MY QUERY, WHERE ARE MY RESULTS? A SEARCH LOG ANALYSIS OF THE EOWEB® GEOPORTAL

Sirko Schindler, *Marcus Paradies*

German Aerospace Center DLR
Institute of Data Science
Jena, Germany
{sirko.schindler,marcus.paradies}@dlr.de

André Twele

German Aerospace Center DLR
German Remote Sensing Data Center
Oberpfaffenhofen, Germany
andre.twele@dlr.de

ABSTRACT

With the rapid growth of available earth observation data and the rising demand to offer web-based data portals, there is a growing need to offer powerful search capabilities to efficiently locate the data products of interest. Many such web-based data portals have been developed with vastly different search interfaces and capabilities. Up to now, there is no general consensus within the community how such a search interface should look like nor exists a detailed analysis of the user's search behavior when interacting with such a data portal.

In this paper we present a detailed analysis of user's search behavior based on a log analysis of a real earth observation data portal and generalize our findings to recommendations for future data portal search frontends to improve the overall user experience and increase the search quality.

Index Terms— Geoportal, Search, Query Log Mining

1 Introduction

Earth Observation (EO) data is growing rapidly in volume and increasingly scientific and commercial users demand efficient and intuitive access to value-added EO data products. EO geoportals offer such functionality by providing advanced search capabilities over the archived data. With the increasing diversification of potential user groups and different levels of EO domain knowledge, this poses a tremendous challenge to offer an intuitive yet powerful search interface that can be successfully operated by multiple user groups. Missing domain knowledge or different vocabularies often lead to underspecified queries (too many results) or unsuccessful searches (no/wrong results).

To better serve user groups with different levels of domain knowledge and experience it is essential to better understand the user's search behavior based on an analysis of a real-world system with real user queries. Real EO data portals typically log all incoming search requests, effectively providing value information about the most commonly used search keywords and additional temporal and geospatial constraints.

In the past, log data from NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC)[1] has already been analysed [3]. However, the authors focused on the technical aspects of processing large amounts of heterogeneous log data and provide only little information about the results of their analysis. Regarding the constraints used by their users, only a list of top ten keywords and their frequency is given.

In this paper we give a detailed analysis over 6 months (April–October 2018) of log data from DLR's EOWeb GeoPortal.[2] We summarize our findings and provide practical guidelines for the development or enhancement of the search interface of EO data portals.

The remainder of the paper is structured as follows. In Section 2 we introduce the EO data portal that was used for the analysis. In Section 3 we describe our analysis setup before we present our log analysis in Section 4. We summarize our findings and sketch specific search enhancement possibilities in Section 5, before we conclude the paper in Section 6.

2 EOWeb GeoPortal

The German Satellite Data Archive (D-SDA) consists of a large collection of Earth Observation (EO) data from both national and international missions maintained by the German Aerospace Center (DLR). The EOWeb GeoPortal (EGP) has been developed as a multi-mission web portal for accessing the heterogeneous data sources of the D-SDA [4]. It provides access via a set of services compliant with the standards of the Open Geospatial Consortium (OGC) as well as an open web interface that allows users to query the archive for its products and services. Users can express their information need using multiple constraints:

Collections: They can achieve a catalog-like browsing of datasets by restricting their search to single or multiple of *collections*, e.g., spotlight images from the TerraSAR-X mission. In addition, collections are ordered in a hierarchical fashion, so users may directly select all TerraSAR-X collections without the need to iterate through all of them manually.

Geospatial: Users can restrict the *spatial* extent of their

[1] https://podaac.jpl.nasa.gov

[2] https://geoservice.dlr.de/egp/

search via a map interface to a specific region of the world. They can either draw a bounding box within the map, upload an own area of interest as a Shape or KML-file, or select one from a predefined list of regions. The list covers most countries in the world as well as selected regions like central Africa.

Temporal: The third option is to restrict the search by a *time* period, which reflects the acquisition time of the satellite scene. Similar to spatial restrictions, a number of predefined values can be selected. Besides including generic time intervals like “last week”, this also allows setting the time frame to the life time of a specific mission like SRTM.

Keyword: EGP allows the specification of *keywords* matching datasets’ content. Predefined keywords can guide novice users through the portal and may provide experienced ones with shortcuts in their workflow. The offered options include keywords derived from thesauri such as the INSPIRE Spatial Data Themes (e.g., “Atmospheric conditions”), uniform resource names (URNs), such as, e.g., “urn:eop:DLR:EOWEB:GOME.TC”, and mission related terms like “MERIS”. The keyword search is modeled as a full-text search over the collection metadata provided by a OGC-compliant CSW (Catalog Service for the Web) interface.

Type: Users can also restrict their search to a specific type of result. EGP offers not only EO collections, so users can focus their search on either datasets, dataset series, or services.

The search result for EO collections can be further limited through additional filter criteria, which are derived from the product metadata. For example, this allows restricting the search to a cloud coverage of less than 20% in case of optical satellite data or a HH-polarization in case of Synthetic Aperture Radar (SAR) data.

After users identified the products of interest, they can directly order them through the EGP interface. For some products there are access restrictions in place, but most of the products are freely accessible after registration and can be downloaded or retrieved via one of the OGC-compliant services (e.g., WMS, WFS, WMTS, and WCS).

3 Methodology

We performed an offline analysis of the EGP log files for the time period between April and October 2018 of the EGP as the main source of information about current users’ requirements. Before the actual analysis, the log files were stripped of any personal identifying information. They were then parsed and stored using an Elastic stack[3] pipeline. In this process each log entry was classified and deconstructed into its components like identifying spatial constraints used or an anonymous session-id to connect different requests of a single search session. This preprocessing allowed us an easy and efficient access to the various aspects of the query log.

Neither ELASTICSEARCH[4] nor KIBANA[5] supports the full extent of our analysis out of the box. Nonetheless, we tried to automate most of the data extraction tasks using manually created scripts and queries.

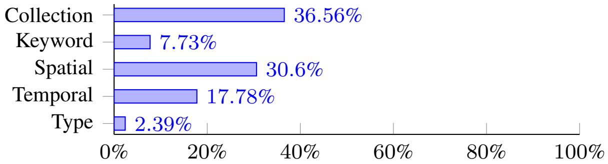

Fig. 1. Frequency of queries using at least the given constraint.

Keyword constraints were decomposed into used concepts. Note at this point the difference between terms, which are most commonly used in query log analysis, and concepts. A concept may consist of a single term, but can also contain an n-gram of terms, e.g., “digital elevation model”. While the individual terms refer to rather general concepts, in conjunction this concept denotes a specific type of value-added product.

Individual concepts were subsequently categorized using information extracted from publicly available resources. In particular, we made use of Wikidata [7], as it covered a wide range of appearing concepts. Further inspection revealed several uncategorized concepts. To increase coverage we employed stemming techniques as well as manually curated mapping files to mitigate the impact of typos and similar mistakes. Uncategorized concepts after this step contain single letters and other unidentifiable sequences of either letters or numbers. We collected those in a separate category “unknown”.

The use of predefined values is not tracked separately in the log files. We attempt to gauge their usage by comparing the respective restrictions with the set of predefined values. Although users may enter the exact value directly, we believe this approach to be sufficiently precise for both keyword and spatial constraints. However, it is not applicable to temporal constraints, as here the generic options like “last week” will not translate to fixed values usable for comparison.

4 Patterns of Use

In our analysis we followed established approaches in query log mining. For an overview we refer the reader to [5]. Unless otherwise noted, the following results refer to initial search requests. Requests that arise from traversing the different result pages are excluded.

In addition to common keyword-based queries, EGP allows users to apply other constraints like spatial or temporal restrictions (cf. Section 2). As shown in Figure 1 almost 31% of all queries contain at least a spatial restriction and 18% a temporal one. On the other side only 8% of queries use keywords, while type restrictions contribute to only 2% overall. The high frequency of collection-based queries is predominantly caused by the order process that requires users to browse collections before ordering.

The use of predefined constraints varies. They are barely used for spatial restrictions: only 0.64% use predefined values. On the other hand, about 21.5% of keyword restrictions make

[3] https://www.elastic.co

[4] https://www.elastic.co/products/elasticsearch

[5] https://www.elastic.co/products/kibana

TerraSAR-X	FIREBIRD
*DLR	*Land Cover
SRTM	Elevation
TanDEM	Terra
DEM	*Climatology, meteorology, atmosphere

Table 1. Most frequent concepts used (* - predefined option).

use of the offered options. Note, that the predefined term “DLR” accounts for more than half of those.

Our main focus then shifted to the keywords used. Prior work [6] establishes rather low numbers for terms used per query. They give an average number of terms per query at about 2.4. For EGP we observed on average 1.02 concepts per query with a standard deviation of 0.17[6]. Also the distribution of terms is highly skewed: Only 1% of the unique concepts contribute to over 25% of the keyword-restricted requests. A list of the ten most frequent concepts is given in Table 1. Kindly recall, that we refer to concepts at this point.

As mentioned in Section 3 we had to cope with various abbreviations and different spellings for some concepts. The concept for TerraSAR-X mission was, e.g., labeled using the following terms (omitting several variations of upper-/lowercase): “TSX”, “TerraSAR-X”, “Terrasar--X”, “TerraSar x”, “TerraSAR”, or “Terra SAR”.

A more comprehensive impression of keyword usage can be gathered from the classes of concepts used. An overview of their respective frequencies is given in Figure 2. The dominant classes relate to the initial gathering of the data with slightly over half of the concepts used referring to specific missions. Furthermore, users looked for instruments (about 4.3%) or specific observational parameters of them (around 1.5%). Also part of these provenance-related classes are organizations (about 12%)[7] and types of products (about 9.6%).

Some users chose to search for applications data products can be used for (around 10.7%). Setting aside the predefined suggestions, the most frequent concept here is “elevation”, followed by “snow”, “flood”, and “water”.

Notable is also the share of location-related information entered as a keyword (about 7.5%). These concepts identify regions of varying size reaching from generic ones like “world” or “global”, over countries and areas like “Romania” or “Baltic” to specific locations like “Moscow” or “Rome”. Most location are referenced by their English name. However, there are some exceptions like the Polish “Warszawa” for Warsaw or the German “Kroatien” for Croatia.

A few users use the keyword field to enter coordinates directly (below 0.5%). Here we observed values like “51.75602 / 14.31971” pointing to a location in Cottbus, Germany, or “N39E068” near the border of Uzbekistan and Tajikistan.

The remainder of concepts includes system-specific IDs (around 2%) – presumably obtained in previous sessions – and rather general terms (around 1%). The later consists of terms like “collection” or “data formats”, which seem to indicate an information need that does not aim at a single data product. Finally, there is a number of concepts we could not assign to a specific class: fragments of terms and arbitrary numbers apparently not referring to any coordinate.

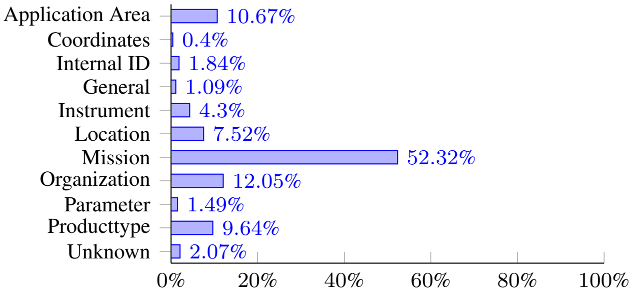

Fig. 2. Frequency of concept classes.

As mentioned before, the search logs showed quite some variation in the keywords used to describe a specific concept. In an ideal world, all those variations would lead to the same result set. However, for many variations we observe substantial differences. One example are the keywords “Ozone” and “O3”. Both denote the same concept, but the former returns 31 results, while the latter one only matches 5.

5 Observations and Directions

The analysis of the EGP log files offered some interesting insights: First, although the predefined spatial restrictions are barely used, users seem to prefer the keyword input for the same purpose. Here, we observe a substantial amount of keywords relating to location or region names.

We can imagine different possible reasons. Independent of the actual techniques used to satisfy user requests, most popular search engines offer a single keyword input field as the default way of interaction. Users familiar with those interfaces might transfer that usage pattern to EGP and describe their information need primarily using keywords.

Another reason might be caused by the current user interface design. Predefined options for spatial restrictions are not available on the default interface itself, but need to be accessed via the “Advanced Map” menu. Users new to the system might not notice that and, hence, resort to the keyword input field.

The final possible reason concerns the selection of predefined options. While those options mostly define the scale of countries or other large regions, many location keywords refer to much smaller areas like specific cities. So users might be lacking the options there to express their search intent.

Most of the keyword-based location queries have no suitable result, as the metadata information does not include the respective terms. This problem could be tackled by adopting databases like OSMNames[8]. They offer a wide selection of geographical entities and their spatial extent. Transparently

[6] Terms per query was slightly higher at an average of 1.20 terms per query with a standard deviation of 0.60.

[7] Note that this class is largely dominated by the concept “DLR”, which is also part of the suggested keywords.

[8] https://osmnames.org

translating from location-keywords to a spatial constraint could significantly improve the quality of search results here.

Another side-effect of using such databases is the support for different languages. Metadata information is usually restricted to a rather low number of languages. However, OSMNames and similar collections offer labels in a wide variety of languages. In particular, OSMNames uses OpenStreetMap [2] as a data source, which has crowdsourced the collection of geographic data including the names of locations. The result is a steady influx of updated data from a multilingual community.

Similar problems arise in other keyword classes as well. The backend currently employs a full-text search engine over the collection metadata. A collection needs to include the exact term as entered by users to appear in the results. This may fail for several reasons, thus preventing users from discovering the datasets they need. The examination of keywords found in the search logs suggests three different categories of reasons:

Typographical Issues: A first category is given by mere typos of different degree. It includes the inconsistent use of space, dash, and similar characters, as well as common misspellings. As example we refer to the different variations denoting the concept TerraSAR-X as mentioned before.

Abbreviation Issues: A second category consists of synonyms and similar relations. Besides traditional synonyms and abbreviations, we also include the various representations in different languages here. The aforementioned pair "O3" and "Ozone" is an example for this category.

Semantic Issues: The final category is comprised of semantically related terms. Metadata authors and end-users oftentimes have different backgrounds and, hence, use different terminologies. This results in a semantic gap that prevents users unfamiliar with the specific vocabulary used in the metadata from finding appropriate datasets. An example here is "height", which probably refers to the concept "elevation" as used throughout the metadata descriptions.

While all these categories will deteriorate a users search experience, there are different techniques to mitigate them. Misspellings can generally be addressed by the use of string similarity measures like Levenshtein distance or stemming/-lemmatization approaches. Similar to the aforementioned resolution of location-related terms, codelists can also counteract the effect of abbreviations by expanding the respective terms, so they concur with the usage within the metadata.

The most challenging category are semantically related terms. A brute-force approach using codelists will soon reach its limits given the vast amounts of terms and relations as well as the effort needed to maintain it. The Semantic Web [1] uses a graph-based knowledge base connecting terms using various relations. It promises to bridge the semantic gap between content creators and consumers beyond the capabilities of traditional search engines.

Beyond the aforementioned keyword-focused aspects, we recognize that other techniques can also improve users' search experience. This includes, but is not limited to using visualizations to represent the results, providing support for explorative search strategies, or recommender engines that are based on users' past interactions. However, we consider their discussion too broad and, hence, out of scope for this paper.

6 Conclusion

EO data grows rapidly in both size and topics addressed. With an increasingly broad range of possible usecases, geoportals serve as the entry point for a diverse group of users coming from a wide range of domains.

As a first step to cater to this expanded audience that might lack knowledge of terms and procedures used in the EO community, in this paper we analyzed the current user behavior in the EOWeb GeoPortal. Based on an analysis of the log files we described different usage patterns and highlighted existing issues with a focus on keyword-based queries. We outlined possible strategies to mitigate those issues and increase user satisfaction and efficiency at finding suitable EO products.

7 Acknowledgments

We would like to thank the EOWeb GeoPortal team for providing us with the log files and their ongoing support in this analysis. We especially thank Gina Campuzano Ortiz, Katrin Molch, Stephan Kiemle, and Daniele Dietrich for their valuable input and technical support.

REFERENCES

[1] T. Berners-Lee et al. The semantic web. *Scientific American*, 2001. doi: 10.1038/scientificamerican0501-34.

[2] M. Haklay and P. Weber. OpenStreetMap: User-generated street maps. *IEEE Pervasive Computing*, 2008. doi: 10.1109/mprv.2008.80.

[3] Y. Jiang et al. Reconstructing sessions from data discovery and access logs to build a semantic knowledge base for improving data discovery. *ISPRS International Journal of Geo-Information*, 2016. doi: 10.3390/ijgi5050054.

[4] H. Rotzoll et al. From Discovery to Download - The EOWEB GeoPortal (EGP). In *PV 2015*, 2015.

[5] F. Silvestri et al. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 2010. doi: 10.1561/1500000013.

[6] A. Spink et al. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 2001. doi: 10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.0.CO;2-R.

[7] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 2014. doi: 10.1145/2629489.