ISPRS Int. J. Geo-Inf. **2014**, *3*, 345-367; doi:10.3390/ijgi3010345

OPEN ACCESS

ISPRS International Journal of

Geo-Information

ISSN 2220-9964

www.mdpi.com/journal/ijgi/

Article

Geo-Enrichment and Semantic Enhancement of Metadata Sets to Augment Discovery in Geoportals

Bernhard Vockner * and Manfred Mittlböck

Research Studios Austria—Studio iSPACE, Schillerstr. 25, Salzburg 5020, Austria;
E-Mail: manfred.mittlboeck@researchstudio.at

***** Author to whom correspondence should be addressed; E-Mail: bernhard.vockner@researchstudio.at;
Tel.: +43-662-90858-5222; Fax: +43-662-90858-5299.

Received: 17 January 2014; in revised form: 18 February 2014 / Accepted: 3 March 2014 / Published: 17 March 2014

Abstract: Geoportals are established to function as main gateways to find, evaluate, and start “using” geographic information. Still, current geoportal implementations face problems in optimizing the discovery process due to semantic heterogeneity issues, which leads to low recall and low precision in performing text-based searches. Therefore, we propose an enhanced semantic discovery approach that supports multilingualism and information domain context. Thus, we present workflow that enriches existing structured metadata with synonyms, toponyms, and translated terms derived from user-defined keywords based on multilingual thesauri and ontologies. To make the results easier and understandable, we also provide automated translation capabilities for the resource metadata to support the user in conceiving the thematic content of the descriptive metadata, even if it has been documented using a language the user is not familiar with. In addition, to text-enable spatial filtering capabilities, we add additional location name keywords to metadata sets. These are based on the existing bounding box and shall tweak discovery scores when performing single text line queries. In order to improve the user’s search experience, we tailor faceted search strategies presenting an enhanced query interface for geo-metadata discovery that are transparently leveraging the underlying thesauri and ontologies.

Keywords: synonym; translation; SKOS; thesaurus; ontology; Wiktionary; geoplatform; catalogue; SDI

1. Introduction

Geoportals have been well established as main entrance points to geographic information. They act as broker between spatial data providers and users. However, we face two major drawbacks working with geoportals opening Spatial Data Infrastructures (SDIs): First, not all existing geographic data sources are documented and are, therefore, not available to a broader community. Second, if datasets are documented in a standard-based manner, its metadata is usually stored in catalogue components of geoportals. Their interfaces currently lack efficient discovery mechanisms.

To overcome the first issue, efforts must be raised to communicate the advantages of sharing information to both resource providers as well as consumers. This is an organizational, not a technical, issue and, therefore, has to be considered the most time-consuming process when establishing an SDI portal as people must be convinced of the advantages of user and resource provider collaboration [1].

The second drawback deals with a mixture of technical implementations, semantic heterogeneity problems, and established user's search strategies and preferences. Over the last years, initiatives, such as Copernicus (formerly known as GMES—Global Monitoring for Environment and Security), GEOSS (Global Earth Observation System of Systems), and SEIS (Shared Environmental Information System) aimed at advancing technical or syntactic interoperability. Semantic interoperability that is dealing with the meaning of terms has been treated less. At least, in all those initiatives semantic interoperability is an important concern. For example, within projects contributing to GEOSS, such as EuroGEOSS, drive forward the approach on how to handle semantic interoperability problems especially within the discovery broker [2].

In this paper, we mainly focus on strategies to overcome the second issue.

Currently, searching for geographic information in geoportal implementations is based on querying title, abstract, and keyword metadata sections in combination with spatial and temporal filters. As keywords are inherently restricted by the ambiguities of natural languages, semantic heterogeneity problems arise [3]. Especially when dealing with multiple languages, conflicts arise due to various interpretations of features depending on social and cultural background [4].

As in many cases, SDI communities cannot share resources in all languages and English as lowest common denominator may not be sufficient, the need for developing a strategy for cross-language information retrieval within geoportals was already noted by Nogueras-Iso *et al.* [5] and Nowak *et al.* [4]. These approaches include translation of query terms to multiple languages and translation of the content of the results.

In addition to language, information domain singularities increase semantic interoperability problems. Geographic communities may even have various different perspectives and according to that, they may use different terminologies to describe the same type of information. In many cases, synonyms exist.

As a result, we have to cope with low recall and low precision in information discovery. Low recall arises when the terminology used by the user differs from that of the provider. Thus, recall identifies the completeness of retrieval. This means, that users cannot discover all relevant information available which they were looking for [6]. Low precision means that not all the discovered information is relevant to the user. It is often combined with a long list of search results.

Wache *et al.* [7] and Buccella *et al.* [8] propose (multilingual) thesauri and ontologies as approaches to reconcile semantic heterogeneity. The major advantage of thesauri and ontologies, however, is that "*they make the semantics of an application domain understandable to both human and machines*" [9]. In that way, they are means to communicate and translate the meaning of terms proposed by the user to the machine and *vice versa*.

The aim of this research paper is not to develop another new multilingual ontology or thesaurus, but instead to use well-established existing controlled vocabularies (thesauri) and ontologies to enhance metadata documents with synonyms and translated terms, as well as location names for an improved discovery and linked-data eligibility. Additional query terms in several languages will improve the process of metadata discovery. Location derived from the bounding box of the metadata sets and converted to a descriptive term offers the possibility for users to search by location in a text-based manner they are used to [10].

2. The Role of Geoportals in SDI

Geoportals have evolved as standard gateways and main access points of spatial as well as non-spatial information for SDIs. The catalogue interface of a geoportal serves as the major component for information retrieval. International, European, and national standards guarantee the technical interoperability of metadata catalogues, while semantic interoperability is only covered for metadata element definitions, not its content. The most widely used standard in the geospatial domain is the Catalogue Service Web (CSW 2.0.2) standard of the Open Geospatial Consortium (OGC). CSW defines an XML (eXtensible Markup Language) structure and operations, which enable the registration of metadata and query of resources in a standardized manner. CSW-compliant clients can search CSW-based catalogue services. They provide enhanced technical querying and maintenance capabilities to a broad audience even across domains.

The design of geoportals is mainly influenced by the principles of the OGC and the concept of Service-Oriented Architectures (SOA). The center of this architectural concept is based on the "publish-find-bind" paradigm. This paradigm states that a service provider registers or publishes a service in a catalogue or repository ("publish") and users subsequently can search and find these resources ("find"). The user also gets all required information to consume the information when provided as a service ("bind"). In this general approach, the discoverability of resources plays a central role: Only when users can discover resources they are able to make use of it.

One central element in the concept of geographic catalogues is that they do not essentially contain the data itself, but rather the descriptive metadata information about the datasets or services. In the ideal case, the metadata also contains a link to the actual data or service.

Metadata in the geospatial domain nowadays is usually stored and maintained based on metadata standards developed by the International Organization for Standardization (ISO) and the OGC. The ISO 191xx series includes standards for metadata descriptions of spatial resources. The most important of these standards is ISO 19115, which provides a framework for the coordinated development of interoperable interfaces for metadata and catalog services. ISO 19139 provides an XML encoding for the ISO 19115 standard.

In addition to these well-suited standards for spatial datasets, in many cases, the geospatial agnostic metadata standard Dublin Core (DC) is used due to interoperability issues as "least common denominator" to share metadata between heterogeneous domains. Dublin Core originates from the library-domain and was at its first stage used for cataloging books. The simple profile consists of a standard set of 15 mandatory metadata elements such as title or abstract of an item. Thus, it is useful for cataloging all different kinds of information resources [11].

In summary, due to the evolvement of approaches over the last few years, the technical basis for future geoportal solutions is prepared. As a major achievement, syntactic interoperability is ensured. When analyzing the results of geoportal queries, we have to state that we have to put additional efforts, work and research to reduce and overcome semantic heterogeneity, which is based on heterogeneity of user communities, different metadata language and domain context.

We propose to minimize and overcome semantic heterogeneity challenges by providing solutions for cross-language information retrieval techniques combined with ontologies and thesauri.

3. Semantic Interoperability and Semantic Heterogeneity Issues

Various authors [3,12,13] have noted the challenges about semantic interoperability and semantic heterogeneity. In that sense, semantic heterogeneity refers to the "*disagreement about the meaning, interpretation or intended use of the same or related data*" [14].

Semantic heterogeneity is the major issue that prevents full interoperability between systems [13]. In contrast to syntactic interoperability, semantic interoperability deals with the actual meaning of elements [3]. Due to the reason that people interpret the same information in a different way, "*detecting and solving semantic heterogeneity are difficult problems*" [12]. This holds especially true for metadata sets if they do not provide enough information to users on how to interpret the resources they are describing correctly [15].

Lutz *et al.* [3] distinguish between three levels of semantic heterogeneity. Semantic heterogeneity occurs at the metadata level (regarding "*discovery of geographic information*"), at the schema level (concerning the *"retrieval of geographic information*"), and at the data content level ("*interpretation, integration and exchange of geographic information*"). In this paper, we focus on semantic heterogeneity at the metadata level. This level arises, for instance, when resource providers and users do not share a common understanding of terms. This originates from having different scientific, societal or cultural backgrounds. When people communicate with each other (orally or in written form), they incorporate their own knowledge when sending and perceiving the words. However, when we interact with computers, the ability to transfer that "knowledge" is more complex: The computer system may not be able to interpret what we want to achieve within a certain context in the way a human being would probably do. Thus, domain knowledge needs to be incorporated in these systems to enable precise exchange of meaning [14].

Semantic interoperability problems increase when having to cope with multiple languages. As already mentioned, cross-language information retrieval is considered to be an efficient means towards overcoming semantic heterogeneity.

4. Cross-Language Information Retrieval in SDI

Cross-language information retrieval is a sub-discipline of information retrieval. "*Information retrieval (IR) is concerned with finding and returning information stored in computers that is relevant to a user*'*s needs (materialized in a request or query)*" [16]. The main challenge of that research area is that "*we are asking the computer to supply the information we want, instead of the information we asked for*" [17].

In our work, cross-language information retrieval (sometimes also referred to as multilingual information retrieval) deals with two aspects: The resources that can be documented in different languages and that the user's query term may differ from the language used to describe the resource [6]. Cross-language information retrieval can be defined as "*retrieving information written in a language different from the language of the user*'*s query*" [18].

Throughout literature in the geospatial domain, most approaches in information retrieval do not consider multilingualism. Instead, they mainly focus on new similarity algorithms, such as de Andrade and de Baptista [9], and Janowicz *et al.* [19] that deal with improving performance of queries. However, search results are not ordered according to suitability for specific questions as they estimate that results have the same relevance to the user, whereas exact matching may be considered more relevant than matching of related terms [9].

Current research extends the concept of information retrieval by adding context of the user [20]. Context may be defined as "*any information that can be used to characterize and interpret the situation in which a user interacts with an application at a certain time*" [21]. Context information that can be incorporated into discovery is for example location (e.g., based on IP-based location) and language (e.g., based on language settings of the web browser or the language of the search term entered).

Within current geoportal implementations, the discovery process is usually restricted to querying title, abstract and keyword using text-based search. This means that the process of information retrieval is restricted to matching of simple text-patterns not taking into account synonyms (different names for the same concepts), homonyms (same name used for different concepts), or different forms of spelling (e.g., American *vs.* British English).

Within the domain of cross-language information retrieval, additional problems occur: These are—among others—naming heterogeneities dealing with different expressions for the same concept (such as "see" in German, "lac" in French, "lago" in Portuguese, for the concept of lake) and cognitive heterogeneities where different concepts are expressed by the same terms (such as "village green" for "meadow", "grassland", "lea", "mead", and "hayfield") [4]. In the latter case, the problem for any computer system is to determine which of the several senses best fits the query and the results the user estimates that can be a quite challenging task [5].

Other problems are for example that within one language two terms may be represented with only one term in the other (e.g., "South Africa" in English and "Südafrika" in German) or no direct counterpart exists (e.g., in the Russian language there is no distinction between "hand" and "arm").

Another noticeable issue that is sometimes referred to as structural heterogeneity is the vagueness of the definition of geographic concepts [7]. For example the definition of a small and a big river—which in most natural languages have different names like the English word "creek" or the German "bach" for small rivers—concerning width of the river as a method to distinguish between the two concepts may vary between different countries or different database representations.

5. The Role of SKOS, Thesauri and Ontologies in Mitigating Semantic Heterogeneity Problems

In order to overcome semantic heterogeneity problems, the usage of (multilingual) thesauri and ontologies has been widely proposed as a possible solution [3,13]. According to Gruber [22], the term ontology has been defined as a "*formal, explicit specification of a shared conceptualization*". This means that ontologies can be considered as some kind of representation of (real world) objects or our conceptualization of these objects [23]. Thesauri are controlled vocabularies containing relationships [24].

As the aim of this research is not to build a new ontology, a combination of well-known and widely-used thesauri and ontologies is carried out. This can also be considered as a hybrid approach to overcome semantic heterogeneity problems [7].

This section contains examples of ontologies and thesauri we considered to use in our approach. Especially within Europe, the multilingual problem is an enormous challenge due to cultural, historical and social differences between or even among states that are sometimes also represented by different languages used within one country (e.g., Spain, Luxemburg, Belgium). As a result, geospatial databases representing different world views using different languages and descriptions of terms exist, which hinder the process of data sharing across countries [4].

The GEMET Thesaurus (GEneral Multilingual Environmental Thesaurus; [25]) is a valuable means dealing with this problem in offering lists of terms in all official languages. It is maintained by the European Union and is available in 29 languages. The SKOS (Simple Knowledge Organization System) version consists of more than 5000 concepts. Another advantage of using GEMET Thesaurus for discovery is that according to the INSPIRE Implementing Rule for Metadata (MD IR) it is mandatory to provide at least one keyword from GEMET. As a result, it can be estimated that at least one keyword coming from GEMET exists in the discoverable resources.

NASA's SWEET Ontology (Semantic Web for Earth and Environmental Terminology; [26]) consists of about 6000 concepts and is downloadable in OWL (Web Ontology Language) format. It is available in the English language only and, thus, not well suited for our cross-lingual approach.

Wordnet [27] is an English database of synonyms. It resembles a thesaurus, as words are grouped together based on their meanings. It consists of 207,000 word-sense pairs in English language.

Openthesaurus [28] is a free German dictionary for synonyms. It is a community-based approach and consists of 106,000 words.

The UNESCO Thesaurus [29] is a structured list of concepts containing the fields of education, culture, natural sciences, social, and human sciences, communication and information. It comprises about 7000 terms in the English, Russian, French, and Spanish languages.

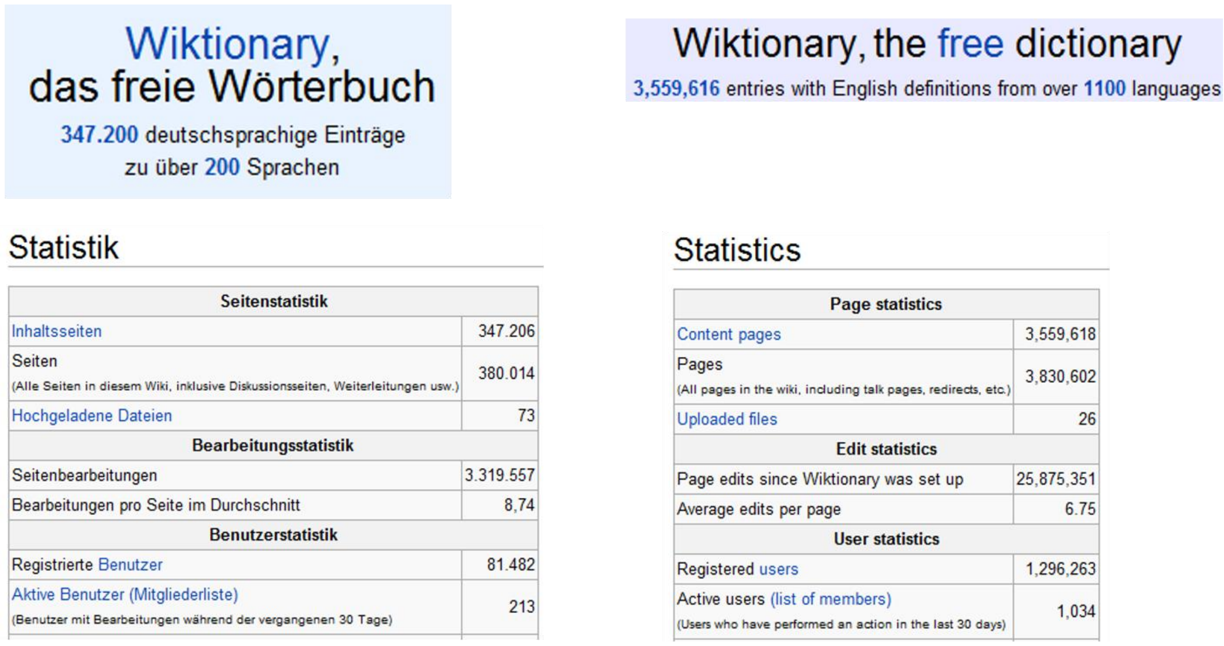
AGROVOC [30] is maintained by the Food and Agriculture Organization of the United Nations. It consists of nearly 40,000 concepts in over 20 languages. It can be downloaded in several formats, including SKOS and OWL, and is free of charge for educational or other non-commercial purposes.

EuroVoc [31] is maintained by the European Union and consists of 22 EU languages. It comprises approximately 6000 terms and is available for download in SKOS/RDF- (Resource Description Framework) or XML format.

Wiktionary [32] is a project intended to provide a free multilingual dictionary. The ambitious overall aim of the project is to create this dictionary for all terms of all languages. Right now, the Wiktionary thesaurus consists of more than 347,200 German words and 3,559,616 English terms

(Figure 1). In contrast to other thesauri, it has the advantage to be maintained and updated regularly by a large user community, as well as having the largest amount of terms compared to other thesauri.

Figure 1. Usage statistics of Wiktionary.



Wiktionary, das freie Wörterbuch
347.200 deutschsprachige Einträge zu über 200 Sprachen

Statistik

Seitenstatistik	
Inhaltsseiten	347.206
Seiten (Alle Seiten in diesem Wiki, inklusive Diskussionsseiten, Weiterleitungen usw.)	380.014
Hochgeladene Dateien	73
Bearbeitungsstatistik	
Seitenbearbeitungen	3.319.557
Bearbeitungen pro Seite im Durchschnitt	8,74
Benutzerstatistik	
Registrierte Benutzer	81.482
Aktive Benutzer (Mitgliederliste) (Benutzer mit Bearbeitungen während der vergangenen 30 Tage)	213

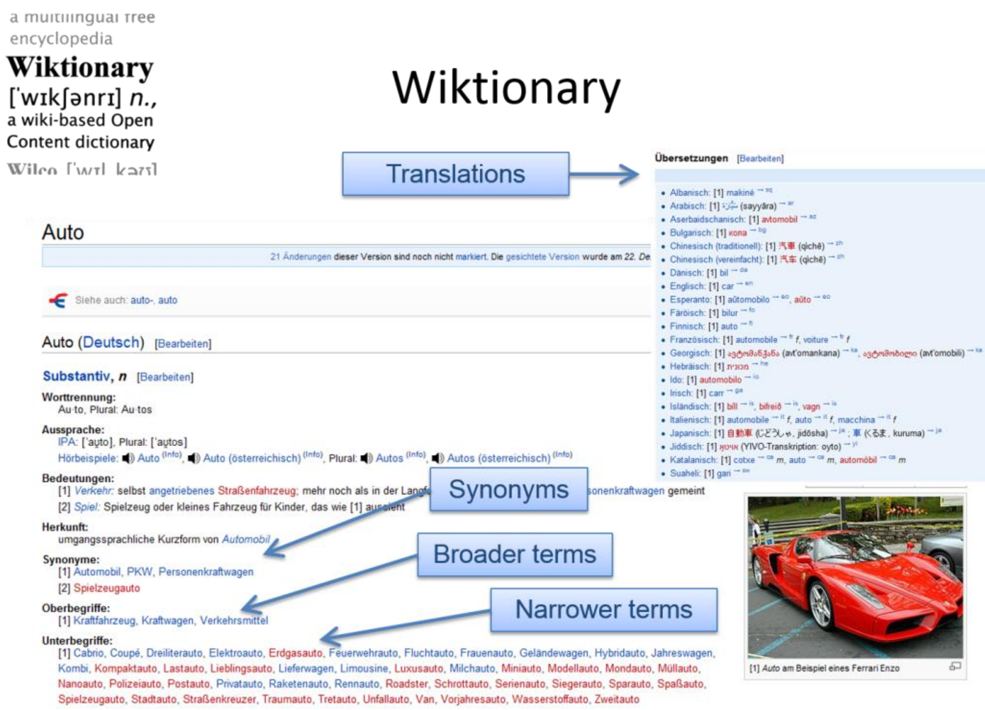
Wiktionary, the free dictionary
3,559,616 entries with English definitions from over 1100 languages

Statistics

Page statistics	
Content pages	3,559,618
Pages (All pages in the wiki, including talk pages, redirects, etc.)	3,830,602
Uploaded files	26
Edit statistics	
Page edits since Wiktionary was set up	25,875,351
Average edits per page	6.75
User statistics	
Registered users	1,296,263
Active users (list of members) (Users who have performed an action in the last 30 days)	1,034

Wiktionary is not available in SKOS format, but as it shares the same principles with SKOS (synonyms, broader, and narrower terms), it can be used like a SKOS when parsing the wiki page through the Wiki API. Figure 2 shows an example of the German Wiktionary and the query for “Auto” (“car” in English), showing synonyms, such as “Automobil” (“automobile” in English), and broader (“PKW”, which is short for “passenger car”), as well as narrower terms, such as “cabrio” (“convertible” in English).

Figure 2. Wiktionary page providing synonyms, broader/narrower terms and translations.



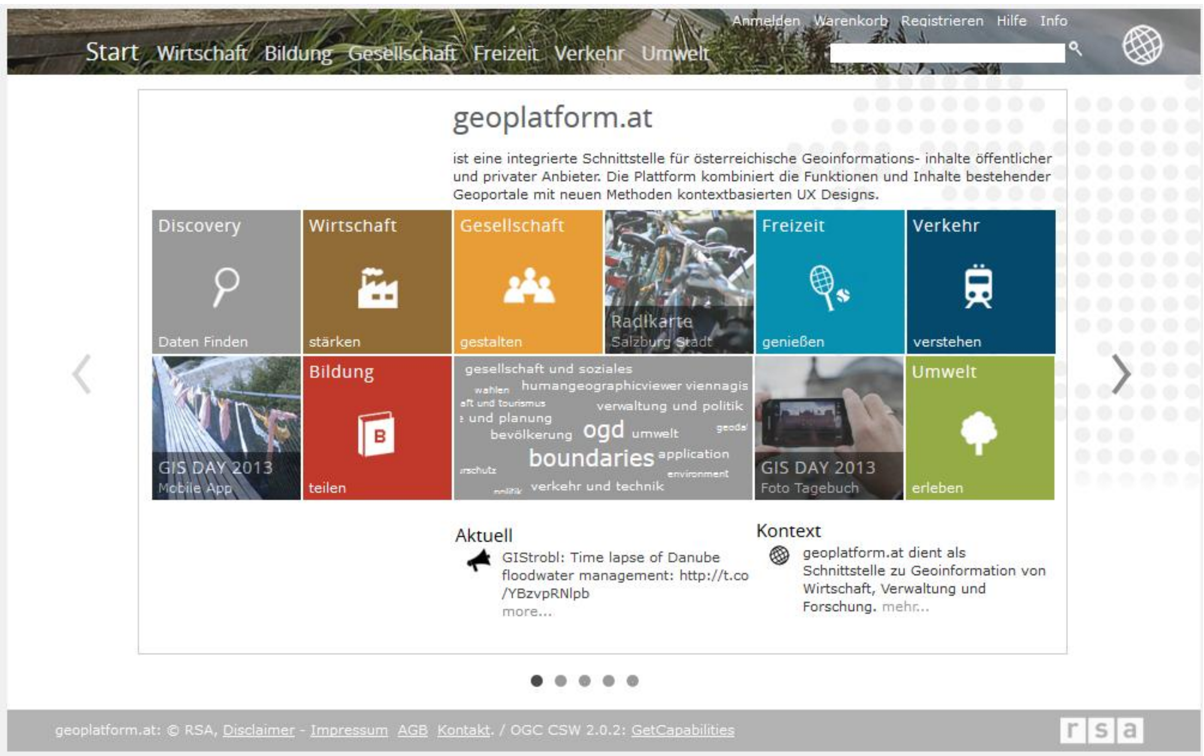
6. Geoplatform.at—A Metadata Catalogue for Geo-Enrichment Semantic Enhancement

Within the domain of geographic information, many valuable sources of information exist. In Austria, several organizations and institutions provide metadata of their geospatial resources. Most of these spatial data providers use geoportals, such as Geoportal Server and Geonetwork, based on the Open Geospatial Consortium's Catalogue Service Web (OGC CSW 2.0.2) standard to share information. As a benefit of standardization approaches in Spatial Data Infrastructure (SDI), it is possible to harvest resources from one metadata catalogue to another.

Within the last few years, a parallel concept of metadata catalogues has evolved coming from the domain of Open Government Data (OGD). The main idea of OGD is to make resources (such as datasets, documents, and web services) of common interest available to the public without any restrictions. As eighty percent of the overall information is considered to be spatially related [33], OGD also provides metadata on geographic information. OGD uses CKAN (Comprehensive Knowledge Archive Network; [34]) as an interface to share their content. Due to harmonization efforts, the Austrian metadata profile ON A 2270:2010 (profil. AT) and the Austrian OGD profile [35] are aligned with each other. Thus, resources can be interchanged between these two domains.

One of these prototypical approaches that use metadata of these heterogeneous resources is geoplatform.at (Figure 3). It serves as the basis for the research approach we present in this paper. It contains resources coming from the "Federal Office for Metrology and Surveying" (BEV), "Land-, forst- und wasserwirtschaftliches Rechenzentrum" (LfRZ), "Geoland", "Cooperation OGD Österreich" and a portal of the federal states of Austria (GIS-Steiermark).

Figure 3. geoplatform.at: A platform to share knowledge [36].



All of the catalogues mentioned above use one of the following technical basis.

6.1. Geoportal Server

Geoportal Server [37] is an open source standard-based solution for storage and discovery of spatial and non-spatial resources. It supports ISO, FGDC, and DC metadata and can be extended with custom or national profiles.

6.2. Geonetwork

Geonetwork [38] provides a user interface to edit metadata as well as a discovery interface to search for geospatial data across multiple catalogs. Like Geoportal Server, it is also capable in using ISO, FGDC, and DC metadata.

6.3. CKAN

CKAN (Comprehensive Knowledge Archive Network; [34]) is a registry or catalogue system for datasets or other "knowledge" resources developed by the "Open Knowledge Foundation" (OKFN). CKAN aims to make it easy to find, share, and reuse open content and data, especially in ways that are machine automatable [39].

CKAN is an Open Data Catalogue that is widely used by governmental institutions in the framework of Open Government Data (OGD), which is—beside others—widely used within the UK (data.gov.uk), Norway (data.norge.no) and mainly by the bigger cities in Austria. Within CKAN, all kinds of resources are documented in a format similar to that of Dublin Core. It gives the "lowest common denominator" of metadata for its packages: author, id, license, user-generated tags, and links, which can be extended by the user in the so-called "extras"-section or predefined by the provider of the CKAN-portal due to domain-specific needs [40].

Geonetwork and Geoportal Server make use of the CSW protocol and thus can be harvested vice-versa. As there is no direct link between OGC's CSW and CKAN's REST API, we developed a Python tool for the harvesting process between the catalogue component of geoportals and CKAN. We query the REST API of CKAN with a library called "ckanclient" [41] and the CSW interface of the geoportal, perform a transformation of the elements to comply with the ISO metadata core and register the resources with Geoportal Server using a CSW insert transaction. The harvesting process itself is, right now, performed once per day, but update frequency can be increased or decreased on demand. Another possible option to harvest metadata from CKAN would have been the "ckanext-spatial" extension. However, as metadata in geoplatform.at is based on the Austrian profile ON A 2270:2010 (profil.AT), we preferred a custom harvesting process with REST because the standard ckanext-spatial extension does not account for all the elements present in the Austrian metadata profile.

Currently, geoplatform.at [36] consists of more than 2200 metadata sets describing datasets, services, and documents of the geospatial domain. The metadata is completely in the German language. As in the initial stage of geoplatform.at users were not satisfied with the results of discovery both in regard of a lack of synonym search capabilities or inefficient discovery of place names, we created and implemented a non-spatial semantic enhancement and geo-enrichment workflow in geoplatform.at.

7. Non-Spatial Semantic Enhancement Workflow

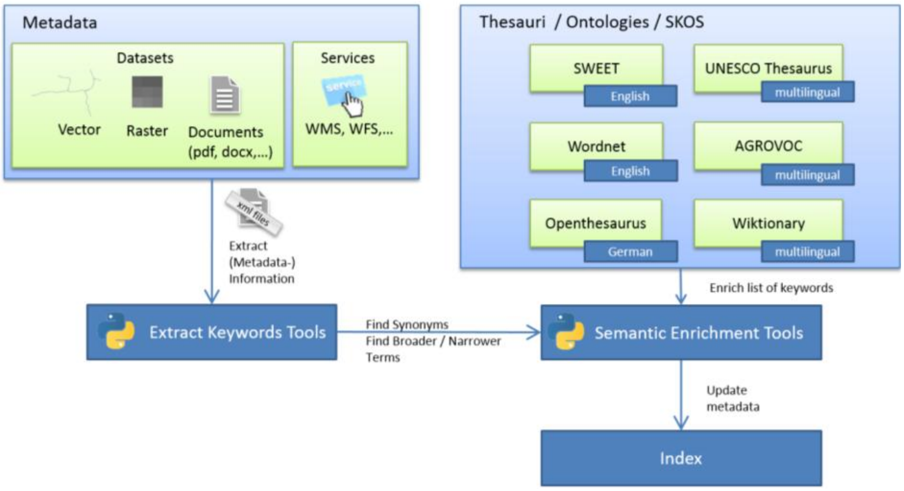
Our non-spatial semantic enhancement workflow relies on thesauri, ontologies and SKOS. The workflow itself starts with metadata stored in the catalogue component of the geoportal (Figure 4). In our prototypical implementation, we use Geoportal Server 1.2.4 as metadata catalogue and Python 2.7 to extract the keywords of the metadata documents. We developed various Python tools to find synonyms, broader and narrower terms, and translations of the user-defined keywords. In our prototype, we can use different sources as input, but we prefer multilingual thesauri, as they better suit our cross-lingual discovery approach. We can use the SWEET ontology (English), Wordnet (English), Openthesaurus (German), the UNESCO Thesaurus (multilingual), AGROVOC (multilingual), and Wiktionary (multilingual). From a technical viewpoint, our prototype is based on several Python modules. They comprise "ztfy.theasaurus" [42], "python-skos" [43], "NLTK" [44], and "requests" [45]. We combined them with each other and added some additional code parts to serve our needs. The implementation of thesauri that came as a web service and offered an API (such as in the case of Openthesaurus) was straightforward. In the case of Wiktionary, we had to query and parse the response site for each keyword in our Python script. Thus, we propose to develop a REST (Representational State Transfer) interface for Wiktionary with a clear API definition to make translations and synonyms, as well as broader and narrower terms, faster and easier accessible for our prototype and other use cases.

Our prototype also makes use of SKOS. The major challenge for parsing SKOS files is to cope with files that do not fulfill the requirements of the SKOS standard. In that case, manual adaptions of the code are necessary and lead to a longer development phase.

Right now, the python prototype runs on a daily basis to enhance the metadata sets with our semantic enhancements. This interval, however, can be in- or decrease on purpose.

As in our research prototype geoplatform.at only German metadata sets exist, we preferred thesauri available in German language. As Wiktionary offers the broadest amount of terms, it is the ideal resource for our approach and we were most satisfied with its results. Thus, Wiktionary was chosen to be used in most cases.

Figure 4. Non-spatial semantic enhancement workflow.



We created a use case to show the benefits of our approach. In our use case, we search for the German term for “convertible”, namely “cabrio” in geoplatform.at (Figure 5). The list of results shows entries such as “Kraftfahrzeugbestand’ (‘stock of cars’). In this metadata entry, the term “cabrio” did not exist before, but was added by our non-spatial semantic enhancement tool.

Figure 5. Synonym search.

As our approach considers translations of terms as well, we can also look for “car” in order to discover the same metadata documents as in the former use case (Figure 6). To give the user at least an idea on the content of the data discovered, we provide an automatic translation of the results (Figure 7). In our prototype, we use the “Microsoft translator” [46], which is currently available in 42 languages. At first sight, we only added English translations to the metadata documents, but other languages can be added in the same manner.

Figure 6. Multilingual search.

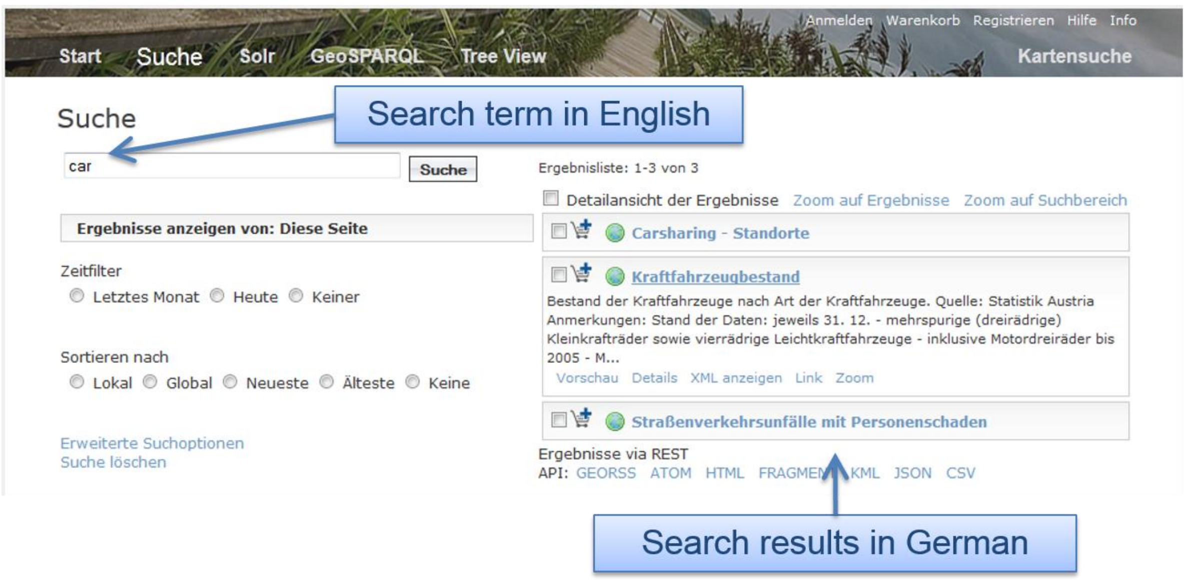
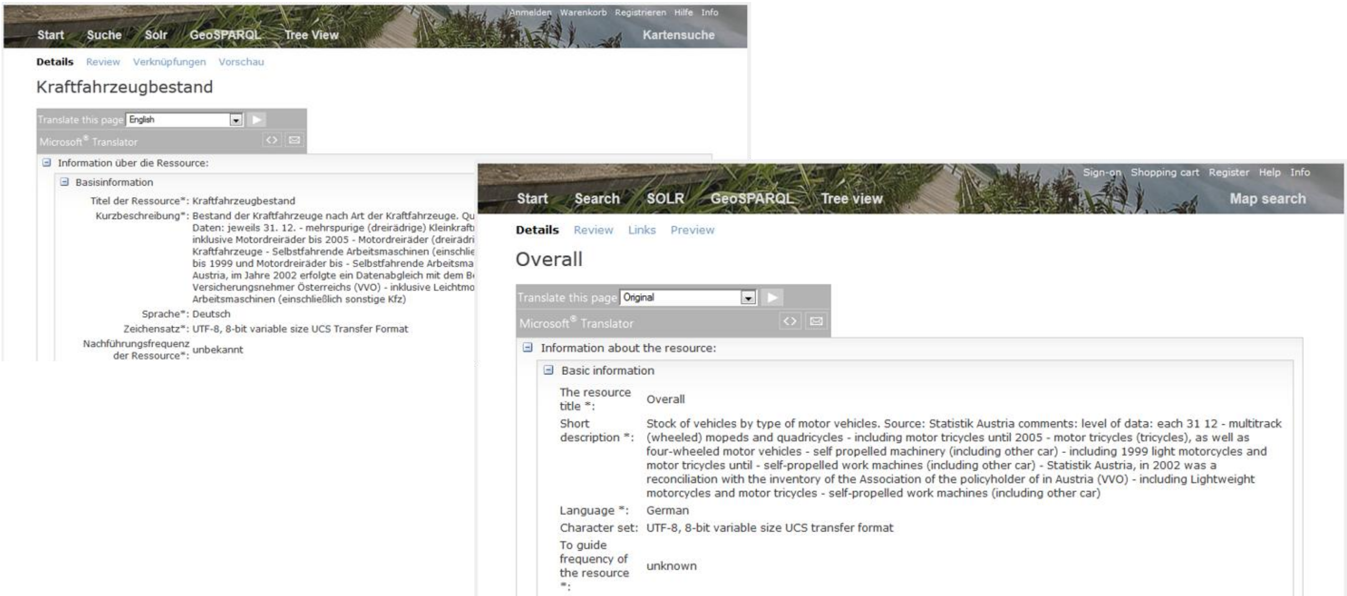


Figure 7. Automatic translation of search results.

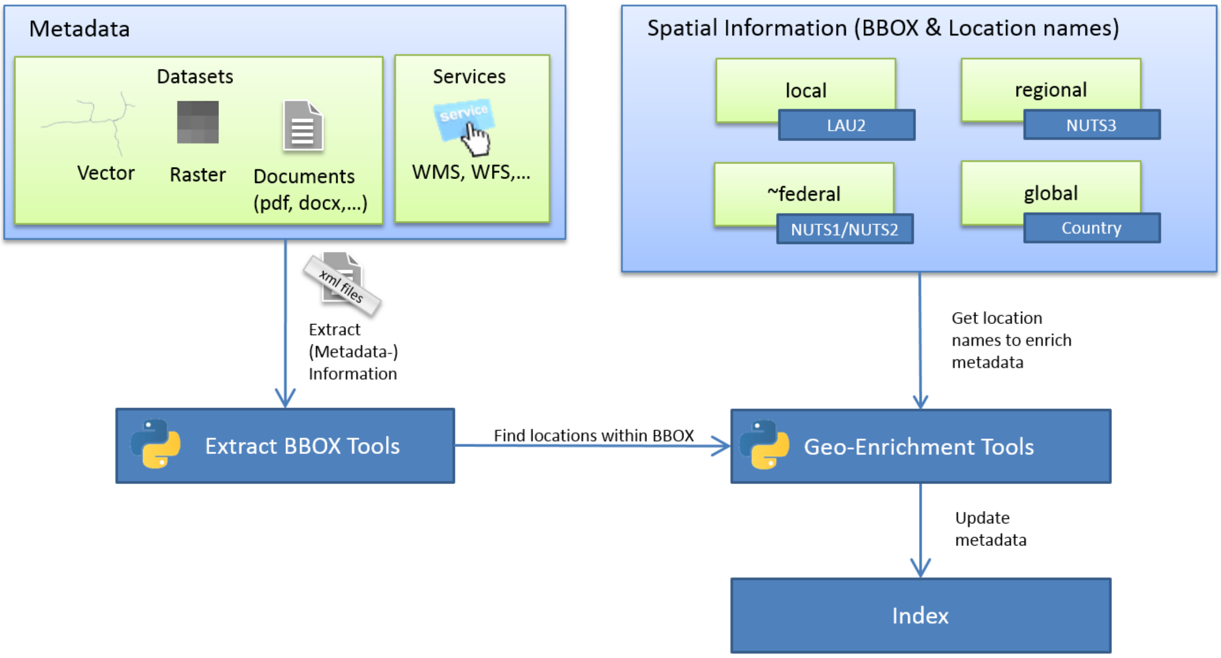


8. Geo-Enrichment Workflow

In addition to being able to query for synonyms, users should be able to look for locations, not only based on a map selection, but also as textual inputs. In our prototype, the location keywords are populated on a daily basis with a python script that uses the NUTS ("Nomenclature des unités territoriales statistiques", "Nomenclature of territorial units for statistics") and LAU 2 ("Local Administrative Unit") regions of the European Union (Figure 8). We use the spatial extent of the resources and annotate them with names of locations that are within the bounding box of the resource automatically. The metadata comes from the catalogue component of our geoportal implementation. In cases where only a location name (which holds especially true for documents, such as pdfs that contain descriptive information about a dataset or project) is present in the metadata, the location name is geocoded with Python and the additional geopy library [47]. With geopy, Google Maps, Yahoo! Maps, Windows Local Live (Virtual Earth), geocoder.us, GeoNames, MediaWiki pages, and Semantic MediaWiki pages can be used. For full texts, we use Named Entity Recognition (NER), such as the Natural Language Toolkit (NLTK; [44]) and an adapted version of the geodict library [48] to extract location keywords.

For example, if users search for "Hallein" (a small town in Austria), they also get results of "Salzburg", because "Hallein" is located within the federal state of "Salzburg". Challenges arise when the same term is used for a city and a federal state. That holds true for Salzburg, which can either be the city of Salzburg or the federal state of Salzburg. In that case, the context can be utilized to handle this issue. For example, if the terms "Salzburg" and "city" are used in combination, the system may interfere that the city of Salzburg is meant.

Figure 8. Geo-enrichment workflow.



9. Enhancing Discovery User Interfaces in SDIs

For discovery purposes, we present three different approaches to query the catalogue. Two rely on Apache Lucene (Geoportal Facets Customization with Apache Solr and GeoSPARQL). One approach uses semantic text matching algorithms in combination with recommender systems that are widely used within online stores like Amazon.com.

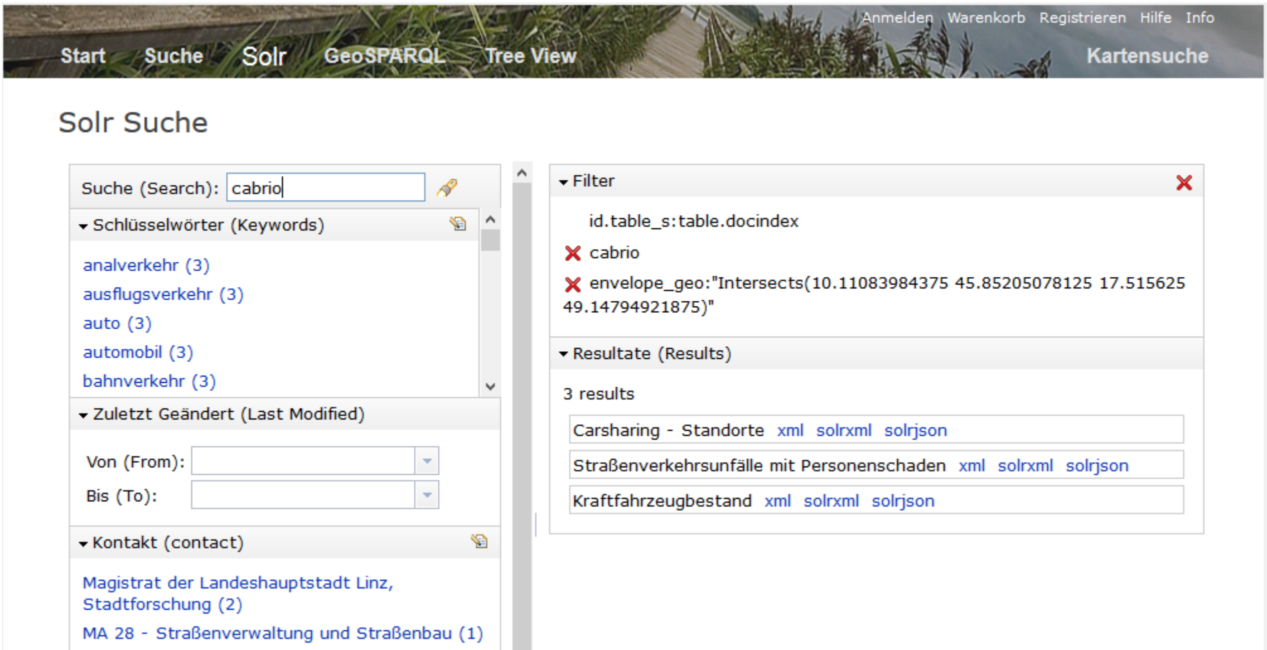
9.1. Extended Search Capabilities Using Apache Solr

Apache Solr offers the possibility for end users to filter their searches based on either keywords and/or spatio-temporal parameters. The major advantage in using Apache Solr is its performance as it uses highly advanced indexing mechanisms. Apache Solr is a Java servlet that can be deployed within a container, such as Apache Tomcat. Full-text indexing and search is based on the Lucene Java search library [49]. To integrate and use it in geoplatform.at, it provides a REST-like API. The Geoportal Facets Customization (GFC) leverages the Apache Solr 4.1.0 index of Geoportal Server and, thus, extends it to be able to use Solr for discovery [50].

Figure 9 shows the user interface of the search page based on Apache Solr. Filters, such as spatial and temporal extent, last modification date, keywords, and organizations, can be used in combination to narrow down search results.

Solr offers the possibility to integrate other systems such as other geoportals in the index, thus providing distributed search mechanisms. This is especially useful for our prototype that uses resources, which originate from different systems. Spatial search can either be performed using the map or searching for locations based on keywords as introduced with our geo-enrichment workflow.

Figure 9. Solr search.



9.2. Extended Semantic Search Capabilities Based on GeoSPARQL for Experts

When trying to establish links between non-related information, RDF as basis for the concept of Linked Data is particularly used in the field of (online) libraries. Since 2010, this concept has also been applied in the geospatial domain, where more and more data of heterogeneous sources, such as databases, spreadsheets, XML, measurements, or crowd sourced data is available due to open data efforts. Conversion of metadata of these resources to RDF can be used to make semantic searches and therefore easier discovery in underlying catalogues possible. The semantics also refer to the heterogeneity problem, when several synonymous terms for the same object exist.

Linking data is especially necessary to be able to answer complex questions such as: "Which circuits are necessary to switch electricity supply from one side of a specific river to the other if the water level increases by more than five meters?"

To formulate such a query, SPARQL can be applied to RDF triples, consisting of subjects, predicates, and objects. The search itself is based on the discovery of patterns that satisfy the query criteria according to filters [51]. The query syntax is similar to that of Structured Query Language (SQL), which is broadly used within the domain of databases. The spatial extension of SPARQL is called GeoSPARQL and facilitates the formulation of spatial queries using well-known concepts of Egenhofer, such as contains, inside, outside, and within [51].

We integrated a GeoSPARQL extension in the open source Geoportal Server. Thus, an expert user familiar with GeoSPARQL can directly use it to query the metadata catalogue. An example for the search syntax of our prototype is provided in Figure 10. The result of the search is shown in Figure 11.

Figure 10. GeoSPARQL search.

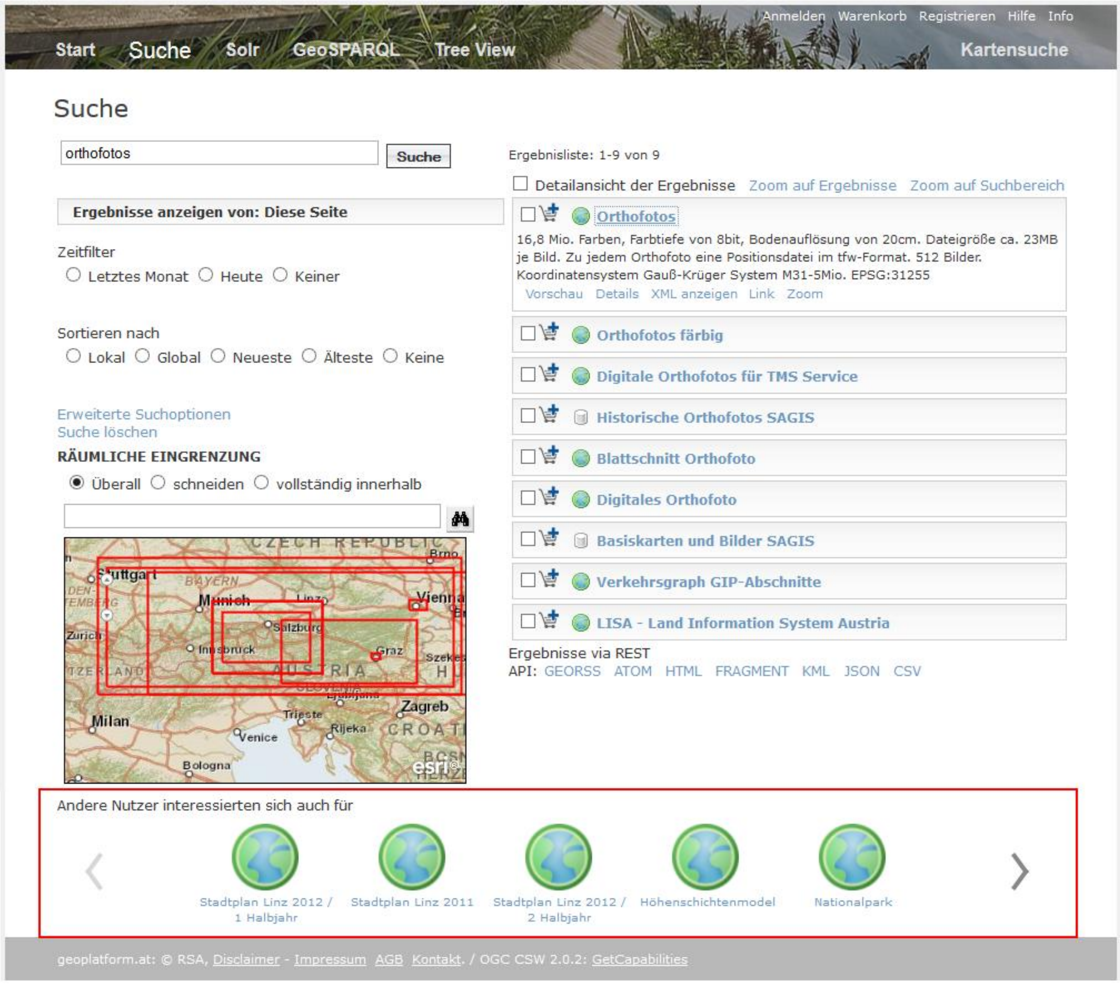
Figure 11. Result of GeoSPARQL search.

9.3. Extended Semantic Search Capabilities for Casual Users

As already presented by Vockner *et al.* [52], we extended the standard search capabilities of the catalogue component of geoplatform.at with the concept of recommendations in combination with a semantic text matching approach that identifies items that are related to each other. The recommendations are calculated based on either which resources other users viewed together, on the ratings of items and on a tool that calculates the semantic similarity of the resource abstracts with an algorithm called Latent Semantic Analysis (LSA; [53]). Simplified, LSA converts texts into n-dimensional vectors, reduces their dimension and discovers the semantic structure of texts by examining the statistical co-occurrence pattern. LSA uses the cosine similarity to determine the similarity of two vectors or texts. A value of 1 shows an exact match, a value of 0 indicates that there is no match.

The user interface showing a scrollable list of recommendations is highlighted in Figure 12.

Figure 12. Recommender-enhanced discovery.



10. Results

To cope with the enormous amount of resources within SDIs, we demanded an efficient and effective strategy for discovery. In order to make resources discoverable, we introduced the concepts of geo-enrichment and semantic enhancement of metadata sets. Thus, users can discover metadata although they may use a different term than the provider of the metadata used to describe the resource. Our approach also enables to discover information in a different language. Automatic translations assist the user in getting knowledge about the content of the resources that were documented in a different language. The overall approach helps users in performing their discovery tasks. They can retrieve (meta-)data they would not have discovered directly with the search term or search language they entered.

Concerning discovery for spatial datasets, users are usually accustomed to a text-based search interface rather than a bounding box selection in a map due to everyday-life experience in performing web searches. Therefore, our geo-enhancement workflow allows the users to formulate their queries in textual form.

We tested our approach in the prototype of geoplatform.at. For this purpose, we queried for terms such as “cabrio”, “convertible”, and “auto” before we applied our semantic enhancement and geo-enrichment workflow. The query did not return any results. Afterwards, we enhanced the metadata sets with synonyms, translations, broader/narrower terms, and toponyms of terms present in the metadata documents. As a result, the search now returned items (e.g., "stock of cars") we were not be able to discover before.

However, our test showed that in some cases results based on the semantically enhanced keywords did not show satisfying results: For example if someone searches for "kokain" (German word for "cocaine") to look for information about drug crime rate, the result shows information such as "snow cover". That is related to the fact, that "snow" is the slang word for cocaine. This shows that our implementation of Wiktionary should at least avoid taking into account any slang words in the next version.

We created three types of user interfaces to discover the resources registered in geoplatform.at. In literature, (Geo-)SPARQL was proposed as an efficient mean to query catalogue interfaces, allowing the formulation of rather difficult queries, which are posed to the catalogue. This approach is especially suited for expert users. For those not familiar with the GeoSPARQL syntax and language, we will develop a graphical user interface that hides the complexity of the query structure from the end user. However, this will also lead to a reduction of the freedom in search as then mainly pre-formulated queries can be posed.

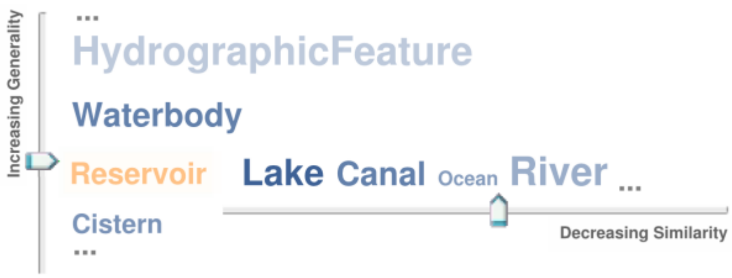
Casual users may prefer our recommendation approach that shows related items to the ones the keyword-based search has returned. This concept has proven to be quite successful in the domain of online stores and was therefore integrated in the geoplatform.at approach we use in this paper. As an additional method to search for information, Apache Lucene and Apache Solr can be used directly from the user interface component of the catalogue. They allow for full-text search, faceted search, rich document handling (e.g., Word, PDF) and geospatial search [49].

11. Outlook and Discussion

We used ontologies and thesauri to enhance metadata and presented three types of user interfaces to search for data. However, to allow users to "fine-tune" the search, we plan to develop a user interface especially suited for similarity searches.

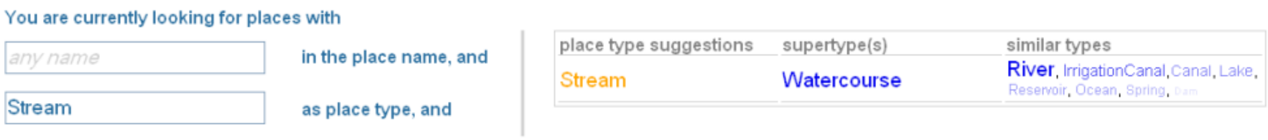
Janowicz *et al.* [54] presented a search interface that allows the user to interact with lists of synonyms. This approach seems to be valuable to extend our approach, as right now only the system but not the user can decide which hierarchy of terms to use in the search process. Their user interface provides scroll bars to increase/descrease generality and similarity of search terms entered by users (Figure 13). The font size and color they use represents similarity values. In contrast to our approach, Janowicz *et al.* [54] do not consider multiple languages. Thus, such a user interface would have to look different in our case.

Figure 13. User Interface to increase/decrease generality and similarity of search terms [54].



In addition, Janowicz *et al.* [54] provide a static search interface, where they show similar place types based on an ontology to the user (Figure 14). This does not allow the user to directly interact with the granularity of terms, but, rather, gives a visual representation of what might be of use to them.

Figure 14. Similarity-based user interface for Web gazetteers [54].



Sinkkilä *et al.* [55] presented a combined concept of auto-completion (a form of textual input tries to complete the word the user enters) of search terms and ontological context navigation (which is a method for selection of terms). In their approach, they present the hierarchy and relationship of concepts to communicate ontologies to the user. This concept could be used in the auto-completion tool we have implemented in geoplatform.at as well, but we would restrict it to less items than Sinkkilä *et al.* [55] use in order to not distract the user. This goes in line with Hyvönen and Mäkelä [56], and Hildebrand *et al.* [57] who also tried to enhance the auto-completion approach with ontologies, coming to the result that auto-completion is a complex feature instead of an easy approach it seems to be at first sight.

Another possibility to enhance our user interface would be to use a tag cloud representation of similar terms. Ostländer and Lutz [58] presented the concept of such a cloud representation of tags in a slightly different manner. There, the size of each concept depends on how often it has been used as a keyword to annotate a resource instead of similarity of the search term and multilingual thesauri entered by the user.

We also experimented with the jOWL JavaScript library [59] to interact with ontologies. It lets the user browse an ontology based on the search text entered to provide more specific or more generic search results (Figure 15).

Figure 15. jOWL JavaScript library to browse ontologies.

The advantages behind the approach of assisting the user in providing broader or narrower terms has already been shown by Croft and Das [60], who reported significant query improvements when users are prompted for additional terms (e.g., in the form of browsing the hierarchy of an ontology) that can be used for searching [6].

We want to introduce a concept of weighting synonyms and translated terms. Right now, we do not assign less weight to synonyms in comparison to user-definded keywords. This is sufficient for our test case but has to be extended for larger amounts of data.

We also plan to integrate the geo-enrichment and semantic enhancement workflows directly in our recommender system so that these workflows also improve the recommendations given to the user by adjusting the ranks of items.

To show the benefits of our approach, benchmark and usability tests to evaluate our approach will be conducted. This is a crucial task because, as van Ossenbruggen *et al.* [61] state, creating and evaluating a user interface containing semantics is quite difficult because the user interface component cannot be separated clearly from the search engine and the quality of the resources used.

12. Conclusions

We propose the integration of geo-enrichment and semantic enhancement of metadata sets in geoportals to overcome semantic heterogeneity problems as the major result of our research. To show the benefits of such an approach, we developed Python tools that run on a daily basis to extract user-defined keywords out of metadata sets and enrich them with synonyms and translations to other languages. This allows for cross-lingual information retrieval where the user is assisted in the discovery task with automatic translation of search results. In addition, we proposed and implemented a workflow on how to geospatially annotate metadata, as users rather search for a location based on textual inputs than bounding boxes. We showed the results of our research as proof-of-concept in geoplatform.at. This prototypical geoportal consists of more than 2200 German metadata sets provided by Austrian metadata organizations that are now semantically and geospatially enriched.

Acknowledgments

Special thanks go to the "Federal Office for Metrology and Surveying" (BEV), "Land-, forst- und wasserwirtschaftliches Rechenzentrum" (LfRZ), "Geoland", "Cooperation OGD Österreich" and a portal of the federal states of Austria (GIS-Steiermark) for providing their metadata for research purposes.

Author Contributions

All authors contributed substantially to the conception of this paper. Bernhard Vockner developed the geo-enrichment and semantic enhancement tools and wrote main parts of the manuscript. Manfred Mittlböck enhanced selected parts substantially and gave guidance and advice during the development process of the paper.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Man, E.D. Understanding SDI; complexity and institutionalization. *Int. J. Geogr. Inf. Sci.* **2006**, *20*, 329–343.
2. The EuroGEOSS Brokering Platform. Available online: http://www.eurogeoss.eu/broker/Pages/TheEuroGEOSSBrokeringPlatform.aspx (accessed on 11 Feburary 2014).
3. Lutz, M.; Sprado, J.; Klien, E.; Schubert, C.; Christ, I. Overcoming semantic heterogeneity in spatial data infrastructures. *Comput. Geosci.* **2009**, *35*, 739–752.
4. Nowak, J.; Nogueras-Iso, J.; Peedell, S. Issues of Multilinguality in Creating a European SDI—The Perspective for Spatial Data Interoperability. In Proceedings of the 11th EC GI&GIS Workshop, ESDI Setting the Framework, Alghero, Italy, 29 June–1 July 2005.
5. Nogueras-Iso, J.; Zarazaga-Soria, F.J.; Lacasta, J.; Tolosana, R.; Muro-Medrano, P.R. Improving Multilingual Catalog Search Services by Means of Multilingual Thesaurus Disambiguation. In Proceedings of the 10th EC GI & GIS Workshop, ESDI State of the Art, Warsaw, Poland, 23–25 June 2004.
6. Smits, P.C.; Friis-Christensen, A. Resource discovery in a European spatial data infrastructure. *IEEE Trans. Knowl. Data Eng.* **2007**, *19*, 85–95.
7. Wache, H.; Vögele, T.; Visser, U.; Stuckenschmidt, H.; Schuster, G.; Neumann, H.; Hübner, S. Ontology-Based Integration of Information—A Survey of Existing Approaches. In Proceedings of the IJCAI-01 Workshop: Ontologies and Information Sharing, Seattle, WA, USA, 4–5 April 2001; pp. 108–117.
8. Buccella, A.; Cechich, A.; Fillottrani, P. Ontology-driven geographic information integration: A survey of current approaches. *Comput. Geosci.* **2009**, *35*, 710–723.
9. Andrade, F.G.D.; Souza Baptista, C.D. Using semantic similarity to improve information discovery in spatial data infrastructures. *J. Inf. Data* **2011**, *2*, 181–194.
10. Renteria-Agualimpia, W.; Lopez-Pellicer, F.J.; Lacasta, J.; Zarazaga-Soria, F.J.; Muro-Medrano, P.R. Identifying Hidden Geospatial Resources in Catalogues. In Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics (WIMS), Madrid, Spain, 12–14 June 2013; pp. 1–7.
11. Nebert, D. GEOSS AIP-2 Polar Ecosystems Biodiversity SBA Engineering Report, 2009. Available online: http://www.ogcnetwork.net/system/files/Final_090721_Polar_ER_ecosystems_v1.pdf (accessed on 14 February 2014).
12. Fileto, R. Issues on Interoperability and Integration of Heterogeneous Geographical Data. Available online: http://www.geoinfo.info/proceedings_geoinfo2001.split/paper6.pdf (accessed on 13 Feburary 2014).
13. Kashyap, V.; Sheth, A. Semantic Heterogeneity in Global Information Systems: The Role of Metadata, Context and Ontologies. In *Cooperative Information Systems: Trends and Directions*; Papazoglou, M., Schlageter, G., Eds.; Academic Press: Waltham, MA, USA, 1997; pp. 139–178.
14. Xu, Z.; Lee, Y.C. Semantic Heterogeneity of Geodata. Available online: http://www.isprs.org/proceedings/XXXIV/part4/pdfpapers/499.pdf (accessed on 11 December 2013).
15. Sheth, A.P.; Larson, J.A. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Comput. Surv.* **1990**, *22*, 183–236.

16. Dominich, S. *The Modern Algebra of Information Retrieval*; Springer Publishing Company: New York, NY, USA, 2008; p. 330.
17. Berry, M.W.; Browne, M. *Understanding Search Engines: Mathematical Modeling and Text. Retrieval (Software, Environments, Tools)*, 2nd ed.; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2005.
18. Manikandan, B.; Shrira, R. A Novel Approach for Cross Language Information Retrieval. In Proceedings of the 3rd International Conference on Electronics Computer Technology (ICECT), Kanyakumari, India, 8–10 April 2011; pp. 34–38.
19. Janowicz, K.; Wilkes, M.; Lutz, M. Similarity-Based Information Retrieval and Its Role within Spatial Data Infrastructures. In Proceedings of the 5th International Conference on Geographic Information Science; Springer-Verlag: Park City, UT, USA, 2008; Volume 5266, pp. 151–167.
20. Keßler, C.; Raubal, M.; Wosniok, C. Semantic Rules for Context-Aware Geographical Information Retrieval. In Proceedings of the 4th European Conference on Smart Sensing and Context; Springer-Verlag: Guildford, UK, 2009; pp. 77–92.
21. Bazire, M.; Brézillon, P. Understanding Context Before Using It. In *Modeling and Using Context*; Dey, A., Kokinov, B., Leake, D., Turner, R., Eds; Springer Berlin Heidelberg: Berlin, Germany, 2005; Volume 3554, pp. 29–40.
22. Gruber, T.R. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.* **1995**, *43*, 907–928.
23. Tripathi, A.; Babaie, H.A. Developing a modular hydrogeology ontology by extending the SWEET upper-level ontologies. *Comput. Geosci.* **2008**, *34*, 1022–1033.
24. Society, I.C. Thesaurus FAQs. Available online: http://www.computer.org/portal/web/tandc/Thesaurus-FAQ (accessed on 23 December 2013).
25. EIONET. GEMET Thesaurus. Available online: http://www.eionet.europa.eu/gemet (accessed on 11 December 2013).
26. SWEET Ontology. Available online: http://sweet.jpl.nasa.gov/ (accessed on 12 December 2013).
27. What is Wordnet? Available online: http://wordnet.princeton.edu/ (accessed on 22 December 2013).
28. Openthesaurus.de-Synonyme und Assoziationen. Available online: http://www.openthesaurus.de/ (accessed on 22 December 2013).
29. UNESCO Thesaurus. Available online: http://databases.unesco.org/thesaurus/ (accessed on 12 December 2013).
30. AGROVOC. Available online: http://aims.fao.org/standards/agrovoc/ (accessed on 12 December 2013).
31. EuroVoc, the EU's Multilingual Thesaurus. Available online: http://eurovoc.europa.eu/.(accessed on 12 December 2013).
32. Wiktionary. Available online: http://www.wiktionary.org/.(accessed on 12 December 2013).
33. Ryttersgaard, J. Spatial Data Infrastructure, Developing Trends and Challenges. In Proceedings of International Conference on Spatial Information for Sustainable Development, Nairobi, Kenya, 2–5 October 2001; p. 8.
34. CKAN, the World'S Leading Open-Source Data Portal Platform. Available online: http://ckan.org/ (accessed on 20 December 2013).

35. OGD Metadata-2.2. Available online: http://reference.e-government.gv.at/Veroeffentlichte-Informationen.2774.0.html (accessed on 16 January 2014).
36. Geoplatform.at. Available online: http://www.geoplatform.at(accessed on 11 December 2013).
37. ESRI Geoportal Server. Available online: http://www.esri.com/software/arcgis/geoportal (accessed on 11 December 2013).
38. GeoNetwork Opensource. Available online: http://geonetwork-opensource.org/ (accessed on 20 December 2013).
39. Comprehensive Knowledge Archive Network (CKAN). Available online: http://lod2.eu/Project/CKAN.html (accessed on 15 December 2013).
40. CKAN: Apt-Get for the Debian of Data. Available online http://events.ccc.de/congress/2009/Fahrplan/events/3647.en.html(accessed on 18 December 2013).
41. Ckanclient. Available online: https://github.com/okfn/ckanclient (accessed on 12 Ocotober 2013).
42. Ztfy.Thesaurus. Available online: https://pypi.python.org/pypi/ztfy.thesaurus (accessed on 13 Feburary 2013).
43. Python-Skos. Available online: https://pypi.python.org/pypi/python-skos (accessed on 13 Feburary 2014).
44. Natural Language Toolkit. Available online: http://www.nltk.org (accessed on 23 December 2013).
45. Requests. Available online: https://pypi.python.org/pypi/requests/2.2.1 (accessed on 13 Febuary 2014).
46. Microsoft Tranlator. Available online: http://www.microsoft.com/en-us/translator/ (accessed on 10 Ocotober 2013).
47. Geopy—A Geocoding Toolbox for Python. Available online: http://code.google.com/p/geopy/ (accessed on 12 December 2013).
48. Geodict—An Open-Source Tool for Extracting Locations from Text. Available online: http://petewarden.com/2010/10/03/geodict-an-open-source-tool-for-extracting-locations-from-text/ (accessed on 12 December 2013).
49. Apache Solr. Available online: http://lucene.apache.org/solr/ (accessed on 11 December 2013).
50. Geoportal Facets. Available online: https://github.com/Esri/geoportal-server/wiki/Geoportal-Facets (accessed on 15 December 2013).
51. OGC GeoSPARQL—A Geographic Query Language for RDF Data. Available online: http://www.opengeospatial.org/standards/geosparql (accessed on 12 December 2013).
52. Vockner, B.; Richter, A.; Mittlböck, M. From geoportals to geographic knowledge portals. *ISPRS Int. J. Geo-Inf.* **2013**, *2*, 256–275.
53. Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Deerwester, S.; Harshman, R. Using Latent Semantic Analysis to Improve Access to Textual Information. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Washington, DC, USA, 15–19 May 1988; pp. 281–285.
54. Janowicz, K.; Schwarz, M.; Wilkes, M. Implementation and Evaluation of a Semantics-Based User Interface for Web Gazetteers. In Proceedings of the Visual Interfaces to the Social and the Semantic Web (VISSW 2009) Workshop in conjunction with the International Conference on Intelligent User Interfaces (IUI 2009), Sanibel Island, FL, USA, 8–11 February 2009.

55. Sinkkilä, R.; Mäkelä, E.; Kauppinen, T.; Hyvönen, E. Combining Context Navigation with Semantic Autocompletion to Solve Problems in Concept Selection. In Proceedings of the SEMMA 2008 Workshop, 5th European Semantic Web Conference 2008, Tenerife, Canary Islands, Spain, 1–5 June 2008.
56. Hyvönen, E.; Mäkelä, E. Semantic Autocompletion. In Proceedings of the First Asian conference on the Semantic Web, Beijing, China, 3–7 September 2006; pp. 739–751.
57. Hildebrand, M.; Ossenbruggen, J.R.V.; Amin, A.K.; Aroyo, L.M.; Wielermaker, J.; Hardman, L. *The Design Space of a Configurable Autocompletion Component*; Stichting Centrum voor Wiskunde en Informatica: Amsterdam, The Netherlands, 2007.
58. Ostländer, N.; Lutz, M. INSPIRE-ing GEMET—Enhancing Metadata Creation and Discovery. In Proceedings of Environmental Informatics and Industrial Ecology Lüneburg, 2008, Lüneburg, Germany, 10–12 September 2008; Moeller, A., Page, B., Schreiber, M., Eds.; Shaker Verlag: Aachen, Germany; pp. 212–214.
59. jOWL Ontology Browser. Available online: http://jowl.ontologyonline.org/ (accessed on 23 December 2013).
60. Croft, W.B.; Das, R. Experiments with Query Acquisition and Use in Document Retrieval Systems. In Proceedings of the 13th Annual International Acm Sigir Conference on Research and Development in Information Retrieval, Brussels, Belgium, 5–7 September 1990; pp. 349–368.
61. Ossenbruggen, J.V.; Hildebrand, A.A.M. Why Evaluating Semantic Web Applications is Difficult. In Proceedings of Semantic Web User Interaction at CHI 2008: Exploring HCI Challenges, Florence, Italy, 5–10 April 2008; pp. 1–4.