Comparing Different Search Techniques for Location-Dependent Document Retrieval

Rolf Grütter and Bettina Waldvogel
Swiss Federal Research Institute WSL, Zürcherstrasse 111, 8903 Birmensdorf, Switzerland

1 INTRODUCTION

In this paper different techniques for location-dependent retrieval of text documents are compared (cf. figure 1). The purpose of the comparison is to answer the question, which technique is best suited for what kind of search. In order to serve this purpose, a set of three use cases is computed.

The database used in the study is the Datacenter Nature and Landscape (DNL) of the Swiss Federal Research Institute WSL. The DNL stores more than 30,000 text documents about biotopes and protected landscapes, attributive data, metadata and GIS perimeters (Bauer-Messmer, 2009).

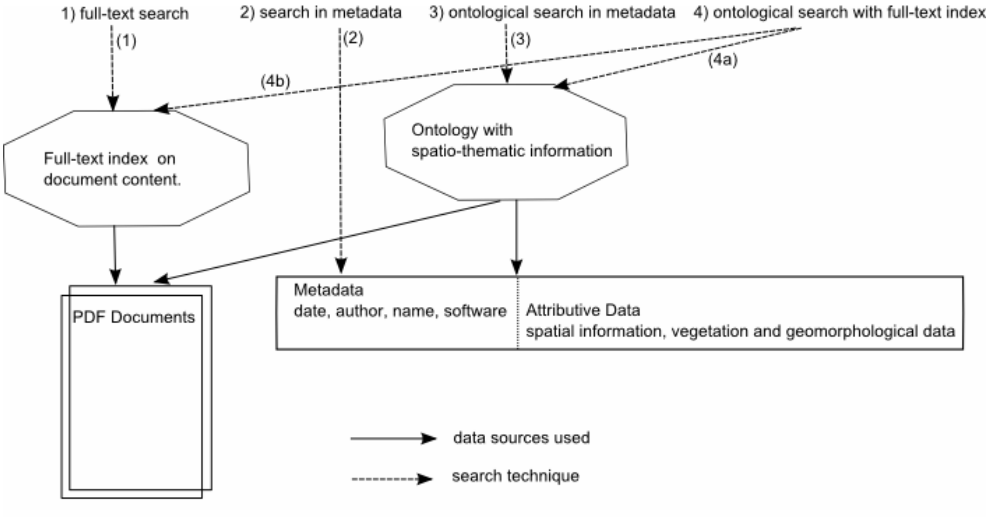

1) full-text search
2) search in metadata
3) ontological search in metadata
4) ontological search with full-text index
(1)
(2)
(3)
(4a)
(4b)
Full-text index on document content.
Ontology with spatio-thematic information
PDF Documents
Metadata
date, author, name, software
Attributive Data
spatial information, vegetation and geomorphological data
data sources used
search technique

Figure 1: Overview of the search techniques and the data sources they use.

2 SEARCH TECHNIQUES

The search techniques compared are:

1. String search in fully indexed text data (i.e. text documents), referred to as *full-text search*;
2. string search in metadata columns of data tables, referred to as *metadata search*;
3. search in metadata columns of data tables using terms of a bilingual vocabulary controlled by an ontology, referred to as *semantic search*;
4. a combination of full-text search (1) with semantic search (3), referred to as *combined search*.

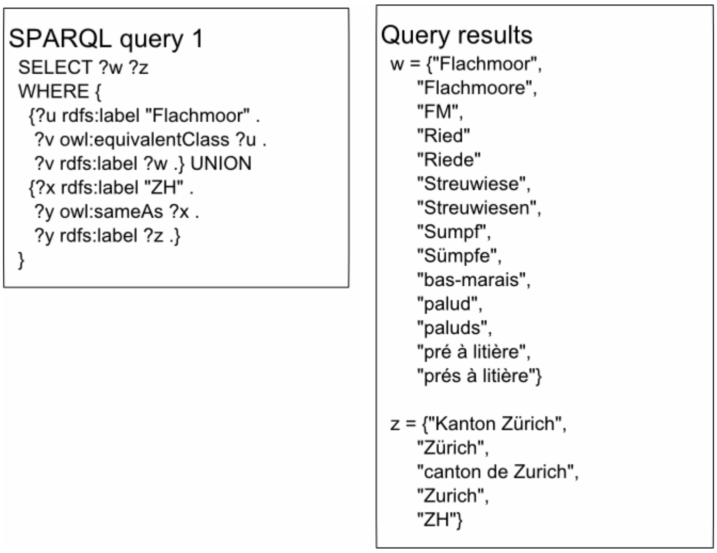
For the full-text search a multi-language `CONTEXT` index was created using Oracle Text (version 10.2.0.3).[1] The metadata search makes use of the SQL `LIKE` operator. The semantic search makes use of Pellet 2.0[2] to preprocess queries in an OWL DL knowledge base (Patel-Schneider, 2004) and of the SQL `LIKE` operator to search in the database.

3 USE CASES

We consider three cases of typical retrieval problems encountered by DNL users:

1. Inventory objects within an administrative region;
2. Wildlife refuges for a rare species;
3. Completing an incomplete query in search of an object.

The retrieval problems in use cases 1 and 2 are: “Retrieve for all objects of the inventory of Fens in the Canton of Zurich all records holding a text document” and “Retrieve for all objects of the inventory of Moorlands, which are (known) reservoirs for the dragonfly *Arktische Smaragdlibelle*, all records holding a text document.” These problems are put in terms of the search strings “Flachmoor” and “ZH”, respectively “bas-marais” and “ZH”, and “Moorlandschaft” and “Arktische Smaragdlibelle”, respectively “paysage de marais” and “Cordulie arctique”, in a Googlized fashion. In order to make sure that only records holding a document (and not a GIS geometry) are returned, the search space is restricted to records containing a document, which are identified by a flag (this also applies to use case 3).



SPARQL query 1

```
SELECT ?w ?z
WHERE {
  {?u rdfs:label "Flachmoor" .
   ?v owl:equivalentClass ?u .
   ?v rdfs:label ?w .} UNION
  {?x rdfs:label "ZH" .
   ?y owl:sameAs ?x .
   ?y rdfs:label ?z .}
}
```

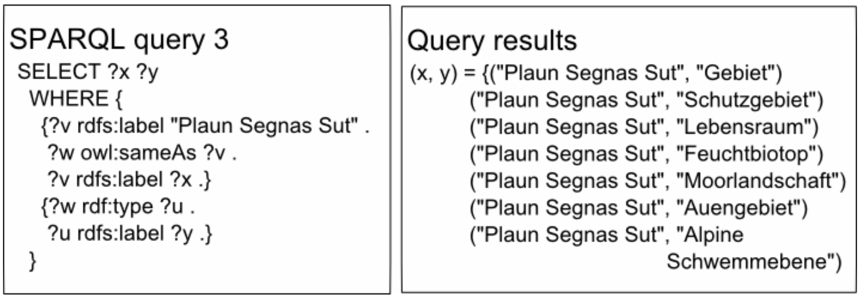
Query results

```
w = {"Flachmoor",
     "Flachmoore",
     "FM",
     "Ried"
     "Riede"
     "Streuwiese",
     "Streuwiesen",
     "Sumpf",
     "Sümpfe",
     "bas-marais",
     "palud",
     "paluds",
     "pré à litière",
     "prés à litière"}

z = {"Kanton Zürich",
     "Zürich",
     "canton de Zurich",
     "Zurich",
     "ZH"}
```

Figure 2: SPARQL query and query results of use case 1.

[1] http://www.oracle.com/pls/db102/homepage
[2] http://clarkparsia.com/pellet



SPARQL query 3

```
SELECT ?x ?y
  WHERE {
    {?v rdfs:label "Plaun Segnas Sut" .
     ?w owl:sameAs ?v .
     ?v rdfs:label ?x .}
    {?w rdf:type ?u .
     ?u rdfs:label ?y .}
  }
```

Query results

```
(x, y) = {("Plaun Segnas Sut", "Gebiet")
          ("Plaun Segnas Sut", "Schutzgebiet")
          ("Plaun Segnas Sut", "Lebensraum")
          ("Plaun Segnas Sut", "Feuchtbiotop")
          ("Plaun Segnas Sut", "Moorlandschaft")
          ("Plaun Segnas Sut", "Auengebiet")
          ("Plaun Segnas Sut", "Alpine
                                Schwemmebene")
```

Figure 3: SPARQL query and query results of use case 3.

In the Oracle Text query used to solve the retrieval problem by the full-text search the strings are related to each other by an `AND` operator. In the `WHERE` clause of the SQL statement used to solve the retrieval problem by the metadata search the strings are fed into an expression in CNF without consideration of synonyms, similar terms, translated terms and abbreviations. The SPARQL query (Prud'hommeaux, 2008) and the results from query preprocessing used to solve the retrieval problem by the semantic and the combined searches in use case 1 are shown in figure 2.

The retrieval problem in use case 3 is: "Retrieve for the moorland *Plaun Segnas Sut* all records holding a text document." This problem is put in terms of the search string "Plaun Segnas Sut" in a Googlized fashion, assuming that *Plaun Segnas Sut* is a moorland.

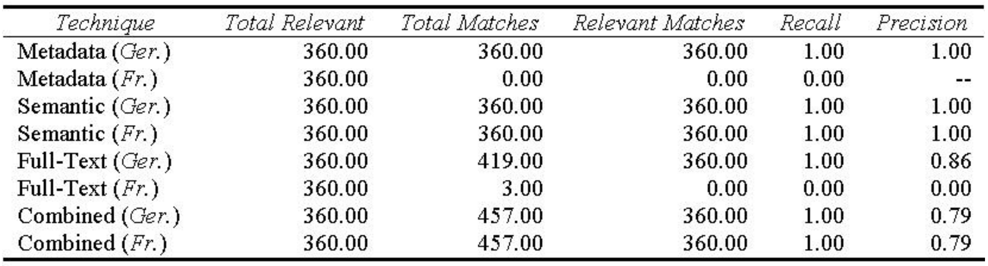
The Oracle Text query and the SQL statement used to solve the retrieval problem by the full-text and the metadata searches are constructed as described for use cases 1 and 2. The SPARQL query and some of the results from query preprocessing used to solve the retrieval problem by the semantic and the combined searches are shown in figure 3.

4 EVALUATION

The number of relevant documents in a result set for the calculation of recall and precision is found by manual counting. For the calculation of *precision*, the count is related to the total number of documents returned. For the calculation of *recall*, the count is related to the total number of relevant documents in the database. For the retrieval problems in use cases 1 and 2 the total numbers of relevant documents are found by spatially selecting overlapping polygons in a Geographic Information System (GIS). Those documents that can be attributed via specific process types to the identified objects are then counted in the database using SQL statements. For the retrieval problem in use case 3 only the latter is performed.

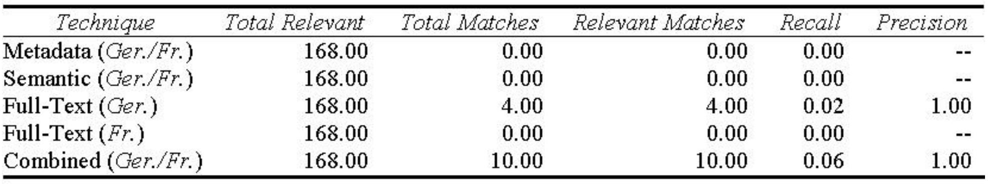
Tables 1–3 show the results of the searches performed in use cases 1–3. The entries in the first columns refer to the search techniques introduced in section 2. The abbreviations *Ger.* and *Fr.* stand for searches using German and French terms.

Table 1: Results of searches for documents about fens in the Canton of Zurich



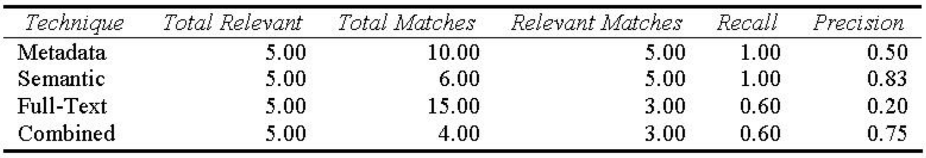
Technique	*Total Relevant*	*Total Matches*	*Relevant Matches*	*Recall*	*Precision*
Metadata (*Ger.*)	360.00	360.00	360.00	1.00	1.00
Metadata (*Fr.*)	360.00	0.00	0.00	0.00	--
Semantic (*Ger.*)	360.00	360.00	360.00	1.00	1.00
Semantic (*Fr.*)	360.00	360.00	360.00	1.00	1.00
Full-Text (*Ger.*)	360.00	419.00	360.00	1.00	0.86
Full-Text (*Fr.*)	360.00	3.00	0.00	0.00	0.00
Combined (*Ger.*)	360.00	457.00	360.00	1.00	0.79
Combined (*Fr.*)	360.00	457.00	360.00	1.00	0.79

Table 2: Results of searches for documents about moorlands with *Somatochlora arctica* occurrences



Technique	*Total Relevant*	*Total Matches*	*Relevant Matches*	*Recall*	*Precision*
Metadata (*Ger./Fr.*)	168.00	0.00	0.00	0.00	--
Semantic (*Ger./Fr.*)	168.00	0.00	0.00	0.00	--
Full-Text (*Ger.*)	168.00	4.00	4.00	0.02	1.00
Full-Text (*Fr.*)	168.00	0.00	0.00	0.00	--
Combined (*Ger./Fr.*)	168.00	10.00	10.00	0.06	1.00

Table 3: Results of searches for documents about the moorland *Plaun Segnas Sut*



Technique	*Total Relevant*	*Total Matches*	*Relevant Matches*	*Recall*	*Precision*
Metadata	5.00	10.00	5.00	1.00	0.50
Semantic	5.00	6.00	5.00	1.00	0.83
Full-Text	5.00	15.00	3.00	0.60	0.20
Combined	5.00	4.00	3.00	0.60	0.75

5 DISCUSSION

Recall and precision of searches for documents about fens in the Canton of Zurich using either German or French terms, the latter in combination with semantic preprocessing, are high (cf. table 1). The reason why the metadata search does not return any results, when using French terms without preprocessing, is that the metadata fields of objects in the inventory of fens do not contain any descriptions in French.

Recall and precision of searches for documents about moorlands with *Somatochlora arctica* occurrences are either zero or undefined except for the full-text search using German terms and for the combined searches using either German or French terms (cf. table 2). Metadata fields of objects in the inventory of moorlands do not contain any information about animal species. Note that the combined searches are not as weak as they seem. They provide information about 10 out of 33 moorlands with *Somatochlora arctica* occurrences. This means that in a subsequent search 30 percent of the 168 relevant documents could be retrieved. Still, all 168 relevant documents are found only after spatially selecting the relevant objects in a GIS.

Recall of all searches for documents about the moorland *Plaun Segnas Sut* and precision of searches with semantic preprocessing are high (cf. table 3). Similar to use case 1, semantic preprocessing improves the searches, although to a more moderate extent. The reason why two relevant documents are not found by the full-text and combined searches is that they are scans and as such were not considered during index creation.

6 CONCLUSION

The following conclusions can be drawn from the comparison:

- Semantic preprocessing tends to improve both metadata and full-text searches as well as searches using incomplete queries. The effect is particularly clear-cut if (i) the metadata and the text documents are not written in the same language as the search terms and (ii) if names used as search terms are not unique.
- In searches for attributes, which are not essential (such as animal species), selecting the relevant biotopes spatially in a GIS before searching the respective text documents outperforms all search techniques compared.
- Recording metadata and searching in metadata is particularly important if (parts of) the documents to retrieve are not textually indexed.

Acknowledgements

This research has been funded by the Swiss Federal Office for the Environment (FOEN).

BIBLIOGRAPHY

Bauer-Messmer, B. et al., The Data Centre Nature and Landscape (DNL): Service Oriented Architecture, Metadata Standards and Semantic Technologies in an Environmental Information System. In V. Wohlgemuth et al., (eds.). Environmental Informatics and Industrial Environmental Protection. Shaker Verlag, Aachen: 101–112, 2009.

Patel-Schneider, Peter F., Hayes, P., Horrocks I., OWL Web Ontology Language: Semantics and Abstract Syntax. W3C Recommendation 10 February 2004. WWW document, http://www.w3.org/TR/2004/REC-owl-semantics-20040210/

Prud'hommeaux, E., Seaborne, A., SPARQL Query Language for RDF. W3C Recommendation 15 January 2008. WWW document, http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/