University of Zurich
Department of Informatics

Dynamic and Distributed Information Systems

Master Thesis August 22, 2012

Tagging Methods for Linked Media Data

David Oggier
of Vevey VD, Switzerland

Student-ID: 05-219-209
d_oggier@hotmail.com

Advisor: Thomas Scharrenbach, PhD

Prof. Abraham Bernstein, PhD
Department of Informatics
University of Zurich
http://www.ifi.uzh.ch/ddis

Acknowledgements

I would like to thank Professor Abraham Bernstein for the opportunity to write my Master's Thesis with the DDIS group of the University of Zurich. I am also very thankful to my advisor Thomas Scharrenbach for his great support and advice and to Jörg-Uwe Kietz for filling in during his absence. Many thanks to the whole IT-OSE team at Hessischer Rundfunk for their help and in particular to Patric Kabus and Haiko Emmel for making this thesis possible. Last but not least, I would like to thank my family and friends for supporting me. Special thanks to my proofreaders Amir and Christoph.

Abstract

In this thesis, a method is presented to tag the media metadata of a broadcasting company with Linked Data concepts. Specifically, a controlled vocabulary in the form of a thesaurus is used as an intermediary between broadcast metadata and Linked Data vocabularies. A method to link this metadata with appropriate thesaurus entries, as well as an algorithm to align the latter with Linked Data concepts are presented and evaluated. Furthermore, it is investigated whether a benefit is gained for user queries by applying faceted search on the resulting semantically enhanced data.

Zusammenfassung

In dieser Arbeit wird ein Vorgehen um Medien-Metadaten einer Rundfunkanstalt mit Linked Data-Konzepten zu verlinken vorgestellt. Dabei wird ein kontrolliertes Vokabular in der Form eines Thesaurus verwendet, um die Verbindung zwischen Inhaltsbeschreibungen von Beiträgen und Linked Data-Datenbeständen herzustellen. Das Vorgehen beinhaltet eine Methode um diese Inhaltbeschreibungen mit Thesaurus-Konzepten zu versehen, sowie einen Algorithmus welcher diese Thesaurus-Konzepte mit Linked Data-Ressourcen verknüpft. Des Weiteren wird untersucht, wie diese Verknüpfung von Inhaltsbeschreibungen mit semantischen Daten durch eine facettierte Suche genutzt werden kann, und ob diese gegenüber herkömmlichen Suchmethoden vorteilhaft ist.

List of Figures

Figure 1: Semantic Web standards stack (Obitko 2007) 7

Figure 2: Example RDF graph 8

Figure 3: The Linked Data cloud 16

Figure 4: Screenshot of Faceted Wikipedia Search 18

Figure 5: The FESADdigital search form 27

Figure 6: The FESADred search page 28

Figure 7: Overview of the implementation steps 31

Figure 8: Workflow of the matching for a particular thesaurus-DBpedia concept pair 35

Figure 9: Screenshot of the faceted search client 41

Figure 10: Aggregate precision values 49

Figure 11: Aggregate recall values 49

List of Tables

Table 1: General alignment statistics 45

Table 2: Frequencies of match types 45

Table 3: Alignment quality results 46

Table 4: Keyword extraction quality results 50

List of Listings

Listing 1: Sample RDF/XML statement 8

Listing 2: Sample RDF/N3 statement 9

Listing 3: Sample SPARQL query 11

Listing 4: Sample SKOS concept definition 13

Contents

Acknowledgements ii
Abstract iii
Zusammenfassung iv
List of Figures v
List of Tables vi
List of Listings vii
Contents viii
1. Introduction 1
1.1 Motivation 1
1.2 Thesis Goals 2
1.3 Thesis Outline 3
2. Background 5
2.1 The Semantic Web 5
2.1.1 Semantic Web Goals 5
2.1.2 Resource Description Framework 6
2.1.3 RDF Schema 10
2.1.4 SPARQL 10
2.1.5 Web Ontology Language 11
2.1.6 Simple Knowledge Organization System 13
2.2 Linked Data 14
2.2.1 Linked Data Principles 14
2.2.2 The Linked Data Cloud 15
2.2.3 Faceted Search 17
2.3 Ontology Learning and Matching 19
2.3.1 Ontology Learning 20
2.3.2 Ontology Alignment 21
3. Linking Media Metadata 23
3.1 Media Metadata 23

3.1.1 Metadata Generation and Consumption Lifecycle 23
3.1.2 Content Description Metadata 25
3.1.3 Current FESAD Access Points 26
3.2 Implementation Details 29
3.2.1 Implementation Overview 29
3.2.2 Thesaurus Conversion 32
3.2.3 Selection of Linked Data Vocabulary 33
3.2.4 Thesaurus Alignment 34
3.2.5 Metadata Conversion 38
3.2.6 Tagging of Media Data 38
3.2.7 Faceted Search Client 39
4. Evaluation 43
4.1 Alignment Evaluation 43
4.1.1 Evaluation Methods 43
4.1.2 Alignment Statistics 44
4.1.3 Alignment Quality 46
4.2 Keyword Extraction Evaluation 47
4.2.1 Evaluation Method 48
4.2.2 Results 48
4.3 Faceted Search Evaluation 50
4.3.1 Evaluation Method 50
4.3.2 Questionnaire 51
4.3.3 User Interview Results 52
5. Related Work 55
5.1 Keyword Extraction Systems 55
5.2 Ontology Alignment Tools 56
5.3 Faceted Search Systems 57
5.4 Creation of Linked Data 58
6. Discussion 60
6.1 Implementation 60
6.1.1 Thesaurus Alignment 60
6.1.2 Keyword Extraction 62
6.1.3 Faceted Search Client 62
6.1.4 Linking Method Discussion 62

6.2 Suitability of Faceted Search 64
7. Conclusion and Outlook 65
8. Bibliography 67
Appendix 70
A.1 CD-ROM Content 70
A.2 Keyword Extraction Evaluation Results 71
A.2.1 Broadcasts 71
A.2.2 Evaluation Results 74
A.3 Thesaurus Alignment Evaluation Results 76
A.3.1 General Thesaurus Alignment Evaluation 76
A.3.2 Geography Thesaurus Alignment Evaluation 81
A.4 User Interview Transcripts 85

1. Introduction

This thesis presents a method to tag the media metadata of a German broadcasting company with Linked Data concepts and investigates the resulting benefits for user queries. In this chapter, the purpose of this approach is explained and specifically why it is expected to be advantageous. Then, the goals of the thesis are presented and its structure is outlined.

1.1 Motivation

A large amount of the content most broadcasting companies show on their program is created by them themselves, usually even on a daily basis. However, at least equally as often is old content reused. For instance, an editor might need footage of a particular bridge for a news report in which it is mentioned. In order to do that, the editor needs to be able to retrieve relevant content quickly, even without ex-ante knowledge of what specific broadcast to look for. This in turn means that accurate metadata, as well as appropriate search facilities, are necessary.

At the broadcasting company that is the subject of this thesis, the broadcast metadata consists, in more or less detail, of freely assigned tags and natural language descriptions. This means that for a specific query to return relevant results, two conditions must be met: First, that the metadata of relevant broadcasts contains the particular search term. Without an enforced vocabulary to regulate broadcast descriptions, it is for example possible that it only contains the word “Burma” and can thus not be found in a query using this country’s other name ”Myanmar”. The second necessary condition to retrieve relevant results is that the search term is not a word whose meaning depends on the context, i.e., a homograph. An often cited example for a homograph is the word “jaguar”. In that case, there would be no way for the query system to know about the search term’s ambiguity, let alone to distinguish the different possible interpretations, which could result in irrelevant search results.

In Linked Data (Bizer et al. 2009), terms can be given a machine-interpretable meaning. For a concept “Burma” for instance, a computer would be able to understand that it is a country and that another

name for it is “Myanmar”. Also, assuming there was another concept named “Burma”, for example a restaurant, a computer would be able to tell them apart because their labels would not be the only information identifying them.

Given this context, it is to be examined if the broadcast metadata can be equipped with such information background in order to provide it with semantics, and also if the result of this process would bear any benefits to user queries against this metadata.

1.2 Thesis Goals

Having media metadata serialized according to Semantic Web principles is in itself not of great use, as the corresponding standards provide just another way to represent data in a well-structured fashion. Where it really gets interesting, is when this format is utilized to increase the data’s information content by linking it to external data sources. The expected benefit of Linked Data over data that is not semantically enhanced was briefly described in the previous section. Linked Data standards provide a method to publish and exchange data with machine-comprehensible semantics. This makes it possible to extend information from one dataset with information from external ones using standard vocabularies. In this thesis, it shall be verified whether this benefit can also be harnessed with semantically enriched media metadata in the context of user queries.

The currently used search facilities provide a great number of options that allow formulating queries with great accuracy. However, they cannot overcome the fact that these queries are based on free-text descriptions and an uncontrolled set of tags. While this is not necessarily a problem when performing a directed search for a specific broadcast, it might be for an exploratory, content-oriented one. This is not only due to possible semantic ambiguities, but also to imprecisely or insufficiently described content. The former can result in a loss of precision, where irrelevant results are retrieved, while the latter may cause poor recall because relevant results cannot be found. This problem is partly addressed at the broadcaster in question, Hessischer Rundfunk, by the use of tagging conventions, which supposedly enables a common understanding of how broadcasts are tagged. The problem with this approach is that it assumes some ex-ante knowledge of these agreed-upon terms, which creates an entry barrier to efficient searching. Furthermore, complex queries cannot be performed over full-text descriptions.

Based on these explanations, the goals of this thesis are the following:

1. to develop a method to interlink media metadata from Hessischer Rundfunk with existing Linked Data using recommended Semantic Web standards,
2. to investigate how this interlinking can be harnessed to perform faceted searches on the media metadata, and
3. to test whether a benefit is gained over traditional search methods by performing queries with a faceted browser, and if such a benefit would address a current need.

The first two goals are addressed by an implementation. In order to achieve the last goal, a practical setting is required as the basis for the tests, although Linked Data could conceivably be advantageous in many applications. In the scope of this thesis, potential benefits for a particular use case involving queries performed by editors will be examined. Such queries frequently involve exploratory queries, for example to find stock footage of a snowy landscape. In this context, and to address Linked Data's strong suit, the two test cases to be supported are queries for an ambiguous term, such as "jaguar", and complex, multi-stage queries.

1.3 Thesis Outline

The rest of this thesis is organized as follows: in Chapter 2, the necessary theoretical background is introduced, which includes an explanation of the general ideas behind the Semantic Web and of its main standards, as well as the introduction of some more specific concepts that are of relevance for this thesis. In particular, Linked Data and ontology alignment are presented.

Chapter 3 first exposes the current situation in terms of media metadata and its utilization, and then cuts to the chase and introduces the proposed solution. All of the implementation's components for interlinking broadcast metadata with Linked Data are explained in individual sections, as well as a proposed faceted search client.

In Chapter 4, an evaluation tests how well the aforementioned goals were achieved by the implementation. Also, the results from user interviews that were conducted to evaluate the faceted search paradigm are presented.

Related work in the domains of keyword extraction, ontology alignment, and faceted search is reviewed in Chapter 5. Furthermore, some projects to create Linked Data are presented in it.

Chapter 6 contains an assessment of the implementation based on the evaluation results, as well as a discussion on the suitability of faceted search in the examined use case.

Finally, Chapter 7 offers a conclusion to the thesis and an outlook on future work.

2. Background

This chapter introduces the theoretical concepts underlying this thesis. From an initially general introduction into the Semantic Web and the closely related Linked Data initiative, the discussion will then go on to focus on ontologies and in particular on the matching of distinct ontologies. The descriptions focus mainly on aspects relevant to the context of this thesis.

2.1 The Semantic Web

The term "Semantic Web" was originally coined in the seminal paper by Berners-Lee et al. (2001). It envisions "an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." The Semantic Web has come a long way since its inception, introducing a number of widely used standards and allowing the interlinking of a large amount of information with machine-readable semantics. These will be introduced after a more thorough look at the Semantic Web initiative's goals.

2.1.1 Semantic Web Goals

The goal of the Semantic Web initiative is to integrate the information present in the World Wide Web by providing standards to structure the underlying data and to interlink it. As opposed to the traditional Web, where content is mainly tailored to human consumption, the Semantic Web also allows computers to access and make use of the meaning of information. It thus introduces a paradigm shift, where not only the layout details of documents are published in a standardized format, but also the meaning of the information contained in them. Consequently, the Semantic Web initiative does not suggest a substitution of current Web standards, but rather an extension of them, allowing this new technology to be fully backward compatible.

The advantages of structuring data and making it computer-readable are numerous. Interpretable information potentially allows retrieving documents with much better precision and recall because the indexing of Web pages involves less ambiguities, and queries become less dependent on the

vocabulary used in the relevant documents. Also, information can be exchanged between agents and reused with much more ease, and thus be integrated across any number of different sources, making complex queries across multiple datasets possible.

For a more concrete illustration of the Semantic Web's value proposition, the limitations of the traditional Web are frequently referred to. In particular, the task of retrieving a relevant piece of information in the ocean of data that the Web represents has been a major challenge ever since its beginnings. The source of the problem lies in the fact that software needs to analyze natural language documents and interpret their semantics in order to index them accurately. However, free text by definition lacks a standardized structure, and thus explicit indications as to which parts of it are most essential and what the meaning of terms are.[1] An often mentioned example is the homograph "jaguar" and a search engine's supposed inability to tell queries for the animal and the car brand apart.

While current search engines have nonetheless managed to implement efficient algorithms to retrieve relevant documents, one other aspect they still struggle with is complex queries. For instance, a user planning a vacation might be interested in hotels located in a particular ski resort, with the additional conditions that it be equipped with an indoor swimming pool, offers rooms within a certain price range, and is situated no more than 1 km from the closest ski slope. Performing such a search on a traditional search engine can be rather cumbersome, as it would probably span multiple individual queries, perhaps even in different languages and on different platforms.

In contrast, an application would be able to easily integrate the corresponding information and allow the user to find a suitable match in a few clicks by applying Semantic Web standards. Some of the more important ones of these standards will be introduced in the following sections.

2.1.2 Resource Description Framework

The Semantic Web initiative is regulated by the World Wide Web Consortium (W3C), which provides a framework of standard data formats and promotes their use by the means of so-called recommendations.[2] With regard to the Semantic Web, these standards are commonly represented as a stack, such as illustrated in Figure 1.

[1] For the sake of the argument, the structured elements of Web documents, such as title, headings, or keywords, are disregarded here.
[2] http://www.w3.org

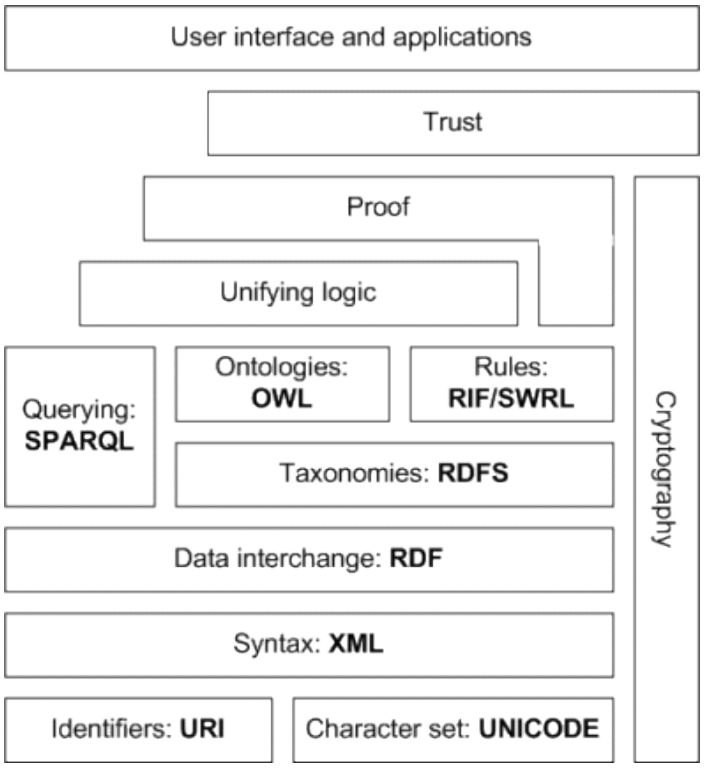

Figure 1 : Semantic Web standards stack (Obitko 2007)

For the syntactical representation of data, W3C proposes the use of XML (eXtensible Markup Language), which is a very well-suited format to structure data in a customizable way.[3] XML does however not prescribe a standardized approach to nesting tags, and can thus not be shared across applications without previously explicitly agreeing on a format.

Consequently, RDF (Resource Description Framework) is used on top of XML.[4] This format is able to convey information on so-called "resources", entities representing any concrete object or person, as well as abstract concepts, such as organizations or scientific disciplines. A resource can be identified by an URI (Unique Resource Identifier), typically an HTTP address that can often also be dereferenced and which is always unique for that particular resource. This enables reusing it at will.

The RDF Data Model

While independent from any particular syntax, RDF data is always modeled in triples, which are positive statements consisting of a set of three resources: a subject, which is the entity described in that statement, a predicate, representing a property of that subject, and an object, the value

[3] http://www.w3.org/XML/
[4] http://www.w3.org/RDF/

corresponding to this property. Such a triple can also be represented as an RDF graph, as depicted in Figure 2, which shows a statement expressing that the broadcast identified by the URI `http://www.hr-online.de/bc_id1234` has the title "hessenschau" and was produced by an employee identified as "`empl_id86`".

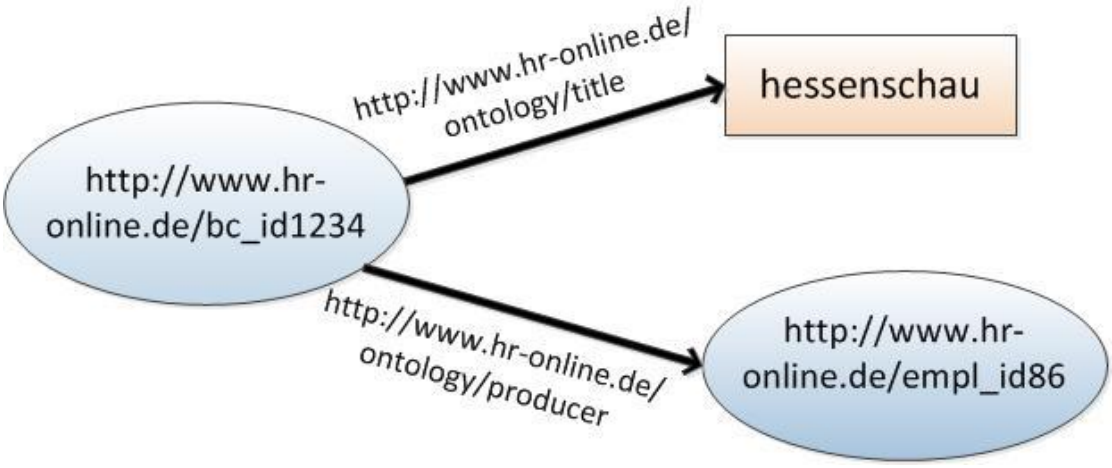

Figure 2: Example RDF graph

This example illustrates a few important aspects of RDF. For one, the object and predicate can only be described with an URI. In the RDF graph model, these are referred to as nodes. The object may however also be a literal, commonly represented as a rectangular box in RDF graphs. The illustration also shows how multiple triples with the same subject can be represented.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:hr="http://www.hr-online.de/ontology>
      <rdf:Description about="http://www.hr-online.de/bc_id1234">
            <hr:title>hessenschau</hr:title>
            <hr:producer rdf:resource="http://www.hr-online.de/empl_id86"/>
      </rdf:Description>
</rdf:RDF>
```

Listing 1: Sample RDF/XML statement

Regarding this aspect, the same principle applies when the RDF statement is serialized with the XML syntax. The code in Listing 1 is an equivalent statement written in RDF/XML.[5] As this example shows, multiple statements about a single resource can be nested within an `<rdf:Description>` tag, which itself declares the subject resource in an `about` attribute. While literals are represented as regular XML text content between two tags, which can optionally be asserted to existing XML data types, object resources are found in `rdf:resource` attributes of corresponding predicate tags.

Further RDF Features

A special feature of RDF is the use of blank nodes. These are anonymous nodes without URI references, and are used to group multiple related statements, such as information on how and where the broadcast was recorded. As they bear no importance to this thesis, blank nodes and other RDF characteristics will not be further elaborated on at this point.

An additional aspect illustrated by Listing 1 is the use of XML namespaces, which allows conveniently abbreviating long URI references. Namespaces are declared in the RDF/XML root element `<rdf:RDF>` and assigned to a prefix, such as "`hr`". These can be assigned freely, although there exist quite a few commonly used prefixes such as "`rdf`", which should not be overwritten. This makes for a more condensed and readable code while still being semantically equivalent to using full URIs.

As previously mentioned, RDF is independent from any serialization format. While XML is the only official recommendation, its frequent use of XML attributes makes it difficult to read for humans. That is why the N-Triple format is also often used. [6] It consists of simple one-line statements for each triple with URIs written between angle brackets and literals in quotation marks, as Listing 2 demonstrates. As shown in this statement, which is equivalent to the one in Listing 1, leading subjects do not need to be repeated. Evidently, namespaces can also be declared in the N-Triple notation, making it even more readable.

```
<http://www.hr-online.de/bc_id1234>
  <http://www.hr-online.de/ontology/title> "hessenschau".
  <http://www.hr-online.de/ontology/producer> <http://www.hr-online.de/empl_id86>.
```

Listing 2: Sample RDF/N3 statement

[5] http://www.w3.org/TR/rdf-syntax-grammar/
[6] http://www.w3.org/2001/sw/RDFCore/ntriples/

2.1.3 RDF Schema

With RDF, statements about resources can be made in a standardized and unambiguous way. However, the exact nature of these resources, and consequentially the predicates that are suited to describe them, are not explicitly stated. That is why a model must be defined on top of RDF in order to make a statement semantically unequivocal. RDF Schema (RDFS) does precisely that.[7] While RDF can be used to describe particular resources, RDFS allows describing what type of resources these are, how they are linked to other resource types, and what predicates are admissible to describe them. In short, RDFS is used for RDF vocabulary description. As a semantic extension of RDF, it has the same underlying data model, the triple.

Classes and properties are the main constructs in RDFS. They themselves are resources that can be identified by URI references. A resource can be declared to be an instance of a class by assigning the predicate `rdf:type` to it with the corresponding class as object. The class itself is declared as such by being an instance of `rdfs:Class`, while a property would be an instance of `rdf:Property`. Furthermore, with `rdfs:range` and `rdfs:domain` the application of a predicate can be restricted in terms of what subjects or objects are admissible. Finally, classes and properties can transitively be subclasses and subproperties, respectively, of other resources of the same type.

Putting all these elements together, RDFS is able to describe complex vocabularies in terms of hierarchies of classes and properties, and can represent the background knowledge for a set of RDF statements. This implicit information can thus be used to make inferences over RDF data, for instance to deduce that the resource `http://www.hr-online.de/empl_id86` must be of type `Person` because this is the declared range of the `hr:producer` property. Such interpretations, derived from implicit information, define the semantics of RDF statements.

2.1.4 SPARQL

SPARQL (SPARQL Protocol and RDF Query Language) is the W3C recommendation for querying RDF data.[8] Its syntax resembles SQL and features many similar query operators, but is specifically tailored to the triple model (Angles & Gutierrez 2008). It functions by means of triple pattern matching, that is, by looking for a portion of the RDF graph that matches the query expression. The example in Listing 3 illustrates this more concretely.

[7] http://www.w3.org/TR/rdf-schema/
[8] http://www.w3.org/TR/rdf-sparql-query/

```
PREFIX  hr:<http://www.hr-online.de/ontology>
SELECT ?producer
WHERE {
             ?broadcast hr:title “hessenschau” .
             ?broadcast hr:producer ?producer .
      }
```

Listing 3: Sample SPARQL query

This SPARQL query matches resources with the title “hessenschau” and returns the corresponding producer. In the `SELECT` part of the query statement, the set of variables to appear in the results is indicated, while the `WHERE` clause is used for the graph pattern to be matched against. In this case, the variable to be matched is named `?producer`. Variable names have to start with a question mark. The `WHERE` clause expresses the graph pattern this variable must match, specifically to be the object in a `hr:producer` statement of a resource with the value “hessenschau” as title.

SPARQL furthermore features the possibility to declare prefixes in order to shorten statements and several options to define how the result set should be filtered or grouped. The latest version, SPARQL 1.1, also adds a further set of functionalities, such as the aggregation of results, the ability to perform nested subqueries within the `WHERE` clause, and the ability to express negations.[9]

2.1.5 Web Ontology Language

Depending on the application, different levels of expressivity and formalism may be needed to describe a domain. With RDFS, it is possible to define a simple structure for concepts and relationships in a particular area of concern, but only to a rather limited extent. In contrast, OWL (Web Ontology Language) is a much more expressive language that allows a user to describe full-blown ontologies.[10]

An ontology is, according to its most frequently cited description, “an explicit specification of a conceptualization” (Gruber 1993), with a conceptualization being all the objects and their properties and relationships contained in a domain of knowledge. This cannot be expressed in RDFS because ontologies require tools to make much more precise statements. However, the more expressive an ontology language is, the more difficult becomes reasoning, the derivation of implicitly stated facts from a specification. In order to accommodate the different needs with regard to this trade-off, the

[9] http://www.w3.org/TR/sparql11-query/
[10] http://www.w3.org/TR/owl2-overview/

current version of OWL comes in three flavors, so called profiles. These sublanguages are OWL2 EL, OWL2 QL, and OWL2 RL.[11]

The main components are nonetheless very similar to the ones in RDFS. Instances, called individuals in OWL, belong to classes and are defined by properties, where a distinction is made between data type and object properties. The former include relations between individuals and RDF literals or XML schema data types, and the latter relations between two individuals. Where OWL really differs is how these entities can be described. A basic OWL statement is called an axiom, of which there are three kinds. TBox axioms add terminological knowledge with statements about how classes related to each other, RBox axioms define relations between properties and restrictions on them, and ABox axioms assert individuals to classes and pairs of individuals to properties. By means of the logical language elements OWL provides, such as conjunction, disjunction and negation, complex classes can be defined as combinations of atomic classes. Furthermore, a wide array of property restrictions is available in order to declare the relation between two sibling individuals to be symmetrical, or to restrict the cardinality of their parent individuals, for instance.

The benefit of modeling such complex knowledge domains is to make them machine-understandable. Software agents that are able to extract implicit knowledge from ontologies are called reasoners. With complex domain specification, the question of decidability can arise for reasoners, that is, whether there is a method to solve a problem in a finite number of steps. Typical reasoning tasks include subsumption, which checks whether a specific concept is always a logical subset of another set, and satisfiability, which verifies that the description of a concept does not necessarily make it an empty concept. Another important problem for OWL reasoners lies in consistency checking, that is, to determine whether ABox statements are not in contradiction to corresponding TBox statements.

OWL is like RDF format-independent, but the syntax most frequently used and endorsed by the W3C is RDF/XML. Beside its expressivity, a major asset of OWL is its flexibility, which it achieves by means of the open world assumption it makes. This implies that missing information is considered to be unknown instead of false, and that the description of a resource can always easily be extended.

[11] For an overview of the expressions allowed in each sublanguages, refer to http://www.w3.org/TR/2009/REC-owl2-profiles-20091027/

2.1.6 Simple Knowledge Organization System

Towards the opposite end of ontologies on the expressivity scale of vocabulary descriptions are thesauri. A thesaurus is essentially just a list of terms connected with each other by some simple relationships, which reflect the hierarchy, as well as equivalencies and general relatedness between terms. Due to this simplicity, there is no need for an explicit schema to define how concepts can relate to each other, unlike it is the case for RDF or OWL.

SKOS (Simple Knowledge Organization System), a formal syntax to describe thesauri, will be of particular interest to this thesis.[12] This standard is built on RDF triples and can thus be serialized using any RDF format. The SKOS data model describes concepts which are similar to resources in RDF and are also identified by an URI. A concept can be described by lexical labels, of which the SKOS vocabulary features three kinds. A preferred label (`skos:prefLabel`) defines the main term to describe a concept. A language tag may be added to it, and there may be at most preferred label per language. To add further labels to a concept, `skos:altLabel` or `skos:hiddenLabel` can be used. Both allow for a better retrieval of the concept by adding synonyms, but the latter label is only visible to applications performing text-based searches and is otherwise hidden. It would for example be used to include misspelled versions of labels.

SKOS concepts are linked among each other with semantic relationships, the most important of which being `skos:narrower` and `skos:broader`. These are used to express which other concepts are more specific (narrower) or more general (broader) in meaning, respectively. These properties are not transitive. Concepts can also be linked with each other by `skos:related`, indicating that they are generally associated in some way.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:skos="http://www.w3.org/2004/02/skos/core#">
  <skos:Concept rdf:about="http://www.example.org/thesaurus#country">
    <skos:prefLabel xml:lang="en">Country</prefLabel>
    <skos:altLabel>Nation</skos:altLabel>
    <skos:hiddenLabel>Coutry</skos:hiddenLabel>
    <skos:broader rdf:resource="http://www.example.org/thesaurus#country"/>
    <skos:narrower rdf:resource="http://www.example.org/thesaurus#province"/>
    <skos:related rdf:resource="http://www.example.org/thesaurus#republic"/>
  </skos:Concept>
</rdf:RDF>
```

Listing 4: Sample SKOS concept definition

[12] http://www.w3.org/TR/skos-reference/

The example in Listing 4 shows a concept definition for the term country, with an English language label, a synonym, and a misspelled version as hidden label. The description is, as usual in RDF, introduced by namespace definitions. The semantic relationships link directly to the corresponding resources' URIs, since their labels may not be unique.

As shown, SKOS is not anywhere as expressive as OWL, but it is a very suitable and interchangeable format to describe a controlled set of terms and their relations.

2.2 Linked Data

Linked Data is a major initiative driven forward by the W3C in the Semantic Web context (Bizer et al. 2009).[13] It is about publishing data in such a way that it can be easily accessed by other data sources, while itself being informed by external information using Semantic Web standards. This section will introduce the Linked Open Data project's core principles, as well as its current state and main components.[14]

2.2.1 Linked Data Principles

The idea of Linked Data was introduced in (Berners-Lee 2006) and defined as "[…] data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets." To achieve this, the use of Semantic Web standards such as RDF and SPARQL to represent and query data was proposed. Furthermore, four basic rules for publishing linked data were outlined, which are the following:

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
- Include links to other URIs, so that they can discover more things

Data published in accordance to these best practices contribute to a structured web of data that can be traversed across different data sources and namespaces using a common set of standards. It

[13] "Linked Data" is capitalized when the initiative is referred to, as it is a proper noun.
[14] http://linkeddata.org/

essentially suggests using whenever possible resources that can be dereferenced, instead of simple full text labels. As described in (Bizer et al. 2007), a resource can be either one of two things. A distinction is made between information and non-information resources, the former being documents as found on the traditional Web, such as HTML pages or images, and the latter identifying essentially all real-world concepts, like people, places, or physical products. This distinction is of importance because when dereferenced, information resources can directly be returned, while non-information resource lookups need to be redirected by the server to an RDF representation of the resource.

This leads to another important aspect of Linked Data, namely vocabularies. In order to describe resources in a uniform and universally understandable manner, some type of convention is necessary. For this reason, there exist numerous well-known vocabularies to describe all kinds of concepts. For instance, a company Web site adhering to the Linked Data principles would be best advised to use the Dublin Core vocabulary to represent its metadata and the Friend-of-a-Friend (FOAF) convention to describe its staff, instead of using terms under its own namespace.[15] [16] For instance, widely used properties from these vocabularies include `dc:title` and `foaf:primaryTopic` to describe a document, or `foaf:name`, `foaf:mbox`, and `foaf:homepage` to describe a person. This enhances the reusability of the represented data by a large degree. It is common practice to use terms from different vocabularies, and only where needed to define an own terminology.

In summary, Linked Data principles in combination with widely used and known vocabularies provide a well-structured framework in which to publish one's data. However, the real value resides in the data itself, since Linked Data is a prime example for a network effect coming into play.

2.2.2 The Linked Data Cloud

An ever-increasing amount of data is published and accessible with Linked Data standards. Due to it all being interlinked, Linked Data forms a huge RDF graph, which is why it is commonly represented as a cloud diagram. The illustration of Figure 3 shows the current size of this cloud.[17] The bubbles depict individual data sets and the arrows between them the directionality and weight of their links. As illustrated by the coloring of the bubbles, all types of domains are represented in the Linked Data cloud, such as media, life sciences, publications, or government.

[15] http://dublincore.org/documents/dces/
[16] http://www.foaf-project.org/
[17] As of September 2011.

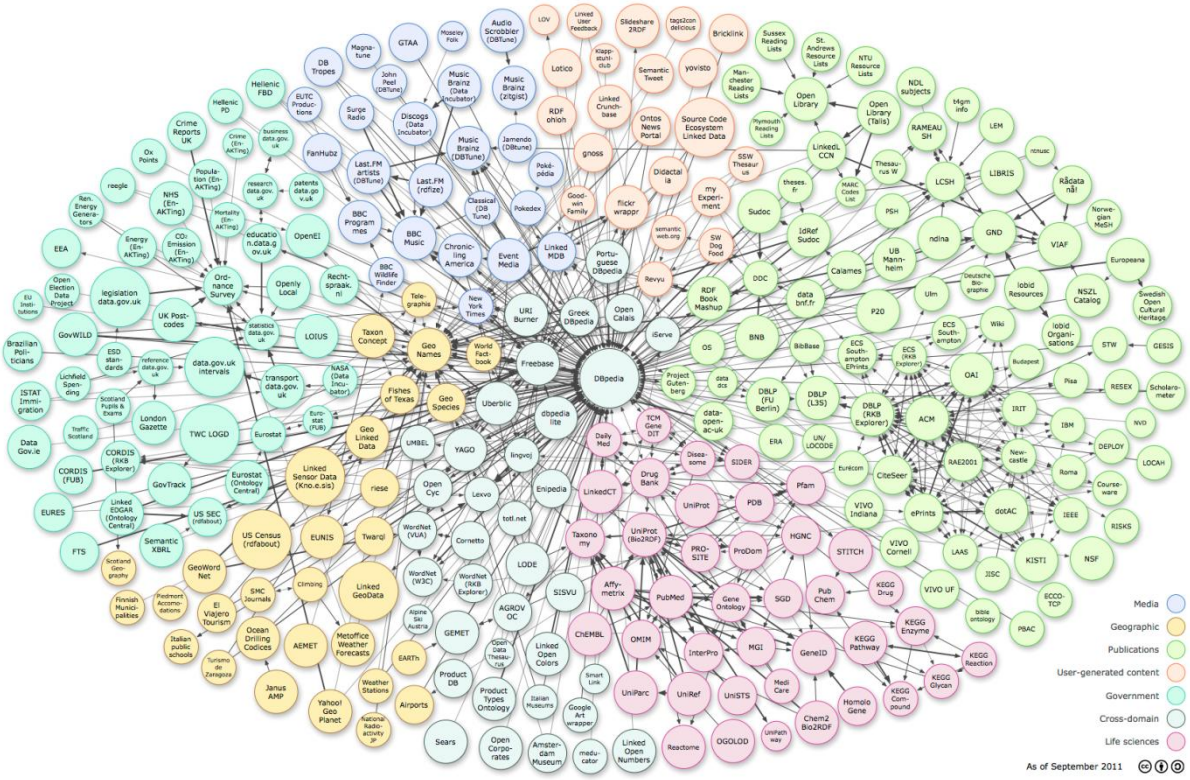

Figure 3: The Linked Data cloud

Another obvious characteristic of the diagram is that, although somewhat interlinked among each other, all the bubbles point to a central dataset, DBpedia.[18] DBpedia contains structured information extracted from the online encyclopedia Wikipedia.[19] More precisely, it features information taken from so-called "infoboxes" on Wikipedia articles, which are HTML tables displaying some key information on the entity described in the article. The reason for this is that they represent the only structured information on Wikipedia and thus allow for an easy extraction. Wikipedia's very broad scope makes it a very suitable choice for the Linked Data centerpiece. According to (Auer et al. 2007), DBpedia currently contains descriptions of more than 1.95 million things, of which for instance over 80'000 are of people and 70'000 of places. The DBpedia dataset is accessible as dereferenceable RDF descriptions via HTTP, through queries over a SPAQL endpoint, and as downloadable RDF dumps, which are serialized in the N-Triple format.[20]

[18] http://dbpedia.org
[19] http://wikipedia.org
[20] http://dbpedia.org/sparql

Unlike DBpedia, Linked Data datasets typically focus on conveying information about specific domains such as art objects (Europeana), geographical features (GeoNames), or proteins (Bio2RDF).[21] [22] [23] Some other datasets will be described in Section 5.4.

2.2.3 Faceted Search

Most Linked Data use cases involve the exchange of information between non-human agents. After all, the essence of semantically enriched data is to be machine-understandable and thus easily usable across applications. However, Linked Data can also benefit the user directly, for instance when formulating queries over it. The common information lookup paradigm on the Web is the matching of search terms with the keywords indexed for documents. But the structure and semantic information inherent to Linked Data allows extending query possibilities in interesting new ways.

Faceted search (also known as faceted browsing or faceted navigation) is an information retrieval paradigm for structured data (Tunkelang 2009). As its name suggests, its central components are so-called facets, which are characteristics of data elements. For Linked Data, this intuitively corresponds to the property-value assertions expressed for resources. Facets make it possible to formulate queries much more precisely. However, in contrast with traditional keyword-based queries, some more input effort is required from the user. A faceted search operation corresponds to a continuous refinement of the result set. It begins with a broadly formulated query expression, often by means of a set of search terms entered by the user. The faceted search system then returns an initial number of results, which may be retrieved using standard search engine techniques, such as inverted indexing. In addition to these results, a set of facets and corresponding facet values are computed and presented to the user in order to allow him to refine the query. The selection of a facet value by the user adds an additional constraint to the query and narrows down the set of results, an operation that can be repeated until a satisfying set of results is left.

[21] http://data.europeana.eu
[22] http://www.geonames.org
[23] http://bio2rdf.org

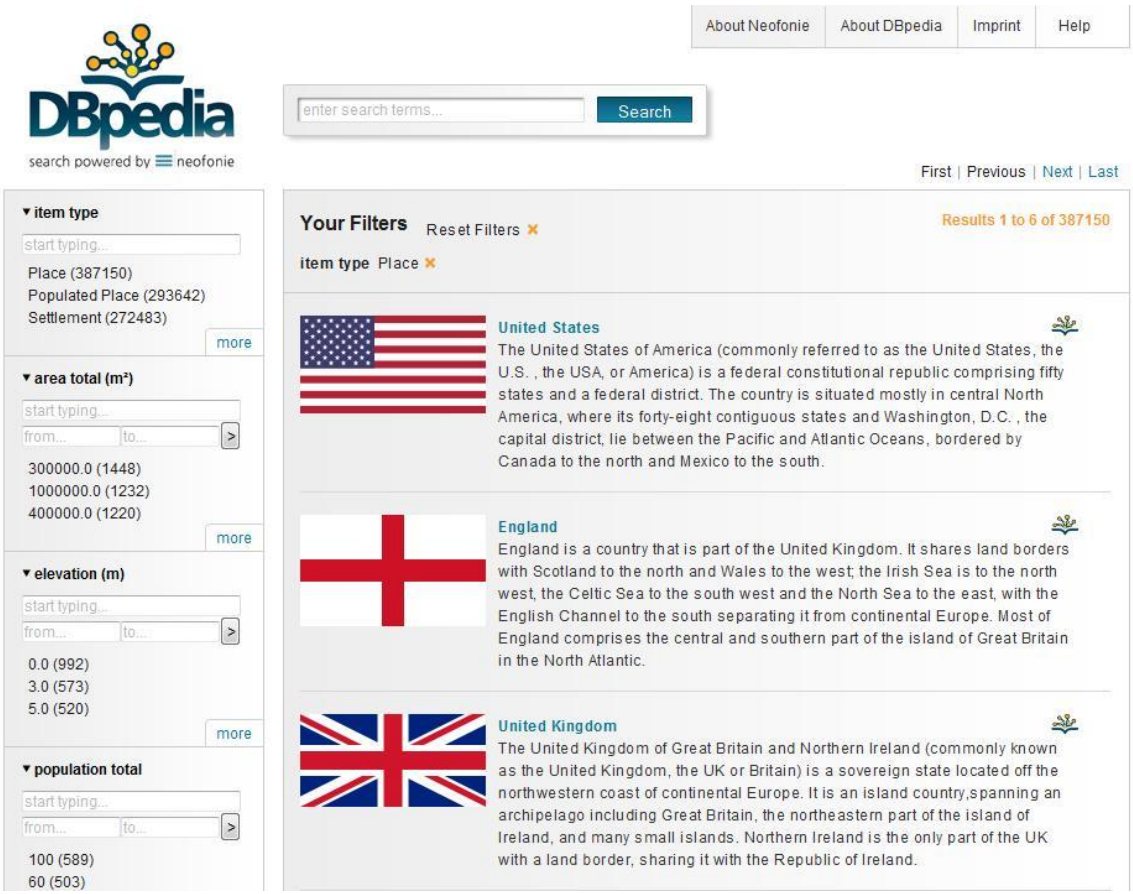

Figure 4: Screenshot of Faceted Wikipedia Search

Facets are computed dynamically as a query is processed. This is necessary because the facets need to correspond to the current result elements' properties, which a static set of facets might not. But since the computed set of facets can be expected to be quite overwhelming in size, especially in the initial stages of a query, only a selection of the most commonly occurring ones can be displayed. A facet usually comes with multiple facet values, which obviously correspond to the value of the property expressed in the facet. In the sample faceted search interface shown in Figure 4, the facet value "Place" was for instance selected for the facet "item type". Facet values are often displayed with a count of the number of result elements satisfying this particular property-value pair.

Precision and Recall

Two important notions in the information retrieval context are precision and recall. Assume a set S of data elements and a set R of relevant data elements for a particular query such that $R \subseteq S$, and that the query retrieves a set of data elements $T \subseteq S$. Precision computes the percentage of relevant data elements in the retrieved ones, or formally

$$Precision = \frac{|(R \cap T)|}{|T|} \quad .$$

The closer the precision value is to 1, the more returned results are relevant to the query. Recall expresses the fraction of relevant data elements that are successfully retrieved, in formal terms

$$Recall = \frac{|(R \cap T)|}{|R|} \quad .$$

A higher recall reflects a higher amount of relevant results that were returned by the query out of all the relevant data elements. The reason why precision and recall are introduced here is because faceted search bears the potential to significantly improve these values. Assuming a set of query results with poor precision, i.e. a large fraction of false positives, a faceted search system offers the possibility to refine it by selecting a suitable facet value, a feature traditional query tools cannot provide. As for recall, all the relevant results should be retrieved, assuming their descriptions are complete and accurate. This condition also holds true for the precision of a query, which can only be as good as the quality of the underlying information.

Faceted Search Examples

Several commercial applications implement the faceted search paradigm. Famous examples include the online retailer Amazon, which allows its users to browse the book catalog by format, author, or language, and the online marketplace eBay, where facets include item categories, price range, or item location. [24] [25] An example for a faceted browsing application over Linked Data is Faceted Wikipedia Search (Hahn et al. 2010) of which a screenshot is shown in Figure 4.[26] It allows browsing DBpedia data using a keyword search and the described resources' properties as facets. A type facet (using `rdf:type` properties) is displayed by default, from which a search can alternatively be started. Further faceted search systems are presented in Section 5.3.

2.3 Ontology Learning and Matching

As introduced previously, ontologies provide a shared understanding of domains and allow people and machines alike to exchange information about them in a concise manner. Yet some questions

[24] http://www.amazon.com
[25] http://www.ebay.com
[26] The faceted search system can be tested at http://dbpedia.neofonie.de/browse.

remain: where does this information come from, and how is it acquired, conceptualized, and formalized? The manual elaboration of an ontology is a daunting and time-consuming task, due to the typically intricate nature of the domain to be represented. To address this difficulty, machine learning techniques have been applied in order to semi-automatically extract information from source documents, a discipline most frequently referred to as “ontology learning”.

A further challenge involving ontologies is to determine equivalencies between a pair of them. The need for this task occurs on the one hand from distinct, yet overlapping ontologies being generated independently from each other, and on the other hand from the necessity to integrate these resources to allow for interoperability between them. The process of finding correspondences between two sets of concepts is called “ontology alignment” or “ontology matching”.

These two disciplines will be discussed in this section, along with some related techniques.

2.3.1 Ontology Learning

In order for the Semantic Web to be a success, mechanisms to generate ontologies cheaply and rapidly need to be available, since they contain the knowledge required for machines to handle meaningful information, as well as the skeleton to represent it. These criteria are not met by the process of manual ontology creation, which requires the involvement of domain specialists and is a very time-consuming and complex task. This phenomenon is commonly recognized as the “knowledge acquisition bottleneck”. For this reason, research has been looking into methods to perform this task in an automatic or semi-automatic manner.

Ontology learning is the task of creating a structured representation of a given domain from some prior knowledge. This is not to be confused with ontology population, which is concerned by the instantiation of an ontology’s concepts. The input for ontology learning can come in various levels of structuration. Highly structured data, which can for instance be extracted from databases, is evidently best suited for translation into an ontology. As the structure is already provided, only the selection of relevant parts of it, as well as a mapping to the target ontology is required for knowledge extraction. However, it is more common that the ontology learning input is not or only partially structured, and consists mainly of natural language content, in which case text mining approaches are used. As pointed out by (Navigli & Velardi 2010), the necessity of implicit knowledge for the understanding of text is a major issue that is difficult to overcome by text mining methods. Since the extraction strategies vary strongly depending on the input format, ontology learning tools typically focus on knowledge bases with either structured or unstructured source.

A tool for the generation of ontologies from textual input is Text2Onto (Cimiano & Völker 2005). It is a framework centered on a so-called “Probabilistic Ontology Model” (POM), which stores the results of different ontology learning algorithms along with confidence measurements. This allows for the tool to let its user decide between different alternatives in the results. After some linguistic preprocessing, different algorithms are run to extract ontology components. For instance, to determine subclass relations between concepts, the hypernym structure of WordNet (Miller 1995) is taken into consideration.

Another approach to learning an ontology from an unstructured input document is implemented in OntoLearn (Navigli & Velardi 2004). It creates an ontology by updating an existing one, by default WordNet, with extracted domain concepts. The terminology is extracted with a linguistic processor and then semantically interpreted using WordNet.

More detailed descriptions and evaluations of ontology learning tools can be found in (Drumond & Girardi 2008) or (Shamsfard & Barforoush 2003).

2.3.2 Ontology Alignment

Closely related to ontology generation from structured knowledge is the discipline of ontology alignment. The distinctive difference however, is that ontology alignment always assumes a very well-structured input, typically serialized in a common ontology language, such as OWL. Also, the primary goal does not lie in the creation of a new ontology, but in the matching of existing ones. More concretely, ontology alignment consists in determining a set of correspondences between the entities (e.g. classes or properties) of two or more distinct ontologies, in order for knowledge expressed with regard to these ontologies to be integrated and consistent. Such an alignment can then, for instance, be used to merge the ontologies, run queries against the integrated knowledge base, or translate data (Shvaiko & Euzenat 2012).

In ontology alignment, a correspondence is a relation between entities from different knowledge representations. Such a relation can for instance declare the entities to be equivalent, disjoint, or to subclass each other. Frequently, a correspondence is also associated with a confidence value to express the likelihood for the relation to be accurate. The implicit assumption when aligning ontologies is that they overlap with regard to the entities they describe. But while the meaning of entities from distinct ontologies may be the same, their description can differ in terms of labeling, as well as in terms of their surrounding taxonomy structure, which is precisely what makes the task of discovering alignments a very challenging one.

Matching strategies

The defining feature of ontology alignment systems rarely lies in the algorithms that are used to actually match ontology entities, but rather in the architecture describing how they are applied and how their results are combined. In fact, the matching strategies proposed by most systems are very similar, and many of them make the same distinction between syntactical and structural matching.

Syntactical matching methods attempt to find similarity in two sets of concepts. Among these methods, edit distance (also known as Levenshtein distance) (Levenshtein 1966) is most frequently cited. The edit distance of two strings expresses their similarity in terms of the minimum number of edit operations required to transform one of them into the other. Another measure of the similarity between two strings is the Jaro-Winkler distance (Winkler 1990), which takes sequences of matching characters into account. Some systems, such as X-SOM or AgreementMaker (which are described in more detail in Section 5.2), also utilize knowledge contained in thesauri (e.g. WordNet) to include synonyms. The problem is that measures of string similarity disregard the meaning of the words, which can vary according to their context.

To take this aspect into account, structural matching strategies are applied to compute similarity measures based on the entities' neighborhood in the ontology graph. The assumption here is that entities from ontologies occurring in similar contexts are likely candidates for an alignment. A concept's context may include its super-class, subclass, properties, and relations. Since each of these contextual entities is labeled, syntactical similarity measures can also be computed for them. That is why the output from syntactical measures is a very suitable input for comparing neighboring entities with structural measures.

3. Linking Media Metadata

This chapter exposes the details of how data compliant with Semantic Web standards shall be created from media metadata and connected to existing Linked Data vocabularies in order to enhance its expressiveness. First, the initial situation concerning the generation and use of media metadata will be outlined. Then, the implementation will be described in detail, including the tools supporting specific tasks.

3.1 Media Metadata

In order to understand how the conversion into and connection with Linked Data can be beneficial to actual use cases, the initial situation around the media metadata is introduced first. This section contains a brief overview of how media metadata is acquired at a broadcasting organization, a description of this metadata's content, and an introduction of the involved software tools.

3.1.1 Metadata Generation and Consumption Lifecycle

The broadcasting organization whose media metadata is to be converted into linked data is ARD.[27] ARD is a German nation-wide public broadcaster, and at the same time a joint organization of nine independent state broadcasters (Deutsches Rundfunkarchiv 2011). The present use case and the implementation's input data originate from one particular regional broadcaster, namely Hessischer Rundfunk (HR).[28] HR is the public broadcaster for the German state of Hesse and has its headquarters in Frankfurt am Main.

As new footage is produced on a daily basis, related metadata is added to a centralized database, mostly manually. This database is called FESAD and is used by most ARD broadcasters.[29] It stores formal and content-related metadata, which is accessible through different access points (some of

[27] **A**rbeitsgemeinschaft der öffentlich-rechtlichen **R**undfunkanstalten der Bundesrepublik **D**eutschland, http://www.ard.de
[28] http://www.hr-online.de
[29] **Fe**rns**e**h**a**rchiv**d**atenbank, in English „television archive database"

which are described in Section 3.1.3). The metadata related to an entry, which can be an entire show or just a report (a part of a broadcast), is acquired in two independent steps.[30]

First, formal metadata is captured, which includes information such as the data of the initial broadcasting, title(s), or technical features.[31] This step is neither complex nor time-consuming. At HR, it is performed by a dedicated team of three people.

While FESAD does not maintain actual contents nor allows managing it, it provides information on where and how records are physically stored or, in the case of digital recordings, interfaces to the corresponding content management systems. In a second, more challenging metadata generation task, the broadcast's content is captured. This is done in the form of a general categorization, a set of tags related to its content and natural language content descriptions.

While the system, and thus the database schema, is shared across ARD broadcasters, its usage is somewhat individual to each one, in particular in terms of content description. Conventions regarding form and vocabulary emerge over time and are communicated between actors of a particular organization, but are limited to this scope. While this facilitates routinely performed queries, it makes integration across broadcasters difficult, for instance when editors wish to broaden their search scope.[32] Although there is a common set of guidelines (Gross et al. 2008) that provides recommendations with regard to the importance of individual content describing data elements and their syntactical format, some liberties are taken in practice.[33] Also, metadata capturing practices tend to evolve over time, leading to further heterogeneity. However, the current conventions internal to a particular organization are quite strictly enforced. For a closer look at the content of the data elements, refer to Section 3.1.2.

Content descriptions are captured by archivists, whose task it is to view broadcasts, or excerpts from them, in order to gain a qualified understanding of their contents. Based on this and the metadata conventions prevalent at their organization, the information is then entered directly into the FESAD database and the corresponding broadcast labeled as "processed". As this task can be somewhat time-consuming and at any given time, a long queue of broadcasts waits to be processed, archivists need to prioritize their efforts. To do that, at HR they are issued a list of show formats listed according to their priority. At the top of this list are formats that are broadcasted daily, such as

[30] The distinction between complete shows ("Sendung", in short SDG) and reports ("Beitrag" BTR) is of relevance in the database, as edited content can be stored in both of these forms. Both terms are generalized as "broadcast" in this thesis.
[31] All the details concerning metadata generation were gathered in an interview with an archivist at HR.
[32] The term "editor" ("Redakteur" in German) in this thesis refers generally to content creators, such as journalists and executive producers.
[33] These guidelines are issued by a joint ARD/ZDF task force

“hessenschau”, the daily evening news program, and “Heimspiel”, a sports show. This is not only due to their high recurrence, but also to the topicality of their content, which may have to be reused as soon as the next day. This criterion is by far the most influential in the priority listing.

Typical consumers of content describing metadata are editors in search of reusable contents. Editors are aware of content description and tagging conventions, and know how to adjust their queries accordingly. An interface tailored specifically to this task allows them to preview broadcasts and order physical or digital representations of them, depending on their availability, and to produce new content with it. Footage is stored separately in its raw and edited form, both of which are kept track of in FESAD, although the content of raw footage is less frequently described.

3.1.2 Content Description Metadata

As the previous section introduced, broadcasts are described both in terms of extrinsic properties and with regards to their content. This distinction is not only important because it represents two different metadata generation tasks, but also because only the latter metadata is suitable to be linked to external datasets and thus relevant to this thesis. As mentioned earlier, broadcast content is outlined by means of a categorization, tags, and textual descriptions.

At Hessischer Rundfunk, categorization is done against a controlled set of labels. Out of the four category dimensions suggested in the guidelines (content, presentation form, genre, and target group), only content categorization (identified as KAI in the database) is always captured.[34] It serves to express the general topic of a non-fictional broadcast, for example “biography”, “history”, or “culture”. Multiple content categories are allowed and recommended where necessary, and subcategories, which are also predefined, can be used for a more specific classification, but are never assigned at HR. A presentation form category (KGP) is only assigned to database entities that represent complete shows, and even then only occasionally. Example presentation forms are “documentary”, “magazine”, and “movie”. The genre category (KGG), which is also rarely used, serves to further specify some presentation form categories. For instance, a television series can be declared to be a science fiction series, or some non-fictional content to be a newscast. A value for the target group (e.g. children, elders, minorities) categorization type (KAZ) is never assigned at HR.

Tags are assigned to a broadcast in order to describe its contents with a set of defining terms. There is no controlled vocabulary for the terms that can be assigned. However, some conventions are

[34] In German, the category dimensions are called “Inhalt”, “Präsentationsform”, “Genre”, and “Zielgruppe”, respectively.

prescribed at HR in order to enforce the use of certain terms in tagging broadcasts. Theoretically, five classes of tags exist: geography (GEO), person (PERS), institution (INST), work of art (WERK), and general tags (STD). The last of these is the default class, and also the only one that is actually used in practice. Only rarely are the tag types geography, person, and work of art used when the corresponding entity is the main subject of the broadcast. Allegedly, tags are not used at HR when performing queries.

There are several types of textual content description available in FESAD. The main one is the data element identified as T (for text), a short abstract describing the essence of the broadcast. The convention observed at HR is to prepend these concise texts with the name of the location of the subject.[35] They generally do not consist of more than one or two sentences or sentence fragments. A broadcast's content may also be summarized more extensively in a press text (PTX) or in an internet text (OTX). Furthermore, a few data fields are provided to reflect the content of different aspects of a broadcast, such as shown photographs (FOT), interviewees or speakers (OT), or music that is played in it (MUS). Last but not least, the action occurring in the broadcast can be expansively described in the data field "image content" (BI). This is again one of the more important descriptions and is frequently captured with great minutia by archivists. Its significance stems from the fact that editors often need to search for suitable footage and are less interested in the broadcast context it originates from. In the image content, each clip appearing in the broadcast is portrayed with brief descriptions, usually sentence fragments, and separated by forward slashes. The descriptions frequently also feature indications on how the content was filmed, for instance by a camera pan, or with a panorama shot, information that is also query-relevant for editors.

3.1.3 Current FESAD Access Points

FESAD represents the centerpiece of many operations at HR. Broadcast metadata plays a role in the work processes of many different actors, who all make use of it in different ways. This entails a certain number of diverging requirements, which are accommodated by various access points to FESAD data. In this thesis, the focus will be directed at the interfaces FESADdigital and FESADred.

FESADdigital, commonly referred to as "FESAD Web Client", is a browser-based client that can be accessed over HR's intranet. Its purpose is to provide extensive search options to retrieve broadcasts and the possibility to order broadcast tapes or files. Its search form is shown in Figure 5.

[35] It is not known whether this convention is also adhered to by other ARD broadcasters.

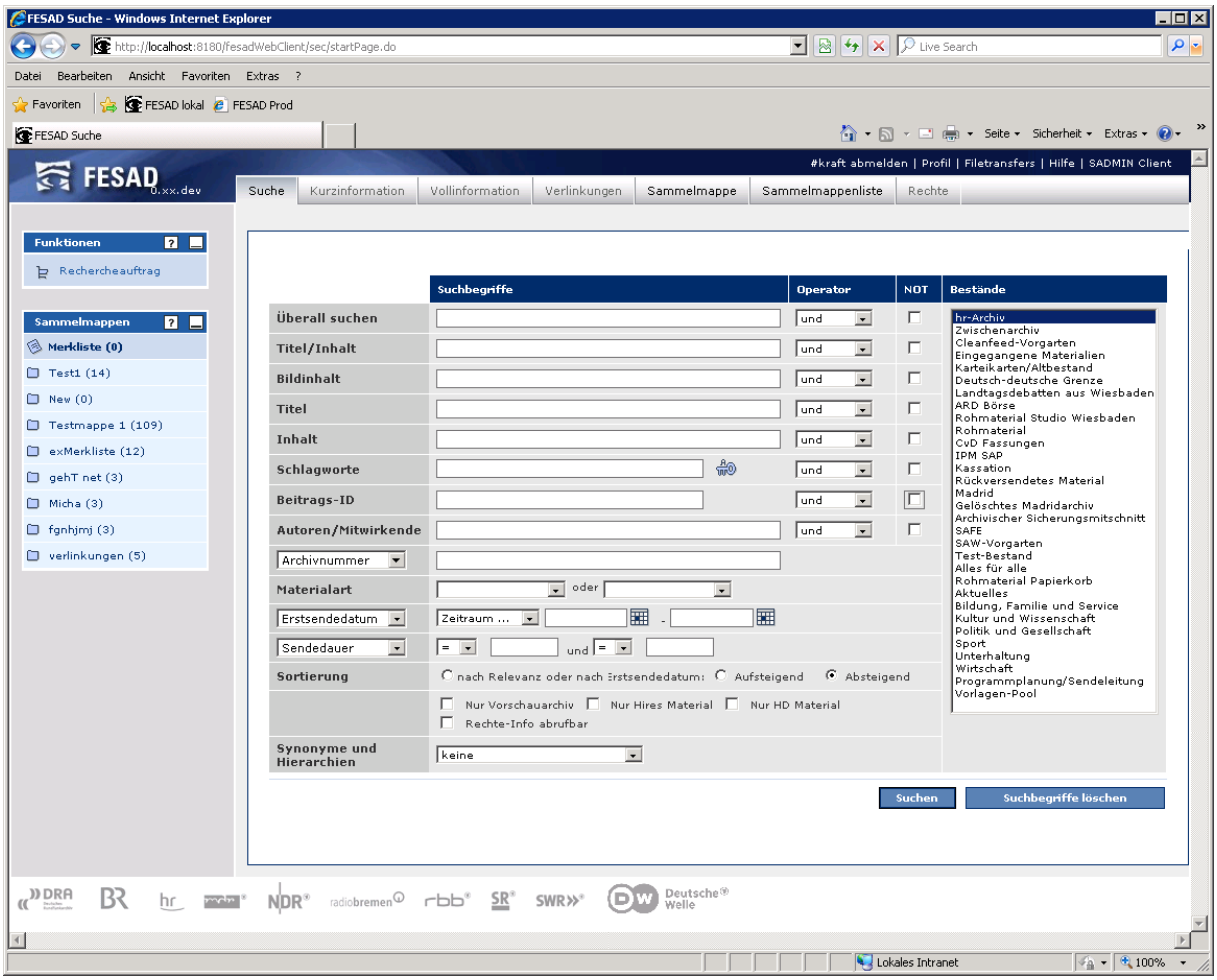

Figure 5: The FESADdigital search form

Multiple text input fields allow the user to base his search on different criteria. Input in the field “Überall suchen” (“search everywhere”) is used to perform a full-text search, i.e. the search terms are matched against all metadata elements. The next text fields can be used for queries based on image content, title, abstract, tags, collaborators, and broadcast ID, respectively. All of these fields allow the use of the wildcard character “*” to match any sequence of characters. Furthermore, broadcasts can be filtered by the date of recording or airing (of the initial or of a repeat broadcasting) and by their duration.

The column to the right of the search form contains dropdown menus with operators for union and conjunction, as well as ones with which the location of search terms can be specified. For instance, “NAHE-5” (“close 5”) finds all documents in which the search terms from the corresponding text field occur with a maximum distance of five words from each other. Finally, the rightmost column lets the user specify in which collection of documents the query is to be performed.

FESADdigital offers the possibility to display results in two ways: a short and a long information view. The former option displays the results as a list, showing only some of the main properties of broadcasts, like title or duration. In the long information view, the full information pertaining to particular broadcasts in the result set is displayed. At this point, the user has the opportunity to watch a preview of the broadcast he is interested in and to order the tape or file it is stored on. Further options offered by FESADdigital include the creation of folders in which broadcasts can be stored in and the possibility to mandate an archivist to perform a specific search for the user.

The second FESAD interface we will take a look at is FESADred. It is more tailored to the needs of content creators. FESADred is also browser-based and essentially provides the same possibilities as FESADdigital. But in comparison with the latter, FESADred's search facilities are somewhat simplified. Figure 6 represents a screenshot of the search page.

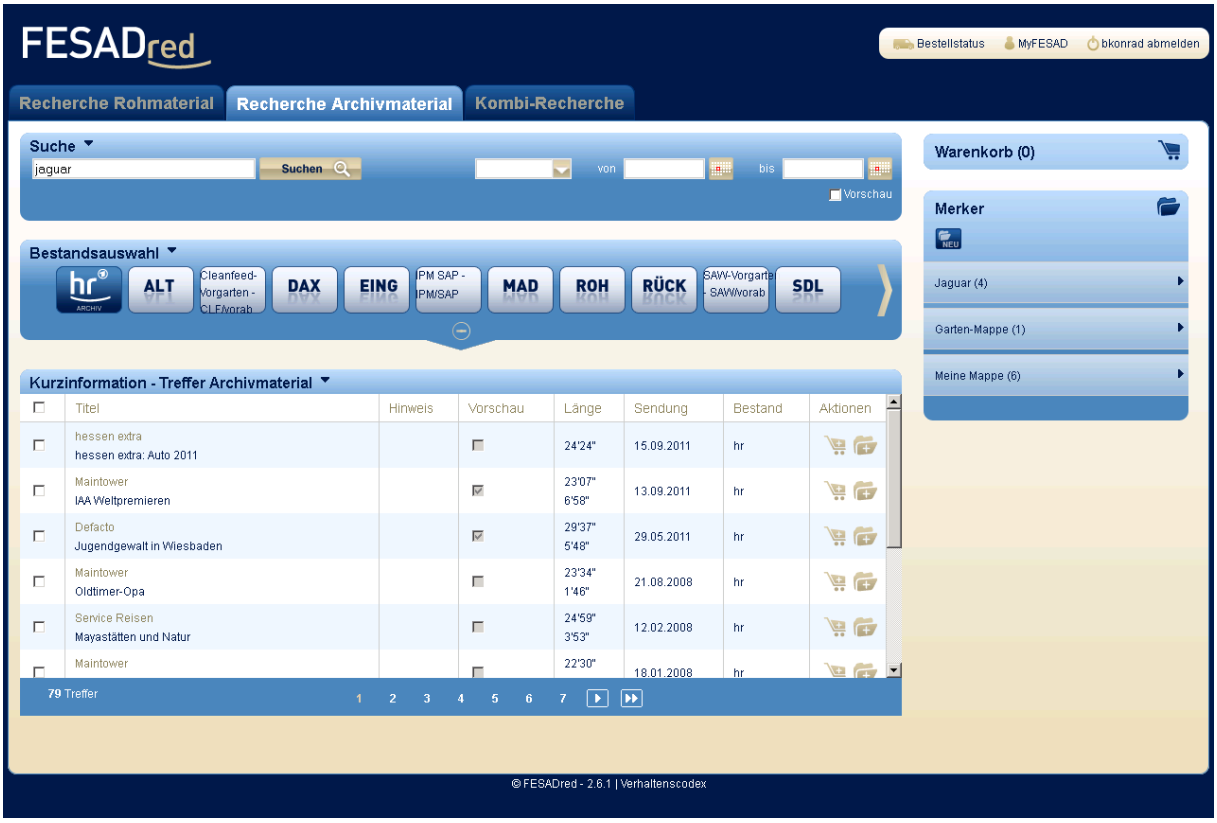

Figure 6: The FESADred search page

Tabs at the top of the page let the user decide whether to search in raw or edited footage, or a combination of the two. Below it is the search bar, whose main component is the single text input

field (as opposed to FESADdigital's nine input fields). It allows for a search with one or more keywords, and operators, such as an "OR" placed between two terms for a logical disjunction, are typed directly into the search bar. The search terms are compared with all of the broadcasts' content description data elements, and there is no option to discriminate between them when formulating queries. At the right side of the search bar, either a specific broadcasting date, or a date range can be specified.

Like in FESADdigital, the user can navigate from a list of results displaying some essential information to a complete information view of a broadcast, including a preview of it that can be watched in an integrated video player. Finally, a user also has the option to organize broadcasts in folders and to order them.

In conclusion, the principal difference between the two interfaces is that FESADred presents a boiled-down version of FESADdigital to the user, covering only the particular needs of editors. The fact that FESADred is widely used, despite both clients being available, attests that it indeed addresses their requirements adequately.

3.2 Implementation Details

This section introduces the implementation of methods to link HR broadcast descriptions with the Linked Data cloud and to browse them using a faceted search system. This is achieved in several steps, which are described individually, along with some corresponding available tools. But first, an overview of the entire set of processes is presented.

3.2.1 Implementation Overview

Aside from evaluating the benefit of connecting media metadata with Linked Data, the implementation serves the purpose of demonstrating a method to perform this connection. Because this task involves a certain level of complexity, it is implemented as individual and mutually independent projects. This decoupling also has the advantage of providing some flexibility. For example, the assignment of keywords to broadcasts is not dependent on their description being serialized in the RDF/XML format, a step that can hence be skipped. The source code can be found on

the enclosed CD-ROM and is explained with frequent inline comments. The implementation is mainly written in Java code, unless it is stated otherwise in its description, and using the Eclipse IDE.[36]

First, an important question for the implementation is what part of the media metadata is to be connected with Linked Data concepts. The answer is not straightforward, because neither broadcasts' content descriptions, nor the tags assigned to them can be directly linked. For the textual content descriptions, the reason for this is evident; without some form of preprocessing, it cannot be made known which parts of the text are crucial and which terms express its essence most accurately. Thus, a text mining operation is needed first in order to boil description texts down to a limited number of keywords. However, at this point the same problem is encountered as if it were attempted to link tags directly, namely the lack of a controlled vocabulary. Without a defined set of keywords to link to external datasets, there would be no control over which terms are linked and how this linking would be done. Semantically equivalent words such as "car" and "automobile", as well as semantically equivocal terms such as "jaguar" could not be distinguished, resulting in redundant or inaccurate links, respectively.

A Controlled Vocabulary

This is why a thesaurus is used as a controlled vocabulary. Only the concepts contained in it are linked to external datasets. This has the advantage that the linking is decoupled from the content descriptions, but also that the thesaurus places the terms in a context and allows identifying them unambiguously. Conveniently enough, ARD maintains a set of thesauri, which includes a large, general thesaurus covering concepts across many topics, and a classification of geography concepts.[37] The concepts contained in these are described by means of main labels, synonyms, as well as related, broader, and narrower terms. In order to connect these concepts with Linked Data entities, the thesaurus needs to be converted to a Semantic Web standard. In this case, the obvious format choice is SKOS (see Section 2.1.6), as it allows to represent all the thesaurus expressions equivalently. Details of this conversion's implementation can be found in Section 3.2.2.

As mentioned, the purpose of the controlled vocabulary is to connect broadcast descriptions with Linked Data. First off, this entails the task of examining Linked Data vocabularies in order to assess whether they are suitable candidates to be linked to (see Section 3.2.3). Then, the entities of both vocabularies must be matched. This means that for every thesaurus concept, an equivalent concept

[36] Eclipse Helios Service Release 2

[37] In German, „Sachklassifikation" and „Geographika", respectively. While ARD maintains further thesauri, for example of persons and institution, only these two were used in the context of this thesis.

in the Linked Data vocabulary must be found. The problem in determining such a correspondence is that it cannot be asserted with certainty on the basis of identical labels alone because of homographs, and equivalent concepts may have different names. An accurate matching is a critical factor for the usefulness of the connection with Linked Data. It is also a good idea to link thesaurus concepts to multiple vocabularies to increase their information content. The implementation of the alignment is explained in Section 3.2.4.

In order to use the broadcast metadata for further processing, it was copied from FESAD as XML files. More precisely, 101 reports were extracted using an SQL view designed to retrieve entries that appeared in a hessenschau broadcast, were fully processed, and were aired after 2006. hessenschau is HR's daily newscast and was selected because of the wide range of topics it covers and because its content is consistently well-documented. The broadcast metadata is then converted to RDF, a step that is strictly speaking not absolutely necessary to link them to thesaurus concept. However, it might be convenient to be able to store all the input data of a faceted search application together in a consistent format and database layout, in this case a triple store. A short overview of the implementation is given in Section 3.2.5.

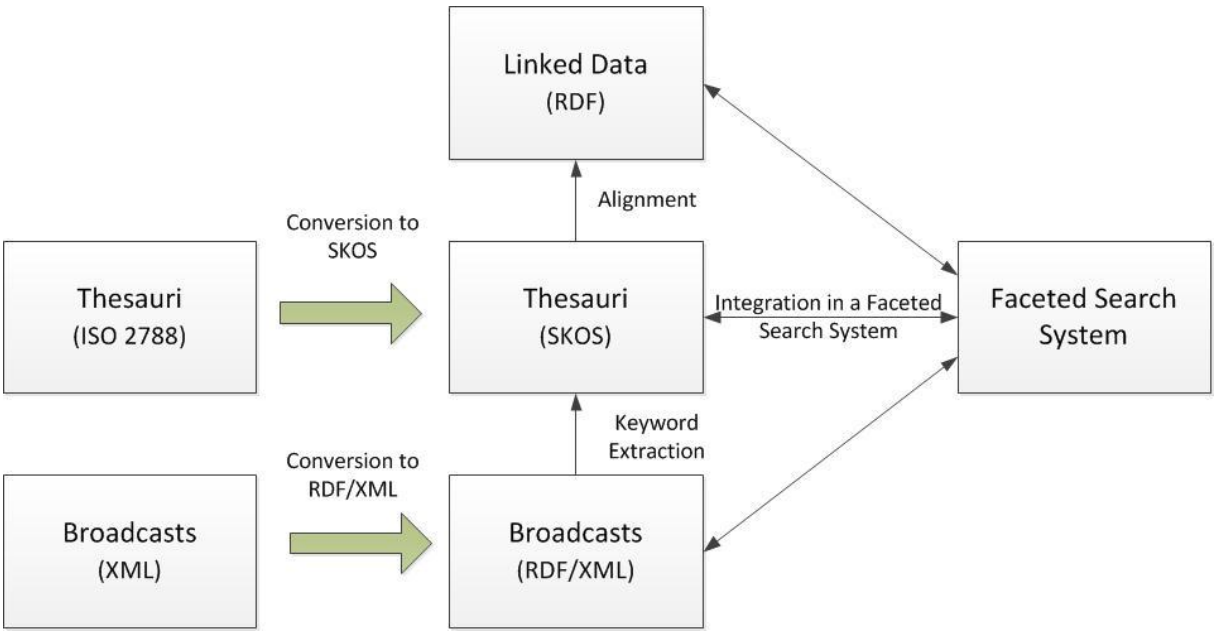

Figure 7: Overview of the implementation steps

Finally, in order to connect the broadcasts to the thesaurus, their descriptions, which include tags, content summary and other textual information, must be translated into thesaurus concepts. The implementation of this keyword extraction is detailed in Section 3.2.6.

In this thesis, a faceted search application to test its suitability for querying broadcasts was also built. Its implementation is documented in Section 3.2.7.

The illustration in Figure 7 shows an overview of all the implementation steps and how they fit together. The rectangles represent the individual datasets and the faceted search system. The thick arrows indicate conversion steps and the thin ones data interconnections with their directionality.

3.2.2 Thesaurus Conversion

The conversion of the thesaurus, or rather thesauri, to SKOS is a straightforward task. The inputs are the general and the geography thesauri, which are available in the ISO 2788 format.[38] All their concepts are labeled only with German language designations. The relations that can be assigned to a concept are synonyms (`SYN`), related terms (`RT`), broader terms (`BT`), and narrower terms (`NT`). Broader and narrower terms are loosely defined as super-concepts and sub-concepts, respectively, and are not transitive. The SKOS format (see Section 2.1.6) is able to express all these relations as well, meaning the task at hand is a 1:1 conversion. The only new aspect that is added to the concepts is URLs, since RDF resources should be identified that way.

The conversion is implemented in the *ThesaurusConverter* project. First, the assertions for each particular concept in the input document are split up and parsed. As a result of this operation, `Concept` objects, which encapsulate all the information relating to a particular thesaurus entry, are created. The extraction of this information is simply done by examining the start of each line of the concept description to find out what it describes and assigning the label in it to the corresponding `Concept` attribute.

Before converting the concepts into SKOS, some redundant information contained in the source thesaurus is removed. Specifically, concepts are originally always referred to by their main label and synonyms. For example, if the concept “automobile maker” is assigned the synonym “car manufacturer”, both of these labels always appear together, that is, “automobile industry” would be asserted to have both of them as narrower terms. If this were converted into SKOS, it would mean that both labels refer to distinct concepts, since a `<skos:narrower>` statement links two SKOS concepts. For this reason, references to synonym labels are removed whenever they appear outside of their concept descriptions.

[38] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=7776

The actual conversion of concepts into SKOS is finally done by writing a file conforming to its syntax. Concept URLs are created by concatenating a base URL chosen by the user with the appropriately modified concept name.[39]

3.2.3 Selection of Linked Data Vocabulary

To start with, an analysis of Linked Data vocabularies is made in order to determine whether the concepts they describe generally correspond to the thesaurus concepts that are to be linked.

As seen in Section 2.2.2, DBpedia is a dataset containing RDF descriptions of a large number of concepts described in Wikipedia and is at the center of the Linked Data cloud. It would evidently be a very good fit for both the general and the geography thesaurus to be aligned with, as it includes a considerable amount of German language labels and descriptions, and is strongly interlinked with other datasets. Its data is regularly extracted from Wikipedia articles, which themselves can be expected to be quite up-to-date.

Another general-purpose knowledge base is Freebase.[40] It is a collection of metadata maintained by an active community and is available in the RDF format.[41] Freebase contains more than 125,000,000 tuples about concepts of 4,000 different types (Bollacker et al. 2008). An option, although it is not implemented in this thesis, would be to link thesaurus concepts to DBpedia as well as to Freebase is performed, to check whether both knowledge bases confirm or complement each other's statements.

For the connection of the geography thesaurus with Linked Data, GeoNames is also a very good candidate.[42] GeoNames is a database of 6.2 million geographical features, which makes it one of the most comprehensive ones (Vatant 2012). Descriptions of places include information such as elevation, population, latitude and longitude, or postal codes, and originate from official public sources. GeoNames concepts have labels in multiple languages including German, making them suitable for matching ARD thesaurus concepts with.

Many Linked Data vocabularies specialize on a particular domain. In the context of linking broadcast metadata with them, the relevance of the information they contain obviously depends on the type of broadcast and the particular purpose of the Linked Data application. For instance, a show frequently

[39] „Appropriate modifications“ includes converting the concept name to a lower case string and replacing blank spaces with underscore ("_") characters.
[40] http://www.freebase.com
[41] http://wiki.freebase.com/wiki/RDF
[42] http://www.geonames.org/

discussing health problems could be linked to a comprehensive knowledge base of concepts pertaining to the field of medicine. The usefulness of this connection would however depend on whether such high information content would be actually needed in practice.

What could potentially be useful is linking HR media metadata with vocabularies describing the same particular media. For instance, movie metadata could be connected with the Linked Movie Database, a highly interlinked set of movie-related information.[43] It might also be worth considering linking broadcast metadata to specialized Linked Data knowledge bases in order to describe specific types of concepts. A broadcast could for example be about a certain painting, which could then be linked to the corresponding description in the Europeana dataset, a Linked Data collection of culturally relevant entities.[44]

3.2.4 Thesaurus Alignment

For the thesaurus' connection with Linked Data, DBpedia and GeoNames were chosen, since these two vocabularies are expected to cover the entirety of the concepts' range, are very comprehensive, and contain a great amount of information and links to other datasets. The next step is to establish this connection by matching every thesaurus concept to an equivalent Linked Data description. Ontology alignment tools (see Section 5.2) were first considered for this task. It was however soon found out that most alignment systems are not suited to the job of comparing SKOS concepts to ontology classes because this format is not supported. With the alignment frameworks that are able to handle SKOS, another issue emerged, namely that due to its size, the DBpedia ontology could not be processed in its entirety. Some major adjustments would have been necessary to perhaps achieve a working solution for at least one Linked Data vocabulary.

Instead, the decision was made to write a custom solution. This implementation is contained in the *OntoCompare* project. A method to match all thesaurus concepts with DBpedia and another one to match only the geography concepts with GeoNames were implemented, although they share most of the code.

The DBpedia data that is used are German concept labels, German long abstracts, type and property assertions, and disambiguation links.[45] They are available in the N-Triple/RDF format. To compare the labels, thesaurus entries and their information are first encapsulated in `Concept` objects, while

[43] http://www.linkedmdb.org/
[44] http://data.europeana.eu
[45] Downloaded from http://wiki.dbpedia.org/Downloads37

DBpedia terms are stored with their URLs as `VocabularyResource` instances. This information is quite simple to extract, since the input is well-structured. Then, the central task of the alignment is executed, the label comparison. For every thesaurus concept, its label, synonyms, and broader terms are checked against the DBpedia labels. This works by means of a direct label comparison. Figure 8 illustrates the actual workflow. Aside from label comparisons, concept names were also compared in their presumed DBpedia URL form, as in practice this approach was expected to be promising in certain cases.[46]

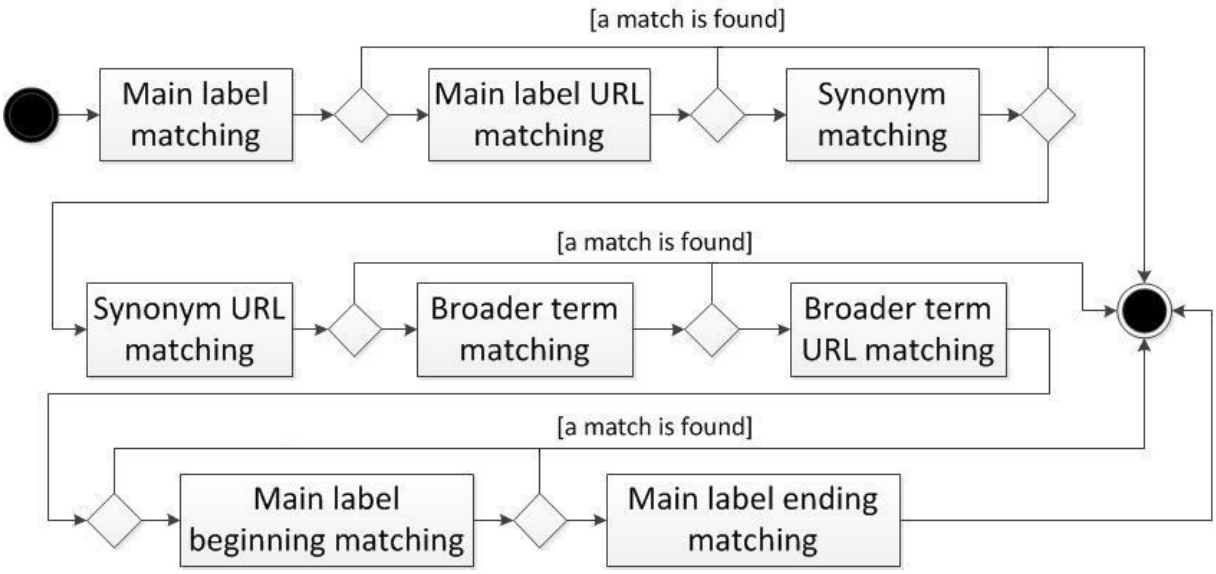

Figure 8: Workflow of the matching for a particular thesaurus-DBpedia concept pair

As Figure 8 illustrates, the matching approaches are used sequentially until a match is found, after which the algorithm moves on to the next Linked Data concept. In the last two steps, concept names with matching beginnings and endings, respectively, are searched. The matches found in the label comparison are kept track of in the corresponding `Concept` objects.

As a next step, disambiguation links are resolved. At this point, some explanations are in order: a disambiguation page shows up in Wikipedia when a user enters a search term that refers to more than one thing (e.g. "Alien" – a movie, a law, a novel, and a band, amongst others), and provides some context and links in order to redirect him to the appropriate page. Since DBpedia mirrors Wikipedia, it contains concepts that are in fact only the equivalent of disambiguation pages. Luckily,

[46] For instance, if the thesaurus concept's name is "University of Zurich" and no direct label match was found, it would be compared as "http://dbpedia.org/resource/university_of_zurich".

DBpedia provides a download of the complete set of links from disambiguating concepts to actual concepts. In the alignment implementation, such disambiguation links are simply resolved by replacing them with all of the "real" DBpedia concepts they point to. Since this could introduce duplicate matches, a method is called to remove them.

Contextual matching

Evidently, it is to be assumed that the matching made up to this point is not yet of great precision. Rather, label comparison is the initial step in a breadth-first search whose result set must now be narrowed down. As existing ontology alignment systems demonstrate, structural information must be taken into consideration in order to assess the adequacy of alignments. On the thesaurus side, structural information comes in the form of broader and narrower terms. DBpedia properties are available in a single downloadable file, with property-value assertions expressed in no particular order. But the situation here is not quite as simple as in the thesaurus because the vocabulary used to express relations is much more extensive. While a 1:1 alignment of ontology properties is thus out of question, a 1:n alignment is practically not feasible in a complete and accurate manner. In summary, the situation is that properties for DBpedia concept are known, but not how they semantically related to broader and narrower terms in the thesaurus.

Since the meaning of the properties asserted to the DBpedia concepts is unknown, a simple solution would be to disregard semantics altogether. In order to nonetheless incorporate thesaurus structure information, the alignment could be based on the assumption that concepts with identical labels and similar neighborhoods are indeed equivalent. In this context, the neighborhood consists in broader and narrower terms for thesaurus concepts, and is loosely defined as the set of concepts to which a relation exists in DBpedia.

While the implementation does rely on the aforementioned assumption, the computationally simpler solution of using concept abstracts was chosen.[47] These are short, natural language descriptions of DBpedia concepts. Concretely, for every connection made between a thesaurus and a DBpedia concept in the label matching phase, the corresponding abstract is fetched and examined. It is checked for occurrence and frequency of terms from the thesaurus entry's neighborhood. Evidently, the occurrence of the concept label has to be valued more highly than one of a neighborhood term. When a particular alignment is examined in the implementation, a value is computed for it on the

[47] It is computationally simpler because, as opposed to properties, only one abstract exists per concept. Also, many properties do not have resource values and when they do, the corresponding label must then be retrieved for the comparison to be made.

basis of the frequency of the thesaurus concept's set of terms and the weights that are assigned to them. The weight for a concept label or synonym is 100, and those of broader and narrower terms are 30 and 20, respectively. A broader term occurrence is valued more highly because it is assumed to be more relevant and thus more frequent in a concept's description.

To briefly illustrate this step, assume a concept "Jaguar" from the thesaurus is matched with a particular DBpedia concept in the label comparison phase. In the thesaurus, the concept has the broader terms "Animal" and "Cat", but no synonym or narrower term. To assert that this match is an accurate one, the DBpedia resource's abstract is retrieved and examined for the occurrence of the terms "jaguar", "cat", and "animal". Let's further assume, "jaguar" is found twice, as well as one occurrence of "cat". This particular alignment would then be assigned the value of 230, calculated by multiplying the frequency of terms with their weight. On the basis of such values, for each thesaurus entry only the best matches that also pass a certain threshold are kept.

Because DBpedia also provides links to other Linked Data vocabularies, references to further equivalent resources could be added at this point, an option that is provided in the implementation but not evaluated.

Alignment with GeoNames Concepts

For the alignment with GeoNames resources, some other information was available at the context comparison stage. These vocabulary's geographic concepts are available as RDF/XML dumps and are described, amongst other, by their names, features, and corresponding country, all of which can be used to compare thesaurus concepts with them. The reason for this is the fact that the concepts in the ARD geography thesaurus frequently have the country they are located in featured in their set of broader terms. In addition, all the concepts are assigned a type in the source thesaurus, which in the geography thesaurus indicates the kind of geographical feature such as "place", "country", or "lake". This information is compared with the one given in the GeoNames descriptions and used to assign values to the individual alignments, like it is done with DBpedia alignments. Feature types are a very useful additional comparison dimension, but it comes at the cost of manually establishing the semantic equivalencies between the two vocabularies.

An evaluation of the alignments can be found in Section 4.1.

3.2.5 Metadata Conversion

The XML broadcast files were converted to RDF/XML using XSLT (eXtensible Stylesheet Language Transformation).[48] The conversion is very straightforward, transforming XML elements by prepending a namespace ("`ard`") to their names, but keeping their original content. This implies that a custom vocabulary is used throughout. Existing RDF vocabularies could also have been used to describe the broadcasts, but since, at least in the context of this thesis, the broadcasts are not made visible to external applications and thus do not need to be semantically interpretable, this is not necessary.

The document structure is also modified, specifically by being flattened. This removes some unnecessary, i.e., semantically irrelevant, intermediate elements which would have had to be translated to blank nodes in the RDF documents and added complexity to queries. The implementation was made in the project *XMLConversion*, which can also be found on the enclosed CD-ROM.

3.2.6 Tagging of Media Data

The goal of media data tagging, also and interchangeably referred to as "keyword extraction" in this thesis, is to determine the set of the most suitable thesaurus terms to describe the contents of the broadcasts with.[49] A keyword extraction tool that would fit the requirement at hand, namely to match the contents of multiple texts, as well as tags, against a controlled vocabulary, could not be found. Here again, a custom implementation was written and has the Java project name *ContentLink*.

Information relative to broadcasts is parsed into `BCDescription` objects. This includes the ID number, tags assigned by archivists, and content descriptions, which in turn include the broadcast's summary and the detailed account of its image content. For these full-text descriptions, punctuation marks and lower case words are removed. This is based on the assumption that the nouns always start with uppercase letters and that they represent the entirety of the descriptions' relevant information.

Tags and the words remaining in the content descriptions are directly compared with all the thesaurus concept labels (including synonyms). If no match is found for a specific tag, a stemmed version of it is compared. The stemming method is a very basic attempt to convert plural nouns to their singular form. All of the matching terms are added to a list, which then obviously needs to be

[48] http://www.w3.org/TR/xslt20/

[49] To anticipate some potential confusion: in this thesis, "tags" most frequently refers to the terms assigned to broadcasts by archivist, while keywords are the controlled vocabulary terms assigned to them as a result of the process described in this section.

narrowed down to the set of most relevant terms. As in the alignment algorithm described in Section 3.2.4, frequency and a weight, which is determined by where the match was found, are used to compute a matched keyword's value. The weights for tag, summary, and image description matches are 5, 2, and 1, respectively. Up to seven keywords are then selected per broadcast based on their values. A keyword is assigned to a broadcast by adding an `<ard:keyword>` RDF statement with the corresponding thesaurus concept's URL as object. An evaluation of the keyword extraction is described in Section 4.2.

3.2.7 Faceted Search Client

Several aspects need to be taken into consideration for the implementation of a faceted search system. On the one hand, there are obviously concerns regarding the design of the user interface. The user's needs have to be accommodated in terms of usability. Also, given the novelty of the faceted search paradigm, an important question is how to keep the query process simple and easily accessible, while still harnessing the advantages of Linked Data. More precisely, in what form is information entered into and returned by the system? On the other hand, implementation decisions have to be made. What background data is used, how should it be accessed, and how is it utilized to gain information from for the faceted search? Also, what technology is used to write the implementation?

Given the scope of this thesis, a full-fledged faceted search interface was not implemented. Some compromises were made both in terms of the front-end design, as well as in terms of the input data extent. The following descriptions thus correspond to what the faceted search client was originally envisioned to include, while the actual implementation is discussed afterwards.

Blueprint for the Faceted Search

The front-end layout was inspired by the Faceted Wikipedia Search (Hahn et al. 2010), which was already mentioned in Section 2.2.3. A text field at the top of the page takes a keyword from the user to start the search off. Two options can be considered for this initial query step: The first one is to use the search terms to perform a full-text query, as done in traditional search engines as well as in the Faceted Wikipedia Search. This would however require the preliminary creation of at least an inverted index of the broadcasts to be queried. The second and simpler option is run the query against thesaurus concepts only. In this case, it would be useful to enforce the use of this controlled

vocabulary, for instance by means of automatic word completion. The result set returned for such an initial query would consist of broadcasts that were tagged with the corresponding keyword.

In a second query step, the user has the possibility to refine the result set by selecting a facet value. Facets and corresponding facet values are displayed at the left hand side of the result list below the aforementioned text field. As introduced in Section 2.2.3, a facet represents a particular RDF property. In this case, the facets would be retrieved from the Linked Data concepts the result broadcasts' keywords are connected with. The selection of a particular facet value would cause the result set to be narrowed down to the broadcasts that are identified with at least one keyword whose Linked Data equivalent has a corresponding property-value assertion. For example, selecting the facet value "animal" for the property "type" leaves only broadcasts with keywords that are linked to a concept with property "type" and value "animal". In practice, it would be best to keep a certain level of control over the facets. A reason for this is that some facets might be unimportant and distracting to the user, such as the coordinates of a location. In order to keep the interface simple, only relevant facets should be displayed. Also, facets should be labeled with meaningful and descriptive names, which is not always the case for RDF properties.

In the result list, some relevant information related to the retrieved broadcasts is displayed. This includes title, a short description, broadcasting dates, and a preview image. If the faceted search were integrated with a FESAD client, the same options for further use of the broadcast could be offered, such as the rental of the corresponding tape.

Actual Faceted Search Client Implementation

As a framework for the implementation, Callimachus was used.[50] Callimachus is a handy tool for the development of Web applications based on Linked Data, as it provides both a stand-alone server and a data store. It also includes facilities to develop Web pages and to query RDF data with SPARQL. However, Callimachus only handles RDF/XML, which is why the DBpedia data, serialized in the N-Triple format, has to be converted.

The search interface was written in HTML and CSS for the layout and JavaScript for the search functionalities.[51] [52] Figure 9 shows a screenshot of the search interface, whose layout is quite rudimentary. In the illustrated example, a query with the term "frankfurt" was performed, which

[50] http://callimachusproject.org
[51] http://www.w3.org/TR/html5/
[52] http://www.w3.org/Style/CSS/

returned the facets corresponding to the properties of that thesaurus keyword's DBpedia concept, as well as the broadcasts with this particular keyword.

Figure 9: Screenshot of the faceted search client

Only a reduced sample of data was loaded into Callimachus, as some performance issues were experienced with large datasets, such as the ones containing DBpedia labels and properties, each of a size of about 100 MB. If the thesaurus also included other concept with the name "frankfurt" (e.g. Frankfurt an der Oder), the facet "name" could for instance be used to make a distinction between them, which would leave only broadcasts about the chosen concept.[53]

To retrieve RDF data, so-called named queries can be used in the Callimachus framework. A named query is in fact a predefined SPARQL query whose results can be retrieved via JavaScript. The result

[53] The thesaurus does in fact contain other "Frankfurts", but not the uploaded sample. Only the data necessary to showcase faceted search abilities (one ambiguous and one complex query) in the user interviews was used in this implementation.

format of such a named query complies with the Google Chart Tools Datascript Protocol.[54] This means that to access the named query results, JavaScript functions from the Google query library have to be used. In the faceted search implementation, named queries play a central role. They are used to retrieve information from the RDF datasets and to structure it.

Concretely, four named queries are used:

- `propertyvalue-query` retrieves all the property-value assertions for the DBpedia concepts
- `skoslabel-query` fetches thesaurus concepts with their labels and DBpedia links
- `broadcastlabel-query` retrieves information relevant to broadcasts
- `type-query` gets the property-value assertions where the property is `rdf:type` for the DBpedia concepts

The last query in this list retrieves information that is also contained in the `broadcastlabel-query` results, but is needed because the concept type property is treated differently in the search interface.

When a query is executed, the input string is compared with the thesaurus concept labels retrieved by the `skoslabel-query`. If a match is found, the DBpedia concept for that entry is looked up with `type-query` and `propertyvalue-query`, and the property-value pairs displayed as facets and facet values. In the same time, the result table from `broadcastlabel-query` is used to look up broadcasts tagged with the keyword in question, whose information is then showed in the result list.

The faceted search system was evaluated offline in a static version, due to implementation issues. Interviews about the faceted search paradigm as a whole were conducted with users. The experimental setup and the results can be found in Section 4.3.3.

[54] https://developers.google.com/chart/interactive/docs/dev/implementing_data_source#responseformat

4. Evaluation

In this chapter, the evaluation results are presented. The evaluation is designed to give an indication on how well the thesis objectives stated in Section 1.2 are achieved. On the one hand, this involves an assessment of the process that is used to connect broadcasts with Linked Data, which includes the connection of these broadcasts with a controlled vocabulary and the alignment of this vocabulary with a Linked Data knowledge base. On the other hand, the usefulness of faceted search in real-life use cases has to be tested. The following three sections offer an overview of the evaluation results, while a discussion of them can be found in Chapter 6.

4.1 Alignment Evaluation

The alignment between thesaurus and Linked Data concepts represents the point through which access to semantically enriched data is gained. It is thus crucial that the controlled vocabulary terms be connected to resources representing the same thing. A poor alignment quality would negate the positive effects of creating a link to semantic information because the wrong information would be used to describe thesaurus concepts. In this section, the methods that are used to evaluate the alignment and the corresponding experimental set-up are introduced, after which the evaluation results are exposed.

4.1.1 Evaluation Methods

Information regarding the alignment was gathered in two ways. First, some measurements were kept track of during the alignment process. They give some insight as to how the alignment performed in purely quantitative terms. This data can be used to analyze what the alignment process did, but gives no indication as to how well it did. To determine the quality of an alignment, frequently used measures are gold standards, which are manually created reference alignments that can be used as a basis for comparisons.[55] For the present case, the creation of a gold standard was deemed

[55] As explained in Section 5.2, the Ontology Alignment Evaluation Initiative makes frequent use of gold standards.

unsuitable. Because DBpedia often inconsistently features multiple entries for the exact same thing, it would have required retrieving every single one of those redundant entries in order to be able to make a qualified statement about the alignments' precision.[56]

Instead, the decision was made to involve experts in the assessment of the individual alignments' quality. For this, random sets of thesaurus concepts, for which at least one match was found, were generated. Specifically, 20 geography concepts and 30 general thesaurus concepts were selected.[57] For every single alignment, the participants had to determine whether it represented an exact correspondence, an approximate correspondence, or no correspondence at all. A three-point Likert scale was chosen because a larger range of options should not be necessary to express the quality of an alignment, nor would it in this case convey any additional information. They also had the option not to rate individual alignments, and to write a remark. Since concept names can be ambiguous and DBpedia resource URLs offer no information at all, some context was provided to the participants to help inform their evaluation. For the thesaurus concepts, this consists of the synonym and broader, narrower, and related terms, and for the DBpedia concepts context information is given by means of the corresponding abstract text.

Three expert users participated in the evaluation, all of which work as archivists at HR. This qualifies them to make an informed evaluation of the alignments, as they deal with metadata on a professional basis. The full results of their evaluation are given in Appendix A.3.

4.1.2 Alignment Statistics

Three pairs of vocabulary alignments were run: general thesaurus concepts with DBpedia, geography concepts with DBpedia, and geography concepts with GeoNames. Table 1 gives an overview of some general alignment statistics.

[56] To illustrate DBpedia's inconsistency, for the city of Zürich, there are entries with the names "Zurich", "Zurich (Zurich)", and "Zurich (city)", amongst others. Apparently, a DBpedia entry is made for every article name, even if the same thing is described each time.

[57] These numbers reflect the size relation between the two thesauri (2:3), so that an extrapolation of the results to the entire concept sets is done by approximately the same factor.

Alignment type	Thesaurus concepts	Linked Data concepts	Matched concepts	Percentage matched
General concepts - DBpedia	38’943	658’519	16’193	41.58
Geography - DBpedia	20’959	658’519	12’216	58.29
Geography - GeoNames	23’387	1’892’028	3’758	16.07

Table 1: General alignment statistics

For the general thesaurus, all of the concepts were used in the alignment, whereas due to time constraints, only about three quarters of geography concepts could be attempted to be aligned, which should however be enough to draw valid conclusions about the alignment process. Almost 42% of the general thesaurus concepts and about 58% of geography concepts could be matched with at least one DBpedia equivalent. For the latter concepts’ alignment with GeoNames, this percentage drops to 16%.

The frequencies of the individual types of labels that were matched, e.g. main label or synonym, was also recorded and are summarized in Table 2.[58]

Alignment type	Thesaurus concepts	Main label	Main label as URL	Synonym	Synonym as URL	Broader term	Broader term as URL	Label beginning	Label ending
General concepts - DBpedia	38’943	11’101	1’440	1’223	157	3’953	721	41’720	25’642
Geography - DBpedia	20’959	9’314	1’450	662	102	3’588	2’262	19’587	18’903
Geography -GeoNames	23’387	9’510	-	723	-	4’505	-	-	-

Table 2: Frequencies of match types

Some of these values may appear to be in stark contradiction with the ones from Table 1. For instance, a much higher amount of matches seem to have been found in the GeoNames alignment than the previous table indicated. The reason for this is that the frequencies of match types, as presented in Table 2, were recorded right as the matches are found. However, after the label matching, duplicate and insufficiently valued alignments are removed before definitely keeping them, which obviously filtered out a certain number of matches and is why many more matches were found than finally kept. Another anomaly seems to occur in the number of thesaurus labels successfully matched with Linked Data concept name beginnings, which is higher than the actual amount of thesaurus concepts. Obviously, this is explained by the fact that multiple matches for

[58] URL-based comparisons were not made for the GeoNames alignment, because GeoNames URL only contains a number identifying the concept.

thesaurus concepts are allowed, and some labels seem to be represented quite often in the first or last letters of DBpedia concept names.

Out of the direct string matches, most were found when comparing the thesaurus concepts' main labels. Concept synonyms were matched significantly less often, only about 10% as frequently. This confirms the assumption that the main labels are indeed the most commonly used ones, especially for geography concepts. Matches based on URL versions of the concept labels were rather rare, as opposed to ones made on the basis of identically beginning or ending strings, which in average occurred almost once per thesaurus concept. As the final number of matches leads one to suspect, many of these initial matches are discarded.

4.1.3 Alignment Quality

As mentioned in Section 4.1.1, the alignment for 30 general thesaurus concepts and 20 geography concepts with DBpedia was evaluated by experts at HR. The results are aggregated in Table 3 and grouped by thesaurus, as well as by concept type (for general thesaurus concepts).

	Concepts	Good alignment	Adequate alignment	Bad alignment	At least one good alignment	At least one adequate al.
Class	3	100%	0%	0%	100%	100%
Descriptor	13	44.4%	22.2%	33.3%	53.9%	84.6%
Free descriptor	14	26.9%	58.7%	14.4%	71.4%	100%
General thesaurus total	30	36.5%	43.7%	19.8%	70%	93.3%
Geography thesaurus total	20	60.9%	25%	14%	90%	100%

Table 3: Alignment quality results

For every group of concepts, the table first shows the overall aggregated distribution of the alignment ratings. As mentioned in Section 4.1.1, the three participants had to rate each alignment, of which there can be more than one per thesaurus concept, on a three-point Likert scale. For the general thesaurus concepts, the found matches were considered to be accurate in 36.5% of the time, and adequate in about 44% of alignments. The alignment of geography concepts did quite a bit better, with about 61% good matches and only 14% unsuitable ones.

The rightmost two columns of the table indicate the percentage of concepts with at least one Linked Data match that was asserted to be accurate or adequate, respectively. In this case, the criterion for a match to be qualified as good is that at least two out of the three experts evaluated it as such. Similarly, for a match to be at least adequate, it must have received a minimum of two corresponding or better votes. In this definition, a good match is thus also an adequate one. For the general thesaurus, for 70% of the concepts a good match was found, and another 23% concepts are above the adequacy threshold. Equally high results were reached in the geography alignment, with 90% and 100% of the concepts with one good, or at least one adequate match, respectively. The divergence between these values and the overall rating statistics can be attributed to a somewhat lower recall.

The alignment evaluation results for general thesaurus concepts are furthermore grouped according to the type of concept. In the general thesaurus, a type is assigned to a concept with regards to how it is described: if a concept has narrower terms, it is a class, and otherwise a descriptor. A large amount of concepts come with no further information than their label; these are free descriptors.[59] For class concepts, all of the matches were evaluated as accurate, although it should also be mentioned that only three class concepts were part of the evaluation set. Descriptor concept matches scored significantly lower; 44% of the matches were good, 22% adequate and 33% unsuitable. For 54% of the descriptor concepts, at least one good match was found, and for 85% of them at least an adequate one. Finally, free descriptors matches were good 27%, adequate 59%, and inaccurate 14% of the time. However, for 71% of them a good match was found and an at least adequate match was made for 100% of free descriptor concepts.

4.2 Keyword Extraction Evaluation

After the thesaurus alignment, the keyword extraction establishes the other part of the connection between broadcasts and Linked Data information. In order to be able to perform useful faceted search of broadcasts, they need to be described accurately by the keywords that represent them. This section describes the evaluation of this accuracy and the corresponding results.

[59] The original German names of the concept are “Klasse”, “Sachdeskriptor”, and “freier Sachdeskriptor”.

4.2.1 Evaluation Method

The task of keyword extraction is to find the set of thesaurus terms that best describe a given broadcast's content in order to use them as tags for the latter. In this context, evaluating its results against a reference alignment is a valid option, although achieving complete recall might be a problem, as it was the case for the alignment evaluation. The gold standard would have to be absolutely accurate in order to be useful. Thus, the keyword extraction was also manually evaluated by experts. The same three participants from the alignment evaluation were involved in the assessment of the quality of the keyword extraction.

In this case, their task consisted in evaluating the tagging with thesaurus terms for 25 randomly selected broadcasts. For each broadcasts, they had to determine both the precision and the recall on a five-point Likert scale. The labels for the precision scale options went from "the keywords match the broadcast content very well" to "the keywords do not match the content at all" and for the recall scale from "the keywords cover the broadcast's complete content" the broadcast content is not at all covered by the keywords". The participants also had the possibility not to give a rating. In terms of context information, the broadcast summaries and content descriptions were given, as well as synonyms and broader terms for the concepts. Again, the complete evaluations are to be found in Appendix A.2.

4.2.2 Results

The graphs in Figure 10 and Figure 11 show the frequencies of the assignment of the precision values and the recall values, respectively. The values on the horizontal axis correspond to the Likert scale options, with 5 standing for the best precision (or recall) and 1 for no precision (or recall) at all. The vertical axis determines the frequency with which the individual ratings were given in relation to the totality of ratings.

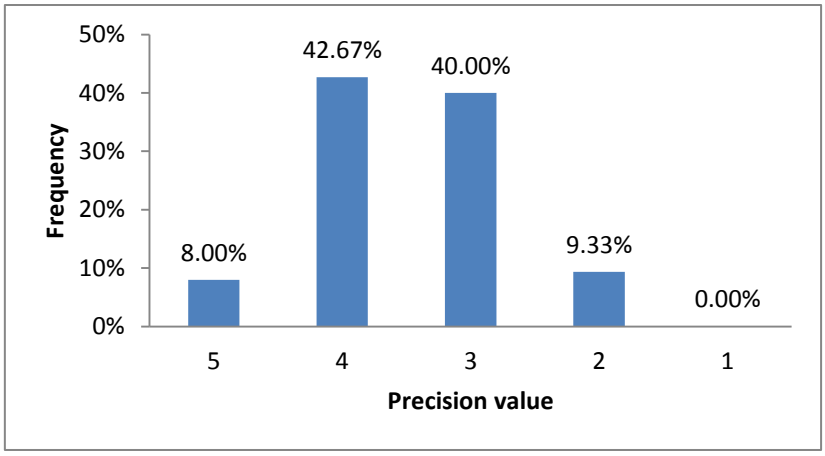

Figure 10: Aggregate precision values

While the set of extracted keywords were rarely found to perfectly describe a broadcast's content (8%), they were never judged to be completely unsuitable, and only 9% of the time did just a few of them match the content. Most often, almost all (43%) or some (40%) keywords were assessed to match the broadcast's content.

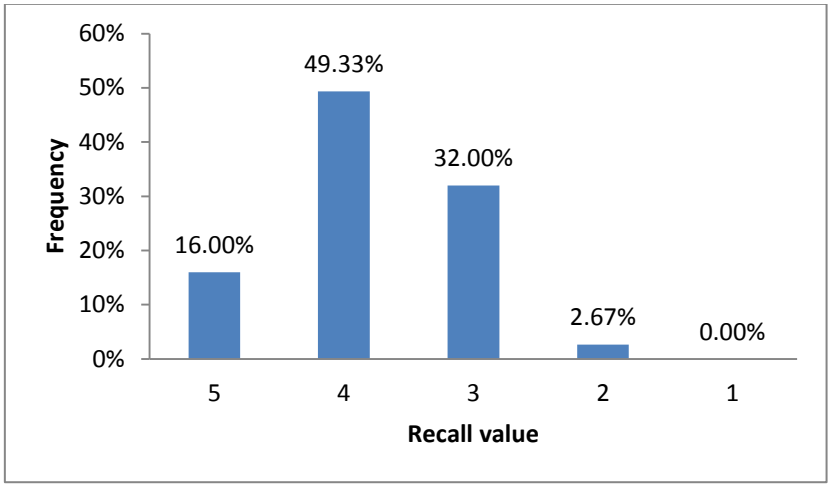

Figure 11: Aggregate recall values

The recall values indicate how exhaustively the keywords describe the content. In 16% of the time, the keywords were determined to completely cover the content they were to describe. Another 49% of the keyword sets were found to include almost all of the content, and 32% to be partly complete. Finally, in less than 3% of the time did the keywords cover almost none of the broadcast content. The lowest recall value was never assigned.

	5	4 or above	3 or above	2 or above	1 or above
Precision	0%	40%	92%	100%	100%
Recall	8%	64%	100%	100%	100%

Table 4: Keyword extraction quality results

Taking a closer look at the evaluation on the individual broadcast level, Table 4 shows the cumulative percentages of broadcasts that scored at or above a certain value, according to at least two out of the three experts. While for no broadcast a perfect precision was achieved, 40% of them scored at the second highest precision level and only 8% below what could still be considered an adequate precision. In terms of recall, for the totality of the broadcasts in the evaluation set did the keywords cover the content at least adequately. In 64% of the cases, the keywords were quite exhaustive, and for 8% of the broadcasts, they covered all of the content.

4.3 Faceted Search Evaluation

The goal of the faceted search evaluation is to test the suitability of this search paradigm for the chosen use case at Hessischer Rundfunk. An offline version of the implemented client was evaluated through a series of interviews with users. This section includes a description of the method and of the questionnaire that were used for this. Finally and most importantly, the results of the evaluation are presented.

4.3.1 Evaluation Method

The user interviews were conducted face-to-face and individually. First, they were asked some personal information regarding the precise nature of their occupation, in particular in relation with FESAD, and their technical skills. The next part of the interview involved the users performing two specific tasks with the FESAD client they typically work with, as well as answering some questions about the tasks and their satisfaction with the results.

The last part of the interview was centered on the faceted search paradigm. The users were first introduced to what exactly faceted search is and how it works. To do so, queries were demonstrated using the Faceted Wikipedia Search and a static version of the faceted search client implemented with Cliopatria. Specifically, the queries corresponded to the ones they previously had to perform on the FESAD client. The goal was to illustrate the potential advantages of a faceted search over a

traditional one. The users were then asked some questions in order to establish whether these benefits would translate positively to their work process. The specifics of the questions that were asked are presented in the next section.

4.3.2 Questionnaire

In this section, the complete questionnaire according to which the interviews were conducted is presented in detail, as well as some explanations regarding the specific purpose of the questions. After some personal questions, the users were asked about their experience with the FESAD client they currently use, in particular with regard to the frequency and purpose of their usage.[60] The users were also inquired about their personal opinion of the FESAD interface. The goal here was to find out how well it generally performed in the support of their day-to-day tasks. They were asked to estimate the FESAD client's precision and usability, as well as whether they had suggestions for improvement.

Next, the users were asked to execute two queries on the FESAD client. In the first task, the users had to find broadcasts about jaguars (the cat species), in the second one broadcasts involving cities in the state of Hesse with a particular local majority party. These two particular queries were chosen because they are assumed to address the strong suits of faceted search, namely the precise search for an ambiguously labeled entity and a complex search involving multiple filtering steps, respectively. The users had to run these queries on the FESAD client in order to establish a basis of comparison with faceted search. The users' approaches in solving these problems were kept track of, as well as their opinions on the quality of the results and on the effort necessary to perform the queries. They were also asked if these types of queries occurred frequently in their daily work.

In a subsequent phase, the idea behind faceted search was explained and shown, using examples on the Faceted Wikipedia Search (see Section 2.2.3) and on the implemented faceted search client. In particular, the latter was used to showcase how the very same queries the users were previously asked to perform on the FESAD client would be done with a faceted search.[61] The "jaguar" ambiguity would be solved by selecting the "animal" type in the proposed facets, and the second task by entering "hessen" in the text field and then filtering the results with the type facet value "city" and by selecting the political party in question for the facet "majority party".

[60] For a comparison with faceted search, specifically which client (FESADred or FESADdigital) is used is not of great importance, since both of them rely roughly on the same search mechanisms.
[61] The reason why the queries were only demonstrated (as opposed to performed by the users themselves) is that they require a very specific input in the offline version of the faceted search client.

Based on this demonstration, the users were asked their opinion on the usefulness of faceted search in their day-to-day, in particular when performing potentially ambiguous or complex queries. Furthermore, they were inquired on how understandable and practical they found faceted search to be, and whether they would use it, if it were an additional query option. Finally, the users had the chance to make a general remark about faceted search. All of the results are presented in the following section. For a more report of the interviews, refer to their transcripts in Appendix A.4.

4.3.3 User Interview Results

In total, seven users with various backgrounds were interviewed. Five of them were editors, the other two archivists. The latter two estimated their own technical skills to be rather good, whereas most others qualified them to be at an average level or limited.

Four out of the seven interviewees, all of them editors, claimed to work most frequently with FESADred, the other three were FESADdigital users. In general, they used their FESAD client at least on a daily basis. In contrast, the replies to the question for the reasons they used it for were less homogeneous. The tasks that were said to be performed on the FESAD client were: To look for movies or shows to include in the broadcasting program, to find possibly content created by other broadcasters, to test and evaluate the system, to look for footage containing a particular subject, or to perform queries that are focused on technical metadata. In summary, three interviewees said to use the FESAD client mainly for directed searches with specific results in mind, where the other four most frequently used it to perform explorative queries, for example about a particular topic.

FESAD Search Client

Six of the interviewees were very satisfied with their FESAD search client's precision. Only one user expressed some discontent with the way topic-oriented queries are handled, as often no results at all are returned. The users however also acknowledged that the precision of a query is strongly dependent on its strategy. In particular for explorative queries some ex-ante knowledge of the metadata is required in order to be able to achieve precise query results. Directed search, however, was said to work very well. With regard to usability aspects, the users were also all pleased. Some minor improvement suggestions were mentioned, but all in all, they found both of the FESAD clients to satisfy all of their needs.

For the first query task, all of the participants started by a simple search with the term “jaguar”. Apparently, the quite large result list also contained broadcasts about subjects other than the animal.[62] The interviewees reacted in various ways to this fact: Four of them directly attempted to filter out bad results, while the other three said they would traverse the individual results manually and look for suitable results. Filtering was done by trying to add inclusive search criteria, such as the addition of the search term “animal” or “zoo”, or by excluding terms such as “car”. While this did narrow down the result set, the interviewees never stated to really be happy with it, because it either contained many false positives, or they felt they excluded many suitable results with their query. Five users described the effort to perform this query to be quite large, and some said that browsing through a large amount of results is evitable with this kind of queries. The same number of interviewees said that such ambiguous searches were a frequently recurring part of their work.

The second task involved the participants finding broadcasts about cities in the state of Hesse with CDU as their local majority party.[63] Many different approaches were taken to try to solve this task. One approach three users said they would use involved consulting an online search engine such as Google to get the required information and narrow down the set of cities, before actually performing the query on the FESAD interface.[64] The other four attempted to express their query in FESAD with combinations of search terms such as “hesse”, “city”, “cdu”, or “majority”.[65] Only one of these participants would proceed to manually browse the resulting broadcasts, the other three evaluated the results to be inconclusive and would use a different information channel to find out what cities to search for. Evidently, all of the interviewees were discontent with the effort this query required and came to the conclusion, that it cannot be performed on the FESAD client. However, only for one user are such complex queries on the FESAD client a regular occurrence, the other participants all already know precisely what to search, in part also because they look up the corresponding information beforehand.

Faceted Search

After seeing a demonstration of faceted search, the participants were asked how useful it would be in their daily work to perform ambiguous or complex queries. For the former type of query, five of them said that it would be of great or very great use to make certain searches simpler and faster. The other two either see no additional benefit in faceted search, or never perform ambiguous queries.

[62] Because the interviewees performed their searches with different tools on different datasets, the size of this result list varied, which is also why no number is given here. Result list sizes ranged from 13 to 360.
[63] CDU stands for Christlich-Demokratische Union. This party was chosen for no particular reason.
[64] http://www.google.com
[65] Obviously, the participants actually used the German-language equivalents of these terms.

Regarding complex queries, four users thought that faceted search would be beneficial, although two of them said it would also depend on the usefulness of the facets. In general, the users found faceted search to be both easily understandable and accessible. But again, some mentioned that the quality of the facets would play an important role, since they would then guide the search. Five interviewees said they would use or at least test faceted search if it were an option. Two of them would however also require it to support all the currently available search options.

Finally, some general remarks about faceted search were made. One interviewee voiced concerns with regard to the validity of the background information, which would be a crucial factor for faceted search's usefulness, in particular when a user does journalistic research with it. Another participant said that the faceted search possibility should not create additional work for archivists in creating content-describing metadata, that is, that the task of extracting keywords should definitely be automated.

5. Related Work

In this chapter, approaches related to the one presented in this thesis are described. More specifically, existing solutions in the domains of keyword extraction from natural language texts, ontology alignment, and faceted search systems are examined, as well as a few initiatives to create and connect existing datasets with Linked Data.

5.1 Keyword Extraction Systems

In the context of this thesis, a method was developed to extract controlled vocabulary terms from natural language texts and corresponding tags simultaneously. This bears many similarities to text mining, which Tan (1999) defines as "[…] the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents". While text mining approaches are useful to determine a text's most essential terms, their application on broadcast descriptions would not have been suitable.

First, because only one such tool which assigns keywords from a given controlled vocabulary could be found, namely KEA++ (Keyphrase Extraction Algorithm). KEA (Medelyan & Witten 2006) uses a set of documents with manually chosen keyphrases (keywords consisting of multiple words) for training. This training involves learning certain features from the reference document, such as the position of a first occurrence of a keyphrase or its length in words. This information is then applied with machine learning techniques to determine the probability for each candidate term in the document that it is a suitable keyphrase. KEA++ was tested with broadcast descriptions, but achieved unsatisfactory results. A reason for this may be that broadcast summaries and image descriptions are rather inconsistently structured and typically short. KEA++ was also determined to be unsuitable because it did not feature the possibility to extract keyphrases from multiple sources and to assign weights to them.

5.2 Ontology Alignment Tools

Ontology alignment is a problem that has been tackled many times by researchers. In order to compare and evaluate all the different approaches, the Ontology Alignment Evaluation Initiative (OAEI) was created.[66] It runs yearly evaluation campaigns with test cases which the submitted alignment systems have to solve. Usually, multiple tracks are proposed in order to evaluate systems according to different metrics. The alignments submitted by participating systems are compared with a reference alignment, a so-called "gold standard", and precision and recall computed on this basis. A few of the systems that were evaluated in recent OAEI campaigns are briefly reviewed here.

Falcon-AO (Hu & Qu 2008) is an ontology matching system consisting of five components which parse input ontologies, run a set of matching algorithms, generate alignments, allow manual configuration of matching algorithms, and store data about the matching process, respectively. Most notable in the set of matchers is V-Doc, which uses local, as well as neighboring descriptions of ontology entities to compute similarities, and GMO, an "iterative structural matcher" which generates a measure of similarity between domain entities and statements using external alignments as input. The similarity measures from such matchers are finally combined in a configurable way and are used in the generation of 1:1 alignments. Falcon-AO participated in OAEI campaigns from 2005 to 2007 and scored particularly well in both precision and recall.

AgreementMaker (Cruz et al. 2009) took part in the three OAEI campaigns since 2009 and was among the best systems for the alignment of most data sets in 2011 (Gracia et al. 2011). Like Falcon-AO, AgreementMaker makes use of a variety of matching methods, the application of which is also configurable. These matchers are organized in a sequence of three steps, called layers. The first layer computes similarity measures for all the entities in the two input ontologies using syntactic and lexical comparison algorithms. While this first layer matching only takes the entities' feature into account, a second layer compares their structural properties. The third layer then combines these results to generate alignments. To optimize users' control of the alignment process, a graphical user interface allows them to manually manage the matching methods and visualizes resulting alignments.

RiMOM (Risk Minimization-based Ontology Mapping) (J. Li et al. 2009) uses Bayesian decision theory to express ontology alignment as a decision problem. Different mapping strategies are regarded as sub-decisions, which compute a risk value for every mapping. That is, for every entity pair and mapping strategy, a risk value is computed and stored in a so-called "predicting cube". These risk

[66] http://oaei.ontologymatching.org/

values in turn come together to inform a final alignment decision, where the mapping with minimal risk is selected. Sub-decisions include mapping strategies based on entity names, instances, textual descriptions, and taxonomy structures. The RiMOM system participated in OAEI campaigns from 2007 to 2010.

Another way to combine matching techniques is implemented in X-SOM (eXtensible Smart Ontology Mapper) (Curino et al. 2007). A three step process is used for the alignment: First, ontology entities are matched with each other using different matching strategies encapsulated as modules. These perform both syntactical and structural comparisons, and since each strategy has its own strengths and weaknesses, the results are weighed by means of a neural network. In a second step, alignments are filtered, a decision that can optionally be delegated to a user for mappings with similarity values below a certain threshold. In a final automated step, inconsistencies resulting from the prior matching and filtering are resolved.

5.3 Faceted Search Systems

Faceted search has been a research interest since before the advent of Linked Data, with the goal of facilitating the browsing across structured information. A notable example of this is the Flamenco Search Interface Project (M. A. Hearst 2000). It includes multiple efforts to study how metadata can be used to guide user searches, for instance in the context of image searches (Yee et al. 2003). This latter effort is highly relevant in the context of this thesis, as it not only presents an approach to create a facet-supported search interface, but also an approach to structuring the metadata in a useful way. A user study was furthermore conducted, which found that participants rated the tested faceted search client very positively, finding it easy to use and to browse, although many were frustrated by its slowness.

The mSpace project (Schraefel et al. 2003) is mainly focused on the design aspects in relation with faceted search. It features an extensive study of users' requirements and difficulties on traditional Web search tools, which discovered that they experienced problems with articulating the query and thus finding useful resources. Furthermore, a facet-based search interface with a multi-column layout is proposed, where columns represent facets and are populated with corresponding facet values.

In the context of Linked Data, there also exist several approaches to performing faceted search. Parallax (Huynh & Karger 2009) is an implementation of a faceted user interface on top of Freebase data.[67] Its layout and query process are very similar to the faceted search system implemented in the context of this thesis. It is started by a keyword search, after which results can be filtered out by the selection of facet values. The Parallax project also features a browser extension which detects concepts in the currently viewed Web page and allows the user to perform faceted search with them using Freebase data.

Faceted Wikipedia Search (Hahn et al. 2010) was already mentioned in this thesis.[68] It implements a classical faceted search interface on top of DBpedia data. Alternatively to a keyword search, it also allows the user to start a query by selecting a facet value from a set of facets that are displayed by default. Facet values are ordered according to their frequency in the current result set.

The previously presented faceted search clients differ from the one implemented for this thesis in that they allow the user to directly query Linked Data, as opposed to content that is indirectly semantically enriched by means of its connection with Linked Data concepts.

5.4 Creation of Linked Data

Some of the efforts with regard to the creation of Linked Data are described in this section.

The creation of DBpedia data from Wikipedia information is described in (Auer et al. 2007). Specifically, the structured descriptions contained in the infoboxes of Wikipedia articles are transformed into RDF using extraction algorithms. The resulting resources are then linked with other datasets in a community effort.

In (Troncy et al. 2010), a method is presented to interlink event information with Linked Data. Three public event directories on the Web (last.fm, upcoming, and eventful) are transformed into OWL using a particular event ontology, LODE (Shaw et al. 2009).[69] Furthermore, a user interface is built on top of the resulting data.

[67] The faceted search client can be tested at http://www.freebase.com/labs/parallax/ .
[68] The Faceted Wikipedia Search can be tested at http://dbpedia.neofonie.de/browse .
[69] http://www.lastfm.com/, http://upcoming.yahoo.com/, http://eventful.com, and http://linkedevents.org/ontology/

Haslhofer & Isaac (2011) present data.europeana.eu, an effort to make the data from Europeana, a collection of metadata for cultural and scientific heritage, available as Linked Data. The data is converted from a well-structured format to RDF, according to the schema defined in the Europeana Data Model (Isaac 2011) using a specific mapping.

Finally, some approaches to connect media metadata to Linked Data should also be reviewed. The Linked Movie Database (Hassanzadeh & Consens 2009) is a project with the goal of creating RDF data about movies and adding it to the Linked Data cloud. Several sources of information, such as IMDb (Internet Movie Database), Freebase, or DBpedia, are integrated and converted to RDF.[70] In order to match movies between datasets, string similarity measures are applied.

BBC Programmes and BBC Music are part of a project to publish the British broadcaster's television program and music-related as Linked Data (Kobilarov et al. 2009). This includes approaches to convert existing data structures to RDF, interlinking them, and presenting the information as XHTML for human interaction. The interlinking step is done by connecting concepts to equivalent DBpedia which was achieved mainly through label matching.

[70] http://www.imdb.com/

6. Discussion

The purpose of this chapter is to determine how well the objectives introduced in Section 1.2 were achieved, namely to develop a method to tag media metadata with Linked Data concepts and to use this connection to support user queries by providing faceted search, as well as to test whether faceted search could bring about benefits in a practical setting. This is done on the basis of the evaluation from Chapter 4. A first part discusses all the aspects relative to the implementation, whereas the second part analyzes how suitable faceted search for the practical purposes described in this thesis.

6.1 Implementation

The implementation of the connection of broadcast metadata with Linked Data features numerous design decisions. While it is outside the scope of this thesis to compare them with other possible designs, the evaluation results should be briefly discussed, as well as some potential improvements regarding the implementation. In this section, the evaluation of the thesaurus alignment and the keyword extraction are reviewed. Then, the implementation of the faceted search client, and finally some general considerations about the linking method presented in this thesis are discussed.

6.1.1 Thesaurus Alignment

Similarly to most examined alignment tools (see Section 5.2), the implementation of the thesaurus alignment was split in a lexical and a structural matching phase. As the evaluation showed, the different label matching approaches encountered various levels of success. Comparing the labels in a URL-converted form showed to be useful, since a certain amount of matches was made this way. Using broader terms to match concepts is however a more debatable decision. On the one hand, they evidently do not directly represent their concept and a match that is made via a broader term will not be entirely accurate. However, on the other hand such a match is better than none at all, since the concept's domain is at least represented in the corresponding Linked Data concept. The comparison of label beginnings and endings apparently resulted in a large number of matches. Although no numbers are available to verify this claim, many of these matches seem to have been

discarded in the second phase of the alignment, raising the question of how accurate this type of match really is.

The overall percentages of thesaurus concepts that were successfully matched vary widely depending on the alignment type. The alignment of geography concepts with the GeoNames vocabulary managed to only find matches for 16% of these concepts. Reasons for this are difficult to pinpoint, but since similar amounts of matches were made in the first alignment phase, the assumption is justified that only a small fraction of them scored above the defined threshold. It remains to be investigated whether the threshold of acceptable match values was set too high or the accuracy of the matches was really that poor.

In contrast, the alignment of geography concepts with DBpedia did much better, as 58% of them were matched successfully. This excludes the possibility that these concepts are insufficiently described in the thesaurus as an explanation for the poor results in the GeoNames alignment. For general thesaurus concepts, 42% could be matched with at least one DBpedia resource. It is difficult to determine whether this percentage is satisfactory. At any rate, an alignment of 100% would be very hard to reach, as many thesaurus concepts are free descriptors, i.e., with no information other than their label, or are of a very specific nature and thus not described in DBpedia.

The suspicion that free descriptors are more difficult to match is confirmed by the qualitative alignment evaluation, according to which only 27% of the matches found for them are completely accurate. But surprisingly enough, for a larger part of them at least one good alignment could be found than for regular descriptors (71% of free descriptors versus 54% of descriptors). The reasons for this phenomenon could not be determined. Concepts of the class type were all matched accurately. This could be due to the nature of the concepts they describe, which are perhaps more general than other concepts, and thus described more exhaustively.

For 70% of the general thesaurus concepts with matches, at least one good alignment was found, meaning that globally, this is true for 29% of all the general thesaurus concepts. Again, this figure appears to be quite low, yet it might be sufficient to link most broadcasts to useful information, assuming a long tail distribution of keyword frequencies. The alignment of geography concepts is however acceptable, with 90% of good and 100% of at least adequate alignments.

Additional possibly beneficial matching strategies that could be investigated in the future are lexical similarity measures such as Levenshtein distance and methods to more intricately compare structural resemblances.

6.1.2 Keyword Extraction

As the evaluation showed, the keyword extraction performs well. While the precision of the keywords was not perfect for any broadcast, almost for all of them were the extracted keywords at least adequate. The recall results were similar, as all the broadcasts were covered at least adequately. One factor in the quality of the results is obviously the implementation of the keyword extraction. Some adjustments could be investigated in terms of mining textual descriptions and weighing the three information sources of a broadcast, which are tags, summary, and content description. But an almost equally important factor is the quality and completeness of the existing textual descriptions and tags. Furthermore, the vocabulary of this information has to correspond to the thesaurus, as it cannot be used otherwise. Hence, an obvious suggestion for improvement would be to consistently assign tags from a controlled vocabulary. This would also avoid any possible ambiguities.

6.1.3 Faceted Search Client

The implemented faceted search system is intended to be a demonstration prototype. While the source code of the markup and of the method to retrieve the relevant data is not necessarily meant to be reused as such, it illustrates a strategy to combine broadcast metadata with Linked Data in a faceted search application. In terms of performance, the chosen approach is yet to be examined for optimization possibilities. Similarly, the layout needs to be adapted to the users' particular requirements, especially regarding how and what broadcast metadata is presented in a result list. In order to better support the users, it might furthermore be of interest to examine which facets are likely to be used the most frequently and which ones should preferably never be displayed.

6.1.4 Linking Method Discussion

Aside from the evaluation of the individual steps towards connecting broadcasts with Linked Data, some general aspects of the linking method need to be discussed. One of them is the choice of Linked Data vocabularies to connect the controlled vocabulary to. DBpedia is a good choice, as its data is up-to-date and covers a wide range of topics. Also, its numerous links can be used to retrieve information from many other knowledge bases. Still, one should also be aware of the fact that DBpedia may have some issues, such as redundant entries (see Section 4.1.1). If the extent or the detail level of the information contained in DBpedia is not sufficient, linking to more specialized vocabularies can also be considered.

However, as seen in this thesis, a generic approach to aligning a controlled vocabulary with Linked Data is difficult to devise. One option would be to further extend the structural matching aspect of the alignment process in order to be less dependent on the particular vocabularies' semantics. The problem is however with the thesaurus, which does not provide a great amount of context, and thus structure for its concepts, if any at all. The approach chosen in this thesis for the DBpedia alignment, namely to use concept abstracts for the structural matching, is not very likely to be suitable for the alignment with other Linked Data vocabularies.

Another important decision is what to use as a controlled vocabulary. The ARD thesauri seem to be a good fit for this task, as they are extensive, well-structured, and can be managed within the organization. Within the ARD vocabularies, further thesauri exist, containing for instance information about people and institutions. Depending on the application, the use of those is also worth considering. This decision would then be faced with a trade-off between the benefits of additional information and the effort involved in the creation of an alignment for another thesaurus.

Practical Considerations

Since the context of this thesis is the tagging of media metadata with linked concepts for a real-life use case at Hessischer Rundfunk, some practical considerations also seem appropriate. If this project were actually to be realized, the end product would obviously not be a stand-alone search client. The faceted search facilities would rather be integrated into the current FESAD environment, not only because of the proximity to the relevant media metadata, but also because the tasks performed with query results would have to remain the same. In this case, some implementation aspects would evidently have to be revisited. For one, the conversion of broadcasts to RDF/XML might not be necessary anymore, as they could probably be used in their native format. However, the question of how to integrate the additionally required data, such as the controlled vocabulary and Linked Data resources, would present itself. In particular, this might be an issue with Linked Data, as not all of it is available as downloadable dumps or in the same RDF serialization format.

Another aspect to take into account is the workflow design. First, an alternative route to go with the tagging of broadcasts with controlled vocabulary keywords is to perform this task manually and instead of the allocation of traditional tags. This would obviously represent a definite commitment to the new search paradigm and a step away from the previous query habits. A semi-automatic approach, where the keyword extraction is verified by human experts, is also conceivable. A second aspect to consider is how the alignment with Linked Data is integrated in the workflow. The best solution for it is to be static and run offline, as it is a very expensive task.

Performance is also an important factor to take into consideration for an application that is meant to be used in a professional environment, in particular for a faceted search system which, in addition to retrieving results, has to generate facets and corresponding values at run-time. In the context of this thesis, it was however not a central concern and remains to be optimized.

6.2 Suitability of Faceted Search

Aside from any aspects pertaining to the implementation, a central objective of this thesis is to determine whether a practical benefit can be gained at HR from using a faceted search rather than a traditional one. In the evaluation, almost all of the users initially stated to be satisfied with their current FESAD search client, apart from a few minor complaints. They however also admitted that some ex-ante knowledge is required to be able to achieve a good result precision. Also, as the query task they were asked to perform showed and the interviewees themselves said, the effort that has to be invested into a query can occasionally be relatively high, especially if it is an explorative one.

From the results, it can be concluded that the FESAD search clients fulfill the needs of their users very well in terms of performing their most frequent task, which is running directed queries. With regard to more complex explorative searches, the users have to come to terms with the fact that large sets of results often have to be browsed to find relevant broadcasts. This undoubtedly causes inefficiencies and is thus an aspect that would need to be improved.

The question is then whether faceted search is the approach that can provide the required improvements. While it is beyond the scope of this thesis to verify it empirically, the results from the user interviews seem to indicate that it is indeed the case. According to four out of the seven participants, faceted search would be useful for complex queries involving multiple criteria, and five participants considered filtering query results with facets to be beneficial addition to current query options.

7. Conclusion and Outlook

In this thesis, a method was described to connect media metadata with Linked Data. Specifically, individual broadcasts were tagged with keywords from a controlled vocabulary, which themselves were aligned with equivalent Linked Data resources. Thus, broadcast descriptions were by proxy interlinked with semantically enhanced information. This method was implemented by using a SKOS thesaurus as the controlled vocabulary. For the keyword extraction, an algorithm was created to find the most suitable set of vocabulary terms to describe each broadcast. As the evaluation indicated, it performed rather well, as it found at least adequate sets of keywords for almost all of the broadcasts.

To align the thesaurus concepts with Linked Data, the particular requirements also dictated the use of a custom implementation. This alignment process consists of a label matching and a context matching phase. In the former, thesaurus and Linked Data concept labels were compared to find candidate alignments. In the second phase, these were tested using contextual information in the form of abstracts for DBpedia concepts and additional features for GeoNames concepts. The DBpedia alignment was evaluated qualitatively by experts, which showed that it achieved its best results with geography concepts, but also found good matches for most general thesaurus concepts.

Furthermore, a method to query broadcasts by making use of their connection with Linked Data was demonstrated with the implementation of a faceted search client. In user interviews, it was found out that the advantages of the faceted search paradigm could also apply for the investigated use case at Hessischer Rundfunk, in particular to improve precision for searches with ambiguous search terms.

Future Work

In future work, methods to improve precision and recall for the alignment need to be investigated. Some possibly promising approaches would be to include further label matching and structural matching techniques or to align the thesaurus concepts with other Linked Data vocabularies. It also remains to be examined whether the inclusion of additional thesauri furthermore benefits the quality of the keyword extraction. With regard to faceted search, its performance needs to be evaluated empirically in order to compare it directly to the traditional search facilities. The usability of the search interface is also an aspect that requires further elaboration and needs to be developed in close collaboration with its users.

Furthermore, if the system were to be implemented for productive use, some practical aspects would require attention, such as a specific solution for the integration into the existing system architecture and the optimization of the query efficiency.

8. Bibliography

Angles, R. & Gutierrez, C., 2008. The Expressive Power of SPARQL. In A. Shet et al., eds. *The Semantic Web - ISWC 2008: 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008, Proceedings*. Springer, pp. 114–129.

Auer, S. et al., 2007. DBpedia : A Nucleus for a Web of Open Data. In *Proceedings of the 6th International Semantic Web Conference, Busan, Korea.*

Berners-Lee, Tim, 2006. Linked Data - Design Issues. Available at: http://www.w3.org/DesignIssues/LinkedData.html [Accessed August 20, 2012].

Berners-Lee, Tim, Hendler, J. & Lassila, O., 2001. The Semantic Web. *Scientific American*, 284(5), pp.34–43.

Bizer, C., Cyganiak, R. & Heath, Tom, 2007. How to publish Linked Data on the Web. Available at: http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/ [Accessed August 20, 2012].

Bizer, C., Heath, T & Berners-Lee, T, 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), pp.1–22.

Bollacker, K. et al., 2008. Freebase : A Collaboratively Created Graph Database For Structuring Human Knowledge. , pp.1247–1249.

Cimiano, P. & Völker, J., 2005. A Framework for Ontology Learning and Data-driven Change Discovery. *Architecture*.

Cruz, I.F., Antonelli, F.P. & Stroe, C., 2009. AgreementMaker : Efficient Matching for Large Real-World. *PVLDB*, 2(2), pp.1586–1589.

Curino, C., Orsi, G. & Tanca, L., 2007. X-SOM : A Flexible Ontology Mapper. *DEXA Int. Workshop on Semantic Web Architectures For Enterprises (SWAE)*.

Deutsches Rundfunkarchiv, 2011. *ARD Flyer*.

Drumond, L. & Girardi, R., 2008. A Survey of Ontology Learning Procedures. In F. Freitas et al., eds. *Proceedings of the 3rd Workshop on Ontologies and their Applications*. São Luis, Brazil.

Gracia, J., Bernad, J. & Mena, E., 2011. Ontology Matching with CIDER : evaluation report for OAEI 2011.

Gross, A. et al., 2008. *Regelwerk Mediendokumentation Fernsehen*.

Gruber, T.R., 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5, pp.199–220.

Hahn, R. et al., 2010. Faceted Wikipedia Search. *Business Information Systems*, 47, pp.1–11.

Haslhofer, B. & Isaac, A., 2011. data.europeana.eu - The Europeana Linked Open Data Pilot. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*. pp. 94–104.

Hassanzadeh, O. & Consens, M., 2009. Linked Movie Data Base Data sources. *Proceedings of the 2nd Workshop on Linked Data on the Web*.

Hearst, M.A., 2000. Next Generation Web Search : Setting Our Sites. *IEEE Data Engineering Bulletin*, 23(3).

Hu, W. & Qu, Y., 2008. Falcon-AO: A practical ontology matching system. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3), pp.237–239.

Huynh, D. & Karger, D., 2009. Parallax and Companion : Set-based Browsing for the Data Web.

Isaac, A., 2011. Europeana Data Model Primer.

Kobilarov, G. et al., 2009. Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections. In *Proceedings of the 6th European Semantic Web Conference*. pp. 723–737.

Levenshtein, V., 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10.

Li, J. et al., 2009. *RiMOM: A Dynamic Multistrategy Ontology Alignment Framework*, Published by the IEEE Computer Society.

Medelyan, O. & Witten, I.H., 2006. Thesaurus based automatic keyphrase indexing. *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries - JCDL '06*, pp.296–297. Available at: http://portal.acm.org/citation.cfm?doid=1141753.1141819.

Miller, G.A., 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11), pp.39–41.

Navigli, R. & Velardi, P., 2010. From glossaries to ontologies: Extracting semantic structure from textual definitions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*. pp. 216–225.

Navigli, R. & Velardi, P., 2004. Learning Domain Ontologies from Web Sites. *Computational Linguistics*, 30(2), pp.151–179.

Obitko, M., 2007. Ontologies and Semantic Web. Available at: http://obitko.com/tutorials/ontologies-semantic-we [Accessed August 20, 2012].

Schraefel, M.C., Karam, M. & Zhao, S., 2003. mSpace: interaction design for user-determined, adaptable domain exploration in hypermedia. In *Proceedings of Workshop on Adaptive Hypermedia and Adaptive Web Based Systems*.

Shamsfard, M. & Barforoush, A.A., 2003. The State of the Art in Ontology Learning : A Framework for Comparison. *Knowledge Engineering Review*, 18(4), pp.293–316.

Shaw, R., Troncy, R. & Hardman, L., 2009. LODE : Linking Open Descriptions of Events. *4th Asian Semantic Web Conference*.

Shvaiko, P. & Euzenat, J., 2012. Ontology matching : state of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 99.

Tan, A., 1999. Text Mining : The state of the art and the challenges Concept-based. In *Proceedings of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases*. pp. 65–71.

Troncy, R., Malocha, B. & Fialho, A., 2010. Linking Events with Media. *6th International Conference on Semantic Systems*, pp.2–5.

Tunkelang, D., 2009. *Faceted Search,* Morgan and Claypool.

Vatant, B., 2012. GeoNames Ontology. Available at: http://www.geonames.org/ontology/documentation.html [Accessed August 20, 2012].

Winkler, W.E., 1990. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods*, pp.354–359.

Yee, K.-P. et al., 2003. Faceted metadata for image search and browsing. *Proceedings of the conference on Human factors in computing systems - CHI '03*, p.401.

Appendix

A.1 CD-ROM Content

The CD-ROM appended to the physical version of this thesis contains the following documents:

- The complete thesis in the PDF format (MastersThesis.pdf)
- The thesis' abstract in English (Abstract.txt)
- The thesis' abstract in German (Zsfsg.txt)
- The folder "Implementation" containing the following eclipse projects:
    - ContentLink contains the implementation of the keyword extraction
    - OntoCompare contains the implementation of the alignment
    - ThesaurusConverter contains the implementation of the conversion of the thesaurus to SKOS
    - ThesaurusLinker contains the implementation of the linking of thesaurus concepts with Linked Data
    - XMLConversion contains the implementation of the broadcast conversion to RDF
- The folder "Alignment Results" containing:
    - The complete results for all three kinds of alignments, including statistics and lists of unmatched concepts
    - The thesauri in their original format and in SKOS

A.2 Keyword Extraction Evaluation Results

A.2.1 Broadcasts

Here are listed the tagged broadcasts that were used for the evaluation of the keyword extraction. The list serves as reference for the precision and recall evaluations which are exposed in detail in the following section. For the evaluation, the users were also given synonyms and broader terms of the thesaurus concepts that were selected as tags for context, which are not part of this list.

Nr	Summary	Image Content	Extracted Keywords
1	Frankfurt. Ausstellung 'Magische Pflanzen' im Palmengarten.	(Name entfernt) erläutert Kräuter für Hexen Flugsalbe, Cannabis, Fliegenpilz, (Name entfernt) /	Gewürzkräuter Wolfgang (Ort) Frankfurt (Stadt) HexePflanze
2	Frankfurt. Kassel. Wiesbaden. Bereits vor seiner Ankunft in Deutschland Demonstrationen gegen den Staatsbesuch von US-Präsident (Name entfernt), der als menschenverachtender Kriegstreiber gesehen wird	Demonstranten = Menschen bei Kundgebung auf Straße (Hauptwache Frankfurt) Protestschilder = Schild "Not welcome (Name entfernt)" bzw "(Name entfernt)swim home" etc / Demonstranten bei Protestmarsch auf Straße (Kassel) benutzen Trillerpfeifen, Spruchband "Wir pfeifen auf (Name entfernt)" / Versammlung Demonstranten auf Luisenplatz (Wiesbaden) bei Protestkundgebung Schild "US World Domination Tour, Bombing a country near you" - Transparente /	Demonstration Strasse Präsident International USA
3	Frankfurt. Macht der Euro die deutsche Wirtschaft kaputt? Muß die Währungsunion zurückgenommen werden? Stellungnahmen.	Befragung Passanten auf Wochenmarkt Bockenheim / Geld Euro Scheine werden gezählt, 2-Euro Münzen / Grafik Euro-Kurs /	Währungsunion Wochenmarkt Befragung Frankfurt (Stadt) Euro
4	Leipzig. Urteil des Bundesverwaltungsgerichts in letzter Instanz, die Wartungshalle für den A380 darf am Flughafen Frankfurt gebaut werden.	Waldgebiet Bannwald / Statements Anwohner / Archivmat: Modell Wartungshalle Werft für A380. Grafik Lageplan Airbushalle. Schwenk von Gelände Flughafen über Wald geplanter Standort /	Luftfahrt Wald Graphik Bannwald Flughafen Ausbau
5	Langen. Restaurant (Name entfernt) bietet nur ein außergewöhnliches Menue, aber mit 26 Gängen, Menuepreis 119 Euro, Weine bis 1700 Euro.	Gebäude Restaurant (Name entfernt)/ Köche in Küche bereiten kleine Speisen / Gäste lesen Speisekarte / Teile des Menue werden serviert, Gäste beim Essen /	Küche Wein Gaststätte Euro Speise Luxus
6	Wiesbaden. Schleppende Umsetzung der beschlossenen Neuregelung des Abgeordnetengesetz hinsichtlich eigenfinanzierte Alterversorgung, bei gleichzeitiger Erhöhung der Diäten	Büro Arbeitsplatz Abgeordneter (Name entfernt) / Grafik "Mindestversorgung hessischer Abgeordneter" / Diskussionsgrundlage Cover Eigenfinanzierte Altersversorgung / Büro Landtagspräsident (Name entfernt)/ Versammlung Abgeordnete CDU Landtagsfraktion / Versammlung Abgeordnete SPD Fraktion /	SPD CDU Graphik Altersvorsorge Diäten
7	(Name entfernt). Sportler der SG	Frau Regina Nase zeigt schwer schließende	Bürgermeister

	(Name entfernt) befürchten den Einsturz der Mehrzweckhalle, das Dach würde sich im Winter absenken, die Halle sei baufällig, der Bürgermeister hält dagegen	Trennwand / Gebäude Sporthalle Schild "Jahrtausendhalle" - Teilnehmer Ortsbegehung, Ortsbesichtigung, in (Name entfernt)/	Einsturz Teilnehmer Winter Sporthalle Sicherheit Sportanlage Bauwerk
8	Berlin. Änderung des Lärmschutzgesetz in erster Lesung im Bundestag beraten, zulässiger Grenzwert um Flughäfen soll drastisch gesenkt werden, Auswirkungen auf Flughafen Frankfurt.	Studiogespräch (Name entfernt)/	Luftfahrt Lärmschutz Auswirkung Grenzwert Flughafen Gesundheitsgefährdung Lärm
9	Büdingen. Der Bildhauer (Name entfernt)nutzt ein altes Stadttor als Museum, 25 Millionen Jahre alte Sandrosen aus Schwerspat, die es weltweit nur in der hessischen Wetterau gibt.	(Name entfernt) mit Hund, Ausstellung Sandrosen, versteinerter Baum, Quarze, Fundstücke im Jerusalemer Tor /	Stadttor Tor Hund Büdingen (Ort) Bildhauer Sand Geologie
10	Kassel. (Name entfernt) gewann zum zweiten Mal einen Preis beim hessenweiten Schülerwettbewerb 'Faszination Technik', diesmal für eine Spargelschälmaschine.	(Name entfernt) bei Vorführung mechanisches Entmoosungsgerät auf Hausterrasse / beim Basteln in seinem Zimmer / zeigt Mitschülern Spargelschälmaschine / Preisverleihung Verein Deutscher Ingenieure an (Name entfernt)/	Auszeichnung Basteln Schülerwettbewerb Kassel (Stadt) Technik Erfindung Wettbewerb
11	Kassel. Zirkus (Name entfernt) gastiert in der Stadt, Reporter versucht sich als Hochseilartist	Seiltänzer = Artisten Trio (Name entfernt)auf Hochseil unter Zirkuskuppel /	Trio Seiltänzer Stadttor Zirkus Kassel (Stadt) Kleinkunst
12	Weiterstadt. Tödlicher Unfall wegen überhöhter Geschwindigkeit, Frontalzusammenstoß Auto stieß mit Linienbus zusammen, Schwerverletzte im Bus.	Unfallstelle Autowrack und beschädigter Bus auf Landstraße, eingepackter Leichnam wird abtransportiert /	Geschwindigkeit Weiterstadt (Ort) Landstrasse Kraftfahrzeug Unfall
13	Hofgeismar. (Name entfernt) betreut den Toilettenwagen auf dem Weihnachtsmarkt, um ihre Rente aufzubessern.	Weihnachtsmarkt vor Gebäude Rathaus Hofgeismar, (Name entfernt) am Toilettenwagen, Schale mit Trinkgeld, reinigt WC, Statements Kunden /	Weihnachten Rente Kunde Hofgeismar (Ort) Weihnachtsmarkt Armut
14	Frankfurt. Wachsendes Interesse an Privatschulen.	(Name entfernt) mit Mutter betritt (Name entfernt)-Schule / Erstklässler bei gemeinsamen Frühstück / Englischunterricht in erster Klasse / Kinder in Klasse beim Malen / Kinder in Nachmittagskurs (Name entfernt)-Schule / Kinder auf dem Weg zur Schule /	Kinder Mutter Frühstück Privatschule Male (Hauptstadt) Schulwesen
15	Flughafen Frankfurt. Vor dem großen Erprobungsprogramm für den Airbus A380.	Landeanflug Flugzeug ljuschin 67, Hobbyfotografen / Archivmat: landendes Flugzeug Airbus A380 / Besucherterrasse am Terminal 2 /	Luftfahrt Terminal Flugzeug Frankfurt (Stadt) Landeanflug
16	Kinderschutz durch Pflicht-Vorsorgeuntersuchungen für Säuglinge mit zentraler Datenspeicherung, jedes Neugeborene soll künftig digital erfaßt und seine Daten über Jahre gespeichert werden, mit einem Strichcode soll sichergestellt werden, daß die Eltern	Bromskirchen; Kerzen und Teddys zum Gedenken an (Name entfernt) vor Wohnhaus / Schild Fachärztin (Name entfernt)/ Gebäude Kreiskrankenhaus (Name entfernt)/ Säugling wird untersucht, gemessen, Blutabnahme /	Frankenberg (Ort) Kinder Daten Kerze Gedenken Sozialpolitik Datenschutz

	der Pflicht zur Vorsorge nachkommen.		Gesundheitsvorsorge
17	Frankfurt. Amateurtanzturnier "Hessen tanzt" - mittendrin (Name entfernt) und (Name entfernt) aus Darmstadt	Tanzpaar Jugendliche (Name entfernt)und (Name entfernt) und viele andere Tänzer auf Parkett in Halle, im Wettkampf Turniertanz, vor Publikum /	Frankfurt (Stadt) Amateursport Sportwettkampf Tänzer Halle (Ort) Tanzen
18	Darmstadt. Europa-Premiere der Tanzoper "Ainadamar" von (Name entfernt)	kostümierte Akteure auf Bühne, Tänzer, Schauspieler, Sänger / (Name entfernt) /	Schauspieler Oper Theateraufführung Tanztheater Sänger Theater
19	Wiesbaden. Eröffnung Maifestspiele mit Oper "Salome" am Staatstheater - Ausblick auf das Programm	Menschenmenge Theaterbesucher in Foyer / Ausschnitte Bühnenszenen, Opernsänger agieren / Regisseur (Name entfernt) auf Bühne während Probe auf Bühne /	Programm Oper Probe Regisseur Wiesbaden (Stadt) Musiktheater Kulturfestival
20	(Name entfernt) aus Hüttenberg hatte mehrere Herzinfarkte und seine Lungenfunktion ist eingeschränkt. Ein falsches Gutachten erkannte seine Behinderung nicht zu 100 Prozent an und sein Antrag auf einen Behindertenparkplatz wurde somit abgelehnt.	-	Behinderung Gutachten Herzinfarkt Parkplatz Antrag
21	Fulda. Der Fuldaer Autoentwickler (Name entfernt) hat ein Hightech-Fahrzeug mit Basalt-Karosserie entwickelt: "Light Car - Open Source".	-	Erfindung Fulda (Ort) Karosserie Elektroauto
22	Niederaula Niederjossa, Finalist im Wettbewerb Dolles Dorf 2009. Reporter (Name entfernt) live vor Ort.	-	Dorf Aktion Wettbewerb Reporter
23	Kassel. Viele in Deutschland verwendete Grabsteine werden von Kindern in Indien hergestellt. Am Internationalen Tag gegen Kinderarbeit macht das Museum für Sepulkralkultur mit der Tagung "Indische Grabsteine und Kinderarbeit" auf das Problem aufmerksam.	-	Tagung Indien (Staat) Sepulkralkultur Tag Grabstein Steinbruch Kinderarbeit
24	Kassel, Lichtenfels. Wenige Tage vor dem Weltklimagipfel in Kopenhagen, findet das Hessisches Klimaschutzforum in Kassel statt. Vertreter der Kommune Lichtenfels stellen dort ihr Konzept für die örtliche nachhaltige Energeigewinnung vor. Auch Großabnehmer wie die (Name entfernt)nutzen schon Strom aus erneuerbaren Energien wie Wasserkraft.	-	Handelsvertreter Energieversorgung Kopenhagen (Hauptstadt) Wasserenergie Lichtenfels (Ort) Klimaschutz Konferenz Veranstaltung
25	Frankfurt. Eine Aschewolke behindert nach einem Vulkanausbruch in Island (Vulkan Eyjafjallajökull) den Luftverkehr in weiten Teilen von Nordeuropa - Flugausfälle auch am Flughafen Frankfurt.	-	Luftfahrt Aschewolke Nordeuropa Island Flughafen Ausbruch Vulkan

A.2.2 Evaluation Results

The following list shows in aggregated form the precision and recall values assigned to the keyword extraction for every broadcast. The precision values have the following meanings:

- 5: All the keywords are suitable for this broadcast
- 4: Most keywords are suitable for this broadcast
- 3: Some keywords are suitable for this broadcast
- 2: Only a few keywords are suitable for this broadcast
- 1: None of the keywords are suitable for this broadcast

The recall values have the following meanings:

- 5: The broadcast's content is completely covered by the keywords
- 4: The broadcast's content is almost completely covered by the keywords
- 3: Some keywords are missing to completely cover the broadcast's content
- 2: Many keywords are missing to completely cover the broadcast's content
- 1: The broadcast's content is not covered at all by the keywords

Nr	Precision					Recall				
	5	4	3	2	1	5	4	3	2	1
1		1	2			2	1			
2		2	1			1		2		
3		1	1	1			2	1		
4		2	1			1	2			
5			2	1			1	2		
6		1	2			1	1	1		
7		3				1	2			
8	1	2				1	2			
9			1	2			1	1	1	
10	1	2				2	1			
11		1	2				2	1		
12	1	1	1			1	2			
13		1	2				1	2		
14			1	2			1	1	1	
15		2	1				2	1		
16		1	2				3			
17		3				1		2		
18	1		2				2	1		
19	1		2				1	2		
20		1	2					3		
21		3					3			
22		2	1				2	1		
23		1	1	1			2	1		
24			3				1	2		

25	1	2				1	2			
Total	**6**	**32**	**30**	**7**	**0**	**12**	**37**	**24**	**2**	**0**
%	**8**	**42.66**	**4**	**9.33**	**0**	**16**	**49.33**	**32**	**26.67**	

A.3 Thesaurus Alignment Evaluation Results

A.3.1 General Thesaurus Alignment Evaluation

The table displayed here contains the information that was given to the experts that performed the evaluation. It contains context information for both thesaurus and Linked Data concepts, although the DBpedia abstracts were somewhat shortened. In the rightmost three columns, the frequencies of the selected accuracy values are aggregated for each individual alignment pair. These values have the following meaning:

- 3: The alignment represents an exact match
- 2: The alignment represents an approximate match
- 1: The alignment represents an incorrect match

At the end of the table, total value frequencies are specified.

Thesaurus concept			DBpedia concept		Evaluation		
Name / Synonyms	**Broader terms**	**Concept type**	**DBpedia resource**	**Abstract**	**3**	**2**	**1**
Präservativ Kondom	Empfängnis-verhütung	Descriptor	http://dbpedia.org/resource/condom (Kondom)	A condom is a barrier device most commonly used during sexual intercourse to reduce the probability of pregnancy and spreading sexually transmitted diseases.	3		
Gastgewerbe Gastwirt-schaft Speisegast-stätte	Freizeitwirt-schaft	Class	http://dbpedia.org/resource/hospitality_industry (Gastgewerbe)	Unter dem Gastgewerbe versteht man die gewerbsmäßige Verpflegung und/oder Beherbergung in einem Beherbergungs- und Gaststättenbetrieb.	3		
Klimatologie	Klima	Descriptor	http://dbpedia.org/resource/climate (Klima)	Das Klima steht als Begriff für die Gesamtheit aller meteorologischen Vorgänge, die für den durchschnittlichen Zustand der Erdatmosphäre an einem Ort verantwortlich sind.		2	1
			http://dbpedia.org/resource/climatology (Klimatologie)	Die Klimatologie ist eine interdisziplinäre Wissenschaft der Fachgebiete Meteorologie, Geographie, Geologie, Ozeanographie und Physik.	3		
Schwager	Verwandt-schaft	Descriptor	http://dbpedia.org/resource/dietmar_schwager (Dietmar Schwager)	Dietmar Schwager ist ein ehemaliger deutscher Fußballspieler und -trainer. Als Spieler agierte er in der Abwehr.			3
			http://dbpedia.org/resource/raymund_schwager (Raymund Schwager)	Raymund Schwager war ein katholischer Theologe und Jesuit.			3
			http://dbpedia.org/resource/sebastian	Sebastian Schwager ist ein ehemaliger deutscher Radrennfahrer.			3

			_schwager (Sebastian Schwager)				
Flugreise Flugtouris-mus	Reisen Tourismus-art Inlandsreise	Descriptor	http://dbpedia.org/resource/air_travel (Flugreise)	Eine Flugreise unternimmt, wer den Flug eines Luftfahrzeugs nutzt, um zu einem meist entfernt liegenden Ort zu gelangen.	3		
Derby	Pferdesport Reitsport Geländeritt	Descriptor	http://dbpedia.org/resource/derby (Derby (Ort))	Derby ist eine Stadt am Fluss Derwent in der Grafschaft Derbyshire in den East Midlands von England.			3
			http://dbpedia.org/resource/apache_derby (Apache Derby)	Apache Derby ist ein Projekt der Apache Software Foundation, das ein gleichnamiges Java-basiertes, relationales Datenbank-Management-System entwickelt.		1	2
			http://dbpedia.org/resource/earl_of_derby (Earl of Derby)	Earl of Derby ist ein erblicher britischer Adelstitel in der Peerage of England, benannt nach der Stadt Derby in England, der nunmehr von der Familie Stanley getragen wird.		1	2
Vogelschutz-warte	Vogelschutz	Descriptor	http://dbpedia.org/resource/bird_conservation (Vogelschutz)	Vogelschutz ist grundsätzlich eine Sammelbezeichnung für sämtliche Maßnahmen, die zur Erhaltung, Förderung oder Ansiedlung von Vögeln geeignet sind.		3	
Natrium	Chemisches Element	Descriptor	http://dbpedia.org/resource/sodium (Natrium)	Natrium ist ein häufig vorkommendes chemisches Element mit dem Symbol Na und der Ordnungszahl 11.	3		
Glasbau Glasarchi-tektur	Bauweise	Descriptor	http://dbpedia.org/resource/architectural_glass (Glasarchitektur)	Als Glasarchitektur bezeichnet man eine über das übliche Maß hinausgehende großflächige Verwendung des Baustoffes Glas als architektonisches Gestaltungselement an Fassaden.	2	1	
Datenbank Database Datensamm-lung	-	Class	http://dbpedia.org/resource/database (Datenbank)	Ein Datenbanksystem (DBS) ist ein System zur elektronischen Datenverwaltung.	3		
Landesverrat Staatsgefäh-rdung	.	Descriptor	http://dbpedia.org/resource/treason (Landesverrat)	Der Landesverrat ist in der Regel als Verbrechen gegen den Staat definiert.	3		
Disziplinar-vergehen Dienstvergeh en	Öffentliches Dienstrecht	Descriptor	http://dbpedia.org/resource/malfeasance_in_office (Dienstvergehen)	Ein Dienstvergehen ist eine schuldhafte Verletzung der Dienstpflichten oder des außerdienstlichen Verhaltens durch Beamte, Soldaten, Zivildienstleistende, Richter oder Notares.	2	1	
Verbraucher-konkurs Restschuld-befreiung	Verschuld-ung Privatver-schuldung Schulden	Descriptor	http://dbpedia.org/resource/debt_relief_order (Restschuldbe-freiung)	A DRO is a simplified, quicker and cheaper alternative to bankruptcy in the United Kingdom.	3		
Arbeitsbe-schaffung Beschäftig-ungspro-gramm	Arbeits-marktpolitik	Klasse	http://dbpedia.org/resource/public_works (Arbeitsbe-schaffung)	Unter Arbeitsbeschaffung versteht man staatliche Investitionen, um Beschäftigungslose wieder in Beschäftigung zu bringen.	3		
Aquada	-	Free descriptor	http://dbpedia.org/resource/gibbs_aquada (Gibbs Aquada)	Der Gibbs Aquada ist ein Amphibienfahrzeug auf Basis des Mazda MX-5.	2	1	

			http://dbpedia.org/resource/aquadag (Aquadag)	Aquadag ist eine wässrige Suspension von kolloidalem Graphit in gelartigen Flüssigkeiten.	1	1	1
Bowls	-	Free descriptor	http://dbpedia.org/resource/bowls (Bowls)	Bowls ist ein britisches Kugelspiel. Das Wort bowls ist der englische Ausdruck für „Kugeln“ (vgl. italienisch Boccia, französisch Boule).	3		
			http://dbpedia.org/resource/lawn_bowls_at_the_2006_commonwealth_games (Commonwealth Games 2006/Bowls)	Bei den Commonwealth Games 2006 in Melbourne fanden sechs Wettbewerbe im Bowls statt.		2	
			http://dbpedia.org/resource/lawn_bowls_at_the_2010_commonwealth_games (Commonwealth Games 2010/Bowls)	Bei den Commonwealth Games 2010 in Neu-Delhi fanden im Bowls sechs Wettbewerbe statt, je drei für Männer und Frauen.		1	2
Fideikommiss	-	Free descriptor	http://dbpedia.org/resource/fee_tail (Familienfidei-kommiss)	Das Familienfideikommiss ist eine Einrichtung des Erb- und Sachenrechts, wonach durch Stiftung das Vermögen einer Familie, auf ewig geschlossen erhalten werden sollte und immer nur ein Familienmitglied allein das Nießbrauchsrecht innehat.	3		
Tesla	.	Free descriptor	http://dbpedia.org/resource/tesla_turbine (Tesla-Turbine)	Eine Tesla-Turbine (auch Scheibenläuferturbine) ist eine schaufellose Turbine, die von Nikola Tesla erfunden wurde.		3	
			http://dbpedia.org/resource/nikola_tesla (Nikola Tesla)	Nikola Tesla war ein Erfinder, Physiker und Elektroingenieur.		3	
			http://dbpedia.org/resource/tesla_coil (Tesla-Transformer)	Ein Tesla-Transformator, auch als Teslaspule bezeichnet, ist ein nach seinem Erfinder Nikola Tesla benannter Resonanztransformator zur Erzeugung hochfrequenter Wechselspannung.		3	
			http://dbpedia.org/resource/tesla_particle_accelerator (International Linear Collider)	Der International Linear Collider (ILC) ist ein zurzeit in der Planungsphase befindlicher Linearbeschleuniger für Elementarteilchen.		1	2
			http://dbpedia.org/resource/tesla_arena (Tesla Arena)	Die Tesla Arena ist eine Mehrzwecksporthalle im Prager Stadtteil Holešovice. Sie wird hauptsächlich für Eishockey, aber auch für Konzerte benutzt.		3	
			http://dbpedia.org/resource/tesla_motors (Tesla Motors)	Tesla Motors, Inc. ist ein US-amerikanisches Unternehmen, das Elektroautos, sowie Komponenten für deren Antriebsstrang entwirft und baut.		3	
			http://dbpedia.org/resource/tesla_roadster (Tesla Roadster)	Der Tesla Roadster ist ein vollständig elektrisch betriebener, zweisitziger Sportwagen. Er ist das erste Serienfahrzeug von Tesla Motors.		3	
			http://dbpedia.org/	(2244) Tesla ist ein Asteroid im		3	

			resource/2244_tesla ((2244) Tesla)	Hauptgürtel zwischen Mars und Jupiter.			
			http://dbpedia.org/resource/tesla_model_s (Tesla Model S)	Das Tesla Model S ist eine Limousine mit elektrischem Antrieb von Tesla Motors, die als Prototyp am 26. März 2009 vorgestellt wurde.		3	
			http://dbpedia.org/resource/nvidia_tesla (Nvidia Tesla)	Tesla ist ein Prozessor mit stark parallelisiertem Design, auch Streamprozessor genannt, der Firma Nvidia.		2	1
			http://dbpedia.org/resource/tpp_nikola_tesla (Wärmekraftwerke Nikola Tesla)	Die Wärmekraftwerke Nikola Tesla sind ein Kohlekraftwerks-Komplex in Serbien, benannt zu Ehren des Erfinders Nikola Tesla.		2	1
Funktionalis-mus	-	Free descriptor	http://dbpedia.org/resource/structural_functionalism (Strukturfunktio-nalismus)	Der Strukturfunktionalismus, eine theoretische Richtung der Soziologie, betrachtet soziale Systeme als ihre eigene Existenz erhaltende Gebilde.		2	1
			http://dbpedia.org/resource/neofunctionalism (Neofunktionalis-mus)	Unter Neofunktionalismus versteht man eine politikwissenschaftliche Theorie im Teilgebiet der internationalen Beziehungen.		3	
			http://dbpedia.org/resource/functionalism_in_international_relations (Funktionalismus (Internationale Beziehungen))	Der Funktionalismus ist eine Theorie innerhalb der Internationalen Beziehungen, nach der das Zusammenwachsen von Staaten über die Delegation staatlicher Souveränität zugunsten über- oder zwischenstaatlicher Institutionen erfolgt.	3		
Eurostar	-	Free descriptor	http://dbpedia.org/resource/eurostar (Eurostar)	Die Eurostar Group Ltd. ist ein europäisches Verkehrsunternehmen mit Sitz in London.	1	2	
			http://dbpedia.org/resource/eurostar_italia (Eurostar Italia)	Eurostar Italia ist eine Zuggattung der italienischen Eisenbahngesellschaft Trenitalia für Hochgeschwindigkeitszüge.		3	
Nesselsucht	-	Free descriptor	http://dbpedia.org/resource/urticaria (Nesselsucht)	Die Nesselsucht (Urtikaria) ist eine krankhafte Reaktion der Haut auf Nahrungsmittel oder Medikamente (Allergien), auf Wärme-, Kälte-, Licht-, Druck- oder Wassereinwirkung, auf Immunologische Phänomene oder auch auf psychischen Stress..	3		
			http://dbpedia.org/resource/aquagenic_urticaria (Wassernessel-sucht)	Wassernesselsucht, Wasserurtikaria oder aquagene Urtikaria ist eine sehr seltene Form der Nesselsucht, die nach Kontakt mit Wasser auftritt.		3	
Probiotikum	-	Free descriptor	http://dbpedia.org/resource/probiotic (Probiotikum)	Ein Probiotikum ist eine Zubereitung, die lebensfähige Mikroorganismen enthält.	3		
Bundesstadt	-	Free descriptor	http://dbpedia.org/resource/federal_city (Bundesstadt)	In der Schweiz wird Bern Bundesstadt genannt, weil es offiziell keine Schweizer Hauptstadt gibt. In Deutschland ist Bundesstadt der Titel einer Stadt, die eine besondere Rolle im Bundesstaat spielt.	3		
Dulcinea	-	Free	http://dbpedia.org/	(571) Dulcinea ist ein Asteroid des	2	1	

		descriptor	resource/571_dulcinea ((571 Dulcinea)	Hauptgürtels, der am 4. September 1905 vom deutschen Astronomen Paul Götz in Heidelberg entdeckt wurde.			
			http://dbpedia.org/resource/dulcinea (Dulcinea del Toboso)	Dulcinea del Toboso ist eine fiktive literarische Figur, die im Roman Don Quijote von Miguel de Cervantes zur imaginären Geliebten Don Quijotes wird.		2	1
Bauernkrieg	-	Descriptor	http://dbpedia.org/resource/list_of_revolutions_and_rebellions (Liste von Bauernaufständen)	Bauernkrieg ist ein Sammelbegriff für eine Reihe von Bürgerkriegen und Aufständen vom ausgehenden 15. bis ins 17. Jahrhundert.	3		
Verschwie-genheit	-	Free descriptor	http://dbpedia.org/resource/confidentiality (Verschwiegen-heitspflicht)	Die Verschwiegenheitspflicht ist die rechtliche Verpflichtung bestimmter Berufsgruppen, ihnen anvertraute Geheimnisse nicht an Dritte weiterzugeben.		3	
			http://dbpedia.org/resource/discretion (Ermessen)	Ermessen ist ein juristischer Fachbegriff. Er räumt einem Entscheidungsträger gewisse Freiheiten bei seiner Entscheidungsfindung ein.		1	2
Saft	-	Free descriptor	http://dbpedia.org/resource/juice (Saft)	Mit Saft wird allgemein eine Flüssigkeit in oder von organischen Körpern bezeichnet.	1	2	
Han-Chinesen	-	Free descriptor	http://dbpedia.org/resource/han_chinese (Han-Chinesen)	Da der Begriff Chinesen im Deutschen nicht eindeutig unterscheidet zwischen Staatsangehörigen Chinas und Angehörigen der Han-Nationalität, also „ethnischen Chinesen“, hat sich die Bezeichnung „Han-Chinesen“ oder kurz Han zunehmend eingebürgert.	3		
Zippo	-	Free descriptor	http://dbpedia.org/resource/zippo (Zippo)	Ein Zippo ist ein benzinbetriebenes Sturmfeuerzeug, das in annähernd unveränderter Bauform seit 1933 von der 1932 gegründeten Zippo Manufacturing Co. in Bradford, Pennsylvania (USA) hergestellt wird.	2	1	
			http://dbpedia.org/resource/zipporah (Zippora)	Zippora ist eine biblische Gestalt. Sie ist die Tochter des Priesters Jitro (oder Jethro) von Midian. Sie wurde die Ehefrau des Mose.			3
			http://dbpedia.org/resource/zippo_200_at_the_glen (Zippo 200)	Das Zippo 200 ist ein Autorennen der NASCAR Nationwide Series, das seit der Saison 1991 auf der Rennstrecke Watkins Glen International ausgetragen wird.		1	2
Mathematik-Olympiade	-	Free descriptor	http://dbpedia.org/resource/international_mathematical_olympiad (Internationale Mathematik-Olympiade)	Die Internationale Mathematik-Olympiade (IMO) ist ein internationaler Schülerwettbewerb im Bereich Mathematik, der seit 1959 (mit einer Ausnahme) jährlich stattfindet.	3		
Total					**61**	**73**	**33**
%					**37**	**44**	**20**

A.3.2 Geography Thesaurus Alignment Evaluation

The table displayed here contains the information that was given to the experts that performed the evaluation. It contains context information for both thesaurus and Linked Data concepts, although the DBpedia abstracts were somewhat shortened. In the rightmost three columns, the frequencies of the selected accuracy values are aggregated for each individual alignment pair. These values have the following meaning:

- 3: The alignment represents an exact match
- 2: The alignment represents an approximate match
- 1: The alignment represents an incorrect match

At the end of the table, total value frequencies are specified.

Thesaurus concept			DBpedia concept		Evaluation		
Name / Synonyms	**Broader terms**	**Concept type**	**DBpedia resource**	**Abstract**	**3**	**2**	**1**
Seeve	-	Place	http://dbpedia.org/resource/seeve (Seeve)	The River Seeve is a tributary of the Elbe in northern Germany. Its source is near Wehlen in the northern part of the Lüneburg Heath.	2	1	
			http://dbpedia.org/resource/seevetal (Seevetal)	Seevetal is a municipality in the district of Harburg, in Lower Saxony, Germany. It is named after the river Seeve.		3	
Mayen-Koblenz	Koblenz	Place	http://dbpedia.org/resource/koblenz (Koblenz)	Koblenz ist eine kreisfreie Stadt im nördlichen Rheinland-Pfalz. Sie ist nach Mainz und Ludwigshafen am Rhein die drittgrößte Stadt dieses Landes und bildet eines seiner fünf Oberzentren.		3	
			http://dbpedia.org/resource/mayen-koblenz (Mayen-Koblenz)	Mayen-Koblenz is a district (Kreis) in the north of Rhineland-Palatinate, Germany.	3		
Schmalensee	Kreis Segeberg	Place	http://dbpedia.org/resource/schmalensee (Schmalensee)	Schmalensee is a municipality in the district of Segeberg, in Schleswig-Holstein, Germany.	3		
Salouf	-	Place	http://dbpedia.org/resource/salouf (Salouf)	Salouf (until 1943 officially called Salux) is a municipality in the district of Albula in the canton of Graubünden in Switzerland. The population is predominately Romansh-speaking.	3		
Taipeh	Taiwan	Capital	http://dbpedia.org/resource/taipei (Taipeh)	Taipeh, häufig in der englischen Schreibweise als Taipei vorzufinden, ist eine Millionenstadt am Fluss Danshui im nördlichen Teil von Taiwan und die Hauptstadt der Republik China.	3		
Bockenheim	-	Place	http://dbpedia.org/resource/stein-	Stein-Bockenheim is an Ortsgemeinde in the Alzey-Worms		2	1

			bockenheim (Stein-Bockenheim)	district in Rhineland-Palatinate, Germany.			
Gaarz	-	Place	http://dbpedia.org/resource/neu_gaarz (Neu Gaarz)	Neu Gaarz is a municipality in the Mecklenburgische Seenplatte district, in Mecklenburg-Vorpommern, Germany.	1	1	1
Heilbronn	-	City	http://dbpedia.org/resource/heilbronn (Heilbronn)	Heilbronn is a city in northern Baden-Württemberg, Germany. The city on the Neckar is a former Imperial Free City and current independent city (i.e. not part of any county) and seat of Heilbronn County.	3		
Hautes-Alpes	-	Place	http://dbpedia.org/resource/hautes-alpes (Hautes-Alpes)	Hautes-Alpes is a department in southeastern France named after the Alps mountain range.	3		
Badenheim	-	Place	http://dbpedia.org/resource/badenheim (Badenheim)	Badenheim is an in the Mainz-Bingen district in Rhineland-Palatinate, Germany. It belongs to the Verbandsgemeinde of Sprendlingen-Gensingen.	2		
Wartburg	-	Place	http://dbpedia.org/resource/wartburg (Wartburg)	Die Wartburg ist eine Burg in Thüringen, über der Stadt Eisenach am nordwestlichen Ende des Thüringer Waldes gelegen.	3		
			http://dbpedia.org/resource/urs_von_wartburg (Urs von Wartburg)	Urs von Wartburg (* 1. März 1937) ist ein Schweizer Leichtathlet, der rund 30 Jahre lang als Speerwerfer aktiv war.		1	2
			http://dbpedia.org/resource/walther_von_wartburg (Walther von Wartburg)	Walther von Wartburg (-Boos) war ein Schweizer Romanist und Sprachwissenschaftler.		1	2
			http://dbpedia.org/resource/wartburg,_tennessee (Wartburg, Tennessee)	Wartburg is a city in Morgan County, Tennessee, United States. The population was 890 at the 2000 census. It is the county seat of Morgan County.	3		
			http://dbpedia.org/resource/wartburg_festival (Wartburgfest)	Wartburgfest ist der Name mehrerer, zumeist studentischer Versammlungen, die jeweils auf der Wartburg bei Eisenach in Thüringen stattfanden.		2	1
			http://dbpedia.org/resource/wartburg,_kwazulu-natal (Wartburg (Südafrika))	Wartburg ist ein kleiner Ort in den Midlands der südafrikanischen Provinz KwaZulu-Natal.	3		
Veldenz	-	Place	http://dbpedia.org/resource/county_of_veldenz (Grafschaft Veldenz)	The County of Veldenz was a principality in the contemporary Land Rhineland-Palatinate. A town of the same name Veldenz and a castle Schloss Veldenz are located in the district of Bernkastel-Wittlich.	2	1	
			http://dbpedia.org/resource/veldenz (Veldenz)	Veldenz ist eine rheinland-pfälzische Ortsgemeinde im Landkreis Bernkastel-Wittlich und früherer Hauptsitz der gleichnamigen Grafschaft.	3		
			http://dbpedia.org/resource/schloss_veldenz	Schloss Veldenz in the Mosel Valley region is a Castle ruin about 1.5 kilometers southeast of Veldenz.		3	

			(Schloss Veldenz)				
			http://dbpedia.org/resource/veldenz_lion (Veldenzer Löwe)	The Veldenz lion is a modified heraldic emblem. As a charge the heraldic lion in the coat of arms has a stationary appearanceAs a representation of the County of Veldenz this lion can be found in many arms in the region		2	1
Mutare	-	Place	http://dbpedia.org/resource/mutare (Mutare)	Mutare (known as Umtali until 1982) is the fourth largest city in Zimbabwe, with a population of around 170,000. It is the capital of Manicaland province.	3		
			http://dbpedia.org/resource/roman_catholic_diocese_of_mutare (Bistum Mutare)	The Roman Catholic Archdiocese of Mutare is a suffragan diocese in the city of Mutare in the ecclesiastical province of Harare in Zimbabwe.		3	
Drelsdorf	Kreis Nordfries-land	Place	http://dbpedia.org/resource/drelsdorf (Drelsdorf)	Drelsdorf is a municipality in the district of Nordfriesland, in Schleswig-Holstein, Germany.	3		
Agra	Uttar Pradesh	Place	http://dbpedia.org/resource/agra	Agra, the erstwhile capital of Hindustan, is a city on the banks of the river Yamuna in the northern state of Uttar Pradesh, India.	3		
Flintbek	Kreis Rends-burg-Eckernförde	Place	http://dbpedia.org/resource/flintbek (Flintbek)	Flintbek is a municipality in the district of Rendsburg-Eckernförde, in Schleswig-Holstein, Germany. It is situated at the Eider River, approx. 10 km southwest of Kiel.	3		
Marbella	-	Place	http://dbpedia.org/resource/marbella (Marbella)	Marbella is a city and municipality in southern Spain, belonging to the province of Málaga in the autonomous community of Andalusia.	3		
			http://dbpedia.org/resource/seat_marbella (Seat Marbella)	The SEAT Panda was a badge-engineered Fiat Panda produced by SEAT from 1980 to 1986.		1	2
			http://dbpedia.org/resource/andalucia_tennis_experience (Andalucia Tennis Experience)	The Andalucia Tennis Experience is a professional women's tennis tournament played on outdoor clay courts.		1	2
			http://dbpedia.org/resource/ud_marbella (UD Marbella)	Unión Deportiva Marbella is a Spanish football team based in Marbella, in the autonomous community of Andalusia.		1	2
Quiberon	-	Place	http://dbpedia.org/resource/quiberon (Quiberon)	Quiberon is a commune in the Morbihan department in Brittany in north-western France. It is situated on the southern part of the Quiberon peninsula.	3		
			http://dbpedia.org/resource/saint-pierre-quiberon (Saint-Pierre-Quiberon)	Saint-Pierre-Quiberon is a commune in the Morbihan department of Brittany in north-western France. The commune occupies the northern part of the Quiberon peninsula, that it shares with the commune Quiberon.	2	1	
			http://dbpedia.org/resource/battle_of_quiberon_bay	The naval Battle of Quiberon Bay took place on 20 November 1759 during the Seven Years' War in	1	1	1

			(Seeschlacht in der Bucht von Quiberon)	Quiberon Bay, off the coast of France near St. Nazaire.			
Sierentz	-	Place	http://dbpedia.org/resource/sierentz (Sierentz)	Sierentz is a commune in the Haut-Rhin department in Alsace in north-eastern France. It is located roughly halfway between Mulhouse and Basel.	3		
Bastogne	-	Place	http://dbpedia.org/resource/bastogne (Bastogne)	Bastogne is a Walloon municipality of Belgium located in the province of Luxembourg in the Ardennes.	3		
			http://dbpedia.org/resource/arrondissement_of_bastogne (Bastogne (Arrondissement))	The Arrondissement of Bastogne is one of the five administrative arrondissements in the Province of Luxembourg, Belgium.	3		
Total					**67**	**28**	**15**
%					**61**	**25**	**14**

A.4 User Interview Transcripts

The following are the complete transcripts of all seven user interviews in German. They were transcribed from audio recordings and observations of the users' search approach made during the interviews.

Nutzerbefragung Nr. 1

Angaben zur Person

Geschlecht: weiblich

Alter: 35-45

Technische Kenntnisse: eher beschränkt

Abteilung: Fernsehunterhaltung

Tätigkeit: Redaktionssekretärin

Erfahrung mit FESAD-Benutzerschnittstelle

Welche FESAD-Schnittstelle verwenden Sie?

FESADred.

Wie häufig verwenden Sie diese Schnittstelle?

Nicht sehr häufig, schätzungsweise 1-2 Mal pro Tag.

Wozu verwenden Sie diese Schnittstelle?

Für die Suche nach Filmen und Sendungen, um das Programm zu gestalten.

Typische Abfrage:

Suche nach Film- oder Sendungstitel und entsprechenden Ausstrahlungsdaten.

Einschätzung der FESAD-Schnittstelle

Wie schätzen Sie die Genauigkeit der Suchergebnisse ein:

Ungefähr 70-80% der Treffer sind durchschnittlich passende Ergebnisse, wobei das auch davon abhängt, ob man den Titel des gesuchten Inhaltes kennt. Wenn man nur nach einem Stichwort sucht sind die Ergebnisse meist schlechter. In diesem Fall muss man genau wissen, wie die Inhalte von den Archivaren mit Stichworten versehen werden. Dazu ist eine gewisse Kommunikation notwendig, zum Beispiel wenn Filme unbenannt werden, welche aber leider oft nicht gegeben ist. Ohne gewisse Vorkenntnisse kann eine Suche daher schwieriger sein.

Bedienung:

Die Bedienung ist eigentlich recht gut, es gibt genügend Suchkriterien, um eine Suche gut differenzieren zu können.

Verbesserungsmöglichkeiten:

Die Suche nach inhaltlich bezogenen Daten könnte verbessert werden, wobei oftmals Inhalte auch nicht mit Schlagworten versehen sind. Oft probiert man es dann mit unterschiedlichen Suchmethoden.

Test-Aufgabe

Finden Sie Beiträge zu Jaguaren (Tier):

Die Eingabe „Jaguar“ im Suchfeld ergibt 79 Treffer, davon scheinbar einige zur Automarke oder zu einer „Jaguar“-Arbeitsgruppe der Polizei. Die Suche wird dann durch die zusätzliche Eingabe des Begriffs „Tier“ weiter eingeschränkt, wodurch nur noch drei Treffer übrig bleiben. An dieser Stelle würde sich die Versuchsperson die Ergebnisse einzeln ansehen, eventuell auch noch die vorherige Ergebnismenge.

Wie schätzen Sie die Qualität dieser Suchergebnisse ein?

Eher dürftig, da in der ersten Ergebnismenge viele falsche Treffer waren und in der zweiten fast gar keine Treffer.

Wie empfinden Sie den mit dieser Suche verbundene Aufwand?

Man muss sich damit abfinden, dass man manchmal grosse Mengen von Ergebnissen durchforsten muss, das ist aber nicht weiter schlimm.

Ist die Suche nach mehrdeutigen Begriffen ein häufiger Anwendungsfall?

Nein, solche Suchen kommen in der Regel nicht wirklich oft vor. Die Versuchsperson würden in einem solchen Fall Kollegen um Hilfe bitten.

Finden Sie Beiträge zu hessischen Städten mit Mehrheitspartei CDU:

Es werden bei der Eingabe „berichte hessische Städte CDU“ keine Ergebnisse gefunden, dafür aber 142 durch die Suche „berichte CDU“. Diese Ergebnisse würde die Versuchsperson dann einzeln durchgehen.

Wie schätzen Sie die Qualität dieser Suchergebnisse ein?

Es ist schwierig, in FESAD eine solche Suche durchzuführen, da solche Informationen nicht mitgespeichert sind. Daher muss man damit leben, dass die Ergebnismenge in solchen Fällen eher ungenau sein kann.

Wie empfinden Sie den mit dieser Suche verbundene Aufwand?

Der Aufwand ist nicht schlimm, die Versuchsperson findet es in Ordnung, unterschiedliche Suchen auszuprobieren und Ergebnisse einzeln durchzugehen.

Sind mehrstufige Suchen ein häufiger Anwendungsfall?

Nein, die Suche beschränkt sich meist auf die Titel von Beiträgen oder Filmen.

Faceted Search

In Bezug auf Ihre Arbeit, wie nützlich schätzen Sie die facettierte Suche nach mehrdeutigen Begriffen ein?

Facettierte Suche erscheint durchaus nützlich zu sein, würde aber im Arbeitsalltag der Versuchsperson nicht wirklich eingesetzt werden müssen.

In Bezug auf Ihre Arbeit, wie nützlich schätzen Sie die facettierte Suche bei mehrstufigen Suchabfragen ein?

Siehe obige Antwort.

Wie schätzen Sie die Verständlichkeit der facettierten Suche ein?

Es ist sehr verständlich, die Informationen sind ja alle angezeigt. Der Aufbau ist logisch.

Wie schätzen Sie die Bedienungsfreundlichkeit der facettierten Suche ein?

Es scheint sehr einfach erlernbar zu sein.

Würden Sie die Möglichkeiten der facettierten Suche nutzen, und wie?

-

Haben Sie weitere Bemerkungen zur facettierten Suche?

-

Nutzerbefragung Nr. 2

Angaben zur Person

Geschlecht: weiblich

Alter: 35-45

Technische Kenntnisse: mittelmässig

Abteilung: Programm-Management

Tätigkeit: Planung des Programms nach 20h15

Erfahrung mit FESAD-Benutzerschnittstelle

Welche FESAD-Schnittstelle verwenden Sie?

FESAD-Webclient.

Wie häufig verwenden Sie diese Schnittstelle?

Täglich mehrmals, wobei Suchen auch des Öfteren an Assistenten delegiert werden.

Wozu verwenden Sie diese Schnittstelle?

Zur Recherche von Programmen bei der Übernahme von anderen Sendern. Dabei ist es auch nützlich, dass man Sendungen thematisch sortieren kann.

Typische Abfrage:

Den Inhalt einer mit Namen bekannten Sendung abfragen, sowie andere Informationen wie Erstsendedatum, Produzent, etc. Häufig kommen aber auch Suchen zu bestimmten Themen vor, z.B. „Olympische Spiele“. Auch werden manchmal Sendungen gesucht, die in einer bestimmten Zeitspanne ausgestrahlt wurden.

Einschätzung der FESAD-Schnittstelle

Wie schätzen Sie die Genauigkeit der Suchergebnisse ein:

Ergebnisse sind recht genau, sofern die Metadaten auch entsprechend eingelesen wurden und der Redakteur weiss, wonach er suchen muss. Zum Beispiel sind Inhalte wie Musiksendungen manchmal nicht mit dem Schlagwort „Musik“ versehen.

Bedienung:

Die Bedienung ist eigentlich sehr gut, einziges Problem ist, dass man über den Zurück-Button des Browsers nicht auf vorher angezeigte Ergebnisse zurückgehen kann.

Verbesserungsmöglichkeiten:

Suchfunktion „Sendedauer“ funktioniert nicht, wäre aber sehr nützlich um Ergebnismenge einzugrenzen.

Test-Aufgabe

Finden Sie Beiträge zu Jaguaren (Tier):

Die Suche nach „Jaguar“ ergibt im FESAD-Webclient sehr viele Ergebnisse. Trotzdem würde die Versuchsperson bereits diese Treffer manuell durchgehen, da sie nicht wüsste, wie man sie mit den gegebenen Mitteln besser filtern können.

Wie schätzen Sie die Qualität dieser Suchergebnisse ein?

Es sind sehr viele falsche Treffer dabei, was die Suche eher schwierig macht, da man diese von Hand aussortieren muss.

Wie empfinden Sie den mit dieser Suche verbundene Aufwand?

Diese Suche wäre mit einem sehr hohen Aufwand verbunden.

Ist die Suche nach mehrdeutigen Begriffen ein häufiger Anwendungsfall?

Ja, es kommt häufig vor, dass man sich genau wegen solcher Mehrdeutigkeiten durch zahlreiche Ergebnisse durchklicken muss und dabei viel Arbeitszeit investiert.

Finden Sie Beiträge zu hessischen Städten mit Mehrheitspartei CDU:

Suche über das „Schlagwort“-Suchfeld nach „CDU stadt hessen“. Da dies keine Ergebnisse liefert, würde die Versuchsperson die notwendigen Informationen zunächst über Google einholen und dann mit diesen einzelne Suchen machen zu den gefunden Städten.

Wie schätzen Sie die Qualität dieser Suchergebnisse ein?

In diesem Fall als schlecht, da keine Ergebnisse gefunden wurden.

Wie empfinden Sie den mit dieser Suche verbundene Aufwand?

Durch die vorgängige Informationssuche wäre der Aufwand entsprechend grösser, v.a. wenn viele Begriffe auf meine Suchkriterien zutreffen.

Sind mehrstufige Suchen ein häufiger Anwendungsfall?

Nicht direkt auf die Suchbegriffe bezogen, aber z.B. bei der Suche nach Musiksendungen, welche im Winter gelaufen sind.

Faceted Search

In Bezug auf Ihre Arbeit, wie nützlich schätzen Sie die facettierte Suche nach mehrdeutigen Begriffen ein?

Der Nutzen wäre sehr hoch, da man dadurch viel Zeit sparen könnte, die man damit verbringen, falsche Treffer auszusortieren.

In Bezug auf Ihre Arbeit, wie nützlich schätzen Sie die facettierte Suche bei mehrstufigen Suchabfragen ein?

Nein, die Suchbegriffe sind meist bekannt und können direkt eingegeben werden. Suchen mit mehreren Suchbegriffen kommen aber häufig vor. Dann müssen die Ergebnisse von Hand sortiert werden, was die Versuchsperson manchmal mit ausgedruckten Listen davon macht.

Wie schätzen Sie die Verständlichkeit der facettierten Suche ein?

Die Anzeige ist sehr logisch und recht leicht verständlich.

Wie schätzen Sie die Bedienungsfreundlichkeit der facettierten Suche ein?

-

Würden Sie die Möglichkeiten der facettierten Suche nutzen, und wie?

Ja, absolut. Weitere Suchfunktionen wären aber auch wichtig, zum Beispiel hessenschau-Beiträge aus der Suche auszuschliessen. Wichtig wäre weiterhin, dass die Qualität der Metadaten gut ist.

Haben Sie weitere Bemerkungen zur facettierten Suche?

-

Nutzerbefragung Nr. 3

Angaben zur Person

Geschlecht: männlich

Alter: 35-45

Technische Kenntnisse: gut

Abteilung: Politik und Zeitgeschichte

Tätigkeit: Redakteur und FESADred-Koordination (zwischen Nutzern und Entwickler)

Erfahrung mit FESAD-Benutzerschnittstelle

Welche FESAD-Schnittstelle verwenden Sie?

FESADred

Wie häufig verwenden Sie diese Schnittstelle?

Mehrmals täglich

Wozu verwenden Sie diese Schnittstelle?

Beitragsrecherche, vorwiegend mit der Absicht, das System zu testen und Nutzerbedürfnisse nachzuvollziehen

Typische Abfrage:

Suche nach spezifischen Beiträgen

Einschätzung der FESAD-Schnittstelle

Wie schätzen Sie die Genauigkeit der Suchergebnisse ein:

In der Regel sind weit über 90% der Ergebnisse passende Treffer, wobei es manchmal auch etwas erstaunliche Ergebnisse gibt.

Bedienung:

Allgemein ist die Bedienung leicht verständlich, wobei manche Sachen nicht sehr logisch sind. Der Zurück-Button ist oben nicht immer vorhanden, es ist in gewissen Fällen nicht möglich, auf frühere Suchen zurückzugehen.

Verbesserungsmöglichkeiten:

Test-Aufgabe

Finden Sie Beiträge zu Jaguaren (Tier):

Suche im Rohmaterial-Bestand nach dem Begriff „Jaguar“ ergibt 79 Treffer, welche die Versuchsperson zunächst von Hand durchgehen würde um die passende Treffer rauszupicken. Da scheinbar wenig Nützliches dabei ist, würde sie dann noch versuchen, Treffer zur Automarke mit „-auto“ auszuschliessen. Dadurch bleiben nur noch 38 Treffer.

Wie schätzen Sie die Qualität dieser Suchergebnisse ein?

Die Ergebnisse sind völlig unbefriedigend, da sehr wenige der zurückgegebenen Treffer nützlich sind.

Wie empfinden Sie den mit dieser Suche verbundene Aufwand?

Der Suchaufwand ist sehr gross. Die Suche nach Beiträgen zu Jaguaren ist ein gutes Beispiel einer sehr mühsamen Suche.

Ist die Suche nach mehrdeutigen Begriffen ein häufiger Anwendungsfall?

Im persönlichen Arbeitsalltag der Versuchsperson nicht, aber für Redakteure, welche Beiträge kompilieren, ist es sehr häufig so, dass sie sich grosse Mengen von Ergebnissen ansehen müssen, um akzeptable Treffer zu finden, was sehr viel Zeit in Anspruch nimmt.

Finden Sie Beiträge zu hessischen Städten mit Mehrheitspartei CDU:

Die Suche nach einem solchen Kriterium einzuschränken, ist über FESAD nicht möglich, die Versuchsperson würde daher vorgängig im Internet nach den entsprechenden Informationen suchen, z.B. beim Städte- und Gemeindebund, bei der CDU, oder beim statistischen Landesamt, eventuell auch über Telefonanrufe. Erst danach würde sie die einzelnen Städte in FESAD recherchieren.

Wie schätzen Sie die Qualität dieser Suchergebnisse ein?

-

Wie empfinden Sie den mit dieser Suche verbundene Aufwand?

Sehr aufwändig, da solche Informationen nicht einmal über Google mit Leichtigkeit gefunden werden können.

Sind mehrstufige Suchen ein häufiger Anwendungsfall?

Ja, absolut.

Faceted Search

In Bezug auf Ihre Arbeit, wie nützlich schätzen Sie die facettierte Suche nach mehrdeutigen Begriffen ein?

Die Möglichkeit, Suchbegriffe nach Typ einzuschränken, wäre sehr nützlich.

In Bezug auf Ihre Arbeit, wie nützlich schätzen Sie die facettierte Suche bei mehrstufigen Suchabfragen ein?

-

Wie schätzen Sie die Verständlichkeit der facettierten Suche ein?

Zur Verständlichkeit gibt es nichts einzuwenden, die Suchmaske ist sehr übersichtlich.

Wie schätzen Sie die Bedienungsfreundlichkeit der facettierten Suche ein?

-

Würden Sie die Möglichkeiten der facettierten Suche nutzen, und wie?

Ja, auf jeden Fall ausprobieren. Die Versuchsperson würde am ehesten so vorgehen, dass sie die Informationen aus der facettierten Suche mit einer Web-Suche überprüfen würde. Sollte sich dann herausstellen, dass dieser Weg zu aufwändig ist, würde sie die facettierte Suche aber wieder sein lassen.

Haben Sie weitere Bemerkungen zur facettierten Suche?

Die Validität der Hintergrundinformationen (DBpedia) ist ein wichtiger Faktor für die Nützlichkeit von facettierter Suche. Der Suchende sollte erwarten dürfen, dass diese Informationen richtig und vollständig sind. Sollte die Datengrundlage nur rudimentär sein, wäre möglicherweise eine Google-Suche nach wie vor besser. Bei journalistischer Recherche muss man den Quellen hohes Vertrauen schenken können, daher würden oft wohl ohnehin zusätzliche Quellen konsultiert werden.

Nutzerbefragung Nr. 4

Angaben zur Person

Geschlecht: weiblich

Alter: 45-55

Technische Kenntnisse: eher schlecht

Abteilung: Sendung „Service Gesundheit“

Tätigkeit: Redaktion; Beiträge für Sendungen zusammenstellen

Erfahrung mit FESAD-Benutzerschnittstelle

Welche FESAD-Schnittstelle verwenden Sie?

FESADred

Wie häufig verwenden Sie diese Schnittstelle?

Etwa 2-3 Mal in der Woche.

Wozu verwenden Sie diese Schnittstelle?

Gezielte Suche nach Drehmaterial, wobei die verwendeten Schlagworte bekannt sind und auch mit den Autoren abgesprochen sind. Thematisch orientierte Suchen kommen auch manchmal vor, wenn nicht bekannt ist, in welchen Sendungen dazu Inhalte existieren.

Typische Abfrage:

Ein Beispiel für eine thematische Suche ist die Suche nach Beiträgen zum Thema „Hüfte“, aber hauptsächlich werden im Vornherein bekannte Inhalte gesucht.

Einschätzung der FESAD-Schnittstelle

Wie schätzen Sie die Genauigkeit der Suchergebnisse ein:

Bei thematischen Suchen gibt es oft keine Ergebnisse, was dann sehr zum Teil erstaunlich ist. Ein anderes Beispiel ist die Suche nach einer bestimmten Graphik, wobei dann alle möglichen Suchergebnisse zu Graphiken angezeigt werden, welche dann aussortiert werden müssen. Die Versuchsperson weiss nicht, wie man da die Suche noch genauer formulieren könnte.

Bedienung:

Allgemein gut. Unbefriedigend ist aber beispielsweise wenn man mehrere Ergebnisse in die Mappe ziehen möchte, diese aber nicht alle gemeinsam selektieren kann.

Verbesserungsmöglichkeiten:

-

Test-Aufgabe

Finden Sie Beiträge zu Jaguaren (Tier):

Im Archivmaterial sucht die Versuchsperson nach „Tier Jaguar“, was 13 Treffer ergibt.

Wie schätzen Sie die Qualität dieser Suchergebnisse ein?

Die Treffer sind ziemlich genau, wobei die Versuchsperson das Gefühl hat, dass wahrscheinlich einige Beiträge fehlen. Daher würde sie eigentlich eher nochmals eine neue Suche starten mit nur dem Begriff „Jaguar“ und die längere Ergebnisliste von Hand durchgehen.

Wie empfinden Sie den mit dieser Suche verbundene Aufwand?

Der Aufwand ist leider schon recht gross, aber dann kann man leider nichts machen.

Ist die Suche nach mehrdeutigen Begriffen ein häufiger Anwendungsfall?

Ja, zum Beispiel müsste an bei der vorhin erwähnten Suche nach einer Graphik präzisieren können, um was für eine Graphik es sich dabei handelt.

Finden Sie Beiträge zu hessischen Städten mit Mehrheitspartei CDU:

Eine erste Suche mit den Begriffen „Städte Hessen Magistrat CDU“ ergibt keine Treffer, ohne „Magistrat“ sind es dann aber 96. Diese würde sie dann manuell aussortieren. Da sie aber keine passenden Treffer findet, würde sie über das Internet weitersuchen, um dann mit diesen Informationen neu zu suchen.

Wie schätzen Sie die Qualität dieser Suchergebnisse ein?

Die Genauigkeit ist gar nicht gut, die Versuchsperson meint aber, dass das vielleicht auch mit ihren Fähigkeiten zusammenhängt.

Wie empfinden Sie den mit dieser Suche verbundene Aufwand?

In diesem Fall wäre der Suchaufwand schon recht gross.

Sind mehrstufige Suchen ein häufiger Anwendungsfall?

Eher nicht, sie macht im Arbeitsalltag nur Suchen mit im Voraus klar definierten Begriffen und muss nicht zusätzliche Informationen einholen.

Faceted Search

In Bezug auf Ihre Arbeit, wie nützlich schätzen Sie die facettierte Suche nach mehrdeutigen Begriffen ein?

Der Mehrwert ist nicht so gross. Die Versuchsperson findet ihre Suchmethode („Tier Jaguar“) gleich gut.

In Bezug auf Ihre Arbeit, wie nützlich schätzen Sie die facettierte Suche bei mehrstufigen Suchabfragen ein?

Das käme dann auf die angezeigten Facets an. Wenn das gesuchte Kriterium darin vorkommt, wäre die facettierte Suche schon sehr praktisch.

Wie schätzen Sie die Verständlichkeit der facettierten Suche ein?

Ja, man müsste die angezeigten Facets aber nachvollziehen können.

Wie schätzen Sie die Bedienungsfreundlichkeit der facettierten Suche ein?

Sofern die benötigten Facets angezeigt werden, gut. Sonst wüsste sie aber nicht, wie weiterzusuchen.

Würden Sie die Möglichkeiten der facettierten Suche nutzen, und wie?

Im Arbeitsbereich der Versuchsperson (Gesundheit) fände sie die facettierte Suche eher weniger nützlich, weil die Suche meist nicht so genau einschränken muss, sondern lieber in der Breite sucht.

Haben Sie weitere Bemerkungen zur facettierten Suche?

-

Nutzerbefragung Nr. 5

Angaben zur Person

Geschlecht: männlich

Alter: 35-45

Technische Kenntnisse: recht gut

Abteilung: Bildung, Familie, Service

Tätigkeit: Redakteur

Erfahrung mit FESAD-Benutzerschnittstelle

Welche FESAD-Schnittstelle verwenden Sie?

FESADred

Wie häufig verwenden Sie diese Schnittstelle?

Ungefähr einmal in der Woche

Wozu verwenden Sie diese Schnittstelle?

Gedrehte Inhalte von Autoren durchsehen und auswählen und Zuständigkeit für Content Management (z.B. Ausleihe von Beiträgen verlängern).

Typische Abfrage:

Meisten gezielte Suche von gedrehten Inhalten (Suche über Autorennamen) oder thematische Suchen nach Beiträgen von anderen Sendern mit Suchbegriffen wie „Herz“ oder „Kreislauf“.

Einschätzung der FESAD-Schnittstelle

Wie schätzen Sie die Genauigkeit der Suchergebnisse ein:

Oft ergeben Suchen schon recht viele Treffer, was aber nicht unbedingt etwas Schlechtes ist, weil man dann eine grosse Auswahl hat. Diese Ergebnisse werden dann einzeln durchgegangen. Wenn man den Autor eines Beitrags aber nicht kennt, also nur via inhaltsbezogene Suchbegriffe suchen muss, kommen oft gar keine Treffer, obwohl man weiss, dass es etwas zum gesuchten Thema gibt.

Bedienung:

Mit der Bedienung gibt es keine Probleme.

Verbesserungsmöglichkeiten:

-

Test-Aufgabe

Finden Sie Beiträge zu Jaguaren (Tier):

Die Versuchsperson sucht zunächst im Rohmaterial-Bestand mit „jaguar“, was keine Treffer ergibt. Im Archivmaterial ergibt danach dieselbe Suche 360 Treffer. Sie weiss nicht ob es eine Möglichkeit, Treffer zur Automarke auszuschliessen, würde es dann aber mit „jaguar zoo“ versuchen, da sie davon ausgeht, dass die meisten damit verbundenen Inhalte in einem Zoo gedreht wurden. Dies lässt noch 30 Treffer übrig. Diese Ergebnismenge würde die Versuchsperson dann genauer durchsuchen und sich einzelne Beiträge auch anschauen. Falls da nicht viel dabei wäre, würde sie eine weitere Suche starten mit „jaguar tier“ und diese Ergebnisse mit den vorherigen vergleichen.

Wie schätzen Sie die Qualität dieser Suchergebnisse ein?

Wahrscheinlich werden mit beiden Suchen einige passende Inhalte ausgeschlossen.

Wie empfinden Sie den mit dieser Suche verbundene Aufwand?

Der Aufwand ist nicht schlimm.

Ist die Suche nach mehrdeutigen Begriffen ein häufiger Anwendungsfall?

Das kommt recht oft vor, wobei dann eben meist bei einer grossen Menge von Ergebnissen versucht wird, diese mit zusätzlichen Begriffen zu reduzieren.

Finden Sie Beiträge zu hessischen Städten mit Mehrheitspartei CDU:

Die Versuchsperson würde zuerst über Google nach entsprechenden Städten suchen.

Wie schätzen Sie die Qualität dieser Suchergebnisse ein?

-

Wie empfinden Sie den mit dieser Suche verbundene Aufwand?

Der ist notwendig, weil die Suche auf Google wohl bessere Informationen liefert. Schöner wäre es aber schon, wenn man dies in FESAD machen könnte.

Sind mehrstufige Suchen ein häufiger Anwendungsfall?

Nein, das solche Informationen eigentlich nicht noch recherchiert werden müssen, der Suchauftrag ist in der Regel klar definiert. Wobei oft auch vorgängigen Informationen auf Google-Suchen gesucht werden, aber nicht zwingend um dann dazu in FESADred auch Inhalte zu finden.

Faceted Search

In Bezug auf Ihre Arbeit, wie nützlich schätzen Sie die facettierte Suche nach mehrdeutigen Begriffen ein?

Die Möglichkeit, ein Suchkonzept zu präzisieren, wäre sehr nützlich. Die Versuchsperson würde es FESAD z.B. gar nicht zutrauen, bei einer Suche nach „Jaguar“ und „Tier“ eine genaue und vollständige Ergebnismenge zu liefern, weil Inhalte gar nicht mit diesen Schlagworten versehen sind.

In Bezug auf Ihre Arbeit, wie nützlich schätzen Sie die facettierte Suche bei mehrstufigen Suchabfragen ein?

Das würde den Suchprozess vereinfachen, da der Umweg über das Internet nicht mehr nötig wäre. Dies würde natürlich aber auch von den vorgegeben Facets abhängig sein, da man dann nicht völlig frei in der Suche ist.

Wie schätzen Sie die Verständlichkeit der facettierten Suche ein?

Problemlos verständlich.

Wie schätzen Sie die Bedienungsfreundlichkeit der facettierten Suche ein?

Die ist eigentlich recht gut, vorausgesetzt man wird nicht durch die angezeigten Facets in den Suchmöglichkeiten eingeschränkt.

Würden Sie die Möglichkeiten der facettierten Suche nutzen, und wie?

Ja. Die Versuchsperson würde dann den Hintergrundinformationen der facettierten Suche auch vertrauen, also diese nicht noch im Internet verifizieren wollen. Wobei sie sich dann ohnehin die zurückgegebenen Beiträge anschauen würde, um herauszufinden, ob die den Bedürfnissen entsprechen. Inhaltliche Recherchen (z.B. zu einem bevorstehenden Sendungsthema) würde sie aber weiterhin nicht auf FESAD machen, da sie dann ja nicht nach Bildern sucht.

Haben Sie weitere Bemerkungen zur facettierten Suche?

-

Nutzerbefragung Nr. 6

Angaben zur Person

Geschlecht: weiblich

Alter: 25-35

Technische Kenntnisse: recht gut

Abteilung: Dokumentation und Archive

Tätigkeit: Recherche von Fernsehbildern, welche von den Redakteuren an die Archivare delegiert wird, Digitalisierung von alten Inhalten.

Erfahrung mit FESAD-Benutzerschnittstelle

Welche FESAD-Schnittstelle verwenden Sie?

FESAD Web-Client

Wie häufig verwenden Sie diese Schnittstelle?

Täglich mehrmals

Wozu verwenden Sie diese Schnittstelle?

Für alle oben genannten Tätigkeiten.

Typische Abfrage:

Suche nach Symbolbildern zu einem Thema, z.B. Bilder von Blitzen. Eher selten kommt die Suche nach spezifischen Sendungen vor.

Einschätzung der FESAD-Schnittstelle

Wie schätzen Sie die Genauigkeit der Suchergebnisse ein:

Eigentlich sehr zufrieden, was dadurch bedingt ist, dass die Versuchsperson genau weiss, wie eine Anfrage durchzuführen ist. Es besteht auch enger Kontakt mit den Dokumentaren, um sich über Schlagworte abzustimmen. Probleme gibt es nur dann, wenn Inhalte schlecht dokumentiert sind.

Bedienung:

Mit der ist sie sehr zufrieden, wobei sie das Gefühl hat, das die Suchmaske des Webclients auch leicht überfordernd sein kann, weil es so viele Suchoptionen gibt. Sie schätzt aber, dass sie im Swing-Client die Suchmaske selber konfigurieren kann und Abfragen speichern kann. Eine gewisse Schulung ist aber notwendig.

Verbesserungsmöglichkeiten:

-

Test-Aufgabe

Finden Sie Beiträge zu Jaguaren (Tier):

Eine erste Suche im Feld „Bildinhalt“ und dem Begriff „Jaguar“ mit einer Nicht-Verknüpfung mit dem Begriff „Auto“ ergibt sehr viele Treffer. Eine weitere verfeinerte Suche mit „Jaguar“ und „Tier“ ergibt keine Treffer, dafür kommen bei „jaguar“ und „zoo“ 27 Ergebnisse, davon einige teilweise passende Treffer. Bei einigen Treffern geht es aber um die IAA (Automobil-Messe), also fehlt bei diesen scheinbar das Schlagwort „Auto“.

Wie schätzen Sie die Qualität dieser Suchergebnisse ein?

Diese Suche ergibt keine wirklich brauchbaren Ergebnisse, man müsste in diesem Fall die Suche auf weitere Bestände ausweiten.

Wie empfinden Sie den mit dieser Suche verbundene Aufwand?

Zum Teil weiss man im Vornherein nicht, welche Begriffe man ausschliessen muss (z.B. bei der Jaguar AG), daher muss man oft eine Suche mit zusätzlichen Suchbegriffen und Operatoren (AND, OR, NOT) neu formulieren. Man muss dann oft ziemlich viel ausschliessen. Mit diesem Aufwand ist die Versuchsperson nicht sehr zufrieden.

Ist die Suche nach mehrdeutigen Begriffen ein häufiger Anwendungsfall?

Das kommt gelegentlich vor, wobei man dann eben über den Ausschluss von nicht gewollten Bedeutungen eines Begriffes die Suche verfeinern muss.

Finden Sie Beiträge zu hessischen Städten mit Mehrheitspartei CDU:

Dazu eignet sich FESAD nicht wirklich. Die Versuchsperson würde da gleich auf eine Web-Suche zurückgreifen.

Wie schätzen Sie die Qualität dieser Suchergebnisse ein?

Wenn man die Namen der Städte kennen würde, wäre diese wohl in Ordnung, aber eine direkte Suche in FESAD ergäbe wohl keine brauchbaren Ergebnisse.

Wie empfinden Sie den mit dieser Suche verbundene Aufwand?

Da eine solche Suche nie vorkommt, stört sie das eigentlich wenig.

Sind mehrstufige Suchen ein häufiger Anwendungsfall?

In der Regel geht sie davon aus, dass die Redakteur solchen Recherchen im Vorfeld tätigen und die Suchaufträge eindeutig sind. Manchmal sind Web-Recherchen schon nötig, aber nicht um solche Hintergrundinformationen zu finden.

Faceted Search

In Bezug auf Ihre Arbeit, wie nützlich schätzen Sie die facettierte Suche nach mehrdeutigen Begriffen ein?

Es wäre sehr nützlich, eine Suche im Vornherein eingrenzen zu können, vor allem bei mehrfach belegten Begriffen.

In Bezug auf Ihre Arbeit, wie nützlich schätzen Sie die facettierte Suche bei mehrstufigen Suchabfragen ein?

Die Versuchsperson könnte sich gut vorstellen, dass dies beispielsweise bei der Suche nach Jahrestagen sehr hilfreich sein könnte. Aktuell sind z.B. Personen gar nicht mit weiteren Informationen beschrieben.

Wie schätzen Sie die Verständlichkeit der facettierten Suche ein?

Es ist sehr verständlich, sie könnte sich sogar zusätzliche Anwendungsmöglichkeiten vorstellen.

Wie schätzen Sie die Bedienungsfreundlichkeit der facettierten Suche ein?

Es praktischer, Vorschläge zur Suche zu bekommen, als sämtliche Begriffe selber eingeben zu müssen, daher wäre facettierte Suche möglicherweise leichter und schneller zu bedienen. Eine solche Suche würde dann auch Spass machen, da man schnell zu einem befriedigenden Ergebnis kommt.

Würden Sie die Möglichkeiten der facettierten Suche nutzen, und wie?

Sie würde die facettierte Suche von Anfang an nutzen, ohne über Google zu gehen. Die ganze Suche in FESAD durchführen zu können, wäre ein grosser Mehrwert.

Haben Sie weitere Bemerkungen zur facettierten Suche?

Facettierte Suche wäre auch für die Redakteure von Nutzen, weil die mehr mit Recherchen zu bestimmten Themen zu tun haben.

Nutzerbefragung Nr. 7

Angaben zur Person

Geschlecht: männlich

Alter: 45-55

Technische Kenntnisse: gut

Abteilung: Dokumentation und Archive

Tätigkeit: Sachgebietsleitung; Anforderungsmanager in der FESAD-Entwicklung

Erfahrung mit FESAD-Benutzerschnittstelle

Welche FESAD-Schnittstelle verwenden Sie?

FESAD Webclient.

Wie häufig verwenden Sie diese Schnittstelle?

Täglich mehrmals.

Wozu verwenden Sie diese Schnittstelle?

Für Bestandsabfragen im Rahmen der Langzeitsicherung der Inhalte. Früher hat die Versuchsperson auch noch Suchaufträge bearbeitet.

Typische Abfrage:

Suche anhand von eher technisch orientierten Kriterien, um Inhalte dann digitalisieren zu lassen.

Einschätzung der FESAD-Schnittstelle

Wie schätzen Sie die Genauigkeit der Suchergebnisse ein:

Man kann die Genauigkeit der Treffer nicht formal überprüfen, aber tendenziell ist diese sehr hoch. Der Sucherfolg ist aber natürlich auch stark abhängig von der Suchstrategie. Persönlich bevorzugt es die Versuchsperson, die Suchabfragen eher grob zu formulieren und sie bei Bedarf erst dann weiter einzuschränken, wenn es wirklich sehr viele Treffer gibt.

Bedienung:

Mit der Bedienung ist die Versuchsperson ziemlich zufrieden. Vielleicht liessen sich gewisse Aspekte noch verbessern, aber im Grossen und Ganzen ist die Bedienung recht gut. Gewohnheit spielt dabei eine grosse Rolle, wenn man täglich damit arbeitet.

Verbesserungsmöglichkeiten:

-

Test-Aufgabe

Finden Sie Beiträge zu Jaguaren (Tier):

Im Feld „Bildinhalt“ sucht die Versuchsperson nach „Jaguar“. Danach schaut sie sich einen Treffer an, in der die Automarke gemeint ist, um herauszufinden, wie da die Verschlagwortung aussieht. Damit würde sie dann eine neue Suche mit einer Not-Verknüpfung des nicht gewollten Begriffes, in diesem Fall „Technik“ machen. Dies ergibt 56 Treffer, welche sie dann einzeln durchgehen würde.

Wie schätzen Sie die Qualität dieser Suchergebnisse ein?

Wie empfinden Sie den mit dieser Suche verbundene Aufwand?

Ist die Suche nach mehrdeutigen Begriffen ein häufiger Anwendungsfall?

Ja. In solchen Fällen muss man bei der Suche kreativ sein und unterschiedliche Möglichkeiten testen und kombinieren.

Finden Sie Beiträge zu hessischen Städten mit Mehrheitspartei CDU:

Die Eingabe „Hessen CDU Mehrheit*“ ergibt keine Treffer. Mit „Hessen CDU“ gibt es dann einige Treffer. Wenn da nichts drin wäre, würde die Versuchsperson zum Pressearchiv des HR gehen, und dort nach der Information fragen.

Wie schätzen Sie die Qualität dieser Suchergebnisse ein?

Die Ergebnisse entsprechend eigentlich gar nicht dem, was ich suche. Nur in FESAD lässt sich eine solche Suche nicht wie gewünscht einschränken.

Wie empfinden Sie den mit dieser Suche verbundene Aufwand?

Der Aufwand für eine solche Suche ist gross, weil sich FESAD aber auch nicht dafür eignet.

Sind mehrstufige Suchen ein häufiger Anwendungsfall?

Für die Versuchsperson persönlich nicht, aber bei Redakteuren und Autoren kommt eine solche Suche häufig vor. FESAD wird aber gegenwärtige nur als Produktions-, und nicht als Informationswerkzeug genutzt.

Faceted Search

In Bezug auf Ihre Arbeit, wie nützlich schätzen Sie die facettierte Suche nach mehrdeutigen Begriffen ein?

Man könnte damit ein Konzept definitiv schneller einschränken, facettierte Suche wäre also durchaus von grossem Nutzen.

In Bezug auf Ihre Arbeit, wie nützlich schätzen Sie die facettierte Suche bei mehrstufigen Suchabfragen ein?

Auch liesse sich damit eine Suche sehr gut eingrenzen.

Wie schätzen Sie die Verständlichkeit der facettierten Suche ein?

Die Suchmethode ist leicht verständlich.

Wie schätzen Sie die Bedienungsfreundlichkeit der facettierten Suche ein?

Der Umgang mit facettierter Suche lässt sich schnell erlernen.

Würden Sie die Möglichkeiten der facettierten Suche nutzen, und wie?

Ja, absolut. Weil die Möglichkeit weiterhin besteht, mit einer breiten Suche anzufangen, und ist je nach Bedarf weiter einzuschränken.

Haben Sie weitere Bemerkungen zur facettierten Suche?

Es sollte für die Archivare nicht zusätzliche Arbeit darstellen bei der Schlagwortvergabe. Wenn dies automatisiert werden könnte, wäre das natürlich kein Problem. Dadurch würden wohl auch neue Suchgewohnheiten entstehen.