A model for information retrieval driven by conceptual spaces

Diana Tanase

Faculty of Social Sciences and Humanities

A MODEL FOR INFORMATION RETRIEVAL DRIVEN BY CONCEPTUAL SPACES

DIANA IRINA TANASE

A thesis submitted in partial fulfillment of the
requirements of the University of Westminster
for the degree of Doctor of Philosophy

February 2015

Abstract

A retrieval model describes the transformation of a query into a set of documents. The question is: what drives this transformation? For semantic information retrieval type of models this transformation is driven by the content and structure of the semantic models.

In this case, Knowledge Organization Systems (KOSs) are the semantic models that encode the meaning employed for monolingual and cross-language retrieval. The focus of this research is *the relationship between these meanings' representations and their role and potential in augmenting existing retrieval models effectiveness*.

The proposed approach is unique in explicitly interpreting a semantic reference as a pointer to a concept in the semantic model that activates all its linked neighboring concepts. It is in fact the formalization of the information retrieval model and the integration of knowledge resources from the Linguistic Linked Open Data cloud that is distinctive from other approaches. The preprocessing of the semantic model using Formal Concept Analysis enables the extraction of conceptual spaces (*formal contexts*) that are based on sub-graphs from the original structure of the semantic model. The types of conceptual spaces built in this case are limited by the KOSs structural relations relevant to retrieval: *exact match*, *broader*, *narrower*, and *related*. They capture the definitional and relational aspects of the concepts in the semantic model. Also, each formal context is assigned an operational role in the flow of processes of the retrieval system enabling

a clear path towards the implementations of monolingual and cross-lingual systems.

By following this model's theoretical description in constructing a retrieval system, evaluation results have shown statistically significant results in both monolingual and bilingual settings when no methods for query expansion were used. The test suite was run on the Cross-Language Evaluation Forum Domain Specific 2004-2006 collection with additional extensions to match the specifics of this model.

Contents

List of Figures

List of Tables

List of Symbols, Nomenclature, or Abbreviations

Symbols

d	document
γ	object to formal concept mapping
G	the term-term correlation matrix with components g_{uv}
MAP	mean average precision
M	the term-document matrix with components m_{ij}
μ	attribute to formal concept mapping
π	precision is the proportion of the retrieved documents which are relevant
q	query
ρ	recall is the proportion of the relevant documents retrieved

Acronyms

CLIR	Cross-Language Information Retrieval
ESA	Explicit Semantic Analysis
FCA	Formal Concept Analysis

GEMET	GEneral Multilingual Environmental Thesaurus
GIRT	German Indexing and Retrieval Test database
GVSM	Generalized Vector Space Model
IR	Information Retrieval
KOS	Knowledge Organization System
LLOD	Linguistic Linked Open Data
LOD	Linked Open Data
MLIA	Multilingual Information Access
MLIR	Multilingual Information Retrieval
MT	Machine Translation
NLP	Natural Language Processing
OWL	Web Ontology Language
RDF	Resource Description Framework
RDFS	RDF Schema
SIR	Semantic Information Retrieval
SKOS	Simple Knowledge Organization System
SM	Semantic Model
SPARQL	A recursive acronym for SPARQL Protocol and RDF Query Language
URI	Uniform Resource Identifier

Declaration

I confirm that this thesis represents my own work; the contribution of any supervisors and others to the research and to the thesis was consistent with normal supervisory practice.

Acknowledgements

This thesis was a challenging journey through an exciting interdisciplinary area of research. In my explorations I was fortunate to have the support of my Director of Studies, Dr. Epaminondas Kapetanios whom has always prompted me to think about the big questions and aim for rigor and clarity in my work. Throughout, I have learned a lot more about myself than I have anticipated. During the highs and lows of this journey my family and friends dotted across the world America, France, Italy, England, Romania and Australia have been incredibly supportive. Cups of tea, advice, books, challenging questions, they all helped push this research one step further.

To my family across the world(s)

Chapter 1

Introduction

1.1 Connecting information retrieval to the Semantic Web of Data

The following thesis is positioned at the intersection of information retrieval, natural language processing, and technologies for the Semantic Web. Considering the growing number of knowledge and language resources published on the Semantic Web platform and the consolidation of a diverse technology stack setup to enable machines to *identify*, *represent* and *operate* with semantics, current research is focusing on finding ways to develop methodologies for building semantically-aware applications. In this case, I investigate how to connect a *semantic model* constructed from Semantic Web resources that explicitly define meanings to an information retrieval system. Towards that aim, I introduce a semantic information retrieval model that employs Knowledge Organization System (KOS) type of resources available in the Web of Data that can be instantiated for retrieval in monolingual and cross-lingual settings. This research was motivated by the growing claims that by building a large, distributed, and shared space of language and knowledge resources using Semantic Web technologies, it is possible to create semantically-aware applications and in particular better information retrieval systems. This investigation shows that in restricted settings with the proposed retrieval model and a mathematical toolbox that enables conceptual analysis, a missing element from the Semantic Web stack, a deeper insight is gained on the complex relations between the KOSs as semantic models and retrieval performance.

1.1.1 A prototypical monolingual retrieval task

A prototypical monolingual retrieval task starts with a query, *a vague expression of a question to which a retrieval system answers with a concise, organized response supplying information based on its understanding of the information need* Berry and Browne [2005]. In order to fill in the gap between a user's lexicalization of a question and its true intent, a search system incorporates a certain model of representation for queries and documents that enable it to quantify the relation between them.

For example, the bag-of-words retrieval models represent text as frequencies of words or phrases disregarding any linguistic information or relations between words. This corresponds to the simplifying assumption that the user is explicit in their request and is indeed looking for documents containing the words specified in a query. Consider the request for the query: *account of Funes the memoirist*. A keyword-based search, as this is referred to, would bring up documents containing a combination of the words in this query. It is apparent that this approach is affected by problems that arise from the synonymy and polysemy of words, where documents containing the synonym *story* of the word *account* are missed. In the case of polysemy, when the word occurrences encountered in text have a different meaning from the initial request, irrelevant documents are included in the result set (e.g. the bank *account* of Funes).

These types of problems coupled with the fact that information seeking is actually a series of interactions between a user and a search interface, motivated the quest in retrieval research towards more expressive representations that would provide a better solution to the Information Retrieval (IR) problem of identifying all *relevant documents* with as few non-relevant documents as possible Baeza-Yates and Ribeiro-Neto [2011]. The notion of relevancy aims to capture how closely a document output by the system matches the user's need, how useful a document is to the user and how satisfied is the user with the documents selected as a whole Robertson and Hancock-Beaulieu [1992].

Therefore, the ideal retrieval system with regards to relevance should *use richer representations that operate at a higher level of abstraction*. One way to achieve this is to use semantic models i.e. external knowledge resources that enable interpreting documents by linking segments of text to unambiguous entries in the chosen resource. A positive side-effect of this approach is that synonyms or similar text would be linked to the same entry in the semantic model. This is the first out of two key characteristics for a semantically-aware retrieval system.

1.1.2 Across the language barrier

Access to information across languages is a reality of this century's globalization. Traveling, expanding businesses, immigration and cultural exchanges are all factors in creating contexts where searching through multilingual information is a necessity Peters et al. [2012].

From a user's perspective, a Cross-Language Information Retrieval (CLIR) system is an environment that opens the access path to multilingual information. In a CLIR setting, a user can pose queries in one language, for example English, and get answers from a collection in another language. For the running query from Section 1.1, let us consider a user is interested in finding some relevant documents in a Spanish collection of documents. To bridge the language gap, one solution is to employ a machine translation system to convert the initial query to *el cuento de Funes el memorioso* and than resubmit the search request to the collection of Spanish documents. The results list points in this case to Spanish documents. For users with large passive vocabulary in Spanish the list can be shown immediately to the user. For monolingual users, the CLIR system would need to perform an extra step and provide back-translation to the source language of the results, in this case to English. This scenario shows that *CLIR is in essence a mix of translation processes plus monolingual retrieval*.

In summary, a CLIR system supports a user in discovering answers to queries from resources in a different language than the initial query with several approaches in bridging the language gap: a) translating the query before submitting it; b) translating all the documents in the collection prior to searching; or c) mapping both queries and documents to a language independent representation i.e. an interlingual index. Out of the three, the query translation has proved more efficient, flexible, and less costly with the major drawback of having to handle translation ambiguities or the lack of translations of its query terms. The document translation requires running machine translation programs on a batch set of documents before queries can be submitted. This approach does not scale for CLIR system supporting many language pairs and a dynamic collection. The third approach requires, the existence of an interlingual index that encodes concepts from a source language and maps them to a target language. This entails there is no translation on-the-fly, just matching text from queries and documents to concepts from the index.

In applications, the interlingual indexes EuroWordNet or MultiWordNet, multilingual wordnets bootstrapped from the computational English lexicon WordNet, achieved mixed results for the improvement of system performance. One reason often stated was the limited conceptual coverage with direct impact on the quality of the representations for queries and documents as concepts from the interlingual index.

The controlled human-assisted process involved in maintaining and extending a resource like EuroWordNet is expensive. Nevertheless, recent projects like the newly created multilingual lexical resource BabelNet Navigli and Ponzetto [2012] have been automatically generated from existing resources like WordNet, the dynamic and ever-growing Wikipedia and Machine Translation (MT) systems. Their approach for creating large-scale multilingual resources employs principles embedded in the Semantic Web: a) interoperability, b) sharing, c) dynamic evolution, d) auto-enriching, and e) deployment on an accessible platform. BabelNet has been made available in the Resource Description Framework (RDF), a general method for conceptual description and data serialization formats. Based on these developments in the creation of language and knowledge resources and the consolidation of the Semantic Web platform, the third approach in bridging the language gap has gained more potential. Therefore, the second characteristic I consider for the realization of a semantically-aware retrieval system that operates in cross-lingual contexts is to *implement the interlingual index approach and connect to large-scale resources that are continually updating, cross-linked and accessible through standardized languages*. The next section provides an overview to existing types of resources published on the Semantic Web, a specification language for Knowledge Organization Systems (KOSs), and an example use case for the application of KOSs in the search explorations of digital libraries.

1.1.3 The Semantic Web – a platform for Language and Knowledge Resources

The *Semantic Web* is an initiative setup by Tim Berners-Lee in 2001, aimed to create *an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in co-operation* Berners-Lee et al. [2001]. More than one decade of research has been channelled into defining methods and suitable modeling languages to encode and port existing knowledge into machine readable representations accessible through web standards and protocols, such that a new set of semantically-aware applications can be built. With the first layers of the infrastructure of this enhanced web currently in place, recent research efforts have been focused on the challenges of developing applications that are connected to the new growing body of factual data, as well as language and knowledge resources i.e. the Linked Open Data (LOD) cloud. The technological interpretation of the Semantic Web Stack is described by Figure 1.1. The bottom layers are the basis of the current Web as well as of the Semantic Web. The Uniform Resource Identifier (URI), provides means for uniquely identifying Semantic Web resources; UNICODE serves to represent and manipulate text in many languages, while XML is a markup language that enables creation of documents composed of structured data. The

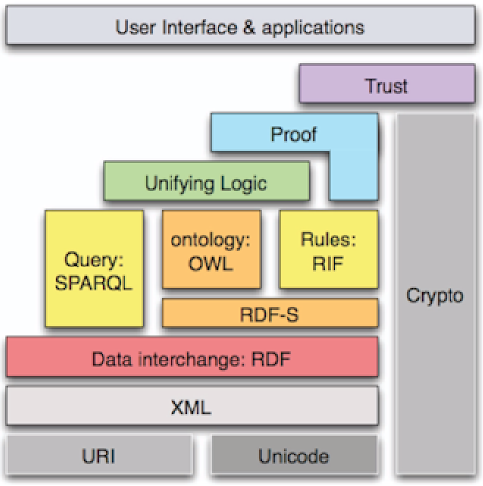


Figure 1.1: Semantic Web Stack [2]

middle layers, standardized by W3C enable building the Semantic Web applications. It includes the framework for representing information Resource Description Framework (RDF) together with the RDF Schema (RDFS) that provides the basic vocabulary for RDF. Using RDFS it is possible to create hierarchies of classes and properties. The Web Ontology Language (OWL) is added to the stack to extend RDFS by adding more advanced constructs to describe the semantics of RDF statements. It allows stating additional constraints, such as for example cardinality, restrictions of values, or characteristics of properties such as transitivity. It is based on description logic and brings reasoning power to the Semantic Web. SPARQL is a RDF query language - it can be used to query any RDF-based data (i.e., including statements involving RDFS and OWL). Querying language is necessary to retrieve information for semantic web applications. The Rule Interchange Format (RIF), the last component of the middle layers, allows describing relations that cannot be directly described using description logic used in OWL.

The top layers of the stack have not been standardized or realized, but indicate the desired qualities of the technologies necessary to complete the Semantic Web vision: trust, proof, unifying logic, and cryptography.

[2] Image Source: http://www.ipgems.com/present/swuidemo/images/layercake200609.png

1.1.3.1 Linguistic Linked Open Data

The Linguistic Linked Open Data (LLOD) cloud is a sub-cloud of the LOD that includes at its center DBpedia and other interlinked monolingual and multilingual language resources (e.g. lexical-semantic resources such as WordNet, corpora, metadata repositories and linguistic data bases). Its emergence has been motivated by the desire to address an old set of issues namely the language and knowledge resources lack of interoperability, challenging creation, maintenance and sharing processes, and their predominantly static nature. The Semantic Web, its principles and affiliated languages and technologies have provided the suitable distribution infrastructure for the resources in the LLOD to be shared, interlinked and enriched. Figure 1.2 is based on the the LLOD cloud diagram by the Open Linguistics Working Group Chiarcos et al. [2012] issued in February 2015.

The high interest in the LLOD cloud stems from its value in enabling and influencing the quality of Natural Language Processing (NLP) tools, systems, applications, and evaluationsCalzolari [2008].

1.1.3.2 Knowledge Organization Systems expressed with SKOS

A particular group of knowledge resources that were migrated to the LLOD cloud are the Knowledge Organization Systems(KOSs). According to the Council on Library and Information Resources [2014]: the term *Knowledge Organization Systems* is intended to encompass all types of schemes for organizing information and promoting knowledge management. Knowledge Organization Systems include classification schemes that organize materials at a general level (such as books on a shelf), subject headings that provide more detailed access, and authority files that control variant versions of key information (such as geographic names and personal names). They also include less-traditional schemes, such as semantic networks and ontologies.

The purpose of KOSs is to organize collections of items (both physical and digital) such that each item can be discovered by users through brows-

[3] Image Source: http://linghub.lider-project.eu/llod-cloud

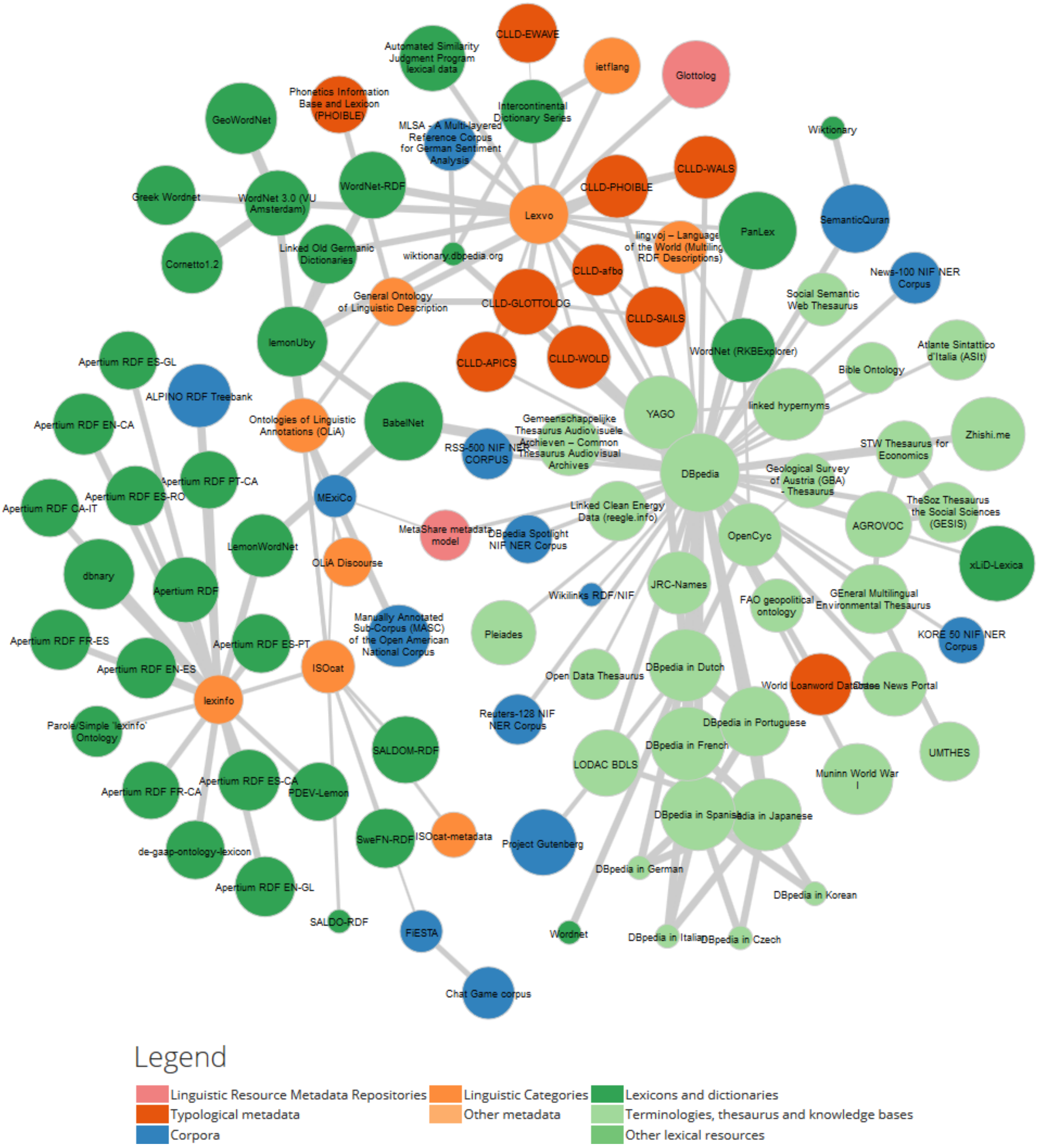


Figure 1.2: LLOD cloud[3]

ing or searching, and most importantly provide paths of discovery for items a user is not aware exist in the collection. It is worth mentioning that there is a cost in organizing collections based on KOSs, and in the library of science field, this was often achieved manually. The critical aspect is providing a balanced meta-description (*metadata*) of an item with both general and specific terms. In contrast in information retrieval, KOSs are employed for indexing collections with the aid of automatic methods of generation of metadata.

Overall, KOSs are extremely valuable for retrieval with an expanding potential due to the changing practices in their creation, maintenance, publication and sharing. In 2009, the W3C announced a new standard: the Simple Knowledge Organization System (SKOS) *a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary* SKOS Primer. This meant that existing Knowledge Organization Systems employed by libraries, museums, newspapers, government portals, and others could now be shared, re-used, interlinked, or enriched. Since then, SKOS has seen growing acceptance in the Linked Data publishers community and more than 20% of existing Linked Open Data is using SKOS relations to describe some aspects of their datasets[4]. A partial listing of existing monolingual or multilingual resources has been compiled[5], with more up-to-date datasets available through the Data Hub portal[6]. Furthermore, the recently updated ISO 25964-1:2011 [2011] and ISO 25964-2:2013 [2013] standards for creation and interlinking of thesauri reflect that the Knowledge Organization Systems publishers community adheres to the principles of Linked Data by using SKOS to publish web-oriented representations of existing and future KOSs.

[4] http://lod-cloud.net/state/

[5] http://www.w3.org/2001/sw/wiki/SKOS/Datasets

[6] http://datahub.io/dataset?q=skos

1.1.3.3 Digital libraries – a use case for exploiting SKOS resources

In applications, a number of museums in particular Rijksmuseum have embraced SKOS as a better way of representing their existing vocabularies used to describe a cultural object. The conversion process detailed in Omelayenko [2008] is not automatic and specific rules of conversion were necessary. Regardless, the benefits are a shareable representation of a cultural repository that is easier to reference and integrate with larger collections.

Another example is Europeana[7], a portal to the digital resources of Europe's galleries, museums, libraries, archives and audiovisual collections. Its current developments[8] are focused on providing its items with a formal semantic context sourced from rich knowledge bases like WordNet[9] and the Art&Architecture Thesaurus[10] represented in SKOS. This contextualization enables exploring the collections beyond keywords-based search by adding more metadata to each digital resource Omelayenko [2010].

This show of support by different parties proves that the group of datasets that employ the SKOS vocabulary as their schema for simple conceptualizations, represents a significant and growing part of the LLOD cloud.

It is also relevant for the next section to emphasize that concept schemes have been used traditionally in Information Science for classifying, searching and finding records (documents) in a collection.

[7] http://europeana.eu

[8] http://eculture.cs.vu.nl/europeana/session/search

[9] http://www.w3.org/TR/wordnet-rdf/

[10] http://www.getty.edu/research/tools/vocabularies/aat/

1.2 Augmenting the classic IR models with representational and translational resources from the LLOD

1.2.1 Problem Statement

Formally, the relationships between KOSs expressed as SKOS datasets structures and IR, respectively CLIR have not been investigated. I setup this research to explore the relevant characteristics of existing SKOS datasets for monolingual and bilingual search settings and describe a retrieval model that integrates such datasets as part of its indexing, matching, and ranking processes. The underlining assumption of this research is that the KOSs investigated are part of the LLOD cloud and are employed for the particular application scenarios of monolingual and bilingual search.

1.2.2 Research Scope

Specifically, this research is concerned with the use of KOSs published in the LLOD as knowledge bases and translation sources for monolingual and cross-language settings within the constraints space presented in Figure 1.3 (figure adapted from Strasunskas and Tomassen [2010]).

- Scope of the search process: this research is applicable to limited domain repositories, restricted Web search, or digital libraries collections
- User input: controlled vocabularies or natural language
- Search goal: monolingual and cross-language information retrieval
- Search processes: indexing (term-based and concept-based), semantic annotation, query processing, query reformulation, matching, ranking
- Architecture: web-services, stand-alone, experimental
- Knowledge representation: primarily multilingual SKOS datasets

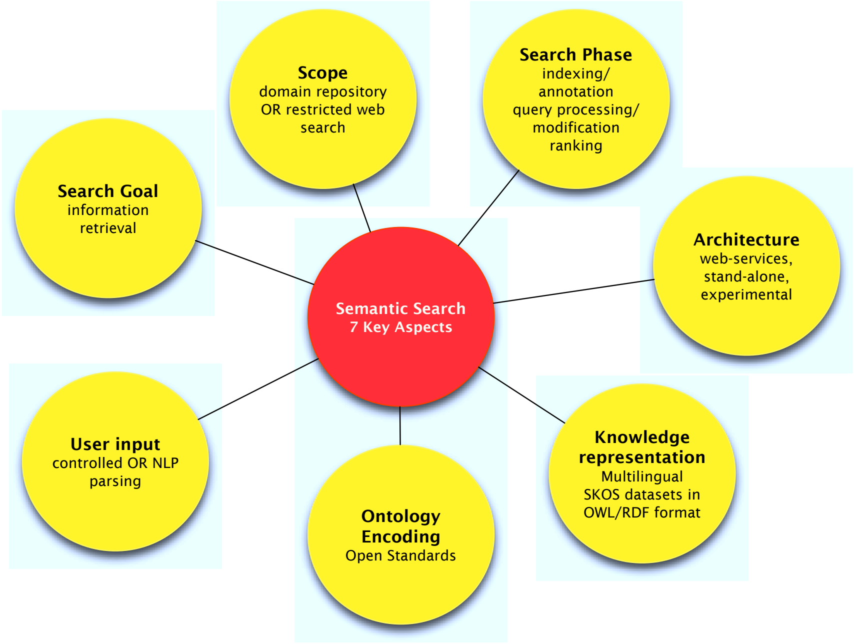


Figure 1.3: Constraints Space

- Ontology encoding: open standards RDF, OWL, SKOS

The search systems that fall within this scope, allow simple keyword search or via controlled vocabularies, and maximize the use of background knowledge from the underlying KOSs (thesauri, taxonomies, etc.) to improve the exploration of the document collection. They also incorporate the two requirements mentioned in Sections 1.1.1 and 1.1.2 for semantically-aware search systems: a) *use richer representations that operate at a higher level of abstraction* and b) *implement the interlingual index approach and connect to large-scale resources that are continually updating, cross-linked and accessible through standardized languages*.

1.2.3 Objectives and Research Questions

Based on the problem space and the scope specified above the objectives of this research are the following:

Main Objective: In the context of information retrieval systems, provide a better understanding of *the relationships between the meanings'* *representations captured by Knowledge Organization Systems (KOSs) expressed as SKOS datasets deployed on the Semantic Web, and their role and potential in augmenting existing retrieval models effectiveness.*

This objective refers to the exploitation of KOSs for semantic annotation and translations. Is the level of detail provided in such datasets, their structure and content useful for integrating them in a IR and CLIR flow of processes? It also points to the fact that research has been focused on how to identify and represent meaning, but there is a missing element in how to operate with these *well-defined meanings* and how does that feed back to the encoding or revisions of KOS type of resources.

Secondary Objective: Describe a retrieval model that explicitly employs such resources for its indexing, matching, and ranking processes.

The secondary objective, spawns from the need to expand the classic Information Retrieval Model to specifically incorporate knowledge resources in all its inner operations.

This leads to the following three groups of research questions (RQ) where each set of questions refers to the semantic model, query and document representation based on the semantic model, and overall performance of the model in different scenarios:

Assessessment of KOS resources as semantic models for monolingual and cross-language retrieval

RQ1: What aspects, more specifically levels of detail of a Knowledge Organization System's representations of meaning are relevant to retrieval processes? How can the lexical bias of KOS resources for its main language (in most cases English) be remedied and more lexical details automatically created for other languages using the cross-schema links between resources in the LLOD cloud?

Queries and documents representations based on the semantic model

RQ2: Considering Formal Concept Analysis as the framework for interpreting the information provided by the semantic model, and that queries and documents are annotated with concepts from the semantic model, what is a suitable representation that is *expressive* enough to capture the connections between documents through their annotating concepts from the semantic model and that *maximizes the exploitation of the semantic model at both lexical and knowledge level*.

Investigating the relationships between the meaning representations from the semantic model and their impact in augmenting existing retrieval models' effectiveness measured based on mean average precision

RQ3: Given the document collection is pre-annotated with concepts from the semantic model, does query expansion with concept labels from the semantic model based on three distinct methods: a) implicit annotation, b) explicit annotation, and c) pseudo-relevance annotation improve retrieval where the baseline is provided by query reformulation based on a local method (weighted terms from top-ranked documents)?

RQ4: Queries and documents are represented as mixed vectors of weighted terms and formal concepts constructed from the semantic model, what is the impact of this representation and ranking parameter α in comparison to existing vectorial represention based only on term-weighting models such as TF-IDF, DLH13, and PL2?

RQ5: The weighted formal concepts part of the defined vectors representing text are language independent, does the bilingual setting outperform machine translation as the baseline?

RQ6: Considering an effective query expansion method how does the formal concept based-expansion of queries and documents perform in comparison?

RQ7: What is the impact of considering all semantic relations in the semantic model in comparison to a retricted set of formal contexts? How do vectorial representations based on core formal concepts impact retrieval?

1.2.4 Research Methodology

The starting point for this research was an exploration into retrieval in general and CLIR issues in particular, with an emphasis on translation. This is described in Section 1.2.4.1. This part of my research led to the consolidation of a new semantic retrieval model, which I tested through comparative studies using the CLEF Domain Specific test-suite outlined in Section 1.2.4.2.

Each stage of this research is characterized by an overall approach: *linguistic*, *user-centered*, and *LLOD*, on how aspects of monolingual and cross-language retrieval can be improved. The solution I arrived to connects existing IR with LLOD resources from the Semantic Web for all the processes in the control flow of a search system.

1.2.4.1 The Exploratory Phase

The linguistic approach In Kapetanios et al. [2006] and Kapetanios et al. [2008], we considered the possibility of building a language-independent query parser that functioned as a Universal Query Language Automaton. The parser was envisioned to incorporate a knowledge base of parametric descriptions of each language (word type order, head directionality, etc.). Thus, the differences between languages were encoded as a set of *parameters*, whose combination according to the Principles and Parameters Theory generates a language's syntax and semantics Baker [2001].

The preliminary validations have shown that at a small scale this approach can lead to improvements in precision and recall of documents retrieved. Unfortunately, a full implementation is not possible due to the limited numbers of language parameters identified for each language. Additionally, no model for the automatic acquisition of parameters Roberts and Holmberg [2005] exists. The Universal Grammar theory is still controversial with very few further developments expected in this direction from the linguistic community where the theory of the pre-specified grammar is considered extreme [Jackendoff, 2003, p.102].

The user-assisted translation approach A line of research in CLIR initiated by Marlow et al. [2008] was to consider the impact of the users' language skills and their search strategies and effectiveness in a multilingual access setting. This led us to consider the involvement of users in a CLIR system translation process. During our participation Tanase and Kapetanios [2008] to iCLEF 2008, the Cross Language Evaluation Forum (CLEF) campaign, we analyzed the logs of a default multilingual search interface on top of the Flickr image collection. The users' task was a known-item retrieval search task that is given a raw unannotated image the users needed to formulate a query such that they can find the image again in the Flickr database.

Our focus for iCLEF has been to determine how the provided query translation assistant was used and if users whom entered their own translations to words into a personal dictionary performed better in the known-item retrieval search task. We could not detect a clear link between the usage of personal dictionaries and the efficiency of the users' search. Nevertheless, the iCLEF experiment as a whole has shown that users will interact with the search system in the translation process regardless of their language abilities.

We further reviewed the potential of user-generated content in the chapter Improving Cross-Language Information Retrieval by Harnessing the Social Web Tanase and Kapetanios [2009]. This revealed that users are willing to become active participants in the creation of web-based multilingual dictionaries by contributing to semantically-enabled social platforms like Wiktionary and OmegaWiki. Yet, these resources are by design ad-hoc and generic and do not include domain specific vocabularies, which are usually part of thesauri or ontologies or other knowledge bases.

These explorations revealed that in the context of the widely adopted Social Web and a growing Web of Data, the future of retrieval systems is strongly-linked to the use of the Web as a lexical resource, as a distribution infrastructure, and as a channel of communication between users.

1.2.4.2 The Problem-Solving Phase

The LLOD approach In order to address the first research question from Section 1.2.3, I investigated if SKOS datasets are suitable for representing meaning to be used by cross-language retrieval applications in *Are SKOS concept schemes ready for multilingual retrieval applications*? Tanase and Kapetanios [2012]. I used as case study the GEneral Multilingual Environmental Thesaurus (GEMET), a core of general terminology for the environment. This investigation is described in full detail in Chapter 3, where I underline the different processes in the flow of a retrieval system and the specific requirements engendered from a SKOS resource. For example, the creation of semantic annotations entails identifying and disambiguating potential occurrences of a thesauri concept in a text. Thus, the concept scheme used requires that its concepts have the approapriate level of lexical detail such that these two NLP operations can be performed. For the situations when the dataset used has more detail in just one language, I have built a set of algorithms that generate a multilingual dataset linking to the original SKOS dataset. The output is another dataset containing more details about the lexical entities that describe concepts. This new dataset, now part of LLOD as *gemet-annotated* dataset, contains specific RDF triples that support concept identification, disambiguation and translation in CLIR.

I thus validated that the characteristics of KOS resources deployed on the Semantic Web, allow them to be customized through a set of algorithms to a richer resource with direct application for retrieval systems.

This allowed me to proceed with the definition of a new retrieval model based on KOSs deployed on the Semantic Web, the space where *things* are assigned a well-defined meaning. In the case of *text* as the *thing*, *semantic annotation* is the technique for determining the meaning of a text by mapping it to a semantic model (SM) like a *thesaurus*, *ontology*, or other type of *knowledge base*. It is *a linking procedure, connecting an analysis of information objects (limited regions in a text) with a semantic model. The linking is intended to work towards an effective contribution to a task of interest to end users* Kamps et al. [2012].

Evaluation There are two lines of evaluation in this research. First, what qualities do KOSs represented as SKOS have to support retrieval, and second, how effective is the proposed retrieval model that extends its specification to incorporate KOS resources for queries and documents representations.

Regarding the potential of KOSs under the new representation language SKOS, I determine it can support different processes in monolingual and bilingual retrieval in Chapter 3, with the condition that the eligible SKOS datasets incorporate three levels of specification: conceptual (relations between concepts), terminological (relations between concepts and labels), and lexical (relations between labels). For the particular case of multilingual SKOS datasets, where there is a dominant language, its lexical specifications can be balanced across different languages with a set of algorithms (described in the same chapter). Also, these algorithms can be applied to any text document not necessarily just the SKOS concepts' definitions.

For the evaluation of the IR prototype system, I approached it as a set of laboratory-style evaluation experiments. This allowed me to compare system configurations on a domain specific test-suite of reusable data (*test collection*, *topics* a.k.a queries, *relevance judgements*) released by the Cross-Language Evaluation Forum (CLEF), an institution with a fifteen years track of organizing information retrieval evaluation campaigns Ferro [2014]. I designed a set of experiments aimed at investigating the impact of representing queries and documents as linear combinations of formal concepts on retrieval. These experiments were run on a benchmark collection from CLEF detailed in Chapter 6.

Each of the experiments contributes towards building a clearer picture of the positive impact of this retrieval model. In retrieval evaluation it is not possible to speak in absolute terms about a model, but for the given setup and by following the theoretical description in Chapter 5 the results obtained were statistically significant in both monolingual and bilingual settings.

1.2.5 A New Semantic Retrieval Model

We cannot hope to make predictions about systems if we cannot reason about their underlying structure, and for this we need some kind of formal-

ity Mooers [1958].

I propose a *semantic information retrieval model* that employs KOSs expressed using SKOS as its sources of conceptual knowledge, interprets them using Formal Concept Analysis (FCA), and describes documents and queries as a linear combination of formal concepts extracted from the semantic model (the chosen KOSs).

In the proposed model, a *semantic annotation* is interpreted as a pointer to a concept in the semantic model that activates all its immediate neighbors that it is linked with. This is important for an IR system, because it expands a document's and query's conceptual fingerprint increasing the chances of a match. This led me to consider the representations of documents and queries as conceptual structures that can capture a concept and its neighbors. This is achieved using formal concepts, whose construction relies on FCA's mathematical toolbox. This model is also described in relation with the Generalized Vector Space Model (GSVM) that aimed to consider the relations between terms in the representations of queries and documents. This retrieval model builds on Miles [2006] investigation of a set-theory based model that uses structured vocabularies expressed in SKOS. In contrast, our model's specification is easier to translate into a system implementation and with the use of FCA provides clear strategies to integrate KOSs within the existing flow of processes of an IR system.

1.2.6 Hypothesis and Predictions

Hypothesis: Expanding queries and documents representations with concepts from semantic models and considering the inner structure of the semantic model allows determining their conceptual overlap and supports matching and ranking for retrieval.

The main advantages of the model can be grouped into the following categories:

A. Extensive exploitation of the semantic model's conceptual and lexical within monolingual and cross-lingual retrieval settings:

- semantic annotation
- disambiguation
- translation
- query and document enhancements
- computation of the conceptual overlap between queries and documents

B. Multiple representational contexts:

- can use several semantic models at the same time
- can preset a bias for a certain semantic model

C. Connects the data-layer of the Semantic Web stack to the application level:

- The Semantic Web provides no assistance in choosing and testing the implications of using certain ontologies or strategies in experimenting with applications, while the suggested model demonstrates the use of Formal Concept Analysis as a framework for interpreting the information in a KOS and mathematical toolbox to build extended representations of documents based on their annotating concepts.

D. Benchmarking of KOSs expressed as SKOS in an application setting:

- It allows to quantify the impact of their structural relations on the performance of a search system. Depending on the experimental results, this can be used as feedback for further revisions of the dataset used as a semantic model.

1.3 Thesis Structure

Related Work This chapter iterates through different monolingual retrieval models proposed throughout time emphasizing their evolving representation of meaning. This demonstrates how mathematics is used for modeling language semantics and its shortcomings.

The gradual introduction of a taxonomy of retrieval models, is followed by the descriptions of their CLIR adaptations, and of a particular type of IR models that employ knowledge bases such as thesauri, ontologies, or others at the core of their model. I also look into the performance issues these models manifested in experiments. Although these latter models have not always outperformed the classic IR models, in domain-specific settings it is clear they are highly suitable for the task.

KOSs expressed using Semantic Web Languages This chapter discusses KOSs representation and formalization before the Semantic Web, the building blocks for describing KOSs in the SKOS language, followed by a discussion of the complex relation between KOSs and retrieval. I investigate closely SKOS specifications encoding of conceptual, terminological, and lexical knowledge. I also introduce a set of algorithms that support the creation of a linked data set that adds a lexical layer to an existing thesaurus. This chapter addresses RQ1.

Formal Concept Analysis: a framework for operational semantics This chapter presents the background notions of Formal Concept Analysis (FCA) necessary for controlling meaning representations defined by semantic models (KOSs) at application level. I introduce FCA's broad spectrum of applications with a long standing history of experimentation in linguistic and information retrieval. This is also demonstrated by the review of existing lattice-based IR models. Based on the FCA's mathematical toolbox presented in this chapter, I build the foundation of a hybrid semantic retrieval model and introduce the process of constructing complex algebraic queries and documents representations that exploit the structural relations within the semantic model (KOSs).

A New Semantic Information Retrieval Model Instance The mathematics behind the proposed retrieval model is detailed with an accompanying example in this chapter. I illustrate how to apply FCA to pre-process a semantic model and how to partition it into *functional formal contexts* and *relational formal contexts*. These types of contexts are instrumental in the main processes in the flow of a retrieval systems: indexing, matching and ranking. Particularly for indexing queries and documents with weighted formal concepts. This chapter addresses RQ2.

Applying the Semantic Information Retrieval model in monolingual and bilingual settings In this chapter, I have brought together all the theoretical elements from Chapter 5 and designed a number of experiments aimed at investigating the impact on retrieval of representing queries and documents as linear combinations of weighted formal concepts.

Four different sets of experiments were run on a benchmark collection from CLEF. Each of the experiments contributes towards building a clearer picture of the positive impact of this retrieval model. For the given setup and by following the theoretical description in the previous chapter, statistically significant results were obtained in both monolingual and bilingual settings. This chapter addresses RQ3, RQ4, RQ5, RQ6, and RQ7.

Conclusions This chapter summarizes the main answers for the research questions I started with, discusses my contribution to knowledge and the future steps towards proving that connecting a retrieval system to KOSs resources can lead to an improved retrieval system in tune with cultural changes of semantics. It also discusses a surprising connection between this research and recent work in cognitive sciences theory of meaning.

Chapter 2

Related Work

This chapter iterates through different retrieval models proposed throughout time emphasizing their evolving representation of meaning in text in monolingual and cross lingual contexts. It demonstrates how mathematics is used for modeling language semantics and the shortcomings of these models.

2.1 Formal description of an IR model

The terminology associated with the description of a retrieval model and a system is well established. In this chapter, I adopt the definitions used in the classic retrieval textbook Baeza-Yates and Ribeiro-Neto [2011].

Definition 1. ***Information Retrieval (IR)*** *Given a collection D containing information items d_i and a keyword query q representing an information need, IR is defined as the task of retrieving a ranked list of information items d_1, d_2,... sorted by their relevance in respect to the specified information need q. In the monolingual case, the content of information items d_i and the keyword query q are written in the same language.*

Definition 2. ***Cross-language Information Retrieval (CLIR)*** *Given a collection D containing documents in language l_D (collection language), CLIR is defined as retrieving a ranked list of relevant documents for a query* *in language l_q (query language), with $l_D \neq l_q$. D is a monolingual collection i.e. all documents in D have the same language.*

Definition 3. ***Multilingual Information Retrieval (MLIR)*** *Given a collection D containing documents in languages $l_1, \ldots, l_n$ for $1 \leq i, j \leq n$, $i \neq j$ then MLIR is defined as the task of retrieving a ranked list of relevant documents for a query in language l_q. These relevant documents may be distributed over all languages $l_1, \ldots, l_n$.*

Definition 4. ***Information Retrieval Model*** *An information retrieval model is a quadruple $<$ $\mathbf{D}$, $\mathbf{Q}$, $\mathcal{F}$, $\mathcal{R}(q_i, d_j)$ $>$:*

- *$\mathbf{D}$ is a set of composed views (representations) for the resources (documents) in the collection.*

- *$\mathbf{Q}$ is a set composed of views (representations) for the user information needs, also called queries.*

- *$\mathcal{F}$ is a framework for modeling resource representations, queries and their relationships.*

- *$\mathcal{R}(q_i, d_j)$ is a ranking function which associates a real number with a query $q_i \in \mathbf{Q}$ and a document representation $d_j \in \mathbf{D}$. Such ranking defines an ordering among the documents with regard to the query q_i.*

In summary, these definitions formalize the retrieval processes that transform both queries and documents based on the mathematical framework into a set of representations that serve as input for the scoring function $\mathcal{R}(q, d)$. This in turn determines the degree of relevance of the document to the user's information need.

Optionally, a user-assisted or automatic feedback process can take place once the list of documents is presented to the user Nie [2010]. This leads to two ways of refining the query representation a) true *relevance feedback*, where the user directly selects documents as relevant or b) *pseudo relevance feedback* (PRF), where it is assumed that the top documents in the list are relevant. In each instance the initial query is expanded and resubmitted.

Definition 5. *An index term is a word or group of consecutive words in a document. In its most general form, an index term is any word in the collection. This is the approach taken by search engine designers. In a more restricted interpretation, an index term is a preselected group of words that represents a key concept or topic in a document. This is the approach taken by librarians and information scientists.*

Definition 6. *Let t be the number of index terms in the document collection and k_i be a generic index term. V = k_1,...,k_t is the set of all distinct index terms in the collection and is commonly referred to as the vocabulary V of the document collection. The size of the vocabulary is t.*

Definition 7. *Let $M = [m_{ij}]$ be a term-document matrix with t rows and N columns, where $m_{ij} = w_{i,j}$ i.e. each entry ij in the matrix is given by the weight associated with the term-document pair (k_i, d_j). Given that M^T is the transpose of M, the matrix $G = MM^T$ is a term-term correlation matrix. Each element $g_{u,v} \in G$ expresses a correlation between terms k_u and k_v, given by*

$$g_{u,v} = \sum_{d_j} w_{u,j} \times w_{v,j} \tag{2.1}$$

2.2 Taxonomy of IR models

There are a number of classic IR models summarized in Table 2.1. In categorizing these models, Baeza-Yates and Ribeiro-Neto [2011] split them into two main groups: *without term-interdependencies* and *with term-dependency* also referred as alternative models.

The first category treats different index terms as independent. This is usually represented in the *vector space models* by the orthogonality assumption of term vectors or in *probabilistic models* by an independency assumption for term variables.

For the other category of models, a certain degree of term interdependency is considered by the model. It is usually directly specified from lexical or knowledge resources or indirectly derived from the co-occurrence of those terms in the whole set of documents Manning et al. [2008].

Table 2.1: Taxonomy of Information Retrieval Models

ℱ	**Description**	**Model**	
		without term-interdependencies	with term-interdependencies
Set-theoretic	Set-theoretic models represent documents as sets of words or phrases. Similarities are usually derived from set-theoretic operations on those sets.	• standard boolean	• extended boolean • fuzzy • set-based
Algebraic	Algebraic models represent documents and queries as vectors, matrices, or tuples. The similarity of the query vector and document vector is represented as a scalar value.	• vector space	• generalized vector space • latent semantic • explicit semantic • mixed • relatedness
Probabilistic	Probabilistic models treat the process of document retrieval as a probabilistic inference. Similarities are computed as probabilities that a document is relevant for a given query.	• binary independence • probabilistic relevance model • uncertainty inference	• language model

2.2.1 Models with term-interdependencies

The *extended boolean model* from the second category was first described in 1983 by Salton et al. [1983]. It aimed to address the basic assumption behind the boolean model that considered relevant only the documents containing all the query terms. In the description of this model term weighting was introduced. This led to the definition of metrics for computing query-document similarity inexistent in the original boolean model.

The *fuzzy set model* assumes that each query term defines a fuzzy set and that each document has a degree of membership in this set. The degree of membership function is constructed from the components of the term-term correlation matrix. An implementation by Ogawa et al. [1991] considers the term relations in the thesaurus to determine the term-term matrix.

In the case of the *set-based model* the standard index terms as basic components are replaced by termsets as subsets of terms occurring within documents and queries. For a query of n terms there are 2^n possible termsets. In order to make this model operational in Pôssas et al. [2005] the authors introduced several restrictions limiting the size of the termsets, employing heuristics to set in place a threshold frequency for terms, and focusing on two types of termsets: *closed* with respect to co-occurence in the same subset of documents, and *maximal* based on the meaningful grouping of query terms.

The key shift in IR models happened in ’85 when Wong et al. [1985] proposed the *generalized vector space model* (GVSM) where the index terms are not considered independent. This was followed by other algebraic models all adding layers of interpretation of text using external resources.

More specifically, *latent models* build meaningful groupings beyond single words using *explicit models* to index texts with respect to generic concepts such as Wikipedia articles Gabrilovich and Markovitch [2007], Cimiano et al. [2009], and *mixed models* that adopt the bag-of-words model, but extend it using taxonomies Woods [1997], ontologies Guarino and Giaretta [1995], networks of concepts Lenat [1998], thesauri School and Priss [2000] or categories derived from WordNet Fellbaum and Vossen [2007]. Furthermore, in the *relatedness model* the semantic relatedness between words is computed using Wikipedia and incorporated into the retrieval ranking process Zesch et al. [2008].

2.2.2 The Generalized Vector Space Model, a reference framework

The original Generalized Vector Space Model (GVSM) presented in Wong et al. [1985] was defined starting with three assumptions. First that an index term k_i is characterized by a set of documents. An index term corresponds to the maximal subset of documents such that every document in the set contains the concept. Second, k_i is unrelated to k_j if the set of documents characterizing k_j does not intersect the set of documents characterizing k_i. Third, the greater the overlap between the document sets characterizing two different k_i and k_j the more similar the two index terms are.

The representations satisfying these assumptions are expressed by the next definition.

Definition 8. *Let document d_j and query q_i be described as weights vectors in the real number space $\mathbf{R}_t$, then in the Generalised Vector Space Model the corresponding ranking function is given by the following:*

i) $\vec{d_j} = (w_{1,j}, w_{2,j}, ..., w_{t,j})$

ii) $\vec{q_i} = (w_{1,i}, w_{2,i}, ..., w_{t,i})$

iii) $\Re(q_i, d_j, G) := \vec{d_j}^T \cdot G \cdot \vec{q_i}$*, where* $G = MM^T$ *is the term-term correlation matrix.*

If $G = I_t$ then any two k_i terms are considered independent and the respective model is a VSM. For the case $G \neq I_t$, a number of instantiations exist that differ in the heuristics used to determine the matrix G.

From a theoretical point of view, all models including the probabilistic ones can be rewritten as instances of $GVSM$ Roelleke [2013]. Two interesting and very different implementations of $GVSM$ are the *topic vector space model* and the *explicit semantic analysis model*.

The Topic Vector Space Model (TVSM) and its variation the enhanced TVSM (eTVSM) Polyvyanyy and Kuropka [2007] considers documents as vectors represented by so-called fundamental topics determined from the content of the documents. These fundamental topics are assumed to be orthogonal and independent from each other. The eTVSM is similar in its assumptions to TSVM, however the documents are now constructed from *interpretation vectors*, where an interpretation is a linking process between a term and a topic. The topics are extracted from the document collection based on a heuristic established during implementation, and together with terms and interpretations are organized as an ontology. This is at best semi-automatic. Despite the challenges of deriving this ontology, the authors consider it essential to model the ontology in close relation to the collection. This can be viewed as a special kind of indexing and a completely different approach to IR models that use knowledge resources defined entirely outside the application space. The Explicit Semantic Analysis model Gabrilovich [2006] considers as index terms concepts from an index collection such as Wikipedia and preserves GVSM's three assumptions Gottron et al. [2011]. More details for this model are described in Section 2.2.3.2.

2.2.3 Adapted models for CLIR

Most of the models in the previous section have been adapted and tested for multilingual settings. In effect, a CLIR scenario is a combination between monolingual retrieval with additional translation processes.

A CLIR model's components depend on the choice of language and knowledge resources used to solve the translation problem. In Peters et al. [2012] these are grouped into: a) machine readable dictionaries (MRD) approaches and b) statistical approaches.

In the first instance, a CLIR system employs resources such as bilingual words and phrase lists or dictionaries, multilingual thesauri and ontologies. The known difficulties in this case are translating multi-word expressions, out-of-vocabulary words, and disambiguating words in order to choose an appropriate translation. The approach has been useful for query translation-based solutions for CLIR implementations such as Hiemstra et al. [2001], Pirkola et al. [1999]. In the evaluation of these particular systems Darwish and Oard [2003], Adriani and Rijsbergen [1999] results underline how critical for the performance of a CLIR system is query formulation and refinement with the user's assistance or through query expansion. Hence, the long standing research track for algorithms that focus on query processing, query expansion, and pre and post translation coupled with the use of machine readable dictionaries.

In contrast, statistical approaches have been developed as a response to the above-mentioned difficulties with MRD. They too can be divided into two categories, the ones that *generate machine-readable dictionary-type resources enriched with statistical probabilities* [Peters et al., 2012, p.65] and others that aim to avoid translations by supporting a unified view of queries and documents and by mapping them to language-independent representations. The details of the latter vary depending on the input resources used. The next sections describe some of the algebraic and probabilistic models from the taxonomy of models in Table 2.1 that have been revisited for CLIR. The more prolific of these being the latent and explicit models, together with the language models.

The generic hypothesis regarding semantics that is incorporated by all these models is the statistical semantics hypothesis, which states that *statistical patterns of human word usage can be used to determine the meaning of text* Turney and Pantel [2010]. Refinements of this hypothesis are stated in each case and a short overview of their main characteristics.

2.2.3.1 Latent Models using corpora for training

The semantics hypothesis: *Distributional hypothesis* relies on the intuition that words with a similar meaning are often used in similar contexts.

The original *latent model* also known as *latent semantic indexing* (LSI) by Deerwester et al. [1990] observed that the term-document matrix disregards that the meaning of words is constrained by its context and solely considers frequencies within a document and the overall collection. Mathematically if the word *red* is encountered in the multiword expression *red tape* or *red flag* there are no differences. The question that emerged was how to capture a word's different contexts of use in more detail.Therefore a new model was instated that replaced the term-document matrix with a word-context matrix where the chosen context was a window of co-occurring words. Sahlgren [2006] investigated questions regarding the size of the window such that expressions like *red tape* will be part of the same window.

The Cross Language Latent Semantic Indexing (CL-LSI) by Rehder et al. [1997] is an extension of LSI. In this case, the word-context matrix is built using parallel documents collections in English, French, and German. A word is represented by a language-independent *vector lexicon* where each component is a number in the semantic space induced by all documents in the three collections. This high-dimensional vector space can be reduced using singular value decomposition. This linear algebra method truncates the number of dimensions considered to reduce noise and enable the discovery of high-order co-occurrences when two words appear in similar contexts.

Advantages

Covers synonymy problems: For any given query it is possible to retrieve documents even if they have no words in common.

Semantics is derived from the training collection: No linguistic or other knowledge resources are required.

Multilingual Retrieval: Due to the language-independent representation of queries and documents the result sets are heterogenous with respect to the document's language. No extra steps for merging results are necessary for multilingual retrieval.

Cost

Dependance on the existence and quality of a parallel or comparable corpora: The LSI and CL-LSI performance depends on the initial training collections used to build the high-dimensional semantic space. The alignment and verification of document collections, and the use of machine translation to bootstrap the parallel collections are difficult processes.

In more recent research, the retrieval models aim to mimic what humans do naturally when communicating. That is use both dictionary definitions, but also derive the sense of words from their context of usage. The expectation is that performance from algorithms that incorporate lexical or knowledge information and take into account words co-occurences will improve such CLIR systems. This is the case for the next set of models.

2.2.3.2 Explicit Models using Knowledge Bases

The explicit models distinguish themselves from latent models by specifically choosing a source of reference for the meaning of words.

The semantics hypothesis: *Referential semantics* suggests that the meaning denoted by a segment of text can be specified by reference to an individual, class, object property or datatype property in some formally defined common sense or domain-specific world knowledge Stuckenschmidt [2012]. It is the equivalent of understanding a text by looking up each word in a dictionary.

The Explicit Semantic Analysis (ESA) is an example model that circumvented the purely statistical techniques employed in LSI by constructing a high-dimensional space of concepts derived from Wikipedia, where a concept is an article in this large encyclopedia Gabrilovich and Markovitch [2009]. A word is represented by a vector of concept frequencies and the composite semantics of a text is a combination of the vectorial representations of the words in the text. This model was applied for computing word relatedness in monolingual and cross-lingual settings Hassan and Mihalcea [2009] with good correlation results to human judgements for classic datasets such as Miller-Charles and WordSimilarity-353. The latter method was tested on six language pairs connecting English, Spanish, Arabic and Romanian and also proved competitive for translations based on direct Wikipedia links versus statistical translation.

An extension of ESA is the Cross-Language ESA (CL-ESA) that proposed indexing documents with respect to their language's Wikipedia articles as in ESA, but the resulting vectors are mapped to vectors to other Wikipedias relying on Wikipedia's cross-lingual structure linking articles to their corresponding articles across languages Cimiano et al. [2009]. Wikipedia has a wide range of articles, with only approximately 7000 concepts common to several languages at the time. Sorg and Cimiano [2012] have improved the cross-language linking used by the CL-ESA and obtained comparable performance for CL-LSI. Also, no additional training corpora is required in this instance.

Advantages

Semantics is derived from an external knowledge resource: It uses existing knowledge resources like Wikipedia that are in a continuous flow of development with increasing coverage of general topics.

Scaleability to other languages: depends on expanding the knowledge resource to new languages.

Cost

Noise: Homonyms from different topics introduce noise and distortion in the composition of vectors and influence the word similarity values.

Weaker performance for specific domains: The idea of using Wikipedia because of its topic range can be counterproductive when addressing a particular topic domain.

2.2.3.3 Language Models for CLIR

The semantics hypothesis: The phrase *language models* was initially defined in speech recognition research to refer to the probability distributions that model the regularities in spoken language and are used to predict the likelihood that the next token in the sequence is a given word.

Language Models (LM) applied in IR define the probability distributions for documents and use them to predict the likelihood of observing the query terms. This enforces the fact that the retrieval system has no knowledge of how queries are generated. In Ponte and Croft [1998] the only assumption made by this model is that the queries are well-formulated by the user with clear discriminatory terms.

The combination of relevance with this model engendered a scoring function that estimates the probability distribution for how often a word is expected to be seen in the set of documents relevant to the query. In its CLIR milestone version by Lavrenko et al. [2002] the authors employ either a parallel corpus or a bilingual dictionary for the estimation. Direct query translation is avoided by creating language-independent representations and performance is characterized by high-precision for the first 5 and 10 results.

Furthermore, Vuliƒá and Moens [2013] fuse topical knowledge and relevance modeling in monolingual and cross-lingual settings into a new model where the estimation function uses a topic model trained on document-aligned bilingual corpus discussing the same events such as Wikipedia articles or news stories. The central idea is to first identify a group of topics (latent variables) and afterwards use these topics to describe documents. For example, from a multilingual collection in English, Italian and Dutch, this model would extract approximately 1000 cross-lingual topics represented by words and their probabilities over documents (e.g. {*tourist, hotel, travel*, ...} in English, {*albergo (hotel), viaggio (journey), viaggiatore (traveller)*, ...} in Italian, and {*reis (travel), toerisme (tourism), hotel (hotel)*, ...} in Dutch). The difficulty is to create semantically coherent topics i.e. where words selected per topic are semantically related.

Advantages

Implicit query expansion and disambiguation

Performance: It achieves performance close to strong mono-lingual baseline in terms of average precision.

Unified model: It is a unified formal model that treats queries and documents in the same way.

Cost

Dependance on the existence and quality of a parallel or comparable corpora: In the instances when corpora needs to be built the use of machine translation software is a necessity; another option for its estimation processes are bilingual dictionaries, but good coverage is essential for this model.

Thus far, the investigated purely theoretical retrieval models do not consider the searcher and his experience. Next, I consider models where search is equivalent to guided explorations. This is primarily supported by models that incorporate knowledge bases. I refer to them as *semantic retrieval models* where the mechanisms for operational semantics are induced by these knowledge bases.

2.3 Words, Facts of the World, and Context

Words *The account of Funes the memoirist* and its equivalent Spanish translation *el cuento de Funes el memorioso*

Facts of the World For the query above a knowledge base would incorporate elements of the following background information. *Funes the memoirist* is the title given in English to a fantasy short story by Jorge Luis Borges about Ireneo Funes who effortlessly learns English, French, Portuguese, Latin from dictionaries and has a perfect mental catalog of everything around him at every moment in time. He also creates an infinite vocabulary that maps natural numbers to words... *nevertheless, he was not very good at thinking. To think is to ignore (or forget) differences, to generalize, to abstract* Borges [1998].

Context An IR task is a sequence of steps to formally determine and represent the meaning of text. A knowledge base adds to the retrieval model a semantic space (a context of interpretation for words) that informs these processes. In short, these models employ external sources that describe common or domain knowledge referred to as *knowledge bases*.

Definition 9. *A knowledge base (KB) is a collection characterized by four types of elements [Sowa, 1999, p.487]: a type hierarchy, a relation hierarchy, a catalog of individuals, and an outermost context (a domain).*

The set of *semantic retrieval models* can be characterized based on what *facts of the world* they incorporate and in what format they are available. In the following section I present how the choice of knowledge resources creates a biased system view of what a relevant document is and how it impacts the actual search experience.

Table 2.2 presents a comparison of how retrieval systems strategies are dependent on the instances of knowledge bases they employ to answer the sample query and how the result sets differ.

In particular, these systems' results set suit a particular type of information seeking namely *sense-making* [Baeza-Yates and Ribeiro-Neto, 2011, p.22], an iterative process of formulating a conceptual representation from a large collection of information. It supports the user for deep analysis and discovery tasks.

In the monolingual retrieval context described in Table 2.2 the prototypical retrieval systems that integrate knowledge bases at their core have the following benefits: a) expansion of the search space, b) user support for an exploratory search behavior for concept and ontology-based search as described by Marchionini [2006], and c) user support for information lookup for Semantic Web Search, where the desired results would be a set of discrete data pieces akin to question answering.

Table 2.2: What would the results set be?

Search Type	**Results set for *account of Funes the memoirist***	**Knowledge Base**
Concept Search	All documents that contain the initial keywords and synonyms of the word *history*.	WordNet
	After identifying a Wikipedia entry *Funes el memorioso* use its associated categories: *Short stories by Jorge Luis Borges, 1942 short stories, Fictional Argentine people* to reformulate the query and expand the search.	Wikipedia
Ontology-based Search	The query is mapped to the SUMO class *FictionalText,* which leads to selecting documents referring to *FictionalText*; it also infers that the query refers to something that falls in one of the subclasses *NarrativeText,* a *MysteryStory*, a *ShortStory* and has all the properties defined for them.	Suggested Upper Merged Ontology (SUMO)
Semantic Web Search	Using DBpedia identify all facts linked to the central entry and build new queries that extract more information about the writer of the book, when he wrote it, etc. The documents retrieved contain more factual aspects directly linked to the central entry.	DBpedia

Table 2.3: A Search System Taxonomy by Knowledge Base

Search	**Knowledge Bases**	**Knowledge Representation Language**
Concept Search	It employs Knowledge Organization Systems (KOS) such as classifications, lexical databases, taxonomies and thesauri that model the underlying semantic structure of a domain by specifying a comprehensive description of the terminological concepts (units of thought) and their semantic relations (narrower, broader, related).	Simple Knowledge Organization System (SKOS)
Ontology-based Search	It uses generic and domain specific ontologies that formally specify a set of classes, properties, and their instances with description logic languages. Classes can be related to each other and assertional axioms make statements about the properties of the instances. The formal axioms constrain the interpretation and well-formed use of ontological entities Gruber [1993].	Web Ontology Language (OWL)
Semantic Web Search	The world is described by a linked set of facts, where each fact or statement expresses the relationship between two resources or a resource and a literal. A resource is uniquely characterized by its associated Uniform Resource Identifier (URI).	Resource Description Framework (RDF)

Table 2.3 describes the three prototypical search systems, differentiated by the type of KB. It also specifies the Semantic Web language available to describe such KBs in anticipation of the discussion in Chapter 3.

The realization of these systems requires that their corresponding retrieval model presents a strategy for mapping documents and queries to elements in the knowledge bases and a representation-induced metric for assessing relevance. Each different semantic model introduces a bias on the interpretation of relevance.

2.3.1 Semantic Information Retrieval

Let us extend the classic description of an Information Retrieval Model by Baeza-Yates and Ribeiro-Neto [2011] to incorporate the semantic model (SM) and its operational mathematical framework for matching queries to documents in an extended definition for semantic information retrieval models.

Definition 10. *Semantic Information Retrieval Model. This information retrieval model is a tuple* $< \mathrm{SM}, D_{\mathrm{SM}}, Q_{\mathrm{SM}}, \mathcal{F}_{\mathrm{SM}}, \mathcal{R}_{\mathrm{SM}}(q_i, d_j) >$ *accompanied by a set of operators defined to map the initial set of queries and documents onto elements of the semantic model.*

- SM *is a knowledge base (ontology, hierarchy of concepts, parallel corpora, or word networks).*
- D_{SM} *is the relative representation to the semantic model of the initial document collection.*
- Q_{SM} *is the relative representation to the semantic model of the initial queries.*
- $\mathcal{F}_{\mathrm{SM}}$ *is the framework in which it is possible to assess that a certain query* q_i *is answered by a document* d_j *(e.g. set-based, algebraic, probabilistic).*

- $\mathcal{R}_{\mathbf{SM}}(q_i, d_j)$ *is a ranking function.*

2.3.2 Instances of SIR models in monolingual and bilingual settings

The following instances of SIR models were extracted from the review of experiments submitted to the Domain-Specific Cross-Language Evaluation Forum (CLEF[1]) track. The initial hypothesis set out to be tested by participating researchers to this track was whether domain-specific enhancements to an IR system provide statistically significant improvements in performance over general information retrieval approaches Kluck [2001]. By domain-specific enhancements this hypothesis referred to resources like thesauri. It directly challenged the potential of knowledge bases for IR. Also, CLEF is one of the major references concerning the evaluation of multilingual information access systems.

These enhancements to IR aimed to handle the *vocabulary disconnect problem* and the *diversity and coverage problem*. The first problem underlines that specific area text content can only be translated by using dictionary or machine translation systems that contain or have been trained for the given domain. The second problem refers to the fact that users expect relevant results that have little overlap and facilitate the full exploration of the collection for a given topic.

The mono-language and cross-language domain-specific track at CLEF studied retrieval on different versions of the German Indexing and Retrieval Test database (GIRT) containing German social science data. Though there were some other smaller corpora provided, GIRT was the largest and most used collection with 151319 documents. Also, GIRT was provided as a parallel corpora in German and English, where each document was enriched with subject metadata from term-based multilingual thesauri in English, German, Russian accompanied by separate files describing bi-directional mappings between these terminologies. Topics in the Text REtrieval Conference (TREC) format i.e. a title query, a description, and a narrative of

[1] http://www.clef-initiative.eu

what are the characteristics of its relevant documents, were offered in English, German and Russian.

Participants investigating in this track chose very different IR models to test their approaches: logistic regression and variations Petras et al. [2005], relevance models Meij and de Rijke [2008], explicit semantic analysis Zesch et al. [2008], language modeling, divergence from randomness, and many more. It is not possible to compare these approaches at system level because not every submission aimed to resolve the same problem, but they shared the same context of experimentation.

The next section describes the limitations of the use of thesauri at this CLEF track due to their terminological nature and how improvements were obtained when more complex structures were constructed from the input set of thesauri.

2.3.3 Application limitations in IR of static thesauri

Figure 2.1 contains a series of entries extracted from three thesauri. Each entry describes a vocabulary term in one or several languages, while alignments between terms in different thesauri are located in separate mapping files created by domain-experts. This basic semantic integration allows switching from the terms of one knowledge system to another and expanding monolingual thesauri such as the mappings between the English Thesaurus of Sociological Indexing Terms (CSA) into terms from the German Thesaurus for Social Sciences (TheSoz).

In the experiments that used one or several of these thesauri, the first task was to devise an algorithm that allowed matching query terms to thesauri terms. The classic method is to identify in the CLEF topics, which are the longest matching entries in the thesaurus Petras et al. [2002] and add them to the initial query. This relies on the explicit mention in the topics of words from the thesauri. If this precondition is not satisfied then the retrieval performance drops. Alternative approaches transform this matching task into a search task. To achieve this, each entry in the thesaurus is handled as a short document and then indexed. Next, each query is submit-

```
TheSoz (bilingual entry)
<entry>
    <german>Absatz</german>
    <german-caps>ABSATZ</german-caps>
    <scope-note-de>nicht im Sinne von
                   Vertrieb
    </scope-note-de>
    <english-translation>sale</english-
translation>
</entry>

CSA-to-TheSoz (EN)
<mapping>
    <original-term>Sales</original-term>
    <mapped-term>selling</mapped-term>
</mapping>

TheSoz-to-CSA (EN)
<mapping>
    <original-term>sale</original-term>
    <mapped-term>Sales</mapped-term>
</mapping>
```

```
TheSoz-to-CSA (DE-EN)
<mapping>
    <original-term>Absatz</original-term>
    <mapped-term>Sales</mapped-term>
</mapping>

INION thesauri
<Descriptor>
    <DE-Russian>продажи</DE-Russian>
    <DE-English>sales</DE-English>
</Descriptor>

INION-to-TheSoz (RU-EN-DE)
<mapping>
   <original-term>продажи</original-term>
   <original-term-eng>sales</original-term-eng>
   <mapped-term>Verkauf</mapped-term>
</mapping>
```

Figure 2.1: Selected entries from the CSA, TheSoz, and INION thesaurus

ted against this new index allowing to retrieve relevant thesaurus entries to be used for query expansion. In Fautsch et al. [2007] all the information available across different thesauri was consolidated based on the separate mappings into single entries wherever it was possible (e.g. Figure 2.2). This produced richer entries containing term lexicalizations such as transliterations, capitalizations, and the conversion of all data to UTF-8 encoding.

In other experiments the initial thesauri entries were expanded with words extracted from the documents where a term was used as metadata and a new structure was created, namely, the Entry Vocabulary Indexing (EVI) used by the University of Berkley in Peters et al. [2005].Thus, a query would get matched to thesauri terms based on the EVI structure. This led to clear improvements in comparison to simple matching Petras [2004].

Other approaches followed in developing methods that connected the given thesauri terminology with their lexicalizations in the documents collections. For example, similarly to EVI, the Meij and de Rijke [2008] use the generative language modeling framework to estimate index term distributions for the thesauri terms, also referred there as thesauri concepts.

```
<entry>
    <german>Absatz</german>
    <german-caps>ABSATZ</german-caps>
    <scope-note-de>nicht im Sinne von Vertrieb</scope-note-de>
    <english-translation>sale</english-translation>
    <german_utf8>Absatz</german_utf8>
    <russian> продажи </russian>
    <translit> prodazhy </translit>
    <mapping>
        <original-term>Absatz</original-term>
        <mapped-term>Sales</mapped-term> </mapping>
        <mapping> <original-term>sale</original-term>
        <mapped-term>Sales</mapped-term>
    </mapping>
</entry>
```

Figure 2.2: Mixed entry obtained from merging information from multilingual thesauri TheSoz, CSA, INION Fautsch et al. [2007]

A query is mapped to concepts and mapped back to query terms after selecting through a special technique called *parsimonization*, the most distinguishing terms given a concept. In effect, the concepts serve as a pivot language in this context. Experimental results have shown that this model significantly outperformed baseline query-likelihood runs, both in terms of mean average precision and early precision on both title-only and title plus narrative queries. Yet the main drawback of any collection-dependent models is the assumption of already annotated documents with concepts from a knowledge base.

Though the results in the eight year run of the CLEF track had been mixed, their legacy is more clarity on what are the elements not investigated and where improvements could be made. In this context, better solutions to the *vocabulary disconnect problem* require to:

1. Identify better matching (annotation) techniques for queries and documents;

2. Use richer resources with linguistic and knowledge data integrated;

3. Exploit the internal semantic relations of the knowledge base such as narrower, broader and related in the case of thesauri;

This echoes Calzolari [2008]'s for interoperable, collaborative creation, dynamic (self-enriching) and distributed language and knowledge resources .

A more recent research example that implements the set of observations above is provided by Bosca et al. [2014]. They demonstrate how to exploit multilingual ontologies for enriching documents representation with multilingual semantic information and on-the-fly mapping of queries to ontology concepts (Edunet portal[1] a resource specifically developed in the context of the project and the domain-specific thesaurus AGROVOC[2]). Also, for any concept enriching a document its parent concepts are also added to the document's representation with a decreasing weight depending on the distance from the concept.

Their results on a multilingual collection of 13000 documents have shown that domain-specific resources led to a significant improvement of CLIR performance, with top-k higher precision (representing the precision obtained after k retrieved documents with k values 5, 10, 20, and 30). Also, as expected, manual annotations of documents have a positive impact on results. These results echo our observations regarding solutions to the vocabulary disconnect problem in the Domain-Specific CLEF track and also the benefits of working with knowledge bases represented using a Semantic Web language such as SKOS.

2.4 Summary

In this related work, a variety of retrieval models were gradually introduced, together with their CLIR adaptations, and a particular type of IR models that employ knowledge bases such as thesauri, ontologies, or others at the core of their model. Though these approaches have not always outperformed the classic IR models, in domain-specific settings it is clear they are highly suitable for the task. I will reinvestigate the Domain-Specific CLEF experimentation space later in Chapter 6, but before that in the next chapter the focus is set on knowledge organization systems (KOSs), their

[1] http://organic-edunet.eu

[2] http://aims.fao.org/standards/agrovoc/about

representations using Semantic Web languages, and their relevance to retrieval.

Chapter 3

KOSs expressed using Semantic Web Languages

Knowledge Organization Systems is a term from the information systems field used in reference to thesauri, ontologies, classification systems, and others. These type of resources are effectively knowledge bases, created to be applied for information search type of applications. This emphasis of the functional aspect of KOSs is missing from the general term knowledge bases. Note that simpler types of KOSs like term lists to be described next, do not qualify as knowledge bases because they do not contain any relationships specifications.

3.1 Before the Semantic Web

Knowledge Organization Systems are first mentioned in the introductory chapter listing the variety of concept schemes encompassed by this terminology. Based on their structure, complexity, and the relationships captured between terms, one possible non comprehensive grouping places authority files, glossaries, dictionaries, gazetteers in the *term lists* group; subject headings, classification schemes, taxonomomies, and categorization schemes in the *classifications and categories* group, and finally, thesauri, semantic networks and ontologies in the *relationship lists* group.

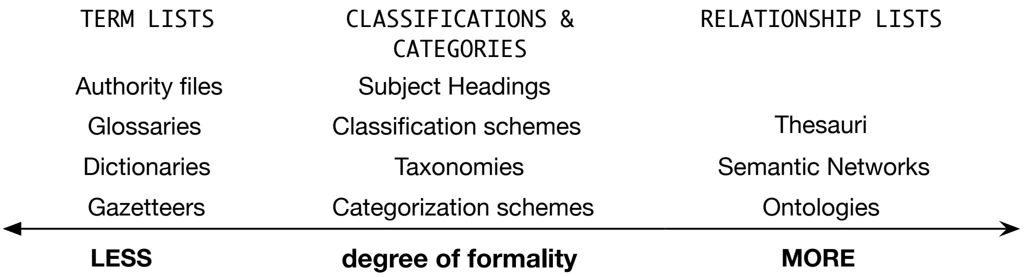


Figure 3.1: Overview of types of KOSs and their degree of formality adapted from Brewster and Wilks [2004]

Figure 3.1 lists them along the degree of formality axis reflecting that the *terms lists* and *classifications and categories* capture and organize the vocabulary of a domain. In the case of the latter group, they only describe a shallow structure between their terms, for example parent-child or is-a-kind-of, while the relationship lists group captures the connections between terms and concepts. Its elements help to *model the underlying semantic structure of a domain for purposes of information retrieval, knowledge discovery, language engineering, and more recently the Semantic Web*. The relationships defined between concepts are richer compared to the classifications and categorizations group. In particular, the associative relationships and broader-narrower relationships are defined in more detail.

In applications such as retrieval, the premise behind creating KOSs is to emphasize a particular view of the world on a document collection. A given document can be characterized in different ways, depending on the KOS that is being used. Selecting what terms or concepts from a KOS should enrich the document is a decision based on the commonality between the item from KOS and the document. This process is either manual or automatic, but the desired outcome is that the added concepts make the document easier to find even when the user is not familiar with the domain or has not enough knowledge of a specific terminology to expand and refine his query.

KOS publishers have continuously revised their representational standards looking forward towards new application fields. At the National Information Standards Organization (NISO) workshop in 1999 entitled *Electronic Thesauri: Planning for a Standard*, three requirements were identified for the future of electronic thesauri: a) persistent identification at the concept level, b) the need for a simple protocol for the distributed querying and response from a KOS, and c) the development of a standard set of metadata attributes for describing a remote KOS. This thread of research was pursued further in the Networked Knowledge Organization Systems Workshop in 2003, where the discussions concentrated on how to change traditional KOSs representations to *support a more semantic-based meaningful Web environment* Soergel [2003].

In the next sections of this chapter, I will demonstrate how the Semantic Web languages and infrastructure have helped realize all these requirements, followed by the investigation of the use case of a KOS expressed as SKOS as the semantic model for monolingual and bilingual settings.

3.2 Semantic Web Languages

The Semantic Web can be viewed as an infrastructure that improves the current Web with formal semantics and interlinked data, enabling flexible, reusable, and open knowledge management systems [Troncy et al., 2011, p.81]. There are many elements to this infrastructure and threads of research from how to port existing knowledge to it, to how to discover it afterwards, search it, query it, reason over it and exploit the formal semantics.

3.2.1 The building blocks

The fundamental data model of the Semantic Web is the Resource Description Framework (RDF). RDF is a language for asserting statements about the world. It uses URIs to identify all resources involved in these assertions, while SPARQL is a language for querying such RDF data. An RDF document consists of (subject, predicate, object) statements or triples such as in Example 3. Each of the examples constructed here represents an RDF description of some of the world facts stated in Section 2.3.

A statement (subject, predicate, object) means the relation denoted by predicate *dbpediaowl:author* holds between the subject *dbpedia:Funes_the _Memorious* and the object *dbpedia:Jorge_Luis_Borges*. Subject, predicate and object are called resources or entities and have unique IDs in this case in the namespace of *DBpedia*. Also, such statements can be seen as directed node-arc-node links in a graph, where an entire RDF document becomes a graph.

Definition 11. ***URI, URL*** *A Uniform Resource Identifier (URI) is a unique identifier according to RFC 2396 Berners-Lee et al. [1998]. A Unified Resource Locator (URL) represents a resource by its primary access mechanism, that is, its network location. A URI can denote any resource. URIs are treated as constants in RDF. Let* $\mathbb{U}$ *be the set of all URIs.*

Example 1.
```
@prefix dbpedia-owl: <http://dbpedia.org/ontology/> .
@prefix dbpedia: <http://dbpedia.org/resource/> .
dbpedia:Funes_the_Memorious dbpedia:author dbpedia:Jorge_Luis_Borges .
```

Definition 12. ***RDF Literal*** *An RDF literal is one of the following: A plain literal of the form <string>(@<lang>), where <string> is a string and <lang> is an optional language tag. A plain literal denotes itself. A typed literal of the form <string>ˆ<datatype>, where <datatype> is a URI denoting a datatype according to XML-Schema2, and <string> is an element of the lexical space of this datatype. A typed literal denotes the value obtained by applying <datatype>'s lexical-to-value mapping to <string> Troncy et al. [2011].*

Let L be the set of all literals and $\mathbb{L}_P$ and $\mathbb{L}_T$ the sets of plain and typed literals.

Example 2.
```
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix dbpprop: <http://dbpedia.org/property/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
dbpedia:Funes_the_Memorious dbpprop:pubDate "1942"^xsd:integer .
```

Example 3.
```
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix dbpprop: <http://dbpedia.org/property/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
```

```
@prefix category: <http://dbpedia.org/resource/Category> .
dbpedia:Funes_the_Memorious dcterms:subject category:1942_short_stories .
```

Definition 13. ***Blank Node*** *A blank node is a unique resource, which is not a URI or a literal. It can only be identified via its properties, and cannot be named directly. Even though some RDF serializations use blank node identifiers, these are just syntactic auxiliary constructs. Blank nodes are treated by RDF as existentially quantified variables. Troncy et al. [2011]*

Example 4. *If apart from the details of this short story, one would like to record its yearly sales across the world, this can be encoded as a blank node defined by year and number of sales on Amazon.com.*

```
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix ex: <http://myexample.org/> .
dbpedia:Funes_the_Memorious ex:sold _:bn .
_:bn ex:year "2013"^xsd:integer .
```

Definition 14. ***RDF Graph*** *An RDF graph G is a set of RDF statements. H is a subgraph of G if $H \subseteq G$. The vocabulary V of G is the set of URIs and literals used in statements in G.*

3.2.1.1 Applying semantics to an RDF graph

The semantics of a set of RDF statements is evaluated through an interpretation function into the domain of discourse. The definition below provided by the latest W3C Recommendation[1] shows that determining an RDF graph's semantics prescribes very basic inferences. All resources and statements constitute the universe of the interpretation $\mathcal{I}$, the predicates are properties of this universe, and the statements can be viewed as the output of three mappings. As an *analogy with natural language, RDF is the alphabet and on its own allows constructing sentences, but without being able to identify what is the meaning of these sentences* [Troncy et al., 2011, p.83].

Definition 15. ***Simple RDF Interpretation*** *A simple RDF interpretation $\mathcal{I}$ of a vocabulary V is a structure consisting of the following Troncy et al. [2011]:*

[1] http://www.w3.org/TR/2014/REC-rdf11-mt-20140225/#simple-interpretations

1. A non-empty set $\Delta^{\mathcal{I}}$ called the universe of the $\mathcal{I}$;

2. A set $\mathbb{P}^{\mathcal{I}}$, called the set of properties of $\mathcal{I}$;

3. A mapping $\cdot^{\mathcal{I}}_{\mathbb{P}} : \mathbb{P}^{\mathcal{I}} \to 2^{\Delta^{\mathcal{I}}}$ of properties into pairs from the domain, defining the extensions of the properties;

4. A mapping $\cdot^{\mathcal{I}}_{\mathbb{U}} : V \cap \mathbb{U} \to \Delta^{\mathcal{I}} \cup \mathbb{P}^{\mathcal{I}}$ from URI references in V into $\Delta^{\mathcal{I}} \cup \mathbb{P}^{\mathcal{I}}$ defining the semantics of URIs in V;

5. A mapping $\cdot^{\mathcal{I}}_{\mathbb{L}} : V \cap \mathbb{L}_T \to \Delta^{\mathcal{I}}$ from typed literals in V into $\Delta^{\mathcal{I}}$

Therefore, two other standardized vocabularies have been introduced to handle the definition of computer-usable meanings: RDF Schema (RDFS) and the Web Ontology Language (OWL). The RDFS allows expressing schema-level information such as class membership, sub-class hierarchies, class attributes (properties), and sub-property hierarchies, while its extension OWL enables richer specification of classes and properties. The RDFS constructs come with entailment rules that can be implemented by inference engines to derive new facts from asserted ones. Apart from this, OWL also enables consistency checks of a given RDF graph with respect to the specifications found in its underlying ontology(ies), such as *dbpedia-owl* that provides the conceptual elements required to formally describe the relation between an author and his book.

OWL has a high level of formal precision that is not appropriate for modeling a wide range of more lightweight vocabularies where relations between concepts are not completely sharp i.e. cannot be described as axioms or facts of the world. A good example of this are all the different KOSs specified earlier in this chapter with the exclusion of ontologies. The alternative specification language proposed is SKOS. An overview of its characteristics, current adoption, and quality issues are part of the next section.

3.2.2 Simple Knowledge Organization Systems (SKOS)

The entry section of this chapter anticipated that the discussions of KOSs publishers in the late 1990s will translate into a new vocabulary and data model that enables KOSs to be ported to a new representation. In this new data model each concept from a given KOS can be individually identified and the whole KOS can be queried and accessed remotely. These aspects and a number of other use cases[1] have been considered by the World Wide Web Consortium (W3C) Semantic Web Deployment Working Group in their iterative development of SKOS into a W3C Recommendation in 2009 as a lightweight intuitive conceptual modeling language for developing and sharing new KOSs.

The following extract from the synopsis of the SKOS Reference [2009] document identifies the key characteristics of using SKOS as a representational model for the concepts of an organizational system:

*Using SKOS, **concepts** can be identified using URIs, **labeled with** lexical strings in one or more natural languages, assigned **notations** (lexical codes), **documented** with various types of note, **linked to other concepts** and organized into informal hierarchies and association networks, aggregated into **concept schemes**, grouped into labeled and/or ordered **collections**, and **mapped** to concepts in other schemes.*

The migration path of an existing KOS to SKOS is not straightforward, but documents such as the SKOS Primer provide a detailed description and examples of usage for the elements in the SKOS vocabulary. Another relevant resource for understanding its components is the *Key Choices in the Design of SKOS* report Baker et al. [2013], which gives an extensive presentation of the decisions in including or excluding certain SKOS components in the final W3C recommendation. For thesauri publishers that follow ISO 25964-1:2011 [2011]; ISO 25964-2:2013 [2013] standards a correspondence table between the two representations is available with the SKOS Primer[2]. Moreover, work reported in Summers et al. [2008], Zapilko and

[1] http://www.w3.org/TR/skosusr

[2] http://www.w3.org/TR/skos-primer/#seccorrespondencesISO

Sure [2009], or Albertoni et al. [2014] underlines the processes of converting an existing resource to SKOS, the use of RDFS extensions and the integration of other Semantic Web vocabularies such as Dublin Core to complement SKOS.

Since SKOS has become a W3C recommendation its usage has expanded across collections. A two year old study on the state of SKOS vocabularies on the web by Abdul Manaf et al. [2012] has identified 478 datasets, where not surprisingly a lot of variety has been observed in the specification of concepts, with some not explicitly declaring concepts as SKOS concepts. It was estimated that a third of these datasets represent term lists, with no linking relations between concepts. Also, the lexical labeling sometimes uses non-SKOS predicates like rdfs:label. Also, SKOS is not the only solution for the representation of conceptual schemes as the authors of Pastor-Sanchez et al. [2009] show. XML, RDF, or the XML Topic Maps specification have potential for this task, but the key advantage of SKOS is that it has already become a W3C recommendation and by its nature is an adaptable specification.

These observations do not exhaustively reflect the state of SKOS adoption, but are a good indicator that before selecting a SKOS dataset for an application a quality check is mandatory. This led towards the development of quality assessment tools like qSKOS and the quality improvement tool Skosify Tool [2011] by Suominen and Mader [2013]. These tools do not assess a SKOS dataset's content from an intellectual point of view, but its compliance with the data model and integrity conditions listed in the SKOS Reference [2009].

3.3 The complex relation between KOSs expressed as SKOS and information retrieval

The introduction of KOSs in this thesis is justified by their role in information search type of applications. Nagy et al. [2011] groups existing applications of KOSs in search settings into six categories: i) filtering, browsing and classification of content, ii) standard indexing by enriching the documents with domain knowledge, iii) autocompletion of a free text query, iv) query formulation and expansion by choosing alternative terms, synonyms for example, or widening and narrowing it using hierarchical relationships, v) recommendation of other documents or query terms based on relationships in the respective KOS, and vi) comprehensive search of the collection using mappings into glossaries to completely describe a domain. The authors attempt to distill the structural requirements for a SKOS dataset depending on the application scenario. It is apparent that the application setting demands that both the structure and the content of a SKOS dataset meets certain prerequisites and next, I pursue the related first research question from Section 1.2.3 by detailing my experiments in adapting a KOS for a monolingual and multilingual retrieval setting.

3.3.1 Use Case: Semantic Search Service Across Mapped Multilingual Thesauri

Search Setting Requirements In 2006 after the SKOS Core Specification Miles and Brickley [2005] was opened for comments, use case scenarios were elicited from the research community and one of the received scenarios was Use Case#3[1] from the AIMS project [2] that focused on the use of a multilingual agricultural thesauri for semantic search, under the assumptions that the resources are indexed by thesauri terms and queries are boolean expressions of concepts. The requirements issued were mostly focused on SKOS supporting existing features of multilingual thesauri: conceptual relations, concept labels (preferred and alternative), concept textual descriptions, multilingual lexical information (e.g. transliteration, acronyms of a concept), and relationships between labels (e.g. translation links or the link between a label and its abbreviation). An application specific feature that was requested was the definition of an indexing relationship.

From the requirements listed above all of them have been realized with

[1] http://www.w3.org/TR/skosucr/#UCAims
[2] http://www.fao.org/aims

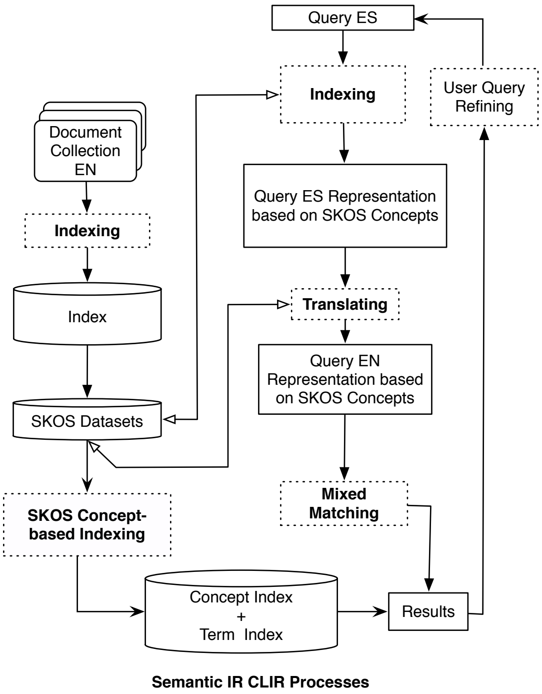

Query ES
Indexing
User Query Refining
Document Collection EN
Query ES Representation based on SKOS Concepts
Indexing
Translating
Index
Query EN Representation based on SKOS Concepts
SKOS Datasets
Mixed Matching
SKOS Concept-based Indexing
Concept Index + Term Index
Results
Semantic IR CLIR Processes

Figure 3.2: CLIR Flow of Processes

the exception of the indexing relationship. The justification of this omission stems from the existence of relationships with the same role in other vocabularies such as the Dublin Core Metadata Elements Set[1] (e.g. dc:subject).

Let us consider the CLIR prototype in Figure 3.2 an instance of a Semantic IR Model. In this scenario the SKOS datasets constitute the reference semantic model and are at the core of several processes: *query indexing* incorporating processing and mapping to interlingual representations, *translation* based on the multilingual labels of a concept or interlinks with other concept schemes, and the *generation of a SKOS concept-based index*. The first two processes are known uses for thesauri, like in the retrieval system created using a SKOS-based astronomical vocabularies by Gray et al. [2009] where queries are built using terminology from the SKOS domain vocabulary. Also, the previous experiments run for the Domain-Specific CLEF mentioned in Section 2.3.2 of Chapter 2 addressed the problem of mapping

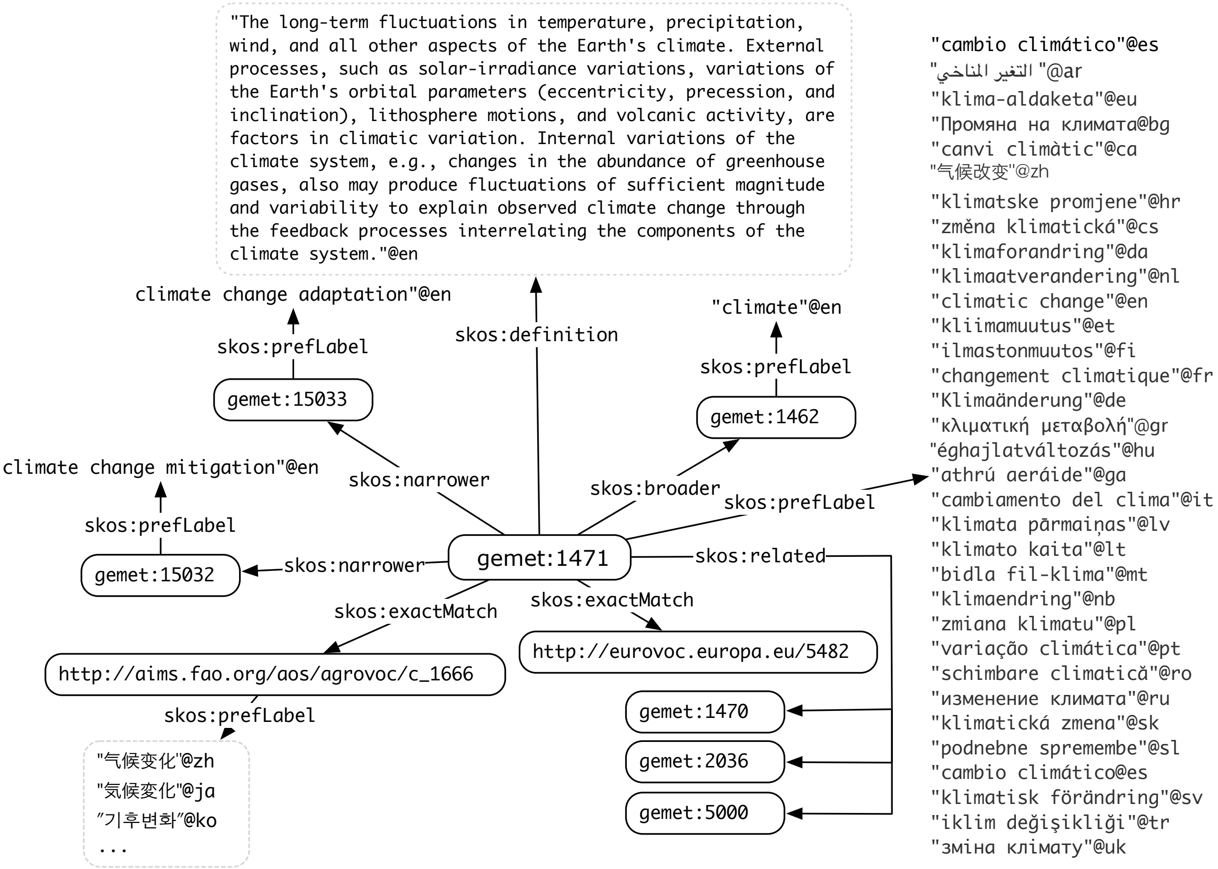
[1] http://dublincore.org/documents/dces/

queries to terms from bilingual thesauri using an exact match (the technique of identifying the longest matching entry term) or fuzzy match (the technique of using a similarity measure to find a suitable candidate entry term). In practice these approaches proved useful only when the concept's label was explicitly used in a text, otherwise there were many queries with no matching concepts which led to poor system-level performance. There were also experiments where concepts were mapped to a construct similar to a frequencies vector of index terms based on the document collection or special training corpora. This last type of approach was more successful than the previous ones and it emphasizes the importance of a concept's textual description to facilitate indexing as opposed to just using labels.

Also, operating on the thesauri as a whole and incorporating all its cross-mappings as part of the evaluation was attempted only by a couple participants of the Domain-Specific CLEF track like Petras [2005] and Savoy and Berger [2006]. This aspect is a side-effect of the traditional ways thesauri were interlinked with terms and relationships specifications being separated across multiple files. In the next paragraphs, I analyze the interplay between the components of a relationship list type of KOS ported to SKOS and information retrieval.

GEneral Multilingual Environmental Thesaurus (GEMET) The GEneral Multilingual Environmental Thesaurus (GEMET), developed by an international consortium, was intended to be used as an indexing, retrieval and control tool for the European Topic Centre on Catalogue of Data Sources (ETC/CDS) and the European Environment Agency (EEA) Albertoni et al. [2014]. It was conceived as a multilingual thesaurus, aimed to define a common general language, a core of general terminology for the environment.

Specifically, its SKOS description of the concept scheme contains a range of basic information about its concepts and the relations between them. As an example, let us refer to Figure 3.3 to illustrate the details captured in GEMET for the concept *climatic change*. The prefixes used are *skos* to denote the namespace *http://www.w3.org/2004/02/skos/core* and *gemet* for



"The long-term fluctuations in temperature, precipitation, wind, and all other aspects of the Earth's climate. External processes, such as solar-irradiance variations, variations of the Earth's orbital parameters (eccentricity, precession, and inclination), lithosphere motions, and volcanic activity, are factors in climatic variation. Internal variations of the climate system, e.g., changes in the abundance of greenhouse gases, also may produce fluctuations of sufficient magnitude and variability to explain observed climate change through the feedback processes interrelating the components of the climate system."@en

Figure 3.3: GEMET Climatic Change Concept

http://www.eionet.europa.eu/gemet/concept/.

Each concept in this dataset has a set of multilingual lexical labels: the unique preferred term, a number of alternative terms, and additional documentation such as definitions and optional notes that describe the concept scheme's domain.

The concepts may be related to one another in a variety of ways. In this example, *climate* (gemet:1462) is a broader concept than *climatic change* (gemet:1471). There are two narrower concepts *climate change adaptation* (gemet:15033) and *climate change mitigation*(gemet:15032), and a number of related concepts with which it shares an unspecified association relation (*climatic alteration* (gemet:1470), *deforestation*(gemet:2036), *man-made climate change*(gemet:5000)).

The *broader* and *narrower* relationships define the hierarchical structure for the concepts, while *related* is used for associations. It should be noted that the broader and narrower terms do not prescribe a subsumption relationship, but are given the definition that any resource annotated via a given term can be retrieved via its broader term. Also, SKOS allows for a loose specification of facts, where *climatic change* narrower than *climate* for example, does not imply that the former is a specialization of the latter.

Another aspect of SKOS resources that can be observed in Figure 3.3 is its support for interconnecting concept schemes. For example, the *gemet:1471* from GEMET with preferred label *climatic change* is an *exact match* to *http://eurovoc.europa.eu/5482* from EuroVoc with preferred label *climate change*.

GEMET specifies mappings between its SKOS concepts and other multilingual datasets such as DBpedia[1],the AGROVOC[2] thesaurus containing specific terms for agricultural digital goods, the UMTHES[3] German-centric thesaurus about environmental protection, and EuroVoc[4] multilingual thesaurus of the European Union.

These mappings represent connection points to the evolving Linked Data

[1] http://wiki.dbpedia.org/DBpediaLive

[2] http://aims.fao.org/website/AGROVOC-Thesaurus

[3] http://data.uba.de/umt/de/concepts/_00014452.html

[4] http://eurovoc.europa.eu/

Table 3.1: GEMET VoID summary description

source	http://www.eionet.europa.eu/gemet
author	European Environment Agency
links:agrovoc-skos	1199
links:dbpedia	3005
links:umthes	3483
namespace	http://www.eionet.europa.eu/gemet
triples	20229105

and in the context of an information access system allow for the exploration of concepts and documents across a concept scheme's boundaries. Examples of mappings are *exact match* (equivalent concepts), *close match* (similar but not equivalent concepts), *broad match* (a more general concept), *narrow match* (a more specific concept), and *related match* (an associated concept).

The Vocabulary of Interlinked Datasets (VoID) description is an optional accompanying document to a SKOS resource. Its aim is to help the discovery of a resource and to summarize some of its characteristics as seen in Table 3.1. Unfortunately, this document is not always updated with the SKOS resource and in this case the VoID document did not contain a count of the EuroVoc or Wikipedia links that are part of the dataset.

In summary, a rich SKOS resource such as GEMET has two levels of structure: a *conceptual level*, where concepts are identified and their interrelationships established; and a *terminological correspondence level*, where terms are associated (preferred or non-preferred) to their respective concepts.

A third level, optional level, can be defined using SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL) SKOS-XL [2009] allowing to define a *lexical level* where lexical relationships are defined to interconnect terms. This lexical level can be further extended to enrich each of the concepts with textual information for Natural Language Processing tasks such as matching concepts to text. In the next sections, the focus is on demonstrating how to automatically add this third level to an existing SKOS resource for this purpose.

3.4 A method for balancing the lexical level information for a multilingual SKOS resource using ESA

Definition 16. *A SKOS concept's signature in a given language is the concatenation of a SKOS concept's textual elements: preferred label, alternative label, and definition in that language. If exact match links to concepts from other concept schemes exist, the equivalent concept's signature is added, following recursively the symmetric equivalence relation across schemes.*

A clarification is necessary regarding the functional role of meaning in a KOS, when trying to explicitly and precisely map text to SKOS concepts. My view is that understanding meaning is a process of constraining what a text refers to by analyzing the words as signifiers of concepts. The result of this process is links i.e. *semantic annotations* that can quantify the about-ness of a text or just establish the existence of a connection between sections in the text and concepts. The process is supported by the evidence provided by a *concept's signature*. Therefore, if more detail is available to construct a *concept's signature* then establishing connections between text and that respective concept improves the outcome of correctly annotating text.

From previous experiments using thesauri or ontologies for semantic annotation it is clear that a SKOS resource needs to support concept matching beyond identification of concept labels. This entails that at a minimum a good SKOS resource should specify for each concept the values for the skos:prefLabel, skos:altLabel, and skos:definition. If the resource does not meet these minimum criteria it is paramount that equivalent concepts from other schemes with that level of detail are used. This process is formalized in Chapter 5.

In the case of a CLIR setting like in Figure 3.2, the SKOS resource should provide multilingual labels or have cross-links to concept schemes in other languages, and overall have the same level of detail in all languages supported to ensure consistent behavior of annotation algorithms. This has been acknowledged as a problem by other researchers when discussing current challenges of the Multilingual Semantic Web Gracia et al. [2012].

Therefore, the tasks are: a) constructing concepts' signatures that exploit the terminological level and the mappings across concept schemes, as well as b) balancing the level of detail in all languages. The extended SKOS resource obtained through these processes will facilitate the concept-indexing stage of the CLIR prototype.

To achieve the first task it is sufficient to prepare and run a series of SPARQL queries to interrogate the SPARQL endpoints the datasets are deployed at. In Chapter 5 this task is formally described as equivalent to computing *functional relational contexts*.

In contrast, the second task requires a heuristic approach for which I devised a two step method: first, enrich an existing SKOS resource using a self-reflection algorithm and second, translate the results.

3.4.1 Step 1: Enriching an existing SKOS resource using a self-reflection algorithm

The aim of this algorithm is to balance the lexical details across different languages for an existing SKOS concept, by exploiting the resource itself. The assumption is that the input resource has a dominant language with concept definitions available. The key requirement in this case is to carefully choose the elements to be translated and added to the concept's signature in another language. All the steps in this algorithm aim to minimize translation errors.

Our proof-of-concept implementation used English as the starting language for the GEMET dataset expressed as SKOS.

Algorithm 1 Generating Annotations

```
INPUT
KOS expressed using SKOS
Select dominant language of the SKOS dataset
for all $c$ SKOS Concept from the resource do
    Index over the concept's definitions content
    Create the term frequency vectors for each concept.
    Semantic annotations
    Using GATE Embedded, tokenize $c$'s definition and identify exact occurrences of other concept labels (preferred or alternate) in the definition

    Phrase extraction
    Using ESA and EN Wikipedia determine content-bearing phrases from $c$'s definition
    Extract groups of words from the definition that have a strong association
    The association function is based on the Language Model metric
end for
OUTPUT
A set of annotations for each $c$ a SKOS Concept from the resource
```

Development Setup The following list describes the main components used in implementing and testing the algorithms described below:

- Search Engine for CLIR: Terrier IR Platform[1]
- Relevant Java Libraries: skosapi[2],
- Natural Language Processing: GATE[3] Embedded is an object-oriented framework for performing Semantic Annotations tasks; APOLDA a GATE Plugin[4]
- Semantic repository: Virtuoso Universal Server[5]
- Other Resources: English Wikipedia

[1] http://terrier.org/

[2] http://skosapi.sourceforge.net/

[3] http://gate.ac.uk/download/

[4] http://apolda.sourceforge.net/

[5] http://virtuoso.openlinksw.com/

- Translation Service: GoogleTranslate
- research-esa[1] an implementation for Explicit Semantic Analysis

Output The algorithm highlights for a given concept other concepts from the resource itself or phrases from the definition. These will become candidates for translation. The intuition behind this is to translate just certain parts of a concept's definition by using the multilingual labels from the resource, thus maintaining the domain of the resource, and short phrases using machine translation. The examples of the concepts identified for *climatic change* are described in Figure 3.4, while the phrases identified for the same concept are in Figure 3.5.

The initial part of this algorithm is to create frequency lists of the terms from the definitions of concepts and keep track of the high frequency words.

The semantic annotation part of the algorithm relies on APOLDA (Automated Processing of Ontologies with lexical Denotations for Annotation) Gate plugin Wartena et al. [2007] that determines annotations based on label matching of GEMET concepts against the text of concepts' definitions. This produced a total of 18120 mentions for the 5208 GEMET concepts that were disambiguated using the algorithm 2.

The task of disambiguating the semantic annotations from the previous algorithm is difficult and the results obtained show how the details specified for each SKOS concept impact the ability to determine if a concept is used in a piece of text in the same sense characterized by the SKOS resource. The algorithm relies on pre-processing the content of the SKOS resource to extract concepts' signatures. For each concept mention, the algorithm measures the relatedness between its concept's signature and the definition of the concept it is annotating. The task of building a well-performing disambiguation algorithm on short texts is out of the scope of this thesis, nevertheless it is mandatory to verify the quality of the semantic annotations process.

This algorithm removed 530 of the 18120 mentions. A third of those re-

[1] http://code.google.com/p/research-esa/

The long-term fluctuations in temperature, precipitation, wind, and all other aspects of the Earth's climate. External processes, such as solar-irradiance variations, variations of the Earth's orbital parameters (eccentricity, precession, and inclination), lithosphere motions, and volcanic activity, are factors in climatic variation. Internal variations of the climate system, e.g., changes in the abundance of greenhouse gases, also may produce fluctuations of sufficient magnitude and variability to explain observed climate change through the feedback processes interrelating the components of the climate system.

Figure 3.4: SKOS concept *climatic change* definition with highlighted semantic annotations

Algorithm 2 Disambiguating Semantic Annotations

```
INPUT
KOS expressed using SKOS
for all $c$ a SKOS Concept do
    Build $c$'s concept signature
    for each of $c$'s neighbors, narrower or broader concepts do
        Add the union of their annotations' labels to $c$'s concept signature
    end for
end for
for all $c$ a SKOS Concept do
    for all $annotation$ identified for $c$ and a SKOS concept do
        Compute the semantic relatedness between the $annotation$'s concept
        signature and $c$'s textual definition
    end for
end for
OUTPUT
Disambiguated set of semantic annotations
```

moved were actually valid annotations but for them there was no definition for that particular concept or the definition was very short (5-6 words) and there were very few neighboring concepts to build the concept's signature. Where applicable, the disambiguation process can be improved by using the cross-links between concept schemes relations such as *exact match* or *close match* to discover further lexical entities and improve a concept's textual signature. Based on this last process the remaining mentions became candidates for translation.

The phrase identification part of the algorithm detects short phrases in a text (2, 3, or 4 words), based on the strength of their association computed by determining the semantic relatedness of their English Wikipedia feature vectors. The respective vectors are computed using research-esa implementation of the Explicit Semantic Analysis algorithm on a local instance of the English Wikipedia[1]. The approach provided good results in identifying generic phrases.

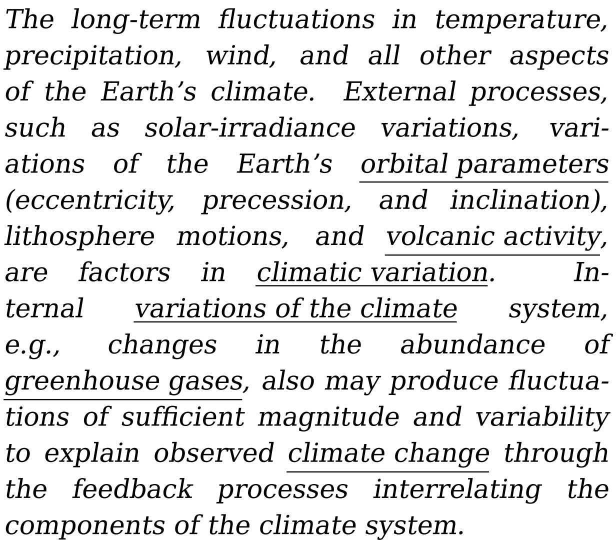
On GEMET it identified 15781 occurrences of 6850 unique content-bearing phrases. These counts are provided after removing any duplicates with the annotations obtained at the previous step. The phrases include multiword-expressions (e.g. *toxic chemical*, *oxygen concentration*, *wind velocity*), named entities (e.g. *New Zealand*), and other phrases (e.g. *pipes supplying water*). The automatically selected phrases have all been manually checked as valid atomic groupings of words (in terms of meaning).

3.4.2 Step 2: Translating Annotations

The output of the enriching stage of this two step method is a mixed set of concepts, phrases, and single words annotating existing concepts from the input KOS. The concepts' labels are either multilingual or monolingual, while phrases and words are monolingual. To create a new layer of lexical information for a chosen target language machine translation is used.

The translation results as expected vary. For single words it is difficult to enforce that the translation matches the domain. In the case of phrase

[1] http://en.wikipedia.org/wiki/Wikipedia:Database_download



The long-term fluctuations in temperature, precipitation, wind, and all other aspects of the Earth’s climate. External processes, such as solar-irradiance variations, variations of the Earth’s orbital parameters (eccentricity, precession, and inclination), lithosphere motions, and volcanic activity, are factors in climatic variation. Internal variations of the climate system, e.g., changes in the abundance of greenhouse gases, also may produce fluctuations of sufficient magnitude and variability to explain observed climate change through the feedback processes interrelating the components of the climate system.

Figure 3.5: *Climatic change* Phrase Identification

translations translated expressions maintain the domain, yet in some instances words have the wrong inflections or the wrong word order. In the particular case of CLIR, words get stemmed during indexing, for example the word *climate* is stemmed to *climat*), thus inflections issues and mixed word order do not affect the particular case of IR as an application domain. By running this algorithm on GEMET, all concepts were enriched with relevant lexical details in Spanish, French and Romanian. In turn, these new annotations can be used to expand the concepts’ signatures in other languages than English.

Serialization The final step in generating a multilingual dataset that links to the original SKOS dataset is to serialize all annotations as RDF triples. The added triples are expressed using SKOS-XL, which provides additional support for identifying, describing and linking lexical entities.

The SKOS data model described in SKOS-XL [2009] defines the property skosxl:labelRelation that links instances of skosxl:Label. It is an extension point, for which I define two object sub properties: *literalTranslation* and *domainTranslation*. The *literalTranslation* is used for handling the machine

Algorithm 3 Serializing Multilingual Annotations

```
INPUT
KOS expressed using SKOS
for all $c$ a SKOS Concept load its annotation maps do
  Source Language: en
  Target Languages: es, fr, ro
  for all $annotation$ a SKOS based annotation do
    Generate label ID
    if $c$ does not have a label for the target language then
      Translate
    end if
    Generate RDF triples description
  end for
  for all $annotation$ a phrase annotation do
    Generate label ID
    Translate phrase
    Generate RDF triples description
  end for
end for
OUTPUT
A new RDF graph of lexical annotations resulted from SPARQL queries
```

translation of a label using Google Translate service, while *domainTranslation* is intended to link labels from different concept schemes, when there exists the transitive relation *exact match* between the concepts the labels refer to. The *domainTranslation* extension is useful to include, since several GEMET concepts have pointers to concepts in the bilingual UMTHES. This is not explored further for now, but is added to the extensions set. These two relations, capturing both translations and context, are in agreement with other work on representing translations for the Semantic Web Montiel-Ponsoda et al. [2011].

I also define a third property, *annotation*. The latter is a sub property of skosxl:hiddenLabel and is used to express a link between the preferred label of a concept and the annotations identified previously from a SKOS concept's definition.

Figure 3.6 details a partial SPARQL query for creating the new lexicalizations dataset. Each skosxl:Label instance is preceded with the string *label* followed by a concept id. In the example query, the id number 1471 points to the *climatic change* concept in GEMET, while ids numbers 1462, 8366, 9327 match respectively *climate*, *temperature*, *wind*. I am using the original ids for creating a GEMET annotated dataset. Note, as expected from the two types of annotations determined in algorithms 1 and 2, it is required to differentiate between the two annotations with a zero or one added to their label. For example, a label like *label_1471_1_en* describes the phrase annotation *climatic variation*, while *label_8366_0_en* describes the semantic annotation with concept *temperature*. The algorithm in this section can be extended to support any number of target languages.

3.5 Summary

In this chapter, I have emphasized the potential of KOSs under the new representation language SKOS to support different processes in monolingual and bilingual retrieval.

For the first part of the research question RQ1 in Section 1.2.3 *What aspects of a Knowledge Organization System's representations of meaning*

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX skosxl: <http://www.w3.org/2008/05skos-xl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX gemet:<http://www.eionet.europa.eu/gemet/gemet-
                                        skoscore.rdf#>

INSERT INTO <http://gemet-annotated> {
        gemet:1471 a skos:Concept ;
                 skosxl:prefLabel gemet:label_1471_0_en ;
                 skosxl:altLabel gemet:label_1471_0_es ;
                 skosxl:altLabel gemet:label_1471_0_ro ;
                 skosxl:altLabel gemet:label_1471_0_fr .

        gemet:label_1471_0_en a skosxl:Label ;
                          skosxl:literalForm "climatic change"@en .
        gemet:label_1471_0_es a skosxl:Label ;
                          skosxl:literalForm "cambio climático"@es .
        gemet:label_1471_0_fr a skosxl:Label ;
                          skosxl:literalForm "changement climatique"@fr .
        gemet:label_1471_0_ro a skosxl:Label ;
                      skosxl:literalForm "schimbare climatică"@ro .

        gemet:1471 gemet:annotation gemet:label_1462_0_en .
        gemet:1471 gemet:annotation gemet:label_8366_0_en .
        gemet:1471 gemet:annotation gemet:label_9327_0_en .

        gemet:1471 gemet:annotation gemet:label_1471_1_en .
        gemet:1471 gemet:annotation gemet:label_1471_1_es .
        gemet:1471 gemet:annotation gemet:label_1471_1_ro .
        gemet:1471 gemet:annotation gemet:label_1471_1_fr .

        gemet:label_1471_1_en  gemet:literalTranslation
                                gemet:label_1471_1_es .
        gemet:label_1471_1_en  gemet:literalTranslation
                                gemet:label_1471_1_ro .
        gemet:label_1471_1_en  gemet:literalTranslation
                                gemet:label_1471_1_fr .
        gemet:label_1471_1_en a skosxl:Label ;
                          skosxl:literalForm "climatic variation"@en .
        gemet:label_1471_1_es a skosxl:Label ;
                          skosxl:literalForm "la variación climática"@es .
        gemet:label_1471_1_ro a skosxl:Label ;
                          skosxl:literalForm "climatice variaţie"@ro .
        gemet:label_1471_1_fr a skosxl:Label ;
                        skosxl:literalForm "les variations climatiques"@fr .


        gemet:1471 gemet:annotation gemet:label_1471_2_en .
        gemet:1471 gemet:annotation gemet:label_1471_2_es .
        gemet:1471 gemet:annotation gemet:label_1471_2_ro .
        gemet:1471 gemet:annotation gemet:label_1471_2_fr .


        gemet:label_1471_2_en  gemet:literalTranslation
                                gemet:label_1471_2_es .
        gemet:label_1471_2_en  gemet:literalTranslation
                                gemet:label_1471_2_ro .
        gemet:label_1471_2_en  gemet:literalTranslation
                                gemet:label_1471_2_fr .

        gemet:label_1471_2_en a skosxl:Label ;
                          skosxl:literalForm "processes"@en .
        gemet:label_1471_2_es a skosxl:Label ;
                          skosxl:literalForm "los procesos de"@es .
        gemet:label_1471_2_ro a skosxl:Label ;
                          skosxl:literalForm "procese"@ro .
        gemet:label_1471_2_fr a skosxl:Label ;
                          skosxl:literalForm "processus"@fr .
}
```

Figure 3.6: SPARQL query to serialize annotations and translations for the *climatic change*

are relevant to retrieval processes? it is now possible to conclude that the eligible SKOS datasets for the retrieval application scenario need to be *relationship list type of resources* and incorporate three levels of specification: *conceptual* (relations between concepts), *terminological* (relations between concepts and labels), and *lexical* (relations between labels). Yet the required third level when present has to have a similar level of detail across all languages.

For the second part of RQ1 *How can the lexical bias of KOS resources for its main language (in most cases English) be remedied and more lexical details automatically created for other languages using the cross-schema links between resources in the LLOD cloud*? I have defined the algorithms in Section 3.4 to add more lexical detail automatically for an existing resource. The goal of the algorithms is the construction of concepts' signatures in all languages supported by the chosen KOS, and for them to be used as basis for NLP processes like matching concepts to text beyond the identification of concept labels.

The output of the application of these algorithms for GEMET was published under the Open Database License to be used as needed (see link[1]).

[1] http://datahub.io/dataset/gemet-annotated

Chapter 4

Formal Concept Analysis: a framework for operational semantics

Formal Concept Analysis (FCA) is an area of applied mathematics that provides *a mathematical theory of concepts and concept hierarchies* Ganter and Wille [1999] allowing for the formal manipulation of conceptual structures. Its usage has extended over the years from domains such as data analysis, knowledge representation and information management towards the interdisciplinary area of information science. This chapter describes the background notions of FCA necessary in controlling meaning representations defined by semantic models at application level. The term *operational semantics* will be used to denote the interpretation within the application space of the meanings described by the chosen semantic models.

4.1 Basic Definitions and Notations

The next sections introduce the basic notions from the FCA's mathematical toolbox, which will enable the presentation of past applications of FCA in three distinct fields namely natural language processing, information retrieval, and Semantic Web. FCA's role in each of these instances is to help construct, discover and explore conceptual structures representations that can improve a certain type of application or task. In this research its applicability will be proved in the next chapter, where FCA is employed to analyze large-scale SKOS datasets that are then integrated into the description of the proposed semantic IR model.

Definition 17. *A formal context $\mathbb{K}$ consists of a triple (G, M, I) where G and M are two sets, and I is a relation between G and M. The elements of G are called objects and the elements of M are called attributes of the context. In order to express that an object g is in a relation with an attribute m, we write gIm or $(g, m) \in I$. This reads g has attribute m.The set of all concepts for this context is denoted by $\mathfrak{B}(G, M, I)$.*

Definition 18. *For a set $A \subseteq G$ of objects the derivational operator* **prime** *defines $A^{'}$:= { $m \in M|$ gIm for all $g \in A$} i.e the set of common attributes for all the objects in A. Correspondingly, for a set $B \subseteq M$ of attributes $B^{'} := \{g \in G|$ gIm for all $m \in B$} i.e. the set of objects which have all attributes in B.*

Definition 19. *A* **formal concept** *in the context (G, M, I) is a pair (A, B) with $A \in G$, $B \in M$, $A^{'} = B$ and $B^{'} = A$. The sets A and B are referred to as the extent, respectively the intent of the formal concept. The set of all concepts of (G, M, I) is denoted by $\mathfrak{B}(G, M, I)$.*

Definition 20. *The ordering of concepts. If (A_1, B_1) and (A_2, B_2) are concepts of a context, (A_1, B_1) is called a subconcept of (A_2, B_2), provided that $A_1 \subseteq A_2$ (which is equivalent to $B_2 \subseteq B_1$). In this case, (A_2, B_2) is a superconcept of (A_1, B_1) i.e. $(A_1, B_1) \leq (A_2, B_2)$. The relation $\leq$ is called the hierarchical order (or simply order) of the concepts. The set of all concepts of (G, M, I) ordered in this way is denoted by $\underline{\mathfrak{B}}(G, M, I)$ and referred to as the* **concept lattice** *of the context (G, M, I).*

Theorem 1. *The Basic Theorem on Concept Lattices. The concept lattice $\underline{\mathfrak{B}}(G, M, I)$ is a complete lattice in which infimum and supremum are given by:*

$$\bigwedge(A_i, B_i) = ((\bigcap A_i), (\bigcup B_i)^{'}) \tag{4.1}$$

$$\bigvee(A_i, B_i) = ((\bigcup A_i)', (\bigcap B_i)) \tag{4.2}$$

A complete lattice V is isomorphic to $\underline{\mathfrak{B}}(G, M, I)$ if and only if there are mappings $\widetilde{\gamma} : G \longrightarrow V$ and $\widetilde{\mu} : G \longrightarrow V$ such that $\widetilde{\gamma}(G)$ is supremum-dense in V, $\widetilde{\mu}(M)$ is infimum-dense in V and gIm is equivalent to $\widetilde{\gamma} \leq \widetilde{\mu}$ for all $g \in G$ and all $minM$. In particular, $V \cong \mathfrak{B}(V, V, \leq)$.

For the special case $V = \underline{\mathfrak{B}}(G, M, I)$, the $\widetilde{\gamma}$ and $\widetilde{\mu}$ that satisfy the conditions of the theorem are defined by γ and μ in the next definitions.

Definition 21. *Relating G and M to $\mathfrak{B}(G, M, I)$. Let (G, M, I) be a context and $\mathfrak{B}(G, M, I)$ its associated set of concepts for which the following mappings are defined:*

- *$\gamma : G \longrightarrow \mathfrak{B}(G, M, I)$ with $\gamma g = \left(\{g\}'', \ \{g\}'\right)$ is the object to formal concept mapping*
- *$\mu : M \longrightarrow \mathfrak{B}(G, M, I)$ with $\mu m = \left(\{m\}', \ \{m\}''\right)$ is the attribute to formal concept concept mapping*
- *$Ext : \mathfrak{B}(G, M, I) \longrightarrow \mathcal{P}(G)$ with $Ext(c) = \{g \in G | \ \gamma g \leq c\}$ is the extension mapping*
- *$Int : \mathfrak{B}(G, M, I) \longrightarrow \mathcal{P}(M)$ with $Int(c) = \{m \in M | \ \mu m \geq c\}$ is the intension mapping*

Apart from these mappings, the *plus operator* allows an alternative mode of conceptualization different from the conjunctive mode of formal concepts Valverde-Albacete and Peláez-Moreno [2011]. For any selected object set the A^+ represents the set union of all the attributes corresponding to objects in A. Similarly, B^+ derives the set union of all the objects for the attribute set B.

Definition 22. *For a set $A \subseteq G$ of objects the derivational operator* **plus** *maps it to $A^+ := \{ \ m \in M | \ \exists g \in A$ with $gIm \ \}$ the union set of attributes for the objects in A. Correspondingly, for a set $B \subseteq M$ of attributes $B^+ := \{g \in G | \ \exists m \in B$ with $gIm\}$ the union set of objects, which have at least one attribute in B.*

It is therefore possible to construct smaller contexts from the original context starting with a kernel object or attribute set, where the size of these contexts depends on the number of times the *plus operator* is applied. These are known as *neighborhood contexts*.

Definition 23. *A neighborhood context of a set $A \subseteq G$ of objects is a smaller context derived from the formal context $\mathbb{K}$ by applying recursively the plus operator. The corresponding neighborhood lattices consist of a concept and its neighbors. A plain n-m-neighborhood starts with an object and has the plus operator applied 2n - 2 times to obtain the set of objects and 2m - 1 times to obtain the set of attributes Priss and Old [2010].*

4.2 Application Domains

The examples presented in the following sections have been selected because they emphasize that FCA takes an intuitive approach to modeling the world in terms of objects (also referred as entities), definitional attributes (also referred as features), and the corresponding relationships between them. This may seem as a simplified way of viewing data, but it is an approach that lends itself to conceptual classifications and partial ordering. Moreover, according to Wille [2005] formal concepts can stand in place of cognitive acts and knowledge units potentially independent of language. In general, formal concepts as a mathematical representation have proven their *plastic adaptability* Ganter et al. [2005] as described in the following sections.

4.2.1 Natural Language Processing

In the context of NLP, the FCA community has investigated formalizing linguistic resources such as thesauri or wordnets Priss [2004] to facilitate visual exploration of such resources, as well as building new linguistic resources bootstrapped using FCA-based algorithms. Furthermore, Jaansen [2002] has argued in favor of a way of structuring the interlingual meanings in a multilingual lexical databases using formal concepts.

$sense_1$: (n) change, alteration, modification (an event that occurs when something passes from one state or phase to another)
direct hyponym $sense_5$: (n) acceleration (an increase in rate of change)
direct hypernym $sense_6$: (n) happening, occurrence (an event that happens)
$sense_2$: (n) variety, change (a difference that is usually pleasant)
$sense_3$: (v) exchange, change, interchange (give to, and receive from, one another)
$sense_4$: (n) a different or fresh set of clothes

Figure 4.1: WordNet descriptions of the different synsets of word *change*

FCA's flexibility stems from being able to choose the objects and attributes to consider to define the formal contexts, based on relevancy to the application domain, yet use established mechanisms, to construct the corresponding concept lattice, where formal concepts determine a clustering of objects and attributes.

For example, the lexical concept **change** extracted from WordNet is specified in its original description using semantic relations (hypernymy, hyponymy) between synsets Miller and Fellbaum [2007], where synsets are grouping of synonym words. All words that form a synset share a sense. The formal context $\mathbb{K}_{change}$ is constructed by extracting a small set of word-sense relations based on the WordNet's specification. The objects in this context are the words, while the attributes are the word senses labeled $sense_{number}$. Each synset defines a facet of the different meanings of the word *change*. In the case of hypernyms words all their hyponyms will also share that sense.

A formal context can be visually represented, using a cross table i.e. a rectangular table, where the rows are headed by the object names and the columns are headed by the attributes names.

From the $\mathbb{K}_{change}$ context, the corresponding concept lattice is derived and represented in Figure 4.2 by a line diagram a.k.a Hasse diagram (realized using ConExp Yevtushenko [2000] and GraphViz[1]). Each circle indicates a concept. Each concept's position in the diagram and its connecting edges indicate a subconcept, respectively superconcept relation between concepts. Note, that the diagram shows next to each circle only the concept's objects and attributes that are not specified by a subconcept or a superconcept.

[1] http://www.graphviz.org/

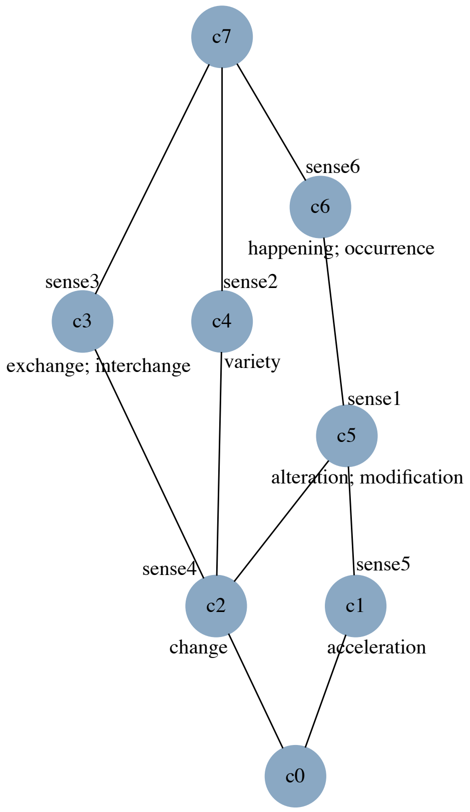


Figure 4.2: The Hasse Diagram for $\mathbb{K}_{change}$ context

Table 4.1: Cross table for context $\mathbb{K}_{change}$ where the objects are the synonyms of word *change*, the attributes are its WordNet senses, and the incidence relation is the semantic relation from WordNet

$\mathbb{K}_{change}$	$sense_1$	$sense_2$	$sense_3$	$sense_4$	$sense_5$	$sense_6$
change	×	×	×	×		×
alteration	×					×
modification	×					×
variety		×				
exchange			×			
interchange			×			
acceleration	×				×	×
happening						×
occurrence						×

One of the most used algorithms for constructing the concept lattice, described by Carpineto and Romano [2004] is reproduced by Algorithm 4 to provide insight in the complexity of the computations. This particular algorithm is based on a top-down iterative process. The concept lattice is built one concept at a time, by finding the neighbors in the line diagram of known concepts, starting from the concept with an empty set of attributes, c_7 in this case, and progressively adding its lower neighbors. Each edge of the line diagram of the concept lattice connects one of the c_i concepts to the concept formed by the meet of c_i with a new object $m^{'}$. The amount of time spent to traverse the entire concept lattice in this way is polynomial in the number of input objects and attributes per generated concept.

Its time complexity is $O(|G|^2|M||\mathfrak{B}(G,M,I)|)$, and its polynomial delay is $O(|G|^2|M|)$ where |G| stands for the cardinality of the set of objects G, |M|, similarly, is the number of all attributes from M and $|\mathfrak{B}(G,M,I)|$ is the size of the concept lattice.

Kuznetsov and Obiedkov [2002] surveyed many algorithms for concept lattices generation and compared their performance. The two key issues considered by all these algorithms are the generation of all the concepts and the construction of a structure, a search tree for example, that helps avoid

Algorithm 4 Finding Lower Neighbours

```
**INPUT** Context $(G, M, I)$ and concept $(X, Y)$ of this context
$lowerNeighbours := \emptyset$
$testedCandidates := \emptyset$
**for all** $m \in M \setminus Y$ **do**
    $X_1 := X \cap m^{'}$
    $Y_1 := X_1^{'}$
    **if** $(X_1, Y_1) \notin testedCandidates$ **then**
        Add $(X_1, Y_1)$ to $testedCandidates$
        $count(X_1, Y_1) := 1$
    **else**
        $count(X_1, Y_1) := count(X_1, Y_1) + 1$
    **end if**
    **if** $(|Y_1| - |Y|) = count(X_1, Y_1)$ **then**
        Add $(X_1, Y_1)$ to $lowerNeighbours$
    **end if**
**end for**
**OUTPUT** The set of lower neighbors of $(X, Y)$ in the concept lattice of $(G, M, I)$
```

repeated concept generations (computations of set closures). One of the better performing algorithms is Close-by-One. It generates concepts in the lexicographical order of their extents assuming that there is a linear order on the set of objects. At each step of the algorithm there is a current object. The generation of a concept is considered canonical if its extent contains no object preceding the current object (the canonical test). Close-by-One's use of the canonicity test allows selecting subsets of a set of objects G and an intermediate structure that helps to compute closures more efficiently using the already generated concepts. Its time complexity is $O(|G|^2|M||\mathfrak{B}(G, M, I)|)$, and its polynomial delay is $O(|G|^3|M|)$.

The In-Close algorithm by Andrews [2009b] used in the following chapters is based conceptually on Close-By-One producing fast results even on large contexts. Despite the theoretical complexity, it is often the case that the formal contexts are sparse, which will prove true in this research context.

4.2.1.1 An experiment in automatically grouping translations by their senses

A bilingual dictionary is one of the resources often used by CLIR systems to perform translation. Even without out-of-vocabulary situations, choosing an appropriate translation from a dictionary requires a strategy to determine which of the possible translations should be used. I set out to re-implement the algorithm defined by Dyvik [1994]. The initial purpose of the algorithm known as the *Semantic Mirrors Method* was the automatic derivation of thesaurus entries from a word-aligned parallel corpus. Their results pointed out that bilingual dictionaries are not a sufficient source for automatically building a thesaurus, but can definitely be used for bootstrapping the process. Using Google Translate, I derived a simple method for building bilingual formal contexts. For example, the bilingual English-French context for the word **climate** is constructed by adding the word itself to the object set of $\mathbb{K}_{Climate_{EN-FR}}$, followed by its synonyms, and other words from the back translations of the word *climate* from French-to-English, while the attributes set contains all translations of the word *climate* and of its synonyms. For each word two actions were carried out: forward translation of *climate* to French and a back-translation of all the attributes obtained in the previous step. The incidence relation corresponds to the existence of a translation in the dictionary between two words in the dictionary. This generates the formal context $\mathbb{K}_{Climate_{EN-FR}}$.

The corresponding concept lattice is described in Figure 4.3. It reflects the distinct senses of the word *climate* and a partitioning of its translations grouped by sense. The hierarchical order can be linguistically interpreted for $c_2 \leq c_3$ that in English *clime* is a hyponym of *climate* and in French *région* is a hyponym of *climat*. Note that this algorithm produces only two level lattices, which are not complete with respect to the translation of the last set of objects added from back-translation. According to Definition 23 the $\mathbb{K}_{Climate_{EN-FR}}$ is a neighborhood context for the word *climate*.

I used the same method to generate the bilingual English-German (EN-DE) formal context for *climate* and constructed the concept lattice in Figure

Table 4.2: Cross table for context $\mathbb{K}_{Climate_{EN-FR}}$ where the objects and attribtutes are determined based on Google Translate forward and back-translation in English and French starting with the word *climate*

$\mathbb{K}_{Climate_{EN-FR}}$	atmosphère	brise	ciel	climat	région	tempête	temps
air		×					
atmosphere	×						
beat							×
blue			×				
breath		×					
breeze		×					
climate				×			
clime				×	×		
days							×
eon							×
era							×
gust		×					
heaven			×				
season							×
sky			×				
tense							×
time							×
times							×
waft		×					
weather		×				×	×

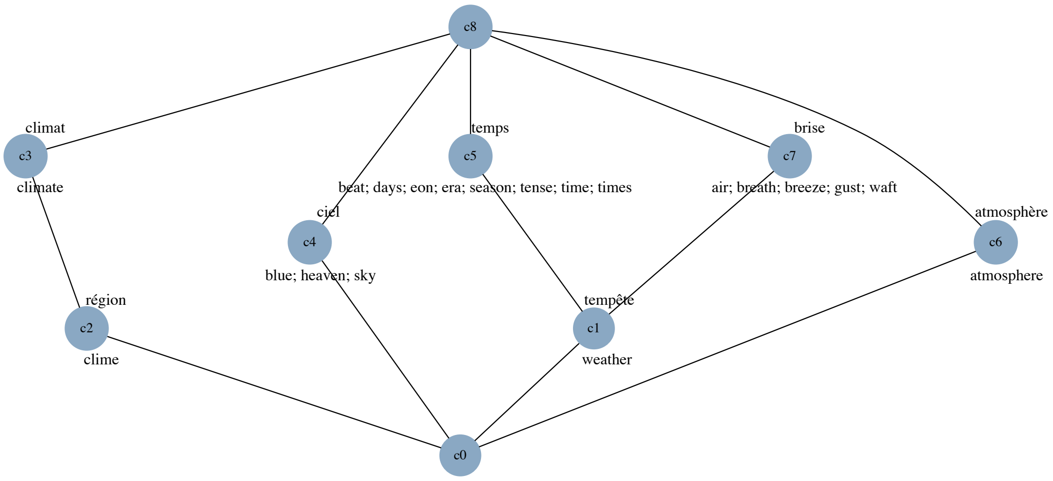


Figure 4.3: The concept lattice for $\mathbb{K}_{Climate_{EN-FR}}$

4.4. If one compares the two concept lattices for the two bilingual contexts, it is noticeable that the German back-translations for *Klima* introduces several new words to the context related to the sense of *atmosphere* of the word *climate*. In English, according to WordNet *climate* has two senses one related to *weather* and the other to *mood*. These two senses partition each concept lattice in two sub-lattices. In practical terms, if it is possible to determine correctly the sense of a word in a piece of text, then using these concept lattices, leads to translations that preserve the meaning of the original words. The difficulty arises in pinning the sense of each word in a piece of text with precision. For the IR and CLIR settings this deep level of representation for text is not scaleable, but by overlapping the concept lattices and removing the concepts from the $\mathbb{K}_{Climate_{EN-DE}}$ that do not have a correspondent in the $\mathbb{K}_{Climate_{EN-FR}}$, it could be possible to automatically construct a generic multilingual lexical database.

4.2.2 Information Retrieval

The beginning of IR as a research field stems from work carried out by Mooers. He investigated several instances of using lattices for modeling the doc-

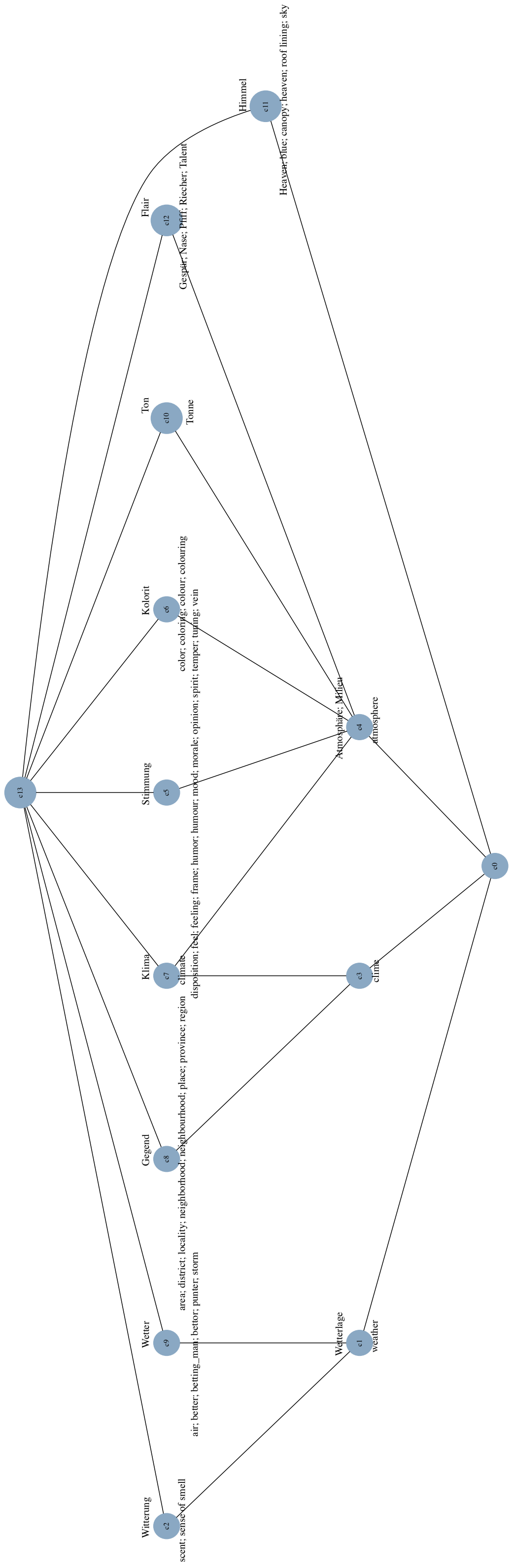

Himmel
c11
Heaven; blue; canopy; heaven; roof lining; sky
Flair
c12
Gespür; Nase; Pfiff; Riecher; Talent
Ton
c10
Tonne
Kolorit
c6
color; coloring; colour; colouring
Atmosphäre; Milieu
c4
atmosphere
Stimmung
c5
disposition; feel; feeling; frame; humor; humour; mood; morale; opinion; spirit; temper; tuning; vein
c13
c0
Klima
c7
climate
c3
clime
Gegend
c8
area; district; locality; neighborhood; neighbourhood; place; province; region
Wetter
c9
air; better; betting_man; bettor; punter; storm
Wetterlage
c1
weather
Witterung
c2
scent; sense of smell

Figure 4.4: The concept lattice for $\mathbb{K}_{Climate_{EN-DE}}$

ument collection, the query space (all possible queries that can be formed using given terms), term hierarchies, and boolean queries and documents Mooers [1958]. Table 4.3 summarizes and categorizes existing lattice-based retrieval models against the established taxonomy of IR models in Chapter 2 using as a point of reference the review carried out by Dominich [2008].

The drawback of some of the early models was computing the concept lattice from the sparse, yet very large term-document matrix. Refinements to this initial work led to more successful models such as the one in the last row of the table Abdulahhad et al. [2013].

4.2.3 Semantic Web

As mentioned in the first chapter, Tim Berners-Lee vision in 2001 was to explicitly add a machine-processable semantic layer to the existing content on the Web. After more than one decade of research channelled into defining methods and suitable modeling languages to encode and port existing knowledge into machine readable representations accessible through web standards and protocols, the research focus has been reset on building semantically-aware applications. There is though a gap between the Web of Data and its potential applications. Before one can make use of such data, an application needs to discover it and have some inbuilt strategies of exploring a particular dataset.

From the Semantic Web community, a solution to this problem was to define VoID[1] an RDF Schema vocabulary for expressing metadata about RDF datasets. It is a specification intended as a bridge between the publishers and users of RDF data, by summarizing the number of triples, the links with other datasets, and other structural metadata. At application level this does not translate in a strategy for how to query or process this data with SPARQL.

Yet FCA can help derive a concept layer connecting the data layers with the application layer of the Semantic Web Stack. The idea as described in Kirchberg et al. [2012] is to partition the data based on relations (predi-

[1] http://www.w3.org/TR/void/

Table 4.3: Overview of lattice-based IR models

Lattice-based IR	Query/Document	Relevance/Scoring	Relates To
Mooers' Model	Theoretical model that defines queries and documents as lattices.	not defined	Boolean Retrieval
FaIR	A query q is mapped to concepts within the facets of a thesaurus. Each facet is represented as a lattice conceptually complete. Each document d is assigned concepts from each facet of a thesaurus.	• *inclusive*, retrieve document d assigned the same or broader concepts to query q • *exclusive*, retrieve document d assigned exactly the same concepts to query q	Set-theoretic
Galois Concept Lattice-Based Models	The classic term-document matrix is interpreted as a formal context.	Retrieval matches the query terms specified as attributes in the context matrix extracting matching documents; ranking is given by the concept order in the term-document lattice.	Set-theoretic
BR-explorer	Extends the Galois model. A query is a set of attributes, which are added to the term-document concept lattice on the fly.	Relevant documents share at least one attribute with the query.	Set-theoretic
Rajapakse-Denham	Documents and queries have individual lattices. Atoms of the lattice are the elements consisting of objects that have identical attributes.	Relevance of a document to a query is determined on the basis of their common concepts.	Set-theoretic
Logic&Lattice Theory	A document d is a logical clause, or equivalently, a conjunction of its terms. Queries are represented in the same way. Representing a document d as a conjunction of its terms, means that: in any model of d, the terms that appear in d must be true and the other terms can be true or false.	The ranking function is a combined estimation of two measures Exhaustivity and Specificity that enable comparing the coordination level between d and q, namely what they have in common.	Probabilistic

cates) equivalently to extracting vertices from a graph that are connected by a set type of edge. Thus, for each predicate a formal context can be built where the object set and attribute set are vertices. Their experiments have shown that the scale of datasets does not pose significant problems for FCA-based algorithms as long as concept lattice computation is offline. In the next chapter, I propose to employ this techniques for analyzing KOSs relationship by relationship. I also emphasize FCA's role in establishing the operational semantics of concepts in KOSs.

4.3 Summary

Thus far, this chapter uncovered FCA's broad spectrum of applications with a long standing history of experimentation in linguistic and information retrieval. The impact of FCA in IR has been limited Valverde-Albacete and Peláez-Moreno [2013] and this has been justified by the fact that FCA was applied for simple tasks in IR. It took a number of years to overcome the idea that FCA's use in IR is purely theoretical, while in the meantime IR has developed independently as a discipline. In our view, it is also because very few approaches participated in formal IR specific evaluation campaigns to give FCA-based approaches more weight.

Despite this mixed picture I consider FCA as a suitable framework for data analysis and integration between KOSs and retrieval applications. This is demonstrated in the next chapter where based on the FCA's mathematical toolbox presented in section 4.1, the foundations of a hybrid semantic retrieval model are laid out.

Chapter 5

A New Semantic Information Retrieval Model Instance

This chapter presents an instantiation of the generic semantic information retrieval (SIR) model from Definition 10, where the framework $\mathcal{F}_{\mathbf{SM}}$ is algebraic and the semantic model is a collection of KOSs expressed as SKOS datasets.

5.1 Rationale for a new retrieval model

Semantic search in the context of the Semantic Web makes two assumptions about its information retrieval model. First, the representations of queries and documents are extended beyond term frequencies using meaning described in external linguistic and knowledge resources from its semantic model. Second, these resources are part of a network of interlinked, dynamic, and evolving set of resources within the Linguistic Linked Open Data (LLOD) cloud. The ideal retrieval model in this context should be *expressive* enough to capture the connections between documents through their annotating concepts from the semantic model and *maximize the exploitation of the semantic model at both lexical and knowledge level*.

Let us consider Figure 5.1 describing three documents d_A, d_B, and d_C annotated with concepts from the semantic model SM=$(:X, :Y, :Z, :W, :T, :V,$

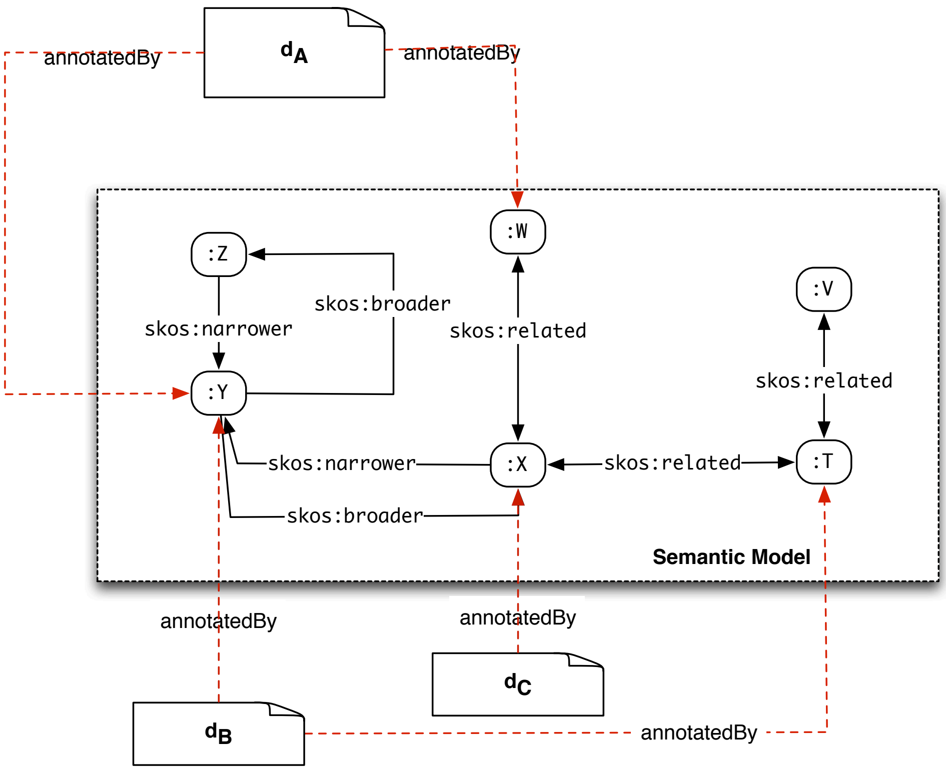


Figure 5.1: Connected documents through the Semantic Model

$skos : narrower, skos : broader, skos : related)$. The SMs considered by SIR models in general are expressed as RDF graphs. A sample RDF statement such as concept $: X \ skos : narrower : Y$. establishes the relation between the concept $: X$ and the concept $: Y$ in the SM.

In Figure 5.1 the document d_A is connected to document d_B through a shared concept $: Y$, while d_B is connected to d_C through an inferred link based on the relation between $: X$ and $: Y$. Therefore, there is a conceptual overlap between the documents parametrized by the semantic model. The question becomes how to measure the conceptual overlap and how does that relate to the bigger problem of information retrieval effectiveness.

To solve this problem the retrieval model presented in this chapter assumes as true the following hypothesis defined by Jardine and van Rijsbergen [1971]:

The cluster hypothesis: Closely associated documents tend to be relevant to the same query requests.

The *closely associated documents* part of the hypothesis is interpreted in this case as the existence of an inferred path through the semantic model between two documents. For example Table 5.1 lists existing paths of different lenghts between the sample documents.

Retrieval systems that use controlled vocabularies to enhance the query and enable an exhaustive browsing of the collection are instances of models where the path equals zero.

While the cases when the path's lenght is one have been explored in query expansion research with WordNet, where the original query is reformulated or expanded to contain weighted synonyms, hypernyms or hyponyms depending on the heuristics of the setup. The difficult aspect in these cases is to determine a good algorithm for setting the weight of the words added to the query such that with increased recall, there is no drop in precision.

For paths longer than one the weighting models make use of the hierarchical structure of the semantic model (e.g. depth of the structure or other apriori weighting).

As seen in the example in Figure 5.1, there are a number of paths between the documents: simple (following the same type or relation) or complex (a combination of relationships). Yet, for a query and a document annotated by concepts the existence of an inferred path through the semantic model under the *cluster hypothesis* its not a sufficient constraint for a match between the two. The longer the path between two concepts the further they are conceptually, therefore in the upcoming model the paths are constrained in their nature and length. The decision on which paths to consider and which lengths is dependent on the relevance to retrieval of the relations from a chosen path.

In the next Section 5.2 I describe the methodology for precomputing paths using Formal Concept Analysis to partition the SM and extract information relevant to NLP processes used by a SIR model such as query and document annotation, concept disambiguation, and translation. This

Table 5.1: Paths through the Semantic Model

Path Length	Path Example
0	d_A annotatedBy : Y, d_B annotatedBy : Y
1	d_B annotatedBy : Y, d_C annotatedBy : X, : X skos:narrower : Y
2	d_A annotatedBy : W, d_B annotatedBy : T, : W skos:related : X, : X skos:related : T

is than followed by the extraction of other sub-graphs which contain concept relational information and support the document indexing in Section 5.3 and the document selection in Section 5.3.2 during the query-document matching phase of retrieval. The reasoning behind the nature of the paths considered in this research is stated in the *retrieval relevance assumptions* throughout Section 5.3. At the end of this chapter in Section 5.4 I discuss further the commonalities and differences between this work and document clustering using knowledge bases, followed by a summary of the features of this model and its advantages.

5.2 Representational Contexts

In this model's description KOS resources are used as semantic models. With FCA as the framework for conceptual clustering, the semantic model's information is partitioned as described in this section into two groups:

a) *functional formal contexts*, where the incidence relations are I=*skos:exactMatch* or I=*skos:closeMatch*

b) a group of *relational formal contexts* where $I \in$ *{skos:broader, skos:narrower, skos:related}*

The first group captures the paths in the graph relevant for the NLP processes, while the second group captures the paths relevant for document indexing and selection during matching. All the semantic relations defined by the SM are considered and this can be viewed as a method for layering the intrepretations of a piece of text.

Overall, the relations between the concepts from the SM characterizing the documents and queries are captured using formal concepts derived from the *relational formal contexts*. This is the distinguishing characteristic between this model and the original Generalized Vector Space Model (GSVM) that was limited to capturing the dependency between the terms in the representations of queries and documents. This retrieval model builds on Miles [2006] investigation of a set-theoretic model that uses structured vocabularies expressed in SKOS. The mathematical grounding of the model to be described makes it easier to translate into a system implementation, and with the use of Formal Concept Analysis, clear strategies can be provided to integrate KOSs within the existing flow of processes of an IR system.

5.2.1 Pre-processing the Semantic Model

Hereafter, the KOSs discussed are assumed to be described using SKOS and refer to their set of concepts and relations as the semantic model SM. Each KOS can be viewed as a family of *formal contexts*, where the contexts are the triples (G, M, I) with G the set of all the concepts in the knowledge base and I a structural relation such as broader, narrower, related, exact match, etc., and M a set of attributes from the range of the incidence relation I. This approach partitions the information provided by a KOS for each of its concepts into groups: *functional formal contexts* where the incidence relations are I=*skos:exactMatch* or I=*skos:closeMatch* and a group of *relational formal contexts* where $I \in$ {*skos:broader, skos:narrower, skos:related*}. The prefix *skos* refers to all predicates defined in the SKOS Core[1] specification.

5.2.1.1 Functional formal contexts

To understand the connection between *functional formal contexts* and meaning definition, let us refer to Dahlberg's meaning triangle and its extension for the Semantic Web. Dahlberg[2]'s review of the classic meaning triangle considers that the specification of the meaning of a concept requires all

[1] http://www.w3.org/2004/02/skos/core#

[2] the founder of the International Society of Knowledge Organisation (ISKO)

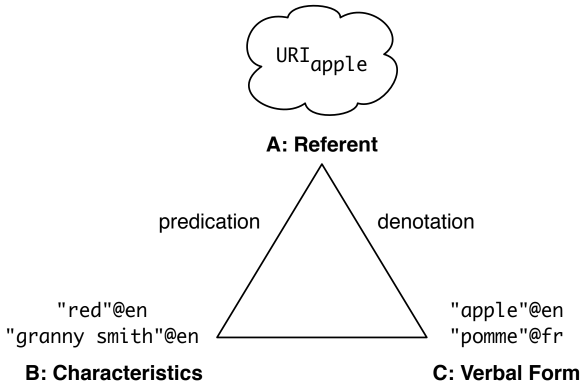


Figure 5.2: An interpretation within the Semantic Web space of Dahlerberg's meaning triangle

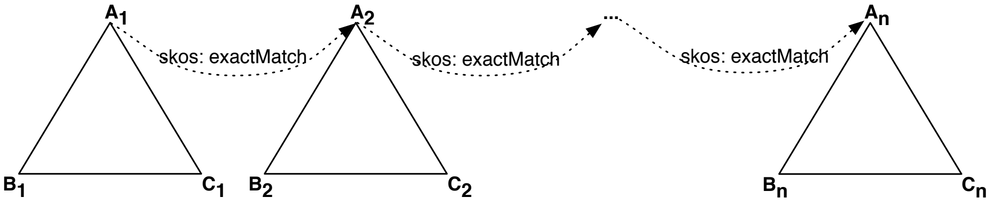


Figure 5.3: Formal concept formation from concept cross-links mappings

three elements of the triangle in Figure 5.2 with **A** the referent (an object, a property, an activity, a topic, something abstract), **B** the necessary statements describing **A**'s characteristics, and **C** the term used to verbalize **A**.

The point of departure in determining what are the knowledge elements that describe a concept also known as a knowledge unit has a long track of discussions in philosophy and knowledge representation Veltman [2006]. The structural aspects of a modern concept-based KOS presented in Chapter 3 reflect this triadic view of the meaning definition.

Thus, a SKOS specification for a concept like *apple* starts with a URI_{apple}, followed by a set of RDF statements detailing an *apple*'s formal characteristics and the different lexicalisations across languages for this concept. In SKOS the formal characteristics are presented descriptively through definitions, scope notes, etc. Yet a concepts' description does not end here, each cross-link to an equivalent or near equivalent concept like in Figure 5.3 adds more detail to the meaning of a concept. An *accurate interpre-*

Table 5.2: Cross table for context $\mathbb{K}_{skos:exactMatch}$ with G=M={GEMET SKOS concepts} and I=*skos:exactMatch*

$\mathbb{K}_{skos:exactMatch}$	gemet:1471	agrovoc:c_1666	eurovoc:5482	...
gemet:1471	×	×	×	
agrovoc:c_1666	×	×	×	
eurovoc:5482	×	×	×	
...				

tation has to incorporate all these details and it can be achieved by constructing a functional formal context where $G = M$ and is the union set of all concepts from the chosen KOS and of all the other concepts from KOSs for which cross schema equivalence or near-equivalence links exist (*skos:exactMatch*, *skos:closeMatch*, or *skos:relatedMatch*). Either of these can be used as the incidence relation for the formal context.

This formalization leads to the formation of formal concepts that cluster all URIs referring to one *knowledge unit*. An example of the outcome based on Figure 3.3 description of skos:exactMatch relations of the concept *gemet:1471* for *climatic change* is listed in Table 5.2, where the objects and attributes correspond to the entire set of concepts in GEMET. By building the concept lattice for this context a natural grouping of all the concepts linked by skos:exactMatch is obtained.

Note that given the axiom description of SKOS predicates in Appendix C, specifically axioms S39 through S45 the skos:relatedMatch, skos:closeMatch and skos:exactMatch are each instances of the owl:SymmetricProperty, hence the corresponding formal context is a symmetric matrix.

Operational role of functional formal contexts in IR applications

For a retrieval application, the process of building *functional formal contexts* has direct application in constructing a *concept's signature* (see Definition 16 in Section 3.4). The sum of all the descriptive elements available for the SKOS concepts in the formal concepts of this type of context constitutes a concept's signature.

In practice this entails writing SPARQL queries to query the KOSs endpoints. It is important not to have a prescriptive approach when writing

SPARQL queries that extract portions from an RDF graph, but focus on identifying the existing predicates that can be considered equivalent in meaning to the SKOS ones (e.g. skos:definition vs. dbpedia-owl:abstract). These issues appear since publishers of data are provided with a set of guidelines on how to port existing KOSs to SKOS and dataset specific constructs are often used.

Once having created a *concept's signature* connections between text and that respective concept can be established. This is achieved by exact or fuzzy concept label matching against the text i.e. *explicit semantic annotations* or by some other methods that perform topic matching against the text i.e. *implicit semantic annotations*. In the first case annotation errors can occur due to homonymy and polysemy. A concept in a KOS is bound to a single meaning, therefore for any application to use these annotations reliably, disambiguation is paramount.

In my view, disambiguation methods and implicit semantic annotations methods depend on the *concept's signature*. One method I have employed for both situations is to create a retrieval setting where an index is created over the collection of concepts' signatures. The result set of a search task using this index is a set of concepts. For disambiguation, the annotated text is submitted as a query to this special index. Heuristically, if the similarity between the annotated text and the annotating concepts's signature is higher than a set threshold the annotations are kept. For *implicit annotations*, a decision on how many of the concepts retrieved should be considered as enrichments of the query is again, a heuristic decision.

5.2.1.2 Relational formal contexts

Using three different knowledge resources: DBpedia Categories, GEMET, and TheSoz three full size relational contexts were obtained as described by Table 5.3 and Table 5.4. This shows that FCA can scale to handle even larger contexts like DBpedia Categories with a total of 26998 triples for related, while GEMET has 5191 triples for broader/narrower, and 2088 for related. TheSoz 13706 broader/narrower triples and 3738 for related.

Table 5.3: SKOS datasets as Formal Contexts

$\mathbb{K}$	relation	# of objects	# of attributes	# of concepts
$\mathbb{K}_{DBpediaCategories}$	$skos:related$	17689	17600	49727
$\mathbb{K}_{GEMET}$	$skos:narrower$	1437	5099	1492
$\mathbb{K}_{GEMET}$	$skos:related$	1428	1428	1182
$\mathbb{K}_{TheSoZ}$	$skos:narrower$	1184	8383	2975
$\mathbb{K}_{TheSoZ}$	$skos:related$	2379	2379	2492

Table 5.4: Associative and Hierarchical Branching

$\mathbb{K}$	relation	avg. # of objects	avg. # of attributes
$\mathbb{K}_{DBpedia}$	$skos:related$	6.3032	3.3035
$\mathbb{K}_{GEMET}$	$skos:narrower$	1.0517	3.5248
$\mathbb{K}_{GEMET}$	$skos:related$	1.5559	1.5533
$\mathbb{K}_{TheSoZ}$	$skos:narrower$	5.9216	2.1015
$\mathbb{K}_{TheSoZ}$	$skos:related$	1.5659	1.5646

Operational role of relational formal contexts in IR applications Another problem with semantic annotations at topical level is granularity. Even if the chosen general topic is correct, let us assume it to be *climate*, the document itself could be talking more about *climate change*, yet this is not manifested at lexical level clearly. This issue can be handled by mapping documents to formal contexts from *relational formal contexts* where $I \in$ *{skos:broader, skos:narrower, skos:related}*. This may seem as a complicated approach, but in effect it allows precomputing for each document its conceptual neighborhood formed as described in Section 5.3. These neighborhoods capture all immediate pathways of exploration through the KOS. Search applications that employ KOSs often provide interfaces that prompt users to explore other documents associated to a concept's generalizations (upper neighbors), specializations (lower neighbors) and categorizations (siblings). The potential of user explorations is captured by these neighborhoods, as well as the fact that a document usually covers several topics. The model proposed considers that a document is characterized by the overlap of several conceptual neighborhoods.

In summary thus far by partioning the semantic model based on information from the terminological level (the *functional formal contexts*) and

Table 5.5: Term-Document Matrix

	a	b	c	d	e	f	g	h	i	j	k	l	m	n
d_1	4	2	2	2	0	0	0	0	0	0	0	0	0	0
d_2	2	0	0	2	1	1	2	2	1	0	0	0	0	0
d_3	0	3	0	2	0	0	2	1	0	1	1	0	0	0
d_4	0	3	0	2	0	0	0	0	0	0	0	3	2	2

the conceptual level (the *relational formal contexts*) this SIR model is set up to *maximize the exploitation of the semantic model*.

5.3 Indexing, Matching, and Ranking

Indexing in classic IR is the key process in setting up an IR system. It requires deciding on what the indexing unit is: a word, a phrase, or a block of text of a certain size. In this case, both terms and concepts from the semantic model are considered. The process of building a term-based retrieval index is known and this chapter's contribution is on creating a new method for building a semantic index where the indexing units are formal concepts. The mathematical formulas will show the natural connection with the general vector space model. Also, the process of index expansion is driven by a series of Retrieval Relevance Assumptions explained in Section 5.3.1.

The classic retrieval index is built from a compressed version of the term-document matrix. Let us consider $D = \{d_1, d_2, d_3, d_4\}$ a collection of documents and {*a, b, c, d, e, f, g, h, i, j, k, l, m, n*} all the terms contained by the documents with Table 5.5 indicating the term frequencies per document.

The Semantic Model Let us consider the following KOS described using SKOS:

```
:T rdf:type skos:Concept ;
   skos:related :X ;
   skos:related :V .
:X rdf:type skos:Concept ;
   skos:narrower :Y ;
   skos:related :T ;
   skos:related :W .
:Y rdf:type skos:Concept ;
   skos:broader :X ;
   skos:broader :Z .
```

```
:Z rdf:type skos:Concept ;
   skos:narrower :Y .
:V rdf:type skos:Concept ;
   skos:related :T .
:W rdf:type skos:Concept ;
   skos:related :X .
```

The Semantic Annotations express links between concepts from the semantic model and the document collections. These links are established a-priori.

d_1 `annotatedBy :X .`
d_2 `annotatedBy :Y .`
d_2 `annotatedBy :T .`
d_3 `annotatedBy :Z .`
d_4 `annotatedBy :W .`

The same information about the documents can be encoded separately using the Open Annotation Core Model [1] which allows a more detailed specification of the actual annotation going beyond establishing a link between a document and a concept. If any two documents are annotated with the same concept, it is possible to specify multiple targets for an annotation and construct a specification that naturally groups documents based on their common concepts.

[1] http://www.openannotation.org/spec/core/core.html#

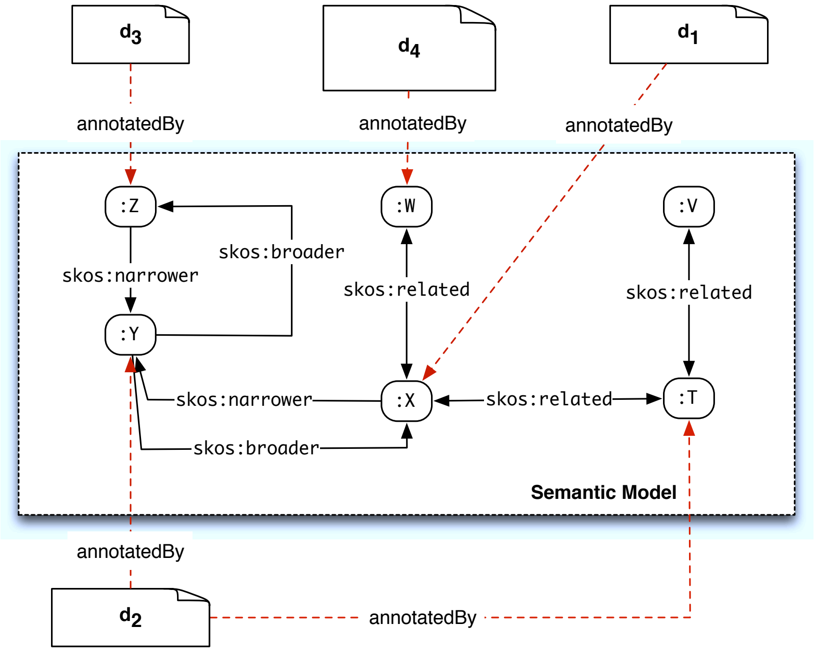


Figure 5.4: Overview of the connections between the sample documents collection and the semantic model

```
<annotation1> a oa:Annotation;
        oa:motivatedBy oa:tagging;
        oa:hasBody :X;
        oa:hasTarget <d1> .
<annotation2> a oa:Annotation;
        oa:motivatedBy oa:tagging;
        oa:hasBody :Y;
        oa:hasTarget <d2> .
<annotation3> a oa:Annotation;
        oa:motivatedBy oa:tagging;
        oa:hasBody :T;
        oa:hasTarget <d2> .
<annotation4> a oa:Annotation;
        oa:motivatedBy oa:tagging;
        oa:hasBody :Z;
        oa:hasTarget <d3> .
<annotation5> a oa:Annotation;
        oa:motivatedBy oa:tagging;
        oa:hasBody :W;
        oa:hasTarget <d4> .
:X a oa:SemanticTag .
:Y a oa:SemanticTag .
:Z a oa:SemanticTag .
:T a oa:SemanticTag .
```

All the N-triples[1] are summarized diagrammatically by Figure 5.4

[1] http://www.w3.org/2001/sw/RDFCore/ntriples/

Pre-indexing conceptual clustering of the semantic model Based on the relations in the semantic model the formal contexts $\mathbb{K}_{skos:broader}$, $\mathbb{K}_{skos:narrower}$, and $\mathbb{K}_{skos:related}$ are built in Table 5.6, Table 5.7 and Table 5.8. These formal contexts preserve the properties of the relations in the semantic model. For example, the property that skos:broader and skos:narrower are inverse of each other is observed by comparing the transpose incidence relation matrix of one context to the matrix of the other. They are equal. While the skos:related is symmetric, the SKOS Reference [2009] specification does not state that skos:related is a reflexive property, neither does it state that skos:related is an irreflexive property. Therefore, in this research X skos:related X is allowed and added to the context. The formal concepts of this context are therefore symmetric and determine closed groups of related concepts. Based on this representation of the original KOS, the formal contexts are consistent with the SKOS data model.

Table 5.6: Cross table for $\mathbb{K}_{skos:broader}$

$\mathbb{K}_{skos:broader}$	X	Y	Z
X			
Y	×		×
Z			

Table 5.7: Cross table for $\mathbb{K}_{skos:narrower}$

$\mathbb{K}_{skos:narrower}$	X	Y	Z
X		×	
Y			
Z		×	

Table 5.8: Cross table for $\mathbb{K}_{skos:related}$

$\mathbb{K}_{skos:related}$	X	T	V	W
X	×	×		×
T	×	×	×	
V		×	×	
W	×			×

This phase of partitioning the semantic model is followed by the computation of their corresponding concept lattices using an existing fast algorithm like In-Close Andrews [2009a]. The computed formal concepts in these lattices are used in the next sections to map documents and queries to formal concepts.

The computed formal concepts derived from the formal contexts of the SM are:

$$\mathfrak{B}(G, M, I)_{skos:broader} = \{(\mathrm{X,Y,Z};\emptyset),(\mathrm{Y;X,Z}),(\emptyset;\mathrm{X,Y,Z})\}$$

$\mathfrak{B}(G, M, I)_{skos:narrower}$={(∅;X,Y,Z),(X,Z;Y),(X,Y,Z;∅)}

$\mathfrak{B}(G, M, I)_{skos:related}$={(∅;T,V,X,W),(X;T,X,W), (T,X,W;X),(T;T,X,V), (T,X;T,X),(T,V;T,V),(T,X,V;T),(X,W;X,W),(T,V,X,W;∅)}

5.3.1 Formal concept indexing of the document collection

The documents, annotations, and computed formal concepts are the input for this process in the semantic IR model, where each document becomes a linear combination of formal concepts.

Two previously defined functions are reintroduced γ, the object concept mapping, and μ, the attribute concept mapping from the previous chapter:

$\gamma : G \longrightarrow \mathfrak{B}(G, M, I)$ with γg=$\big(\{g\}'', \{g\}'\big)$ is the object to formal concept mapping

$\mu : M \longrightarrow \mathfrak{B}(G, M, I)$ with μm=$\big(\{m\}', \{m\}''\big)$ is the attribute to formal concept concept mapping

These functions support *mapping documents to core formal concepts*, followed by *mapping to formal concepts from the conceptual neighborhoods.* In the first mapping step a document's representation incorporates all other concepts one link away in the semantic model from the original concepts that annotate the document, while in the second step this is extended to paths longer than one.

5.3.1.1 Mapping documents to core formal concepts

Let us define $\mathbb{K}_{SM}$ as the disjoint union of all formal contexts used in enhancing the documents vectorial representation:

$$\mathbb{K}_{SM} = \mathbb{K}_{skos:broader} \cup \mathbb{K}_{skos:narrower} \cup \mathbb{K}_{skos:related} \cup ... \tag{5.1}$$

The process of extending the documents representations starts by building a core of formal concepts for each document with the $\{c_1, c_2, ..c_k\} \subseteq KOS$ that annotate them. The core of formal concepts captures all immediate links of the concepts that annotate a document. Thus, allowing documents that share a path of length one to be explicitely grouped together. In the implementation of this model it can be decided on a case-by-case basis which of the semantic relations need to be included. For example, if the annotating concepts denote topic-level information about the whole document by using the $\mathbb{K}_{skos:narrower}$ a document is automatically enhanced with more specific concepts. This can be viewed as a way of balancing the semantic annotations with both general and specific concepts.

Therefore, $\mathbb{K}_{SM}$ in the theoretical description is the union of all formal contexts, but in practice the formal contexts from the union need to be chosen based on the granularity level of the annotation.

$$d \xrightarrow[\mathbb{K}_{SM}]{c_1,c_2,..c_k} \bigcup \gamma(c_i) \cup \mu(c_i) \tag{5.2}$$

Retrieval Relevance Assumption For a query annotated with a concept X from the semantic model, all documents annotated with X and its semantically related concepts are considered relevant. If the query is matched to several concepts, the relevant documents set is the union of all documents sets obtained for each concept.

In the following operations the functions γ and μ have the range defined by the mapping's subscript specification of the context (e.g. $\mathbb{K}_{skos:broader}$). For simplicity : X's prefix indicator is removed and X is used instead.

$$d_1 \xrightarrow[\mathbb{K}_{skos:broader}]{X} \gamma(X) \cup \mu(X) = (X, Z; Y) \tag{5.3}$$

$$d_1 \xrightarrow[\mathbb{K}_{skos:narrower}]{X} \gamma(X) \cup \mu(X) = (Y; X, Z) \tag{5.4}$$

$$d_1 \xrightarrow[\mathbb{K}_{skos:related}]{X} \gamma(X) \cup \mu(X) = (X; T, X, W), (T, X, W; X) \tag{5.5}$$

Since $T \notin \mathbb{K}_{skos:broader}$ or $\mathbb{K}_{skos:narrower}$ the document d_2 is not mapped to a formal concept in this instance. For consistency, we write:

$$d_2 \xrightarrow[\mathbb{K}_{skos:broader}]{T} \gamma(T) \cup \mu(T) = \emptyset \tag{5.6}$$

$$d_2 \xrightarrow[\mathbb{K}_{skos:narrower}]{T} \gamma(T) \cup \mu(T) = \emptyset \tag{5.7}$$

$$d_2 \xrightarrow[\mathbb{K}_{skos:related}]{T} \gamma(T) \cup \mu(T) = (T; T, X, V), (T, X, V; T) \tag{5.8}$$

$$d_2 \xrightarrow[\mathbb{K}_{skos:broader}]{Y} \gamma(Y) \cup \mu(Y) = (X, Z; Y) \tag{5.9}$$

$$d_2 \xrightarrow[\mathbb{K}_{skos:narrower}]{Y} \gamma(Y) \cup \mu(Y) = (Y; X, Y) \tag{5.10}$$

Since $Y \notin \mathbb{K}_{skos:related}$ the document d_2 is not mapped to a formal concept in this instance.

$$d_2 \xrightarrow[\mathbb{K}_{skos:related}]{Y} \gamma(Y) \cup \mu(Y) = \emptyset \tag{5.11}$$

$$d_3 \xrightarrow[\mathbb{K}_{skos:broader}]{Z} \gamma(Z) \cup \mu(Z) = (X, Z; Y) \tag{5.12}$$

$$d_3 \xrightarrow[\mathbb{K}_{skos:narrower}]{Z} \gamma(Z) \cup \mu(Z) = (Y; X, Z) \tag{5.13}$$

$$d_3 \xrightarrow[\mathbb{K}_{skos:related}]{Z} \gamma(Z) \cup \mu(Z) = \emptyset \tag{5.14}$$

$$d_4 \xrightarrow[\mathbb{K}_{skos:broader}]{W} \gamma(W) \cup \mu(W) = \emptyset \tag{5.15}$$

$$d_4 \xrightarrow[\mathbb{K}_{skos:narrower}]{W} \gamma(W) \cup \mu(W) = \emptyset \tag{5.16}$$

$$d_4 \xrightarrow[\mathbb{K}_{skos:related}]{W} \gamma(W) \cup \mu(W) = (X, W; X, W) \tag{5.17}$$

Considering all the relationships these documents have with concepts from the semantic model, the process above allowed to determine for each document a set of formal concepts that incorporate the immediate connected concepts in the semantic model.

The output of this phase is:

$$d_1 \xrightarrow[\mathbb{K}_{SM}]{X} \{(X, Z; Y), (Y; X, Z), (X; T, X, W), (T, X, W; X)\} \tag{5.18}$$

$$d_2 \xrightarrow[\mathbb{K}_{SM}]{Y,T} \{(X, Z; Y), (Y; X, Z), (T; T, X, V), (T, X, V; T)\} \tag{5.19}$$

$$d_3 \xrightarrow[\mathbb{K}_{SM}]{Z} \{(X, Z; Y), (Y; X, Z)\} \tag{5.20}$$

$$d_4 \xrightarrow[\mathbb{K}_{SM}]{W} \{(X, W; X, W)\} \tag{5.21}$$

5.3.1.2 Mapping documents to formal concepts from the conceptual neighborhood

The next step is to determine for each generated formal concept other neighboring concepts and add them to the descriptions of each document. The idea behind extending a document's mapping is to capture the possibilities of browsing the information space by following semantic relations between concepts in the semantic model one link forward than in the previous section. A concept's neighborhood based on Definition 23 can be built by starting with one concept and then retrieving other items the first item is related to and repeat.

2-1-neighborhood contexts are built using the following steps in dual

Algorithms 5 and 6. These contexts' corresponding concept lattices are computed afterwards and all these concepts are added to the description of the document. Also, these concepts are concepts in the larger lattice $\underline{\mathfrak{B}}(G, M, I)$.

Retrieval Relevance Assumption For a query annotated with a concept X from the semantic model, all documents annotated with X and its semantically related concepts found by advancing through the semantic model at depth two are considered relevant. If the query is matched to several concepts, the relevant documents set is the union of all documents set obtained for each concept.

Algorithm 5 Constructing a 2-1 Lower Neighborhood

```
INPUT Context $(G, M, I)$ and $g \in G$
$G_{neighbors}$ := $\emptyset$
$M_{neighbors}$ := $\emptyset$
for all $m \in M \ \ where \ \ gIm$ do
    Add $m$ to $M_{neighbors}$
end for
for all $m \in M_{neighbors}$ do
    Add $Ext(\mu m)$ to $G_{neighbors}$
    Add $Int(\mu m)$ to $M_{neighbors}$
end for
OUTPUT The $\mathbb{K}_{g:neighborhood}$ derived from $(G, M, I)$ for object $g$
```

For example, for the concept X linked to d_1 its formal concept is $(X; T, X, W)$. Computing $\mu T, \mu X, \mu W$ generates the $\mathbb{K}_{X:LowerNeighborhood}$ = $(T, X, W, V, T, X, W, skos : related)$ with the concepts {(X;T,X,W), (T,X,W;X),(T;T,X,V),(T,X;T,X),(T,X,V;T), (X,W;X,W)}. In this case the upper and lower neighborhood contexts are the same. Note that a formal concept that contains X, the initial annotation concept, in either its object set or the attribute set is part of its 2-1 neighborhoods. Also, the neighborhood context will include the core formal concepts.

Algorithm 6 Constructing a 2-1 Upper Neighborhood

```
INPUT Context $(G, M, I)$ and $m \in M$
$G_{neighbors}$ := $\emptyset$
$M_{neighbors}$ := $\emptyset$
for all $g \in G \;\; where \;\; gIm$ do
    Add $g$ to $G_{neighbors}$
end for
for all $g \in G_{neighbors}$ do
    Add $Ext(\gamma g)$ to $G_{neighbors}$
    Add $Int(\gamma g)$ to $M_{neighbors}$
end for
OUTPUT The $\mathbb{K}_{m:neighborhood}$ derived from $(G, M, I)$ for attribute $m$
```

$$d_1 \xrightarrow[\mathbb{K}_{SM}]{X} \{(X, Z; Y), (Y; X, Z), (X; T, X, W), (T, X, W; X), (T; T, X, V), (T, X; T, X), (T, X, V; T), (X, W; X, W)\} \quad (5.22)$$

$$d_2 \xrightarrow[\mathbb{K}_{SM}]{Y,T} \{(X, Z; Y), (Y; X, Z), (T; T, X, V), (T, X, V; T), (X; T, X, W), (T, X, W; X), (V, T; V, T)\} \quad (5.23)$$

$$d_3 \xrightarrow[\mathbb{K}_{SM}]{Z} \{(X, Z; Y), (Y; X, Z)\} \quad (5.24)$$

$$d_4 \xrightarrow[\mathbb{K}_{SM}]{W} \{(X, W; X, W)\} \quad (5.25)$$

The algebraic equivalent of these operations can be obtained by considering a matrix M^{SM}, where each row of the matrix represents a formal concept from the different formal contexts generated from the SM. The rows represent formal concepts that are derived using the *object or attribute to* *formal concept mapping* from the c_i concepts from the SM.

$$m_{ij} = \begin{cases} 1 & \text{if } c_j \text{ belongs to the extension or the intension of} \\ & \text{the formal concept representing row i} \\ 0 & \text{otherwise} \end{cases} \tag{5.26}$$

In this case:

$$M^{SM} = \left(\begin{array}{lcccccc} & T & X & Y & Z & V & W \\ (Y;X,Z) & 0 & 1 & 1 & 1 & 0 & 0 \\ (X,Z;Y) & 0 & 1 & 1 & 1 & 0 & 0 \\ (X;T,X,W) & 1 & 1 & 0 & 0 & 0 & 1 \\ (T,X,W;X) & 1 & 1 & 0 & 0 & 0 & 1 \\ (X,W;X,W) & 0 & 1 & 0 & 0 & 0 & 1 \\ (T;T,X,V) & 1 & 1 & 0 & 0 & 1 & 0 \\ (T,X,V;T) & 1 & 1 & 0 & 0 & 1 & 0 \\ (T,X;T,X) & 1 & 1 & 0 & 0 & 0 & 0 \\ (T,V;T,V) & 1 & 0 & 0 & 0 & 1 & 0 \end{array} \right) \tag{5.27}$$

In the definition of M^{SM}, sibling and directly connected concepts are captured by one row. It also introduces redundancy, which reflects that specifying both broader and narrower in a semantic model is not necessary, as long as the model, SKOS in this case, defines them as the inverse of each other.

For transforming a document from a concept-based representation to a formal concept-based one, the following equation is introduced:

$$\vec{d}^{SM} = M^{SM}\vec{d} \tag{5.28}$$

Based on this transformation only rows for formal concepts with at least one concept in its intension or extension annotating document d will be different from zero. These are excluded from computation by using ${M_d}^{SM}$ to indicate a reduced matrix. For example, for d_1 the matrix $M_{d_1}^{SM}$ has only eight rows after removing the row for $(T,V;T,V)$. In the final vector the component for $(T, V; T, V)$ will be zero.

The cf_i^c denotes the frequency of concept $c \in SM$ in document d_i.

$$\begin{aligned}
\vec{d_1}^{SM} &= M_{d_1}{}^{SM} \times (cf_1^T, cf_1^X, cf_1^Y, cf_1^Z, cf_1^V, cf_1^W)^\top \\
&= M_{d_1}{}^{SM} \times \begin{pmatrix} T & X & Y & Z & V & W \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}^\top \\
&= \begin{pmatrix} (Y; X, Z) & 1 \\ (X, Z; Y) & 1 \\ (X; T, X, W) & 1 \\ (T, X, W; X) & 1 \\ (X, W; X, W) & 1 \\ (T; T, X, V) & 1 \\ (T, X, V; T) & 1 \\ (T, X; T, X) & 1 \end{pmatrix}
\end{aligned} \tag{5.29}$$

Doing the computation for the non-zero components gives the following results for the other documents:

$$\begin{aligned}
\vec{d_2}^{SM} &= M_{d_2}{}^{SM} \times (cf_2^T, cf_2^X, cf_2^Y, cf_2^Z, cf_2^V, cf_2^W)^\top \\
&= M_{d_2}{}^{SM} \times \begin{pmatrix} T & X & Y & Z & V & W \\ 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}^\top \\
&= \begin{pmatrix} (Y; X, Z) & 1 \\ (X, Z; Y) & 1 \\ (T; T, X, V) & 1 \\ (T, X, V; T) & 1 \\ (X; T, X, W) & 1 \\ (T, X, W; X) & 1 \\ (V, T; V, T) & 1 \end{pmatrix}
\end{aligned} \tag{5.30}$$

$$\begin{aligned}\vec{d_3}^{SM} &= M_{d_3}{}^{SM} \times (cf_3^T, cf_3^X, cf_3^Y, cf_3^Z, cf_3^V, cf_3^W)^\top \\ &= M_{d_3}{}^{SM} \times \begin{pmatrix} T & X & Y & Z & V & W \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}^\top \\ &= \begin{pmatrix} (Y;X,Z) & (X,Z;Y) \\ 1 & 1 \end{pmatrix}^\top \end{aligned} \tag{5.31}$$

$$\begin{aligned}\vec{d_4}^{SM} &= M_{d_4}{}^{SM} \times (cf_4^T, cf_4^X, cf_4^Y, cf_4^Z, cf_4^V, cf_4^W)^\top \\ &= M_{d_4}{}^{SM} \times \begin{pmatrix} T & X & Y & Z & V & W \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}^\top \\ &= \begin{pmatrix} (X,W;X,W) \\ 1 \end{pmatrix} \end{aligned} \tag{5.32}$$

5.3.2 Matching a query to a document

Let $\vec{q} = (tf_1, tf_2, tf_3, ..., tf_n)$ where tf_k denotes the term frequency within the query of term t_k and $\vec{q} = (cf_q^{c_1}, cf_q^{c_2}, ..., cf_q^{c_m})$, where $cf_q^{c_j}$ denotes the frequency of concept $c_j \in SM$ for query q.

Let q_* be the following sequence $\{a, b, d, n\}$. Considering only the columns corresponding to $a, b, d,$ and n in the term-document matrix in Table 5.5, then the retrieval of documents uses the following reduced dimension transposed document vectors considering : $\vec{d_1} = (\frac{4}{6}, \frac{2}{8}, \frac{2}{8}, 0)$, $\vec{d_2} = (\frac{2}{6}, 0, \frac{2}{8}, 0)$, $\vec{d_3} = (0, \frac{3}{8}, \frac{2}{8}, 0)$, and $\vec{d_4} = (0, \frac{3}{8}, \frac{2}{8}, \frac{2}{2})$. Each component is the frequency of the index terms in the documents divided by their total frequency in the collection.

The retrieval status value [Peters et al., 2012, p.23] defined as $\mathcal{R}(q, d) = (\vec{d})^\top \cdot \vec{q}$ is:

$$\mathcal{R}(q_*, d_1) = 1 * \frac{4}{6} + 1 * \frac{2}{8} + 1 * \frac{2}{8} = \frac{7}{6} \tag{5.33}$$

$$\mathcal{R}(q_*, d_2) = 1 * \frac{2}{6} + 1 * \frac{2}{8} = \frac{7}{12} \tag{5.34}$$

$$\mathcal{R}(q_*, d_3) = 1 * \frac{3}{8} + 1 * \frac{2}{8} = \frac{5}{8} \tag{5.35}$$

$$\mathcal{R}(q_*, d_4) = 1 * \frac{3}{8} + 1 * \frac{2}{8} + 1 * \frac{2}{2} = \frac{13}{8} \tag{5.36}$$

$$\tag{5.37}$$

Thus, the ranking based on this metric is d_4, d_1, d_3, d_2.

Without any given concepts from the SM to characterize the query, the first two document d_4, d_1 are assumed relevant and the concepts that characterize them W and X are employed as annotations for the query. If the documents share any concepts the corresponding concept frequency is a count of how often a certain concept appears in the description of the pseudo-relevant documents. This method does not insure perfect annotation of the query and can be replaced with other methods depending on the application context.

In this model the query is treated similarly to a document.

$$\begin{aligned}\vec{q_*}^{SM} &= M_{q_*}{}^{SM}\vec{q_*} \\ &= M_{q_*}{}^{SM} \times \begin{pmatrix} T & X & Y & Z & V & W \\ 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}^{\top} \\ &= \begin{pmatrix} (Y; X, Z) & 1 \\ (X, Z; Y) & 1 \\ (X; T, X, W) & 2 \\ (T, X, W; X) & 2 \\ (X, W; X, W) & 2 \\ (T; T, X, V) & 1 \\ (T, X, V; T) & 1 \end{pmatrix}\end{aligned} \tag{5.38}$$

After these computations the corresponding vectors for the documents and query have different dimensions. Using the inclusion mapping, each of the vectors obtained are projected to the

$\mathbb{R}_{|\mathfrak{B}(G,M,I)_{skos:broader}|+|\mathfrak{B}(G,M,I)_{skos:narrower}|+|\mathfrak{B}(G,M,I)_{skos:related}|}$, where a component is different from zero if there exists a non-zero value in the vectors computed above.

5.3.3 Ranking a mixed metric

The retrieval status value for the semantic retrieval model

$$\mathcal{R}({q_*}^{SM}, d^{SM}) = (\vec{d^{SM}})^\top \cdot \vec{q_*}^{SM} = d^\top \cdot {M^{SM}}^\top \cdot M^{SM} \cdot q_*, \tag{5.39}$$

with $\mathcal{R}({q_*}^{SM}, {d_1}^{SM})$ = 10, $\mathcal{R}({q_*}^{SM}, {d_2}^{SM})$ = 8, $\mathcal{R}({q_*}^{SM}, {d_3}^{SM})$ = 2,
$\mathcal{R}({q_*}^{SM}, {d_4}^{SM})$ = 2

Therefore, the final ranking according to this metric is d_1,d_2, followed by d_3 and d_4 with the same rank. Considering the query is annotated by X and W and based on the semantic references in 5.4 this reflects that d_1 is easily found via highly connected concept X or W and that d_2 has more connections (paths) to the query concepts than d_4 via Y and T, which are connected to X and W by $skos : broader$, respectively $skos : related$ relation. Thus, this example underlines that the proposed semantic retrieval model boosts the ranking of documents with a higher conceptual overlap.

So far, the formal procedure of expanding the initial concepts annotating the document collection into sets of formal concepts determined using FCA operators were described. For the user, these enrichments enable selecting documents at retrieval time based on the structure of the conceptual information space.

Retrieval Relevance Assumption For a query annotated with concepts $X, Y, Z, ...$ from the semantic model a document is considered as relevant based on a numerical function that determines the probability of a document described by core and neighboring formal concepts to be close to the focus of the query. This is a reinterpretation of Miles [2006] *quantified assumption of relative relevance*. The metric complying with this assumption is described by Equation 5.42.

It is possible to consider a weighting scheme based on TF-IDF where the weight per document of a formal concept is

$$fcw_d = (1 + \log(n(fc, d)) * \log(1 + N/n(d, fc)) \tag{5.40}$$

with $n(fc, d)$ the number of formal concepts for a document (an indication how discriminatory a fc is for that collection), $n(d, fc)$ the number of documents indexed with the formal concept whose weight is being calculated, and N the size of the collection of documents. This weight is sensitive to details in the structure of the index and of the semantic model. There is no differentiation between formal concepts weights from the core set or from the neighborhood set.

The weight per query of a formal concept is

$$fcw_q = (1 + \log(n(fc, q))) * \log(1 + N/n(d, fc)) \tag{5.41}$$

with $n(fc, q)$ the number of occurrences in the query of the selected formal concept.

As an effect of using FCA, the result set for a given query are clusters of documents that have the same ranking within the cluster. In retrieval terms this leads to higher recall, but a serious drop in precision. This problem is handled by defining a mixed ranking function:

$$\mathcal{R}(q^{SM}, d^{SM}) = \alpha * (\vec{d}^{SM})^\top * \vec{q}^{SM} + (1 - \alpha) * \mathcal{R}(q, d) \tag{5.42}$$

where parameter α is a number in the interval $[0, 1]$.

In the case when several KOSs are used for annotation, it would be possible to encode a weight bias towards a certain KOS. For example, when a domain specific KOS resource is combined with a general knowledge one. This way documents with semantic enrichments from a particular domain are extracted.

5.4 Discussion

The Definition 8 of the Generalized Vector Space Model the matrix G is the key element in the representation of queries and documents and of the dependencies between terms. In the SIR model described in this chapter the $G^{SM} = M^{SM}(M^{SM})^{\top}$ matrix describes the dependecies between concepts based on the union of formal contexts $\mathbb{K}_{SM}$. Therefore, the model fullfils the *expressiveness* requirement of capturing the connections between documents through their annotating concepts from the semantic model. Considering all possible paths between documents is computationally expensive, but more importantly in the case of KOSs as the semantic model this does not make sense. It would lead to documents' representations that are too broad and too specific at the same time, introducing too much noise during search.

Compared to other models such as the Topic Vector Space Model (TVSM) the topics are the connecting elements between documents, with topics being derived from the document collection itself not an external resource. A closer model is proposed by Tsatsaronis and Panagiotopoulou [2009] where the matrix G is created based on the semantic relatedness of the terms described by WordNet. The measure introduced uses both the path length between concepts and an apriori defined weighting scheme for the path edges. The model considers G as a dependecy matrix between the terms of the documents, while in the SIR model built here G^{SM} is a dependecy matrix only for the annotating concepts.

In the beginning of this chapter, the *cluster hypothesis* is reinterpreted in this context as the existence of an inferred path through the semantic model between two documents, where the *retrieval relevance assumptions* set the constraints regarding the nature and length of these paths. In the text clustering research by Hotho et al. [2003] a similar idea is explored, where the documents vectorial representations is extended using lexical entries from the WordNet by mapping terms to synsets. Thus, a document vector is a mix between terms frequencies and synsets frequencies. There are several strategies considered for pruning and modifying the document vectors before using a partioning algorithm to cluster the documents. The documents can belong to several clusters, and a further conceptual clustering is performed using FCA helping to determine the commonalities and distinctions of different clusters. A formal context is constructed where each cluster is an object while the important attributes are derived from the centroid documents of each cluster. This is also another example of incorporating term-to-term dependency in computing the similarity of documents. In both the proposed model and in this the work described in Hotho et al. [2003] it is possible to know exactly what are the common aspects of the documents grouped together.

Overall, these approaches differ from the described model on several counts. First, documents and queries are combinations of frequencies of terms and formal concepts. A formal concept is derived from a formal context constructed based on the semantic model and its relevance to search. Its a concise way of capturing that if d_1 is relevant to a query and given it is annotated by concept X, documents annotated by W, T, and V (due to transitivity) should also be included in the pool of results. Second, the incidence relation of a formal context can capture any semantic path such as author-book-author in a linked dataset. Therefore, this approach is flexible and can be extended by deriving conceptual spaces using other structural elements of the semantic model. The KOS-based semantic model can be replaced with other knowledge resources as long as its structure and content are relevant to information retrieval.

The dependency between documents via their annotating concepts could also be computed by defining a semantic relatedness metric between concepts. Such a metric would require an apriori set weighting scheme or a mechanism to derive it based on the document collection. In the case when the KOS is WordNet such a weighting scheme exists already and it is possible to validate any new metric against a golden standard like the Miller-Charles Miller and Charles [1991] and WordSimilarity-353 Finkelstein et al. [2001]. This model could be refined by creating a more sophisticated weighting scheme for the formal concepts, but this issue remain open for further research.

Finally, if the KOS is WordNet or a similar semantic network, the model can also be viewed as a generalized instance of existing research in document and query expansion with synonyms, hypernyms or hyponyms based on the background knowledge. The concepts considered as objects and attributes for the formal contexts are actually synsets and the incidence relations match the semantic relations between synsets.

5.5 Summary

The proposed *semantic information retrieval model* employs KOSs expressed using SKOS as its source of representational context, describes documents and queries as a combination of concepts from KOS resources, and determines the operational semantics using Formal Concept Analysis. This is achieved by constructing a higher representation as linear combinations of weighted formal concepts extracted from the semantic model.

In this chapter I introduced a semantic retrieval model with the following features:

1. It exploits the rich conceptualizations available within the semantic model;

2. It is not restricted to a particular domain, with the possibility to map documents and queries to any number of semantic models;

3. A semantic model is not necessarily completely describing the domain, and the mixed ranking metric accounts for this by using keyword-based retrieval models as a fallback mechanism;

4. It is a model that scales even when large knowledge bases are used, because of the process of building clusters using Formal Concept Analysis;

5. The interlinked nature of semantic models deployed on the Semantic Web allows extracting missing knowledge from one semantic model to support processes for another such as annotation or disambiguation;

6. The explicit methods used for connecting a semantic model to a document collection can be used as basis for the standardization of evaluation measures and benchmarks for semantic retrieval systems operated within a similar scope with the one defined in Section 1.2.2;

7. It allows to identify and quantify structural characteristics of the KOS models used that correlate with the observed retrieval performance in practical applications.

Chapter 6

Applying the model in monolingual and bilingual settings

6.1 General Requirements

After more than ten years of information retrieval campaigns run by CLEF[1], TREC[2], and NCTIR[3], a report was issued by Braschler and Gonzalo [2009] synthesizing a set of recommendations for the development of future Multilingual Information Access (MLIA) Systems. For this research, the recommendations for both monolingual and bilingual retrieval are particularly valuable and the following list summarizes the ones that influenced the prototype retrieval system's setup.

Recommendations for MLIA systems Braschler and Gonzalo [2009]:

1. Use a retrieval system that supports term weighting and ranked retrieval.

2. Use one of the consistently high-performing weighting schemes such as Okapi-BM25, LM, or Divergence From Randomness framework like

[1] http://www.clef-campaign.org
[2] http://trec.nist.gov
[3] http://research.nii.ac.jp/ntcir

PL2, DLH13, etc.

3. Select high coverage translation resources and add domain-specific resources where possible.

4. Use interlingua in cases where direct translation resources have questionable quality.

5. Use pseudo-relevance feedback as an option to boost recall, expanding the queries with new terms from the top-ranking documents.

6.2 The Flow of Processes in Semantic Information Retrieval

Multilingual search systems incorporate three distinct processes in their setup: indexing, translation, and matching with no fixed order between them Braschler and Gonzalo [2009]. Figure 6.1 presents a comparative image of the flow of interactions between a classic retrieval system and a SIR instance. In both sections of the diagram, during a live run, the system starts by processing a query from the source language, Spanish in this case, followed by its translation to the target language, English. The indexing in the prototype part of the diagram requires pre-processing of the textual information and mapping it to a language-independent *concept-based representation* using knowledge from the SKOS datasets. There are several possible strategies for achieving this, and the evaluation section pinpoints the different results obtained for two particular methods i.e. implicit and explicit annotation. Once the query has been formalized within the representational contexts considered by the system, the matching phase starts, looking to filter the documents relevant to the query, while the relevance score value enables computing a hard number based on which of the final results listings are created.

Apart from these live interactions my SIR system includes several offline phases specific to the theoretical approach presented in Chapter 5. Next, I discuss the pre-processing of the semantic model, the construction of the

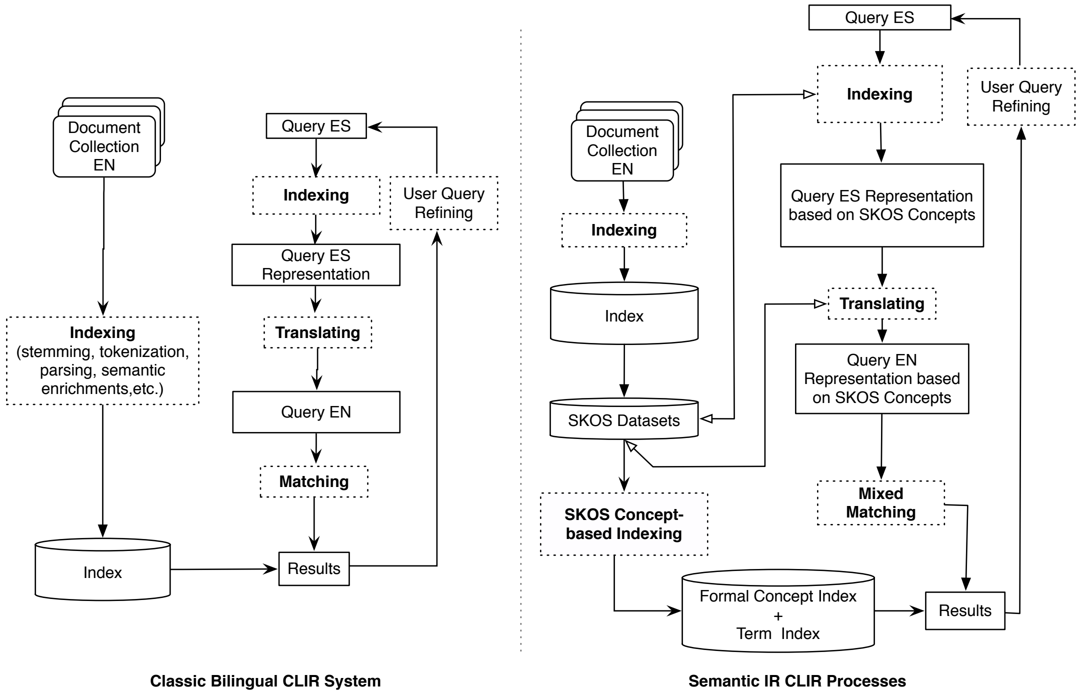

Document Collection EN
Query ES
Indexing
User Query Refining
Query ES Representation
Translating
Indexing (stemming, tokenization, parsing, semantic enrichments,etc.)
Query EN
Matching
Index
Results
Classic Bilingual CLIR System
Query ES
Indexing
User Query Refining
Document Collection EN
Query ES Representation based on SKOS Concepts
Indexing
Index
Translating
Query EN Representation based on SKOS Concepts
SKOS Datasets
Mixed Matching
SKOS Concept-based Indexing
Formal Concept Index + Term Index
Results
Semantic IR CLIR Processes

Figure 6.1: Classic vs SIR Processes

formal concepts index, the mixed sources used for translation, and the relevance considerations incorporated by the matching phase.

Pre-processing the Semantic Model Phase In the experimental setup, I used the domain specific Thesaurus for the Social Sciences (TheSoz)[1] that has been released in SKOS format as described in Zapilko and Sure [2009]. An entry in TheSoz is shown in Figure 6.2. First, the *functional formal contexts* are created with *skos:exactMatch* as the incidence relation.

The actual formal concepts do not need to be built using concept lattice computation algorithms. They can be derived by constructing template SPARQL queries to extract a section of the TheSoz's RDF graph by following the *skos:exactMatch* links. This is where FCA proves its operational role, by helping guide the exploration of the semantic model, TheSoz in this case.

[1] http://datahub.io/dataset/gesis-thesoz

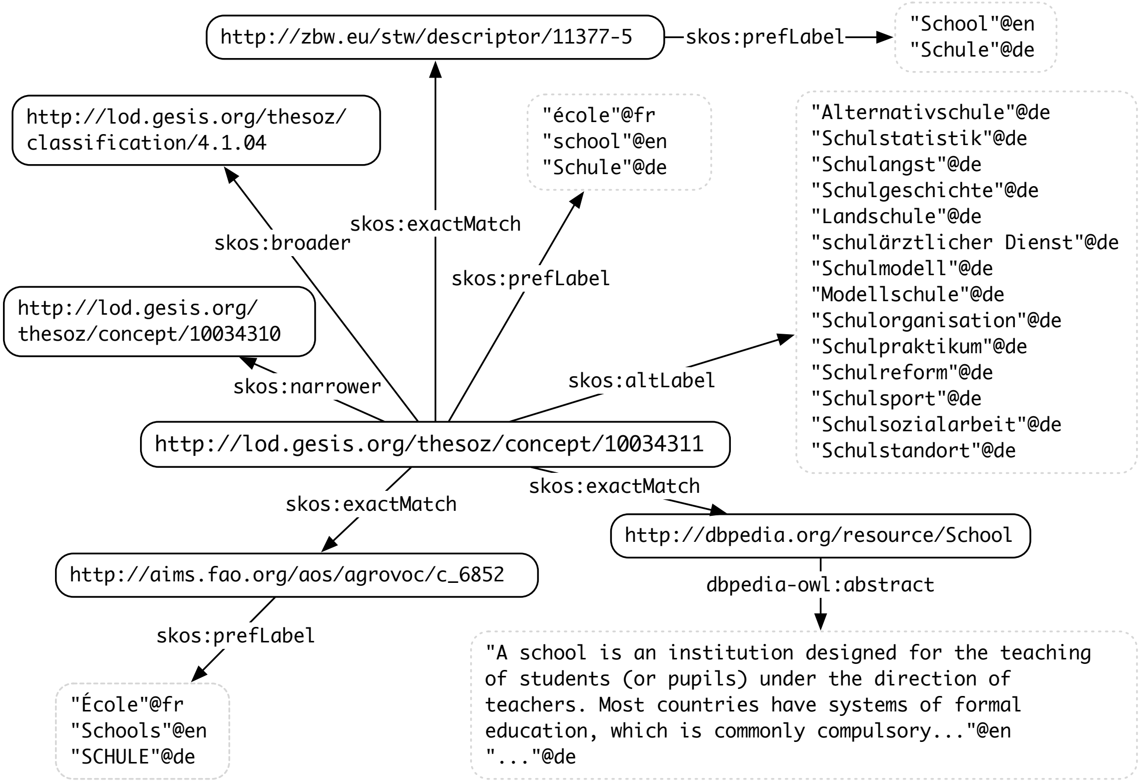


Figure 6.2: TheSoz *school* SKOS Concept

Table 6.1: Cross table for context $\mathbb{K}_{skos:exactMatch}$ with G=M={TheSoz SKOS concepts} and I=*skos:exactMatch*

$\mathbb{K}_{skos:exactMatch}$	thesoz:10034311	agrovoc:c_6852	stw:11377-5	dbpedia:School	...
thesoz:10034311	×	×	×	×	
agrovoc:c_6852	×	×	×	×	
stw:11377-5	×	×	×	×	
dbpedia:School	×	×	×	×	
...					

For each cluster of concepts i.e. formal concept such as (*thesoz:10034311, agrovoc:c_6852, stw:11377-5, dbpedia:School; thesoz:10034311, agrovoc:c_6852, stw:11377-5, dbpedia:School*), its *concept signature* is compiled. In TheSoz, concepts do not have a definition, but do have preferred and alternative labels in several languages. For the experiments below, separate English and German signatures were created. In this case the semantic model incorporates many of the SKOS predicates, which made querying its endpoint straightforward, but in other cases it is necessary to use dataset specific predicates that contain the same type of information as SKOS ones. Thinking in terms of creating contexts and formal concepts helps, as opposed to having a set of SPARQL queries ready that may fail to work.

Second, the two *relational formal contexts* are constructed, where $\mathbb{K}_{skos:narrower}$ has 1184 objects, 8383 attributes, and 13652 relations between the two. After, computing the concept lattice 2975 formal concepts were generated. Similarly, the $\mathbb{K}_{skos:related}$ has 2379 objects, 2379 attributes, with only 3712 relations, resulting in a total of 2492 formal concepts. The counts for the incidence relations show that the two contexts are sparse, therefore the lattices' computations were not problematic. Once, the formal concepts are obtained, they are serialized in a relational database. This is a pipeline process, starting with custom code to build the formal contexts as comma-delimited lists of object and attribute pairs saved to *.con* files, which are converted using FCAStone[1] to a special FCA format *.cxt* that is passed as input to the In-Close algorithm Andrews [2009b].

Indexing Phase The document collection indexing phase takes place offline. A classic term-index is extracted from the collection by pre-processing text using language dependent tools. These include *tokenization* that segments text in sentences and words, *stop word removal* which disregards certain words from text based on language-specific lists, and *stemming* that reduces words to a base form without prefixes or affixes. These language dependent tools can negatively influence a retrieval system's performance

[1] http://fcastone.sourceforge.net/

and several studies looked at the best combinations configurations. An extensive report on the strong link between language resources and multilingual information access was created by Moreau [2009]. For the experimental document collection used in the experiments below both stop word removal and stemming were applied after testing configurations with and without their application.

In addition to these, semantic tools are used to extend documents with metadata describing the topic of a document or annotations recognizing information units (e.g. name of a person, city, time, topic). In this research semantic annotation is outside the scope, therefore the assumption is that the document collection has had the metadata already added. If this is not the case the work presented in Chapter 3 provides a good basis for automatically enhancing the documents.

The theoretical SIR model, I proposed has a two step sequence of deriving the representational set of formal concepts for each document. These sets are formed by *core* and *neighboring* concepts. In practice, it is not necessary to build the M^{SM} matrix, but use γ and μ to map the initial concept annotations into the set of *core formal concepts*. For each element of this set, the upper and lower neighbors are computed and added to the set using the Algorithm 6, respectively Algorithm 5. In practice, the concepts corresponding to an empty object set (the $\mathbf{0}$ zero concept) or empty attribute set (the $\mathbf{1}$ unit concept) are not added to these neighborhoods.

Translation Phase This phase for the CLIR settings provides a transfer mechanism between languages that is suitable for search, and not in the stricter linguistic sense of rendering text in a new language, while preserving the original meaning as accurately as possible Braschler and Gonzalo [2009]. This means that Machine Translation systems like Google Translate and also the multilingual labels in the semantic model can be used. The translation mechanism for a search system has to answer three questions identified by He and Wang [2007], in this case the following answers hold:

What are the translation units? What words or phrases should be translated?

SKOS concepts or phrases, words as a fallback mechanism.

What are the suitable resources for translation: bilingual dictionaries, corpora, and other knowledge and lexical resources to handle out-of-vocabulary situations.

SKOS resources and machine translation

How is the translation knowledge used when words or phrases have several translations?

Select the appropriate SKOS resource based on search systems's application domain keeping the translation of a word or phrase within the domain. Use statistical-based machine translation as a fallback mechanism.

Matching Phase In the proposed semantic information retrieval model, a document matches a query if their concept neighborhoods overlap. Determining a ranking between documents is based on Equation 5.42. The second component of this mixed metric $\mathcal{R}(q,d)$ is obtained using high-performing weighting schemes from the Divergence From Randomness family, known to perform well in monolingual settings: PL2 and DLH13 (see Appendix B for more details on these models).

6.3 Retrieval Evaluation

The evaluation of IR systems has relied on laboratory-style evaluation experiments for many years. It allows comparing systems on a test-suite of reusable data (*test collection*, *topics* a.k.a queries, *relevance judgements*). The CLEF project has been using a comparative evaluation approach, where a control task is predefined. This corresponds to testing the function of a complete system or of a single component. This framework does not allow any evaluation of a user's satisfaction with the system or his information seeking behavior Robertson [2001]. It just measures effectiveness of computing *precision* and *recall* for each query, and for an overall system measurement, the *mean average precision* (MAP).

6.3.1 Measures of Retrieval Effectiveness

The following definitions describe established measures of effectiveness based on [Peters et al., 2012, p.147]. For $D = \{d_j| \ d_j \in D\}$ a set of N documents and $Q = \{q_i| \ q_i \in Q\}$ queries. The relevant documents for a query q_i are denoted with $D^{rel}(q_i) \subset D$. The rank parameter r, where a small r denotes a focus on precision, while a large r indicates an exhaustive search. $D_r(q_i)$ the answer set of the first r documents helps define the $D_r^{rel}(q_i) = D^{rel}(q_i) \cap D_r(q_i)$ as the subset of relevant retrieved documents.

Precision, π, is the proportion of the retrieved documents which are relevant.

$$\pi_r(q_i) = \frac{D_r^{rel}(q_i)}{D_r(q_i)} \tag{6.1}$$

Recall, ρ is the proportion of the relevant documents which has been retrieved.

$$\rho_r(q_i) = \frac{D_r^{rel}(q_i)}{D^{rel}(q_i)} \tag{6.2}$$

Average precision, AP, is given by:

$$AP_i = \frac{1}{|D^{rel}(q_i)|} \sum_{r=1}^{|D|} \rho_r(q_i) * rel(r), \tag{6.3}$$

where $rel(r)$ = 1 if the document at rank r is relevant to query q_i and $rel(r)$ = 0 otherwise.

Mean average precision:

$$MAP = \frac{1}{|Q|}(\sum_{q_i \in Q} AP_i) \tag{6.4}$$

6.3.2 Experimental Setup

Test Document Collection The experiments carried out rely on the CLEF Domain Specific 2004-2008 test-suite distributed by the European Language Association (ELRA)[1]. The document collection is the German Indexing and

[1] http://catalog.elra.info/

Retrieval Test database (GIRT) and a set of topics from the CLEF 2004-2006 Domain-Specific (DS) track. GIRT consists of two parallel corpora in EN and DE composed of bibliographic records extracted from various sources in the social sciences domain, each with 151319 documents. The documents are in XML format (see Figure 6.3 for a German example) and consist of a unique identifier (tag <DOCNO>), title (tag <TITLE-DE>), author name (tag <AUTHOR>), document language (tag <LANGUAGE-CODE>), publication date (tag <PUBLICATION-YEAR>) and abstract (tag <ABSTRACT-DE>). Manually assigned descriptors and classifiers are provided for all documents. In the German corpus all documents consist of a title and an abstract. Additionally, a typical record also contains manually assigned metadata terms from TheSoz (tags <CONTROLLED-TERM-DE>, <CLASSIFICATION-TEXT-DE>, <METHOD-TEXT-DE>, and <METHOD-TERM-DE>).

The English collection is a human-translated version of the German collection. The tags maintain the same meaning for the English records (see Figure 6.4 for an example). However, abstracts are available for only around 15% of the English records Dolamic and Savoy [2010].

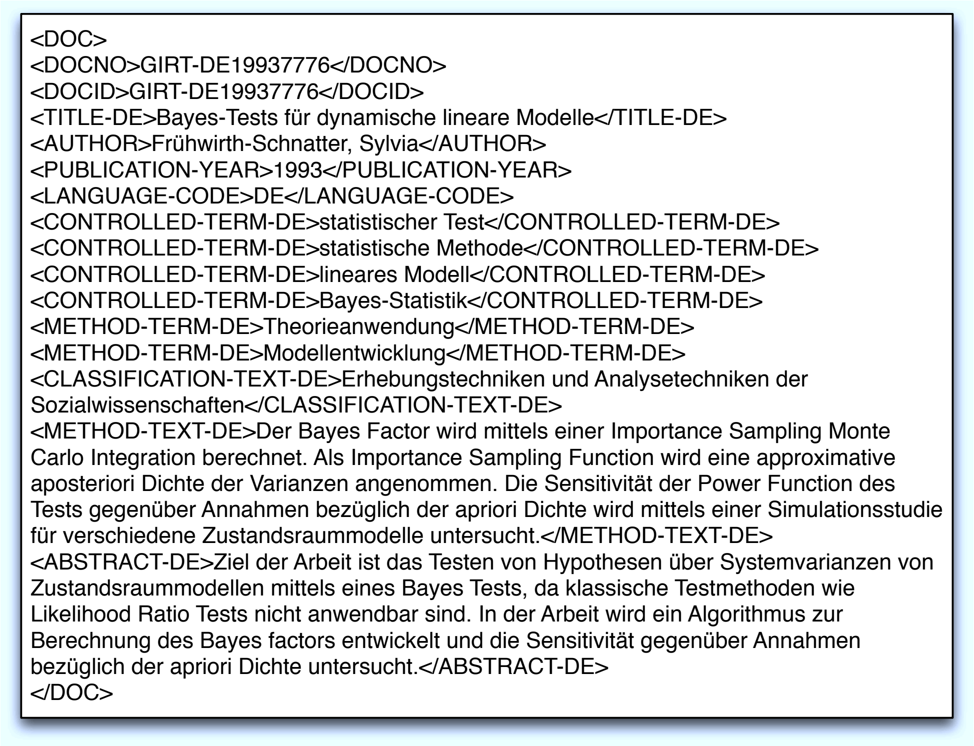
From the original collection of XML files, I have built an RDF graph of documents and annotations using the Open Annotation Core Model, where the following triples are generated for GIRT-DE19937776 document. URIs <thesoz:*>[1] match the SKOS concepts for the controlled terms assigned to the document (e.g. statistischer Test, statistische Methode, etc.)

```
<GIRT-DE19937776> annotatedBy <thesoz:10051231> .
<GIRT-DE19937776> annotatedBy <thesoz:10052184> .
<GIRT-DE19937776> annotatedBy <thesoz:10057920> .
<GIRT-DE19937776> annotatedBy <thesoz:10037769> .
```

Similarly, the following triples were extracted from the English records:

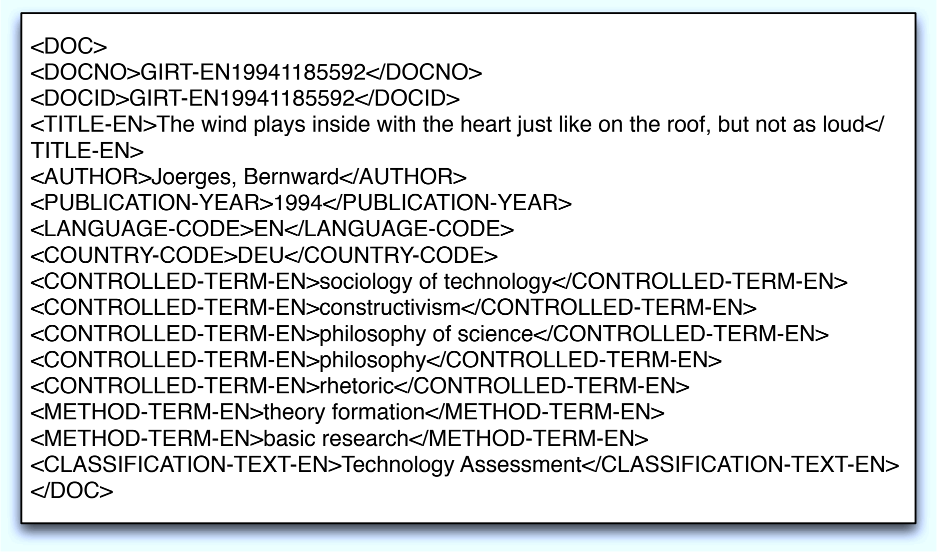
```
<GIRT-EN19941185592> annotatedBy <thesoz:10063417> .
<GIRT-EN19941185592> annotatedBy <thesoz:10049622> .
```

[1] http://lod.gesis.org/thesoz/concept/



```
<DOC>
<DOCNO>GIRT-DE19937776</DOCNO>
<DOCID>GIRT-DE19937776</DOCID>
<TITLE-DE>Bayes-Tests für dynamische lineare Modelle</TITLE-DE>
<AUTHOR>Frühwirth-Schnatter, Sylvia</AUTHOR>
<PUBLICATION-YEAR>1993</PUBLICATION-YEAR>
<LANGUAGE-CODE>DE</LANGUAGE-CODE>
<CONTROLLED-TERM-DE>statistischer Test</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>statistische Methode</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>lineares Modell</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Bayes-Statistik</CONTROLLED-TERM-DE>
<METHOD-TERM-DE>Theorieanwendung</METHOD-TERM-DE>
<METHOD-TERM-DE>Modellentwicklung</METHOD-TERM-DE>
<CLASSIFICATION-TEXT-DE>Erhebungstechniken und Analysetechniken der
Sozialwissenschaften</CLASSIFICATION-TEXT-DE>
<METHOD-TEXT-DE>Der Bayes Factor wird mittels einer Importance Sampling Monte
Carlo Integration berechnet. Als Importance Sampling Function wird eine approximative
aposteriori Dichte der Varianzen angenommen. Die Sensitivität der Power Function des
Tests gegenüber Annahmen bezüglich der apriori Dichte wird mittels einer Simulationsstudie
für verschiedene Zustandsraummodelle untersucht.</METHOD-TEXT-DE>
<ABSTRACT-DE>Ziel der Arbeit ist das Testen von Hypothesen über Systemvarianzen von
Zustandsraummodellen mittels eines Bayes Tests, da klassische Testmethoden wie
Likelihood Ratio Tests nicht anwendbar sind. In der Arbeit wird ein Algorithmus zur
Berechnung des Bayes factors entwickelt und die Sensitivität gegenüber Annahmen
bezüglich der apriori Dichte untersucht.</ABSTRACT-DE>
</DOC>
```

Figure 6.3: Sample of GIRT German Document



```
<DOC>
<DOCNO>GIRT-EN19941185592</DOCNO>
<DOCID>GIRT-EN19941185592</DOCID>
<TITLE-EN>The wind plays inside with the heart just like on the roof, but not as loud</
TITLE-EN>
<AUTHOR>Joerges, Bernward</AUTHOR>
<PUBLICATION-YEAR>1994</PUBLICATION-YEAR>
<LANGUAGE-CODE>EN</LANGUAGE-CODE>
<COUNTRY-CODE>DEU</COUNTRY-CODE>
<CONTROLLED-TERM-EN>sociology of technology</CONTROLLED-TERM-EN>
<CONTROLLED-TERM-EN>constructivism</CONTROLLED-TERM-EN>
<CONTROLLED-TERM-EN>philosophy of science</CONTROLLED-TERM-EN>
<CONTROLLED-TERM-EN>philosophy</CONTROLLED-TERM-EN>
<CONTROLLED-TERM-EN>rhetoric</CONTROLLED-TERM-EN>
<METHOD-TERM-EN>theory formation</METHOD-TERM-EN>
<METHOD-TERM-EN>basic research</METHOD-TERM-EN>
<CLASSIFICATION-TEXT-EN>Technology Assessment</CLASSIFICATION-TEXT-EN>
</DOC>
```

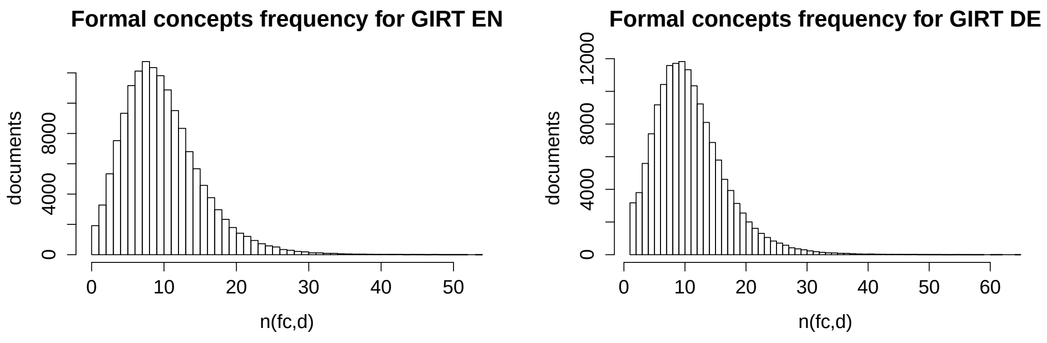
Figure 6.4: Sample of GIRT English Document

Table 6.2: Formal Concepts to Document Connections

Formal Concept Context	Collection	Formal Concepts Links Count
$\mathbb{K}_{skos:narrower}$	GIRT DE	1437159
$\mathbb{K}_{skos:narrower}$	GIRT EN	1250980
$\mathbb{K}_{skos:related}$	GIRT DE	655743
$\mathbb{K}_{skos:related}$	GIRT EN	580240

```
<GIRT-EN19941185592> annotatedBy <thesoz:10035530> .
<GIRT-EN19941185592> annotatedBy <thesoz:10045191> .
<GIRT-EN19941185592> annotatedBy <thesoz:10056758> .
```

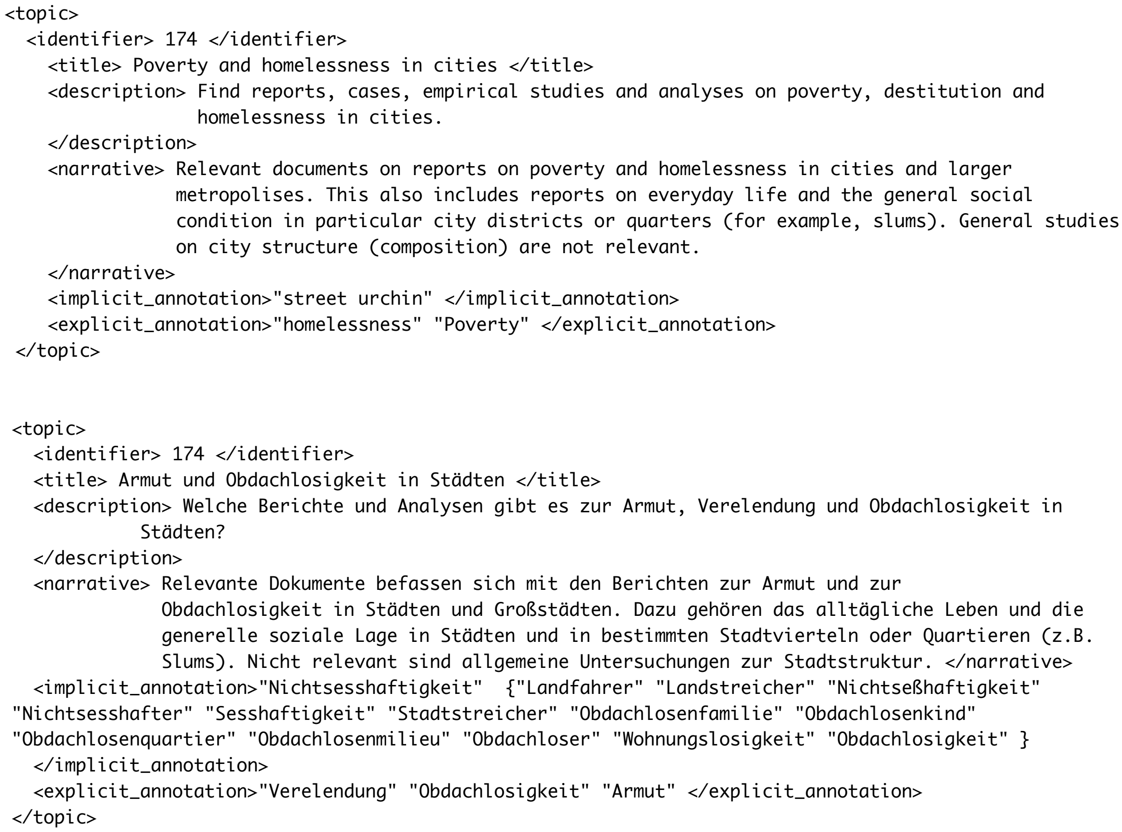
All the fields available were used to create a term-based index for the original document collection. For the RDF graph derived from the collection only the <CONTROLLED-TERM-*> tag is used. The formal concepts-based index is created following the two step procedure in Chapter 5, constructing a semantic core and a neighborhood for each document. Figure 6.5 shows the distribution of formal concepts per document with a mean of 10 concepts per document. The links established totals are specified in Table 6.2.



(a) (b)

Figure 6.5: Formal concepts frequency per documents

Topics The experiments used 75 topics in English (EN) and German (DE) as in Figure 6.6.



```
<topic>
  <identifier> 174 </identifier>
    <title> Poverty and homelessness in cities </title>
    <description> Find reports, cases, empirical studies and analyses on poverty, destitution and
                  homelessness in cities.
    </description>
    <narrative> Relevant documents on reports on poverty and homelessness in cities and larger
                metropolises. This also includes reports on everyday life and the general social
                condition in particular city districts or quarters (for example, slums). General studies
                on city structure (composition) are not relevant.
    </narrative>
    <implicit_annotation>"street urchin" </implicit_annotation>
    <explicit_annotation>"homelessness" "Poverty" </explicit_annotation>
 </topic>


 <topic>
   <identifier> 174 </identifier>
   <title> Armut und Obdachlosigkeit in Städten </title>
   <description> Welche Berichte und Analysen gibt es zur Armut, Verelendung und Obdachlosigkeit in
             Städten?
   </description>
   <narrative> Relevante Dokumente befassen sich mit den Berichten zur Armut und zur
               Obdachlosigkeit in Städten und Großstädten. Dazu gehören das alltägliche Leben und die
               generelle soziale Lage in Städten und in bestimmten Stadtvierteln oder Quartieren (z.B.
               Slums). Nicht relevant sind allgemeine Untersuchungen zur Stadtstruktur. </narrative>
   <implicit_annotation>"Nichtsesshaftigkeit"  {"Landfahrer" "Landstreicher" "Nichtseßhaftigkeit"
 "Nichtsesshafter" "Sesshaftigkeit" "Stadtstreicher" "Obdachlosenfamilie" "Obdachlosenkind"
 "Obdachlosenquartier" "Obdachlosenmilieu" "Obdachloser" "Wohnungslosigkeit" "Obdachlosigkeit" }
   </implicit_annotation>
   <explicit_annotation>"Verelendung" "Obdachlosigkeit" "Armut" </explicit_annotation>
 </topic>
```

Figure 6.6: Example CLEF Topic

A topic is a textual statement of a user need, identified by a unique topic number, and is provided in three different lengths:

- *title*, a short formulation of a few keywords;
- *description*, a somewhat longer formulation consisting of one or two sentences;
- *narrative*, a lengthy formulation detailing specific preferences;

The narrative forms the basis on which the relevance assessments are established. In some experiments both *title* and *description* are often used. In our case all the experiments use only the *title*.

Relevance assessments These are lists of document identifiers, complete with information of relevance with regard to specific topics. The CLEF project provides binary relevance assessments: documents are either relevant or irrelevant with respect to a specific topic. These are also referred to as *qrels*.

Development Setup The following list describes the main components used in implementing and determining the results below.

- Search Engine: Terrier 4.0 IR Platform[1]
- Knowledge Bases: TheSoz (DE,EN,FR), DBpedia, AGROVOC(19 languages), STW (mainly DE)
- Natural Language Processing: GATE[2] Embedded is an object-oriented framework for performing semantic annotations tasks; APOLDA a GATE Plugin[3]
- Semantic repository: Virtuoso Universal Server[4]
- Translation Service: GoogleTranslate

[1] http://terrier.org/

[2] http://gate.ac.uk/download/

[3] http://apolda.sourceforge.net/

[4] http://virtuoso.openlinksw.com/

6.4 Experiments

The design of the experiments aimed to investigate in stages the impact of representing queries and documents as linear combinations of weighted formal concepts in comparison with the classic term frequency based representations. The underlying research questions are RQ3-RQ6, how effective is the formalization of queries and documents that takes into account their concept annotations and the semantic relations between these concepts as described by the semantic model.

To determine this I studied component by component modifications in the prototype system's performance. All experiments have been run on the same document collection and set of query topics. The prototype system used Terrier 4.0.2 as the underlying search system with extension code to handle the proposed model's specific phases.

In all the experiments the queries go through the following pipeline of processes: stop words removal, followed by decompounding for the German queries only and stemming using the PorterStemmer specific to each language.

The method used for pseudo-relevance feedback in the term weighting model Bo1 that is defined in the Divergence From Randomness (DFR) framework Amati [2003]. The weighting model infers the informativeness of a term by the divergence between its distribution in the top-ranked documents and a random distribution. It is considered the most effective DFR term weighting model and its formula is described by Equation 6.5.

$$w(k) = tf_{x=3}(k)\log_2((1+n(k,N)/N) * N/n(k,N)) + \log(1+n(k,N)/N) \quad (6.5)$$

where $tf_x(k)$ is the frequency of the query term k in the x top-ranked documents, $n(k,N)$ is the frequency of the query term k in the collection, and N is the number of documents in the collection.

There are four groups of experiments each addressing one or several of the research questions:

- Experiment 1: RQ3
- Experiment 2: RQ4, RQ5
- Experiment 3: RQ6
- Experiment 4: RQ7

6.4.1 Experiment 1: Query expansion using concept labels

Scope *Improving system effectiveness*

Objective *Performance focused on query semantic annotation*

Test collection *GIRT collection and GIRT RDF graph*

Topics *Original CLEF DS 2004-2006 Topics with added fields for implicit and explicit annotations, Topics RDF graph*

Relevance assessment *CLEF Domain Specific Track qrels for DE and EN*

Effectiveness measure *Mean Average Precision*

This experiment aims to investigate the impact of expanding queries based on the lexicalization of the concepts that annotate the queries. Three methods are proposed: a) *explicit annotation* equivalent to matching concepts to text by label; b) *implicit annotation* where a text's topic is identified based on the similarity between the query and the concept's signature; and c) *pseudo-relevance annotation* where the queries are expanded based on the labels of the concepts annotating the top-ranking documents.

6.4.1.1 Finding literal occurrences of concepts in a text (explicit semantic annotation)

The explicit semantic annotation aspect of these experiments relies on APOLDA (Automated Processing of Ontologies with Lexical Denotations for Annotation) Gate plugin Wartena et al. [2007] to determine annotations based on the SKOS-converted-to-OWL of the initial KOS resource. This plugin provides a scalable solution for basic text annotation using TheSoz's concepts labels. The output of this process is a new field *explicit_annotation* to all the topics built from the annotations of the topic's title and description. It was not necessary to disambiguate the annotations, since the topics and TheSoz are from the same subject domain. For example, the Topic 174 with title *Poverty and homelessness in cities* is linked to the TheSoz concepts with the labels *homelessness* and *poverty*.

6.4.1.2 Beyond literal occurrences (implicit semantic annotation)

In order to create a topic level annotation, beyond exact matches of labels in text, TheSoz's links to other SKOS datasets particularly DBpedia are used. For each of the concepts with an exact match to other concept schemes a set of SPARQL queries were run, exploring the other datasets looking for preferred and alternative labels. This is where the *functional formal context* was used. Note, TheSoz concepts do not have definitions, but their exact DBpedia counterparts do. Thus, the definitions for the 5024 linked TheSoz concepts were extracted from DBpedia. The outcome is a set of textual signatures for each of the dataset's concepts. Some concepts have longer signatures than others and this impacted the quality of the implicit annotations.

To produce the implicit semantic annotations the topics are used as queries against a term-index built using Terrier over the set of concepts signatures. Therefore, the results set in this case will be a list of concepts. BM25 Baeza-Yates and Ribeiro-Neto [2011] was used as the matching model and the top-ranking concept in the ranking list is extracted. They were added to the *implicit_annotations* field together with any alternative labels found across the schemes (marked by curly brackets in Figure 6.6). For Topic 174 a successful match was *street urchin* for the English set of topics, while for its German counterpart it was *Nichtsesshaftigkeit* (*vagrancy*). It is noticeable from Figure 6.6 that the dominant German language across TheSoZ and STW introduces a bias in this annotation process.

Overall, on manual examination approximately 25% of the topics' both implicit and explicit annotations are complementing each other and circumventing their intent (e.g. Topic 174: *Poverty and homelessness in cities* with implicit annotation *street urchin* and explicit annotations *poverty, homelessness*. Yet, perfect and balanced annotations are hard to achieve automatically, and as the results in this experiment will show, these annotations led to varied retrieval results.

6.4.1.3 Pseudo-relevance based annotation

This method relies on an existing pseudo-relevance method to identify the first three top-ranking documents for each query. In this case the documents are all annotated with concepts. The concepts from the top-ranking documents are identified and the ones that occur more than once are used in expanding the query. No relationships between concepts are considered in this case. The assumption is that the concepts are all independent of each other.

6.4.1.4 Results

For all the runs in this experiment described in Tables 6.3, 6.4, 6.5, and 6.6 language-specific stop word lists and stemmers were used. The queries were formulated as combinations, considering at turn pairings between the title (T) or the title and description(TD) and annotations (implicit annotations - IA, explicit annotations - EA, and pseudo-relevance annotation - PRFA). The matching model used was PL2 (Poisson estimation for randomness)[1] and for the baseline query expansion model the Bo1 model in Equation 6.5. For the bilingual runs, the annotations were translated based on TheSoz's multilingual labels, while the topics' *titles* were translated using Google's Translate service.

The results show that the explicit annotations (T+EA) runs outperformed the implicit annotations (T+IA) and in most instances the baselines (T and

[1] http://terrier.org/docs/v3.5/configure_retrieval.html#cite1

Table 6.3: MAP results for English annotated query topics

	Baseline (T)	T+IA	T+EA	T+IA+EA	T+PRFA
EN	37.00	32.06	**39.17**	36.11	33.35
	Baseline (TD)	TD+IA	TD+EA	TD+IA+EA	TD + PRFA
EN	39.24	36.63	**40.02**	38.44	33.72

Table 6.4: MAP results for German annotated query topics

	Baseline (T)	T+IA	T+EA	T+IA+EA	T+PRFA
DE	**42.17**	35.35	40.29	37.44	38.48
	Baseline (TD)	TD+IA	TD+EA	TD+IA+EA	TD + PRFA
DE	41.38	38.93	**41.70**	40.45	40.87

Table 6.5: MAP results for German to English annotated and translated query topics

	Baseline (T)	T+IA	T+EA	T+IA+EA	T+PRFA
DE-EN	35.51	34.55	**38.79**	38.57	34.18
	Baseline (TD)	TD+IA	TD+EA	TD+IA+EA	TD + PRFA
DE-EN	36.71	38.60	37.81	**39.61**	37.05

Table 6.6: MAP results for English to German annotated and translated query topics

	Baseline (T)	T+IA	T+EA	T+IA+EA	T+PRFA
EN-DE	38.67	31.50	**38.73**	34.09	38.34
	Baseline (TD)	TD+IA	TD+EA	TD+IA+EA	TD + PRFA
EN-DE	38.89	37.64	39.05	38.43	**40.40**

TD). The problem with EA is that it requires the queries to have an accompanying description as is the case with TREC-style queries in a lab setting, but not in a live retrieval system setting. Also, the expansion based on pseudo-relevance concepts did not improve the MAP results across the different settings. This echoes the fact that document annotations are not consistent, even though two documents may be very close to each other conceptually one could be annotated with a more general concept, while the other is annotated with a more specific concept. Therefore the search based on concept expansion would not necessarily select them both at the same time.

These mixed results demonstrate that query expansion using just concept lexicalization and ignoring concept relations does not lead to improved results. Therefore for **RQ3** in Section 1.2.3 the conclusion is that query expansion with concept labels does not improve the baselines.

6.4.2 Experiment 2: Formal Concepts based indexing and mixed ranking

Scope *Improving system effectiveness*

Objective *FCA retrieval model compared to other IR models in monolingual and bilingual settings*

Hypothesis *The proposed FCA-based representation impacts positively on performance*

Test collection *GIRT collection and GIRT RDF graph*

Topics *Original CLEF DS 2004-2006 Topics*

Relevance assessment *CLEF Domain Specific Track qrels for DE and EN*

Effectiveness measure *Mean Average Precision*

In this experiment group pseudo-relevance feedback is used to determine topic annotations. Each topic's title was submitted as a query and based on the top three documents retrieved their TheSoz annotating concepts were considered annotations for the query. In monolingual settings, as in this context, pseudo-relevance feedback is frequently used to improve recall. In the next two experiments, I also use this technique to automatically assign concepts to the search topic. Once the thesaurus concepts are selected, the formal concepts representation is constructed and used to filter out the documents that do not match the query in terms of conceptual overlap, followed by the application of the ranking function in Equation 5.42.

I extracted from the output file during this evaluation session the formal concepts generated for Topic 174. To make it readable I converted the TheSoz concept URIs into labels *(honorarium, livelihood, income;income), (subsistence level, low income, immiseration, poverty, combating poverty, state of destitution; poverty)*. These formal concepts are assigned their corresponding weights and integrated in the ranking function.

To judge the difference between runs I used the paired test. The hypothesis is set as a one-tailed t-test, aiming to determine if the new retrieval model impacts positively on the overall performance of the system. 75 topics were used for each run.

H_0: The systems producing the two runs have the same retrieval characteristics and any difference between the runs occurred by random chance.

H_1: The FCA-based retrieval system in implementation outperforms the classic IR models.

Several runs were executed with the following parameters: term weighting model (TF-IDF, PL2, DLH13), language setting (EN, DE, DE-EN, EN-DE), and α with values 0.0, 0.25, 0.50, 0.60, 0.65, 0.70, 0.75, and 1.0.

In most instances when α ranges between 0.50 and 0.75, the t-test rejected the null hypothesis H_0, which based on statistical inference concludes that the alternative hypothesis H1 is true. The results of pair testing with $p < 0.05$ in Tables 6.7, 6.8, and 6.9 are marked by the symbol * when statistically significant improvements occurred between the performance of the proof of concept implementation of the SIR model in the previous chapter and the baseline ($\alpha = 0$) a standard system implementing the classic vector space model using one of the following term weighting schemes TF-IDF, DLH13 and PL2.

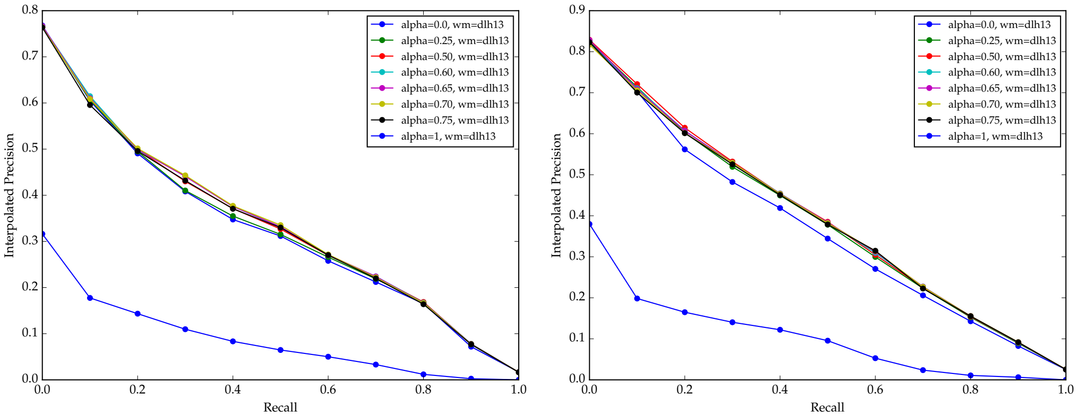
RQ4: Queries and documents are represented as mixed vectors of weighted terms and formal concepts constructed from the semantic model, what is the impact of this representation and ranking parameter α in comparison to existing vectorial represenions based only on term-weighting models such as TF-IDF, DLH13, and PL2?

In Figures 6.7, 6.8, 6.9 the results for precision and recall for all 75 topics are plotted demonstrating that for $\alpha = 1.0$ the proposed model underperforms significantly. This was to be expected because the model does not introduce a method to compare formal concepts, but to cluster as a response to a query all documents that share the same concepts or are connected through paths in the semantic model. The role of the weighted term-based part of the vectorial representation is to support in the final ranking of documents. Also, the performance of $\alpha = 1.0$ is improved when the $\mathbb{K}_{SM}$ does not include all formal contexts derived from the semantic model like in Tables 6.12, 6.13, and 6.14 where only $\mathbb{K}_{narrower}$ and $\mathbb{K}_{related}$ are used.

For $\alpha = 0.0$ the plots describe the baselines of the experiments depending only on the term-weighting models. Several models were considered because it allows investigating if the increase in performance for $\alpha > 0.0$ is consistent. For all configurations there are several instances for α when statistical significance is obtained.

RQ5: The weighted formal concepts part of the defined vectors representing text are language independent, does the bilingual setting outperform machine translation as the baseline?

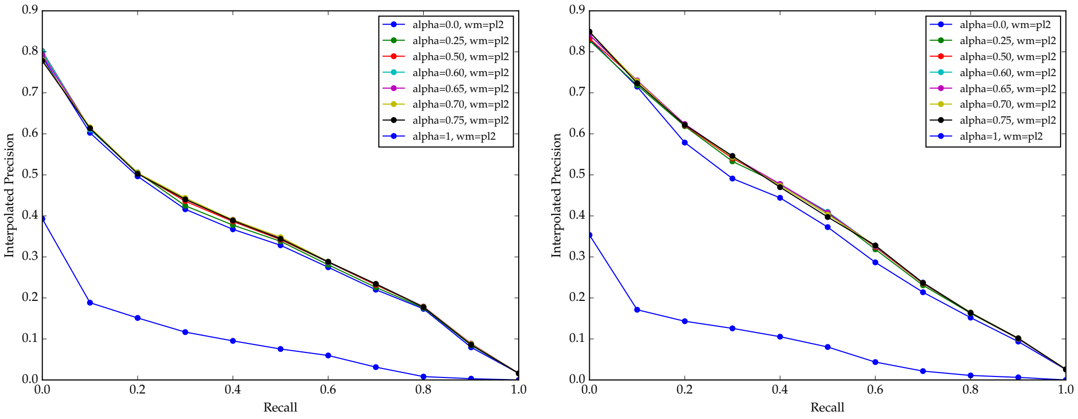
Figures 6.10, 6.11, and 6.12 summarize the bilingual performance of the SIR model proposed. In a bilingual retrieval setting the query is annotated based on pseudo-relevance feedback by concepts, which are mapped to their corresponding formal concept using the μ operator. This allows determining the formal concepts. Note that the formal concepts are independent of language allowing the selection of documents from the target



(a) EN

(b) DE

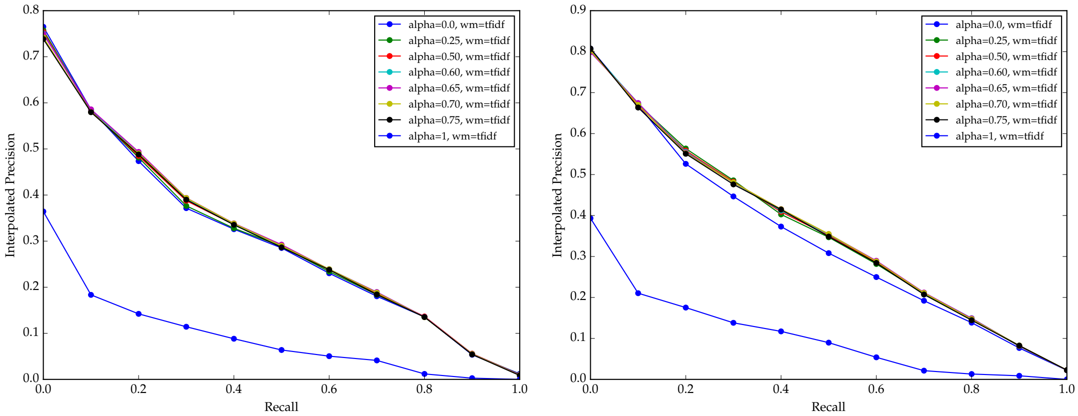
Figure 6.7: Monolingual retrieval performance with weighting model DLH13



(a) EN

(b) DE

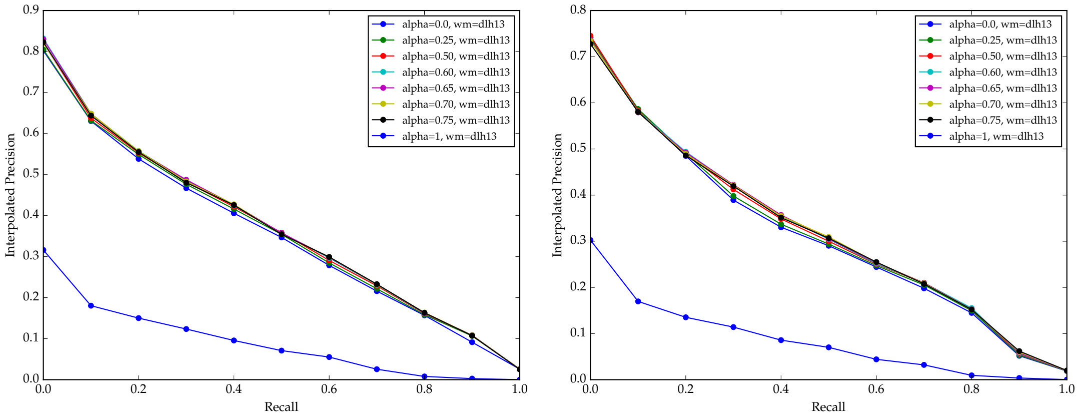
Figure 6.8: Monolingual retrieval performance with weighting model PL2



(a) EN (b) DE

Figure 6.9: Monolingual retrieval performance with weighting model TF-IDF

language to be added to the pool of documents to be ranked.



(a) EN-DE (b) DE-EN

Figure 6.10: Bilingual retrieval performance with weighting model DLH13

Table 6.7: MAP results for weighting model DLH13, with $\alpha = 0.0$ as baseline, and $\mathbb{K}_{SM} = \mathbb{K}_{skos:broader} \cup \mathbb{K}_{skos:narrower} \cup \mathbb{K}_{skos:related}$

WM=DLH13	α**=0.0**	α**=0.25**	α**=0.50**	α**=0.60**	α**=0.65**	α**=0.70**	α**=0.75**	α**=1.0**
EN	32.32	32.68*	33.34*	33.51*	33.54*	33.57*	33.08	7.51
EN + QE	35.53	35.77*	36.03*	36.11	36.09	35.99	35.75	7.51

DE-EN	30.67	31.10*	31.53*	31.68*	31.69*	31.68*	31.57	7.20
DE-EN + QE	34.46	34.65*	34.86*	34.87	34.80	34.64	34.29	7.20
DE	35.61	37.74*	38.20*	38.03*	38.05*	37.91*	37.75*	8.85
DE + QE	40.99	41.26*	41.40	41.09	40.80	40.53	40.19	8.85
EN-DE	34.45	35.25*	35.73*	36.01*	36.09*	35.93*	35.86*	7.49
EN-DE + QE	37.56	37.86*	38.25*	38.44*	38.27	38.30	38.16	7.49

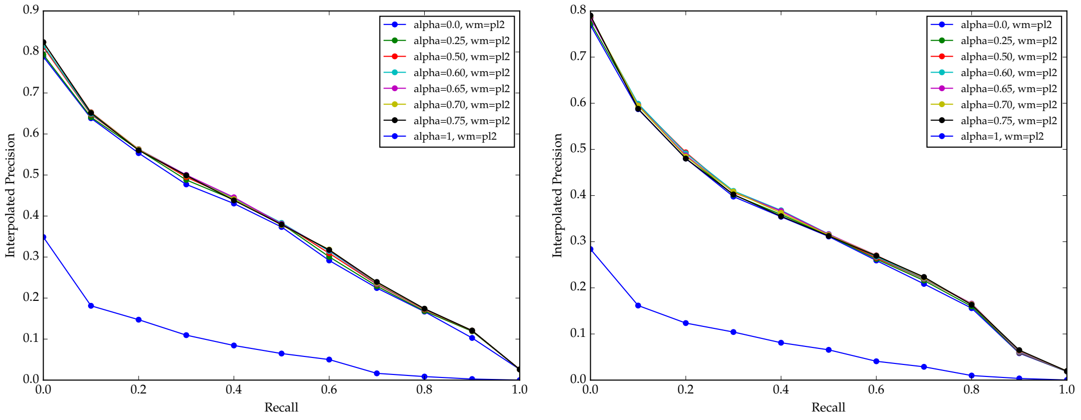
As seen in Tables 6.7, 6.8, and 6.9 the rows for bilingual settings are marked by EN-DE or DE-EN and for α between 0.25 and 0.70 statistically significant improvements were obtained. The baseline in this case is given by the machine translation of the queries using Google Translate. This is a strong baseline achieving 95% of monolingual performance.

Table 6.8: MAP results for weighting model PL2, with $\alpha = 0.0$ as baseline, and $\mathbb{K}_{SM} = \mathbb{K}_{skos:broader} \cup \mathbb{K}_{skos:narrower} \cup \mathbb{K}_{skos:related}$

WM=PL2	α**=0.0**	α**=0.25**	α**=0.50**	α**=0.60**	α**=0.65**	α**=0.70**	α**=0.75**	α**=1.0**
EN	33.35	34.03*	34.34*	34.43*	34.44*	34.43*	34.26	8.31
EN + QE	37.00	37.19*	37.31	37.29	37.18	37.00	36.71	8.31
DE-EN	31.73	32.13*	32.51*	32.59*	32.56*	32.48*	32.30	6.72
DE-EN+QE	35.51	35.66*	35.76	35.72	35.57	35.33	35.01	6.72
DE	36.99	39.06*	39.34*	39.47*	39.46*	39.41*	39.23*	7.76
DE + QE	42.17	42.34	42.35	42.22	42.07	41.81	41.39	7.76
EN-DE	35.69	36.49*	36.88*	37.02*	37.11*	37.08*	36.92*	7.29
EN-DE + QE	38.69	38.93*	39.24*	39.32*	39.30	39.23	38.82	7.29

Table 6.9: MAP results for weighting model TF-IDF, with $\alpha = 0.0$ as baseline, and $\mathbb{K}_{SM} = \mathbb{K}_{skos:broader} \cup \mathbb{K}_{skos:narrower} \cup \mathbb{K}_{skos:related}$

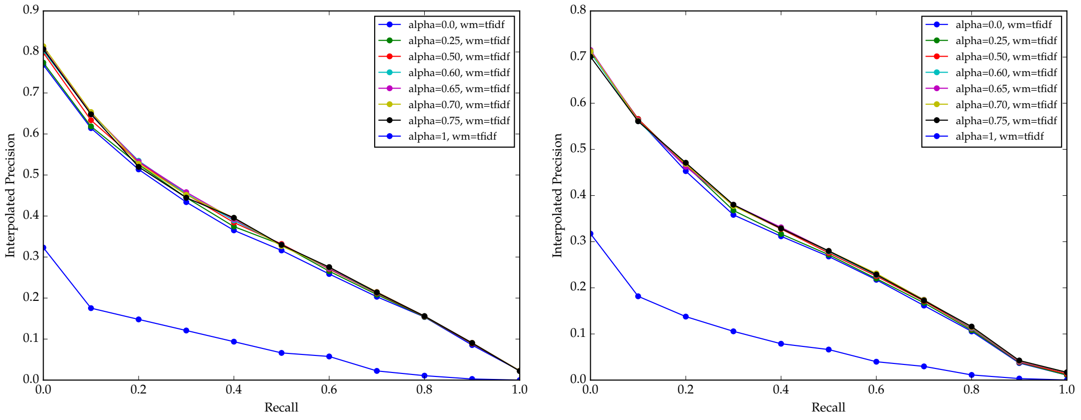
WM=TF-IDF	α**=0.0**	α**=0.25**	α**=0.50**	α**=0.60**	α**=0.65**	α**=0.70**	α**=0.75**	α**=1.0**
EN	29.68	30.02*	30.39*	30.43*	30.56*	30.38	30.15	7.81
EN+QE	31.91	32.23*	32.62*	32.69	32.65	32.55	32.16	7.81
DE-EN	27.89	28.43*	28.91*	29.05*	29.11*	29.08*	28.96	7.16
DE-EN + QE	31.01	31.33*	31.72*	31.83*	31.79*	31.69	31.36	7.16
DE	33.03	35.10*	35.15*	35.21*	35.23*	35.19*	34.92	9.06
DE + QE	37.10	37.39*	37.19	37.09	36.96	36.73	36.41	9.06
EN-DE	32.38	33.12*	33.66*	33.96*	34.11*	34.03*	33.71	7.40
EN-DE + QE	34.57	34.94*	35.17	35.33	35.38	35.43	35.27	7.40



(a) EN-DE

(b) DE-EN

Figure 6.11: Bilingual retrieval performance with weighting model PL2



(a) EN-DE

(b) DE-EN

Figure 6.12: Bilingual retrieval performance with weighting model TF-IDF

6.4.3 Experiment 3: Query Expansion Component

Scope *Improving system effectiveness*

Objective *FCA retrieval model compared to other IR models in monolingual and bilingual settings when using Query Expansion*

Hypothesis *The proposed FCA-based representation impacts positively on performance*

Test collection *GIRT collection and GIRT RDF graph*

Topics *Original CLEF DS 2004-2006 Topics*

Relevance assessment *CLEF Domain Specific Track qrels for DE and EN*

Effectiveness measure *Mean Average Precision*

Query expansion is a two step technique. The first step is to extract a set of feedback documents from the first submission of the query. In this case the first three documents. In the second step, all terms are ranked in descending order of their $tf \cdot idf$ weights and a fixed number of them (in this case 10) are added to the query to be re-submitted for search. Query expansion does not always perform well, if the feedback documents cover a wide variety of topics, indadvertedly introducing noise in the results set He and Ounis [2009].

H_0: The systems producing the two runs have the same retrieval characteristics and any difference between the runs occurred by random chance.

H_1: The FCA-based retrieval system in implementation outperforms the classic IR models when using Query Expansion.

RQ6: Considering an effective query expansion method how does the formal concept based-expansion of queries and documents perform in comparison?

In this case, the results obtained did not reject the null hypothesis H_0. In this particular setup query expansion performs well because the topics and the collection are semantically from the same domain. For cases where this level of agreement between queries and documents is missing and query expansion is not a viable option the proposed model can have more impact as seen in Experiment 2.

Therefore, in the context of **RQ6** in only a few settings α enables a marginal statistical significance.

Table 6.10: MAP results for Query Expansion Experiment for $\alpha = 0.60$

DLH13+QE	**EN**	**DE**	**EN-DE**	**DE-EN**
Baseline($\alpha = 0.0$)	35.53	40.99	37.56	34.46
$\mathbb{K}_{SM}$	36.11	41.09	$38.44^{\star}$	34.87
PL2 + QE	**EN**	**DE**	**EN-DE**	**DE-EN**
Baseline($\alpha = 0.0$)	37.00	42.17	37.56	35.51
$\mathbb{K}_{SM}$	37.29	42.22	$39.32^{\star}$	35.72
TF-IDF+QE	**EN**	**DE**	**EN-DE**	**DE-EN**
Baseline($\alpha = 0.0$)	31.91	37.10	34.57	31.01
$\mathbb{K}_{SM}$	32.69	37.09	35.33	$31.83^{\star}$

6.4.4 Experiment 4: Limited exploration restricting the representational contexts

Scope *Improving system effectiveness*

Objective *FCA retrieval model performance when considering limited exploration of the semantic model*

Hypothesis *The proposed FCA-based representation impacts positively on performance*

Test collection *GIRT collection and GIRT RDF graph*

Topics *Original CLEF DS 2004-2006 Topics*

Relevance assessment *CLEF Domain Specific Track qrels for DE and EN*

Effectiveness measure *Mean Average Precision*

This experiment explores how each type of relationship hierarchical or associational impacts the results produced by the prototype system. This is not possible to determine with classic IR models, but based on the suggested model at each run only neighborhoods generated by one type of relationship are considered. Intuitively, this experiment wants to assess the value of each type of relationship for retrieval in this context. These results cannot be generalized to state for example that navigating using related links will always produce better results than following narrower links, but it is a method of testing a resource used in conjunction with a system without users' interaction. On this collection the optimal combination are using formal concepts from the $\mathbb{K}_{skos:narrower}$ and $\mathbb{K}_{skos:related}$ formal contexts.

Tables 6.12, 6.13, and 6.14 describe the results obtained for restricting the relations considered to narrower and related, while Tables 6.15, 6.16, and 6.17 show the results for the case when only formal concepts from the core are used in the document and query representations. There is a small drop in the overall performance in each compared to the results in Experiment 2.

For a further closer look Table 6.11 demonstrates that for $\alpha = 0.60$ the $\mathbb{K}_{SM} = \mathbb{K}_{skos:broader} \cup \mathbb{K}_{skos:narrower} \cup \mathbb{K}_{skos:related}$ produces the statistically significant results and that by comparison the $\mathbb{K}_{skos:related}$ produces close MAP results.

Table 6.11: MAP results for restricted contexts

DLH13	**EN**	**DE**	**EN-DE**	**DE-EN**
Baseline($\alpha = 0.0$)	32.32	35.61	34.45	30.67
$\mathbb{K}_{skos:broader}$	32.24	36.64	35.05	31.01
$\mathbb{K}_{skos:narrower}$	33.28	37.35	35.60	31.40
$\mathbb{K}_{skos:related}$	32.94	37.90	35.65	31.51
$\mathbb{K}_{SM}$	$33.51^{\star}$	$38.03^{\star}$	$36.01^{\star}$	$31.68^{\star}$
PL2	**EN**	**DE**	**EN-DE**	**DE-EN**
Baseline($\alpha = 0.0$)	33.35	36.99	35.69	31.73
$\mathbb{K}_{skos:broader}$	33.43	38.13	36.44	31.97
$\mathbb{K}_{skos:narrower}$	34.11	38.46	36.94	32.32
$\mathbb{K}_{skos:related}$	33.86	39.19	36.87	32.34
$\mathbb{K}_{SM}$	$34.43^{\star}$	$39.47^{\star}$	$37.02^{\star}$	$32.59^{\star}$
TF-IDF	**EN**	**DE**	**EN-DE**	**DE-EN**
Baseline($\alpha = 0.0$)	29.68	33.03	32.38	27.89
$\mathbb{K}_{skos:broader}$	29.16	33.85	32.85	28.30
$\mathbb{K}_{skos:narrower}$	30.24	34.49	33.46	28.73
$\mathbb{K}_{skos:related}$	30.10	35.35	33.42	29.00
$\mathbb{K}_{SM}$	$30.43^{\star}$	$35.21^{\star}$	$33.96^{\star}$	$29.05^{\star}$

So far the results described are elucidating the first part of **RQ7** What is the impact of considering all semantic relations in the semantic model in comparison to a retricted set of formal contexts? In the context of Experiment 4, a combination of formal concepts from all formal contexts provides the best results, but a decision on what the union context $\mathbb{K}_{SM}$ is empirical and the SIR model in Chapter 5 describes the methodology behind it.

The answer for the second part of **RQ7** How do vectorial representations based on core formal concepts impact retrieval? is based on the last three tables (Table 6.15, Table 6.16, and Table 6.17) in this chapter. The MAP results for reduced formal concepts index show a small decrease in MAP compared to results in Table 6.12, Table 6.13, and Table 6.14, but a better performance for $\alpha = 1.0$ meaning that expanding document descriptions to include formal concepts outside the conceptual neighborhoods needs to be controlled either through a set threshold or a new weighting schemes for formal concepts. This is an aspect open to further research.

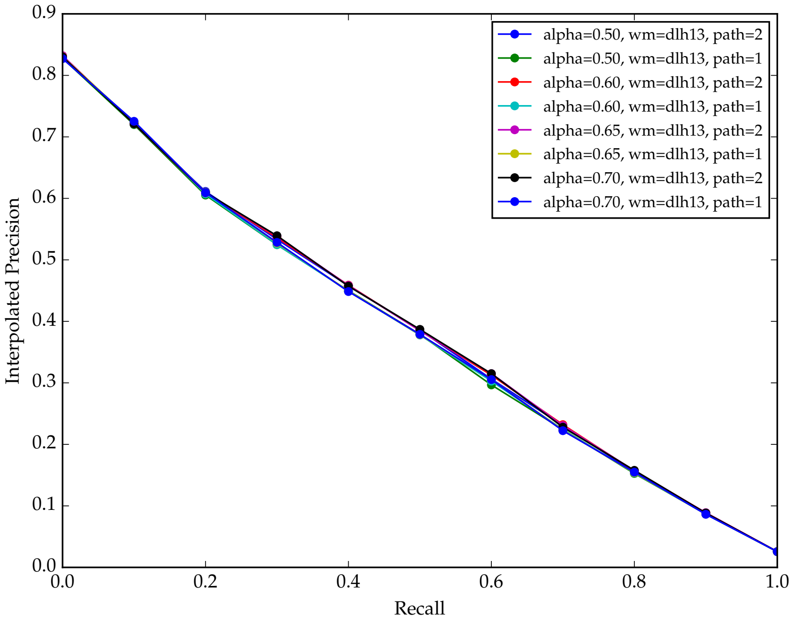


Figure 6.13: Monolingual German retrieval considering with parameters α and path length

Table 6.12: MAP results for weighting model DLH13, with $\alpha = 0.0$ as baseline, and $\mathbb{K}_{SM} = \mathbb{K}_{skos:narrower} \cup \mathbb{K}_{skos:related}$

WM=DLH13	**α=0.0**	**α=0.25**	**α=0.50**	**α=0.60**	**α=0.65**	**α=0.70**	**α=0.75**	**α=1.0**
EN	32.32	32.67*	33.31*	33.36*	33.28*	33.17	32.43	10.94
EN+QE	35.53	35.79*	36.08*	36.10	35.97	35.77	35.37	11.07
DE-EN	30.67	31.07*	31.50*	31.58*	31.60*	31.57*	31.40	13.41
DE-EN+QE	34.46	34.65*	34.81	34.75	34.68	34.40	34.09	14.34
DE	35.61	37.87*	38.28*	38.41*	38.44*	38.43*	38.34*	14.01
DE+QE	40.99	41.27*	41.47*	41.40*	41.27*	41.06	40.63	14.16
EN-DE	34.45	35.42*	35.85*	36.11*	36.18*	36.19*	36.18	13.11
EN-DE+QE	37.56	37.82*	38.18*	38.26*	38.29*	38.27	38.20	12.99

Table 6.13: MAP results for weighting model PL2, with $\alpha = 0.0$ as baseline, and $\mathbb{K}_{SM} = \mathbb{K}_{skos:narrower} \cup \mathbb{K}_{skos:related}$

WM=PL2	**α=0.0**	**α=0.25**	**α=0.50**	**α=0.60**	**α=0.65**	**α=0.70**	**α=0.75**	**α=1.0**
EN	33.35	34.00	34.21*	34.25*	34.19	34.09	33.74	11.73
EN + QE	37.00	37.18*	37.30	37.20	37.08	36.87	36.47	12.11
DE-EN	31.73	32.12*	32.54*	32.64*	32.64*	32.57*	32.27	11.00
DE-EN + QE	35.51	35.66*	35.81	35.70	35.58	35.37	34.95	11.89
DE	36.99	39.26*	39.55*	39.73*	39.79*	39.75*	39.60*	12.47
DE+QE	42.17	42.37*	42.45	42.34	42.21	42.05	41.40	12.64
EN-DE	35.69	36.69*	37.13*	37.38*	37.45*	37.50*	37.48*	14.31
EN-DE + QE	38.67	38.93*	39.22*	39.24	39.40	39.27	38.96	14.19

Table 6.14: MAP results for weighting model TF-IDF, with $\alpha = 0.0$ as baseline, and $\mathbb{K}_{SM} = \mathbb{K}_{skos:narrower} \cup \mathbb{K}_{skos:related}$

WM=TF-IDF	**α=0.0**	**α=0.25**	**α=0.50**	**α=0.60**	**α=0.65**	**α=0.70**	**α=0.75**	**α=1.0**
EN	29.68	30.04*	30.64*	30.33	30.25	30.00	29.64	13.51
EN + QE	31.91	32.23*	32.93*	32.89	32.79	32.23	31.77	13.99
DE-EN	27.89	28.50*	28.93*	29.01*	28.98*	28.80	28.68	13.30
DE-EN + QE	31.01	31.36*	31.78*	31.80*	31.69	31.37	30.94	11.89
DE	33.03	35.24*	35.60*	35.73*	35.74*	35.76*	35.52*	13.99
DE+QE	37.10	37.41*	37.54	37.46	37.30	37.11	36.75	14.14
EN-DE	32.38	33.21*	33.71*	33.90*	33.79*	33.78	33.64	15.99
EN-DE + QE	34.57	34.94*	35.22	35.31	35.41	35.29	35.16	15.93

Table 6.15: MAP results for reduced formal concepts index, weighting model DLH13, with $\alpha = 0.0$ as baseline, and $\mathbb{K}_{SM} = \mathbb{K}_{skos:narrower} \cup \mathbb{K}_{skos:related}$

WM=DLH13	**α=0.0**	**α=0.25**	**α=0.50**	**α=0.60**	**α=0.65**	**α=0.70**	**α=0.75**	**α=1.0**
EN	32.32	32.61*	32.94*	33.00*	32.95*	32.85	32.60	10.83
EN + QE	35.53	35.78*	35.98*	35.93	35.86	35.64	35.28	10.87
DE-EN	30.67	31.13*	31.64*	31.88*	31.96*	31.97*	32.01*	13.88
DE-EN + QE	34.46	34.88*	34.88*	34.93	34.86	34.73	34.55	14.66
DE	35.61	37.58*	37.97*	38.11*	38.13*	38.03*	37.93*	15.80
DE+QE	40.99	41.26*	41.50	41.44	41.34	41.15	40.77	15.94
EN-DE	34.45	35.37*	35.69*	35.84*	35.91*	35.90*	35.85*	12.03
EN-DE + QE	37.56	37.75*	37.99*	38.05	38.08	38.09	38.04	11.90

Table 6.16: MAP results for reduced formal concepts index, weighting model PL2, with $\alpha = 0.0$ as baseline, and $\mathbb{K}_{SM} = \mathbb{K}_{skos:narrower} \cup \mathbb{K}_{skos:related}$

WM=PL2	**α=0.0**	**α=0.25**	**α=0.50**	**α=0.60**	**α=0.65**	**α=0.70**	**α=0.75**	**α=1.0**
EN	33.35	33.65*	34.23*	34.26*	34.19	34.11	33.81	11.96
EN + QE	37.00	37.19*	37.27	37.23	37.10	36.89	36.49	12.31
DE-EN	31.73	32.21*	32.70*	32.80*	32.86*	32.85*	32.87*	11.69
DE-EN + QE	35.51	35.69*	35.90*	35.84	35.73	35.57	35.34	12.43
DE	36.99	38.90*	39.18*	39.28*	39.36*	39.32*	39.19*	13.30
DE+QE	42.17	42.35*	42.44	42.35	42.25	42.02	41.39	13.46
EN-DE	35.69	36.64*	37.00*	37.15*	37.18*	37.23*	37.21*	13.25
EN-DE + QE	38.67	38.89*	39.08*	39.07	39.12	39.08	38.79	13.14

Table 6.17: MAP results for reduced formal concepts index, weighting model TF-IDF, with $\alpha = 0.0$ as baseline, and $\mathbb{K}_{SM} = \mathbb{K}_{skos:narrower} \cup \mathbb{K}_{skos:related}$

WM=TF-IDF	**α=0.0**	**α=0.25**	**α=0.50**	**α=0.60**	**α=0.65**	**α=0.70**	**α=0.75**	**α=1.0**
EN	29.68	30.03*	30.21*	30.22	30.21	30.11	29.88	12.17
EN + QE	31.91	32.12	32.31	32.26	32.16	31.92	31.57	12.44
DE-EN	27.89	28.56*	29.08*	29.28*	29.38*	29.39*	29.31*	14.08
DE-EN + QE	31.01	31.37*	31.73*	31.83*	31.71*	31.46	29.69	14.19
DE	33.03	34.94*	35.30*	35.42*	35.45*	35.44*	35.20*	14.65
DE+QE	37.10	37.43*	37.64*	37.53	37.34	37.09	36.66	14.77
EN-DE	32.38	33.17*	33.57*	33.74*	33.79*	33.58*	33.40	13.53
EN-DE + QE	34.57	34.88*	35.05	35.05	35.02	34.98	34.89	13.47

6.5 Summary

This chapter has brought together all the theoretical elements from Chapter 5 and I devised a number of experiments aimed at investigating the impact of representing queries and documents as linear combinations of formal concepts on retrieval. Four different sets of experiments were run on a benchmark collection from CLEF:

- Experiment 1: Performance focused on query semantic annotation
- Experiment 2: FCA retrieval model compared to other IR models in monolingual and bilingual settings
- Experiment 3: FCA retrieval model compared to other IR models in monolingual and bilingual settings when using Query Expansion
- Experiment 4: FCA retrieval model performance when considering limited exploration of the semantic model

Each of the experiments contributes towards building a clearer picture of the impact of this retrieval model. In retrieval evaluation is not possible to speak in absolute terms about a model, but for the given setup and by following the theoretical description in the previous chapter statistically significant results were obtained in both monolingual and bilingual settings when no query expansion methods were used. The difficult aspect of testing this model is the multitude of requirements: a document collection with trusted semantic enrichments from a semantic model i.e. a KOS, which is available in RDF/SKOS format, a set of queries from the domain of the KOS, and relevance judgements for each of these queries. This is actually one of the very few examples of evaluation of an FCA-based IR model at scale (an exception is Abdulahhad et al. [2013] but experiments are limited to TF-IDF as matching model, which is not a high-performing weighting model as PL2 or DLH13).

Chapter 7

Conclusions

7.1 Research Questions Summary

For research question RQ1-RQ7 the following conclusions were reached:

7.1.1 RQ1

What aspects, more specifically levels of detail of a Knowledge Organization System's representations of meaning are relevant to retrieval processes? How can the lexical bias of KOS resources for its main language (in most cases English) be remedied and more lexical details automatically created for other languages using the cross-schema links between resources in the LLOD cloud?

For the first part of this research question, based on the exploration in Chapter 3 the eligible SKOS datasets for the retrieval application scenario need to be *relationship list type of resources* that incorporate three levels of specification: *conceptual* (relations between concepts), *terminological* (relations between concepts and labels), and *lexical* (relations between labels). An extra constraint is that the required third level has to have a similar amount of detail across all languages to prevent biases in performance.

The algorithms in Section 3.4 address the second part of this RQ. The proposed algorithms can be used to add more lexical detail automatically for an existing resource. The goal of the algorithms is the construction of concepts' signatures in all languages supported by the chosen KOS, and for them to be used as basis for NLP processes like matching concepts to text beyond the identification of concept labels.

7.1.2 RQ2

Considering Formal Concept Analysis as the framework for interpreting the information provided by the semantic model, and that queries and documents are annotated with concepts from the semantic model, what is a suitable representation that is expressive *enough to capture the connections between documents through their annotating concepts from the semantic model and that* maximizes the exploitation of the semantic model at both lexical and knowledge level?

Chapter 5 described the process of mapping documents and queries into vectorial linear combinations of weighted formal concepts, where the matrix M^{SM} captures the possible pathways between concepts in the semantic model. This representation can be viewed as a concept-based approximation of each document and query.

$$d \xrightarrow[\mathbb{K}_{SM}]{c_1,c_2,..c_k} \bigcup \gamma(c_i) \cup \mu(c_i)$$

$$\vec{d}^{SM} = M^{SM}\vec{d}$$

The retrieval model presented is an instance of the Generalized Vector Space Model with $G^{SM} = M^{SM}(M^{SM})^{\top}$. It captures the dependencies between concepts based on the union of formal contexts $\mathbb{K}_{SM}$ derived from the SM with the support of FCA's mathematical framework. Therefore, the model fullfils the *expressiveness* requirement of capturing the connections between documents through their annotating concepts from the semantic model.

By partioning the semantic model based on information from the terminological level (the *functional formal contexts*) and the conceptual level (the *relational formal contexts*) and their operational role in the flow of IR processes, this SIR model is set up to *maximize the exploitation of the semantic model*.

The model is flexible and the KOS-based semantic model can be replaced with other knowledge resources from the LOD as long as its structure and content are relevant to information retrieval. Formal contexts can be built to reflect a search path reflected by patterns in query sessions and the documents indexed by more complex search paths. The objects and attributes would be selected based on the beginning and the endpoint of the path in the knowledge resource.

7.1.3 RQ3

Given the document collection is pre-annotated with concepts from the semantic model, does query expansion with concept labels from the semantic model based on three distinct methods: a) implicit annotation, b) explicit annotation, and c) pseudo-relevance annotation improve retrieval where the baseline is provided by query reformulation based on a local method (weighted terms from top-ranked documents)?

In Experiment 1 in Section 6.4.1 I investigate the impact of expanding queries based on the lexicalization of the concepts that annotate the queries. Three methods are proposed: a) *explicit annotation* equivalent to matching concepts to text by label; b) *implicit annotation* where a text's topic is identified based on the similarity between the query and the concept's signature; and c) *pseudo-relevance annotation* where the queries are expanded based on the labels of the concepts annotating the top-ranking documents. The experiments showed a mixed picture demonstrating that query expansion using just concept lexicalization and ignoring concept relations does not lead to improved results, despite the different approaches. Therefore for this research question the conclusion is that query expansion with concept labels does not improve the baselines. Yet, it is possible that collection dependent weighting models and adjustments could further ameliorate the results in this setting.

7.1.4 RQ4

Queries and documents are represented as mixed vectors of weighted terms and formal concepts constructed from the semantic model, what is the impact of this representation and ranking parameter α in comparison to existing vectorial represention based only on term-weighting models such as TF-IDF, DLH13, and PL2?

In Chapter 6 through the various experimental setups the behavior of the new model is investigated. For the benchmark setup and by following the theoretical description in Chapter 5, the results obtained were statistically significant in both monolingual and bilingual settings when no methods for query expansion where used. The difficult aspect of testing this model is the multitude of requirements: a document collection with trusted semantic enrichments from a semantic model i.e. a KOS, which is available in RDF/SKOS format, a set of queries from the domain of the KOS, and relevance judgements for each of these queries.

Based on the results of the second experiment in Section 6.4.2, it is possible to generalize that the ranking function consistently augments the performance of classic retrieval models, but depends on the parameter α.

7.1.5 RQ5

The weighted formal concepts part of the defined vectors representing text are language independent, does the bilingual setting outperform machine translation as the baseline?

The representations created for documents and queries have a language independent component. In the proposed model the aim is to actively avoid translation and focus on mapping text to concepts from the semantic model. As seen in Tables 6.7, 6.8, and 6.9 for α between 0.25 and 0.70 statistically significant improvements were obtained. The baseline in this case is given by the machine translation of the queries using Google Translate. This is a strong baseline achieving 95% of monolingual performance.

7.1.6 RQ6

Considering an effective query expansion method how does the formal concept based-expansion of queries and documents perform in comparison?

Query expansion is a two step technique. The first step is to extract a set of feedback documents from the first submission of the query. In this case the first three documents. In the second step, all terms are ranked in descending order of their $tf \cdot idf$ weights and a fixed number of them (in this case 10) are added to the query to be re-submitted for search. Query expansion does not always perform well, if the feedback documents cover a wide variety of topics, indadvertedly introducing noise in the results set He and Ounis [2009].

H_0: The systems producing the two runs have the same retrieval characteristics and any difference between the runs occurred by random chance.

H_1: The FCA-based retrieval system in implementation outperforms the classic IR models when using Query Expansion.

In this case, the results obtained did not reject the null hypothesis H_0. For this particular setup query expansion performs well because the topics and the collection are semantically from the same domain, while for cases where this level of agreement between queries and documents is missing and query expansion is not a viable option the proposed model can have more impact as seen in Experiment 2 from Chapter 6.

7.1.7 RQ7

What is the impact of considering all semantic relations in the semantic model in comparison to a retricted set of formal contexts? How do vectorial representations based on core formal concepts impact retrieval?

On the first part of **RQ7** in the context of Experiment 4 I identified that a combination of formal concepts from all formal contexts provides the best results. The decision on what the union context $\mathbb{K}_{SM}$ is, is in my view, empirical and the SIR model in Chapter 5 describes the methodology behind it. It is to be expected that KOS resources that have a small hierarchical structure would not benefit from using all contexts that can be derived from the semantic model.

The answer for the second part of **RQ7** is based on the last three tables (Table 6.15, Table 6.16, and Table 6.17) from Chapter 6. The MAP results for reduced formal concepts index show a small decrease in MAP compared to results in Table 6.12, Table 6.13, and Table 6.14, but a better performance for $\alpha = 1.0$ meaning that expanding document descriptions to include formal concepts outside the conceptual neighborhoods needs to be controlled either through a set threshold or a new weighting schemes for formal concepts. This is an aspect open to further research.

7.2 Contribution to knowledge

Our main objective for this research was to investigate *the relationships between the meanings' representations captured by Knowledge Organization Systems expressed as SKOS datasets deployed on the Semantic Web, and their role and potential in augmenting existing retrieval models effectiveness*.

I defined a new semantic retrieval model that formalizes the investigation of KOSs' impact on retrieval effectiveness with Formal Concept Analysis, providing the mathematical toolbox to interpret KOSs semantics. By using two types of formal contexts: *functional* and *relational* the model captures the definitional and relational aspects of concepts. Each formal context is assigned an operational role in the flow of processes of a retrieval system enabling a clear path towards implementations.

In summary, its main characteristics are:

1. It exploits the rich conceptualizations available within the semantic model;

2. It is not restricted to a particular domain, with the possibility to map documents and queries to any number of semantic models;

3. A semantic model is not necessarily completely describing the domain, and the mixed ranking metric accounts for this by using keyword-based retrieval models as a fallback mechanism;

4. It is a model that scales even when large knowledge bases are used, because of the process of building clusters using Formal Concept Analysis;

5. The interlinked nature of semantic models deployed on the Semantic Web allows extracting missing knowledge from one semantic model to support processes for another such as annotation or disambiguation;

6. The explicit methods used for connecting a semantic model to a document collection can be used as basis for the standardization of evaluation measures and benchmarks for semantic retrieval systems operated within a similar scope with the one defined in Section 1.2.2;

7. It allows to identify and quantify structural characteristics of the KOS models used that correlate with the observed retrieval performance in practical applications.

7.3 Points of difference from existing FCA's use in IR

A recent survey on Poelmans et al. [2011] titled *FCA-Based Information Retrieval Research* showed that since 1982, only 103 papers where published on this subject. The authors clustered the articles based on the different aspects of retrieval these papers aimed to solve. Four areas are relevant to this thesis: *Knowledge Representation and Browsing with FCA*, *Query Result Improvement with FCA*, *Domain Knowledge in Search results*, and *Image, Software and Knowledge Base Retrieval*.

Each of these groups has chosen to use the retrieval models that best served their application context. For the first group, documents and their semantic annotations were used to build a large formal context. Afterwards, the corresponding concept lattice was derived. A query is approximated by a formal lattice in the navigational structure of the concept lattice. Browsing becomes moving along the hierarchy of the concept lattice. For the second group formal contexts are built on the fly with the query as the intent and the search results documents as the extent. Again, the corresponding concept lattice is computed and the user navigates this new structure instead of a flat list, being able to fine tune the query by exploring upper and lower neighbors. In the case of domain knowledge in search results, FCA is used for the hierarchy and presentation of results.

In the last group, researchers use existing IR models and concentrate on creating a query's formal context to support query refinement operations closely coupled to the search terms used Ducrou and Eklund [2007]. The construction of this conceptual space (i.e. formal context) relies both on search results obtained at a first pass and on the knowledge base information.

In conclusion the points of difference with the proposed model of this thesis are:

FCA is used to construct concept lattices for the multiple contexts incorporated within a KOS, where each semantic relation (skos:broader, skos:narrower, skos:related, skos:exactMatch, etc.) determines a context.

FCA is used to guide the process of building conceptual signatures instrumental in annotation and disambiguation processes.

Indexing: queries and document collections are expressed as linear combinations of weighted formal concepts.

Matching: queries and documents are matched based on their conceptual overlap.

Ranking: a mixed metric connecting filtered documents to the ranking results from a classic retrieval model.

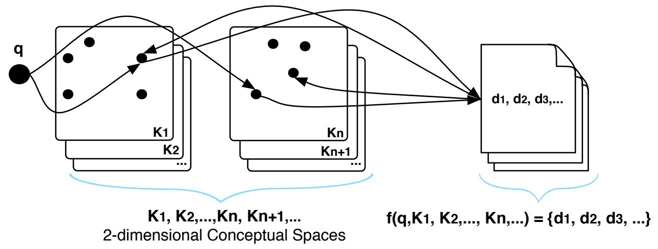

q
K1
K2
...
Kn
Kn+1
...
d1, d2, d3,...
K1, K2,...,Kn, Kn+1,...
2-dimensional Conceptual Spaces
f(q,K1, K2,..., Kn,...) = {d1, d2, d3, ...}

Figure 7.1: From query to documents, a transformation process driven by conceptual spaces

7.4 From query to documents, the transformation process

An idea inherited from Mooers [1958] on the theory of information retrieval is that the retrieval system transforms a query, the user's *prescription*, into a set of documents. The question is: what drives this transformation? For semantic information retrieval type of models this transformation is driven, in my view, by the content and structure of the semantic models.

In Figure 7.1 the query q is a point in space projected on different formal contexts (two-dimensional concept spaces) $K_1, K_2, ..., K_n$. All the projection points (formal concepts) determine the document set relevant to the query.

The proposed approach is unique in explicitly interpreting a semantical reference as a pointer to a concept in the semantic model that activates all its immediate linked concept neighbors. And most importantly it is the formalization of the IR model and the integration of knowledge resources from the LLOD that is distinctive from other approaches. The pre-processing of the semantic model using Formal Concept Analysis enables the creation of two-dimensional concept spaces (formal contexts) that extract sub-graphs of the original structure of the semantic model. The type of conceptual spaces built in my case was limited by the KOSs semantic relations relevant to retrieval: *exact match*, *broader*, *narrower*, and *related*.

Gärdenfors [2014] in his recent book *The Geometry of Meaning* describes *conceptual spaces as a representational level that serves as an anchoring mechanism between language and reality*. As humans we use language and each symbolic description (sequence of words) activates the triangle of meaning for concepts in one of several conceptual spaces. Three cognitive processes take place: linking to a concept, determining the properties associated with that concept, and connecting it to its neighboring concepts.

The retrieval model defined in this thesis matches at computational level these steps, with the distinction that the conceptual spaces extracted from KOS resources are naive versions of what the human brain encodes.

7.5 Outlook

7.5.1 Open Question

On a more pragmatic note, I believe that the current IR models operate fairly successfully at scale, but what is missing is a retrieval model that can use knowledge resources in an analytical way similar to the model introduced in this thesis.

For many, this translates into Semantic Web Search in the sense that the world can be described through the accumulation of a linked set of facts and search is equivalent to clustering *finding all the data* linked to a query.

In contrast, I see this just as an incremental step towards IR models that operate with knowledge, not data. The idea is poignantly expressed by Veltman [2006]: *We need bridging and mapping devices that allow us to move dynamically through different languages, different levels of vocabularies, different chronologies (in the sense of time systems), different cartographical methods and policies (such that we can see how maps of a country such as Poland not only change with time but also differ from those of Russia or Germany for the same area). Such dynamic lists of knowledge will allow us to trace changes of interpretation over time, have new insights and help us to discover new patterns in knowledge.*

To address this problem, a potential refinement would be possible if a time component is encoded with the concept's specification in the semantic model, it would then be possible to construct conceptual spaces restricted to a particular segment in time, and this could lead to time-sensitive retrieval relevant for search in cultural heritage collections.

7.5.2 Future Research

In 2013, I participated[1] in the CHiC pilot lab evaluation campaign Petras et al. [2013]. This was setup to test ad-hoc multilingual retrieval systems and techniques for semantic enrichments. The test collection was the Europeana's Cultural Heritage content from 2012 of 23,300,932 documents. The aim was to experiment with a large collection and establish a good baseline setup for future experiments pursuing the open question above.

Europeana already enriches about 30% of its metadata objects with concept names and places. It uses the following vocabularies for its semantic enrichments: GeoNames[2] for geospatial information, GEMET, and DBPedia. At this stage the collection needs to have a larger distribution of semantic enrichments to allow further evaluation of our prototype system. Europeana's collection is being continuously improved and it will be a relevant use case study to advance this research.

7.5.3 Closing Remarks

In this thesis, I brought together several strands of research from information retrieval, natural language processing, and technologies for the Semantic Web. I considered the growing number of knowledge and language resources published on the Semantic Web platform and investigated how to connect a *semantic model* constructed from Semantic Web resources that explicitly define meanings for an information retrieval system.

I was motivated by the growing claims that by building a large, distributed, and shared space of language and knowledge resources using Semantic Web technologies, it is possible to create semantically-aware applications and in particular better information retrieval systems.

[1] The ranked outcome of the participation is part of this overview paper located at: http://www.clef-initiative.eu/documents/71612/82b4444b-9a3c-4d8e-a986-6a184012991e

[2] http://www.geonames.org/ontology/documentation.html

After our investigation, I believe that the process of encoding meaning needs further refinements and for the LLOD to fulfill its support role, explicit mechanism for quality check should be put in place, going beyond RDF syntax checks.

Formal Concept Analysis is a natural candidate to connect the data layer and the application layer. Currently in FCA, further techniques and tools for knowledge discovery are being developed (e.g. FCART[1]) and scalability of formal contexts manipulation is of high priority.

In the larger context, semantics is dynamic by nature, thus each step taken towards understanding how resources like KOSs in a simplified representation impact retrieval applications will feed back into the KOSs lifecycle.

[1] http://ami.hse.ru/issa/Proj_FCART

Appendix A: TheSoz, from thesaurus to SKOS dataset

In Zapilko and Sure [2009] the authors described how TheSoz was re-encoded in the SKOS format. I extract here some of the most relevant aspects of this transformation. Table A1 gives an overview of TheSoz as an RDF dataset after this transformation.

The main structural characteristics of the TheSoz specified in Zapilko and Sure [2009] are captured in Table together with the detailed correspondence between predicates supported by TheSoz and SKOS classes, properties, and relationships. This uncovers that each LLOD resource though available in a simplified format like SKOS still maintains its initial complexity. Therefore SPARQL constructs like **ASK** or **DESCRIBE** do very little in determining the semantics of the data. It is therefore fundamental to review the Semantic Web Stack to add a mediating concept analysis level with a suitable mathematical toolbox. This thesis points towards FCA to enable application developers to analyze the data without breaking the link

Table A1: TheSoz VoID description summary

source	http://lod.gesis.org/thesoz/
author	GESIS - Leibniz Institute for the Social Sciences
links:agrovoc-skos	846 (840 exact matches, 6 close matches)
links:dbpedia	5024 (all exact matches)
links:stw-thesurus-for-economics	4927 (2844 exact matches, 631 related matches, 1418 broad matches, 34 narrow matches)
namespace	http://lod.gesis.org/thesoz/
triples	425124

between the original thesaurus encoding of knowledge and the LLOD data manipulated.

Description TheSoz contains overall about 11,600 keywords and covers all topics and sub-disciplines of the social sciences. Additionally terms from associated and related disciplines are included in order to support an accurate and adequate indexing process of interdisciplinary, practical-oriented and multi-cultural documents.

Thesaurus characteristics The Thesaurus for the Social Sciences contains about 12,000 keywords, of which more than 8000 are descriptors (authorized keywords) and about 4000 are non-descriptors. Relationships between these keywords are expressed as broader, narrower or related terms as well as there are also “use instead” and “use combination” relations and their counterparts (“used for” and “used for combination”). Additionally a classification hierarchy is provided and each thesaurus term is dedicated to one or more classification terms. The TheSoz contains a special type of non-descriptor called “AD” (for alternative descriptor) which differs from the international standard norms for thesauri and holds more than one “use instead” and/or “use combination” relation at the same time for general or ambiguous terms. There are about 200 of such “AD” terms in the TheSoz.

An exception In case of the TheSoz the “use combination” relation, when a term is defined as the combination of other two terms in the thesaurus, has been modeled via grouping the affected terms as multiple *skos:member* in a *skos:Collection*. This is than included in one *skos:prefLabel*. But as mentioned above, the TheSoz also contains a special type of non-descriptor called “AD” which holds more than one “use instead” and/or “use combination” relation at the same time. Modeling such a term to SKOS would invoke more than one *skos:prefLabel* in one single concept. Therefore these relations were modeled backwards via their “used for” and/or “used for combination” relations in the associated descriptors and a small loss of information could not be avoided with this solution. To avoid a complete loss of this relevant information these relations were included in additional *skos:editorialNotes* until there is a satisfying way to model them correctly with SKOS.

The SKOS version of the thesaurus contains two types of URIs, one for the descriptors and non-descriptors of the thesaurus and one for the terms of the classification hierarchy.

Table A2: Detailed description of TheSoz's structural relations and correspondence to SKOS specification

Extension	**Description**
thesoz:Descriptor	Descriptors of the TheSoz, which are defined as subclasses of skos:Concept.
thesoz:Classification	Notation of the classification hierarchy of the TheSoz, which is defined as a subclass of skos:Concept.
thesoz:EquivalenceRelationship	An equivalence relationship between two terms, where the terms are assigned via thesoz:use and thesoz:usedFor properties. This is a subclass of skosxl:Label.
thesoz:CompoundEquivalence	A compound equivalence between terms. For constructing "use combination" and "used for combination" relations between terms. The non-preferred term is assigned by the thesoz:compoundNonPreferrdTerm property, the preferred terms by the thesoz:preferredTermComponent property. This is a subclass of skosxl:Label.
thesoz:use	Use relation, which is defined as a subproperty of skosxl:labelRelation.
thesoz:usedFor	Used for relation, which is defined as a subproperty of skosxl:labelRelation.

Continued on next page

Table A2 – *Continued from previous page*

Extension	**Description**
thesoz:preferredTermComponent	A preferred term as a component for a "use combination" and "used for combination" relation. This property is defined as a subproperty of skosxl:labelRelation.
thesoz:compoundNonPreferredTerm	The non-preferred term as a component for a "use combination" and "used for combination" relation. This property is defined as a subproperty of skosxl:labelRelation.
thesoz:isPartOfEquivalenceRelationship	Relation from a term to the class thesoz:EquivalenceRelationship.
thesoz:isPartOfCompoundEquivalence	Relation from a term to the class thesoz:CompoundEquivalence.
thesoz:hasTranslation	Relation between different languages of a term, which is defined as a subproperty of skosxl:labelRelation.
thesoz:isTranslationOf	Inverse property of thesoz:hasTranslation.

Appendix B: CHIC 2013 Lab Report

Using the Divergence from Randomness Framework The following lab notes describe our experiments for the multilingual ad-hoc retrieval task organized by PROMISE (Participative Research Laboratory for Multimedia and Multilingual Information Systems Evaluation). The task involved retrieving relevant documents from the CHiC multilingual Europeana collection for the 50 topics provided in 13 languages. For this first participation to the CHiC Lab, we focused on understanding the challenges of working with a collection of cultural heritage objects with short textual descriptions and on how to fine-tune a set of weighting models from the probability models based on Divergence From Randomness (DFR) Amati and Van Rijsbergen [2002] to perform uniformly in monolingual and multilingual scenarios. The official runs submitted used PL2 as the retrieval model and query expansion for four monolingual runs for English and Italian, and two multilingual runs against an English-Italian collection. Our best results were obtained in the unofficial runs using DLH13 with stemming and stopwords removal.

In the next sections we present a summary of retrieval results and the combination of experimental settings we worked with. The results obtained in the official runs are modest, with substantial improvements in the unofficial runs that use DLH13.

Experimental Setup The retrieval models we chose for these experiments are PL2 and DLH13. They are DFR models obtained by instantiating the three components of the framework: selecting a basic randomness model, applying the first normalization and than normalizing the term frequencies. The mathematical formulas Macdonald et al. [2005] describe that terms with informative value abide by the distributional rule *the more the divergence of the within-document term-frequency from its frequency within the collection, the more the information carried by the word t in the document d* [1]. Our decision to consider DFR models was also based on the results reported by Akasereh et al. [2012], where similar retrieval performances are obtained across languages with DFR models.

PL2 weighting model a Poisson model with Laplace after-effect and second normalization for resizing the term frequency by document length.

DLH13 weighting model – a generalization of the hypergeometric model in a binomial case (parameter free):

$$score(d,Q) =$$

$$\sum_{t\in Q} qtw \cdot \frac{1}{tf+0.5} \cdot \left(\log_2 (\frac{tf \cdot avg\_l}{l} \cdot \frac{N}{F}) + (l-tf)\log_2(1-f) + 0.5 * \log_2\left(2\pi tf(1-f)\right)\right) \tag{7.1}$$

where the normalized term frequency is:

$$tfn = tf \cdot \log_2(1 + c \cdot \frac{avg\_l}{l}) \tag{7.2}$$

We used only two of the 13 collections made available: the English collection with 1107176 documents and the Italian Collection with 2120059 documents. Prior experiments at CHiC were performed using Lucene, Solr, Indri, or Cheshire Petras et al. [2012], while in this setup we used Terrier Retrieval Platform Ounis et al. [2005]. After indexing, using the English tokeniser, respectively the UTF tokeniser we obtained two indexes. The English index had 338248 index terms, while the Italian had 274009 index terms, with a much larger number of tokens for Italian.

[1] http://terrier.org/docs/v3.5/dfr_description.html

Notations:
tf is the within-document frequency of t in d
avg_l is the average document length in the collection
l is the document length of d, which is the number of tokens in d
N is the number of document in the whole collection
F is the term frequency of t in the whole collection
nt is the document frequency of t
tfn is the normalized term frequency given by relation 7.2,
where c is a free parameter
λ is the variance and mean of a Poisson distribution. It is given by F/N and F is much smaller than N
qtw is the query term weight given by qtf/qtf_{max}
qtf is the query term frequency and qtf_{max} is the maximum query term frequency among the query terms

Table A3: CHiC Ad-Hoc Multilingual Official Runs

Model	Query Expansion	Stemming	Stopwords	Run	MAP
PL2	-	x	x	EN-EN	4.82
PL2+Bo1	x	x	x	EN-EN	4.75
PL2	-	x	x	IT-IT	2.55
PL2+Bo1	x	x	x	IT-IT	2.89
PL2	-	x	x	EN - Mixed EN/IT	6.30
PL2+Bo1	x	x	x	IT - Mixed EN/IT	5.97

Official Runs Our results presented in Table A3 are also described in finer detail in Ferro and Masiero [2013]. The MAP was computed for the multilingual scenario, where a topic is in one source language and the relevant documents can be in any of the different language collections. We noticed that the query expansion did not always have a positive impact on performance. This is a known issue with query expansion only working well for queries which have a good top-ranked document set returned by the first-pass retrieval. Also, based on query average precision 10 topics from the name topic category had precision zero in the Italian runs (e.g. *isola di madeira, isole falkland,sesame street*).

Overall, our submission is slightly worse than the 5th best result obtained in the multilingual ad-hoc evaluation (MAP 6.43%) and the results submitted only used the English and Italian document collections. We merged

Table A4: Summary Results of the Monolingual EN & IT Unofficial Runs

Model	Query Expansion	Stemming	Stopwords	Query Enrichment	MAP_{EN}	MAP_{IT}
DLH13	-	-	-	-	**36.25**	8.42
DLH13	x	-	-	-	34.97	7.45
DLH13	-	-	-	x	25.76	6.08
DLH13	x	-	-	x	25.44	6.49
DLH13	-	x	x	-	35.19	32.44
DLH13	x	x	x	-	33.75	29.34
DLH13	-	x	x	x	25.87	24.09
DLH13	x	x	x	x	25.70	21.43

the result lists from monolingual retrievals and ordered them based on the $score(d, Q)$ values. This was possible in this instance because the collections had a comparable number of terms.

Monolingual Explorations The PL2 is a parametric model, so the parameter we set a-priori could not be tuned without relevance assessments, and for a second set of experiments we opted for the DLH13 weighting model a parameter-free weighting model, with all its variables being set automatically from the collection statistics.

In the unofficial runs, we varied the conditions for each of them by using light NLP processing (stemming, stopwords removal), query expansion, and query enrichment by adding new terms for each query based on Google's auto-complete feature.

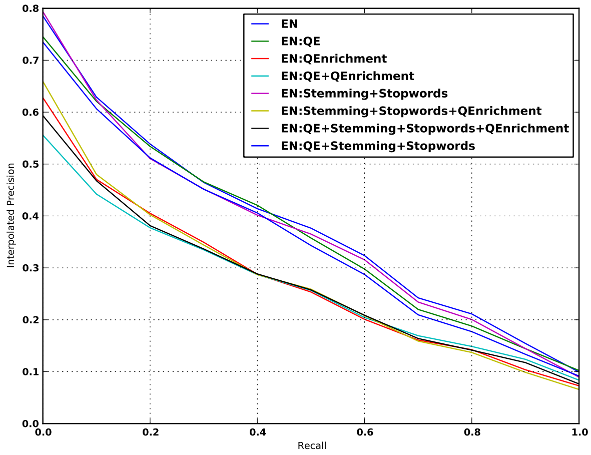


Figure A2: CHiC Ad-Hoc EN Monolingual

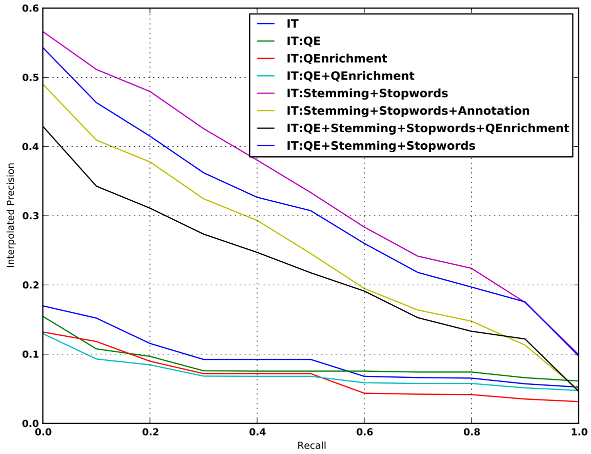


Figure A3: CHiC Ad-Hoc IT Monolingual

Across the different setups (see Figure A2 and Figure A3), we noticed that the stemming and stopwords removal with DLH13 produces the most consistent results. We repeated the multilingual retrieval obtaining an improved MAP of 8.73% with only topic CHIC–91 (*navi di colombo*) having precision zero, an elusive query-topic with a 1.86 mean statistics for the number of relevant retrieved documents.

Conclusions The CHiC Lab 2013 Ad-Hoc Multilingual Task allowed us to experiment with two probabilistic models from the DFR family. The DLH13 outperformed PL2 in this instance, but with further tuning of the parameters for PL2 this could be reversed. We will continue to further our work using the topics and the Europeana collection having acquired the necessary baseline experience to expand to more languages from the collection.

Precision at 1 : 0.6400
Precision at 2 : 0.6600
Precision at 3 : 0.6467
Precision at 4 : 0.6250
Precision at 5 : 0.5920
Precision at 10 : 0.5380
Precision at 15 : 0.5013
Precision at 20 : 0.4740
Precision at 30 : 0.4400
Precision at 50 : 0.3848
Precision at 100 : 0.3066
Precision at 200 : 0.2148
Precision at 500 : 0.1146
Precision at 1000 : 0.0665

Average Precision: 8.73

Precision at 0%: 1.4081
Precision at 10%: 0.6428
Precision at 20%: 0.2661
Precision at 30%: 0.1178
Precision at 40%: 0.0436
Precision at 50%: 0.0082
Precision at 60%: 0.0000
Precision at 70%: 0.0000
Precision at 80%: 0.0000
Precision at 90%: 0.0000
Precision at 100%: 0.0000

R-Precision: 14.30

Figure A4: CHiC Ad-Hoc Multilingual using DLH13, stemming, stopwords removal from EN, IT collections

Appendix C: SKOS and SKOS-XL Axiom Specification

This is a summary of SKOS and SKOS-XL Axiom Specification based on the report *Key Choices in the Design of SKOS* Baker et al. [2013], which contains an extensive presentation of the decisions in including or excluding certain SKOS components in the final W3C recommendation.

Table A5: SKOS Class and Property Definition Axioms

Axiom	**Content**
S1	skos:Concept is an instance of owl:Class.
S2	skos:ConceptScheme is an instance of owl:Class.
S3	skos:inScheme, skos:hasTopConcept and skos:topConceptOf are each instances of owl:ObjectProperty.
S4	The rdfs:range of skos:inScheme is the class skos:ConceptScheme.
S5	The rdfs:domain of skos:hasTopConcept is the class skos:ConceptScheme.
S6	The rdfs:range of skos:hasTopConcept is the class skos:Concept.
S7	skos:topConceptOf is a sub-property of skos:inScheme.
S8	skos:topConceptOf is owl:inverseOf the property skos:hasTopConcept.
S10	skos:prefLabel, skos:altLabel and skos:hiddenLabel are each instances of owl:AnnotationProperty.
S11	skos:prefLabel, skos:altLabel and skos:hiddenLabel are

Continued on next page

Table A5 – *Continued from previous page*

Axiom	**Content**
	each sub-properties of rdfs:label.
S12	The rdfs:range of each of skos:prefLabel, skos:altLabel and skos:hiddenLabel is the class of RDF plain literals.
S15	skos:notation is an instance of owl:DatatypeProperty.
S16	skos:note, skos:changeNote, skos:definition, skos:editorialNote, skos:example, skos:historyNote and skos:scopeNote are each instances of owl:AnnotationProperty.
S17	skos:changeNote, skos:definition, skos:editorialNote, skos:example, skos:historyNote and skos:scopeNote are each sub-properties of skos:note.
S18	skos:semanticRelation, skos:broader, skos:narrower, skos:related, skos:broaderTransitive and skos:narrowerTransitive are each instances of owl:ObjectProperty.
S19	The rdfs:domain of skos:semanticRelation is the class skos:Concept.
S20	The rdfs:range of skos:semanticRelation is the class skos:Concept.
S21	skos:broaderTransitive, skos:narrowerTransitive and skos:related are each sub-properties of skos:semanticRelation.
S22	skos:broader is a sub-property of skos:broaderTransitive, and skos:narrower is a sub-property of skos:narrowerTransitive.
S23	skos:related is an instance of owl:SymmetricProperty.
S24	skos:broaderTransitive and skos:narrowerTransitive are each instances of owl:TransitiveProperty.
S25	skos:narrower is owl:inverseOf the property skos:broader.
S26	skos:narrowerTransitive is owl:inverseOf the property skos:broaderTransitive.
S28	skos:Collection and skos:OrderedCollection are each instances of owl:Class.
S29	skos:OrderedCollection is a sub-class of skos:Collection.
S30	skos:member and skos:memberList are each instances of owl:ObjectProperty.
S31	The rdfs:domain of skos:member is the class skos:Collection.
S32	The rdfs:range of skos:member is the union of classes skos:Concept and skos:Collection.

Continued on next page

Table A5 – *Continued from previous page*

Axiom	**Content**
S33	The rdfs:domain of skos:memberList is the class skos:OrderedCollection.
S34	The rdfs:range of skos:memberList is the class rdf:List.
S35	skos:memberList is an instance of owl:FunctionalProperty.
S36	For any resource, every item in the list given as the value of the skos:memberList property is also a value of the skos:member property.
S38	skos:mappingRelation, skos:closeMatch, skos:exactMatch, skos:broadMatch, skos:narrowMatch and skos:relatedMatch are each instances of owl:ObjectProperty.
S39	skos:mappingRelation is a sub-property of skos:semanticRelation.
S40	skos:closeMatch, skos:broadMatch, skos:narrowMatch and skos:relatedMatch are each sub-properties of skos:mappingRelation.
S41	skos:broadMatch is a sub-property of skos:broader, skos:narrowMatch is a sub-property of skos:narrower, and skos:relatedMatch is a sub-property of skos:related.
S42	skos:exactMatch is a sub-property of skos:closeMatch.
S43	skos:narrowMatch is owl:inverseOf the property skos:broadMatch.
S44	skos:relatedMatch, skos:closeMatch and skos:exactMatch are each instances of owl:SymmetricProperty.
S45	skos:exactMatch is an instance of owl:TransitiveProperty.

Table A6: SKOS Integrity Condition Axioms

Axiom	**Content**
S9	skos:ConceptScheme is disjoint with skos:Concept.
S13	skos:prefLabel, skos:altLabel and skos:hiddenLabel are pairwise disjoint properties.
S14	A resource has no more than one value of skos:prefLabel per language tag.
S27	skos:related is disjoint with the property skos:broaderTransitive.
S37	skos:Collection is disjoint with each of skos:Concept and skos:ConceptScheme.
S46	skos:exactMatch is disjoint with each of the properties skos:broadMatch and skos:relatedMatch.

Table A7: SKOS XL Axioms

Axiom	**Content**
S47	skosxl:Label is an instance of owl:Class.
S48	skosxl:Label is disjoint with each of skos:Concept, skos:ConceptScheme and skos:Collection.
S49	skosxl:literalForm is an instance of owl:DatatypeProperty.
S50	The rdfs:domain of skosxl:literalForm is the class skosxl:Label.
S51	The rdfs:range of skosxl:literalForm is the class of RDF plain literals.
S52	skosxl:Label is a sub-class of a restriction on skosxl:literalForm cardinality exactly 1.
S53	skosxl:prefLabel, skosxl:altLabel and skosxl:hiddenLabel are each instances of owl:ObjectProperty.
S54	The rdfs:range of each of skosxl:prefLabel, skosxl:altLabel and skosxl:hiddenLabel is the class skosxl:Label.
S55	The property chain (skosxl:prefLabel, skosxl:literalForm) is a sub-property of skos:prefLabel.
S56	The property chain (skosxl:altLabel, skosxl:literalForm) is a sub-property of skos:altLabel.
S57	The property chain (skosxl:hiddenLabel, skosxl:literalForm) is a sub-property of skos:hiddenLabel.
S58	skosxl:prefLabel, skosxl:altLabel and skosxl:hiddenLabel are pairwise disjoint properties.
S59	skosxl:labelRelation is an instance of owl:ObjectProperty.
S60	The rdfs:domain of skosxl:labelRelation is the class skosxl:Label.
S61	The rdfs:range of skosxl:labelRelation is the class skosxl:Label.
S62	skosxl:labelRelation is an instance of owl:SymmetricProperty.

Appendix D: CLEF Domain Specific 2004-2006 Results

This is a selected subset of the best runs submitted by different research groups between 2004-2006. For each year, the participants had a new set of queries they have tested their systems. A direct comparison with each submission would have been extremely laborious, therefore we computed an indicated mean average precision for each track (EN, DE, EN-DE, DE-EN) is based on the formula:

$$MAP_{track} = \frac{1}{|Q_1| + |Q_2| + ... + |Q_k|}(\sum_{q_i \in Q} AP_i) * 25$$

where $|Q_i|$ is the number of topics for each run listed in Table .

Table A8: MAP from CLEF Domain Specific Track

Participant	Year	Track	MAP
Berkley	2004	EN	39.85
IRIT	2004	EN	38.55
Unine	2004	EN	50.65
Berkeley_2	2004	EN	46.97
Berkley	2004	DE	42.8
Hagen	2004	DE	24.82
Ricoh	2004	DE	23.81
Berkley	2004	EN-DE	38.68
Ricoh	2004	EN-DE	12.61

Continued on next page

Table A8 – *Continued from previous page*

Participant	**Year**	**Track**	**MAP**
Berkley	2004	DE-EN	40.53
FU Hagen	2004	DE-EN	3.69
Unine	2005	EN	50.65
Berkeley_2	2005	EN	46.97
Univ. Glasgow	2005	EN	34.84
Berkeley	2005	EN	32.91
Irit	2005	EN	32.35
Berkeley_2	2005	DE	49.36
Unine	2005	DE	49.21
Univ. Glasgow	2005	DE	30.29
Hagen	2005	DE	30.31
Berkeley	2005	DE	23.14
Berkeley_2	2005	EN-DE	42.01
Hagen	2005	EN-DE	23.99
Univ. Glasgow	2005	EN-DE	19.46
Hildesheim	2005	EN-DE	17.79
Berkeley	2005	EN-DE	16.87
Berkeley_2	2005	DE-EN	47.43
Univ. Glasgow	2005	DE-EN	38.99
Berkeley	2005	DE-EN	23.98
Unine	2006	EN	43.03
Berkeley	2006	EN	41.36
Tuchemniz	2006	EN	35.53
Tuchemniz	2006	DE	54.54
Unine	2006	DE	50.51
Berkeley	2006	EN	39.17
Hagen	2006	EN	35.39
Hagen	2006	EN-DE	24.48
Berkeley	2006	EN-DE	23.66
Berkeley	2006	DE-EN	33.01

Table A9: Averaged Results of CLEF DS 2004-2006 with Title and Description for Query Formulation

DS 04-06	**EN**	**DE**	**DE-EN**	**EN-DE**
Average MAP for past runs	44.00	37.77	31.27	24.39

References

Nor Azlinayati Abdul Manaf, Sean Bechhofer, and Robert Stevens. The current state of skos vocabularies on the web. In *Proceedings of the 9th International Conference on The Semantic Web: Research and Applications*, ESWC'12, pages 270–284, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-30283-1. doi: 10.1007/978-3-642-30284-8_25. URL http://dx.doi.org/10.1007/978-3-642-30284-8_25. 53

Karam Abdulahhad, Jean-Pierre Chevallet, and Catherine Berrut. Revisiting exhaustivity and specificity using propositional logic and lattice theory. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, ICTIR '13, pages 21:93–21:100, New York, NY, USA, 2013. ACM. 83, 146

Mirna Adriani and C.J. Rijsbergen. Term similarity-based query expansion for cross-language information retrieval. In Serge Abiteboul and Anne-Marie Vercoustre, editors, *Research and Advanced Technology for Digital Libraries*, volume 1696 of *Lecture Notes in Computer Science*, pages 311–322. Springer Berlin Heidelberg, 1999. ISBN 978-3-540-66558-8. doi: 10.1007/3-540-48155-9_20. URL http://dx.doi.org/10.1007/3-540-48155-9_20. 30

Mitra Akasereh, Nada Naji, and Jacques Savoy. Unine at clef 2012. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, 2012. ISBN 978-88-904810-3-1. URL http://dblp.uni-trier.de/db/conf/clef/clef2012w.html#AkaserehNS12. 164

Riccardo Albertoni, Monica De Martino, Sabin Di Franco, Valentina De Santis, and Paolo Plini. Earth: An environmental application reference thesaurus in the linked open data cloud. *Semantic Web*, 5(2):165–171, 2014. 53, 56

G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389, 2002. 163

Gianni Amati. *Probability Models for Information Retrieval based on Divergence from Randomness*. PhD thesis, Department of Computing Science University of Glasgow, June 2003. 128

Simon Andrews. In-close, a fast algorithm for computing formal concepts. In *the Seventeenth International Conference on Conceptual Structures*, 2009a. 98

Simon Andrews. In-close, a fast algorithm for computing formal concepts. In *International Conference on Conceptual Structures*, 2009b. 78, 119

Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern the concepts and technology behind search*. Addison-Wesley, edition, 2011. ISBN 978-0-321-41691-9. 2, 23, 25, 36, 39, 130

Mark C. Baker. *The Atoms of Language: The Mind's Hidden Rules of Grammar*, chapter Toward a Periodic Table of Languages, pages 157–198. Basic Books, 2001. 15

Thomas Baker, Sean Bechhofer, Antoine Isaac, Alistair Miles, Guus Schreiber, and Ed Summers. Key choices in the design of simple knowledge organization system (skos). *CoRR*, abs/1302.1224, 2013. 52, 169

Tim Berners-Lee, Roy Thomas Fielding, and Larry Masinter. Uniform resource identifiers (uri): Generic syntax and semantics. 1998. 49

Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web: Scientific american. *Scientific American*, May

2001. URL http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&#38;pageNumber=1&#38;catID=2. 5

Michael W. Berry and Murray Browne. *Understanding Search Engines: Mathematical Modeling and Text Retrieval (Software, Environments, Tools), Second Edition*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2005. ISBN 0898715814. 2

Jorges Luis Borges. *Collected Fictions*, chapter Funes, His Memory. Penguin Group, 1998. 36

Alessio Bosca, Matteo Casu, Mauro Dragoni, and Chiara Francescomarino. Using semantic and domain-based information in clir systems. In Valentina Presutti, Claudia d‚ÄôAmato, Fabien Gandon, Mathieu d‚ÄôAquin, Steffen Staab, and Anna Tordai, editors, *The Semantic Web: Trends and Challenges*, volume 8465 of *Lecture Notes in Computer Science*, pages 240–254. Springer International Publishing, 2014. ISBN 978-3-319-07442-9. doi: 10.1007/978-3-319-07443-6_17. URL http://dx.doi.org/10.1007/978-3-319-07443-6_17. 44

Martin Braschler and Julio Gonzalo. Best practices in system and user oriented multilingual information access. Technical report, TrebleCLEF, October 2009. 115, 116, 120

C. Brewster and Y. Wilks. Ontologies, taxonomies, thesauri: Learning from texts. In *Proceedings of the Workshop on the Use of Computational Linguistics in the Extraction of Keyword Information from Digital Library Content*, Kings College, London, UK, 2004. viii, 47

Nicoletta Calzolari. *Large-Scale Knowledge Resources. Construction and Application*, volume 4938, chapter Initiatives, Tendencies and Driving Forces for a Lexical Web as Part of a Language Infrastructure, pages 90–105. Springer Berlin/Heidelberg, 2008. 7, 43

Claudio Carpineto and Giovanni Romano. *Concept data analysis: Theory and applications*. John Wiley & Sons, 2004. 77

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. Linking linguistic resources: Examples from the open linguistics working group. In *Linked Data in Linguistics*, pages 201–216. 2012. 7

Philipp Cimiano, Antje Schultz, Sergej Sizov, Philipp Sorg, and Steffen Staab. Explicit versus latent concept models for cross-language information retrieval. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, July 2009. 27, 33

Council on Library and Information Resources. Knowledge Organization Systems: An Overview, May 2014. URL `http://www.clir.org/pubs/reports/pub91/1knowledge.html`. 7

Kareem Darwish and Douglas W. Oard. Probabilistic structured query methods. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 338–344, New York, NY, USA, 2003. ACM. 30

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990. 31

Ljiljana Dolamic and Jacques Savoy. Retrieval effectiveness of machine translated queries. *J. Am. Soc. Inf. Sci. Technol.*, 61(11):2266–2273, November 2010. ISSN 1532-2882. doi: 10.1002/asi.v61:11. URL `http://dx.doi.org/10.1002/asi.v61:11`. 123

Sándor Dominich. *The Modern Algebra of Information Retrieval (The Information Retrieval Series)*. Springer, 1 edition, April 2008. ISBN 3540776583. 83

Jon Ducrou and Peter W. Eklund. Searchsleuth: The conceptual neighbourhood of an web query. In *CLA*, 2007. 154

Helge Dyvik. Exploiting structural similarities in machine translation. *Computers and the Humanities*, 28(4-5):225–234, 1994. 79

Claire Fautsch, Ljiljana Dolamic, Samir Abdou, and Jacques Savoy. Domain-specific ir for german, english and russian languages. In *CLEF*, pages 196–199, 2007. viii, 42, 43

Christiane Fellbaum and Piek Vossen. Connecting the universal to the specific: Towards the global grid. In *Intercultural Collaboration I : Lecture Notes in Computer Science, Springer-Verlag*, 2007. 28

Nicola Ferro. Clef 15th birthday: Past, present, and future. *SIGIR Forum*, 48(2):31–55, December 2014. 18

Nicola Ferro and Ivano Masiero. Appendix CHiC 2013 Evaluation Lab. http://www.promise-noe.eu/documents/10156/8f6af376-8095-48c1-badf-e317c4efdd46, 2013. 165

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM, 2001. 112

Evgeniy Gabrilovich. *Feature generation for textual information retrieval using world knowledge*. PhD thesis, Israel Institute of Technology, 2006. 29

Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, 2007. 27

Evgeniy Gabrilovich and Shaul Markovitch. Wikipedia-based semantic interpretation for natural language processing. *J. Artif. Intell. Res. (JAIR)*, 34:443–498, 2009. 32

Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis. Mathematical Foundations*. Springer-Verlag, 1999. 71

Bernhard Ganter, Gerd Stumme, and Rudolf Wille. *Formal Concept Analysis. Foundations and Applications*. Springer-Verlag, 2005. 74

Peter Gärdenfors. *The geometry of meaning, semantics based on conceptual spaces*. The MIT Press, 2014. 155

Thomas Gottron, Maik Anderka, and Benno Stein. Insights into explicit semantic analysis. In *CIKM*, pages 1961–1964, 2011. 29

Jorge Gracia, Elena Montiel-Ponsoda, Philipp Cimiano, Asunción Gómez-Pérez, Paul Buitelaar, and John McCrae. Challenges for the multilingual web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11(0):63 – 71, 2012. 61

Alasdair J. G. Gray, Norman Gray, and Iadh Ounis. Searching and exploring controlled vocabularies. In *ESAIR '09: Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 1–5, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-430-0. doi: http://doi.acm.org/10.1145/1506250.1506252. 55

Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, June 1993. ISSN 1042-8143. doi: 10.1006/knac.1993.1008. URL http://dx.doi.org/10.1006/knac.1993.1008. 38

N. Guarino and P. Giaretta. Ontologies and Knowledge Bases: Towards a Terminological Clarification. *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, pages 25–32, 1995. URL http://www.csee.umbc.edu/771/papers/KBKS95.pdf.Z. 27

Samer Hassan and Rada Mihalcea. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Singapore, 2009. 32

Ben He and Iadh Ounis. Studying query expansion effectiveness. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 611–619, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-00957-0. doi: 10.1007/978-3-642-00958-7_57. URL http://dx.doi.org/10.1007/978-3-642-00958-7_57. 140, 151

Daqing He and Jianqiang Wang. *Information Retrieval: Searching in the 21st Century*, chapter Cross-Language Information Retrieval. John Wiley & Sons, 2007. ISBN 0470027622. 120

Djoerd Hiemstra, Wessel Kraaij, Ren&#233;e Pohlmann, and Thijs Westerveld. Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In *CLEF '00: Revised Papers from the Workshop of Cross-Language Evaluation Forum on Cross-Language Information Retrieval and Evaluation*, pages 102–115, London, UK, 2001. Springer-Verlag. 30

Andreas Hotho, Steffen Staab, and Gerd Stumme. Text clustering based on background knowledge. Techreport 425, University of Karlsruhe, Institute AIFB, 76128 Karlsruhe, Germany, April 2003. 111, 112

ISO 25964-1:2011. *Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval*. ISO, Geneva, Switzerland, 2011. URL `http://www.iso.org/iso/catalogue_detail.htm?csnumber=53657`. 9, 52

ISO 25964-2:2013. *Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies*. ISO, Geneva, Switzerland, 2013. 9, 52

Martin Jaansen. *SiMuLLDA a Multilingual Lexical Database Appication using a Structured Interlingua*. PhD thesis, Utrecht University, 2002. 74

Ray Jackendoff. *Foundations of Language*. Oxford University Press, 2003. 15

Nick Jardine and Cornelis Joost van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information storage and retrieval*, 7(5): 217–240, 1971. 87

Jaap Kamps, Jussi Karlgren, Peter Mika, and Vanessa Murdock. Fifth workshop on exploiting semantic annotations in information retrieval: ESAIR'12. In *Proceedings of the 21st ACM international conference on*

Information and knowledge management, CIKM ’12, pages 2772–2773, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2398761. URL http://doi.acm.org/10.1145/2396761.2398761. 17

Epaminondas Kapetanios, Vijayan Sugumaran, and Diana Tanase. Multilingual web querying: A parametric linguistics based approach. In *NLDB*, pages 94–105, 2006. 15

Epaminondas Kapetanios, Vijayan Sugumaran, and Diana Tanase. A parametric linguistics based approach for cross-lingual web querying. *Data Knowl. Eng.*, 66(1):35–52, 2008. 15

Markus Kirchberg, Erwin Leonardi, Yu Shyang Tan, Sebastian Link, Ryan K. L. Ko, and Bu-Sung Lee. Formal concept discovery in semantic web data. In *ICFCA*, pages 164–179, 2012. 83

Michael Kluck. The domain-specific task of clef - specific evaluation strategies in cross-language information retrieval. In *In C. Peters(Ed.), Proceedings of the CLEF 2000 evaluation forum*, pages 48–56. Springer, 2001. 40

Sergei O. Kuznetsov and Sergei A. Obiedkov. Comparing performance of algorithms for generating concept lattices. *J. Exp. Theor. Artif. Intell.*, 14 (2-3):189–216, 2002. 77

Victor Lavrenko, Martin Choquette, and W. Bruce Croft. Cross-lingual relevance models. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’02, pages 175–182, New York, NY, USA, 2002. ACM. ISBN 1-58113-561-0. doi: 10.1145/564376.564408. URL http://doi.acm.org/10.1145/564376.564408. 34

D. Lenat. *The Dimensions of Context-Space*. Cycorp, 1998. 27

Craig Macdonald, Ben He, Vassilis Plachouras, and Iadh Ounis. University of at 2005: Experiments in terabyte and enterprise tracks with terrier. In *In Proceedings of TREC-05*, 2005. 164

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008. ISBN 978-0-521-86571-5. 25

Gary Marchionini. Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006. ISSN 0001-0782. doi: 10.1145/1121949.1121979. 36

Jennifer Marlow, Paul Clough, JuanCigarrán Recuero, and Javier Artiles. Exploring the effects of language skills on multilingual web search. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and RyenW. White, editors, *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, pages 126–137. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-78645-0. doi: 10.1007/978-3-540-78646-7_14. URL http://dx.doi.org/10.1007/978-3-540-78646-7_14. 15

Edgar Meij and Maarten de Rijke. Concept models for domain-specific search. In *CLEF*, pages 207–214, 2008. 41, 42

Alistair Miles. A theory of retrieval using structured vocabularies. Master's thesis, 2006. URL http://isegserv.itd.rl.ac.uk/retrieval/. 19, 90, 110

Alistair Miles and Dan Brickley. SKOS Core Guide. World Wide Web Consortium, Working Draft http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102, November 2005. 54

George Miller and Christiane Fellbaum. Wordnet then and now. *Language Resources and Evaluation*, 41(2):209–214, 2007. 75

George A Miller and Walter G Charles. Contextual correlates of semantic similarity. *Language Cognitive Processes*, 6(1):1–28, 1991. URL http://eric.ed.gov/ERICWebPortal/recordDetail?accno=EJ431389. 112

Elena Montiel-Ponsoda, Jorge Gracia, Guadalupe Aguado-De-Cea, and Asunción Gómez-Pérez. Representing translations on the semantic web. In *The 10th International Semantic Web Conference*, October 2011. 68

Calvin N. Mooers. A mathematical theory of language symbols in retrieval. In *Proceedings of the International Conference on Scientific Information*, volume 16-21, November 1958. 19, 83, 155

Nicolas Moreau. Best practices in language resources for multilingual information access. Technical report, TrebleCLEF, October 2009. 120

Helmut Nagy, Tassilo Pellegrini, and Christian Mader. Exploring structural differences in thesauri for
skos
-based applications. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 187–190, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0621-8. doi: 10.1145/2063518.2063546. URL http://doi.acm.org/10.1145/2063518.2063546. 53

Roberto Navigli and Simone Paolo Ponzetto. BabelNetXplorer: a platform for multilingual lexical knowledge base access and exploration. In *Proceedings of the 21st international conference on World Wide Web (WWW), Comp. volume*, pages 393–396, 2012. 4

Jian-Yun Nie. Cross-language information retrieval. In *Cross-Language Information Retrieval*, 2010. 24

Yasushi Ogawa, Tetsuya Morita, and Kiyohiko Kobayashi. A fuzzy document retrieval system using the keyword connection matrix and a learning method. *Fuzzy Sets Syst.*, 39(2):163–179, February 1991. ISSN 0165-0114. doi: 10.1016/0165-0114(91)90210-H. URL http://dx.doi.org/10.1016/0165-0114(91)90210-H. 27

Borys Omelayenko. Porting cultural repositories to the semantic web. In *Proceedings of the First Workshop on Semantic Interoperability in the European Digital Library (SIEDL-2008)*, pages 14–25, June 2008. 10

Borys Omelayenko. Loading Europeana Metadata into Semantic Repository, Europeana White Paper. Technical report, 2010. URL http://borys.name/papers/EuropeanaMetadataRepository.pdf. 10

I. Ounis, G. Amati, Plachouras V., B. He, C. Macdonald, and Johnson. Terrier Information Retrieval Platform. In *Proceedings of the 27th European Conference on IR Research (ECIR 2005)*, volume 3408 of *Lecture Notes in Computer Science*, pages 517–519. Springer, 2005. ISBN 3-540-25295-9. 164

Juan-Antonio Pastor-Sanchez, Francisco Javier Martinez Mendez, and Jose Vicente Rodriguez-Munoz. Advantages of thesaurus representation using the simple knowledge organization system (skos) compared with proposed alternatives. *Information Research: An International Electronic Journal*, 14(4), December 2009. ISSN 1368-1613. URL http://www.editlib.org/p/54656. 53

Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors. *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, volume 3491 of *Lecture Notes in Computer Science*, 2005. Springer. ISBN 3-540-27420-0. 42, 186

Carol Peters, Martin Braschler, and Paul Clough. *Multilingual Information Retrieval - From Research To Practice*. Springer, 2012. ISBN 978-3-642-23007-3. 3, 29, 30, 107, 122

Vivien Petras. GIRT and the use of subject metadata for retrieval. In Peters et al. [2005], pages 298–309. ISBN 3-540-27420-0. 42

Vivien Petras. How one word can make all the difference-using subject metadata for automatic query expansion and reformulation. *Working Notes for the CLEF 2005 Workshop*, pages 21–23, 2005. 56

Vivien Petras, Natalia Perelman, and Fredric C. Gey. Using thesauri in cross-language retrieval of german and french indexed collections. In *CLEF*, pages 349–362, 2002. 41

Vivien Petras, Fredric C. Gey, and Ray R. Larson. Domain-specific CLIR of english, german and russian using fusion and subject metadata for query expansion. In *CLEF*, pages 226–237, 2005. 41

Vivien Petras, Nicola Ferro, Maria Gäde, Antoine Isaac, Michael Kleineberg, Ivano Masiero, Mattia Nicchio, and Juliane Stiller. Cultural Heritage in CLEF (CHiC) Overview 2012. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, 2012. ISBN 978-88-904810-3-1. URL `http://dblp.uni-trier.de/db/conf/clef/clef2012w.html#PetrasFGIKMNS12`. 164

Vivien Petras, Toine Bogers, Elaine Toms, Mark Hall, Jacques Savoy, Piotr Malak, Adam Pawlowski, Nicola Ferro, and Ivano Masiero. Cultural heritage in clef (chic) 2013. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *CLEF*, volume 8138 of *Lecture Notes in Computer Science*, pages 192–211. Springer, 2013. ISBN 978-3-642-40801-4. 157

Ari Pirkola, Heikki Keskustalo, and Kalervo Järvelin. The effects of conjunction, facet structure, and dictionary combinations in concept-based cross-language retrieval. *Information Retrieval*, 1(3):217–250, 1999. ISSN 1386-4564. doi: http://dx.doi.org/10.1023/A:1009939707058. 30

Jonas Poelmans, Paul Elzinga, Stijn Viaene, Guido Dedene, and Sergei O. Kuznetsov. Text mining scientific papers: a survey on FCA-based information retrieval research. In Petra Perner, editor, *Industrial Conference on Data Mining - Poster and Industry Proceedings*, pages 82–96. IBaI Publishing, 2011. ISBN 978-3-942954-06-4. URL `http://dblp.uni-trier.de/db/conf/incdm/incdm2011p.html#PoelmansEVDK11`. 153

Artem Polyvyanyy and Dominik Kuropka. *A quantitative evaluation of the enhanced topic-based vector space model*. Universitätsverlag Potsdam, 2007. ISBN 3939469955. URL http://www.worldcat.org/isbn/3939469955. 29

Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM. ISBN 1-58113-015-5. doi: 10.1145/290941.291008. URL http://doi.acm.org/10.1145/290941.291008. 34

Bruno Pôssas, Nivio Ziviani, Wagner Meira, Jr., and Berthier Ribeiro-Neto. Set-based vector model: An efficient approach for correlation-based ranking. *ACM Trans. Inf. Syst.*, 23(4):397–429, October 2005. ISSN 1046-8188. doi: 10.1145/1095872.1095874. URL http://doi.acm.org/10.1145/1095872.1095874. 27

Uta Priss. Linguistic applications of formal concept analysis. In *the First International Conference on Formal Concept Analysis*. Springer, 2004. 74

Uta Priss and L.John Old. Concept neighbourhoods in lexical databases. In L©onard Kwuida and Barƒ±ü Sertkaya, editors, *Formal Concept Analysis*, volume 5986 of *Lecture Notes in Computer Science*, pages 283–295. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-11927-9. doi: 10.1007/978-3-642-11928-6_20. URL http://dx.doi.org/10.1007/978-3-642-11928-6_20. 74

Bob Rehder, Michael L. Littman, Susan T. Dumais, and Thomas K. Landauer. Automatic 3-language cross-language information retrieval with latent semantic indexing. In *TREC*, pages 233–239, 1997. 31

Ian Roberts and Anders Holmberg. *Organizing Grammar. Linguistic Studies in Honor of Henk van Riemsdijk*, chapter On the role of parameters in universal grammar: a reply to Newmeyer, pages 538–553. Mouton de Gruyter, Berlin, 2005. 15

Stephen Robertson. Evaluation in information retrieval. In Maristella Agosti, Fabio Crestani, and Gabriella Pasi, editors, *Lectures on Information Retrieval*, volume 1980 of *Lecture Notes in Computer Science*, pages 81–92. Springer Berlin Heidelberg, 2001. ISBN 978-3-540-41933-4. doi: 10.1007/3-540-45368-7_4. URL http://dx.doi.org/10.1007/3-540-45368-7_4. 121

Stephen E. Robertson and Micheline Hancock-Beaulieu. On the evaluation of ir systems. *Inf. Process. Manage.*, 28(4):457–466, 1992. 3

Thomas Roelleke. *Information Retrieval Models: Foundations & Relationships*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2013. ISBN 9781627050784, 9781627050791. 29

Magnus Sahlgren. *The Word-space model*. PhD thesis, University of Stockholm (Sweden), 2006. 31

Gerard Salton, Edward A. Fox, and Harry Wu. Extended boolean information retrieval. *Commun. ACM*, 26(11):1022–1036, 1983. 27

Jacques Savoy and Pierre-Yves Berger. Monolingual, bilingual, and girt information retrieval at clef-2005. In Carol Peters, FredricC. Gey, Julio Gonzalo, Henning Müller, GarethJ.F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *Accessing Multilingual Information Repositories*, volume 4022 of *Lecture Notes in Computer Science*, pages 131–140. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-45697-1. doi: 10.1007/11878773_14. URL http://dx.doi.org/10.1007/11878773_14. 56

Uta Priss School and Uta Priss. Lattice-based information retrieval. *Knowledge Organization*, 27:132–142, 2000. 27

SKOS Primer. SKOS Simple Knowledge Organization System Primer, World Wide Web Consortium, 18 August 2009. http://www.w3.org/TR/skos-primer, February . URL http://www.w3.org/TR/skos-primer/. 9, 52

SKOS Reference. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, World Wide Web Consortium, 18 August 2009 . http://www.w3.org/TR/skos-reference/, February 2009. URL http://www.w3.org/TR/skos-reference/. 52, 53, 98

SKOS-XL. SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL). March 2009. URL http://www.w3.org/TR/skos-reference/skos-xl.html. 59, 66

Skosify Tool. Skosify Tool. http://www.w3.org/2001/sw/wiki/Skosify, May 2011. URL http://www.w3.org/TR/skos-primer/. 53

Dagobert Soergel. Building a more meaningful web: From traditional knowledge organization systems to new semantic tools. *SIGIR Forum*, 37(2):65–72, sep 2003. ISSN 0163-5840. doi: 10.1145/959258.959271. URL http://doi.acm.org/10.1145/959258.959271. 48

P. Sorg and P. Cimiano. Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data and Knowledge Engineering*, 74(0):26 – 45, 2012. 33

John F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, CA, August 1999. 36

Darijus Strasunskas and Stein L. Tomassen. On variety of semantic search systems and their on variety of semantic search systems and their evaluation methods. In *International Conference on Information Management and Evaluation*, 2010. 11

Heiner Stuckenschmidt. Data semantics on the web. *Journal on Data Semantics*, 1(1):1–9, 2012. 32

Ed Summers, Antoine Isaac, Clay Redding, and Dan Krech. Lcsh, skos and linked data. In *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*, DCMI '08, pages 25–33. Dublin

Core Metadata Initiative, 2008. URL http://dl.acm.org/citation.cfm?id=1503418.1503422. 52

Osma Suominen and Christian Mader. Assessing and improving the quality of skos vocabularies. *Journal on Data Semantics*, 2(2), 2013. URL http://eprints.cs.univie.ac.at/3707/. 53

Diana Tanase and Epaminondas Kapetanios. Evaluating the impact of personal dictionaries for cross-language information retrieval of socially annotated images. In *Working Notes for the CLEF 2008 Workshop*, September 2008. 16

Diana Tanase and Epaminondas Kapetanios. *Handbook of Research on Web 2.0, 3.0, and X.0: Technologies, Business, and Social Applications*, chapter Improving Cross-Language Information Retrieval by Harnessing the Social Web. Number 16 in Advances in E-Business Research Series (AEBR). IGI Global, 2009. 16

Diana Tanase and Epaminondas Kapetanios. Are skos concept schemes ready for multilingual retrieval applications? In *Proceedings of the 8th International Conference on Semantic Systems*, I-SEMANTICS '12, pages 149–156, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1112-0. doi: 10.1145/2362499.2362520. URL http://doi.acm.org/10.1145/2362499.2362520. 17

Raphael Troncy, Benoit Huet, and Simon Schenk. *Multimedia semantics: metadata, analysis and interaction*. Wiley, Chichester, West Sussex, U.K., 2011. ISBN 9780470747001 (cloth). 48, 49, 50

George Tsatsaronis and Vicky Panagiotopoulou. A generalized vector space model for text retrieval based on semantic relatedness. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, EACL '09, pages 70–78, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1609179.1609188. 111

Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. 2010. 30

Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. Extending conceptualisation modes for generalised formal concept analysis. *Information Science*, 181(10):1888–1909, May 2011. ISSN 0020-0255. doi: 10.1016/j.ins.2010.04.014. URL http://dx.doi.org/10.1016/j.ins.2010.04.014. 73

Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. Systems vs. methods: an analysis of the affordances of formal concept analysis for information retrieval. In *Proceedings of Formal Concept Analysis meets Information Retrieval (FCAIR), worlshop co-located with ECIR-2013*, Moscow, 03/2013 2013. 85

Kim Veltman. Towards a semantic web for culture. *Journal of Digital Information*, 4(4), 2006. ISSN 1368-7506. URL http://journals.tdl.org/jodi/index.php/jodi/article/view/113. 91, 156

Ivan Vuliƒá and Marie-Francine Moens. A unified framework for monolingual and cross-lingual relevance modeling based on probabilistic topic models. In Pavel Serdyukov, Pavel Braslavski, SergeiO. Kuznetsov, Jaap Kamps, Stefan Rºger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz, editors, *Advances in Information Retrieval*, volume 7814 of *Lecture Notes in Computer Science*, pages 98–109. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-36972-8. doi: 10.1007/978-3-642-36973-5_9. URL http://dx.doi.org/10.1007/978-3-642-36973-5_9. 34

Christian Wartena, Rogier Brussee, Luit Gazendam, and Willem-Olaf Huijsen. Apolda: A practical tool for semantic annotation. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications*, DEXA '07, pages 288–292, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-2932-1. doi: 10.1109/DEXA.2007.39. URL http://dx.doi.org/10.1109/DEXA.2007.39. 63, 129

Rudolf Wille. *Formal Concept Analysis. Foundations and Applications*, chapter Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies, pages 1–33. Springer-Verlag, 2005. 74

S. K. M. Wong, Wojciech Ziarko, and Patrick C. N. Wong. Generalized vector spaces model in information retrieval. In *SIGIR '85: Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 18–25, New York, NY, USA, 1985. ACM. 27, 28

William A. Woods. Conceptual indexing: A better way to organize knowledge. Technical report, Mountain View, CA, USA, 1997. 27

S. A. Yevtushenko. System of data analysis Concept Explorer. In *Proceedings of the 7th national conference on Artificial Intelligence KII-2000*, pages 127–134, 2000. 75

Benjamin Zapilko and York Sure. Converting thesoz to skos. Technical report, GESIS – Leibniz-Institut für Sozialwissenschaften, Bonn, 2009. URL http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2009/technicalreport_09_07.pdf. GESIS-Technical Reports 2009|07. 52, 117, 159

Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, 2008. 28, 41