ERA-NET CHIST-ERA: FWF 1097-N23

uComp: Embedded Human Computation for Knowledge Extration and Evaluation

D4.1
uComp Knowledge Extraction and Modelling Services

Editor(s)	Gerhard Wohlgenannt
Responsible Partner	WU
Status-Version:	Final-1.0
Date:	July 15, 2015
Project Number:	FWF 1097-N23
Project Title:	uComp: Embedded Human Computation for Knowledge Extration and Evaluation

Title of Deliverable:	uComp Knowledge Extraction and Modelling Services
Date of delivery to the EC:	July 15, 2015

Workpackage responsible for the Deliverable	WP4
Editor(s):	Gerhard Wohlgenannt
Contributor(s):	Gerhard Wohlgenannt
Approved by:	All Partners

Abstract	This deliverable report (Report on Knowledge Extration and Modeling) presents our efforts in Task 4.1, which includes coping with multilingual data, impact refinement in the ontology learning system resulting from human computation feedback, and system optimization in combination with Human Computation results.

Keyword List:	ontology learning, ontology evolution, impact refinement, optimization, multilinguality, human computation

Document Revision History

Version	Date	Description	Author
First draft	22/01/2015	initial draft	Gerhard Wohlgenannt[1]
Second draft	12/04/2015	extended draft	Gerhard Wohlgenannt[1]
1.0	31/07/2015	final version	Gerhard Wohlgenannt

Contents

1 Executive Summary

This deliverable tackles a couple of goals and issues described in the uComp project proposal. First of all, to "deal with noisy and heterogeneous" data. We put a lot of work into system extensions and experiments in this area. There have been extensive experiments to balance the input evidence sources, and to measure which number of sources, and which number of evidences is necessary to benefit from redundancy between sources. Furthermore we did experiments to optimize the impact values (source impact vector) to find a near-optimal combination of input sources with respect to the accuracy of the concept candidates generated by the system. Another important topic is multilinguality. We added components to learn ontologies from German data, which include text corpora in German, but also using German equivalents to WordNet and other system components. Other issues, for example *conflict mediation*, were tackled by using the spreading activation algorithm to choose concepts and relations, or voting rules in the area of crowdsourcing. Crowdsourcing is tightly coupled with our ontology learning system, we use it to validate parts of the ontology and also as feedback with is used to automatically tune the system.

2 Introduction

Ontologies are a cornerstone of the Semantic Web. As the manual construction of ontologies is expensive and cumbersome, systems for (semi-automatic) learning of ontologies have been created, which bootstrap the ontology construction process using data-driven methods. The D4.1 deliverable is centered around a couple of main areas which are tackled in the remainder of this document:

- Dealing with noisy and heterogeneous data: We did extensive system extensions and experiments to best make use of heterogeneous data sources.
- Using confidence values (the source impact vector) to optimize system quality, and also as a tool to tightly integrate ontology learning and crowdsourcing components (impact refinement).
- Providing support for multilinguality, by extending the system to also support input data in German language.
- Furthermore we provided advanced tools and visualizations to management, inspect and analyse the ontology learning framework, in this deliverable we will also briefly describe the Ontology Learning Frontend and APIs to access the data.

3 The Ontology Learning Framework

All experiments presented in the paper are based on an existing ontology learning system which evolved over the years. This section can only give a quick overview of the framework, for details on the workings of the system see Wohlgenannt et al. [18], or Liu et al. [9], who present the original version of the framework.

In a nutshell, the system starts from a (typically small) seed ontology, which it extends with additional concepts and relations. So the main tasks are the selection of new concept candidates and the positioning of concepts with regards to the seed ontology. The resulting ontologies are lightweight.

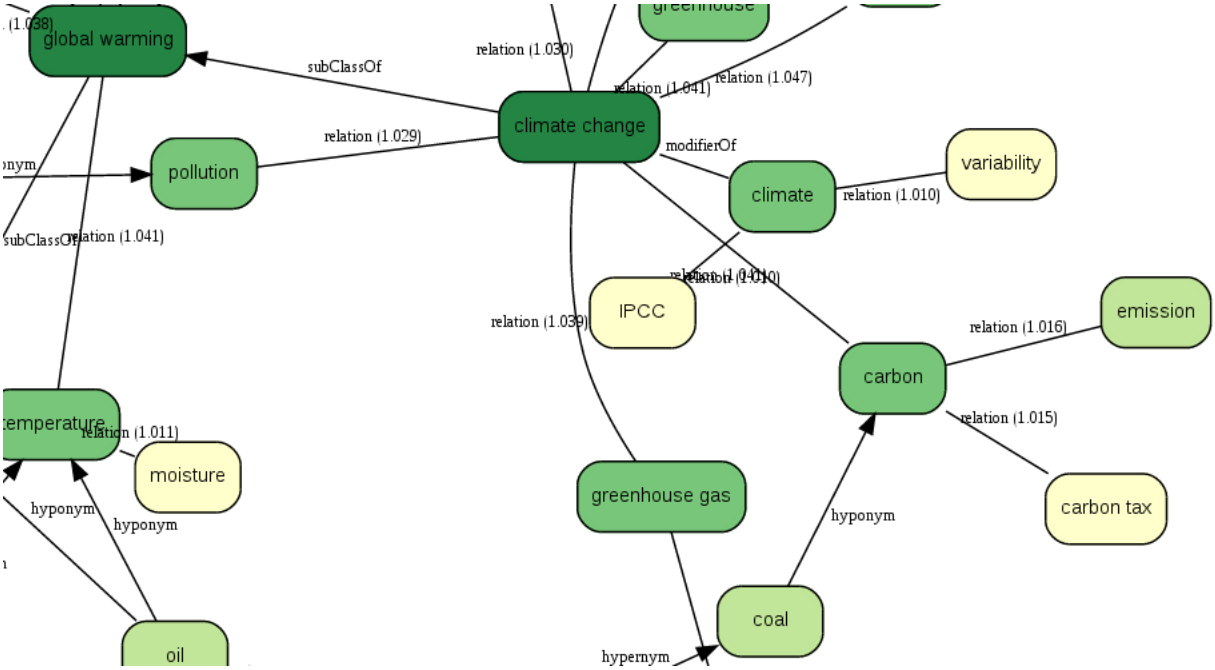

Figure 1: Part of a sample ontology in the domain of climate change after three stages (levels) of extension

We compute new ontologies for the domains in question in regular intervals (monthly) to trace the evolution of the domains. Figure 1 shows (parts of) the graphical representation of an example ontology learned in the *climate change* domain on data from January 2014. The concept colors indicate the ontology extension stage, the darker, the earlier the concepts were introduced. Before explaining the system in more detail, we take a look at the evidence sources used in the process.

3.1 The Evidence Sources

As the name suggests, the *evidence sources* provide the data needed to extend (learn) ontologies. In general, the input to evidence acquisition is a term (typically the label of a seed concept), and the result is a list of terms related to the seed term, and optionally significance values. So, for example, the system sends the seed concept label "CO2" to the Flickr API[1], and gets a list of related terms – which will then be used in the ontology learning process together with information from all other sources of evidence.

The evidence sources are heterogeneous regarding various aspects, such as (i) the number of evidences returned; (ii) the average quality (ie. domain relevance) of terms provided; (iii) the update frequency of the source – some sources are dynamic, eg. social sources and news

[1] www.flickr.com/services/api/flickr.tags.getRelated.html

media, some rather static, eg. WordNet and DBpedia; (iv) the availability of a significance score for the terms; and (v) the type of underlying data (text, structured, etc.).

	Extraction Method		
Data sources			
domain text from:	Keywords/page	Keywords/sentence	Hearst patterns
US news media	**1**	2	**3**
UK news media	4	**5**	6
AU/NZ news media	7	**8**	9
Other news media	**10**	11	12
Social media – Twitter	13	-	**14**
Social media – Youtube	**15**	-	16
Social media – Facebook	17	-	18
Social media – Google+	19	-	20
NGOs Websites	21	22	**23**
Fortune 1000 Websites	**24**	25	26

Table 1: The first 26 evidence sources are based on domain-specific text collected from the Web.

	Method				
Data source:	hypernyms	hyponyms	synonyms	API	SPARQL
WordNet	**27**	**28**	29	-	-
DBpedia	-	-	-	-	**30**
Twitter	-	-	-	**31**	-
Flickr	-	-	-	**32**	-

Table 2: The other 6 sources of evidence, based on WordNet, DBpedia and Social Media APIs.

By default, the ontology learning system currently uses 32 sources of evidence, which are listed in Tables 1 and 2. The first batch of experiments (see Section 6.4) conducted in year 2014 is based on all 32 sources, the second batch (year 2015) is based on the sources marked with boldface fonts. A major fraction of evidence sources is made up by the 16 sources based on keywords computed with co-occurrence statics on documents published and mirrored in the respective period of time, and filtered with a domain-detection service. The system computes page- and sentence-level keywords for documents collected from: US news media, UK news media, AU/NZ News Media, Websites of NGOs, Fortune 1000 company Websites, Twitter tweets, Youtube postings, Google+ postings, and public Facebook pages and postings. Furthermore, we use Hearst patterns [8] on those corpora, which constitutes further 10 evidence sources. Currently, the system includes two evidence sources based on calls to APIs of Social Media sites (Twitter, Flickr). Structured evidence sources contribute the remaining 4 sources of evidence, ie. hypernyms, hyponyms and synonyms from WordNet, and related terms from DBpedia. For more details on evidence sources used see Wohlgenannt et al. [19].

Table 3 includes example data for term extraction, it presents a short snippet of page-level keywords generated for the seed concept "CO2" from UK news media text (evidence source no. 4) in July 2013. This demonstrates the typical characteristics of evidence acquisition: some terms are relevant to the *climate change* domain, some are not relevant, some are too specific. A full listing of evidence for a seed term ("CO2") and examples of ontology run results is found at `https://ai.wu.ac.at/~wohlg/conf_data`. A demo portal of the underlying OL system is available at `http://hugo.ai.wu.ac.at:5050`.

Term	Significance	Term	Significance
carbon price floor	164.85	emission	110.48
sec	135.54	air	99.99
fertilisation	133.63	waste	90.17
PM10	123.45	0-62mph	89.12
environment committee	121.27	flame	86.74
member state	114.62	carbon tax	78.53

Table 3: Example evidence (keywords and their χ^2 co-occurrence significance) for the seed concept "CO2".

3.2 The Ontology Learning Process

Having described the evidence sources, we can introduce the basic workflow of the ontology learning system: The ontology learning run starts with a small seed ontology (in the climate change usecase we use two concepts, namely *climate change* and *global warming*). The system collects evidence for the seed concepts from the 32 (or 14) evidence sources. After integrating all this evidence into a semantic network, a transformation process using the source impact vector (SIV, see Section 6) converts the semantic network into a spreading activation network. The spreading activation algorithm yields new concept candidates. We currently pick the 25 candidates with the highest level of activation. The concept candidates are evaluated for domain relevance by domain experts, this is the only part of the system that involves human intervention. Finally, the system positions new concepts rated as relevant with regards to the existing ontology. For the next ontology extension step the framework uses the result from the previous iteration as new seed ontology, and further extends it. We typically do three ontology extension iterations. Figure 7 gives on overview of the workflow.

4 The Web Frontend

For managing the increasing amount of ontology computations calculated each month the front-end of the Ontology Learning System was developed. Its features are: management of ontology tasks (i.e. creating, running, changing, and deleting ontologies), displaying various high-level and low-level evaluation metrics, performing source impact vector **Optimization**, management of the **Database** backing the system, **Monitoring** the host systems resources (CPU, memory, disk space), providing a machine usable API, and processing evaluation results from Crowdflower or the Facebook Game. The main page, as shown in Figure 3, lists the

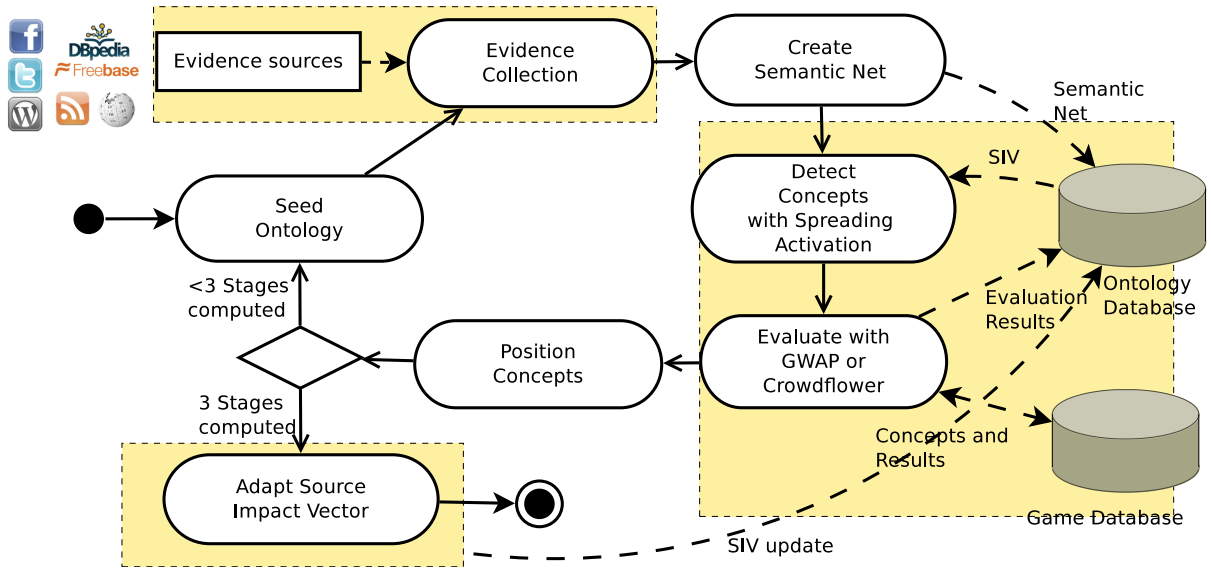

Figure 2: The Ontology Learning Framework

available ontologies found on disk and offers filtering options to the user. The ontology tasks that are currently running are shown and can be started or terminated manually from this page. However, as the front-end was designed to reduce the amount of maintenance needed for automatic ontology generation, it is usually not necessary to create and launch ontology calculations manually. Ontologies are created on automatically on a monthly basis from a predetermined list of ontologies in the database. Afterwards the system automatically launches the computations. Ontologies may be calculated in series or parallel depending on the settings of maximum parallel computations in the system's configuration. The progress of an ontology can be seen in the column *Stages* at the main page. A green bar represents a successful computation stage, orange indicates warnings and red represents errors which may have cause a calculation to fail. The log files for each stage can be viewed by clicking on the respective stage in the progress bar. Further details for an ontology can be seen when clicking on it name in the table. Evaluation comparing mainly concepts to one another can be found on the pages listed in the **Low-Level Evolution** section. The **High-Level Evolution** pages are focused on comparing entire domains to one another and on evaluating the development of domains over time. In Figure 4 shows an example of an ontology computed in July of 2015, which was evaluated via a Game With A Purpose (GWAP) on Facebook. The concepts accepted after the evaluation are highlighted in green, the rejected ones in red. The accepted concepts proceed to become seed concepts for the next ontology stage after positioning them in the Semantic Network of the previous stage. The ontology created by the learned concepts and their relations to the seed concepts are serialized in the Web Ontology Language (OWL) and Simple Knowledge Organization System (SKOS) format. For easy access the web interface provides interactive visualizations for these formats. Figure 5 shows the visualization tool **WebVOWL**, which has been integrated into the ontology learning front-end. WebVOWL implements the Visual Notation for OWL Ontologies (VOWL) [10] by providing graphical representations for elements in OWL in a Web based interface. It is released under the MIT license and available at `http://vowl.visualdataweb.org`. Similarly to WebVOWL the application **SKOS Play**

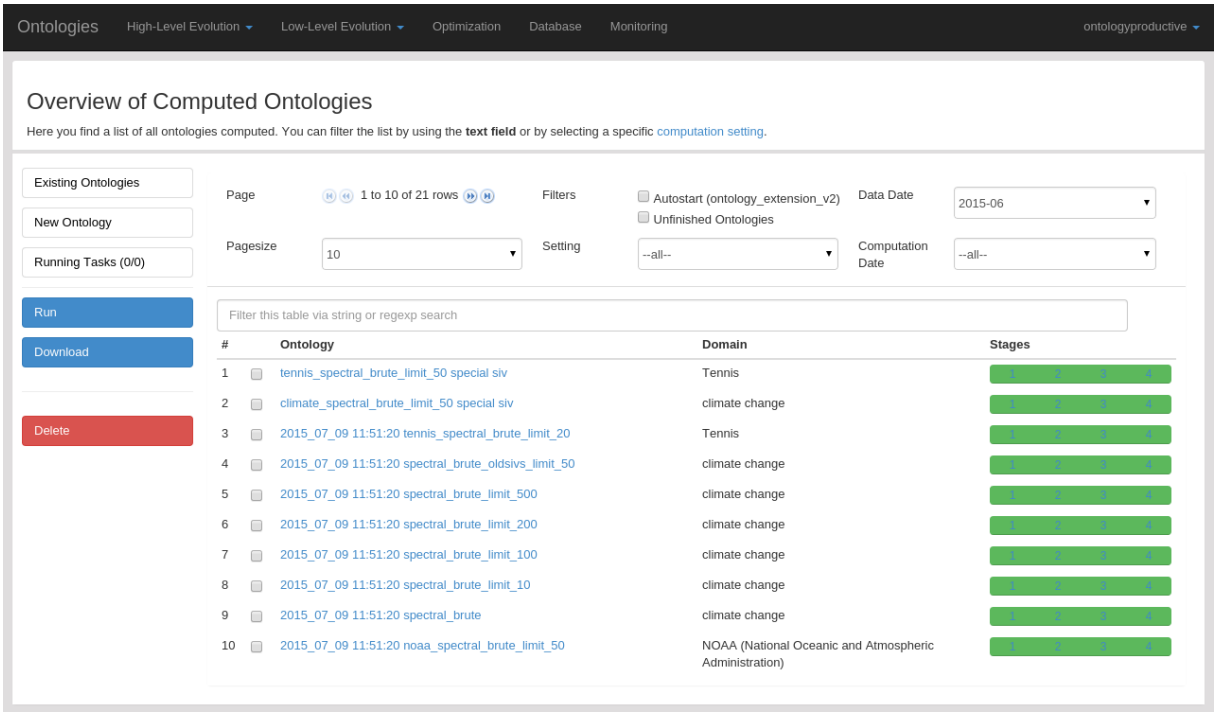

Figure 3: Ontology Learning Front-End - Main Page

was incorporated into the front-end for visualization of the SKOS format. It is open-source software and can be obtained from `http://labs.sparna.fr/skos-play/`.

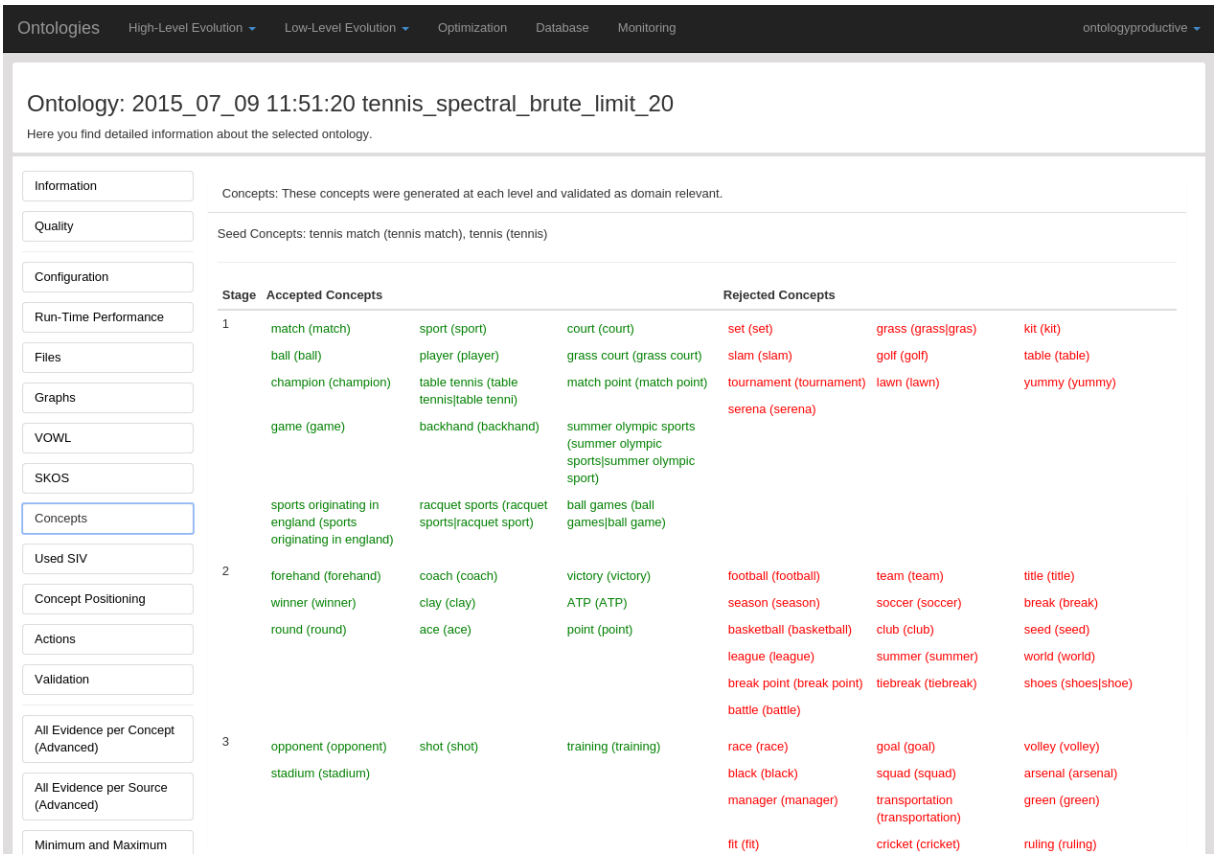

Figure 4: Ontology Learning Front-End - Concepts

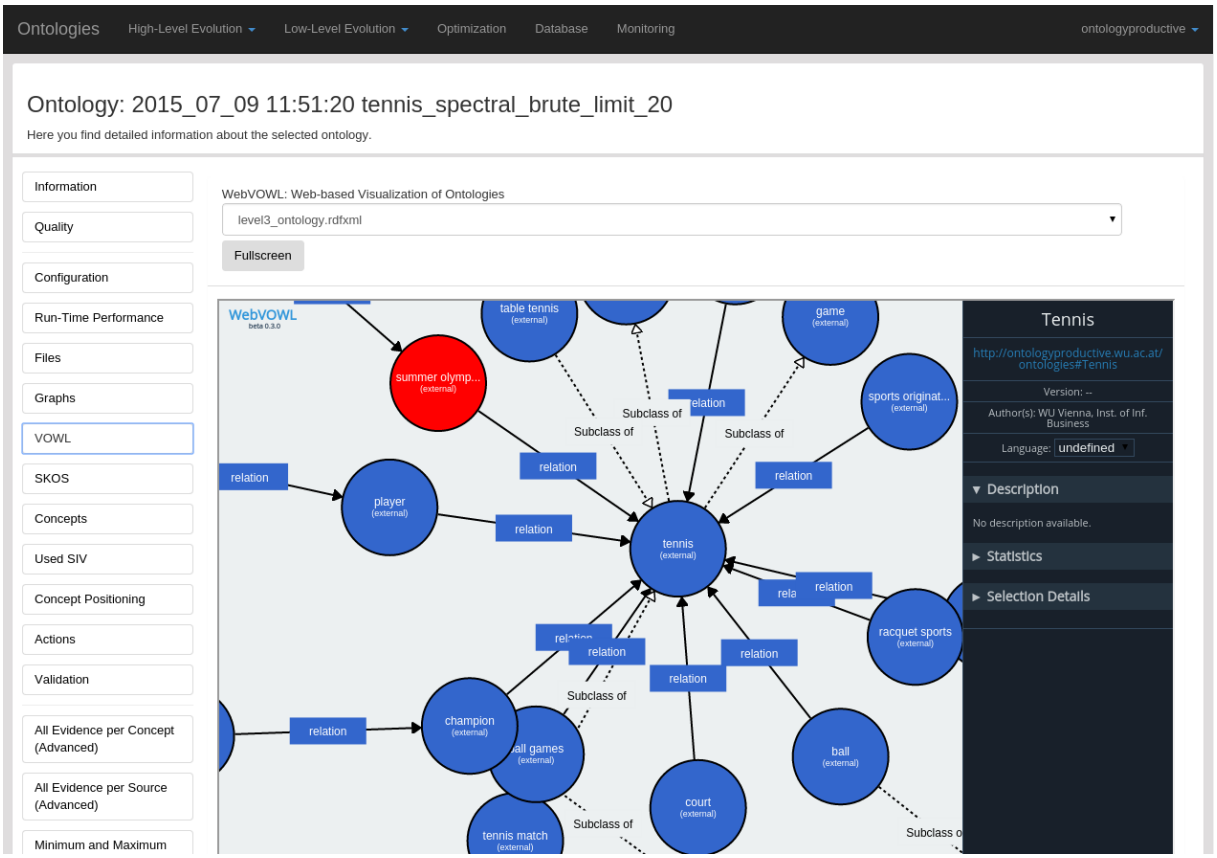

Figure 5: Ontology Learning Front-End - VOWL

4.1 API access

For easier and machine readable access to relevant data an API is also provided by the system, which is publicly accessible under the adress `http://ontologyproductive.wu.ac.at/api`. As of July 2015, the API consists of 19 different functions as shown in Table 4, with some having mandatory and optional parameters to customize the results returned. Further details for each function and it's parameters can be found at the API's help page[2]. Table 4 shows the

get_concept_connections	Returns the relations for a given concept in an ontology.
get_concept_occurrence	Returns the months in which the given concept occurred.
get_default_settings	Returns the default settings for one or more domains.
get_domain_quality	Returns quality metrics for a given domain.
get_manual_siv	Returns a SIV based on the source qualities of previous months.
get_ontology_quality	Returns quality metrics for a given ontology.
get_owl	Returns the Semantic Network of an ontology in OWL format.
get_setting_quality	Returns quality metrics for a given setting.
get_setting_quality_monthly	Returns quality metrics for a given setting over time.
get_skos	Returns the Semantic Network of an ontology in SKOS format.
get_source_quality	Returns the quality of a source for a give ontology, setting or domain.
get_trending_concepts	Returns a list of emerging and vanishing concepts for a domain.
help	Shows all API functions with their mandatory and optional parameters.
list_concepts	Lists all concepts in the database.
list_domains	Lists all domains in the database.
list_ontologies	Lists all ontologies in the database.
list_settings	Lists all settings in the database.
position_concept	Positions a concept within an ontology.
position_concept_in_range	Positions a concept within multiple ontologies over time.

Table 4: API Functions.

list of available API functions and a description of their basic behaviours. However most of these functions have mandatory or optional parameters that can customize the behaviour of the function, the data taken into consideration and therefore the results returned. The results returned by the API can be retrieved in JSON or CSV format for machine readability and in HTML format for human usage, with the global parameter *format* (json, csv or html as value) being applicable to all functions.

[2] `http://ontologyproductive.wu.ac.at/api/help?format=html`

5 Support for Multilinguality

One of the major goals in WP4.1 was to include support for multilinguality into our ontology learning system, this is learning ontologies from (text) data which is not English. We added support for German language corpora and services, which will be explained in more detail in the following.

Corpora Large text corpora crawled from the Web are the basis for keyword generation and the extraction of taxonomic structures with Hearst patterns. The corpora used in the English version of the ontology learning system had to be replaced by German ones for German ontology learning. They are provided by backend services from other workpackages, which crawl data from Austrian and German news, blog and social media sites for their Web intelligence platform.

Appositions The appositions used in the English version had to translated from English into German while maintaining the intended semantics of the single appositions. They are in the form of triples consisting of a regular expression, the position at which the concept can be found in the result of the regular expression and the relation type that is suggested by the apposition. The English appositions can be seen below:

In German these appositions listed above result in more complex regular expressions since in German most often there are articles in front of nouns which change according to the gender and numeric scope of the noun. The list of German appositions can be seen below:

These regular expressions are stored in files which are then dynamically selected by the apposition module based on the language setting of the ontology to be computed.

Open Thesaurus The Open Thesaurus module provides additional concepts and their relations during the calculation of an ontology. It is the counterpart of the Wordnet module which also provides concepts and their relations but only in the English language.

Short comparison of the Wordnet and the Openthesaurus Module: The Wordnet module is able to deliver information about the following relationships in the English language:

- **hypernym relations:** (also called superordinates) are general words; a word with a broad meaning constituting a category into which words with more specific meanings fall
- **hyponym relations:** hyponyms are subdivisions of more general words.
- **synonyms:** a word or phrase that means exactly or nearly the same as another word or phrase in the same language

The Open Thesaurus Module is able to deliver information about the following relationships in the German language:

- **hypernym relations:** (also called superordinates) are general words; a word with a broad meaning constituting a category into which words with more specific meanings fall
- **hyponym relations:** hyponyms are subdivisions of more general words.

The Open Thesaurs Module accesses the freely available web API from https://www.openthesaurus.de/. The API returns a XML document, which is parsed by the Open Thesaurus Module for hypernym and hyponym relations. The hypernym and hyponym contents of the XML document contain additional information that has to be cleaned or removed by the module. Only this cleaned content is used. Depending on the language setting set in the configuration of an ontology either the wordnet or the openthesaurus module is used to gather additional concepts and their relations for an ontology.

Challenges:

- The evidence collection module had to be extended in order to be able to dynamically switch evidence sources based on a specific language setting.
- The XML parser used in the Openthesaurus Module had to be error tolerant.
- The Open Thesaurus API restricts access to 60 requests per minute. The Open Thesaurus Module has to take the API limit in consideration and acts accordingly.

DBpedia This module collects evidence from the ontology offered by DBpedia . In particular it gathers information about acronyms, other names, and subjects for a given concept. The English version is hosted at http://dbpedia.org while versions for other languages are usually preceeded by a lanuage prefix. I.e http://de.dbpedia.org in the case of the German DBpedia page. The existing code had to be adapted to select the correct URL is based on the language setting of the ontology. Other possible languages are French, Italian, Spanish, Dutch, and Czech as DBpedia also hosts pages for these languages.

Twitter The Twitter API offers a language parameter to be sent with a request for trending keywords associated with a concept. There were no major issues with changing the existing Twitter module to also send the ontology language parameter. However as Twitter is a social media platform for short messages written in mostly colloquial language the results are of mixed quality. As English and German are often used together in such tweets the results from calling the API may contain German as well as English keywords.

6 Confidence Management & Impact Refinement

Ontology learning evolved from working on static domain text to Web sources, and more recently there are a few approaches that make use of multiple and heterogeneous data sources (see next section for more details). The introduction of heterogeneous sources into the learning process offers the potential for higher levels of accuracy, on the other hand there are challenges regarding the meaningful integration and balancing (of the impact) of sources. Manzano-Macho et al. [5] list some of the reasons for increased accuracy when using heterogeneous evidence sources: (i) redundancy of information in different sources represents a measure of relevance and trust, and (ii) additional sources can provide complementary data and valuable information that the other sources did not detect.

The question arising is to quantify the gains in accuracy in various OL tasks when using heterogeneous evidence sources. In this deliverable we take a detailed look on gains in the

concept detection task. So, the research question is: How does the number and the characteristics of heterogeneous evidence sources affect accuracy (ie. the ratio of relevant concept candidates) in concept detection? In other words, the problem is as follows: We start with an OL system that includes a number of (heterogeneous) evidence acquisition methods, which basically provide terminology (heterogeneous lists of terms). These are the input, the output of concept detection are a number of domain concept candidates. In the evaluation section we study the impact of the (i) number of evidence sources, (ii) number of evidences per source, (iii) heterogeneity and quality of sources and (iv) the balance between sources on the accuracy of concept detection.

The evidence used in the OL system is heterogeneous in various respects. It originates from different sources such as Web documents, social Web APIs, and structured sources, and from different extraction methods applied. This leads to heterogeneity regarding the quality of evidence, the vocabulary used, the number of evidences and the dynamics of the source (see Section 3.1).

The experiments are conducted with our OL system (see Section 3) that generates lightweight ontologies using the spreading activation algorithm [4] to integrate evidence. For the experiments, the architecture generated lightweight ontologies in two different domains ("climate change" and "tennis") in monthly intervals from scratch. As spreading activation is a simple and intuitive way to integrate heterogeneous evidence, the results can largely be generalized to other OL systems and integration logics for heterogeneous evidence which use a similar approach.

Additionally to mere balancing experiments of evidence sources, we also conducted experiments to optimize the system using the source impact vector. Build on work on balancing sources, we aim to further optimize system accuracy by using an optimization algorithm (Tabu search [7]) to find the best combination of input weights for the individual evidence sources.

The research questions are as follows: (i) How well can an ontology learning system be optimized by adapting source input weights? – especially if quality of evidence varies between sources. (ii) What is the influence of the number of sources used in the system on the optimization results? (iii) What other findings and guidelines can be extracted from the data collected in the optimization runs?

To address the research questions, we did two batches of optimization runs. The first one was conducted in 2013 with all 32 evidence sources used in our system, which was not very well tuned at that point. The second set of optimization runs was done in 2015, then with a better tuned system, and a reduced set of evidence sources (according to our findings in our previous work [17] that a limited number of sources is sufficient for high accuracy).

6.1 Related Work on Balancing and Optimization

Most OL systems learn ontologies from only one source, typically domain text, e.g. Text2Onto [3] or OntoLearn Reloaded [13]. Some authors, e.g. Sanchez and Moreno [11], combine corpus-based methods with Web statistics for ontology learning tasks. Others exploit structured data present in the current Semantic Web, e.g. Alani [1], who proposes a method for ontology building by cutting and pasting segments from online ontologies. More recently, some systems start to make use of heterogeneous evidence sources in OL. Using only one evidence source typically results in modest levels of accuracy [5], the combination of several

sources may partially overcome this problem.

Manzano-Macho et al. [5] present an architecture which learns from multiple sources using a number of methods. In the acquisition layer the system learns hypotheses about candidate elements (the core terminology of the domain) which include a probability of relevance and relations to other candidate elements. Acquisition uses statistical methods as well as NLP tools and visual (HTML layout-based) methods. Furthermore, the system filters for domain relevance, detects domain concepts and taxonomic relations, and evaluates the resulting ontology against a pre-selected reference ontology. OntoElect [12] is methodology for ontology engineering, which applies term extraction to papers by domain experts. They also describe termhood saturation experienced when extending the collection of papers. Among the few papers which focus on OL from heterogeneous sources is also an approach by Cimiano et al. [2] to learn taxonomic relations. This method converts evidence into first order logic features, and then uses standard classifiers (supervised machine learning) on the integrated data to find good combinations of input sources. The input sources include data from lexico-syntactical pattern matching, head matching and subsumption heuristics applied to domain text. Völker et al. [14] propose a similar approach which uses the confidence scores of several heterogeneous methods as features in a classifier, aiming to enrich existing ontologies with disjointness axioms. Manzano-Macho et al. [5] focus on small corpora of high quality domain text, our system however uses noisy and evolving data from the Web and also includes more diverse sources such as APIs from social media Websites and a linked data source (DBpedia). In terms of evaluation, we employ user-based evaluation with domain experts (see below), whereas Manzano-Macho et al. [5] compare their results against a reference ontology. Gacitua and Sawyer [6] present a quantitative comparison of technique combinations for concept extraction. Although the goal is similar to our work, they investigate which process pipeline of NLP techniques is most helpful for term extraction from a domain corpus, whereas we study the balancing of term lists stemming from heterogeneous evidence sources.

As mentioned, the skillful combination and balancing of evidence sources is a crucial factor to leverage the potential of heterogeneous sources. Spreading activation, which is a method for searching semantic networks and neural networks, is the key tool to integrate evidence sources in our framework. Spreading activation is also frequently used in information retrieval. In his survey Crestani [4] concludes that spreading activation is capable of providing good results in associative information retrieval.

6.2 Experiments for Confidence Balancing

This section includes evaluation results of ontology learning (OL) experiments conducted between July 2013 and December 2014. Starting from the seed ontology the system generated 75 concept candidates (3 runs of 25 concepts each) per ontology – this fixed number of 75 concept candidates per ontology was used in all upcoming experiments, irrespective of the number of evidence sources used. After Section 6.2.1 provides details about the evidence (term lists) used, Section 6.2.2 describes the experiments for integrating heterogeneous evidence. Section 6.2.3 discusses concept relevance assessment.

6.2.1 Characteristics of Evidence Sources

In order to get a meaningful interpretation of evidence balancing and integration, first the characteristics of the underlying input data need to be investigated. Two properties greatly vary between evidence sources: the number of evidences provided (for a seed concept), and the average term quality per evidence acquisition method. Term quality was measured as the ratio of terms supplied by the respectively method which label a relevant domain concept. Domain experts and crowdsourcing workers manually evaluated sufficiently large term lists for different seed concepts and methods – resulting in a few thousand terms – to assess term quality.

		Term Quality		
Method:	Avg. Num. of Evid.	Top 25	Top 100	Top 500
Keywords/page	400	0.31	0.26	0.12
Keywords/sentence	200	0.27	0.19	0.10
Hearst Patterns	18		0.15	
API Twitter	70		0.10	
API Flickr	16		0.18	
WordNet (Hypernyms)	15		0.24	
WordNet (Hyponyms)	17		0.21	
DBpedia	13		0.27	

Table 5: Average number of evidence and evidence quality per extraction method.

Table 5 gives an overview of these characteristics. It lists the methods described previously, and gives the rough average numbers of evidences per seed concept which the evidence sources provide (*Avg. Num. of Evid.*). Furthermore, the table includes term quality in the remaining columns. Only co-occurrence statistics-based terms (*keywords*) have a significance value assigned (and are thereby ordered), for these we evaluated the top 25, top 100, and top 500 most significant terms. For all other sources we evaluated all terms supplied. Table 5 shows (i) that the average number of evidences greatly differs between sources, and also that term quality varies to a large extent. Term quality is high for the 25 most significant keywords per seed concept, and also for terms provided by WordNet and DBpedia. Keywords of low significance, and social sources (esp. Twitter) yield low quality terms on average. Hearst patterns generate rather sparse results which are of moderate quality.

One aspect of using heterogeneous sources is that they provide **complementary input** to better cover the domain of interest. In our system, corpus-based techniques (mostly keywords) account for the base layer of evidence. Apart from text-based input, social sources add very recent and emotional terminology, helpful to improve results and capture dynamic aspects of the domain [15], but also include a large share of noise, typos, etc. WordNet typically offers general and high quality input, which also helps to build the taxonomic backbone, but does not reflect dynamic aspects of domain evolution. The current version of SPARQL queries against DBpedia returns specific and technical terms, but also many terms which are too specific or not relevant to the domain.

Balancing the number of evidences. As seen in Table 5, if not limited, the number of evidences (terms) supplied strongly varies between evidence sources. Whereas the number of keywords for a concept sometimes exceeds 1000 terms, other sources provide comparably few results. In the upcoming section we present experiments where the number of evidences per source is either not limited, or limited to a maximum number of evidences per source to (i) balance the influence of sources on the resulting ontology and (ii) study which impact the amount of evidence has on the quality of concept candidates suggested by the OL system.

6.2.2 Leveraging and Balancing Sources and Evidences

As already mentioned, the goal of this research is to provide hints and insights on the combination and integration of heterogeneous evidence sources in OL (specifically for the concept detection phase) which can be generalized.

$$Accuracy = \frac{Relevant\ concept\ candidates\ generated}{All\ concept\ candidates\ generated} \quad (1)$$

In this section, we measured the accuracy of the system by the ratio of relevant concept candidates resulting from the OL system, see Equation 1. Other aspects, such as the positioning of new concepts in the ontology and the detection and labeling of relations are not part of this study, some of these points are covered in [16].

The Number of Evidences per Source. The first question to address is the impact of the number of evidences per source on the quality of concept candidates. Table 6 summarizes experiments where every of the 32 evidence sources was limited to suggest only 5, 10, etc. evidences per seed concept. As discussed in the previous section, some sources like WordNet or DBpedia typically provide very few evidence, whereas keyword-based sources produce up to 1000 terms per source. Obviously, limiting all sources to (for example) 10 evidences per seed concept, will reduce the impact of keyword-based sources. Using limits i) balances to number of evidence between sources, ii) saves computation time, but also iii) removes data which might be helpful in the spreading activation (ie. evidence integration) process. We use two domains in the experiments, *climate change* and *tennis*. The *climate change* ontologies were generated from scratch in every month between July 2013 and November 2014, the *tennis* ontologies between July 2014 and November 2014. The accuracy numbers in Table 6 are based on 17 ontologies computed per respective setting for *climate change*, which leads to 1275 ($75 * 17$) concept candidates per setting. In the *tennis* domain, we have 5 ontologies per setting with 375 concept candidates. If not stated otherwise, these numbers also apply to upcoming tables later in this section.

With very few evidences per source (*limit=5*), the benefits of redundancy and integration of heterogeneous sources are small (poor accuracy), although using only the best (most significant) keywords. Only in interactive systems where run-time is a very critical issue such a setting should be considered. On the other hand, in our experiments with a limit of 200 or more evidences per seed concept, the number of evidences per source is unbalanced, and more and more keywords with low significance are added to the spreading algorithm network, negative effects exceed the benefits of additional evidence data. Accuracy is lower in the *tennis* domain, we attribute this to the underlying data used, which are general domain-agnostic

No. of Evidences	Acc. CC	Acc. Tennis	Acc. Random Keyw. CC
limit=5	56.44	46.80	52.72
limit=10	64.05	55.53	56.51
limit=20	67.57	60.27	60.98
limit=50	68.68	59.87	61.64
limit=100	67.79	58.27	62.73
limit=200	67.87	58.53	65.13
limit=500	66.39	57.88	66.01
no limit	66.29	57.34	66.29

Table 6: Accuracy of concept detection (percentage of relevant concept candidates) for the domains of *climate change* (CC) and *tennis* depending on the number of evidences per source, with default and random selection of keyword evidence.

(news) media corpora, whereas for *climate change* the system uses domain-specific corpora.

In contrast to our initial expectations that more evidence is always better regarding resulting ontology quality (although it will be computationally expensive), even low numbers (*limit=10*) allow high accuracy if evidence is ranked by expected quality. In our system keywords are ranked by their significance value. Our experiments suggest that in the range of 20 to 50 terms per evidence source very good or even best results can be expected. However, this is only true while using a sufficient number of evidence sources (see below). A remark: the differences in accuracy in Table 6 are statistically significant, eg. with $p = 0.009$ between accuracy of *limit=10* and *limit=20*.

A more detailed look at the ontologies exhibits a more frequent occurrence of specific and exotic (but still relevant) concepts when using a low limit (such as *limit=5*), while a high limit promotes more general terms. This fact, which is in favor of high *limit* settings is not reflected by the data in Table 6.

Out of curiosity we also experimented with choosing keywords randomly from the list of all keywords (instead using of the most significant), see column *Acc. Random Keyw. CC* in Table 6. As expected this lowers the accuracy with low *limits*, and gives a more realistic picture for systems where evidence per source is not ordered. Therefore, in a machine learning environment where the expected quality of evidence is unknown and there is no explicit grading, it is advisable to use more evidence per source to fully benefit from redundancy. Another experiment, in which the keyword significance (as yielded by co-occurrence statistics) was not used at all, gave very poor results. This confirms that the quality of sources is important, and that low-quality evidence cannot be compensated by using multiple sources entirely.

In summary, it is important to have enough data to benefit from redundancy and aggregation. Additional evidence beyond this point can even have a negative impact if the balance between sources is lost, or the quality of additional evidence is not sufficient.

The Number of Evidence Sources Used. Not only the number of evidences per source is important, also the influence of the number of heterogeneous sources on the learning algorithm has to be taken into consideration. We evaluated the impact of using (i) only one source

%Relevant	1 (Twitter)	1 (UK-KW-page)	5 srcs	15 srcs	32 srcs
CC limit=50	16.54	48.80	59.52	68.28	68.84
CC limit=200	19.85	49.78	57.48	67.73	67.64
Tennis limit=50	21.15	50.67	52.25	56.88	57.87
Tennis limit=200	23.17	52.78	54.33	57.74	58.33

Table 7: Accuracy (percentage of relevant concept candidates) of concept detection regarding the number of evidence sources ("srcs") used – for two limit-settings, in the domains of *climate change* and *tennis*.

which yields rather low quality terms (*1 Twitter*), (ii) only one source with high quality input (page-level keywords from UK media – *1 UK-KW-page*), (iii) five random sources (*5 sources*), (iv) *15 sources*, (v) all sources (*32 sources*). Table 7 presents the results for these five variants, it shows the outcome for limit settings with the number of evidences (terms) not exceeding 50 and 200, respectively.

When relying on a single source, the quality of evidence of that source is essential, obviously – see *1 Twitter* and *1 UK-KW-Page*. In our experiments, *5 sources* of mixed quality are sufficient to see the benefits of using multiple sources. Around *15 sources* can be enough to gain the full advantage of heterogeneous evidence integration and redundancy. This means that a small and computationally efficient spreading activation network with a sum of a few thousand terms (suggested by 10-15 sources) can be quite enough to get best results. The difference between 5 and 15 evidence sources is statistically significant e.g. for the *climate change* domain as confirmed with a binomial test ($p \approx 0.0006$ for both *limit* settings).

The know-how regarding the minimal number of sources necessary can be helpful in various situations: (i) when setting up a new system, (ii) when there is need to scale down an existing system that is too slow or consumes too many resources, (iii) when there is need to use only a subset of evidence sources for a particular application. For example, we plan ontology evolution and trend detection experiments in which we will only use sources which are highly dynamic, and omit more static sources such as WordNet.

The Number of Seed Concepts. Finally, we investigated the impact of the type and number of seed concepts for which evidence is collected. Our system learns ontologies in 3 iterations of extension. In the first iteration (*Stage1*) there are only very few seed concepts (in the climate domain: "climate change" and "global warming"), which are obviously very relevant to the domain. The seed concepts in *Stage2* are the expert confirmed concepts learned in *Stage1*, in *Stage3* the system uses the concepts acquired in *Stage2*. The concepts in *Stage2* are typically more general than in in *Stage3*, in which the ontology gets more granular. Table 8 presents the ratio of relevant concepts suggested regarding the stage (the number and granularity of seed concepts) and the number of evidences used.

A combination of a low number of seed concepts in *Stage1* and low number of evidences (*limit=5*) does not provide spreading activation with enough data to produce good results. Such a setting creates a network with only a few hundred evidences (2 SC $* 32$ sources $* 5$ evidences per source). This is well below the critical number of evidences of a few thousand

	Stage1 – 2 SC	**Stage2 – ca. 18 SC**	**Stage3 – ca. 35 SC**
limit=5	54.67	61.87	56.53
limit=50	80.30	69.96	55.56
limit=200	78.83	68.33	56.22

Table 8: Accuracy depending on seed concepts (SC) and evidence limit applied.

(according to our experiments) which is needed for high accuracy. On the other hand, when the number of seed concepts is high, then a high number of evidences per seed concept offers no additional benefit, accuracy for *limit=50* and *limit=200* is very similar. The best results are achieved in *Stage1*, which uses domain concepts of high relevance and generality, and enough evidence to exploit redundancy (limit $\geq$ 50). The accuracy in Stages 2 and 3 is diminishing, because the seed concepts tend to get less domain-relevant with increasing distance from the initial seed ontology.

Observations. A list of key observations and hints concludes this section: (i) It is critical to ensure having enough evidence to benefit from redundancy at every step of the learning cycle. In our system enough evidence corresponds to at least a few thousand pieces of evidence (terms). Additional evidence beyond this points only slows the system down while providing little use. (ii) When there is no order (regarding quality) in evidence data, then more evidence will be needed to get the best results. (iii) Using our evidence integration method and settings, around 10-15 sources of heterogeneous evidence are sufficient to gain the full effect of evidence integration. (iv) Balancing input from evidence sources is typically more important than the raw number of evidence per source.

And interesting strain of future work will be the attempt to optimize the source impact vector (SIV), which controls the influence of a particular source on the learning process. In this research we use a uniform source impact for all evidence sources as the goal is to study the balancing of evidence. In future work we will try to find an (almost) optimal configuration of impact of evidence sources in the spreading activation network. Preliminary studies show that this will lead to a significantly higher accuracy of the system.

6.2.3 Relevance Assessment

This section, which concludes the evaluation, takes an alternative view at the judgments on concept relevance made by the domain experts and crowd workers, especially on concept candidates rated *non-relevant*.When rating concept candidates, human workers had only two choices: *relevant* or *non-relevant* to the domain at the given level of granularity. We took a more detailed look at concept candidates that were rated as *non-relevant*. From 100 candidates rated *non-relevant* to the domain of *climate change*, 61% were in fact at least partly relevant to the domain, but very generic or too specific. Only the remaining 39% were not relevant at all, but according to the human workers not relevant for this level of granularity. For this reason, depending on point of view and granularity, the accuracy of the OL system is higher than stated in the evaluation data. Among the 61% of candidates partly relevant to the domain of *climate change* are mostly terms that are too generic (for example: "impact", "mitigation", "issue",

"policy", etc.). The 39% of clearly non-relevant terms include fragments from the phrase detection algorithm such as "change conference" or candidates simply unrelated ("century", "level", "wave").

Conclusions The integration of heterogeneous evidence sources can improve accuracy in ontology learning and other areas which use similar machine learning techniques and multiple evidence sources. In this paper we study how a system needs to be set up to gain the desired results, and give hints and insights on the impact on accuracy of the number of evidences per source, the number of evidence sources, of quality per evidence source, etc. Among the key findings and contributions is the surprising fact that a limited number of evidences – a few thousand terms from heterogeneous sources – provides results of similar quality compared to using much higher numbers. In addition, in our experiments around 10-15 evidence sources were sufficient to gain full benefits of redundancy and evidence aggregation. Heterogeneous sources of evidence not only help to raise accuracy, but also offer complementary vocabulary to cover the domain.

Future work will apply the presented experiments to similar systems. We expect similar results as the basic characteristics of evidence integration do not change. Furthermore, we will further optimize the system using the source impact vector (SIV) by (i) adapting the SIV over time according to the quality of concept candidates suggested by the source to increase the impact of sources that consistently suggest a high ratio of relevant concepts, and (ii) conducting optimization experiments of find an optimal configuration for the SIV.

6.3 Experiments for Impact Optimization

The goal of research presented in this subsection is to improve the ratio of relevant to non-relevant concept candidates, ie. to improve the output of the spreading activation algorithm. The SIV is a key factor in this optimization process, as it determines – in combination with significance scores provided by the evidence sources – the weights in the spreading activation network.

6.3.1 The source impact vector

As mentioned, evidence sources are heterogeneous in number and quality of terms provided, we use a so-called Source Impact Vector (SIV) to manage the influence of a particular evidence source on the ontology learning process. Equation 2 demonstrates that the SIV consists of one impact value per evidence source (and point in time). The impact value is in the interval $[0.0, 1.0]$, a value of 1.0 results in high impact in the learning processes, wheres 0.0 in fact omits evidence suggested by the respective source.

$$SIV_t = \left[\begin{array}{cccc} I_{es_1,t} & I_{es_2,t} & \cdots & I_{es_n,t} \end{array} \right] \tag{2}$$

The SIV is used to set the weights in the spreading activation network (see next subsection for details), which selects new concept candidates for the ontology. Initial versions of the system ([9], [16]) applied a manually picked and static source impact, in this paper we propose novel ideas and experiments to optimize the ontology learning system via the SIV.

The optimization process aims to find a configuration of the SIV which maximises the ratio of relevant new concept candidates suggested by the system.

6.3.2 The Optimization Process

Although a spreading activation network has the fundamental characteristics of a neural network, we did not find a way to apply classic neural network learning techniques to optimize the output for a number of reasons:

- The spreading activation network doesn't have an explicit output layer, the *results* of the spreading activation algorithm are the activation levels of nodes all over the network.
- We select a preset number of nodes (eg. 25) with the highest activation level as concept candidates. The use of an error function (as used eg. in backpropagation) is not straightforward, as we only assess the preset number of nodes with the highest activation, but any other node might be a relevant domain concept as well. So there is no distinct correct output of the spreading activation network that could be used.
- The learning algorithm can not freely optimize the weights in the network, as values of the SIV are only factors in the connection weights. First of all, when multiple SIV factors make up a connection weight, it is not clear which specific SIV factor should be changed. And more importantly, if a specific SIV value is changed for one connection, it needs to be changed simultaneously everywhere in the spreading activation network wherever used, leading to unpredictable effects.

The characteristics described above led us to experiment with heuristics to improve the output of the ontology learning framework based on the modification of the SIV. This includes a baseline with a static SIV (Section 6.3.2), and a model that aims to optimize the SIV (Section 6.3.2).

Overall, the crucial factor which has an impact on the results of the ontology learning process are not so much the absolute values in the SIV, but the differences between evidence sources. Higher source impact for an evidence source results in increased activation levels and therefore a higher chance of being a candidate concept for evidence suggested by the particular source.

Static Source Impact Values The simplest way to use the SIV is to have static values for any source, not changing over time or across domains. We use a source impact of 0.2 for all 32 evidence sources. This uniform source impact has been used in the experiments regarding the number and balancing of evidence sources presented by Wohlgenannt [17], and provides good results, which we use as a baseline and starting point of the optimization experiments.

Optimization With this strategy, instead of having a single static SIV, the system investigates different SIV settings and their results. In the first batch of experiments, we set the source impact for every evidence source to values in the interval $[0.0, 1.0]$ with a step-size of 0.1, i.e. *eleven* values per source. With 32 evidence sources, this leads to an enormous number (11^{32}) of potential permutations. As a single ontology learning run (depending on

settings) takes around four hours of computation time, we decided to use the Tabu Search heuristic [7], and simply optimize every evidence source by itself, with settings for other sources constant. The leads to 352 ($11 * 32$) ontology learning runs. In the second batch, we used a step-size of 0.2 and 14 evidence sources, resulting in 84 ($6 * 14$) ontology learning runs.

The following Pseudocode shows the Tabu Search-based optimization strategy:

Algorithm 1 Optimize SIV with Tabu Search

```
Initial solution ← Static SIV
# do for all 32 evidence sources
for each evidence source e do
  # create neighborhood
  for X in interval [0.0, 1.0] size 0.1 do
    # evaluate every neighbor
    SIV_e ← X
    compute ontology (all 3 extension steps) using SIV
    Q_x ← Evaluate quality of ontology
    Remember result (X, Q_x)
  end for
  # keep value X with best result – skip the rest
  SIV_e ← pick best result from neighborhood
  Put all other solutions from neighborhood on Tabu list
end for
```

Basically, the heuristic looks for the *best* source impact value for a single evidence source, and uses this value when optimizing the other evidence sources. One of the downsides of this method is that the order in which evidence sources are processed obviously has an impact on the result. The system randomizes the order of evidence sources before every optimization run. It will typically not find a global optimum, but hopefully a good solution with a limited number of permutations. Furthermore, the optimization helps to visualize and understand how specific SIV settings contribute to ontology quality.

6.4 Evaluation

This section summarizes the findings of optimization runs performed to gain insights about the improvements of accuracy which can be reached by optimizing the combination (the impact) of evidence sources.

Evaluation Setup In previous experiments conducted in year 2014 (see Section 6.4 for results) we used a step-size of 0.1 in a source impact interval of 0.0 to 1.0, and 32 evidence sources. For the recent batch of evaluation experiments we used a more computationally efficient setup, with optimization runs for 14 evidence sources, and a step-size of 0.2. Previous work shows that 10-15 evidence sources are sufficient to have good results [17].

Domain	Static SIV	Optimized	Improvement
Climate Change (en)	67.15%	76.88%	9.73%
Tennis (en)	44.57	54.42%	9.85%

Table 9: System accuracy and gains by optimizing the SIV as compared to a static SIV, in two different domains.

Relevance assessment of concept candidates is being done by domain experts and crowd workers via the uComp API. The accuracy values used in this section are simply the number of concept candidates rated as domain-relevant by the human workers divided by all concept candidates suggested by the system. We decided to use the ratio of relevant concepts as evaluation metric because (i) the relevance of domain concepts is critical to generating useful domain ontologies, and (ii) relevance assessment for concept candidates is the only part of the system where manual input is applied.

Recent Optimization Experiments These experiments were conducted with the latest version of the ontology learning system in the first half of year 2015. We compared the results of using a static SIV (uniform source impact of 0.2 for all evidence sources) to optimizing source impact.

Table 9 summarizes the results for two different domains, the domains of *climate change* and of the sport *tennis*. The values in the table represent the average accuracy for ontology generation runs with a static SIV and for the optimization processes. The data indicates a substantial improvement in accuracy of around 10% which can be reached by optimizing the SIV.

The difference in accuracy between the two domains can be attributed to the following reasons: (i) In the *climate change* domain we are supplied with much bigger and domain-specific corpora, whereas with *tennis* we use general news corpora which are then filtered for tennis-related documents ex-post. Besides corpus-size and quality, (ii) the domain of tennis has a lot of overlap with other sports domains. Concepts such as ball, tournament, etc. attract related but not domain-relevant terms from other sports, whereas climate change seems to be more "closed".

However, the most interesting fact is the improvement in accuracy, which is statistically significant, as confirmed with a binomial test.

Previous Experiments This section discusses experiments done in early 2014 with an older version of the system. The system wasn't as well tuned then, and general accuracy was a bit lower, which is reflected by the accuracy values for *Static SIV* in Table 10. We did the evaluation for two settings, where we used either up to 50 (*limit=50*) evidences per source and concept, or up to 200 (*limit=200*)[3] – for details on these settings see Wohlgenannt [17].

In year 2014 we started from a lower baseline (around 63-64%), and experience improvements from optimization between 13-15%, more than in the recent batch of experiments. Our interpretation of the results is the following:

[3]The more recent evaluations in Section 6.4 were conducted with *limit=50* settings.

Domain	Static SIV	Optimized	Improvement
Climate Change (en) – *limit=50*	63.33%	78.18%	14.85%
Climate Change (en) – *limit=200*	64.13%	77.33%	13.20%

Table 10: System accuracy of the previous system version, in the domain of *climate change*, for two settings.

- The lower baseline leaves more room for improvement.
- In the 2014 experiments we used a *step-size* of 0.1, which resulted in higher computational cost, but also a more fine-tuned optimization.
- The number of evidence sources was much higher (32 sources), therefore the potential for fine-grained optimization and combination of sources was higher.

With regards to the research questions posed, the evaluation shows that system accuracy can be raised substantially by optimization using the SIV. It helps to have a high number of evidence sources and also a fine-grained step-size, this allows for a more precise optimization process.

Analysis of Evidence Sources The evidence sources provide terms and relations of different quality to the learning algorithms. Wohlgenannt [17] discusses the quality and characteristics of evidence sources in some detail. In a nutshell, the number and quality (domain relevance) of evidence is very heterogeneous. Keyword-based sources typically provide a high number of terms, with good quality for the terms with highest co-occurrence significance, but degrading with more terms added having a lower significance. Terms for structured sources such as DBpedia and WordNet generally offer good quality, but low term numbers. In our experiments, APIs of social sources such as Twitter and Flickr yield mostly low quality terms – but we still have them included to (i) benefit from the effect of redundancy between sources, and (ii) as they often provide very recent and complementary terminology.

Figure 8 visualizes the influence of source impact (SI) settings for some individual sources on system accuracy. The data is taken from one of the optimization runs in the domain of *climate change*, and helps to explain the characteristics and experiences with SIV optimization.

Usually evidence sources fall into one of the following categories:

- *Increasing the SI raises accuracy*. These evidence sources obviously yield relevant terms and helpful contributions to the ontology learning system. With higher impact of the source the accuracy goes up. `keywords:page:UK_media` and `keywords:page:climate_ngos` in Figure 8 fall into this category.
- *Increasing the SI lowers accuracy*. This applies to sources which do not contribute much helpful data. For example `keywords:sent:Fortune1000`.
- *Accuracy independent of SI*. This usually happens when a source provides a very low number of evidences, `social:Flickr` in our example.

- *Erratic*. As with `Hearst:Australian_media`, sometimes the effects of the SI are rather erratic. Such cases are the biggest challenge for the optimization algorithm.
- A mix of the basic categories described above.

Erratic behavior or a mix of the categories described above results from the fact that the system selects the 25 concept candidates with the highest spreading activation level. Raising the influence of a single evidence source gives more importance to all its evidence, relevant or not. The Tabu search heuristic will not find an optimal, but typically good, combination of sources (ie. the SIV).

Conclusions When using multiple and heterogeneous sources, balancing and optimizing the influence of evidence sources is crucial. In this paper, we introduce and evaluate a strategy for optimizing such ontology learning systems, and see improvements in accuracy (in the concept detection phase) of ca. 10-15%. The contributions are as follows: (i) Presenting a novel method to configure and optimize ontology learning systems using the source impact vector and the Tabu-search heuristic, and (ii) experiments in two domains to estimate the accuracy gains from this optimization technique. Future work includes the repetition of experiments in other domains, also based on corpora in other languages, and the application of alternative optimization strategies.

6.5 Conflict Meditation

Conflict mediation is on interesting and important topic in systems that learn formal structures like ontologies. Our system is focused on generating lightweight ontologies, where logical conflicts are not such an important issue, conflict meditation is centered rather around topics like:

- Evidence sources provide evidence for a huge amount of potential concept candidates, but our system only selects 25 concept candidates per ontology extension run. Which candidates should actually be selected?
- Concept candidates that have been selected by the system are assessed for domain relevance by using crowd workers. Usually multiple (default=5) crowd workers assess the relevance of a single concept candidates, which naturally leads to conflicting opinions about concept relevance.

The ontology learning system does candidate selection based on the activation level of nodes in the spreading activation network, we select the 25 nodes with the highest activation level. Therefore spreading activation can be understood as a simple tool for conflict meditation. The source impact vector (SIV), which determines the impact of individual evidence sources on the learning algorithm, also has an aspect of conflict meditation to it, as it favors evidence sources which provided useful information in the past.

When multiple workers assess a concept candidate (default number of votes is five), then we currently apply majority voting to come to a aggregated judgment. There has been some interesting work in the Human Computation community on using conflicting opinions as source of information. We display all votings, including their history, in the Web frontend. Future

work will (i) track the evolution of concept relevance, which can obviously change over time, and (ii) have a more detailed look at and analysis of concept candidates for which crowd workers have made conflicting assessments.

References

[1] Alani, H.: Position paper: ontology construction from online ontologies. In: Carr, L., Roure, D.D., Iyengar, A., Goble, C.A., Dahlin, M. (eds.) Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006. pp. 491–495. ACM (2006)

[2] Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S.: Ontology Learning from Text, chap. Learning Taxonomic Relations from Heterogeneous Sources of Evidence, pp. 59–76. IOS Press, Amsterdam (2005)

[3] Cimiano, P., Völker, J.: Text2onto. In: Montoyo, A., Muñoz, R., Métais, E. (eds.) NLDB. Lecture Notes in Computer Science, vol. 3513, pp. 227–238. Springer (2005)

[4] Crestani, F.: Application of spreading activation techniques in information retrieval. Artificial Intelligence Review 11(6), 453–482 (1997)

[5] David Manzano-Macho, A.G.P., Borrajo, D.: Unsupervised and domain independent ontology learning: Combining heterogeneous sources of evidence. In: Nicoletta Calzolari, e.a. (ed.) LREC'08. ELRA, Marrakech, Morocco (May 2008)

[6] Gacitua, R., Sawyer, P.: Ensemble methods for ontology learning - an empirical experiment to evaluate combinations of concept acquisition techniques. In: ICIS 08. pp. 328–333. IEEE Publishing (5 2008)

[7] Glover, F., Laguna, M.: Tabu Search. Kluwer, Norwell, MA, USA (1997)

[8] Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of COLING'92. pp. 539–545. Nantes, France (1992)

[9] Liu, W., Weichselbraun, A., Scharl, A., Chang, E.: Semi-automatic ontology extension using spreading activation. Journal of Universal Knowledge Management 0(1), 50–58 (2005)

[10] Lohmann, S., Link, V., Marbach, E., Negru, S.: WebVOWL: Web-based visualization of ontologies. In: Proceedings of EKAW 2014 Satellite Events. LNAI, vol. 8982, pp. 154–158. Springer (2015)

[11] Sanchez, D., Moreno, A.: Learning Non-taxonomic Relationships from Web Documents for a Domain Ontology Construction. Data and Knowledge Engineering 64(3), 600–623 (2008)

[12] Tatarintseva, O., Ermolayev, V., Keller, B., Matzke, W.E.: Quantifying ontology fitness in ontoelect using saturation- and vote-based metrics. In: Ermolayev, V., Mayr, H., Nikitchenko, M., Spivakovsky, A., Zholtkevych, G. (eds.) Information and Communication Technologies in Education, Research, and Industrial Applications, CCIS, vol. 412, pp. 136–162. Springer (2013)

[13] Velardi, P., Faralli, S., Navigli, R.: OntoLearn Reloaded: A Graph-based Algorithm for Taxonomy Induction. Computational Linguistics 39(3), 665–707 (2013)

[14] Völker, J., Vrandečić, D., Sure, Y., Hotho, A.: Learning disjointness. In: Proceedings of the 4th European Conference on The Semantic Web: Research and Applications. pp. 175–189. ESWC '07, Springer-Verlag, Berlin, Heidelberg (2007)

[15] Weichselbraun, A., Wohlgenannt, G., Scharl, A.: Augmenting lightweight domain ontologies with social evidence sources. In: Tjoa, A.M., Wagner, R.R. (eds.) 9th International Workshop on Web Semantics, 21st International Conference on Database and Expert Systems Applications (DEXA 2010). pp. 193–197. IEEE Computer Society Press, Bilbao, Spain (August 2010)

[16] Weichselbraun, A., Wohlgenannt, G., Scharl, A.: Refining non-taxonomic relation labels with external structured data to support ontology learning. Data & Knowledge Engineering 69(8), 763–778 (2010)

[17] Wohlgenannt, G.: Leveraging and balancing heterogeneous sources of evidence in ontology learning. In: Fabien Gandon, e.a. (ed.) ESWC 2015, Portoroz, Slovenia, May 31 - June 4, 2015. LNCS, vol. 9088, pp. 54–68. Springer (2015)

[18] Wohlgenannt, G., Weichselbraun, A., Scharl, A., Sabou, M.: Confidence management for learning ontologies from dynamic web sources. In: Proceedings of KEOD 2012. pp. 172–177. SciTePress, Barcelona, Spain (October 2012)

[19] Wohlgenannt, G., Weichselbraun, A., Scharl, A., Sabou, M.: Dynamic integration of multiple evidence sources for ontology learning. Journal of Information and Data Management (JIDM) 3(3), 243–254 (2012)

(%s) especially (\w+ ?\w*)	2	hypernym
(\w+ ?\w*) especially (%s)	1	hyponym
such (%s) as (\w+ ?\w*)	2	hyponym
such (\w+ ?\w*) as (%s)	1	hypernym
(%s) such as (\w+ ?\w*)	2	hyponym
(\w+ ?\w*) such as (%s)	1	hypernym
(%s) including (\w+ ?\w*)	2	hyponym
(\w+ ?\w*) including (%s)	1	hypernym
(%s), (?:a|an|the) (\w+ ?\w*)	2	synonym
(\w+ ?\w*), (?:a|an|the) (%s)	1	synonym
(\w+ ?\w*), (?:a|an|the) (%s)(?:\.|\?\!)	1	synonym
(%s), (?:a|an|the) (\w+ ?\w*)(?:\.|\?|!)	2	synonym
(\w+ ?\w*), (%s) (?:is|was|are|were)	1	synonym
(%s), (\w+ ?\w*) (?:is|was|are|were)	2	synonym
(\w+ ?\w*) (?:is|was|are|were|) (?:a|an|the) (%s)	1	synonym
(%s) (?:is|was|are|were|) (?:a|an|the) (\w+ ?\w*)	2	synonym
(%s) (\(\w+ ?\w*\))	2	synonym
(\w+ ?\w*) \(?:(%s)\)	1	synonym
(%s) (?:and|or) other (\w+ ?\w*)	2	hypernym
(\w+ ?\w*) (?:and|or) other (%s)	1	hyponym
(\w+ ?\w*), which (?:is|was|are|were) (%s)	1	synonym
(%s), which (?:is|was|are|were) (\w+ ?\w*)	2	synonym

Figure 6: Hearst patterns (appositions) and type of relation used for the English language

(%s) speziell (?:der|die|das|dem|den) (\w+ ?\w*)	2	hypernym
(\w+ ?\w*) speziell (?:der|die|das|dem|den) (%s)	1	hyponym
(%s) besonders (?:der|die|das|dem|den) (\w+ ?\w*)	2	hypernym
(\w+ ?\w*) besonders (?:der|die|das|dem|den) (%s)	1	hyponym
(%s) im Speziellen (?:der|die|das|dem|den) (\w+ ?\w*)	2	hypernym
(\w+ ?\w*) im Speziellen (?:der|die|das|dem|den) (%s)	1	hyponym
(%s) im Besonderen (?:der|die|das|dem|den) (\w+ ?\w*)	2	hypernym
(\w+ ?\w*) im Besonderen (?:der|die|das|dem|den) (%s)	1	hyponym
genau (%s) wie (?:der|die|das|dem|den) (\w+ ?\w*)	2	hyponym
genau (\w+ ?\w*) wie (?:der|die|das|dem|den) (%s)	1	hypernym
(%s) genauso wie (?:der|die|das|dem|den) (\w+ ?\w*)	2	hyponym
(\w+ ?\w*) genauso wie (?:der|die|das|dem|den) (%s)	1	hypernym
(\w+ ?\w*) einschließlich (?:der|die|das|dem|den) (%s)	1	hypernym
(%s) einschließlich (?:der|die|das|dem|den) (\w+ ?\w*)	2	hyponym
(%s), (?:ein|einer|eines|der|die|das) (\w+ ?\w*)	2	synonym
(\w+ ?\w*), (?:ein|einer|eines|der|die|das) (%s)	1	synonym
(\w+ ?\w*), (?:ein|einer|eines|der|die|das) (%s)(?:\.|\?|!)	1	synonym
(%s), (?:ein|einer|eines|der|die|das) (\w+ ?\w*)(?:\.|\?|!)	2	synonym
(\w+ ?\w*), (%s) (?:ist|war|sind|waren)	1	synonym
(%s), (\w+ ?\w*) (?:ist|war|sind|waren)	2	synonym
(\w+ ?\w*) (?:ist|war|sind|waren|) (?:ein|einer|eines|der|die|das) (%s)	1	synonym
(%s) (?:ist|war|sind|waren|) (?:ein|einer|eines|der|die|das) (\w+ ?\w*)	2	synonym
(%s) (\(\w+ ?\w*\))	2	synonym
(\w+ ?\w*) \(?:(%s)\)	1	synonym
(%s) (?:und|oder) andere (\w+ ?\w*)	2	hypernym
(\w+ ?\w*) (?:und|oder) andere (%s)	1	hyponym
(\w+ ?\w*), (?:welcher|welches|welche) (?:ist|war|sind|waren) (%s)	1	synonym
(%s), (?:welcher|welches|welche) (?:ist|war|sind|waren) (\w+ ?\w*)	2	synonym

Figure 7: Hearst patterns (appositions) and type of relation used for the German language

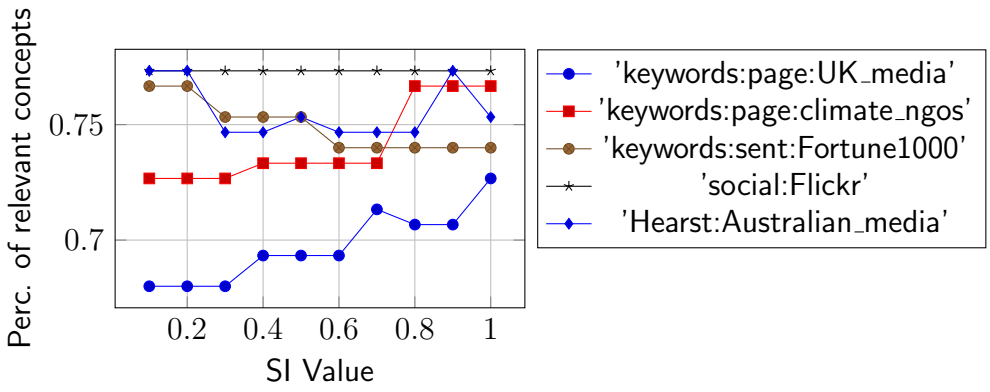

Figure 8: Influence of Source Impact settings for a number of selected evidence sources.