

EnviLOD WP5: Quantitative Evaluation of LOD-based Semantic Enrichment on Environmental Science Literature

Kalina Bontcheva^{1,*}, Niraj Aswani¹, Johanna Kieniewicz², Stephen Andrews², Michael Wallis³

1 Department of Computer Science, University of Sheffield, Sheffield, UK

2 The British Library, London, UK

3 HR Wallingford, Wallingford, UK

* E-mail: K.Bontcheva@dcs.shef.ac.uk

Abstract

This EnviLOD deliverable reports on the quantitative evaluation of the LODIE automatic semantic enrichment algorithm. Firstly, we outline the LODIE approach for LOD-based semantic enrichment of metadata and full-text scientific articles. Secondly, the methods are evaluated, both quantitatively and qualitatively by our British Library users. For the latter, semantic search queries are compared against the full-text search capabilities of the Envia British Library information discovery tool and tentative benefits of LOD-based semantic enrichment are identified.

The resulting richer content underpins the EnviLOD semantic search interface, which was then evaluated much more extensively with users at two different workshops. The results from the latter evaluation appear in the EnviLOD WP2 User Feedback report.

1 Introduction

Semantic annotation is the process of tying semantic models, such as ontologies, and scientific articles together. It may be characterised as the dynamic semantic enrichment of unstructured and semi-structured documents and linking these to relevant domain ontologies/knowledge bases. From a text mining perspective, semantic annotation is about annotating in texts all mentions of concepts from the ontology (i.e., classes, instances, properties, and relations), through metadata referring to their Unique Resource Identifiers (URIs) in the ontology.

Semantic annotation with respect to ontological resources can be broken down into two phases: *candidate selection* and *resource linking* (also called reference disambiguation) [RMD13, JG11]. Candidate selection is concerned with identifying in text all candidate mentions of instances from a knowledge base (e.g. DBpedia). The resource linking/disambiguation step then uses contextual information from the text, coupled with knowledge from the ontology, to choose the correct instance URI. If there is no such corresponding instance, then a NIL value needs to be returned (an open domain assumption). In particular, the disambiguation step needs to handle name variations (instances can be referred to in many different ways) and ambiguity (the same string can refer to more than one instance) [JG11, RMD13].

LOD resources, in particular DBpedia and GeoNames, have emerged as key sources of large-scale ontological knowledge, as well as being used as target entity knowledge bases for semantic

enrichment and resource linking. They offer:

- cross-referenced domain-independent hierarchies with thousands of classes and relations and millions of instances;
- an inter-linked and complementary set of resources with synonymous lexicalisations;
- grounding of their concepts and instances in Wikipedia entries and other external data.

The rich class hierarchies are used for fine-grained classification of named entities while the knowledge about millions of instances and their links to Wikipedia entries are used as features in the resource linking and disambiguation algorithms. However, as noted by [GNP⁺09], the large-scale nature of LOD resources also makes resource linking particularly challenging, due to the ambiguity introduced by the presence of millions of instances.

This deliverable presents an overview of the LODIE semantic enrichment tool, adapted in EnviLOD to scientific articles, reports, and metadata. Our LOD-based semantic enrichment algorithm is designed to improve recall, a common problem of other state-of-the-art approaches. Disambiguation is carried out on the basis of string, semantic, and contextual similarity, coupled with a popularity metric. The algorithm is developed on a general purpose, shared news-like corpus and evaluated on environmental science papers and metadata records from the British Library. A preliminary user-based analysis of the impact of semantic enrichment on information discovery in scientific literature is also presented.


1.1 Environmental Science Context

The environmental information landscape is both wide-ranging and disparate. The interdisciplinary nature of environmental science as a subject matter, the breadth of types of materials that are published, and the lack of a widely used controlled vocabulary (e.g. the Unified Medical Language System (UMLS) in biomedical sciences [SPMM11]) means that information discovery within the field of environmental sciences can often be a fraught and difficult process [KSN11]. A British Library survey of 107 flooding researchers and practitioners found that beyond the information access barriers of time and the costs associated with subscription journal articles and databases, they struggle with the process of information filtering. Searches tend to return either far too many or too few results [RP05]; a case is thus made for improved discovery of environmental information. Semantic technologies, as are discussed in this deliverable, offer new opportunities to improve the process of information discovery, providing researchers more meaningful search results.


A new British Library information discovery tool for environmental science, Envia [KSN11], is used as a case study to test the use of semantics towards enhancing information discovery and management. It provides a case through which to examine the value of LOD-based semantic enrichment from both the perspective of end-users interested in information discovery, as well as that of information managers. Envia is particularly suited for these purposes, as it features a mixed corpus of content, including datasets, journal articles, and grey literature, with accompanying metadata records (see Figure 1). Most Envia entries have very little metadata on import, so one immediate benefit of text mining is the automatic enrichment with LOD terms and entities, based on the semantic annotation algorithms described here.

FRM Act annual report to Parliament ... : Flood Risk Management (Scotland) Act 2009
 EDINBURGH : SCOTTISH GOVERNMENT

DOWNLOAD DOCUMENT







LINK

 <http://www.scotland.gov.uk/Publications/Recent>

CONTENT TYPE
 continuing
 Electronic

DATE
 2010

SUBJECT

-  Flood damage prevention--Scotland--Periodicals
-  Floods--Risk assessment--Scotland--Periodicals
-  Flood control--Scotland--Planning--Periodicals
-  Other social problems and services











-  Flood
-  Act
-  Scotland
-  risk
-  plans
-  Risk Management
-  Scottish Government
-  local authorities
-  National Park
-  flood protection

Figure 1: **An example Envia record.** The highlighted metadata has been added automatically and includes terms and locations in this case.

As a service still in development, Envia has an initial content focus on information pertaining to flooding, an environmental science subject domain that spans hydrology, geology, civil engineering and planning, and is relevant to a broad cross-section of the environmental community, including academics, government agencies, local authorities and charities. Use cases for semantic technologies within this context must therefore account for the the needs of this wide-ranging community. To this end, it is necessary to understand the types of queries that users are likely to make, the way they are likely to phrase their queries, and their expectations in terms of the types of search results that are returned.

We carried out an online survey of 36 members of the flooding community. The results showed that researchers and practitioners could roughly be split into 'high level' users interested in policy, guidance and funding, and 'local level' users, interested in areas at risk, practical implementation, and technology. Amongst both groups, geographic information was of primary importance and comes across in questions, such as 'Where has flooding taken place since 2007'; or 'What is the annual expenditure on flood risk management in The Netherlands?'. As such, location, proximity, and measurement were identified as important requirements in establishing the potential of LOD vocabularies to address such user queries.

1.2 Related Work

1.2.1 Semantic Enrichment of Environmental Science Literature

Within the sphere of environmental science, the area with the greatest legacy of semantic enrichment is that of geospatial information [JSPHss], with applications including GIS environments/Spatial Data Infrastructures (SDI), environmental sensor networks and geotagging (see papers in [PSS11]). These approaches all identify interdisciplinary datasets, as are commonly found in environmental science, as a particularly fruitful area for LOD exploration. In these contexts dataset metadata is semantically enriched in order to improve search and enable correct use of data [SPMM11]. The LOD GEMET thesaurus underpins the EU INSPIRE directive¹, which aims to establish a digital infrastructure for spatial information in Europe in order to support environmental research, policy and decision-making. This ties into the Open Data movement and `data.gov.uk` which is being used as a vehicle through which the UK might comply with INSPIRE requirements for making environmental data available and discoverable [SWC⁺11].

Although progress is being made in environmental informatics with respect to enabling the discovery and better use of datasets and geographic information within the GIS/SDI context, LOD vocabularies have not as yet been applied in the context of text mining of environmental science literature. This contrasts with the biomedical sciences where text mining has been enabled by the Unified Medical Language System, a meta-thesaurus provided by the US National Library of Medicine, which acts as a comprehensive thesaurus and ontology of biomedical concepts [HvMS⁺10]. LOD resources offer an opportunity to realise the potential of environmental vocabularies in text mining by drilling down into the lexical meaning of phrases using inter-linked vocabularies, ultimately resulting in search results that are more relevant.

1.2.2 LOD-based Methods for Semantic Enrichment

There are a number of state-of-the-art methods for semantic annotation and linking to DBpedia (e.g. DBpedia Spotlight [MJGSB11]), YAGO (e.g. [SWLW12]), and MusicBrainz (e.g. [GNP⁺09]). These LOD-based entity linking approaches have their roots in methods that enrich documents with links to Wikipedia articles (e.g. [MW08, RMD13, JG11]). In addition, commercial web services such as AlchemyAPI, OpenCalais, and Zemanta are also relevant. A recent evaluation [RT11] of all state-of-the-art LOD-based methods and tools, showed that DBpedia Spotlight and Zemanta have the best accuracy on annotating texts with the corresponding URIs from DBpedia.

Our own evaluation of DBpedia Spotlight and Zemanta (see the top two rows of Table 3) indeed confirmed that they both produce high accuracy results on assigning DBpedia URIs to mentions of person names, organisations, and locations. Spotlight however, suffers from low coverage (i.e. recall), which is not ideal, given that our task is to enrich environmental science literature with links to DBpedia and GeoNames. Zemanta's coverage is better than Spotlight, but still not as high as our environmental science use case requires.

However, both Zemanta and DBpedia Spotlight are only available as semantic annotation services over the web. Therefore, we implemented a modular, LOD-based semantic enrichment

¹<http://inspire.jrc.ec.europa.eu>

method, which is designed specifically to maximise the recall of the initial candidate entity generation step. This tends to result in higher entity ambiguity, thus making the disambiguation task harder and leading to somewhat lower precision. In order to address this problem, we combine the outputs of the high-precision Zemanta service with the semantic annotations produced by our own method, leading to best overall results and more balanced accuracy and coverage. More details on this approach follow next.

2 The LODIE Semantic Enrichment Tool

2.1 LOD Resources and Annotation Types

Based on the survey results, we identified five key types of entities, that need to be identified automatically:

1. **Location:** these include not just the place name itself (e.g. Norwich), but also the implied reference to the levels 1, 2, and 3 sub-divisions from the Nomenclature of Territorial Units for Statistics (NUTS)². For Norwich, these are East of England (UKH – level 1), East Anglia (UKH1 – level 2), and Norfolk (UKH13 – level 3).
2. **Organisation:** names of companies, government organisations, committees, agencies, universities, and other organisations.
3. **Person:** names of Johanna Kieniewicz², of people who authored publications in Envia, as well as person names mentioned within the full-text content itself (e.g. committee members).
4. **Date:** absolute dates like ‘October 2012’ or ‘2007’, as well as relative dates, such as ‘last year’. Where not available, document dates need to be assigned automatically and used for the normalisation of relative dates.
5. **Measurements:** research in environmental science is highly dependent on quantitative measurements, ranging from measurements of rainfall, sea level rise, and stream flow (velocity) to measurements pertaining to geographic areas and proximity (e.g. 8,596 km², 1 km, one fifth). Particularly in the area of flooding, percentages and probabilities are also important (e.g. 1% annual probability, 200 to 1 chance, 10% or greater chance of extreme rainfall).

Given the target kinds of entities, we identified as most relevant two general-purpose large-scale LOD resources (DBpedia and GeoNames);, as well as several domain-specific ones (GEMET and the OS Hydrology ontology). The latter are used for the recognition of domain-specific terms. In more detail:

- **DBpedia** [BLK⁺09] encodes knowledge about 3.5 million entities, amongst which 410,000 places, 310,000 persons, and 140,000 organisations, which map directly to the first three target entity types above. Entity name variants, a textual abstract, and reference(s) to

²For the UK classification see http://en.wikipedia.org/wiki/NUTS_of_the_United_Kingdom

corresponding Wikipedia page(s) are also included, as well as entity-specific properties (e.g. latitude and longitude for places).

- **Geonames** represents 2.8 million populated places and 5.5 million alternate names, and is thus relevant only to the semantic annotation of locations. It also includes knowledge about NUTS country sub-divisions, which we use for enrichment of recognised locations with the implied higher-level country sub-divisions.
- **GEMET**: The General Multilingual Environmental Thesaurus³, developed for the European Environment Agency, is the leading environmental thesaurus, created through the compilation and linking of sector-specific and language-specific environmental vocabularies. It contains over 6000 environmental topics in 27 different languages and is available via the web in Linked Data format.
- **Ordnance Survey Hydrology**: The Ordnance Survey (OS) Hydrology Ontology was developed to support the use of their data by their customers and provide an unambiguous semantic framework for the description of and relations between inland hydrological features on OS maps [DK10].

Domain-specific LOD resources, such as the GEMET thesaurus⁴ and the Ordnance Survey Hydrology ontology [DK10], contain mostly environmental science terminology and are thus used for term-based enrichment. This, however, is outside the scope of this deliverable, which is on evaluation of entity-based enrichment.

The recognition of dates and measurements is also out of the scope of this deliverable, since we reused already existing pattern-matching grammars [AAB⁺08].

The focus of this report is therefore on the semantic enrichment of environmental science literature with knowledge from DBpedia and GeoNames. Figure 2 shows an online government report, indexed in Envia, where locations and organisations have been annotated automatically with such LOD-based semantic information.

The rest of this section presents the underlying entity annotation and disambiguation algorithm, as implemented in the LODIE semantic enrichment tool.

2.2 Identifying Candidate LOD Instances

The first step is to identify all candidate instance URIs from DBpedia, which are mentioned in the given document. This candidate generation step uses the lexical information associated with LOD instances, in order to build very large gazetteer lists. These are then used to perform lookups of n-grams derived from the document text. The DBpedia lexicalisation properties used in our experiments are `rdfs:label`, `db:name`, `foaf:nick`, `db:nickname`, `db:official_name`.

The DBpedia-based candidate selection was implemented using the open-source GATE Large Knowledge Gazetteer (LKB)⁵ [CMB⁺11]. LKB performs fast string lookup and assigns URIs to words/phrases in the text. It is initialised using a SPARQL query evaluated against the target

³<http://www.eionet.europa.eu/gemet/>

⁴<http://www.eionet.europa.eu/gemet/>

⁵<http://gate.ac.uk/userguide/sec:gazetteers:lkb-gazetteer>

Type	Set	Start
Sem_Location		57
Sem_Location		97
Sem_Location		97
Sem_Location		97
Sem_Location		149
Sem_Location		160
Sem_Location		160

alternateName	South Gloucestershire
caption	South Gloucestershire
count	2
countryCode	GB
geonamesURI	http://sws.geonames.org/3333198/
inst	http://dbpedia.org/resource/South_Gloucestershire
latitude	51.5
longitude	-2.41667
lookupRule	fullString
matched	South Gloucestershire
name	South Gloucestershire
parentAdminURI	http://sws.geonames.org/6269131/, http://sws.geonames.org/3333198/
parentCountryInst	http://sws.geonames.org/2635167/
popularitySimilarity	1.0
randomIndexing	0.0
specificitySimilarity	0.0
string	South Gloucestershire
stringSimilarity	0.2688679
structuralSimilarity	0.0

Figure 2: **An example location annotation and associated semantic information.** The URIs of the respective GeoNames instance, as well as GeoNames information on latitude, longitude, subsuming administrative areas and parent country, are added.

LOD resource endpoint (DBpedia in this case), including only instances of the target entity classes. In our case, the target high-level classes are Place, Organization, and Person, and all their sub-classes. The latter are given as parameters in the SPARQL query, which makes the algorithm customisable, similar to DBpedia Spotlight [MJGSB11].

A sample snippet used to initialise the Person, Location and Organisation LKB gazetteer is:

```
dbpedia:Paris dbpedia-ont:City "Paris"
dbpedia:Paris,_Texas dbpedia-ont:City "Paris"
dbpedia:Paris_Hilton dbpedia-ont:Person "Paris"
```

Each mention of the string *Paris* in text would then be marked as having these candidate instance URIs.

For DBpedia (which has mappings to Wikipedia pages), additional lexicalisations for each URI are acquired from link anchor texts, disambiguation pages, and redirect pages from Wikipedia. This has been shown [JG11,RMD13] to help improve the recall of the candidate selection phase. For instance, many abbreviations and acronyms are acquired in this way.

We have also used the open-source ANNIE Information Extraction system [CMB⁺11] to assign entity types for entity candidates. ANNIE also resolves within-document coreference, so that mentions of the same entity within a document are linked together. For example, *European Environmental Agency* and *EEA* would be marked as referring to the same entity. This coreference information is used to restrict the textual context considered in the subsequent entity disambiguation stage.

In addition, the ANNIE results are used to filter out incompatible entity candidates. Firstly, if the NER component assigns the type Location to the string Paris in a given document, then only instances of class `Place` and its subclasses are retained as candidates for disambiguation. In our example above, this means that only `db:Paris` and `db:Paris,_Texas` will remain. Secondly, tokens and noun phrases produced by ANNIE, are used to filter out entity candidates which do not align with word boundaries and/or do not contain a noun phrase. Lastly, ANNIE helps with identifying mentions of locations, people, and organisations, which do not appear in DBpedia.

2.3 Entity Disambiguation

The entity disambiguation algorithm uses the textual context, in which a given candidate entity appears, in order to calculate a number of similarity metrics. These are combined in a weighted sum, to produce an overall score for each candidate URI. The three metrics are:

- *String similarity*: edit distance between the text string (such as *Paris*), and the lexicalisations of the entity URIs (e.g. *Paris* and *Paris, Texas*).
- *Semantic (structural) similarity*: calculated based on the ontology and instance property values in the LOD resource.
- *Contextual similarity*: the probability that two words have a similar meaning, based on random indexing.
- *Commonness*: a normalised frequency metric against Wikipedia.

Tie-breaks, i.e. candidate URIs for the same textual mention and with the same overall score, are resolved based on which one has the highest commonness score. For example, if the overall score is the same, `db:Paris` will be chosen over `db:Paris,_Texas`, since it appears much more frequently in Wikipedia articles. If nevertheless more than one candidate remains, the instance which has a more specific class in the LOD ontology is preferred.

2.3.1 String Similarity

For each candidate URI, string similarity is calculated using a context of 30 tokens on both sides of the candidate, including all sentences from any co-reference chain. For efficiency reasons, only named entities are used. After some experiments with different string similarity metrics (Levenstein, Jaccard, and MongeElkan [JBGG09]), the Levenstein (or string edit distance metric) was chosen.

In a nutshell, the Levenstein score of two strings is equivalent to the number of substitutions and deletions needed to transform one string into the other. More formally, let \mathbf{s} be the source string and let \mathbf{t} be the target string. The distance is the number of deletions, insertions, or substitutions required to transform \mathbf{s} into \mathbf{t} . If \mathbf{s} and \mathbf{t} are identical, then $LD(\mathbf{s}, \mathbf{t}) = 0$, because no transformations are needed. If \mathbf{s} is “matrics” and \mathbf{t} is “metrics”, then $LD(\mathbf{s}, \mathbf{t}) = 1$, because one substitution is sufficient to transform \mathbf{s} into \mathbf{t} . The greater the Levenshtein distance, the more different the strings are.

String similarity measures the distance between two text strings. For example, a spelling error could be discovered by comparing *metrics* with *matrics*, as the two strings are very similar.

There are several widely adapted algorithms each of which is suitable for different kinds of tasks. We decide to use the following:

Levenshtein Distance or Edit Distance operates between two input strings, returning a score equivalent to the number of substitutions and deletions needed in order to transform one string into another. Let s be the source string and let t be the target string. The distance is the number of deletions, insertions, or substitutions required to transform s into t . If s and t are identical, then $LD(s,t) = 0$, because no transformations are needed. If s is "matrics" and t is "metrics", then $LD(s,t) = 1$, because one substitution (change "a" to "e") is sufficient to transform s into t . The greater the Levenshtein distance, the more different the strings are.

"This is a character-based measure as it considers the strings to be compared merely as character sequences, which makes this approach affordable when the strings to be compared are single words having misspellings, typographical errors, OCR errors, or even some morphological variations." [JBGG09].

Jaccard Similarity is defined as the quotient between the intersection and the union of the pairwise compared variables among two objects. When applied to two strings, it can be defined as the quotient between the intersection and the union of the pairwise compared variables among two set of tokens. In other words, the two strings are first tokenised, and then the Jaccard similarity is calculated by dividing the number of tokens shared by the strings by the total number of tokens.

"The token-based measures compare text strings as sequences of tokens instead of sequences of characters. Such an approach is successful when it is used to compare text strings with many tokens and with different order of the tokens or missing tokens." [JBGG09]

MongeElkan is a hybrid string similarity measure as it compares tokens using an internal static character-based measure [JBGG09]. Hence it preserves the properties of the internal character-based measure such as dealing with misspellings, but it also deals with missing or disordered tokens. This measure is especially useful when comparing long URIs or camelCased local names that contain useful strings, with meaningful words. For example, the similarity between *cityPopulation* and *population* will be high according to MongeElkan, while it will be low with Jaccard due to the dissimilarity between tokens. In fact, the Monge-Elkan method is a general and recursive token similarity method that can combine any token comparison measure, which captures semantics, translations, etc."

To illustrate the differences between the three different string similarity metrics we illustrate an example in Table 1.

Note that in the given example, the Jaccard similarity operates on the token level, where in this example we used the white space tokeniser, meaning that the punctuation is ignored. For more accurate results, it is possible to use any other tokeniser.

	Levenshtein	Jaccard	MongeElcan
“Paris” and “Paris Hilton”	0.42	0.5	1.0
“Paris” and “Paris, Ontario”	0.36	0.0	1.0
“Paris Hilton” and “Paris, Ontario”	0.43	0.0	0.63

Table 1: The differences between string similarity metrics

2.3.2 Semantic (Structural) Similarity

Semantic (structural) similarity is calculated based on whether the ambiguous candidate NE has a relation with any other NE from the same sentence or document. For example, if the document mentions both *Paris* and *France*, then semantic similarity assigns the highest score to `db:Paris`, as the two are connected directly via the `db:country` property. On the other hand, if *Paris* appears in the context of *USA*, the semantic similarity metric will assign higher score for `db:Paris,_Texas`. The latter is derived by combining the DBpedia knowledge that Paris, Texas is part of Lamar County and the latter has country United States.

We use DBpedia as the knowledge base as it is updated regularly with the latest content from Wikipedia, and it is also a good source for named entities, without being overly verbose. The DBpedia ontology gives a reasonably granular hierarchy of concepts which can be useful when identifying what exactly the URI refers to (e.g. whether it is a location, city or a capital).

Semantic similarity scores are first computed against any unambiguous URI instances, found in the context. If this fails to produce results, other NE candidates from the context are prioritised, based on how close they are to the ambiguous entity (measured in number of intermediate tokens) and on which side of the entity they are (left context vs right context).

For efficiency reasons, context size is limited to 60 tokens around the candidate (30 to the left and 30 to the right), as well as all sentences including co-referent mentions of this candidate (as determined by ANNIE). Nevertheless, on larger document sets this could lead to a large number of SPARQL queries needing to be fired. This was optimised through caching.

To clarify the semantic similarity scoring, we give a high level pseudo code:

```

Identify all NE candidates in a document
for each ambiguous NE candidate = NE_c:
  1. Identify any coreference chain NE_c belongs to (given by ANNIE)
  2. Define context as a set of sentences S which cover the coreference chain
     and the 30 tokens to the left and right of NE_c
  3. Find contextURIs: identify all other NE URIs in the context S
  4. Order contextURIs as follows:
     - unambiguous URIs to the left of NE_c
     - unambiguous URIs to the right of NE_c
     - sort the remaining URIs by distance to NE_c
     - where two URIs are equidistant, put the left context one first
  4. relationCount=0;
  5. until relationCount>0 or contextURIs.isEmpty(){
     for each contextURI in contextURIs {
       relationCount=
         select count(?relation) WHERE {
           (NE_c.URI ?relation contextURI)

```

```

        UNION
        (
            NE_c.URI ?relation1 ?anyResource.
            ?anyResource ?relation contextURI
        )
    }
}

```

For our example, if the two concepts `Paris...France` appear in the same sentence then `relationCount` will become equal to 1 in the first iteration, and the algorithm will stop due to DBpedia having the triple:

```
dbpedia:Paris dbpedia:capitalOf dbpedia:France
```

If the two concepts do not appear in the same sentence but the reference to Paris appears in another sentence with France, the algorithm will take more iterations, but will eventually return true.

2.3.3 Contextual Similarity

For calculating contextual similarity, we use Random Indexing (RI) [Sah05] and calculate the scores using cosine similarity. For efficiency reasons, we indexed only DBpedia abstracts as *context* for each URI which refers to either `dbpedia:Person`, `dbpedia:Organisation` or `dbpedia:Place`. This means that our initial *term* × *document* matrix looks as follows:

	term1	term2	...	termN
dbpedia:Paris	5	4
dbpedia:London	3	12
...				

RI can be seen as an approximation to Latent Semantic Analysis (LSA) [DDF⁺90], which is one of the pioneer methods for finding synonyms automatically. The assumption behind it is that words which appear in a similar context (with the same set of other words) are synonyms. Synonyms tend not to co-occur with one another directly, so indirect inference is required to draw associations [CSW09]. LSA has been shown to approximate human performance in many cognitive tasks such as the Test of English as a Foreign Language (TOEFL) synonym test, the grading of content-based essays and the categorisation of groups of concepts (see [CSW09]). However, one problem with LSA is scalability: it starts by generating a *term* × *document* matrix which grows with the number of terms and the number of documents and will thus become very large for large corpora. For finding the final LSA model, Singular Value Decomposition (SVD) and subsequent dimensionality reduction is commonly used. This technique requires the factorization of the term-document matrix which is computationally costly and does not scale well. Also, calculating the LSA model is not easily end efficiently doable in an incremental or out-of-memory fashion.

The Random Indexing method [Sah05] circumvents these problems by avoiding the need for matrix factorization. It has been shown to reach results, similar to LSA [KS01, CH08]. RI can

be updated incrementally and also, the *term * document* matrix does not have to be loaded in memory at once – loading one row at a time is enough for computing the context vectors. Instead of starting with the full term-document matrix and then reducing the dimensionality, RI starts by creating almost orthogonal random vectors (index vectors) for each document. This random vector is created by setting a number of randomly selected dimensions to either +1 or -1. Each term is represented by a term vector, which is a combination of all index vectors of the document in which it appears. For an object consisting of multiple terms (e.g. a document or a search query), the vector of the object is the combination of the respective term vectors.

Random Indexing relies on the Johnson-Lindenstrauss lemma:

Lemma 2.1 *Given $0 < \epsilon < 1$, a set X of m points in R^N , and a number $n > n_0 = O(\frac{\log(m)}{\epsilon^2})$, there exists a mapping $f : R^N \rightarrow R^n$ such that $(1 - \epsilon)\|u - v\| \leq \|f(u) - f(v)\| \leq (1 + \epsilon)\|u - v\|$, for all $u, v \in X$.*

and particularly on the proof provided by Johnson and Lindenstrauss in their 1984 article [JL84], where they show that if one chooses at random a rank n orthogonal projection, then, with positive probability, the projection restricted to X will satisfy the condition in the Lemma. RI relies on the observation that, in a high dimensional space, a random set of vectors is always almost orthogonal.

There are several parameters which can influence the process of generating semantic index, or vectors using the RI method:

- **Seed length:** Number of +1 and -1 entries in a sparse random vector.
- **Dimensionality** of the semantic vector space – predefined number of dimensions to use for the sparse random vectors.
- **Minimum frequency** for a term to be included in the index.

Generating and searching through the semantic space is computationally costly and in order to make it more efficient we pre-processed the abstracts and included only proper and common nouns in the *term × document* matrix. The corpus contained 3 million terms and 3.5 million documents. We used dimensionality of 150, seed length of 4 and minimum term frequency of 3, which reduced the semantic space to around 1.2 million terms. The selection of these parameters is based on our earlier experiments with DBpedia [DSL12].

Once the semantic space was computed, we used it to find terms related to the specific documents (URIs). When calculating the contextual similarity score, for a given candidate URI, we first retrieve the top 20 related terms, and then calculate cosine similarity with the context of the ambiguous NE.

The high level pseudo code looks as follows:

1. For ambiguous NE: search the semantic space where input is the candidate NE URI, and the output is the top 20 related terms for that URI
2. Extract the context of the NE: all proper and common nouns in the sentences in which this NE appears (including any co-reference chains)
3. Calculate the cosine similarity between the 20 terms and the context

2.4 Commonness

The commonness metric reflects the assumption that if a named entity is mentioned frequently in Wikipedia, then it will be also more common within other corpora. Due to the one-to-one mapping between English Wikipedia URLs and DBpedia URIs, the commonness score for a candidate URI is assigned using the *commonness* metric defined by [MW08] for Wikipedia pages. This has also been referred to as popularity [RMD13, ACJ⁺09]. However, unlike [RMD13], for efficiency reasons we do not use Google queries as additional evidence.

The pseudo code is as follows:

Pre-processing stage:

```
for each DBpedia_Instance_URI of class Place, Person or Organization {
  map DBpedia_Instance_URI to corresponding Wikipedia_Page_URL
  FreqCountURI = number of Wikipedia link anchors pointing to Wikipedia_Page_URL
  index <DBpedia_Instance_URI, FreqCountURI>
}
```

Run-time calculations:

```
for each ambiguous NE, take all candidate DBpedia URIs (candidateURIs)
  for each candidateURI {
    commonness_candidateURI= FreqCountURI / (sum of FreqCountURIs for all candidates)
  }
```

2.5 Semantic Enrichment

The result of the semantic annotation and disambiguation algorithm is text, enriched with mentions of DBpedia URIs – one URI per named entity mentioned. However, DBpedia and GeoNames contain a lot of relevant knowledge about these URIs, e.g. latitude and longitude for locations. The semantic enrichment process brings such additional relevant knowledge and associates it with the URI mentions in the text.

In more detail, all mentions of DBpedia class `Place` are enriched with additional knowledge from the corresponding GeoNames instance. The latter is found as the value of the `same_as` property, pointing to a GeoNames URI. This GeoNames URI is then used in a SPARQL query to obtain the country code, parent country, latitude and longitude, and all parent administrative regions. For example, Figure 2 shows this additional semantic information for the mention of the entity South Gloucestershire.

The rationale behind the semantic enrichment stage is that it enables better location-based searches (for details, see the Discussion section). The need to use GeoNames as the knowledge source, instead of DBpedia alone, is that it contains richer information, especially about administrative regions.

3 LODIE Evaluation Results

3.1 Development and Evaluation Datasets

We made use of two entity disambiguated corpora, containing diverse text genres.

The first one is the open-domain TAC-KBP 2010 corpus [JG11], which contains diverse genres: broadcast news, conversations, speech, newswire, and web text. It is created by the Knowledge Base Population (KBP) track of the Text Analysis Conference (TAC). Given a query that consists of a name string (person (PER), organization (ORG), geo-political entity (a location with a government), or unknown) - and a background document ID, the system is required to provide the ID of the KB entry to which the name refers; or NIL if there is no such KB entry [JG11].

Table 2: **Entity candidate selection statistics**

	TAC-KBP					EnvTest		
	PER	LOC	ORG	UKN	TOTAL	LOC	ORG	TOTAL
Entities	89	361	141	274	865	175	87	262
Avg. number of tokens	1.91	1.20	2.12	1.87	1.78	1.38	2.9	2.14
Candidate URIs	9,427	9,553	9,502	14,649	43,131	1,554	720	2,274
Avg. number cand. URIs	105.02	26.46	67.39	53.46	49.86	8.88	8.28	8.68
Unambig. candidates	3	10	3	43	59	13	23	36

We mapped manually the IDs of the KB entries onto DBpedia URIs, where these existed. This resulted in 375 unique URI references (see Table 2). Some of these have multiple mentions in the corpus (865 in total).

The TAC-KBP corpus was used for algorithm development, testing and weight parameter tuning. We also used it to compare the performance of our LOD-based semantic annotation method against other state-of-the-art approaches.

Secondly, since there is no pre-existing corpus of semantically enriched environmental science literature, we chose at random 100 documents from Envia and annotated them manually to create a gold standard for evaluation (referred to as EnvTest). For the semantic annotation and enrichment algorithm, EnvTest constituted unseen data used purely for quantitative evaluation.

In more detail, the EnvTest corpus contains 90,278 tokens, 450 mentions of entities (177 Organisations, 273 Location), and 262 unique entity URIs (175 locations and 87 organisations). In other words, EnvTest annotations refer to 262 different entities in DBpedia and each entity is mentioned on average 1.7 times in the corpus. The EnvTest gold standard currently does not include person URIs from DBpedia, since the person names appearing in the documents are not present in DBpedia. These are predominantly author names for reports, papers, and references to other papers, so are recognised as person annotations by ANNIE, but cannot be disambiguated to a DBpedia URI.

3.2 Development and Parameter Tuning on TAC-KBP

During development, we measured the recall of the entity candidate generation step on the TAC-KBP corpus. The right-hand side of Table 2 shows the detailed statistics. In a nutshell, fewer than 7% of all entities in the corpus are unambiguous with respect to DBpedia. On average, the candidate selection stage produces over 49 URI candidates per entity. Moreover, the ambiguity of the different entity types (PER, LOC, ORG, UNK) varies significantly. Overall, recall is 90%,

i.e. for 90% of the entities the algorithm produced candidates, which contained the correct URI from the TAC-KBP corpus.

Next, the four disambiguation metrics were run independently as baselines, and their performance was measured using precision, recall, and F1-score (the harmonic mean of precision and recall) [MW08]. Since DBpedia can have more than one URI for the same entity, a list of wikiRedirect URIs was obtained. If both the predicted URI and the gold standard URI appear in this list, then the predicted URI is counted as a correct annotation.

Table 3 shows the breakdown of results per entity type and overall. Similar to the findings of [MW08], the commonness baseline achieves high scores, ranging between 0.53 and 0.83, depending on entity type.

Since these four metrics measure different kinds of similarities (string, contextual, semantic, and popularity), we also evaluated the weighted combination of these four scores (referred to as LODIE). The optimal weights were derived through 5-fold cross validation and are as follows: 0.4 (string similarity), 0.35 (contextual similarity), 0.25 (semantic similarity), and 0.00 (commonness similarity). In cases where two or more candidate URIs have an equal overall score, the URI with the highest commonness score is returned. These weights are used unchanged in the evaluation on the unseen EnvTest corpus.

One could perhaps argue that having a weight of 0 for the commonness score is somewhat counter-intuitive, since it is the highest scoring baseline. What this shows is that the best URI disambiguation results are achieved when string, contextual, and semantic similarity are combined together. Only where they fail to produce a conclusive result, should the most common candidate URI be used as a tie-breaker.

LODIE outperforms the best commonness baseline with 6% improvement in overall F_1 score. Table 3 shows a breakdown by entity type, precision, and recall. While overall precision is lower than Zemanta’s (0.71 vs 0.90), overall recall (0.74) is 6% higher.

The next question we investigated is whether Zemanta and LODIE make the same mistakes. In the first instance, we examined the intersection of the results (shown as $Zemanta \cap LODIE$ in Table 3), i.e. only URIs predicted by both systems are retained. The results have very high precision on all four entity types (between 0.95 and 1.00), which overall is 7% higher than Zemanta’s precision on its own (between 0.82 and 0.96). However, this comes at the expense of much lower recall (between 0.42 and 0.74), which overall is 14% lower than Zemanta’s and 20% lower than LODIE’s. This also demonstrates that there is complementarity between the two systems, in terms of entities they find and classify correctly.

Consequently, our last experiment was to investigate a union, where Zemanta’s high-precision results are augmented with any additional entities suggested by LODIE. Since LODIE’s precision is lower than Zemanta’s, this predictably led to lower overall precision ((see the $Zemanta \cup LODIE$ row in Table 3). However, thanks to the complementarity between the two systems and the higher recall of LODIE, the combined results have the highest recall (0.81) amongst all tested methods. Therefore, $Zemanta \cup LODIE$ is the method which achieves the highest overall performance (0.82) on our development set.

Lastly, it must be noted that these precision and recall figures are for the task of detecting mentions of DBpedia entities in the TAC-KBP documents and then disambiguating these to the correct DBpedia URIs. This semantic enrichment task is different from the knowledge base population task for which the TAC-KBP corpus was originally created [JG11]. Therefore, the

results reported here are not directly comparable to those for knowledge base population on the same dataset.

3.3 Quantitative Evaluation Results on EnvTest

Based on the results and similarity metric weights obtained on TAC-KBP, we evaluated the four best methods on EnvTest: Zemanta, LODIE, the high-precision intersection of the two, and the high-recall union of the two.

As shown in Table 4, again the best precision is achieved by the intersection method (98%) – similar to the precision on the TAC-KBP development data. The drop in recall however is significant (down to 0.37% from 0.54%). Similar drops in recall are observed for the other methods too. Our error analysis showed that all methods missed systematically abbreviations (e.g. EU, DTI, NE England), which were very common in EnvTest.

Again, the best results in terms of recall (0.73) and overall F_1 score (0.69) are obtained by taking the union of the entities found by Zemanta and LODIE. The increase in recall is again due to the complementarity between the two systems. Zemanta is missing many locations (recall only 0.45 vs LODIE’s 0.68), while being slightly better at recognising organisation names (0.53 recall vs LODIE’s 0.79).

We investigated in more detail the reasons behind the lower recall on the EnvTest corpus. Firstly, this is due to the differences in genre between EnvTest (long scientific articles and short metadata records) and the more news-oriented TAC-KBP. The short metadata records in EnvTest are particularly challenging for semantic enrichment, due to the limited context that they provide. Secondly, in EnvTest publisher names appear in scientific references, which are harder to annotate correctly. Thirdly, the target DBpedia entities are themselves different, since EnvTest is mostly UK and European content, whereas TAC-KBP mentions many US-related entities and content. Lastly, there are differences between the two corpora in the way entities are referred to linguistically. EnvTest contains longer location and organisation names (see Table 2) as well as more abbreviations, which resulted in boundary detection errors and misses. To take one example, “Forestry Commission Scotland” is mistakenly annotated as two entities – one organisation and one location.

Next we discuss how, from a user perspective, LOD-based semantic enrichment helps with information discovery.

4 Discussion

Environmental science researchers from within The British Library and HR Wallingford carried out information discovery searches on a small subset of metadata and full-text documents from Envia (1000 metadata files and 150 full-text papers), which were enriched semantically with our best performing system. The search queries used for this small-scale experiment came directly from our survey of target Envia users.

The purpose of this user assessment was to gain insight into how semantic enrichment and LOD-based semantic search can improve information discovery. In particular, we examined:

1. how semantic enrichment helps enhance the metadata in Envia, by populating automatically the Dublin Core Subject field with selected annotations, and
2. how the more complex search queries from our survey can be answered by combining full-text search with LOD-based semantic queries.

4.1 Impact of Semantic Enrichment on Article Metadata

The automatically added LOD-based semantic annotations were manually checked in each of the documents, to assess their accuracy and relevance to the types of searches requested by the environmental science researchers in our survey. The focus was on enhancing the Envia metadata by populating the Dublin Core Subject field.

The benefit of the semantic enrichment in this case, is that by surfacing annotated terms derived from the full-text content, concepts buried within the body of the paper/report can be highlighted. The addition of terms affects the relevance ranking in full-text searches. Moreover, searches can be made more specific by limiting the search criteria to the Subject field (e.g. through faceted search). This is similar in principle to the use of Medical Subject Headings (MeSH) within the Medline and PubMed databases, where the content of the original document is described through the use of key terms added to the bibliographic record.

For each semantically annotated full-text document, the metadata enrichment algorithm retained the top five locations and organisations with DBpedia entity URIs and the associated location-related knowledge from GeoNames. Additional terms were recognised separately, on the basis of environmental science ontologies (outside the scope of this paper). This automatically acquired metadata was incorporated into the Subject fields of the document (see the highlighted terms at the bottom of Figure 1).

Once the enrichment process was complete, the enhanced metadata was loaded and indexed in a separate full-text search repository. Differences in retrieval were measured by comparing the results across the annotated and the non-annotated versions of the corpus using structured search queries. Examples of the ontology-derived domain-specific terms that populated the Subject field of one particular record were ‘Environment Agency’, ‘Scottish Government’, and ‘Scotland’. In this, and a number of other cases, these automatically generated terms provided additional contextual information to the user, particularly useful in those instances where the original metadata is sparse and there is no abstract present.

4.2 Impact of Semantic Search on User Query Results

In addition to populating the metadata Subject fields, the semantic annotations derived from both the metadata and article full-text were indexed into a GATE Mimir (Multi-paradigm Information Management Index and Repository) semantic repository [CTR⁺11]. GATE Mimir is a multi-paradigm information management index and repository which can be used to index and search over text, annotations, semantic schemas (ontologies), and Linked Open Data endpoints/repositories. It supports queries that arbitrarily mix full-text, structural, linguistic and semantic constraints and scales up to terabytes of text through federated indexing.

The rationale for choosing GATE Mimir is its transparent support for semantic search constraints, expressed as SPARQL queries against Linked Open Data. In particular, our aim was

to determine whether the more complex search needs of environmental researchers could be met better through semantic queries that make use of the additional knowledge from DBpedia and GeoNames.

4.2.1 Removing False Positives through Semantic Restrictions

The first benefit observed by the users, was that the semantic annotations were making the search results more precise, i.e. removed false positives.

In particular, a frequent literature search query involves environmental science terms (e.g. flooding) coupled with a geo-location (e.g. Oxford). Such queries in Envia often return false positives, due to location names being ambiguous. For instance, the query “flooding Oxford” returns 8 documents, 4 of which mention other locations (e.g. Oxford Road Mill – an industrial site) and organisations (e.g. University of Oxford, Oxford University Press).

The corresponding semantic search query can be made much more precise, by specifying explicitly in the query that Oxford is a location or even a city. In our example, the corresponding Mimir query is `flooding AND ({Sem_Location} OVER ‘Oxford’)`, which filters out the 4 false positives, where Oxford is part of an organisation name.

Other similar queries we tested, included “flood management Northern Ireland” (spurious hits returned due to documents published by the Northern Ireland Rivers Agency) and queries involving city names, where there are documents published by eponymous city councils (e.g. Gloucester vs Gloucester City Council).

4.2.2 Improving Search Results through LOD Knowledge

The second observed benefit was improved search coverage, through knowledge added from GeoNames and DBpedia during the semantic enrichment phase. In particular, location-based search queries in Envia are frequently phrased at the county (e.g. Oxfordshire) or regional (e.g. South East England) level, whereas the full-text documents mention much more specific place names. Therefore, such full-text searches in Envia tend to have poor recall, due to the lack of explicit mentions of the county and the region. For example, the query “climate change Oxfordshire” returns no results.

In contrast, the corresponding Mimir query (for documents mentioning “climate change” and a location within Oxfordshire) returns two relevant documents about soil CO₂ efflux and Wytham Woods (see Figure 3).

4.2.3 Adding Semantic Search Constraints

The third major benefit of semantic enrichment and its combination with semantic search constraints, is in giving users answers to queries involving implicit knowledge. Examples from our user survey included: “flooding in the last 10 years”, “flooding since 2007”, “flood defence spending in non UK countries”, “where are the main rivers”, and “where is the floodplain near Aylesbury”.

The first two kinds of queries are answered based on the automatically recognised and normalised dates in the full-text content. Mimir does offer some support for negation, so queries like the one above are possible, even though very complex to write. Semantic constraints based

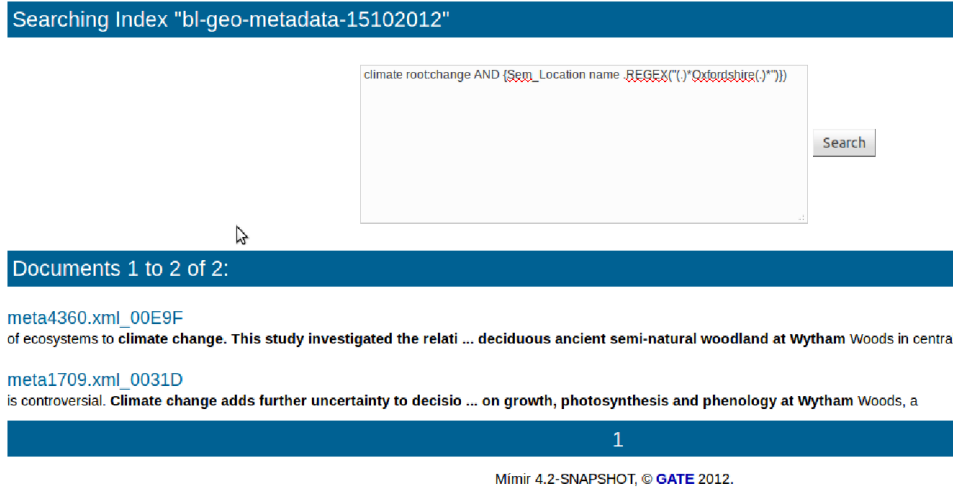


Figure 3: Semantic search example for documents related to climate change and Oxfordshire.

on knowledge in DBpedia and GeoNames are essential for the last two types of queries, which are more about facts, rather than documents. From within Mimir these need to be formulated as SPARQL queries, fired against the respective LOD endpoints, and the matching DBpedia and GeoName instances returned in response.

Let us take as an example a user query about documents on flooding in countries with population density greater than 500 people per square kilometre. Since none of the Envia documents contain any information on population density, this type of query cannot be answered through a standard full-text search engine. The corresponding MIMIR query is:

```
root:flood AND {Sem_Location sparql="select distinct ?inst
    where {?inst rdf:type :Country.
        ?inst :populationDensity ?popDensity.
        FILTER(?popDensity > 500)}"}
```

In this case documents containing the stemmed word ‘flood’ (‘flood’, ‘flooding’, ‘flooded’, etc.) are retrieved along with any words in the document that have been annotated as a location, by the semantic disambiguation algorithm. An additional constraint on these matching DBpedia location URIs is that they need to be of type `Country` and the value of the `populationDensity` property needs to be more than 500.

Our last example query is the full-text search query ‘river flooding’. In Mimir this can be formulated as a query for documents containing the stemmed word ‘flood’ and a location, which is of DBpedia class ‘River’. This query indeed retrieved metadata records relating to the Thames that were not found using the full-text search terms in Envia. In this case the semantic enrichment algorithm had tagged the Thames with the corresponding DBpedia URI and the SPARQL semantic constraint checked against DBpedia that indeed it is of type river.

```
root:flood AND {Sem_Location sparql="select distinct ?inst
```

```
where {?inst rdf:type :River}"}
```

4.3 Future Work

The focus of this deliverable was on studying the benefits of using Linked Open Data (LOD) vocabularies, automatic semantic enrichment methods, and semantic search, to improve information discovery on environmental science literature. Firstly, we described an approach for LOD-based semantic enrichment of metadata and full-text documents. Secondly, the results were evaluated, both quantitatively and with users. For the latter, we compared semantic search queries against the full-text search capabilities of the Envia British Library information discovery tool.

Specifically, we tested the usefulness of knowledge from DBpedia and GeoNames, to enhance information discovery and management of environmental science literature. The conclusion is that accessing knowledge from Linked Data allows for generalisations and, thus, answering more complex information needs, such as identification of documents that refer to water levels at the Thames barrier as relevant to a search for flooding in south-east England.

The ontologies selected for this experiment in semantic enrichment and search were found to be sufficient for the purposes of the initial evaluation. Despite its generic nature DBpedia proved a useful LOD resource. GeoNames was used to enrich further the DBpedia entities annotated in document content with location-based knowledge.

The main focus of ongoing and future work is to develop and evaluate an intuitive user interface for semantic search, that can hide the complexities of the SPARQL semantic search constraints, while, at the same time, allowing environmental researchers to benefit from the more powerful semantic search capabilities.

A parallel line of research is on improving the precision and recall of the semantic enrichment algorithm, both in general and specifically on environmental science literature. The first priority is improving the candidate selection and disambiguation of abbreviations and longer organisation names. We will also be working on giving the methods some credit for minor classification mistakes (e.g. NE England being recognised as England, which is its parent administrative division). At present, such errors are considered completely incorrect, whereas we plan to adopt an ontology-based evaluation metric (e.g. [MPL08]), which would give some credit in cases where a super-class is chosen instead of the correct class.

Last but not least, we will undertake a larger user-based evaluation experiment, including participants from our original user survey, as well as newly recruited environmental science researchers. The Envia tool will be undergoing further development in the next three years, which, over time, will give us access to user query logs and allow us to identify and improve iteratively the quality of the semantic enrichment and search algorithms.

Acknowledgments

The EnviLOD project was funded by JISC, under the JISC Research Tools (#restools) programme, Strand B3.

We would like to thank the programme managers, Torsten Reimer and Christopher Brown, for their support and helpful advice.

We wish to thank Ontotext for providing invaluable help and advice on their scalable OWLIM semantic repository and its use with DBpedia and GeoNames. EnviLOD built on and extended the LODIE information extraction algorithm, developed originally in the TrendMiner EC-funded project (<http://www.trendminer-project.eu>).

References

- [AAB⁺08] M. Agatonovic, N. Aswani, K. Bontcheva, H. Cunningham, T. Heitz, Y. Li, I. Roberts, and V. Tablan. Large-scale, parallel automatic patent annotation. In *Proceedings of the 1st ACM workshop on Patent information retrieval (PaIR '08, 30 October 2008, PaIR '08*, pages 1–8, New York, NY, USA, October 2008. ACM.
- [ACJ⁺09] Eneko Agirre, Angel X. Chang, Daniel S. Jurafsky, Christopher D. Manning, Valentin I. Spitzkovsky, and Eric Yeh. Stanford-ubc at tac-kbp. In *Proceedings of the Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA, November 2009.
- [BLK⁺09] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia – a crystallization point for the web of data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 7:154–165, 2009.
- [CH08] K. Bretonnel Cohen and Lawrence Hunter. Getting started in text mining. *PLoS Comput Biol*, 4(1):e20, 01 2008.
- [CMB⁺11] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, and I. Roberts. *Text Processing with GATE (Version 6)*. 2011.
- [CSW09] T. Cohen, R. Schvaneveldt, and D. Widdows. Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, 2009.
- [CTR⁺11] Hamish Cunningham, Valentin Tablan, Ian Roberts, Mark A. Greenwood, and Nijaraj Aswani. Information Extraction and Semantic Annotation for Multi-Paradigm Information Management. In Mihai Lupu, Katja Mayer, John Tait, and Anthony J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, volume 29 of *The Information Retrieval Series*, pages 307–327. Springer Berlin Heidelberg, 2011.
- [DDF⁺90] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.
- [DK10] Anusuriya Devaraju and Werner Kuhn. A process-centric ontological approach for integrating geo-sensor data. In *Proceedings of the Sixth International Conference on Formal Ontology in Information Systems (FOIS)*, pages 199–212, 2010.

- [DSL12] D. Damjanovic, M. Stankovic, and P. Laublet. Linked Data-based Concept Recommendation: Comparison of Different Methods in Open Innovation Scenario. In *Proceedings of the 9th Extended Semantic Web Conference (ESWC)*, 2012.
- [GNP⁺09] D. Gruhl, M. Nagarajan, J. Pieper, C. Robson, and A. Sheth. Context and Domain Knowledge Enhanced Entity Spotting in Informal Text. In *Proceedings of the 8th International Semantic Web Conference (ISWC'2009)*, 2009.
- [HvMS⁺10] K. Hettne, E. van Mulligan, M. Schuemie, B. Schijvennaars, and J. Kors. Rewriting and suppressing UMLS terms for improved biomedical term identification. *Journal of Biomedical Semantics*, 1(5):1–14, 2010.
- [JBGG09] S. Jimenez, C. Becerra, A. Gelbukh, and F. Gonzalez. Generalized mongue-elkan method for approximate text string comparison. In *Proc. of CICLing*, 2009.
- [JG11] H. Ji and R. Grishman. Knowledge base population: Successful approaches and challenges. In *Proc. of ACL'2011*, pages 1148–1158, 2011.
- [JL84] W. B. Johnson and J. Lindenstrauss. Extensions to lipschitz mapping into hilbert space. *Contemporary Mathematics*, 26, 1984.
- [JSPHss] Krzysztof Janowicz, Simon Scheider, Todd Pehle, and Glen Hart. Geospatial semantics and linked spatiotemporal data past, present, and future. *Semantic Web Interoperability, Usability, Applicability*, In Press.
- [KS01] J. Karlgren and M. Sahlgren. From words to understanding. In Y. Uesaka, P. Kanerva, and H. Asoh, editors, *Foundations of Real-World Intelligence*, pages 294–308. 2001.
- [KSN11] J. Kieniewicz, A. Sudlow, and E. Newbold. Coordinating improved environmental information access and discovery: Innovations in sharing environmental observations and information. In W. Pillman, S. Schade, and P. Smits, editors, *Proceedings of the 25th International EnviroInfo Conference*, 2011.
- [MJGSB11] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8, 2011.
- [MPL08] D. Maynard, W. Peters, and Y. Li. Evaluating evaluation metrics for ontology-based applications: Infinite reflection. In *Proc. of 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008.
- [MW08] D. Milne and I. H. Witten. Learning to link with Wikipedia. In *Proc. of the 17th Conf. on Information and Knowledge Management (CIKM)*, pages 509–518, 2008.
- [PSS11] W. Pillman, S. Schade, and P. Smits. *Innovations in sharing environmental observations and information, Proceedings of the 25th EnviroInfo Conference*. Shaker-Verlag, 2011.

- [RMD13] D. Rao, P. McNamee, and M. Dredze. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, Multi-lingual Information Extraction and Summarization*. Springer, 2013.
- [RP05] Robert G. Raskin and Michael J. Pan. Knowledge representation in the semantic web for earth and environmental terminology (sweet). *Computers and Geosciences*, 31(9):1119–1125, 2005.
- [RT11] Giuseppe Rizzo and Raphaël Troncy. NERD: A framework for evaluating named entity recognition tools in the Web of data. In *ISWC 2011, 10th International Semantic Web Conference*, Bonn, Germany, 2011.
- [Sah05] M. Sahlgren. An introduction to random indexing. In *Proc. of the Methods and Applications of Semantic Indexing Workshop*, Copenhagen, Denmark, 2005.
- [SPMM11] H. Schentz, J. Peterseil, B. Magagna, and M. Mirtil. Semantics in ecosystems research and monitoring. In W. Pillman, S. Schade, and P. Smits, editors, *Proceedings of the 25th International EnviroInfo Conference*, 2011.
- [SWC⁺11] Arif Shaon, Andrew Woolf, Shirley Crompton, Robert Boczek, Will Rogers, and Mike Jackson. An open source linked data framework for publishing environmental data under the UK location strategy. In *Terra Cognita 2011: Foundations, Technologies and Applications of the Geospatial Web*, pages 62–74, 2011.
- [SWLW12] W. Shen, J. Wang, P. Luo, and M. Wang. LINDEN: Linking named entities with knowledge base via semantic knowledge. In *Proceedings of the 21st Conference on World Wide Web*, pages 449–458, 2012.

Table 3: **Entity disambiguation results on the TAC-KBP development corpus. The highest precision, recall and F_1 scores are marked in bold.**

		PER	LOC	ORG	UKN	Overall	
DBpedia	Precision	0.97	0.82	0.86	0.82	0.85	
	Spotlight	Recall	0.40	0.46	0.31	0.34	0.39
	F_1 -score	0.57	0.59	0.46	0.48	0.54	
Zemanta	Precision	0.96	0.89	0.82	0.90	0.90	
	Recall	0.84	0.62	0.57	0.76	0.68	
	F_1 -score	0.90	0.73	0.68	0.82	0.78	
Commonness	Precision	0.833	0.636	0.535	0.759	0.678	
	Baseline	Recall	0.833	0.631	0.531	0.749	0.668
	F_1 -score	0.833	0.634	0.533	0.754	0.673	
String Similarity	Precision	0.578	0.322	0.368	0.620	0.448	
	Baseline	Recall	0.578	0.319	0.365	0.612	0.441
	F_1 -score	0.578	0.321	0.367	0.616	0.445	
Semantic	Precision	0.578	0.322	0.368	0.620	0.448	
	Similarity	Recall	0.578	0.319	0.365	0.612	0.441
	Baseline	F_1 -score	0.578	0.321	0.367	0.616	0.445
Contextual	Precision	0.578	0.314	0.368	0.617	0.443	
	Similarity	Recall	0.578	0.311	0.365	0.608	0.437
	Baseline	F_1 -score	0.578	0.313	0.367	0.612	0.440
LODIE:	Precision	0.81	0.73	0.56	0.75	0.71	
	Weighed	Recall	0.82	0.76	0.59	0.77	0.74
	Combination	F_1 -score	0.82	0.75	0.58	0.76	0.73
Zemanta \cap	Precision	1.00	0.95	0.97	0.97	0.97	
	LODIE	Recall	0.74	0.45	0.42	0.66	0.54
	F_1 -score	0.85	0.61	0.58	0.79	0.69	
Zemanta \cup	Precision	0.94	0.77	0.72	0.82	0.82	
	LODIE	Recall	0.93	0.76	0.71	0.81	0.81
	F_1 -score	0.94	0.77	0.71	0.82	0.82	

Table 4: Evaluation results on the EnvTest corpus

	LOC			ORG			TOTAL		
	P	R	F_1	P	R	F_1	P	R	F_1
Commonness Baseline	0.64	0.67	0.66	0.61	0.78	0.69	0.63	0.71	0.67
String Similarity Baseline	0.64	0.67	0.66	0.61	0.77	0.68	0.63	0.70	0.67
Semantic Similarity Baseline	0.64	0.67	0.65	0.61	0.78	0.69	0.63	0.70	0.67
Contextual Similarity Baseline	0.46	0.48	0.47	0.49	0.62	0.55	0.48	0.53	0.50
LODIE: Weighted Combination	0.68	0.68	0.68	0.69	0.79	0.74	0.68	0.72	0.70
Zemanta	0.79	0.45	0.57	0.84	0.53	0.65	0.81	0.46	0.59
Zemanta \cap LODIE	0.97	0.33	0.49	0.99	0.46	0.63	0.98	0.37	0.54
Zemanta \cup LODIE	0.68	0.72	0.70	0.68	0.77	0.72	0.65	0.73	0.69