

Terra Cognita 2011 Workshop

Foundations, Technologies and Applications of the Geospatial Web

October 23, 2011

Bonn, Germany

Organizers

Rolf Grütter, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

Dave Kolas, Raytheon BBN Technologies, U.S.A.

Manolis Koubarakis, National and Kapodistrian University of Athens, Greece

Dieter Pfoser, Institute for the Management of Information Systems (IMIS),
Athens, Greece

In Conjunction with the 10th International Semantic Web Conference (ISWC 2011)

Introduction

The wide availability of technologies such as GPS, map services and social networks, has resulted in the proliferation of geospatial data on the Web. In addition to material produced by professionals (e.g., maps), the public has also been encouraged to make geospatial content, including their geographical location, available online. The volume of such user-generated geospatial content is constantly growing. Similarly, the Linked Open Data cloud includes an increasing number of data sources with geospatial properties.

The geo-referencing of Web resources and users has given rise to various services and applications that exploit it. With the location of users being made available widely, new issues such as those pertaining to security and privacy arise. Likewise, emergency response, context sensitive user applications, and complex GIS tasks all lend themselves toward Geospatial Semantic Web solutions.

Researchers have been quick to realize the importance of these developments and have started working on the relevant research problems, giving rise to new topical research areas such as "Geographic Information Retrieval", "Geospatial (Semantic) Web", "Linked Geospatial Data", "GeoWeb 2.0". Similarly, standardization bodies such as the Open Geospatial Consortium (OGC) have been developing relevant standards such as the Geography Markup Language (GML) and GeoSPARQL.

The goal of this workshop is to bring together researchers and practitioners from various disciplines, as well as interested parties from industry and government, to advance the frontiers of this exciting research area. Bringing together Semantic Web and geospatial researchers helps encourage the use of semantics in geospatial applications and the use of spatial elements in semantic research and applications thereby advancing the Geospatial Web.

Topics Of Interest

Original, high-quality work related (but not limited) to one of the following research topics is welcome. Submissions must not be published nor must they be submitted for publication elsewhere.

- Data models and languages for the Geospatial Web
- Systems and architectures for the Geospatial Web
- Linked geospatial data
- Ontologies and rules in the Geospatial Web
- Uncertainty in the Geospatial Web
- User interface technologies for the Geospatial Web
- Geospatial Web and mobile data management
- Security and privacy issues in the Geospatial Web
- Geospatial Web applications
- User-generated geospatial content
- OGC and W3C technologies and standards in the Geospatial Web

Organizing Committee

This workshop is organized by members of the Spatial Ontology Community of Practice (SOCoP), and European projects TELEIOS and Geocrowd.

SOCoP (<http://www.socop.org/>) is a geospatial semantics interest group currently mainly with members from U.S. federal agencies, academia, and business. SOCoP's goal is to foster collaboration among users, technologists, and researchers of spatial knowledge representations and reasoning towards the development of a set of core, common geospatial ontologies for use by all in the Semantic Web.

TELEIOS (<http://www.earthobservatory.eu/>) is an FP7/ICT project with the goal of building an Earth Observatory. TELEIOS concentrates heavily on geospatial data (sattelite images, traditional GIS data, geospatial Web data).

GEOCROWD - Creating a Geospatial Knowledge World (<http://www.geocrowd.eu>) is an Initial Training Network (ITN) project with the goal to promote the GeoWeb 2.0 vision and to advance the state of the art in collecting, storing, processing, and making large amounts of semantically rich user-generated geospatial information available on the Web.

Organizers

Rolf Grütter, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland,
rolf.gruetter@wsl.ch

Dave Kolas, Raytheon BBN Technologies, U.S.A., dkolas@bbn.com

Manolis Koubarakis, National and Kapodistrian University of Athens, Greece,
koubarak@di.uoa.gr

Dieter Pfoser, Institute for the Management of Information Systems (IMIS), Athena, Greece,
Athena, pfoser@imis.athena-innovation.gr

Program Committee

Oscar Corcho, Universidad Politecnica de Madrid, Spain
Stavros Vassos, National and Kapodistrian University of Athens, Greece
Jans Aasman, Franz Inc.
Alia Abdelmoty, Cardiff University, UK
Ola Ahlqvist, Ohio State University, USA
Gustavo Alonso, ETH Zurich, Switzerland
Thomas Barkowsky, University Bremen, Germany
Abraham Bernstein, University of Zurich, Switzerland
Isabel Cruz, University of Illinois at Chicago, USA
Mihai Datcu, German Aerospace Center (DLR), Germany
Mike Dean, BBN Technologies, USA
Stewart Fotheringham, National University of Ireland at Maynooth, Ireland
Christian Freksa, University of Bremen, Germany
Alasdair J G Gray, University of Manchester, UK
John Goodwin, Ordnance Survey, UK
Glen Hart, Ordnance Survey, UK
Martin Kersten, CWI, The Netherlands
Werner Kuhn, University of Muenster, Germany
Sergei Levashkin, National Polytechnic Institute, Mexico
Joshua Lieberman, Traverse Technologies, USA
Michael Lutz, European Commission-DG Joint Research Center, Italy
Stefan Manegold, CWI, The Netherlands
Ralf Möller, Hamburg University of Technology, Germany
Alexandros Ntoulas, Microsoft Research
Mathew Perry, Oracle
Euripides Petrakis, Technical University of Crete, Greece
Florian Probst, SAP Research, Germany
Thorsten Reitz, Fraunhofer Institute for Computer Graphics, Germany
Timos Sellis, IMIS, Research Center Athena, Greece
Spiros Skiadopoulos, University of the Peloponnese, Greece
Fabian Suchanek, Max Planck Institute for Informatics, Germany
Agnes Voisard, Free University Berlin and Fraunhofer ISST, Germany
Nancy Wiegand, University of Wisconsin, USA
James Wilson, James Madison University, USA
Stefan Woelfl, University of Freiburg, Germany

Program

9:00 – 10:30

User Generated Geo-Content

Chris De Rouck, Olivier Van Laere, Steven Schockaert and Bart Dhoedt
Georeferencing Wikipedia pages using language models from Flickr

Jens Ortmann, Minu Limbu, Dong Wang and Tomi Kauppinen
Crowdsourcing Linked Open Data for Disaster Management

Efthymios Drymonas, Alexandros Efentakis and Dieter Pfoser
Opinion Mapping Travelblogs

10:30 – 11:00

Coffee Break

11:00 – 12:30

Linked Geospatial Data

Rahul Parundekar, José Luis Ambite and Craig Knoblock
Aligning Unions of Concepts in Ontologies of Geospatial Linked Data

Eetu Mäkelä, Aleksi Lindblad, Jari Väättäinen, Rami Alatalo, Osma Suominen and Eero Hyvönen
Discovering Places of Interest through Direct and Indirect Associations in Heterogeneous Sources - The TravelSampo System

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

12:30 – 14:00

Lunch Break

14:00 – 16:00

Semantics and Geospatial Web

Sven Tschirner, Ansgar Scherp and Steffen Staab

Semantic access to INSPIRE - How to publish and query advanced GML data

Jan Oliver Wallgrün and Mehul Bhatt

An Architecture for Semantic Analysis in Geospatial Dynamics

Iris Helming, Abraham Bernstein, Rolf Grütter and Stephan Vock

Making close to suitable for web searches - a comparison of two approaches

Juan Martín Salas and Andreas Harth

Finding spatial equivalences across multiple RDF datasets

16:00 – 16:30

Coffee Break

16:30 – 18:00

Geographic Information Retrieval and Data Mining

Lars Döhling and Ulf Leser

EquatorNLP: Pattern-based Information Extraction for Disaster Response

Rui Candeias and Bruno Martins

Associating Relevant Photos to Georeferenced Textual Documents through Rank Aggregation

Closing Session

All participants: Summary and Discussion

Table of Contents

Chris De Rouck, Olivier Van Laere, Steven Schockaert and Bart Dhoedt <i>Georeferencing Wikipedia pages using language models from Flickr</i>	3
Jens Ortmann, Minu Limbu, Dong Wang and Tomi Kauppinen <i>Crowdsourcing Linked Open Data for Disaster Management</i>	11
Efthymios Drymonas, Alexandros Efentakis and Dieter Pfoser <i>Opinion Mapping Travelblogs</i>	23
Rahul Parundekar, José Luis Ambite and Craig Knoblock <i>Aligning Unions of Concepts in Ontologies of Geospatial Linked Data</i>	37
Eetu Mäkelä, Aleks Lindblad, Jari Väättäinen, Rami Alatalo, Osma Suominen and Eero Hyvönen <i>Discovering Places of Interest through Direct and Indirect Associations in Heterogeneous Sources - The TravelSampo System</i>	50
Arif Shaon, Andrew Woolf, Shirley Crompton, Robert Boczek, Will Rogers and Mike Jackson <i>An Open Source Linked Data Framework for Publishing Environmental Data under the UK Location Strategy</i>	62
Sven Tschirner, Ansgar Scherp and Steffen Staab <i>Semantic access to INSPIRE - How to publish and query advanced GML data</i>	75
Jan Oliver Wallgrün and Mehul Bhatt <i>An Architecture for Semantic Analysis in Geospatial Dynamics</i>	88
Iris Helming, Abraham Bernstein, Rolf Grütter and Stephan Vock <i>Making close to suitable for web searches - a comparison of two approaches</i>	101
Juan Martín Salas and Andreas Harth <i>Finding spatial equivalences across multiple RDF datasets</i>	114
Lars Döhling and Ulf Leser <i>EquatorNLP: Pattern-based Information Extraction for Disaster Response</i>	127
Rui Candeias and Bruno Martins <i>Associating Relevant Photos to Georeferenced Textual Documents through Rank Aggregation</i>	139

Georeferencing Wikipedia pages using language models from Flickr

Chris De Rouck¹, Olivier Van Laere¹, Steven Schockaert², and Bart Dhoedt¹

¹ Department of Information Technology, IBBT, Ghent University, Belgium,
{chris.derouck,olivier.vanlaere,bart.dhoedt}@ugent.be

² Department of Applied Mathematics and Computer Science, Ghent University,
Belgium, steven.schockaert@ugent.be

Abstract. The task of assigning geographic coordinates to web resources has recently gained in popularity. In particular, several recent initiatives have focused on the use of language models for georeferencing Flickr photos, with promising results. Such techniques, however, require the availability of large numbers of spatially grounded training data. They are therefore not directly applicable for georeferencing other types of resources, such as Wikipedia pages. As an alternative, in this paper we explore the idea of using language models that are trained on Flickr photos for finding the coordinates of Wikipedia pages. Our experimental results show that the resulting method is able to outperform popular methods that are based on gazetteer look-up.

1 Introduction

The geographic scope of a web resource plays an increasingly important role for assessing its relevance in a given context, as can be witnessed by the popularity of location-based services on mobile devices. When uploading a photo to Flickr, for instance, users can explicitly add geographical coordinates to indicate where it has been taken. Similarly, when posting messages on Twitter, information may be added about the user’s location at that time. Nonetheless, such coordinates are currently only available for a minority of all relevant web resources, and techniques are being studied to estimate geographic location in an automated way.

For example, several authors have applied language modeling techniques to find out where a photo was taken, by only looking at the tags that its owner has provided [9, 10]. The main idea is to train language models for different areas of the world, using the collection of already georeferenced Flickr photos, and to subsequently use these language models for determining in which area a given photo was most likely taken. In this way, implicit geographic information is automatically derived from Flickr tags, which is potentially much richer than the information that is found in traditional gazetteers. Indeed, the latter usually do not contain information about vernacular place names, lesser-known landmarks, or non-toponym words with a spatial dimension (e.g. names of events), among others. For the task of assigning coordinates to Flickr photos, this intuition

seems to be confirmed, as language modeling approaches have been found to outperform gazetteer based methods [5].

For other types of web resources, spatially grounded training data may not be (sufficiently) available to derive meaningful language models, in which case it seems that gazetteers would again be needed. However, as language models trained on Flickr data have already proven useful for georeferencing photos, we may wonder whether they could be useful for finding the coordinates of other web resources. In this paper, we test this hypothesis by considering the task of assigning geographical coordinates to Wikipedia pages, and show that language models from Flickr are indeed capable of outperforming popular methods for georeferencing web pages. The interest of our work is twofold. From a practical point of view, the proposed method paves the way for improving location-based services in which Wikipedia plays a central role. Second, our results add further support to the view that georeferenced Flickr photos can provide a valuable source of geographical information as such, which relates to a recent trend where traditional geographic data is more and more replaced or extended by user contributed data from Web 2.0 initiatives [2].

The paper is structured as follows. Section 2 briefly reviews the idea of georeferencing tagged resources using language models from Flickr. In Section 3 we then discuss how a similar idea could be applied to Wikipedia pages. Section 4 contains our experimental results, after which we discuss related work and conclude.

2 Language models from Flickr

In this section, we recall how georeferenced Flickr photos can be used to train language models, and how these language models subsequently allow to find the area that most likely covers the geographical scope of some resource. Throughout this section, we will assume that resources are described as sets of tags, while the next section will discuss how the problem of georeferencing Wikipedia pages can be cast into this setting.

As training data, we used a collection of around 8.5 million publicly available photos on Flickr with known coordinates. In addition to these coordinates, the associated metadata contains tags attributed to each photo, providing us with a textual description of their content, as well as an indication of the accuracy of the coordinates as a number between 1 (world-level) and 16 (street level). As in [10], we only retrieved photos with a recorded accuracy of at least 12 and we removed photos that did not contain any tags or whose coordinates were invalid. Also, following [9] photos from bulk uploads were removed. The resulting dataset contained slightly over 3.25 million photos. In a subsequent step, the training data was clustered into disjoint areas using the k -medoids algorithm with geodesic distance. Considering a varying number of clusters k , this resulted in different sets of areas \mathcal{A}_k . For each clustering, a vocabulary V_k was compiled, using χ^2 feature selection, as the union of the m most important tags (i.e. the tags with the highest χ^2 value) for each area.

The problem of georeferencing a resource x , in this setting, consists of selecting the area a from the set of areas \mathcal{A}_k (for a specific clustering k) that is most likely to cover the geographic scope of the resource (e.g. the location of where the photo was taken, when georeferencing photos). Using a standard language modeling approach, this probability can be estimated as

$$P(a|x) \propto P(a) \cdot \prod_{t \in x} P(t|a) \quad (1)$$

where we identify the resource x with its set of tags. The prior probability $P(a)$ of area a can be estimated as the percentage of photos in the training data that belong to that area (i.e. a maximum likelihood estimation). To obtain a reliable estimate of $P(t|a)$ some form of smoothing is needed, to avoid a zero probability when encountering a tag t that does not occur with any of the photos in area a . In this paper, we use Jelinek-Mercer smoothing ($\lambda \in [0, 1]$):

$$P(t|a) = \lambda \cdot \frac{O_{ta}}{\sum_{a' \in \mathcal{A}_k} O_{ta'}} + (1 - \lambda) \cdot \frac{\sum_{a' \in \mathcal{A}_k} O_{ta'}}{\sum_{a' \in \mathcal{A}_k} O_{ta'} \sum_{t' \in V_k} O_{t'a'}}$$

O_{ta} is the number of occurrences of tag t in area a while V_k is the vocabulary, after feature selection. In the experiments, we used $\lambda = 0.7$ although we obtained good results for a wide range of values. The area that is most likely to contain resource x can then be found by maximizing the right-hand side of (1). To convert this area into a precise location, the area a can be represented as its medoid $med(a)$:

$$med(a) = \arg \min_{x \in a} \sum_{y \in a} d(x, y) \quad (2)$$

with $d(x, y)$ being the geodesic distance. Another alternative, which was proposed in [10] but which we do not consider in this paper, is to assign the location of the most similar photo from the training data which is known to be located in a .

3 Wikipedia pages

The idea of geographic scope can be interpreted in different ways for Wikipedia pages. A page about a person, for instance, might geographically be related to the places where this person has lived throughout his life, but perhaps also to those parts of the world which this person's work has influences (e.g. locations of buildings that were designed by some architect). In this paper, however, we exclusively deal with finding the coordinates of a Wikipedia page about a specific place, such as a landmark or a city. It is then natural to assume that the geographic scope of the page corresponds to a point.

While several Wikipedia pages already have geographic coordinates, it does not seem feasible to train area-specific language models from Wikipedia pages

with a known location, as we did in Section 2 for Flickr photos. The reason is that typically there is only one Wikipedia page about a given location, so either its location is already known or its location cannot be found by using other georeferenced pages. Moreover, due to the smaller number of georeferenced pages (compared to the millions of Flickr photos) and the large number of spatially irrelevant terms on a typical Wikipedia page, the process further complicates. One possibility to cope with these issues might be to explicitly look for toponyms in pages, and link these to gazetteer information. However, as we already have rich language models from Flickr, in this paper we pursue a different strategy, and investigate the possibility of using these models to find the locations of Wikipedia pages.

The first step consists of representing a Wikipedia page as a list of Flickr tags. This can be done by scanning the Wikipedia page and identifying occurrences of Flickr tags. As Flickr tags cannot contain spaces, however, it is important that concatenations of word sequences in Wikipedia pages are also considered. Moreover capitalization should be ignored. For example, an occurrence of “Eiffel tower” on a page is mapped to the Flickr tags “eiffeltower”, “eiffel” and “tower”.

Let us write $n(t, d)$ for the number of times tag t was thus found in the Wikipedia page d . We can then assign to d the area a which maximizes

$$P(a|d) \propto P(a) \cdot \prod_{t \in V_k} P(t|a)^{n(t,d)} \quad (3)$$

where V_k is defined as before and the probabilities $P(a)$ and $P(t|a)$ are estimated from our Flickr data, as explained in the previous section. Again (2) can be used to convert the area a to a precise location.

Some adaptations to this scheme are possible, where the scores $n(t, d)$ are defined in alternative ways. As Wikipedia pages often contain a lot of context information, which does not directly describe the location of the main subject, we propose two techniques for restricting which parts of an article are scanned. The first idea is to only look at tags that occur in section titles (identified using HTML tags of the form `<h1>`), in anchor text (`<a>`) or in emphasized regions (`` and ``). This variant is referred to as *keywords* below. The second idea is to only look at the abstract of the Wikipedia page, which is defined as the part of the page before the first section heading. As this abstract is supposed to summarize its content, it is less likely to contain references to places that are outside the geographical scope of the page. This second variant is referred to as *abstract*. Note that in both variants, the value of $n(t, d)$ will be lower than when using the basic approach.

4 Experimental results

In our evaluation, we used the Geographic Coordinates dataset of DBPedia 3.6 to determine an initial set of georeferenced Wikipedia pages. To ensure that all articles refer to a specific location, we only retained those pages that are

Table 1. Comparison of the Flickr language models for different numbers of clusters k and Yahoo! Placemaker (P.M.). We report how many of the Wikipedia pages are correctly georeferenced within a 1km radius, 5km radius, etc. Accuracy refers to the percentage of test pages for which the language models identified the correct cluster.

k	1 km	5 km	10 km	50 km	100 km	Acc
50	20	156	262	745	1470	76.52
500	334	1060	1385	2993	4195	69.17
2500	736	1636	2139	4020	4995	57.98
5000	892	1857	2377	4194	5075	51.62
7500	1008	1996	2557	4396	5239	49.85
10000	1052	2086	2670	4471	5233	47.80
12500	1103	2131	2697	4528	5263	45.73
15000	1129	2154	2743	4551	5212	44.45
17500	1159	2213	2783	4578	5243	43.71
P.M.	313	1583	2395	4257	5056	–

mentioned as a “spot” in the GeoNames gazetteer. This resulted in a set of 7537 georeferenced Wikipedia pages, whose coordinates we used as our gold standard.

Using the techniques outlined in the previous section, for each page the most likely area from \mathcal{A}_k is determined (for different values of k). To evaluate the performance of our method, we calculate the accuracy, defined as the percentage of the test pages that were classified in the correct area, i.e. the area actually containing the location of page d . In addition, we also look at how many of the Wikipedia pages are correctly georeferenced within a 1km radius, 5km radius, etc.

Our main interest is in comparing the methods proposed in Section 3 with the performance of Yahoo! Placemaker, a freely available popular webservice capable of geoparsing entire documents and webpages. Provided with free-form text, Placemaker identifies places mentioned in text, disambiguates those places and returns the corresponding locations. It is important to note that this approach uses external geographical knowledge such as gazetteers and other undocumented sources of information. In contrast, our approach uses only the knowledge derived from the tags of georeferenced Flickr photos.

In a first experiment, we compare the results of language models trained at different resolutions, i.e. different numbers of clusters k . Table 1 shows the results for k varying from 50 to 17 500, where we consider the basic variant in which the entire Wikipedia page is scanned for tag occurrences. There is a trade-off to be found, where finer-grained areas lead to more precise locations, provided that the correct area is found, while coarse-grained areas lead to a higher accuracy and to an increased likelihood that the found location is within a certain broad radius. In [10], it was found that the optimal number of clusters for georeferencing Flickr photos was somewhere between 2500 and 7500, with the optimum being higher for photos with more informative tags. In contrast, the results from Table 1 reveal that in the case of Wikipedia pages, its is beneficial

Table 2. Analysis of the effect of restricting the regions of a Wikipedia article that are scanned for tag occurrences (considering $k = 17500$ clusters). We report how many of the Wikipedia pages are correctly georeferenced within a 1km radius, 5km radius, etc.

k	1 km	5 km	10 km	50 km	100 km
article	1159	2213	2783	4578	5243
abstract	1194	2163	2707	4419	5051
keywords	1200	2361	3018	5052	5778

to further increase the number of clusters. This finding seems to be related to the intuition that Wikipedia pages contain more informative descriptions than Flickr photos. Comparing our results with Placemaker, we find a substantial improvement in all categories, which is most pronounced in the 1km range, where the number of correct locations for our language modeling approach is 3 to 4 times higher than for the Placemaker.

In a second experiment, we analyzed the effect of only looking at certain regions of a Wikipedia page, as discussed in Section 3. As the results in Table 2 show, when using the abstract, comparable results are obtained, which is interesting as this method only uses a small portion of the page. When only looking at the emphasized words (method *keywords*), the results are even considerably better. Especially for the 50km and 100km categories, the improvement is substantial. This seems to confirm the intuition that tag occurrences in section titles, anchor text and emphasized words are more likely to be spatially relevant.

5 Related work

Techniques to (automatically) determine the geographical scope of web resources commonly use resources such as gazetteers (geographical indexes), and tables with locations corresponding to IP addresses, zipcodes or telephone prefix codes. These resources are often handcrafted, which is time-consuming and expensive, although this results in accurate geographical information. Unfortunately, many of these sources are not freely available and their coverage varies highly from country to country. If sufficiently accurate resources are available, one of the main problems in georeferencing web pages is dealing with the high ambiguity of toponyms [6]. For example, when an occurrence of *Paris* is encountered, one first needs to disambiguate between a person and a place, and in the case it refers to a place, between different locations with that name (e.g. Paris, France and Paris, Texas).

It is only recently that alternative ways have been proposed to georeference resources. In [8], names of places are extracted from Flickr tags on a subset of around 50000 photos. Also, as studied by L. Hollenstein in [4], collaborative tagging-based systems are also useful to acquire information about the location of vernacular places names. In [1], methods based on Wordnet and Naive Bayes classification are compared for the automatic detection of toponyms within articles. To the best of our knowledge, however, approaches for georeferencing

Wikipedia pages, or webpages in general, without using a gazetteer or other forms of structured geographic knowledge have not yet been proposed in the literature.

An interesting line of work aims at automatically completing the infobox of a Wikipedia page by analyzing the content of that page [11]. This work is related to ours in the sense that semantic information about Wikipedia pages is made explicit. Such a strategy can be used to improve semantic knowledge bases, such as YAGO2 [3], which now contains over 10 million entities derived from Wikipedia, WordNet and GeoNames. Similarly, in [7], a gazetteer was constructed based on geotagged Wikipedia pages. In particular, relations between pages are extracted from available geographical information (e.g. *New York* is part of the *United States*). Increasing the number of georeferenced articles may thus lead to better informed gazetteers.

6 Conclusions

In this paper, we investigated the possibility of using language models trained on georeferenced Flickr photos for finding the coordinates of Wikipedia pages. Our experiments show that for Wikipedia pages about specific locations, the proposed approach can substantially outperform Yahoo! Placemaker, a popular approach for finding the geographic scope of a webpage. This is remarkable as the Placemaker crucially depends on gazetteers and other forms of structured geographic knowledge, and is moreover based on advanced techniques for dealing with issues such as ambiguity. Our method, on the other hand, only uses information that was obtained from freely available, user-contributed data, in the form of georeferenced Flickr photos, and uses standard language modeling techniques.

These results suggest that the implicit spatial information that arises from the tagging behavior of users may have a stronger role to play in the field of geographic information retrieval, which is currently still dominated by gazetteer-based approaches. Moreover, as the number of georeferenced Flickr photos is constantly increasing, the spatial models that could be derived are constantly improving. Further work is needed to compare the information contained implicitly in such language models with the explicit information contained in gazetteers.

References

1. D. Buscaldi and P. Rosso. A comparison of methods for the automatic identification of locations in wikipedia. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, pages 89–92, 2007.
2. M. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69:211–221, 2007.
3. J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, and G. Weikum. YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th International Conference on World Wide Web*, pages 229–232, 2011.

4. L. Hollenstein. Capturing vernacular geography from georeferenced tags. Master's thesis, University of Zurich, 2008.
5. M. Larson, M. Soleymani, and P. Serdyukov. Automatic tagging and geotagging in video collections and communities. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2011.
6. J. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, University of Edinburgh, 2007.
7. A. Popescu and G. Grefenstette. Spatiotemporal mapping of Wikipedia concepts. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, pages 129–138, 2010.
8. T. Rattenbury and M. Naaman. Methods for extracting place semantics from flickr tags. *ACM Transactions on the Web*, 3(1):1–30, 2009.
9. P. Serdyukov, V. Murdock, and R. van Zwol. Placing Flickr photos on a map. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 484–491, 2009.
10. O. Van Laere, S. Schockaert, and B. Dhoedt. Finding locations of Flickr resources using language models and similarity search. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2011.
11. F. Wu and D. S. Weld. Autonomously semantifying wikipedia. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, pages 41–50, 2007.

Crowdsourcing Linked Open Data for Disaster Management

Jens Ortmann¹, Minu Limbu^{*1}, Dong Wang¹, and Tomi Kauppinen¹

Institute for Geoinformatics, University of Muenster, Germany
(jens.ortmann, wang.dong, tomi.kauppinen)@uni-muenster.de
minulimbu@gmail.com

Abstract. This paper shows how Linked Open Data can ease the challenges of information triage in disaster response efforts. Recently, disaster management has seen a revolution in data collection. Local victims as well as people all over the world collect observations and make them available on the web. Yet, this crucial and timely information source comes unstructured. This hinders a processing and integration, and often a general consideration of this information. Linked Open Data is supported by number of freely available technologies, backed up by a large community in academia and it offers the opportunity to create flexible mash-up solutions. At hand of the Ushahidi Haiti platform, this paper suggests crowdsourced Linked Open Data. We take a look at the requirements, the tools that are there to meet these requirements, and suggest an architecture to enable non-experts to contribute Linked Open Data.

1 Introduction

The world has recently seen a number of environmental disasters, both natural and man-made. The most severe ones in the last two years were the earthquake that hit Haiti in January 2010 and the earthquake that hit Japan in March 2011. In both cases, information technologies contributed to increasing global awareness of these disasters. Modern communication channels and services have enabled people around the world to spread information, thereby changing the landscape of geographic information.

Crucial parts of disaster management are the acquisition, assessment, processing and distribution of information. In the mentioned disasters, so-called crowdsourced [9] information was massively generated. Crowdsourced information is information that is generated by a large heterogeneous crowd: People in the disaster-struck area share reports online; people all over the world help to

* Minu Limbu has worked more than four years with the United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA) in the field of Information Coordination/Management and was in Haiti for the International Federation for Red Cross and Red Crescent Movement (IFRC) Shelter Cluster immediately after the earthquake in January 2010.

process and distribute information. Crowdsourcing has proven to be an efficient approach to quickly generate huge amounts of near real-time data on a given situation [12,15]. We conducted a short survey targeting disaster management experts that revealed persisting problems of data processing and information integration. Relief organizations working in the disaster-struck area were unable to cope with unstructured and unconfirmed reports. They simply lack the time to integrate these data into existing information systems. This is not to say that crowdsourced data was not used at all. Yet, its potential is far bigger than the actual impact it made, especially in the early phase of a disaster response.

This paper describes an approach based on Linked Open Data [5] to alleviate the integration problems of crowdsourced data and to improve the exploitation of crowdsourced data in disaster management. We suggest to engage people in processing unstructured observations into structured RDF¹-triples according to Linked Open Data principles. Thereby, a substantial part of the integration problem is left to be solved by the crowd.

Hence, the goal of the efforts described in this paper is nothing less than to allow the use of linked open crowdsourced data for disaster management. This will increase the impact of crowdsourced data in disaster management and help humanitarian agencies to make informed decisions.

In Sect. 2 we shortly tell the story so far of crowdsourced data in disaster management and of Linked Open Data in disaster management. We then (in Sect. 3) walk through the envisioned architecture and discuss the requirements. Finally, the conclusion is drawn in Sect. 4.

2 The Story So Far

2.1 Crowdsourcing in Disaster Management

In this subsection we summarize the results of a small survey that we conducted with 14 experts in disaster management, as well as the results of Zook et al. [15].

Our survey comprised seven multiple-choice questions and one free text field for comments. Two questions assessed the participants' background, five questions targeted their experience with crowdsourcing services. The questionnaire was answered by experts from, including but not limited to, agencies like the United Nations, Red Cross, non-governmental organizations and donor communities. We asked the participants about their awareness of certain platforms that allow the crowd to contribute or process information. Fig. 1 shows the results for Twitter², Google Map Maker³ (GMM), Open Street Map⁴ (OSM) and Ushahidi⁵. The two mapping services GMM and OSM are known by about half the participants and also largely used when known. OSM and GMM provide

¹ Resource Description Framework (RDF), see <http://www.w3.org/RDF/>

² <http://www.twitter.com>

³ <http://www.google.com/mapmaker>

⁴ <http://www.openstreetmap.org/>

⁵ see for example the Ushahidi platform for Haiti: <http://haiti.ushahidi.com>

structured map-based information and an immediate visualization. In the case of the Haiti Earthquake these two platforms were a big success in terms of contributions and still a success in terms of use [15]. However, there exist compatibility problems between OSM and GMM and with other Geographic Information Systems (GIS) [15]. One of the authors witnessed first-hand how PhD students at the geography department at the University of Buffalo (USA) mapped damaged buildings in their desktop GIS, creating layers that were compatible with their research group’s disaster response efforts.

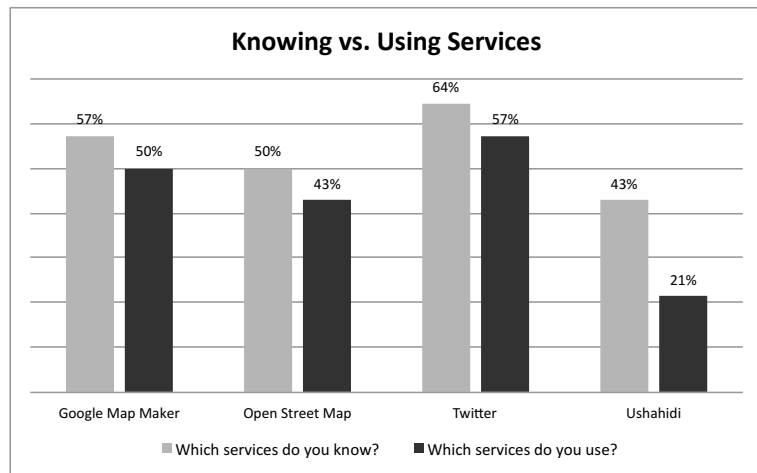


Fig. 1. Bar chart comparing knowledge of a service with its usage. All values are percentages of the 14 participating experts.

Twitter and Ushahidi provide natural language reports along with structural elements to capture formal aspects like time, location and tags or categories. Twitter is widely used to broadcast short messages. The survey responses surprisingly indicate that disaster managers use Twitter more than other platforms. However, one particular expert stated that Twitter is used mostly to broadcast information rather than capturing tweets from the crowd. Hence, the indicated usage does not necessarily reflect the efficiency of Twitter as information source.

Ushahidi is the most recent service. It is remarkable that nearly half the survey participants are aware of it. Unfortunately, Ushahidi shows the biggest drop from knowing to using the service. Out of six experts that knew Ushahidi only three answered that they actually use it. It is worth noting that one of the experts also commented that during early stages of a disaster, humanitarian workers are “too busy to think about such platforms during emergency response”. Another disaster expert noted that crowd information have heterogeneous and incompatible formats making it very difficult to integrate them into humanitarian agency specific information systems.

Morrow et al. [10] identified a general inconsistency between aggregated data in Ushahidi and requirements of relief organizations. Our survey revealed that many information managers see problems mostly in uncertainty (57% of the experts), trust (50%), and semantic⁶ problems (50%). On the upside, disaster managers acknowledged that crowdsourced information is collected near real-time and for free. Furthermore, experts see the greatest need for improvement in filtering of information (64%), training of volunteers (57%), ranking of relevant information (50%), structural compatibility (43%) and compliance with standard terminologies (36%).

Currently, two worlds of information infrastructures exist in parallel. On the one hand, relief organizations have their own information systems. On the other hand, ad-hoc information infrastructures emerge in the social web, which are mostly fed by crowdsourced data. Integration across these two worlds is only possible manually, and in the case of Haiti a full integration did not take place. Yet, despite the flaws of crowdsourced information, many disaster managers are willing to learn how crowdsourced information can be efficiently and effectively integrated into decision making processes.

2.2 Linked Open Data

In this section we outline the choices made to create Linked Open Data⁷ and give examples. Linked Open Data [2,5] is about using web techniques to share and interconnect pieces of data by following a few simple principles. Linked Open Data builds on Semantic Web technologies, wherein data is encoded in the form of <subject,predicate,object> RDF-triples. In the case of Haiti Data, real-time information about a crisis such as Twitter messages, news, articles and weather forecasts are identified, accessed and linked through URIs.

In disaster management, authorities like the relief organizations try to publish their data to people. Usually, these data come in non machine-understandable formats as PDF or CSV, or stored in different information systems. In some sense, these data are all Open Data, as they can be found and used somehow, but they are hard to access without knowing how the CSV-files should be interpreted. Since Linked Open Data is used as a framework of loose integration of data [7], it provides a method to make data really open by interlinking the data sources flexibly in the Semantic Web. In this work, we used the Ushahidi reports from Haiti as input and triplified them into RDF to make the information in the report accessible and linked.

It is considered that the World Wide Web has an abundance of data resources related to disaster management, either authoritative data possessed by a relief organization or crowdsourced data in the social web. The advantages of Linked Open Data as easy manipulation and loose integration make it a potential way to interlink the observed data from volunteers to existing systems.

⁶ we asked whether they see “Difficulties in interpreting the information”.

⁷ We do not distinguish between Linked Data and Linked Open Data here.

3 The Lifecycle of a Crowdsourced Emergency Report

In the lifecycle of a human observation we identified three different personas. There is Jean⁸, the local observer, who is immediately affected by the earthquake. Then, there is Desiree⁹, a web-user with some knowledge of Linked Open Data. Finally, there is Paul¹⁰, the information manager working for a relief organization.

Jean, the Local Observer. The people affected by the earthquake can be seen as a set of human sensors [8]. Equipped with mobile phones or access to the internet, this set of sensors turns into one big human sensor network.

Jean lives in Jacmel, a town in the department Sud-Est in Haiti. To communicate his observations there exist several channels. Maybe the most well-known channel is Twitter. With hashtags (like #haiti), the Twitter message can be tagged. In recent disasters people used hashtags to mark their messages as related to the disaster. This facilitated the triage of Twitter messages. Other examples of communication channels are websites that allow reporting or special emergency report numbers that mobile providers made available. To send this message Jean needs the following:

1. A device to access a communication network.
2. A communication network to access services.
3. A service that allows sharing human observations.

We have hardly any influence on meeting Jean's needs, only the service can be provided externally. There are specific platforms such as Ushahidi, but also more general solutions like Twitter or Facebook can be used.

Jean sends a message via Twitter, to report shortage of food, water and medication. Jean wrote the message in French Creole, his native language:

Nou bezwen aide en urgence nou nan place en face palais justice la .gen
anpil ti bb, nou bezwen dlo , mangé , médicament¹¹

Reporting Platform. Jean's message is broadcast on Twitter. Based on the location information that Twitter provides, the message is identified as coming from Haiti. A volunteer from the Ushahidi Haiti team enters the message into the system [10]. Desiree, who knows some French Creole and English, translates the message:

We need help. We are located at the place in front of the Palais de Justice
There's a lot of babies. We need water, food and medication.

⁸ Jean can be found at <http://personas.mspace.fm/wiki/Jean>

⁹ Desiree can be found at <http://personas.mspace.fm/wiki/Desiree>

¹⁰ Paul can be found at <http://personas.mspace.fm/wiki/Paul>

¹¹ This is a message taken from Ushahidi, it can be found at: <http://haiti.ushahidi.com/reports/view/3815>

The message can be interpreted in terms of the Ushahidi schema. Ushahidi uses ten fields to describe a report. The fields contain the message, a title, the date, the location (place name and coordinate) and categories. The categories are defined by the Ushahidi Haiti team, each report can be in one or more categories. The message above was put into the categories “Water Shortage”, “Food Shortage” and “Hospital Operating”. Additionally, Ushahidi uses two fields to flag approval and verification. The reporting platform Ushahidi makes the reports available as CSV file and as RSS feed. Additionally, Ushahidi publishes an interactive map of reports online.

Paul, the Information Manager. Paul, an Information Manager from an emergency cluster¹² is overwhelmed by the requirement to acquire and process information for both local office and the headquarters.

Information managers from a humanitarian agency struggle to gather ground reality information. To support timely informed decision making process and save more lives, the primary task of Paul as humanitarian cluster information manager is to gather information about the following fundamental questions [14]:

- What type of disasters occurred when and where?
- How many people are affected/have died/are injured?
- Which humanitarian agency is currently working in the region?
- Which agency is doing what kind of humanitarian response, where and when?
- What are the most urgent life saving humanitarian needs and what are the gaps that need urgent attention?

However, even though Paul is aware of the Ushahidi service for Haiti, Paul simply does not have the time to integrate Ushahidi’s CSV files or RSS feeds into his information system. Furthermore, he struggles with the meaning of categories and the problem of trusting this new and rather unknown information source.

The lifecycle up to now reflects the situation in Haiti in January 2010. Fig. 2 wraps up the lifecycle of an emergency report mediated through Ushahidi.

Linked Open Data. To overcome Paul’s problems, the authors transformed the Ushahidi Haiti dataset into Linked Open Data. A simple Java program based on JENA¹³ read the CSV file and wrote the RDF graph. The triples employ standard vocabularies like Dublin Core¹⁴ where possible. When there are no vocabularies to express required information, we created new ones¹⁵. An example of a triplified report is depicted in Listing 1.

¹² Emergency Clusters eg. Food, Shelter or Health <http://www.humanitarianinfo.org/iasc/pageloader.aspx?page=content-focalpoint-default>

¹³ see <http://jena.sourceforge.net/>

¹⁴ The Dublin Core Metadata Element Set: <http://dublincore.org/documents/dces/>

¹⁵ cf. <http://observedchange.com/moac/ns/> and <http://observedchange.com/tisc/ns/>

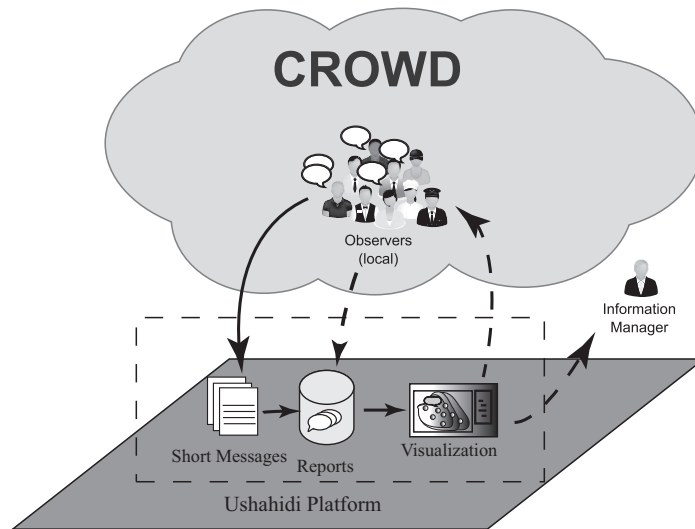


Fig. 2. A report by local observers is entered into the Ushahidi database and subsequently considered in map-visualizations. An information manager has access to the reports and to the visualizations, but can not integrate these sources into his information system. The dashed lines indicate channels that are not fully utilized.

Listing 1. The tripleset for an Ushahidi Report in Turtle form. We left out the prefix specification here.

```
@base <http://haiti.ushahidi.com/reports/view> .
</3815> a </UshahidiReport> ;
  tisc:temporalDistance 9 ;
  dc:subject moac:WaterShortage ;
    moac:FoodShortage ;
    moac:HospitalOperating ;
  dc:coverage "Coordinates are for Palais de Justice in
  Jacmel" ;
  geo:latitude 18.233695 ;
  geo:longitude -72.536217 ;
  sioc:content "We need help.We are located at the place in
  front of the Palais de Justice There\\\\\\' s a lot of
  babies.We need water,food and medication. \\t
```

```
Nou bezwen aide en urgence nou nan place en face palais
justice la .gen anpil ti bb, nou bezwen dlo , mang?, m?
dicament Time:2010-01-21 19:58:06" ;
  moac:verified false ;
  moac:approved true ;
  dc:title "Water, Food, Medicine Needed at Palais de
  Justice, Jacmel" ;
  dc:date "2010-01-21T19:58:00-05:00"^^xsd:dateTime .
```

The reports are made available online on a website and can be inserted into an etherpad¹⁶ for collaborative editing. The RDF content of the etherpad can be retrieved directly from other websites through the export plain text function.

Desiree, the Web User. Desiree studies Geoinformatics, and tutors a course on Linked Open Data. On Twitter she heard about the needs of the people in Haiti. She also heard of the information integration problems and of our effort to crowdsource Linked Open Data. She downloads a report and starts editing it. She detects that the category “Hospital Operating” was assigned falsely, she modifies the respective triple to point to the category “Medical Equipment And Supply Needs”. She then introduces two new triples. Desiree’s new triples are shown in Listing 2.

Listing 2. Two additional triples that enrich the original tripleset of report 3815.
`</3815> tisc:locatedAt <http://www.geonames.org/3723779/jacmel.html> .
dc:subject <http://dbpedia.org/page/Jacmel> .`

These triples lead to more information about Jacmel. Among others, the population figures and the current mayor are available there. With her contribution Desiree helps to extend the triples with information that cannot be harvested automatically from Ushahidi. Furthermore, the introduction of an additional crowdsourcing layer for processing adds some quality control.¹⁷ The misclassification of medical needs to “hospital operating” in this example can have fatal consequences in practice. There are the following requirements on Desiree’s task:

4. Access to the Internet.
5. Access to triples with the basic information.
6. Basic knowledge about triples, linked data and RDF.
7. Basic knowledge about disaster management and its terminologies.
8. A service to edit and validate the triples.
9. Vocabularies to edit and create triples.

We take the access to the Internet for granted. Desiree is not in Haiti but processes the report remotely. Access to the triples can be ensured through a website that makes the reports available in triple format. Desiree needs some basic knowledge to create RDF triples. This limits the crowd. However, many crowdsourced processing tasks require certain skills. We think that given an initial set of triples and a short overview of basic turtle syntax, the addition of triples in turtle format is not expected too much of a proficient web user. The point to consider is the additional interpretation that is made by us here when creating the basic set of triples. To describe the report, we use predicates and

¹⁶ see <http://etherpad.org/> for more information.

¹⁷ Desiree’s tasks involves translation between natural languages and from a natural to a formal language. This can be considered human computation [13]. In fact, many of the tasks described here as crowdsourcing may lay at the intersection of crowdsourcing and human computation as described in [13].

objects that come from shared vocabularies. There exist several well-established vocabularies like Dublin Core or SIOC to describe the formal aspects of the report, but to annotate the content Desiree needs domain specific vocabularies or ontologies. To our knowledge, no vocabularies or ontologies exist that specifically allow describing observations made by victims of disasters. The Ushahidi team came up with a taxonomy of incidents. To describe these incidents in triple format we translated the Ushahidi categories into an RDF vocabulary. However, as also [6] pointed out, further work is required to establish vocabularies and ontologies for disaster management. The lack of vocabularies in this domain constitutes one obstacle to information integration in crisis management. Etherpads are one option to edit triples collaboratively. Etherpad is an open source text editing service. It is easy to handle and the content, e.g. the triples, can be directly accessed from a static URL that returns the content of the etherpad as plain text¹⁸. Etherpads also allow creating different versions that can be individually referenced. However, so far Etherpads only allow writing plain text, there is no syntax highlighting or validation of triples. To validate the edited tripleset Desiree has to resort to another service, for example *sindice inspector*¹⁹.

Hence, editing and validating triples is possible, but there is no integrated service that meets all requirements. Furthermore, necessary vocabularies are still missing.

Visualization. With the reports given in triple form we can take the next step. We have set up a website²⁰ that uses SIMILE Exhibit²¹ to visualize the triples in different forms, for example as timeline, on a map or as thumbnails. Fig. 3 shows a screenshot of a map with report locations.

The illustration shows only one way of using the reports in Linked Open Data. There are not only many more tools for analysis and visualization available, but also further Linked Open Data can be integrated with the Ushahidi reports. Linked Open Data is supported by a growing community of users. The processing, analysis and visualization of Linked Open Data can be seen as an additional crowdsourcing task. However, this requires knowledge of web frameworks that exceeds the typical web user's expertise.

Paul, the Information Manager. Given the reports in Linked Open Data, Paul can access the reports far easier. The linked structure allows for an easier filtering of information, by category, by time, by location and so on. Paul works in an emergency cluster for medical support, the change that Desiree made to the report is important when he decides to distribute the resources he has.

Within 24 hours of the disaster, Paul needs to provide evidence based "humanitarian needs and gaps" figures in a disaster affected area. The chances for

¹⁸ for example <http://pad.ifgi.de/ep/pad/export/haiti-ushahidi-report3815/latest?format=txt>

¹⁹ see <http://inspector.sindice.com/>

²⁰ see <http://www.observedchange.com/demos/linked-haiti/>

²¹ see <http://www.simile-widgets.org/exhibit/> for more information.



Fig. 3. A screenshot of the SIMILE Map Widget. The map illustrates the reports, coloured by category. A click on the marker shows the full tripleset for the report.

a quick fact finding rapid assessment [11] are very low. The reports by local observers are coming in nearly immediately after the earthquake.

For instance, Paul finds out that he can identify at least what the urgent life saving humanitarian concerns are and where. As information manager Paul knows and understands that trust and quality of crowd information are a concern. However, such quick access to crowd information and visualization within the first 2-3 days of a disaster can help better plan the fact finding mission or possible future assessments to further confirm the humanitarian needs and possible intervention in the areas. Paul's basic requirements are:

10. A portal that gathers and integrates crowdsourced information sources.
11. A way to connect his information system to information sources in RDF.

We have exemplary set up a facet-based browsing facility to access the triples we used at <http://www.observedchange.com/demos/linked-haiti/>. The idea is that it is easily possible to extend the portal and to connect it to further information sources. However, the problem remains unsolved on the side of Paul's humanitarian information system. The problem has been recognized [6] and first solutions (e.g., [1]) are under development.

Summary. The new lifecycle of Jean's reports is illustrated in Fig. 4. An additional step of crowdsourcing Linked Open Data refines the data. Linked Open Data allows not only Paul to access the data according to well-established principles, but it also enables members of a world-wide Linked Open Data community to create mash-ups of various sources and tools.

4 Conclusion

This paper suggested crowdsourcing Linked Open Data as the next step towards a full exploitation of crowdsourced information in disaster management. Due to

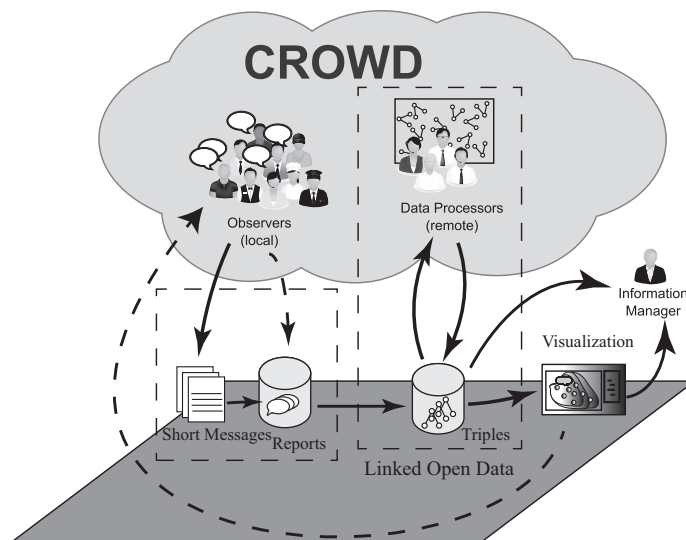


Fig. 4. Workflow including crowdsourced Linked Open Data. An intermediate crowdsourcing task of enriching triples is implemented now. The dashed lines indicate channels that are available but not at the focus of this paper.

the amount of information required and the short time available, crowdsourcing is a promising candidate for disaster management. Linked Open Data serves as common exchange format. Its simplicity and the large community behind it make it well suited for crowdsourcing efforts. The suggested approach has the potential to solve the problems of structural and semantic interoperability [4]. Nonetheless, the issues of uncertainty and trust remain unsolved in our example.²² However, in the response phase immediately after a disaster occurs, when there is no information available, trust and uncertainty issues of crowdsourced information are accepted. Furthermore, to the authors' knowledge, no standardized and formalized vocabulary for disaster management exists. Finally, it is up to the relief organizations to make their systems and data ready for Linked Open Data.

Acknowledgements

This work has been partly supported through the International Research Training Group on Semantic Integration of Geospatial Information by the DFG (German Research Foundation, GRK 1498), by the China Scholarship Council (CSC) and through the Erasmus Mundus Master Program in Geospatial Technologies (contract no. 2007-0064/001/F.R.A.M.E. MUN123).

²² see [3] for an example of a trust and reputation system for human observations of water availability in dry regions.

References

1. Babitski, G., Bergweiler, A., Grebner, O., Oberle, D., Paulheim, H., Probst, F.: Soknos using semantic technologies in disaster management software. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) *The Semantic Web: Research and Applications*. Lecture Notes in Computer Science, vol. 6644, pp. 183–197. Springer Berlin / Heidelberg (2011)
2. Berners-Lee, T.: Linked Data. W3C (2006), <http://www.w3.org/DesignIssues/LinkedData>
3. Bishr, M.: A Trust and Reputation Model for Evaluating Human Sensor Observations. Ph.D. thesis, University of Muenster, Germany, Muenster, Germany (2011), url = http://miami.uni-muenster.de/servlets/DerivateServlet/Derivate-6032/diss_bishr.pdf
4. Bishr, Y.: Overcoming the semantic and other barriers to gis interoperability. *International Journal of Geographical Information Science* 12(4), 299–314 (1998)
5. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data The Story So Far. *International Journal on Semantic Web and Information Systems* 5(3), 2–9 (2009)
6. Di Maio, P.: An open ontology for open source emergency response systems, see <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.93.1829>
7. Feridun, M., Tanner, A.: Using linked data for systems management. In: *NOMS'10*. pp. 926–929 (2010)
8. Goodchild, M.: Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4), 211–221 (2007)
9. Howe, J.: The rise of crowdsourcing. *Wired magazine* 14(6), 1–4 (2006)
10. Morrow, N., Mock, N., Papendieck, A. Kocmich, N.: Independent evaluation of the ushahidi haiti project. Tech. rep., The UHP Independent Evaluation Team (2011)
11. OCHA: Assessment and classification of emergencies (ace) project. Tech. rep., United Nations Office for the Coordination of Humanitarian Affairs (2009)
12. Okolloh, O.: Ushahidi, or 'testimony': Web 2.0 tools for crowdsourcing crisis information. *Participatory Learning and Action* 59(1), 65–70 (2009)
13. Quinn, A.J., Bederson, B.B.: Human computation: a survey and taxonomy of a growing field. In: *Proceedings of the 2011 annual conference on Human factors in computing systems*. pp. 1403–1412. CHI '11, ACM, New York, NY, USA (2011)
14. UNICEF: Emergency Field Handbook: A guide for UNICEF staff. The United Nations Childrens Fund (UNICEF) (2005), url = http://www.unicef.org/publications/files/UNICEF_EFH_2005.pdf
15. Zook, M., Graham, M., Shelton, T., Gorman, S.: Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. *World Medical & Health Policy* 2(2), 2 (2010)

Opinion Mapping Travelblogs

Efthymios Drymonas, Alexandros Efentakis, and Dieter Pfoser

Institute for the Management of Information Systems
Research Center Athena
G. Mpakou 17, 11524 Athens, Greece
{edrimon|efentakis|pfoser}@imis.athena-innovation.gr

Abstract. User-contributed content represents a valuable information source provided one can make sense of the large amounts of unstructured data. This work focusses on geospatial content and specifically on travelblogs. Users writing stories about their trips and related experiences effectively provide geospatial information albeit in narrative form. Identifying this geospatial aspect of the texts by means of applying information extraction techniques and geocoding, we relate portions of texts to locations, e.g., a paragraph is associated with a spatial bounding box. To further summarize the information, we assess the opinion (“mood”) of the author in the text. Aggregating this mood information for places, we essentially create a geospatial opinion map based on the user-contributed information contained in the articles of travelblogs. We assessed the proposed approach with a corpus of more than 150k texts from various sites.

1 Introduction

Crowdsourcing moods and in our specific case opinions from user-contributed data, has recently become an interesting field with the advent of micro-blogging services such as, e.g., Twitter. Here, blog entries reflect a myriad of different user opinions that when integrated can give us valuable information about, e.g., the stock market [4]. In this work, our focus is on (i) extracting the user opinion about places from travel blog entries, (ii) aggregating such opinion data, and, finally, (iii) visualizing it.

The specific contributions in this work are as follows. In an initial stage several travelblog Web sites have been crawled and over 150k texts have been collected. Figure 5 shows such an example travelblog entry. The collected texts are then geoparsed and geocoded to link placename identifiers (toponyms) to location information. With paragraphs as the finite granularity for opinion information, texts are then assessed with the OpinionFinder tool and are assigned a score for each paragraph ranging from very negative to very positive. Scores are linked to the bounding box of the paragraph and are aggregated using a global grid, i.e., the score of a specific paragraph is associated with all intersecting grid cells. Aggregation of opinions is then performed simply by computing the average of all scores for each cell. Finally, the score can be visualized by assigning colors to each cell.

While to the best of our knowledge there exists no work aiming at extracting opinions about places from travel blogs, we can cite the following related work. The concept of information visualization using maps is gaining significant interest in various

research fields. As examples, we can cite the following works [26] [23] [29] [12]. For the purpose of recognizing toponyms, the various approaches use ideas and work from the field of Natural Language Processing (NLP), Part-Of-Speech (POS) tagging and a part of Information Extraction related tasks, namely Named Entity Recognition (NER) [13]. These approaches, can be roughly classified as rule-based [6] [7] [8] [21] [30] and machine learning - statistical [14] [16] [26] [22]. Once toponyms have been recognized, a toponym resolution procedure resolves geo-ambiguity. There are many methods using a prominence measure such as population combined with other approaches [21] [25]. With respect to geocoding, we can exemplary cite [17], one of the first works on geocoding and describing a navigational tool for browsing web resources by geographic proximity as an alternative means for Web navigation. Web-a-Where [1] is another system for geocoding Web pages. It assigns to each page a geographic focus that the page discusses as a whole. The tagging process targets large collections of Web pages to facilitate a variety of location-based applications and data analyses. The work presented in [15] is identifying and disambiguating references to geographic locations. Another method that uses information extraction techniques to geocode news is described in [26]. Other toponym resolution strategies involve the use of geospatial measures such as minimizing total geographic coverage [14], or minimizing pairwise toponym distance [16]. An approach for the extraction of routes from narratives is given in [9]. The proposed IE approach has been adapted to fit the requirements of this work. While statistical NER methods can be useful for analysis of static corpora, in the case of continuously user contributed travel narratives they are not well-suited, due to their dynamic and ever-changing nature [25]. For this purpose, we rely on a powerful rule-based solution based on a modular pipeline of distinct, independent and well-defined components based on NLP and IE methods, as we will see in the next section. Regarding related work on opinion classification and sentiment analysis [20], we can find methods basically relying on streaming data [18] [10] [11] [19]. Recently [2] discusses the challenges that Twitter streaming data poses. The work focusses on sentiment analysis and proposes the sliding-window Kappa statistic as an evaluation metric for data streams.

The remainder of this work is organized as follows. Section 2 describes the information extraction techniques employed in our approach dealing specifically with the aspects of geoparsing and geocoding travel blog entries. Section 3 outlines a method for computing user sentiment scores from travel blog entries. Section 4 outlines how such scores can be aggregated and visualized based on geospatial locations. In addition some specific examples are shown to give an initial validation of the proposed approach. Finally, Section 5 presents conclusions and directions for future work.

2 Information Extraction

In what follows, we describe in detail the processing pipeline, which overall uses an HTML document as input (travel blog article) and produces a structured XML file containing the various entities and their respective attributes (toponyms and coordinate information for the text).

The pipeline consist of four parts (cf. Figure 1), (i) the HTML parsing module, (ii) the linguistic pre-processing, (iii) the main IE engine system (semantic analysis) and

(iv) the geocoding-postprocessing part. In the next section, we describe the first part of our processing pipeline, i.e., the collection of HTML texts, the parsing and their conversion to plain text format, in order to prepare the documents for the forthcoming step of linguistic-preprocessing.

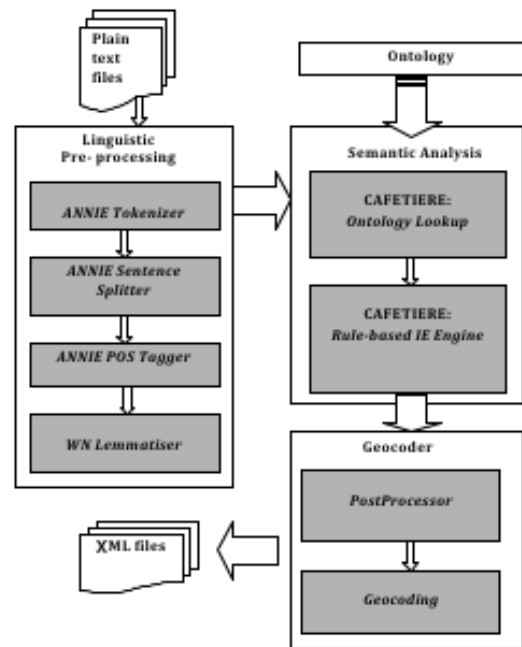


Fig. 1. IE architecture pipeline

2.1 Web Crawling

For collecting travel blog articles containing rich geospatial information, we crawled Web sites providing travelblog authoring services. Each Web site has its own HTML layout and isolating text of interest from crawled and parsed HTML pages is done by hand. Thus, there was a need for Web sites with massive amounts of such type of documents. For this purpose we crawled travelpod.com, travelblog.org, traveljournals.net and worldhum.com, resulting in more than 150,000 documents. For crawling the web sites, we used Regain crawler¹, which creates a Lucene² index for indexing the documents' information, while for HTML parsing and the extraction of useful plain text narratives, we used Jericho HTML parser³.

¹ <http://regain.sourceforge.net/>

² <http://lucene.apache.org/>

³ <http://jericho.htmlparser.net/>

2.2 Linguistic pre-Processing

To prepare the input for the core IE engine for extracting objects of interest, the parsed plain text documents must be prepared accordingly. Such preparation includes linguistic pre-processing tools that analyze natural language documents in terms of distinct base units (i.e., words), sentences, part-of-speech and morphology . We are using the ANNIE tools, contained in the GATE release⁴, to perform this initial part of analysis. To this task, our processing pipeline comprises of a set of four modules: (i) the ANNIE tokenizer, (ii) the (ANNIE) Sentence Splitter, (iii) the ANNIE POS Tagger and (iv) the WordNet Lemmatiser.

The intermediate processing results are passed on to each subsequent analysis tool as GATE document annotation objects. The output of this analysis part is the analyzed document and it is transformed in CAS/XML format⁵, which will be passed to the subsequent semantic analysis component as input, Cafetiere IE engine [3]. Cafetiere combines the linguistic information acquired by the pre-processing stage of analysis with knowledge resources information, namely the lookup ontology and the analysis rules to semantically analyze the documents and recognize spatial information, as we will see later in this section.

The first step in the pipeline process is *tokenization*, i.e., recognizing in the input text basic text units (tokens), such as words and punctuation and orthographic analysis and the association of orthographic features, such as capitalization, use of special characters and symbols, etc. to the recognized tokens. The tools used are ANNIE Tokenizer and Orthographic Analyzer.

Sentence splitting, in our case the ANNIE sentence splitter aims at the identification of sentence boundaries in a text.

Part-of-speech (POS) tagging is then the process of assigning a part-of-speech class, such as Noun, Verb etc. to each word in the input text. The ANNIE POS Tagger implementation is a variant of Brill Transformation-based learning tagger [5], which applies a combination of lexicon information and transformation rules for the correct POS classification.

Lemmatisation is used for text normalisation purposes. With this process we retrieve the tokens base form e.g., for words: [travelling, traveler, traveled], [are, were], the corresponding lemmas are: travel, be. We exploit this information in the semantic rules section. For this purpose we implement the JWNL WordNet Java Library API⁶ for accessing the WordNet relational dictionary. The output of this step is included it in GATE document annotation information.

2.3 Semantic Analysis

Semantic analysis relates the linguistic pre-processing results to ontology information, as we will see in the next subsection about ontology lookup and applies semantic analysis grammar rules, i.e., documents are analyzed semantically to discover spatial concepts and relations.

⁴ <http://gate.ac.uk/>

⁵ CAS is an XML scheme called Common Annotation Scheme allowing for a wide range of annotations, structural, lexical, semantic and conceptual.

⁶ <http://sourceforge.net/projects/jwordnet/>

For this purpose we used Cafetiere IE engine, whose objective is to compile a set of semantic analysis grammar rules in a cascade of finite state transducers so as to recognize in text the concepts of interest. Cafetiere IE Engine combines all previously acquired linguistic and semantic information with contextual information. We modified Cafetiere and implemented it as a GATE pipeline module (GATE creole) for the purpose of performing ontology lookup and rule-based semantic analysis on information acquired from previous pipeline modules, in the form of GATE annotation sets. The input to this process are the GATE annotation objects resulted from the linguistic pre-processing stage stored transformed in Cafetiere needed format, in CAS/XML format for each individual document.

Cafetiere Ontology Lookup The use of knowledge lexico-semantic resources assists in the identification of named entities. These semantic knowledge resources may be in the form of lists (gazetteers) or more complex ontologies providing mappings of text strings to semantic categories, such as in general male/female person names, known organizations and known identifiers of named entities. In our case, the named entities we want to extract with IE methods are location based information. For example, a gazetteer for location designators might have entries such as “Sq.”, “blvd.”, “st.” etc. that denote squares, boulevards and streets accordingly. Similarly there are various sorts of gazetteers available for given person names, titles, location names, companies, currencies, nationalities etc. Thus, the named entity (NE) recognizer can use gazetteer information so as to classify a text string as denoting an entity of a particular class. However, in order to associate specific individual entities object identifiers are required as well as class labels, enabling aliases or abbreviations to be mapped to a concrete individual. For example, for an entity such as “National Technical University of Athens” the respective abbreviation “NTUA” could be included in the knowledge resource as an alias for the respective entity. Thus, more sophisticated knowledge resources than plain gazetteers in the form of ontologies may be used to provide this type of richer semantic information and allow for the specification and representation of more information, if necessary, than identity and class inclusion.

In this way, Cafetiere Ontology lookup module accesses a previously built ontology to retrieve potential semantic class information for individual tokens or phrases. All types of conceptual information, related to domain specific entities, such as terms or words in general that denote spatial concepts or properties and relations of domain interest are pre-defined in this ontology. For example, consider the partial ontology shown in Figure 2. Class “LOCVERB” stores verbs that when matched to a text phrase are likely to indicate a spatial relationship between the corresponding referenced concepts. We label as semantic any classification of tokens according to their meaning in the field of the application, in our case, geosemantics. This could be done, on a broad coverage level, by reading information from a comprehensive resource such as WordNet lexicon about most content words. However, the practice in information extraction applications as discussed in previous paragraph, has been to make the processing application-specific by using lists of the semantic categories of only relevant words and phrases, done by hand. The ontology used in our experimentation was created by manually analyzing a large number of texts and iteratively refining the ontology with words (e.g., verbs) that

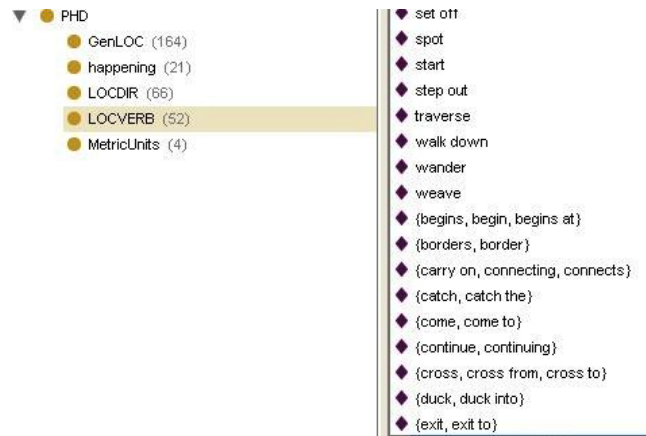


Fig.2. Sample ontology contents (Protg ontology editor)

when matched to a text phrase are likely to indicate a spatial relationship between the corresponding referenced concepts. Summarizing, the lookup stage of analysis:

1. Supplies semantic classes (concepts) corresponding to words and phrases.
2. Supplies object identifiers for known instances, including where aliases and abbreviations name the same instance (For example “National Technical University of Athens”, “NTUA”).
3. Supplies properties of known instances, for example the country of which a city is the capital.
4. Uses verbs of interest to the application in order to identify inside the phrase potential unknown instances.

Cafetiere Information Extraction engine The approaches to Named Entity recognition with IE methods can be divided into two main categories:

- Linguistic/rule-based approaches: in these approaches the Named Entity recognition is based on linguistic/semantic rules defining the possible linguistic patterns denoting Named Entity concepts, such as for example the approaches adopted by ANNIE⁷, and Cafetiere [3]. These approaches can achieve better results than most statistics or machine learning approaches, but they require extensive human effort for the development of the necessary knowledge resources (rules and lexico-semantic resources, like ontologies, described in Cafetiere ontology lookup section). For this reason the adaptation of rule-based systems to new domains is a slow and laborious process.

⁷ <http://gate.ac.uk/>

- Machine learning/statistics-based approaches: these approaches view Named Entity recognition as a classification problem and, they have gained increased popularity due to their relatively rapid development/ domain customization and the reduced amount of human effort required.

Cafetiere is a rule-based system for IE. A set of linguistic patterns (i.e., extraction rules) is written taking into account the lookup ontology and all previously acquired information from linguistic pre-processing. The semantic analysis rules, are developed as a set of context-sensitive/context-free grammar (CSG/CFG) rules and are compiled in a cascade of finite state transducers so as to recognize the concepts of interest in plain texts.

2.4 PostProcessing

In this part, all information regarding each object of interest for each document in the collection is imprinted as a GATE annotation set object. For each document, we have collected information about all extracted entities, along with their respective paragraph, sentence and character offset in this document. During the HTML parsing process we keep the scope that each document is referred to in order to use this information for geocoding each extracted entity. For geocoding, we initially implemented YAHOO! Placemaker⁸ and used in combination with Cafetiere’s output, in order to deliver better results. We observed that PlaceMaker worked well for disambiguating some entities, but it identified significant fewer place entities than our IE engine. Thus, in the remaining entities extracted by Cafetiere, we applied YAHOO! Placefinder⁹ to geocode this place information passing the scope information described below for delivering more accurate results.

Finally, for each HTML travel blog entry (narrative), we created a collection of extracted referred geo-entities, some of them not being able to geocode. For each of these entities there is specific information (acquired from each of the previous pipeline steps) about where they were encountered in the respective document, namely, sentence, paragraph and offset character. Additionally, for each document, we calculated the mean coords and standard distance from all geocoded points extracted. All this information, along with the local parsed text file path and the respective URL of the document, are stored lastly into XML format for each corresponding plain text narrative. Samples of plain text narrative and the corresponding structured XML file are shown in Figure 3 and Figure 4 respectively. The XML tags in Figure 4 are denoting either statistical information, like the mean center and the standard distance of all geocoded locations for each document, or information related with each extracted entity, i.e., the offset characters, the sentence and paragraph ID.

3 Opinion Mapping

Having geocoded the travel blog entries, we, in the following step, want to assign sentiment information (“mood”) to text. To this effect, we use OpinionFinder [28], a sys-

⁸ <http://developer.yahoo.com/geo/placemaker/>

⁹ <http://developer.yahoo.com/geo/placefinder/>

```

On Christmas Eve the Spanish eat a big dinner which usually
includes seafood. And "Papa Noel" is gaining popularity, but
it's more traditional to give gives on "El Día de los Reyes
Magos" (The Day of the Wise men), which falls on January 5th
this year. I have heard that I have a better chance of
seeing snow in Salamanca, so...
Plaza de Fonseca, this small plaza is located right behind
the Cathedral.
Plaza de Obradoiro
Very impressive.
Nativity

```

Fig. 3. Sample plain text

```

-<Document filename="data/texts1/729.txt" placeReferred="Spain"
meanCoords="42.018,-6.07" standardD="564.67" totalNumOfTokens="524"
totalGeocodedSuccessfully="11" totalExtracted="12">
  <poi name="Salamanca" startOffset="2509" endOffset="2518" sentenceID="30"
paragraphID="15" coords="40.9642,-5.66385" accurCode="0"/>
  <poi name="Plaza de Fonseca" startOffset="2558" endOffset="2574" sentenceID="31"
paragraphID="16" coords="40.579929,-6.584242" accurCode="0"/>
  <poi name="is located right" startOffset="2592" endOffset="2608" sentenceID="31"
paragraphID="17" coords="ISRELATION"/>
  <poi name="Cathedral" startOffset="2620" endOffset="2629" sentenceID="31"
paragraphID="17" coords="NULL" accurCode="NULL"/>
  <poi name="Plaza de Obradoiro" startOffset="2631" endOffset="2649" sentenceID="32"
paragraphID="18" coords="39.256985,-5.806305" accurCode="0"/>

```

Fig. 4. Resulting XML file

tem that performs *subjectivity analysis*, automatically identifying when opinions, sentiments, or speculations are present in text. It aims to identify subjective sentences, as also marking various aspects of subjectivity in these sentences, including the source (holder) of the subjectivity and words that are included in phrases expressing positive or negative sentiments.

OpinionFinder operates as one large pipeline. Conceptually, the pipeline can be divided into two parts. The first part performs mostly general purpose document processing (e.g., tokenization and part-of-speech tagging). The second part performs the subjectivity analysis. The results of the subjectivity analysis are returned to the user in the form of SGML/XML markup of the original documents.

For the first part, OpinionFinder takes any incoming text source and removes HTML or XML meta info. Sentences are split and POS tagged using OpenNLP¹⁰, the open source solution providing a variety of java-based NLP tools which perform sentence detection, tokenization, pos-tagging, chunking and parsing, named-entity detection, and coreference using the OpenNLP Maxent machine learning package. Next, stemming is accomplished using Steven Abneys' SCOL v1K stemmer program¹¹. SUNDANCE (Sentence UNDERstanding And ConceptExtraction) [28], is used to provide semantic

¹⁰ <http://opennlp.sourceforge.net/>

¹¹ <http://www.vinartus.net/spa/>

class tags, identify extraction patterns needed by the sentence classifiers, identifying the source of subjective content and distinguishing author statements from related or quoted statements. A final parse in batch mode establishes constituency parse trees, which are converted to dependency parse trees for Named Entity and subject detection.

At this point, for the second part, a Naive Bayes classifier identifies subjective sentences. The classifier is trained against subjective and objective sentences generated by two additional rule-based classifiers drawing from large corpora [27]. Next, a direct subjective expression and speech event classifier tags the direct subjective expressions and speech events found within the document using WordNet¹². The final step applies actual sentiment analysis to sentences that have been identified as subjective. This is accomplished with two classifiers that were developed using the BoosTexter [24] machine learning program and trained on the MPQA Corpus¹³.

4 Mapping Opinion Scores

OpinionFinder produces sentiment information assigned to paragraphs of texts. In the following, we describe how this information can be aggregated for specific locations.

4.1 Aggregating Sentiments

OpinionFinder was applied to all texts of our collection of 150k travel blog entries assigning sentiment data to each paragraph of the collection. In the analysis that follows, only paragraphs containing geospatial data were retained. For each of these paragraphs we keep the total referred positive and negative sentiment scores as computed by OpinionFinder.

Each paragraph contains zero, one or multiple geographic entities that were suitably geocoded. In order to show the spatial extent of a paragraph, we chose to spatially visualize only paragraphs in which the MBR of the contained toponyms does not exceed 0.5 degrees in either dimension (e.g., $Max\ latitude - Min\ latitude \leq 0.5$ AND $Max\ longitude - Min\ longitude \leq 0.5$). Consequently, only paragraphs of limited and focused spatial extent are visualized, thus preventing paragraphs that refer to larger geographic entities (e.g., Europe) to dominate in the results.

We used five different categories for mapping opinion scores. The categories and respective color are given in Table 1, where each category scales from negative (red) to positive (green).

The proposed approach is clarified by the following example. A sample document¹⁴ (Figure 5) contains several paragraphs mentioning Washington D.C. and its landmarks. For each of this document's paragraphs, a MBR covering the discovered toponyms was created and each paragraph was assigned a category according to Table 1. Therefore, this document may be spatially visualized on a map as shown on Figure 6.

Although this approach is viable when there is a limited number of documents and paragraphs, we need to overcome the following problem. Multiple paragraphs from different documents and different scores may partially target the same area, e.g., we need

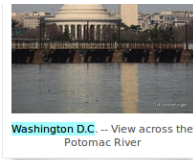
¹² <http://wordnet.princeton.edu/>

¹³ <http://nrrc.mitre.org/NRRC/publications.htm>.

¹⁴ <http://www.travelpod.com/travel-blog-entries/dfumblefinger/1/1269627251/tpod.html>

Result (positive - negative)	Colour
≤ -3	Red
$=-1$ OR $=-2$	Orange
0	Yellow
$=1$ Or $=2$	Olive
≥ 3	Green

Table 1. Opinion mapping to colour representation



The **National Mall** is a **rectangular area** which forms the heart of Washington and which is a unit of the **National Park** Service. Anchored on its east end by the **Capitol building** and on the west end by the **Lincoln Memorial**, it encompasses the land between Constitution and Independence Avenues. This is the one place you absolutely must visit if you have only one or two days in town and its possible to spend weeks here and not see everything. No matter how much time you have, its worth trying to take in as much history as you can. This requires the use of your feet so if you're able, walk the entire stretch and explore its historic richness.

At the Center of the Mall is the **Washington Monument**. Standing over 555 feet tall it easily is the most dominant structure in the city. Built as a tribute to our first (and probably best) President, George Washington, the monument

Fig. 5. Washington D.C. - sample document and toponyms

to visualize partially overlapping MBRs with different scores (colors). To do that, we split each paragraph MBR into small cells of a regular grid of 0.0045 degrees (corresponding to 500m) in each dimension. For each of those cells we sum up the sentiment score from all the containing paragraph MBRs. With this approach, instead of trying to visualize *overlapping* paragraph MBRs with different scores (colors), we visualize *distinct* small cells with each being assigned a unique score (and color). Consequently, it is easy to visualize the overall sentiment scores independent of how many paragraphs target the same area.

4.2 Opinionmap Examples

Further examples shown in the following include the geospatial opinion map of Amsterdam of Figure 7. It is interesting to observe that while most of the city is shaded green, the area around the train station and the Red Light district are shown in red, i.e., expressing rather negative sentiment.

Figure 8 gives a geospatial opinion map of Central Europe indicating the areas mentioned in the travel blogs. What can be observed is that positive sentiments are associated with areas in Switzerland and also Italy, while urban areas such as Brussels overall attract more negative sentiments.

4.3 Summary

Our initial experiments with the creation of geospatial opinion maps derived from subjective travelblog entries show that there is a clear bias for certain geographic areas shared by people. However, since in this work we only performed a simple aggregation of the scores generated by the OpinionFinder tool, it will require more in-depth analysis of the results to generate accurate statements and trends.

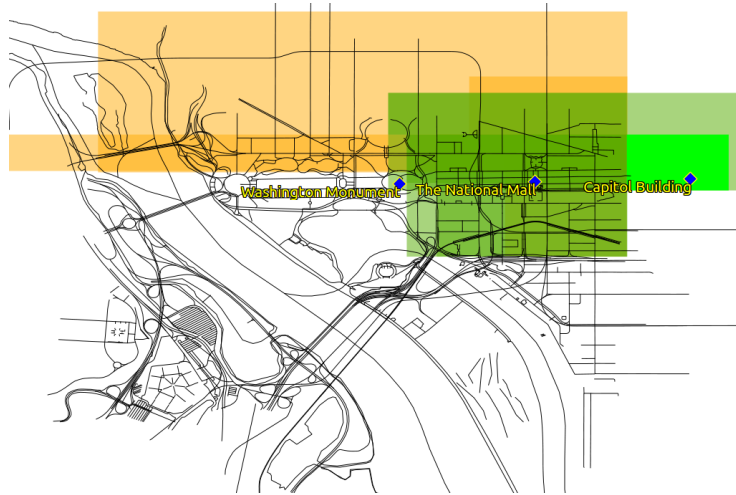


Fig. 6. Washington D.C. - geospatial opinion visualization

5 Conclusions

Aggregating opinions is important for utilizing and assessing user-generated content. This work provides a means of visualizing sentiments for specific geographic areas as derived from travel blog entries. To demonstrate the approach, several travel blog sites were crawled and a total of more than 150,000 pages/articles were processed. Using (i) geoparsing and geocoding tools the content was geo referenced and (ii) sentiment information was derived using the OpinionFinder tool. In the proposed approach, sentiment information from various articles relating to the same geographic area is aggregated and visualized accordingly by means of a geospatial heat map. Directions for future work are as follows. The current approach for aggregating user sentiment for geographic areas is rather simple and a more in-depth analysis of the results is needed to generate accurate statements and trends. An obvious improvement will also be to examine/include microblogging content streams. Here, sentiment information will be updated live and thus represent an accurate picture of the situation of a specific geographic area over time. Finally, OpinionFinder is a general purpose tool for deriving user sentiment. More involved approaches exist and need to be examined/developed for the case of geospatial data.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme - Marie Curie Actions, Initial Training Network GEOCROWD (<http://www.geocrowd.eu>) under grant agreement No. FP7-PEOPLE-2010-ITN-264994.

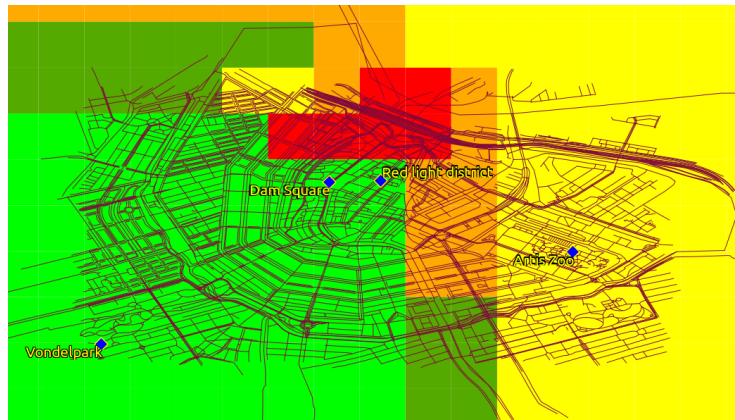


Fig. 7. Amsterdam, The Netherlands - geospatial opinion visualization

References

1. E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '04*, pages 273–280, New York, NY, USA, 2004. ACM.
2. A. Bifet and E. Frank. Sentiment knowledge discovery in twitter streaming data. In *Proceedings of the 13th international conference on Discovery science, DS'10*, pages 1–15, Berlin, Heidelberg, 2010. Springer-Verlag.
3. W. J. Black, J. McNaught, A. Vasilakopoulos, K. Zervanou, B. Theodoulidis, and F. Rinaldi. Cafetiere: Conceptual annotations for facts, events, terms, individual entities and relations. Technical report, Jan 2005. Parmenides Technical Report, TR-U4.3.1.
4. J. Bollen, H. Mao, and X.-J. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.
5. E. Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21:543–565, 1995.
6. D. Buscaldi and P. Rosso. A conceptual density-based approach for the disambiguation of toponyms. *Int. J. Geogr. Inf. Sci.*, 22:301–313, January 2008.
7. P. Clough. Extracting metadata for spatially-aware information retrieval on the internet. In *Proceedings of the 2005 workshop on Geographic information retrieval, GIR '05*, pages 25–30, New York, NY, USA, 2005. ACM.
8. J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Data Bases, VLDB '00*, pages 545–556, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
9. E. Drymonas and D. Pfoser. Geospatial route extraction from texts. In *DMG '10: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Data Mining for Geoinformatics*, pages 29–37, New York, NY, USA, 2010. ACM.
10. A. Go, R. Bhayani, and L. Huang. Twitter Sentiment Classification using Distant Supervision. Technical report, Stanford University.
11. B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Micro-blogging as online word of mouth branding. In *Proceedings of the 27th international conference extended abstracts on*

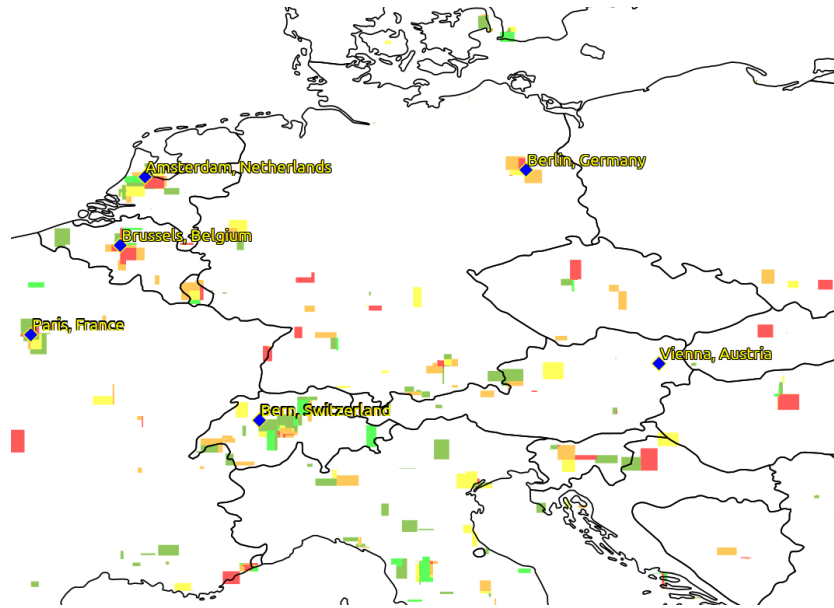


Fig. 8. Europe - geospatial opinion visualization

- Human factors in computing systems*, CHI EA '09, pages 3859–3864, New York, NY, USA, 2009. ACM.
12. R. Jianu and D. Laidlaw. Visualizing gene co-expression as google maps. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, R. Chung, R. Hammound, M. Hussain, T. Kar-Han, R. Crawfis, D. Thalmann, D. Kao, and L. Avila, editors, *Advances in Visual Computing*, volume 6455 of *Lecture Notes in Computer Science*, pages 494–503. Springer Berlin / Heidelberg, 2010.
 13. D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall, 1 edition, 2000.
 14. J. L. Leidner. Toponym resolution in text: annotation, evaluation and applications of spatial grounding. *SIGIR Forum*, 41:124–126, December 2007.
 15. M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *International Conference on Data Engineering*, pages 201–212, 2010.
 16. M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. Steward: architecture of a spatio-textual search engine. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, GIS '07, pages 25:1–25:8, New York, NY, USA, 2007. ACM.
 17. K. S. McCurley. Geospatial mapping and navigation of the web. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 221–229, New York, NY, USA, 2001. ACM.
 18. B. O'Connor, R. Balasubramanian, B. Routledge, and N. Smith. From tweets to polls: Linking text sentiment to public opinion time series, 2010.

19. A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).
20. B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135, January 2008.
21. R. S. Purves, P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, and B. Yang. The design and implementation of spirit: a spatially aware search engine for information retrieval on the internet. *Int. J. Geogr. Inf. Sci.*, 21:717–745, January 2007.
22. G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman. Determining the spatial reader scopes of news sources using local lexicons. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 43–52, New York, NY, USA, 2010. ACM.
23. R. E. Roth, K. S. Ross, B. G. Finch, W. Luo, and A. M. MacEachren. A user-centered approach for designing and developing spatiotemporal crime analysis tools. Zurich, Switzerland, 14-17th September, 2010 2010. GIScience.
24. R. E. Schapire and Y. Singer. BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning*, 39(2/3):135–168, 2000.
25. N. Stokes, Y. Li, A. Moffat, and J. Rong. An empirical study of the effects of nlp components on geographic ir performance. *Int. J. Geogr. Inf. Sci.*, 22:247–264, January 2008.
26. B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. Newsstand: a new view on news. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, GIS '08, pages 18:1–18:10, New York, NY, USA, 2008. ACM.
27. J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)*, pages 486–497, Mexico City, Mexico, 2005.
28. T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, pages 34–35, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
29. J. Zhang, H. Shi, and Y. Zhang. Self-organizing map methodology and google maps services for geographical epidemiology mapping. *Digital Image Computing: Techniques and Applications*, 0:229–235, 2009.
30. W. Zong, D. Wu, A. Sun, E.-P. Lim, and D. H.-L. Goh. On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '05, pages 354–362, New York, NY, USA, 2005. ACM.

Aligning Unions of Concepts in Ontologies of Geospatial Linked Data

Rahul Parundekar, José Luis Ambite, and Craig A. Knoblock

Information Sciences Institute and Department of Computer Science
University of Southern California
4676 Admiralty Way, Suite 1001, Marina del Rey, CA 90292
{parundek, ambite, knoblock}@usc.edu

Abstract. It is evident from the recent growth in Geospatial Linked Data that even though the number of instances being generated and linked has increased drastically, the ontologies behind these sources remain disconnected. Though we can agree that the instances being linked are equivalent, the alignments that are extrapolated from these links between the concepts may or may not agree with our intuitions. It is important to investigate how the concepts in the sources are actually aligned. Our previous work was successful in finding alignments, such as equivalence and subset relations, between concepts of two sources, using the instances that are linked as equal. Such alignments need not be trivial, however, as a concept in the ontology might not have an exact equivalent class in the other source. In this paper we propose a method that uses the subset and equivalence relations between *restriction classes* found by our previous work to find new alignments, where one (larger) concept of a source is aligned to the union of multiple (smaller) concepts from another source. We also show that we can use these alignments to find inconsistencies and use them to identify the instances that may be erroneously aligned.

1 Introduction

The Web of Linked Data has seen huge growth in the past few years. As of September 2010, the size of the Linked Open Data Cloud was about 28.5 billion triples with around 20.6% of the triples belonging to the geospatial domain.¹ As of June 2009, the cloud had recorded an overall growth of about 300% with 91% growth in the geospatial domain.² Out of the 16 geospatial data sources covered in the September 2010 count, there are around 16.5 million outgoing links to other sources. The sources of Geospatial Linked Data are most popularly connected using the *owl:sameAs* property, linking instances that are the same. As more alignments are generated in the Web of Linked Data at the instance level, a pattern of inter-linked data arises where the ontologies behind the sources

¹ <http://www4.wiwiss.fu-berlin.de/lodcloud/state/>

² <http://events.linkedata.org/ldow2011/slides/ldow2011-slides-intro.pdf>

remain un-linked. As described in our previous papers on Linking and Building Ontologies of Linked Data [7] and Aligning Ontologies of Geospatial Linked Data [6], an extensional technique can be used to generate alignments between the ontologies behind these sources. In these papers, we introduce a concept of *restriction classes*, which is similar to that of single value constraints on property restrictions of the *Web Ontology Language (OWL)* to increase the expressivity of sources with a rudimentary ontology. By looking at the set containment relationships of the instance sets of these *restriction classes*, we find equivalent and subset alignments between the two sources. Though the equivalent alignments found are precise in finding similar concepts between the two sources, the subset relations found, though informative, are too numerous to be effectively used.

Reviewing these subset relations we discovered that there are potential equivalent alignments not found by our previous work, linking a larger concept to a union or aggregation of one or more of its subsets. Using this as motivation, the work described in this paper builds on the ontology alignment method of [7]. Picking up where we left off, the approach described in this paper uses the subset relations as hints to create a union of smaller *restriction classes*, by virtue of a common property and *restriction classes* with only a single *property-value pair*, which guides the aggregation and then performs set containment operations with the larger *restriction class* from the other source. Using this method, we explore three Geospatial Linked Data sources - *GeoNames*, *DBpedia*, & *LinkedGeoData* and try to find *new* alignments between *GeoNames* & *DBpedia* and *LinkedGeoData* & *DBpedia*, where a larger subsuming *restriction class* from one source can be explained by an aggregation of smaller *restriction classes* from the other source.

The scope of this paper is in the domain of Geospatial Linked Data, where we find alignments between three sources: *GeoNames*, *DBpedia* and *LinkedGeoData*. We first find equivalences and subset relations as described in our previous work, and then use these to find the new *union alignments*. The nature of each of the three sources investigated is briefly mentioned here and they are described in more detail in [7]. *GeoNames* is a geographic source with a flat-file like ontology where all instances belong to a single concept of *Feature* and have associated *Feature Class* & *Feature Code* property to identify the instances as mountains, lakes, etc. Although *DBpedia* is a Linked Data source that covers domains other than the geospatial domain, there are a large number of instances from *GeoNames* linked to those in *DBpedia* using the *owl:sameAs* property. We also try to find alignments between the ontologies behind *LinkedGeoData* and *DBpedia*. RDF data in *LinkedGeoData* is derived from the *Open Street Map* initiative and has links to *DBpedia*.³

This paper is organized as follows. We first describe briefly our alignment algorithm from [7] along with the limitations of the results that were generated. We then explain our approach to finding alignments between a larger concept from one source and the union set of multiple smaller concepts from the other source. This is followed by identifying the outliers of these alignments that high-

³ <http://linkedgeodata.org/Datasets>

light the inconsistencies and the instances that are erroneously linked. We then describe the experimental results that contain the new alignments discovered in these data sources, along with their outliers. Finally, we describe other related work and conclude with our observations and future work.

2 Aligning geospatial ontologies on the Web of Linked Data

The work described in this paper follows our previous work on aligning ontologies of Linked Open Data, which uses an extensional approach to find alignments between *restriction classes* in two different sources. Though the results generated by our previous algorithm found equivalent alignments between the two sources, a large number of subset alignments were also found. A pattern was observed in these results, where a group of concepts from one source were subsets of the same larger concept from the other source. In many cases these smaller concepts taken together were able to completely explain the larger source. We used this insight as motivation for consuming the subset relations, which were too numerous to be useful by themselves, to find alignments between the larger concept and the union of the group of concepts. Our approach uses this group of smaller concepts and introduces a disjunction operator on these subsets to try to define the common subsuming concept.

2.1 Our previous work on linking and building ontologies of Linked Data

Ontologies of Linked Data sources can be quite rudimentary. For example, *GeoNames* only has a single concept (*Feature*) to which all of its instances belong. On the other hand, in *DBpedia*, we find a rich ontology with a hierarchy of concepts and well-defined properties. In the traditional sense of ontology alignment, we would have found at most a single alignment between *Feature* on the *GeoNames* side and a similar broad concept from *DBpedia*. In order to get a richer set of alignments, we introduced the concept of a *restriction class*. A *restriction class* is a concept that is derived extensionally and defined by the set of instances obtained by restricting a single property to a single value (called a *property-value pair* and represented by $(p_i = v_i)$) in a source. For example, a *restriction class* for schools can be constructed in *GeoNames* by forming a set of instances that have their *geonames:featureCode* restricted to ‘*S.SCH*’. This *restriction class* is represented as *geonames:featureCode=S.SCH*. The scope of the definition of a *restriction class* includes the conjunction operator, which produces a more specialized set of instances, constructed using two or more *restriction classes*. Thus, a *restriction class* $\{geonames:featureCode=S.SCH \ \& \ geonames:countryCode=US\}$, built from the *restriction classes* *geonames:featureCode=S.SCH* and *geonames:countryCode=US*, can be defined by the intersection of the two sets and forms a concept extensionally described by the set of schools in the US in *GeoNames*.

Our algorithm aligns *restriction classes* from two sources, using an extensional technique, as follows. A pre-processing step first performs an inner-join

on the two sources to be aligned based on an instance equivalence property like *owl:sameAs*. As inverse functional properties can only result in *restriction classes* with a single instance belonging to it, the pre-processing step eliminates them. The crux of the algorithm uses a top-down tree exploration of the space of alignment hypotheses. At the topmost level, a seed hypothesis is generated by aligning a *restriction class* with one *property-value pair* from the first source with another *restriction class* with one *property-value pair* from the second source. At each level in the search space, a new *restriction class* is formed from one *restriction class* of one of the sources by adding another *property-value pair* constraint on that *restriction class*. A new alignment hypothesis is thus constructed from the new *restriction class* and the *restriction class* from the other source. Each alignment hypothesis is tested for set containment relations between the intersection set of the *restriction classes* from both sources. This is done with the help of two scoring functions - P & R . If r_1 and r_2 are the two *restriction classes* in the alignment hypothesis, we first define $\text{Img}(r_1)$ as the set of instances in the second source that instances of r_1 are linked to. We then define P as $\frac{|\text{Img}(r_1) \cap r_2|}{|r_2|}$, and R as $\frac{|\text{Img}(r_1) \cap r_2|}{|\text{Img}(r_1)|}$. We mark the relation of the alignment hypothesis as either i) equivalent ($P = 1, R = 1$), ii) subset, with the *restriction class* from the first source as extensionally subsuming the *restriction class* from second source ($R = 1$), iii) subset, with *restriction class* from second source extensionally subsuming the *restriction class* from first source ($P = 1$) or iv) no relation between the two *restriction classes*. To compensate for missing and misaligned instances, we relax our subset scores by defining P' and R' that reduce the required fraction of support to be greater than 0.9 instead of equal to 1. For an optimal exploration of the search tree, we employ certain pruning mechanisms that include i) using ordered exploration to avoid exploring a node twice, ii) pruning a node if the intersection set of the *restriction classes* of the hypothesis has size less than a minimum support size (we used 10 in our experiments), iii) pruning a node if the added *restriction class* does not change the set of instances, etc. After the brute-force exploration of the search space of alignment hypotheses, we use a post-processing step on the results generated, which removes redundant assertions by virtue of set containment of instances of two hypotheses where one is the immediate parent of the other in the search tree.

At the end of the above three steps of processing, the algorithm was able to find equivalent relations between *restriction classes* from two sources as well as subset relations in either direction. As this algorithm was not specific to any particular domain, we explored candidate sources for alignments in three domains: Geospatial, Genetics and Zoology. In these three domains, our algorithm found alignments of 5 pairs of sources. For example, we were able to find alignments between *GeoNames* and *DBpedia* in the Geospatial domain. One such alignment was the equivalent relation between $\{\text{geonames:countryCode}=ES\}$ and $\{\text{dbpedia:country}=Spain\}$ (i.e. correctly aligning the concepts for the country Spain). We also found subset relations like $\{\text{geonames:featureCode}=S.SCH\}$ subset of $\text{rdf:type}=\text{dbpedia:EducationalInstitution}$. More such results are described in [7].

Limitations The approach above produced a large number of equivalent alignments that gave an exact mapping between the two *restriction classes* from the two sources. It also, however, produced a large number of subset relations that were not as useful. This was mainly because the subset relations, by themselves, did not contribute to a useful equivalence alignment between two classes. In all, in the *GeoNames* and *DBpedia* alignment, there were 1647 subset relations found. Though it is understandable that in many cases there might never exist an exact equivalence between two *restriction classes*, because they were auto-generated using *property-value pairs*, we decided to look for additional useful alignments, if any, that these subset relations might be able to provide us. For example, in the *GeoNames* and *DBpedia* alignment, we found that $\{geonames:featureCode=S.SCH\}$, $\{geonames:featureCode=S.SCHC\}$ and $\{geonames:featureCode=S.UNIV\}$ (i.e. Schools, Colleges and Universities from *GeoNames*) are all subsets of $\{rdf:type=dbpedia:EducationalInstitution\}$. Taken individually, though each of these alignments are correct and insightful, they are not particularly useful in understanding the relationships between *GeoNames* and *DBpedia*. Taken together, however, we found that the union of these three *restriction classes* completely define $rdf:type=dbpedia:EducationalInstitution$. The limitation of our approach was in the expressivity of our *restriction classes*. Though it included *restriction classes* containing single *property-value pairs* and the conjunction operator on those *restriction classes*, it did not include a disjunction operator and hence was unable to make use of the subset relations.

2.2 Identifying spatial concept coverings

As explained above, we were able to identify a pattern where a group of *restriction classes* from one source were aligned as subsets of a common concept from the other source. By using these alignments as hints, we were able to construct the union of the smaller *restriction classes* and detect if the union was able to define the larger class entirely. The following section describes this method in detail. In those cases where we are not able to define the larger class entirely, our approach is also able to find and explain the missing instances (*outliers*).

Mapping a *restriction class* from one source with a union of smaller *restriction classes* from the other source Since the problem of finding alignments with conjunctions and disjunctions of *property-value pairs* of *restriction classes* is combinatorial in nature, we focus only on subset relations where both *restriction classes* have a single *property-value pair* and where one is a subset of the other. This helps us find the simplest definitions of concepts and also makes the problem tractable. Alignments generated by our previous work that satisfy the single *property-value pair* constraint are first grouped according to the subsuming *restriction classes*. We then identify a strategy for selecting the smaller *restriction classes* from within such a group to form the union that best describes the larger *restriction class*. Since *restriction classes* are constructed by forming a set of instances that have one of the properties restricted to a single value, aggregating *restriction classes* from the group according to their

properties builds a more intuitive definition of the union. We can now define the disjunction operator that constructs the union concept from the smaller *restriction classes* in these sub-groups. The disjunction operator is defined for *restriction classes*, such that *i)* the concept formed by the disjunction of the *restriction classes* represents the union of their set of instances, *ii)* each of the *restriction classes* that are aggregated contain only a single *property-value pair* and *iii)* the property is the same for all those *property-value pairs*. We then try to find the alignment between the larger common *restriction class* and a set of *restriction classes* from the other source that are aggregated by the disjunction operator by using an extensional approach similar to our previous paper. We call such an alignment as *union alignment*.

We first build candidates for aggregation using the results from our previous algorithm as hints. We group alignments by the larger common *restriction class*. Grouping the subset relations is trivial. Equivalence relationships are subsets in both directions and thus are easily integrated into the groups. For each alignment, $\{p_1=v_1\}$ is the r_1 part and $\{p_2=v_2\}$ forms the r_2 part (each with a single *property-value pair*) as explained in the previous section. Sub-groups are formed by aggregating according to the property of the *property-value pairs* of the smaller *restriction classes*. Such a sub-group is identified by $\{Property\ of\ the\ larger\ restriction\ class(p_1),\ Value\ of\ the\ larger\ restriction\ class(v_1),\ property\ of\ the\ smaller\ restriction\ classes(p_2)\}$. Values of the different smaller *restriction classes* can be denoted by a list $List(v_2s)$. The disjunction of the smaller *restriction classes* creates a set of instances that extensionally identifies the union concept. We can now either confirm or refute the hypothesis that the larger *restriction class* is equivalent to the union concept. We can do this by using a scoring mechanism similar to the use of P & R in our previous paper. Using the same terminology, U_A is defined as the set of disjunctive instances (i.e. $Union(Img(r_1) \cap r_2)$), U_L is defined as the set of instances of the larger class taken by itself (i.e. $Img(r_1)$) and U_S is defined as the set of instances that is the union of individual smaller *restriction classes* (i.e. $Union(r_2)$). The scoring mechanism defines P_U as $\frac{U_A}{U_S}$ and R_U as $\frac{U_A}{U_L}$. P'_U & R'_U are defined as fractions with relaxed scoring assumptions similar to P' & R' from our previous paper.

For example, our previous algorithm finds that $\{geonames:featureCode = S.SCH\}$, $\{geonames:featureCode = S.SCHC\}$, $\{geonames:featureCode = S.UNIV\}$ are subsets of $\{rdf:type=dbpedia:EducationalInstitution\}$. In this case, the sub-group can be identified as $\{rdf:type, dbpedia:EducationalInstitution, geonames:featureCode\}$ and list as $(S.SCH, S.SCHC, S.UNIV)$. As can be seen in the Venn diagram of Figure 1, U_L is the *restriction class* $Img(\{rdf:type = dbpedia:EducationalInstitution\})$, U_S is $\{geonames:featureCode = S.SCH\} \cup \{geonames:featureCode = S.SCHC\} \cup \{geonames:featureCode = S.UNIV\}$ and U_A is:

$$\begin{aligned} & \{Img(\{rdf:type = dbpedia:EducationalInstitution\}) \cap \{geonames:featureCode \\ & = S.SCH\}\} \cup \{Img(\{rdf:type = dbpedia:EducationalInstitution\}) \cap \{geonames:featureCode \\ & = S.SCHC\}\} \cup \{Img(\{rdf:type = dbpedia:EducationalInstitution\}) \cap \{geonames:featureCode \\ & = S.UNIV\}\} \end{aligned}$$

Ideally, for an exact equivalence alignment, P'_U & R'_U should both be 1.0, if the larger *restriction class* covers the union of the smaller *restriction classes* completely and vice-versa. However, similar to the relaxed score assumption from our previous paper to accommodate errors in the dataset, we consider it a complete coverage when the score is greater than a relaxed score of 0.9. (i.e. the *union alignment* is considered to be equivalent if $P'_U > 0.9$ & $R'_U > 0.9$). Due to the minimum support score constraint for subsets from our previous paper, we are assured that $\frac{U^A}{U^S}$ i.e. P'_U is always going to be greater than 0.9.⁴ Thus, we can say that a *union alignment* is equivalent if $R'_U > 0.9$. With the educational institutions example, R'_U for the alignment of *dbpedia:EducationalInstitution* to the union of *S.SCH*, *S.SCHC* & *S.UNIV* is 0.98. We can thus confirm the hypothesis and consider this *union alignment* equivalent. The scores for other *union alignments* found are described in the results section.

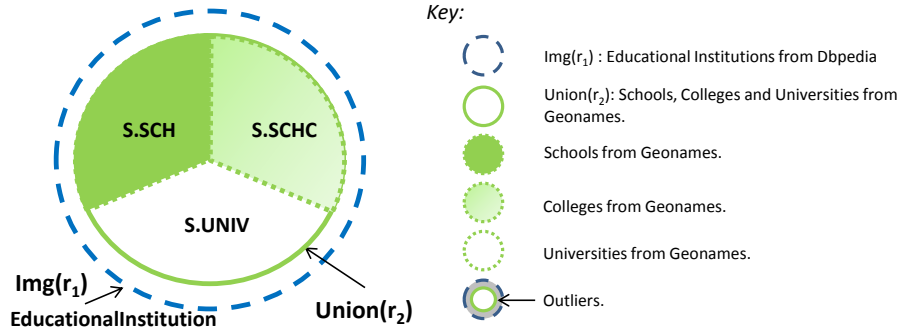


Fig. 1. Spatial covering of Educational Institutions from *DBpedia*

Using mappings to identify outliers As mentioned above, the score for the alignment of $\{rdf:type = dbpedia:EducationalInstitution\}$ to the union of $\{S.SCH, S.SCHC \text{ \& } S.UNIV\}$ is approximately 0.98. For $\{rdf:type = dbpedia:EducationalInstitution\}$, 396 instances out of the 403 Educational Institutions were accounted for as having their *geonames:featureCode* as one of *S.SCH*, *S.SCHC* or *S.UNIV* to give this score. An interesting question to pose then is, how are the remaining 2% of the *dbpedia:EducationalInstitutions* (i.e. 7 instances) classified in *GeoNames*?

While calculating the disjuncted *restriction classes*, we also keep track of other instances with the same $\{p_1, v_1, p_2\}$ but not previously considered as subsets. These had been pruned in the exploration stage as they either had a size of less than the minimum support size constraint of ten instances or had P' less than 0.9. For the first type of *restriction classes*, those with low support size but yet having P' greater than 0.9 are now re-classified as subsets. The

⁴ It should also be noted that each of the smaller subsets also satisfy the minimum support size of 10 instances.

re-classification of the relation as a subset can now be justified due to increased evidence in suggesting subsumption as other values for the same property are also aligned as subsets of the larger *restriction class* from the first source.

The second type of *restriction classes* that had P' less than 0.9 along with the ones that were not re-classified above (i.e. with less than 10 instances and P' less than 0.9) form the *outliers*. For example, as mentioned before, schools, colleges and universities from *GeoNames* make up 396 out of 404 Educational Institutions from *DBpedia*. From the other eight instances, 7 have their feature codes as either S.BLDG (3 buildings), S.EST (1 establishment), S.HSP (1 hospital), S.LIBR (1 library) or S.MUS (1 museum). The eighth instance does not have a *geonames:featureCode* property asserted. The P' score of these *restriction classes* is less than 0.9. One of the instances classified as *dbpedia:EducationalInstitution* in *DBpedia* is linked to an instance in *GeoNames* that has *geonames:featureCode* as ‘S.HSP’.⁵ There are 31 instances in $\{\textit{geonames:featureCode}=\textit{S.HSP}\}$, however, and because this *restriction class* does not meet the relaxed subset score threshold, it cannot be considered in the union of *restriction classes*. Another example of outliers was found in the $\{\textit{dbpedia:country} = \textit{Spain} \equiv \textit{geonames:countryCode} = \textit{ES}\}$ alignment. This equality was found using the relaxed subset assumption, where 3917 of the 3918 instances of *dbpedia:country=Spain* were accounted for as having *geonames:countryCode=ES*, resulting in a subset score of 0.9997. The one instance not having country code ES was actually classified as having country code IT (Italy). This single instance needs to be inspected further and it needs to be determined if the *owl:sameAs* link is correct. It is evident from the above examples that the outliers help in understanding the nature of the sources more explicitly, showing why the alignments failed to completely describe the larger *restriction class*. These, along with a few other examples, are described in detail in the next section.

3 Experimental Results

From the approach described in Section 2.2, we were able to get a total of 752 union alignments for the *GeoNames-DBpedia* alignment and 5843 for the *LinkedGeoData-DBpedia* alignment. From the 752 in *GeoNames-DBpedia*, 318 are such that the larger *restriction class* is from *DBpedia*, while the other 434 have the larger *restriction class* from *GeoNames*. Similarly, 3097 from the 5843 *union alignments* in *LinkedGeoData-DBpedia* have the larger *restriction class* from *DBpedia*, while the other 2746 have the larger *restriction class* from *GeoNames*. Tables 1, 2, 3, & 4 list a few interesting examples of these *union alignments* between *GeoNames-DBpedia* and *LinkedGeoData-DBpedia* (in either direction), which we describe here. The tables are organized as follows. Column 2 describes the sub-group, i.e. (p_1, v_1, p_2) . Column 3 contains the list of the value part of the *property-value pairs* in the *restriction classes* of the smaller sets (i.e. $\text{List}(v_2)$). The score of the union is noted in column 4 ($R'_U = \frac{|U_A|}{|U_I|}$) followed by $|U_A|$ and

⁵ Intuitively, it would make sense to the reader that this instance might perhaps be a hospital of a medical school.

$|U_L|$ in columns 5 and 6. Column 7 describes the outliers, i.e. values of v_2 that form *restriction classes* that aren't direct subsets of the larger *restriction class*. Each of these values also has a fraction with the number of instances that do belong to the larger *restriction class* of the total number of instances of the *restriction class* (or $\frac{|Img(r_1)|}{|r_2|}$). It can be seen that the fraction is less than our relaxed subset score. If the value of this fraction was greater than the relaxed subset score (i.e. 0.9), the set would have been included in column 3 instead. The last column mentions how many of the total U_L instances we were able to explain using U_A and the outliers. For example, the *union alignment* of #1, is the Educational Institution example described before. It shows how educational institutions from *DBpedia* can be explained by schools, colleges and universities in *GeoNames*. Column 4, 5 and 6 explain the alignment score R'_U (0.98), the size U_A (396) and the size of U_L (404). The seven of the eight outliers found (S.BLDG, S.EST, S.LIBR, S.MUS, S.HSP) are mentioned along with their P' fractions in column 7.

We also found some other interesting alignments. #2 shows the details of the Spain example mentioned briefly in Section 2.2. #3 shows a union alignment that aligns smaller sets or parts from *GeoNames* to a complete set. The region of Basse-Normandie in France is made up of three departments. The *restriction classes* of these three regions are constrained by the *geonames:parentADM2* property. #4 shows that Airports and Airbases make up 99% of the airports in *DBpedia*. From its outliers, one might argue that Airfields (S.AIRF) should also be included, but it was not as its P' score was lower than the threshold. Outliers also show that there is a Hill in *geonames* that has been classified as an airport. Even though this instance may be an airport in the hills, ontologically it doesn't make sense that a hill can be an airport. A similar case is observed in #8 where we find that there is at least one water tower in *LinkedGeoData* that is aligned with an Educational institution in *DBpedia*.

The *union alignment* #5 should have been as straightforward as alignment #2. Our approach was able to detect a pattern, however, that might have been overlooked after looking at individual instances. Netherlands from *GeoNames*, for example, should be aligned with the country Netherlands from *DBpedia*. However we have possible alias names, such as *The Netherlands and Kingdom of Netherlands*, as well a possible linkage error to *Flag of the Netherlands.svg* generated while importing Wikipedia data into *DBpedia* (the error seems systematic, see Jordan in #6).

Alignment #7 was able to explain 8 of the 10 license plate codes in the state (bundesland) of Saarland⁶. The ones that it missed were Ottweiler (OTW) and the police vehicle codes (SAL). Since the vehicle code SAL is not associated with any populated places in Saarland, it is quite possible that it does not get mentioned in *LinkedGeoData*. Our approach thus provides a deeper insight into the nature of the sources. #9 tries to find the composition of the state of New Jersey. 100% of the instances in New Jersey from *LinkedGeoData* can be accounted

⁶ <http://www.europlates.com/publish/euro-plate-info/german-city-codes>

for in the 9 counties. New Jersey actually has 21 counties⁷. This suggests that instances in New Jersey in *LinkedGeoData* that are linked to *DBpedia* are not a complete representation resulting in an equivalent alignment. The quality of the results generated by our extensional approach are tied to the quality of the instances in the dataset. We find, however, that such alignments, even though they might be partially incorrect, give an accurate representation of the actual instances in the dataset and highlight the practical quality of the links in the Web of Linked Data.⁸ Finally, alignment #10 describes how the concept Waterways in *LinkedGeoData* can be defined as the union concept of Streams and Rivers in *DBpedia*. The complete set of alignments discovered by our algorithm are available on our group page.⁹

4 Related Work

Ontology alignment has been a well explored area of research since the early days of ontologies. It has received renewed interest in recent years with the rise of the Semantic Web. Euzenat & Shvaiko [3] provide a comprehensive discussion on Ontology Matching approaches. A closely related area of study to ontology alignment is schema matching. Bernstein et al. [1] summarize the developments in this field in the past ten years. Though most work done in the Web of Linked Data is on linking instances across different sources, an increasing number of authors have looked into aligning the sources ontologies in the past couple of years. Jain et al. [4] describe the BLOOMS approach which uses a central forest of concepts derived from topics in Wikipedia. An update to this is the BLOOMS+ approach [5] that aligns Linked Open Data ontologies with an upper-level ontology called Proton. Though we employ a simple set subsumption technique to identifying alignments, our use of *restriction classes* is able to find a large set of alignments in cases like aligning *GeoNames* with *DBpedia* or Proton, while BLOOMS & BLOOMS+ are unable to find alignments because of the small number of classes in *GeoNames* that have vague declarations. Cruz et al. [2] describe a dynamic ontology mapping approach called *AgreementMaker* that uses similarity measures along with a mediator ontology to find mappings using the labels of the classes. Building ontologies of Linked Data sources using a statistical method has also been described in Völker et al. [8]. This work induces schemas for RDF data sources by generating OWL 2 axioms using intermediate associativity table of instances and concepts (called *transaction datasets*) and mining associativity rules from it.

⁷ http://en.wikipedia.org/wiki/List_of_counties_in_New_Jersey

⁸ In [7] we compared the extensional versus intensional perspective on ontology alignment. In a nutshell, the extensional alignment gives a precise characterization of the current relationship between the data in the sources, regardless of the intended meaning of the concept definitions. For example, a source may define instances as universities, but linkage can show that it only contains American universities.

⁹ <http://www.isi.edu/integration/data/UnionAlignments>

Table 1. Example alignments from the *GeoNames* and *DBpedia* datasets, with larger sets from *DBpedia* and smaller sets from *GeoNames*

#	Sub-group $\{p_1, v_1, p_2\}$	List(v_2)	$R'_U = \frac{ U_A }{ U_L }$	U _A	U _L	Outliers	# Explained Instances
1	{rdf:type, dbpedia:EducationalInstitution, geonames:featureCode}	S.SCH, S.SCHC, S.SUNIV	0.9801	396	404	S.BLDG (3/122), S.EST (1/13), S.LIBR (1/7), S.HSP (1/31), S.MUS (1/43)	403
2	{dbpedia:country, dbpedia:Spain, geonames:countryCode}	ES	0.9997	3917	3918	IT (1/7635)	3918
3	{dbpedia:region, dbpedia:Basse-Normandie, geonames:parentADM2}	geonames:2989247, geonames:2996268, geonames:3029094	1.0	754	754		754
4	{rdf:type, dbpedia:Airport, geonames:featureCode}	S.AIRB, S.AIRP	0.9924	1981	1996	S.AIRF (9/22), S.FRMT (1/5), S.SCH (1/404), S.STNB (2/5) S.STNM (1/36), T.HLL (1/61)	1996

Table 2. Example alignments from the *DBpedia* and *GeoNames* datasets, with larger sets from *GeoNames* and smaller sets from *DBpedia*

#	Sub-group $\{p_1, v_1, p_2\}$	List(v_2)	$R'_U = \frac{ U_A }{ U_L }$	U _A	U _L	Outliers	# Explained Instances
5	{geonames:countryCode, NL, dbpedia:country}	dbpedia:Netherlands, dbpedia:The_Netherlands, dbpedia:Flag_of_the _Netherlands.svg	0.9802	1939	1978	dbpedia:Kingdom_of _the_Netherlands	1940
6	{geonames:countryCode, JO, dbpedia:country}	dbpedia:Jordan dbpedia:Flag_of_Jordan.svg	0.95	19	20		20

Table 3. Example alignments from the *LinkedGeoData* and *DBpedia* datasets, with larger sets from *DBpedia* and smaller sets from *LinkedGeoData*

#	Sub-group $\{p_1, v_1, p_2\}$	List(v_2)	$R'_U = \frac{ U_A }{ U_T }$	$ U_A $	$ U_L $	Outliers	# Explained Instances
7	{dbpedia:bundesland, Saarland, lgd:OpenGeoDBLicensePlateNumber}	HOM, IGB, MZG, NK, SB, SLS, VK, WND	0.93	46	49		46
8	{rdf:type, dbpedia:EducationalInstitution, rdf:type}	lgd:Amentiy, lgd:K2543, lgd:School, lgd:University, lgd:WaterTower	0.9901	2609	2610		2609

Table 4. Example alignments from the *LinkedGeoData* and *DBpedia* datasets, with larger sets from *LinkedGeoData* and smaller sets from *DBpedia*

#	Sub-group $\{p_1, v_1, p_2\}$	List(v_2)	$R'_U = \frac{ U_A }{ U_L }$	$ U_A $	$ U_L $	Outliers	# Explained Instances
9	{lgd:gnissST_alpha, NJ, dbpedia:subdivisionName}	Atlantic, Burlington, Cape May, Hudson, Hunterdon, Monmouth, New Jersey, Ocean, Passaic	1.0	214	214		214
10	{rdf:type, lgd:Waterway, rdf:type}	dbpedia:Stream, dbpedia:River	0.97	33	34	dbpedia:Place(1/94989)	34

5 Conclusions and Future Work

We described an approach to identifying *union alignments* in geospatial data sources on the Web of Linked Data. By extending our definition of *restriction classes* with the disjunction operator, we were able to find alignments of union concepts from one source to larger concepts from the other source. Our approach produced *union alignments* as results that found that concepts at different levels in the ontologies of two sources can be mapped even when there was no direct equivalence. We were also able to find outliers that enable us to identify inconsistencies in the instances that are linked by looking at the alignment pattern. The results provide deeper insight into the nature of the alignments of Geospatial Linked Data.

Though the scope of this paper is the geospatial domain, our algorithm can be used in other domains as well. Our next step is to explore other domains like zoology and genetics for *union alignments*. Other possible future work is in the mapping and understanding of the properties in the sources. Our preliminary findings show that the results of this paper can be used to find patterns in the properties. For example, the *countryCode* property in *GeoNames* is closely associated with the *country* property in *DBpedia*, though their ranges are not exactly equal. We believe that an in-depth analysis of the alignment of ontologies of sources is warranted with the recent rise in the links in the Linked Data cloud. This is an extremely important step for the grand Semantic Web vision.

Acknowledgements

This research is based upon work supported in part by the National Science Foundation under award number IIS-1117913.

References

1. Bernstein, P., Madhavan, J., Rahm, E.: Generic schema matching, ten years later. *Proceedings of the VLDB Endowment* 4(11) (2011)
2. Cruz, I., Palmonari, M., Caimi, F., Stroe, C.: Towards on the go matching of linked open data ontologies. In: *Workshop on Discovering Meaning On The Go in Large Heterogeneous Data*. p. 37 (2011)
3. Euzenat, J., Shvaiko, P.: *Ontology matching*. Springer-Verlag (2007)
4. Jain, P., Hitzler, P., Sheth, A., Verma, K., Yeh, P.: Ontology alignment for linked open data. *The Semantic Web—ISWC 2010* pp. 402–417 (2010)
5. Jain, P., Yeh, P., Verma, K., Vasquez, R., Damova, M., Hitzler, P., Sheth, A.: Contextual ontology alignment of lod with an upper ontology: A case study with proton. *The Semantic Web: Research and Applications* pp. 80–92 (2011)
6. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Aligning geospatial ontologies on the linked data web. In: *Proceedings of the GIScience Workshop on Linked Spatiotemporal Data*. Zurich, Switzerland (2010)
7. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Linking and building ontologies of linked data. In: *Proceedings of the 9th International Semantic Web Conference (ISWC 2010)*. Shanghai, China (2010)
8. Völker, J., Niepert, M.: Statistical schema induction. *The Semantic Web: Research and Applications* pp. 124–138 (2011)

Discovering Places of Interest through Direct and Indirect Associations in Heterogeneous Sources — The TravelSampo System

Eetu Mäkelä, Aleksi Lindblad, Jari Väättäinen, Rami Alatalo, Osma Suominen, and Eero Hyvönen

Semantic Computing Research Group (SeCo),
Aalto University and University of Helsinki, Finland
first.last@aalto.fi, <http://www.seco.tkk.fi/>

Abstract. Linked data related to places offers a superior collection of information to base search and recommendation functionality on in eTourism visit planning as well as location-aware mobile applications. Besides places interesting in themselves, through linked data it is possible to discover places interesting only through association, such as being the venue for a concert by an artist with an interesting genre. However, in order to harness this collective data source, challenges relating to data heterogeneity, quality, scale, and indexing and querying complexity must be resolved. In this paper, the TravelSampo visit planning and mobile application is presented, which tackles these issues. Using the system, queries describing both simple and complex interests can be run over some 17 million places of interest from over 20 vastly heterogeneous sources.

1 Introduction

Location-aware mobile devices are becoming increasingly commonplace. This has led to a multitude of mobile applications to search for e.g. events, places of interest or services near the user's physical location. On the other hand, many eTourism web applications also now allow people to design travel plans online, picking sites to visit and exporting visit lists to their phone's navigator software.

The TravelSampo project is an attempt to harness linked data as a source of material for an application to help travellers find content relevant to them, both in planning as well as during a trip. As compared to existing non-linked data solutions as well as similar linked data systems such as DPBedia Mobile [2], mSpace Mobile [13] and SmartMuseum [11], it tries to improve upon the state of the art in being able to integrate both massively more heterogeneous material, as well as to provide more intelligent services on top of it.

Particularly, the TravelSampo system takes into account that there are multiple ways in which a location may be of interest to a user. First, the place itself may have some quality of interest, such as being a church, or being a church in the gothic style. On the other hand, the place may also be of interest only

through a more or less direct association, such as being the venue for an interesting event or having been the birthplace of a painter with a style of interest. In addition, a place may be of interest by virtue of the services offered there, such as Internet access.

This variety of ways in which data can be both interesting as well as encoded necessitates a flexible architecture for querying locations of interest. The strength of this is that the application should ultimately be able to cater to a wide variety of interests, from people looking for nearby museums through music fans interested in concerts by Norwegian heavy metal bands to freegans searching for dumpsters near big supermarkets without nearby surveillance cameras. At the same time, this scale of heterogeneity causes severe problems in both integrating the content as well as providing efficient and intelligent search and recommendation services and user interfaces on top of it.

In the current demonstration system of TravelSampo, some 17 million locations have been loaded into the system, integrating information from over 20 vastly different datasets of places, places of interest, and content making places interesting through association, such as fiction taking place in real-world locations, or the birthplaces of famous artists. Included are for example the huge datasets of DBPedia [3] and the LinkedGeoData.org [1] version of OpenStreetMap, but also fast-changing, dynamically converted datasets such as four different sources for current events and exhibitions in Finland.

In this paper, the TravelSampo application is presented first through its user interface. After that, the challenges faced and solutions developed in integrating, mapping and making usable the disparate heterogeneous data sources are described. Finally described are the indexing and querying interfaces created that make possible the complex queries required to provide the advanced functionality of the TravelSampo application.

2 The TravelSampo Application

The TravelSampo application has two distinct interfaces. The web interface is used to plan the trip beforehand and to examine and share the trip afterwards. The mobile interface is used during the trip to find the destinations and to get more information about them.

2.1 The Visit Planning Interface

A typical user would be someone who is going on a trip to a new city. Before the trip he can use the planning interface to find out what kind of cultural destinations and events the city offers during his trip. The destinations can be searched with different levels of complexity. In the simplest case our user is interested in churches, which are places themselves. Our user is also interested in wall climbing, which is a service located in a place. And finally he's interested in modern art, which is a topic of an exhibition held in a place. The application can handle all these searches.

The user has found a couple of churches, a sport center with wall climbing and an exhibition of modern art and now he can save them to his destination shelf which can be accessed during the trip in the mobile interface. The filters used to find these destinations can also be saved to be used on other trips either in the planning interface or the mobile interface. The visit planning interface is not yet implemented but the shelves and filters can be produced and used in the mobile interface.

2.2 The Mobile Interface

During his trip the user can use the mobile application to find the previously saved destinations. The starting page of the mobile interface, depicted in figure 1 has a list of nearby destinations (1.A) with their types, distance to them and interesting instances associated to them which can be filtered using the filter menu (1.B). This menu contains the users destination shelves that are relevant in his current location as well as his personal filters and some predefined ones.

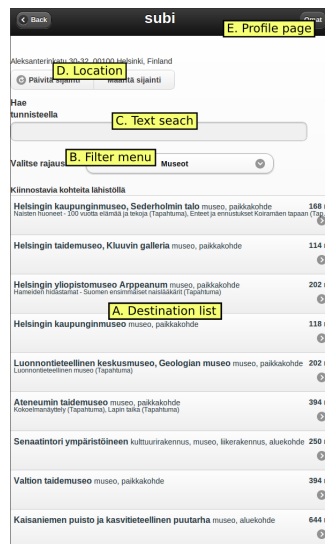


Fig. 1. Main screen of the mobile interface

The user can now use the shelf containing his destinations to access their pages and get a route map to the destination as well as information about it, links to associated instances and a button to mark the destination for future reference. As the user is in the destination page in the context of the destination shelf there are also buttons to browse through other destinations of the shelf.

If the user finds he has more time than he expected he can go back to the main page and use his saved filters to find for example more churches or in-

stances related to modern art. This can also be used to quickly find interesting destinations or useful services on the vicinity without prior planning.

There will also be a possibility to use free text search to find destinations (1.C). The location area (1.D) shows the user's current address and allows him to reload it using the geolocation capabilities of the mobile device or set it manually. On the top right corner there is a link to the user's profile page or a login page if the user hasn't already logged in (1.E).

3 Data Sources and Modeling

As already stated, the TravelSampo data repository contains some 17 million locations sourced from over 20 vastly different datasets. These sources are described in table 1, stating the general type of data sourced from each provider, the number of location and reference items in the data sources, as well as some example content types in the data thus gained. As can be seen from the table, the TravelSampo system contains a truly heterogeneous mix of data of different types, sources and schemas.

Particularly interesting in analyzing the data sources integrated is the fact that the boundaries between geographic place names, places of interest and services are not crisp. For example, the general place name registries contain not only hills and swamps, but also areas of sporting services, churches and abandoned police posts, while on the other hand the locations in the Espoo nature site point of interest database are precisely hills and swamps of special interest. Further, the Helsinki City service database for example contains both office entities such as childcare services as well as located services such as swimming pools, which also feature in point of interest databases.

Based on this observation, the TravelSampo system was designed not to discern between these sources for locations at all, but to treat all location information as equal. This puts the burden of discovering whether a location is a point of interest to someone on the information gathered for that location. Among direct indicators of interest, primary among them is the type of the location. To be able to use this across all the place data in the TravelSampo system, it was decided to attempt to build a single unified place and place of interest type ontology POIO from all the type ontologies used in the various data sources.

This was done semiautomatically so that first all place type labels were compared automatically, already yielding several hundred equivalency mappings. Then, these mappings were examined by hand, and a large number of spurious mappings rejected, while an equally large amount of new mappings and subclass relations were curated, until all place types could be found under a single root. Table 2 relates the numbers of distinct place types in the constituents of the POIO ontology, as well as the size of the final ontology. In total, of the 2499 concepts in the final ontology, 62% (1539) were found to be shared between at least two source ontologies.

Besides differing in place types, the datasets also differed vastly in terms of modeling and level of content description. For example, the RKY database of

Table 1. TravelSampo Data Sources

Type	Source	Size	Example types	Description
Places of Interest	LinkedGeoData	~3.1 million	Statue, tunnel, crossing, school, ruins, bench	Entities with non-amenity types in the RDF conversion [1] of the open mapping project OpenStreetMap
	DBPedia	~432,000	Site of earthquake, tv-mast, bridge	RDF conversion [3] of structured information in Wikipedia
	MJREKI	25,343	Hill fort, rock tomb, place of village	Finnish ancient monuments
	RKY	1,850	Manor, powerplant, cemetery, rectory	Finnish nationally cultural-historically significant milieu
	Espoon rakenmukset	80	Villa, rectory, school, jugend building	Cultural-historically significant buildings in the city of Espoo
Services	Espoon luontokohteet	282	Swamp, glacial erratic, meadow, nature reserve	Nature sites in the city of Espoo
	LinkedGeoData	~3.7 million	Pub, pharmacy, restaurant	Entities tagged as amenities in the RDF conversion [1] of the global open mapping project OpenStreetMap
	Pääkaupunkiseudun palvelukartta	2,929←4350	Swimming pool, daycare, museum, child protection services	Public services provided by the cities in the Helsinki metropolitan area
	EvenemaX	~1480←4300	Pub quiz, concert, exhibit	Finnish commercial cultural event aggregator
	Turku 2011	~305←3400	Concert, exhibition, circus	Events included in Turku's year as European culture capital
Events	Museot.fi	~80←140	Exhibition	Current exhibitions at Finnish museums
	Agricola	~60←70	Lecture, seminar, exhibition	Event-calendar of the Agricola network of Finnish historians
	Tarinoiden Helsinki	1629←4091	Book, music, film, fact	Fiction and facts relating to places in Helsinki
Place-Related Content	CultureSampo	~20,000←600,000	Poem, photograph, painting, video, skill	An integrated portal of some 30 different content types from some 20 different institutions [6]
	DBPedia	~432,000←3.6 million	Person, organization, invention, event	Every conceivable notable aspect of human existence that linked to a place
	SUO	~800,000	Hill, swamp, meadow, glacial erratic, area of sporting services	Finnish registry of place names
Places	SAPO	1,261	Administrative area	A spatio-temporal ontology of historical Finnish counties [7]
	GNS	~4.2 million	Oil field, ramp, hill, abandoned police post, church, glacier	Geonet Names Server, a US government place database
Total	GeoNames	~6.9 million	Oil field, ramp, hill, abandoned police post, church, glacier	Open database of geographical names, using the same feature codes as GNS
	TGN	~895,000	Aqueduct, mausoleum, sinkhole, earldom, Nicaraguan center, dynasty	Getty thesaurus of historic and current locations
	Karelian places	37,476	Village, house	Historical places in the Karelia region of Finland and Russia
Total		~17 million←300 million		

Table 2. Number of distinct place types in the constituents of the POIO ontology

Name	Size
OpenStreetMap	1506
TGN	1737
GNS/Geonames	648
SUO	648+142 ^a
POIO	2499

^a GNS types+additions

culturally significant milieu contains areas of cultural interest defined as polygons. However, most of these areas are actually collections of multiple points of interest, which are not modeled separately at all. On the other hand, most of the other databases listed do not model areas at all, but only provide centerpoints for even large features. Even worse, it is often difficult to automatically deduce when a location actually refers to a notable mass of land, such as an amusement park, instead of a small point, such as a statue.

As regards services, in the vast majority of data and data sources used in the TravelSampo project, the services described are those that can be described indirectly through place type, such as being a restaurant or a pharmacy. However, in the Helsinki City data source the services offered at a particular location are described separately, for example noting if a particular library offers Internet access, has a scanner or loans AV equipment.

In the case of the TravelSampo system, particularly as it was making use of many automatically converted and dynamically updating data sources, it was decided that these heterogeneities in content modeling could not be unified, at least without losing information or the expressivity of the original data, but would have to be resolved at the query construction level. Fortunately, it seems that some quite general mapping rules could be made to facilitate this, for example linking services and events described as separate resources to the places they are provided or help in, or linking a culture site with no direct description to the compound description of the larger area it was found in.

As can be gleamed from the table listing the TravelSampo data sources, events such as concerts and exhibitions were identified as a particularly interesting non-direct element signaling a place of interest. That is, particularly for cultural applications, often one is for example not interested in a museum per se, but in the exhibitions that are on display in that museum at present.

Now, events are particularly dynamic sort of data. At the time of creating the TravelSampo system, there were no sources for current and coming event information in RDF. However, there were multiple sources from which such could be gleamed in other formats, such as comma separated values, JSON or RSS. The event content for TravelSampo thus comes from a converter pipeline that is capable of being run at regular intervals, or by request. This pipeline is actually

a more general one, called Harava [12], created in the FinnONTO project¹ [5] as a Semantic Web infrastructure tool by which data can be harvested, converted, enriched and validated to be published as quality Linked Data for anyone to use².

A major problem in the event data sources to be used in the project however was that none of them contained any machine-processable descriptions of the topics related to the event, such as the style of an exhibition or the artist. This problem was also evident in some of the point of interest data sources. For example, in the OpenStreetMap data on Helsinki, there is an object of type “memorial”, which only in its textual description says that 1) it is actually a statue and 2) it depicts the runner Paavo Nurmi.

To overcome these limitations, automated information extraction services were integrated into the TravelSampo architecture and the Harava pipeline, which could then extract relevant entities such as people, organizations and places as well as general content keywords from the textual descriptions of the events and other data items.

Because the information extraction tools were configured to use the whole vast TravelSampo database as a source for keywords, they are usually able to pick up a huge number of potentially related instances. The problem then became more of filtering these potential instances to the most important and factual ones. Fortunately, here the project could make use of the open source Maui information extraction tool [10], which has been previously shown to be human-competitive in selecting primary topic keywords from text.

4 Place and Event Instance Mapping

After getting all the different data sources together, one finds a large amount of overlap between them. Besides the same places occurring in many of the place databases, also events typically are entered in more than one of our dynamically updated event sources.

The indexing system used in the TravelSampo project is capable of inferring and resolving ontology language equivalency statements transparently. Thus, mapping between the different datasources does not need to happen at indexing time, but can be done centrally and iteratively in the global TravelSampo data space through generating RDFS, OWL and SKOS equivalencies. Actually, this task becomes one with the general task of mapping different RDF materials to each other, and could use any of the readily available ontology and instance mapping tools [8, 4] for doing just that.

Due to this, all but certain mappings are also relegated into this late stage of processing, so that for example all keywords found in the data sources during pipeline processing are created as resources in the data source’s own namespace, instead of being equated with ontology concepts directly. This also makes sure

¹ <http://www.seco.tkk.fi/projects/finnonto/>

² The source code of Harava has been released under a MIT style open source license and published as a Google Code project³.

that no information is lost and no errors introduced in indexing, due to e.g. the keywords used not being found in the reference vocabularies, or being translated to a wrong concept based on improper fuzzy reasoning.

As already said, the semantic enrichment done to the materials through information extraction tools is also done in this global data space. This ensures that, for example, when searching for concepts from textual descriptions, the algorithms have a maximal amount of content available from which to draw matches.

The transparent resolution of equivalency statements also means that any erroneous mappings can be undone easily after the fact by just removing the RDF triple specifying the bad mapping. The tasks of verifying and improving the resource mappings generated, as well as verifying automatic enrichments, can be done in the TravelSampo ecosystem through the SAHA metadata editor [9] created in the FinnONTO project, which has special support for going through annotations marked as suspect. The marking of such annotations can either be done originally in the enrichment process, or at a later date by utilizing heuristic or schema-based quality assessment rules.

For this latter task, the FinnONTO architecture contains the semantic content validator service VERA⁴. The output that VERA produces is not a list of errors per se, but rather a list of possible problems that an expert user can assess, and modify the schema or data as needed. The report also contains general statistics about the data, such as language definition usage, so it can also be used for a general analysis instead of validation.

In this way, the dynamically updated content of the TravelSampo portal can be iteratively improved and corrected as the system is running, right when problems are discovered, allowing focusing on the areas most critical to efficient use.

5 Indexing and Querying

Our stated choice of semantic integration by mapping properties and resources in RDF required that the triple-store used had to support easy and efficient resolution of both equivalency as well as subsumption relations, as those were the primary means used to map content.

In fact, in the custom triple-store implemented for TravelSampo, both of these are done transparently. As an example, a query for “?s rdfs:label ?o” would return also all skos:prefLabel and skos:altLabel triples, as well as any custom schema properties marked as equivalent to any of these. A query for “?s rdf:type foaf:Agent” on the other hand returns also instances of all the subclasses of foaf:Agent. For ease in additional processing, a unified view to the data is also provided, where all URIs in an equivalency set in the source are replaced with a single canonical version. This way, anyone processing the results of such inferred queries need not themselves repeat the equivalency calculation.

⁴ <http://www.seco.tkk.fi/services/vera/>

In the material used for TravelSampo, a total of 11 million equivalency sets were discovered, touching 25 million resources out of a total 350 million.

In addition to subsumption and equivalency inference, the triple-store of TravelSampo also includes support for quickly discovering all location resources annotated with geo-coordinates inside a specified bounding-box, as well as all other resources related to those locations. The same is done for any temporal entity resources such as event times and further resources related to them. These are all functionalities that were needed in the various user interfaces of the TravelSampo system. Similarly, efficient text search is provided for searching 1) objects by their labels, 2) objects by their literal attributes and 3) objects by the labels associated with their object attributes. The last index is used in the general text search interface of TravelSampo, so that one can for instance query by the string “Pyhäjärvi” and be quickly returned all objects that relate to any of the 50 or so lake Pyhäjärvis of Finland.

In order to cater to the complexity and heterogeneity of the data sources used in TravelSampo, the indexing and querying system also has to be able to efficiently query quite complex patterns. For example, the intent is that for example the query “Finnish electronica near Helsinki in the following week” would match a concert by Jimi Tenor at the Helsinki Ice Hall, because Jimi Tenor is a Finnish electronica artist playing there in the specified timeframe. However, inside the data model, this is quite a complex pattern, as visualized in figure 2.

To answer the query, first, each resource with a label matching any of the keywords must be found, as well as those matching the temporal and spatial constraints. This results in a result set with (among others) the nationality Finnish, the genre Electronica, the concert event and the location of Helsinki ice hall. Then, all resources relating to these or their subconcepts must be added to the result set. This results in (among others) the artist Jimi Tenor (who is Finnish) and the album Intervision (which has a genre of downtempo, which is a subgenre of electronica).

Finally, all resources that are not already locations must be mapped to any that they refer to, and finally an intersection taken between all locations found to reveal the final result. In this case, such mappings must also be done iteratively. While the concert event relates directly to the location, but the artist and the album are still two and three steps away, respectively. To obtain the final result, one must follow first the link from album to the artist, and then from the artist to the event, which then finally leads to the location.

To resolve this, the search functionality in the TravelSampo backend was split into multiple stages, each taking in SPARQL queries. First, multiple “select” queries are run, one for each incoming keyword, temporal and spatial constraint, acting on a dedicated index. Using the index, it is easy to efficiently return not only resources matching the spatial and temporal constraints, but also any resource that is related in any way to them, or a literal or another resource with a label matching a particular text query. In addition, this index also performs subclass inference. Thus, from this stage, in the case of the example queries one

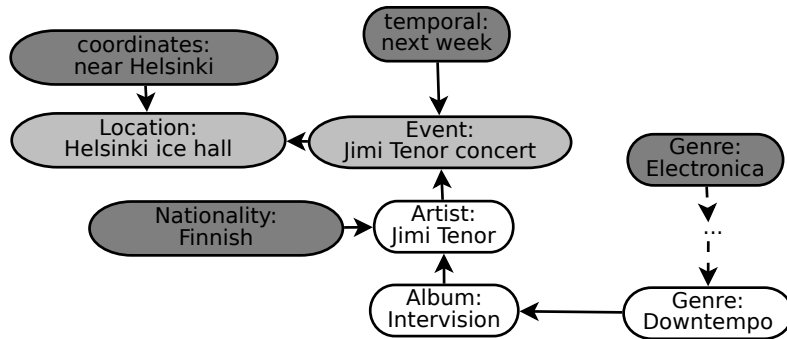


Fig. 2. Mapping to a final search result after keyword and spatiotemporal matching in TravelSampo. Dark grey resources are those returned from matching, while the light grey resources are the final search result.

would already get the artist Jimi Tenor and the album Intervision, in addition to the more direct resource hits.

Then, “mapping” queries are run separately and iteratively for each select query result set. In the example, these would map for example any albums to artists, artists to events and events to locations. After this, the system automatically takes an intersection of the mapped results returned from each select query. A further “filter” query is also run. In TravelSampo, this makes sure that only locations ever make it to the final result set returned.

After the result set is finally obtained, it is paged and returned. This can still be manipulated by a “grouping” query. This can be used to ensure that for example a set amount of both event locations and culturally significant locations matching a particular query are returned. To make sure all information to be shown in the search listing for each matched resource is included (such as images, event details, etc.), the system still runs any given “describe” queries for each returned resource, before finally returning answers.

Because of the efficient indexes of TravelSampo as well as caching of e.g. the mapping query results, the average processing time for even these complex queries is still 100-400 milliseconds on a modern desktop server.

6 Contributions

While still a work in progress, the TravelSampo system already demonstrates the potential for a much richer way of searching for points of interest. In developing the system, multiple issues were identified.

Firstly, locations may be of interest not only through their immediate properties, but through quite long chains of associations. Secondly, it is hard to isolate points of interest from other general locations.

In processing actual databases for use in the TravelSampo system, the lack of machine-processable content keywords in most currently available datasets was

identified to be a major problem. In the TravelSampo system, this was addressed by integrating state of the art information extraction tools into the system.

In order to enhance precision and recall in searching the heterogeneous datasets, key class ontology level reference resources in the TravelSampo system such as point of interest types were mapped to each other by hand. However, another requisite part of an integration architecture such as TravelSampo is still the support for iterative, automatic mapping of the instances and keyword concepts in the different datasets pouring in, sometimes dynamically each day. An equally important feature is the ability of human editors to correct these mappings.

Finally, the TravelSampo system and the datasets loaded into it highlight the complexity of queries needed to cater to complex needs, while demonstrating that answering such queries efficiently even on massive data sources is still quite possible.

Acknowledgements This research is part of the Semantic Ubiquitous Services Project (SUBI) 2009–2011, funded by the Finnish Funding Agency for Technology and Innovation (Tekes) and a consortium of 18 companies and public organizations, pre-eminently Turku – European Capital of Culture 2011. Some work has been funded also by the Finnish Cultural Foundation.

References

1. Auer, S., Lehmann, J., Hellmann, S.: LinkedGeoData: Adding a spatial dimension to the web of data. In: The Semantic Web-ISWC 2009. pp. 731–746. Springer (2009), <http://www.springerlink.com/index/j63221026432x374.pdf>
2. Becker, C., Bizer, C.: DBpedia Mobile : A Location-Enabled Linked Data Browser. In: Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008). Beijing (2008)
3. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - a crystallization point for the web of data. Web Semantics: Science, Services and Agents on the World Wide Web 7(3), 154–165 (2009), <http://www.sciencedirect.com/science/article/B758F-4WS9BS0-1/2/83cd58f9b584b76ccaa85cda59cca3a2>
4. Choi, N., Song, I.Y., Han, H.: A survey on ontology mapping. SIGMOD Rec. 35, 34–41 (September 2006), <http://doi.acm.org/10.1145/1168092.1168097>
5. Hyvönen, E.: Developing and using a national cross-domain semantic web infrastructure. In: Sheu, P., Yu, H., Ramamoorthy, C.V., Joshi, A.K., Zadeh, L.A. (eds.) Semantic Computing. IEEE Wiley - IEEE Press (May 2010)
6. Hyvönen, E., Mäkelä, E., Kauppinen, T., Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K., Kuittinen, H., Viljanen, K., Tuominen, J., Palonen, T., Frosterus, M., Sinkkilä, R., Paakkarinen, P., Laitio, J., Nyberg, K.: CultureSampo – Finnish culture on the semantic web 2.0. Thematic perspectives for the end-user. In: Proceedings, Museums and the Web 2009, Indianapolis, USA (April 15–8 2009)
7. Hyvönen, E., Tuominen, J., Kauppinen, T., Väätäinen, J.: Representing and utilizing changing historical places as an ontology time series. In: Ashish, N., Sheth, A. (eds.) Geospatial Semantics and Semantic Web: Foundations, Algorithms, and Applications. Springer-Verlag (2011, forth-coming)

8. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. *Knowl. Eng. Rev.* 18, 1–31 (January 2003), <http://portal.acm.org/citation.cfm?id=975027.975028>
9. Kurki, J., Hyvönen, E.: Collaborative metadata editor integrated with ontology services and faceted portals. In: Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010, Heraklion, Greece. CEUR Workshop Proceedings (June 2010)
10. Medelyan, O.: Human-competitive automatic topic indexing. Ph.D. thesis, University of Waikato, Department of Computer Science (2009)
11. Ruotsalo, T., Mäkelä, E., Kauppinen, T., Hyvönen, E., Haav, K., Rantala, V., Frosterus, M., Dokoohaki, N., Matskin, M.: Smartmuseum: Personalized Context-aware Access to Digital Cultural Heritage. In: Proceedings of the International Conferences on Digital Libraries and the Semantic Web 2009 (ICSD2009). Trento, Italy (2009)
12. Suominen, O., Hyvönen, E.: Expressing and Aggregating Rich Event Descriptions. In: Proceedings of the 6th Workshop on Scripting and Development on the Semantic Web (2010)
13. Wilson, M., Russell, A., Smith, D.A., Owens, A., Schraefel, M.: mSpace mobile: A mobile application for the semantic web. In: Proceedings of the ISWC 2005 End User Semantic Web Interaction Workshop (2005), <http://eprints.ecs.soton.ac.uk/11101>

An Open Source Linked Data Framework for Publishing Environmental Data under the UK Location Strategy

Arif Shaon¹, Andrew Woolf², Shirley Crompton¹, Robert Boczek¹, Will Rogers¹, Mike Jackson³

¹ e-Science Centre, The Science And Technology Facilities Council, Rutherford Appleton Laboratory, Didcot, OX11 0QX, UK

{arif.shaon, shirley.crompton, will.rogers, robert.boczek}@stfc.ac.uk

² The Bureau of Meteorology, Canberra, Australia

A.Woolf@bom.gov.au

³ The Software Sustainability Institute, The University of Edinburgh, Edinburgh, UK
michaelj@epcc.ed.ac.uk

Abstract. Linked data offers a novel and more flexible means of sharing complex geospatial datasets by breaking away from the traditional domain-specific technologies used for accessing and integrating geospatial data with heterogeneous sources and disparate formats. In 2010, the UK Cabinet Office released a set of draft guidelines for exposing geospatial data as linked-data in support of the UK Open Data initiative. These draft guidelines have been proposed under the UK Location Strategy in specific recognition of the importance of geospatial data, and also with a view to promote linked-data within the EU INSPIRE community. This paper presents a customisable open-source linked-data framework developed by the GeoTOD-II project that implements these guidelines. The framework provides an efficient means for exposing both existing and new data sources in the linked-data form. We also attempt to articulate and address a number of issues and hidden assumptions with these guidelines identified during the development of the framework.

Keywords: linked-data, geospatial data, INSPIRE, DEFRA, GeoTOD

1 Introduction and Motivation

The UK government's "data transparency agenda" aims to make public sector data freely accessible on the web as linked-data. This was greatly inspired by Tim Berners-Lee's invitation in 2009 [1] to publish government data online in light of the emergence of the Linked Open Data movement. While the primary goal of this initiative is to increase accountability associated key public sector datasets, it will, more importantly, enable harmonisation of heterogeneous datasets in a standardised manner by creating a "web of data", thus supplementing the knowledge base of individuals as well as society.

For geospatial information in particular, the linked-data approach offers the potential for developing more flexible means of data sharing and accessibility. In

essence, this could help solve the traditional problems of harmonising geospatial data with heterogeneous sources and disparate formats through standardised but complex web-services. For example, an RDF¹-based linked-data representation of a climate research dataset identified by a unique HTTP URI could be seamlessly linked to another related but external dataset also exposed as linked-data in RDF, through an appropriate vocabulary e.g. RDFS² 'seeAlso'. This would enable a user (whether an application or human) to seamlessly access both of these datasets through their respective resolvable URIs and/or interrogate the datasets using the linked-data recommended query language, SPARQL³ without being constrained by the query language or access mechanism(s) specific to the underlying geospatial web-service(s) (e.g., an OGC Web Feature Service⁴ instance) responsible for serving up these datasets.

There are however several caveats to effectively sharing linked resources using URI and RDF. The chief amongst these is the necessity of a specific community data model, or 'RDF vocabulary'. While RDF provides the base representation for linked-data, this is not enough to specify the internal structure of any specific dataset (much as HTML provides a flexible structure for a huge variety of web page content). As noted by Tim Berners-Lee [2], "Different communities have specific preferences on the vocabularies they prefer to use for publishing data on the Web. The Web of Data is therefore open to arbitrary vocabularies being used in parallel. Despite this general openness, it is considered good practice to reuse terms from well-known RDF vocabularies..." Unfortunately the most well-known RDF vocabularies have little to do with climate research data – they are concerned with social networking (FOAF⁵), blogs/wikis (SIOC⁶), thesauri (SKOS⁷), software projects (DOAP⁸), etc.

It is also important with linked-data to strike the right balance with URI structure between completely opaque identifiers and excessive human-readable semantics⁹. To address this issue, the UK Cabinet Office has released a set of draft recommendations [3] for designing URI identifiers for location data in support of the UK Open Data initiative. These draft guidelines extend more general ones [4] for publishing public sector data (under data.gov.uk), and have been proposed under the UK Location Strategy in specific recognition of the importance of geospatial data, and also

¹ Resource Description Framework (RDF) - <http://www.w3.org/RDF/>

² RDF Schema - <http://www.w3.org/TR/rdf-schema/>

³ SPARQL Query Language for RDF - <http://www.w3.org/TR/rdf-sparql-query/>

⁴ Open Geospatial Consortium (OGC) Web Feature Service - <http://www.opengeospatial.org/standards/wfs>

⁵ The Friend of a Friend (FOAF) vocabulary - <http://xmlns.com/foaf/spec/>

⁶ Semantically-Interlinked Online Communities (SIOC) - <http://sioc-project.org/ontology>

⁷ Simple Knowledge Organization System Reference (SKOS) - <http://www.w3.org/TR/swbp-skos-core-spec>

⁸ <http://code.google.com/p/baetle/wiki/DoapOntology>

⁹ W3C, "Cool URIs don't change", <http://www.w3.org/Provider/Style/URI>

recognising parallel work at the European level on deploying the INSPIRE¹⁰ ‘spatial data infrastructure’ (which uses web services, but not linked-data principles).

In addition, a linked-data service should integrate (e.g. as a layer over or an additional component) with existing data sources (e.g. web services, databases) without the need to make substantial changes to the underlying infrastructure. For example, it may not be desirable to significantly modify an existing Web Feature Service serving up external data from a third party database; or to replace it with a linked-data service to provide linked-data representations of these data. What might be more efficient and practical in this scenario is to implement a linked-data service that wraps the Web Feature Service and leverages it as a “proxy” data source for exposing linked-data.

This paper presents an open-source geospatial linked-data framework developed by the GeoTOD-II¹¹ project that implements the UK Cabinet Office’s draft guidelines for exposing geospatial data as linked-data. This framework provides an efficient means for exposing existing data sources as linked-data using the approach proposed above.

2 Key Concepts

In this section, we provide an overview of the key concepts pertinent to the work of the GeoTOD-II project presented in this paper.

2.1 Linked-data

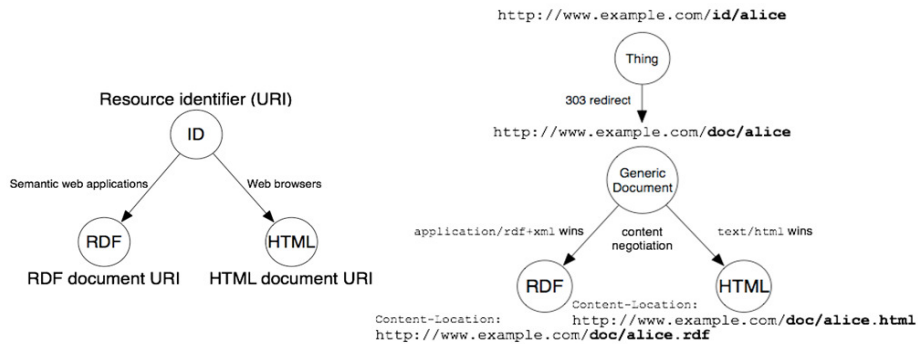


Fig. 1. Linked-data principles: client-dependant resource identification through HTTP URI and resource retrieval through content negotiation. (Source: <http://www.w3.org/TR/cooluris/>)

The success of today’s web results from two core functionalities: the ability to identify and link documents using the HTTP protocol. Simple to implement, widely deployed, and with ubiquitous client support, these two elements provide an obvious model for moving beyond text and documents to a web of data. The ‘linked data principles’ [5] adopt this model by using URIs to identify data objects (or the real-

¹⁰ INSPIRE Directive - <http://inspire.jrc.ec.europa.eu/>

¹¹ Geospatial Transformation with OGSA-DAI (GeoTOD-II), SourceForge <http://geotod.sourceforge.net/about.html>

world ‘things’ that they represent), and creating a data web by linking together related data objects. While HTML provides the *lingua franca* for the web of documents, RDF plays that role for data (Fig. 1). Common to both is the use of HTTP to access information (linked-data also recommends a human-readable representation e.g., HTML, if accessed via a web browser, using ‘content negotiation’¹² – Fig. 1). The adoption of the four elements of linked data (URIs, RDF, HTTP, links between data) has already led to a massive ‘linked data cloud’¹³ connecting hundreds of datasets and billions of individual data items.

2.2 Designing ‘URI Sets’ for Location

The European Union’s INSPIRE Directive requires public authorities across Europe to provide access to their environmental datasets through the adoption of a common framework for uniquely identifying the datasets within a pan-European ‘spatial data infrastructure’. The UK Cabinet Office’s guidelines for “Designing URI Sets for Location” [3] “is focussed on the use of http: URI by the UK public sector to meet that INSPIRE objective”. To that end, the guidelines define three different types of resources identified by three different Uniform Resource Identifier (URI) schemes.

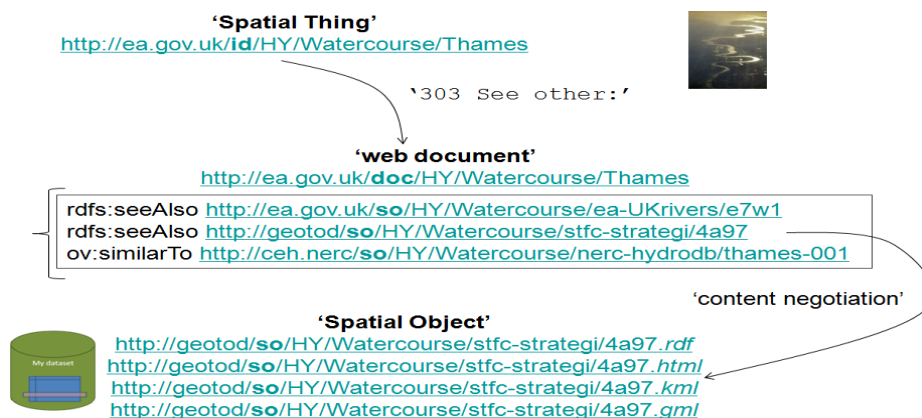


Fig. 2. HTTP URI identifiers for ‘Spatial Things’ and ‘Spatial Objects’

Spatial Thing. The guidelines define this as anything that has a spatial extent, i.e. size, shape or position, and is a subset of ‘real-world’ phenomena associated with a location, e.g. the ‘River Thames’. To uniquely identify a Spatial Thing, the guidelines recommend the following URI scheme, referred to as the “Id” URI:

[http://location.data.gov.uk/id/{INSPIRE theme}/{concept}\[/codeset\]/\[reference\]/\[version\]](http://location.data.gov.uk/id/{INSPIRE theme}/{concept}[/codeset]/[reference]/[version])
([] denotes an optional term).

¹² A web server returns a representation of a resource based on the HTTP-Accept header of a client request.

¹³ <http://linkeddata.org>

An example URI for the ‘River Thames’:

<http://location.data.gov.uk/id/HY/Watercourse/Thames/v1>

According to the guidelines, the URI for a Spatial Thing, if de-referenced, should be re-directed to a web document containing metadata about the Spatial Thing (Fig. 2). The granularity of the metadata is implementation or provider specific. The URI pattern for identifying this metadata document should be that for the Spatial Thing but with the term “id” replaced with the term “doc” – hence, it is referred to as the “Doc” URI.

An example “Doc” URI for the ‘River Thames’:

<http://location.data.gov.uk/doc/HY/Watercourse/Thames/v1>

In addition, a metadata document about a Spatial Thing could also include a list of relevant, known, Spatial Objects (described below) through appropriate vocabulary (e.g. RDFS ‘seeAlso’ or OpenVocab¹⁴ ‘similarTo’ – Fig. 2).

Spatial Object. This is essentially a concrete digital representation of a ‘real-world’ phenomenon associated with a specific geographical location. Notably, this is a direct proxy of the INSPIRE definition of a Spatial Object¹⁵. The guidelines propose the following URI scheme (referred to as the “So” URI) for uniquely identifying a Spatial Object in an INSPIRE compliant way:

<http://location.data.gov.uk/so/{INSPIRE theme}/{Ontology Class}/{Ontology Namespace}/{local id}/{version id}/{rendition}>

For example, the URI for a Spatial Object representation of the ‘River Thames’ could be:

<http://mydata.co.uk/so/HY/Watercourse/hy-p/1234/v1>

As illustrated in Fig.2, multiple representations of the same Spatial Object could be provided by using an efficient content negotiation mechanism.

Ontology. In addition, the guidelines also define a number of URI patterns for querying the concepts used within a description of a Spatial Thing or Spatial Object. These concepts are essentially the Classes and their associated properties defined within an OWL Ontology¹⁶ or RDF Schema representation of an INSPIRE conceptual model (typically formulated in UML¹⁷) underpinning the description of a Spatial Thing or Spatial Object. These URI schemes (referred to as the “Def” URIs) are:

- URI for a class

¹⁴ <http://open.vocab.org/docs/similarTo>

¹⁵ INSPIRE Glossary item 67 in <https://inspire-registry.jrc.ec.europa.eu/registers/GLOSSARY/items>

¹⁶ OWL 2 Web Ontology Language Document Overview - <http://www.w3.org/TR/owl2-overview/>

¹⁷ Unified Modelling Language (UML) - <http://www.uml.org/>

[http://location.data.gov.uk/def/{theme}\[/package\]\[/{conceptclass}\[/version\]/{class}](http://location.data.gov.uk/def/{theme}[/package][/{conceptclass}[/version]/{class})

- URI for a property exclusively associated with the given class
[http://location.data.gov.uk/def/{theme}\[/package\]\[/{conceptclass}\[/version\]/{class}\[/property\]](http://location.data.gov.uk/def/{theme}[/package][/{conceptclass}[/version]/{class}[/property])
- URI for a shared or re-usable property
[http://location.data.gov.uk/def/{theme}\[/package\]\[/{conceptclass}\[/version\]/{property}](http://location.data.gov.uk/def/{theme}[/package][/{conceptclass}[/version]/{property})

So, to access the definition of the ‘Watercourse’ class in the above ‘River Thames’ Spatial Object example, the URI would be:

<http://mydata.co.uk/def/HY/rivers-package/Watercourse>

3 Key Issues and Challenges

3.1 Pragmatic Interpretation of the “Designing URI Sets for Location” guidelines

The key challenge faced by the GeoTOD-II project was interpreting the Cabinet Office’s guidelines in a pragmatic and implementable fashion, as there had so far been little practical application of these guidelines. There are a number of issues and hidden assumptions in the guidelines that needed to be articulated by the project.

For instance, a key question raised by the URI scheme proposed for identifying a ‘Spatial Thing’ is: in an operational context what information should be available at the ‘Doc’ URI which describes a Spatial Thing? We choose to regard this as a ‘master catalogue’ of individual Spatial Objects available from different providers and which relate to the associated Spatial Thing, e.g. all registered representations of the River Thames.

Similarly, it is necessary to clarify matters relating to governance and ownership of concepts: e.g. who is the owner of the concept ‘River Thames’ with responsibility to maintain the ‘id/Doc’ URI? There is an implied ‘registration’ process – all owners of ‘River Thames’ objects must register them with the owner of the concept ‘River Thames’.

3.2 Legacy Geospatial Data Sources

Another issue is how to achieve linked-data representation of legacy geospatial data sources with minimal cost to data providers. As highlighted before, the recommended practice is for linked-data representations to co-exist with any current data sources and representations in order for it to be useful. Therefore, a linked-data solution would effectively sit on top of existing data sources and be configured to use those data sources without changing their underlying data structures or storage formats.

3.3 Ontology Representations of the INSPIRE Conceptual Models

Additionally, there is also the question of representing geospatial data in RDF, which requires developing RDF ontologies based on the UML conceptual information models adopted by INSPIRE (and their underpinning ISO standards) to describe these legacy data sources. There are a number of issues related to the “mappings” between the INSPIRE UML models and their OWL/RDF Ontology/Schema representations. For instance:

- There is a need to define a canonical transformation from geospatial UML conceptual models to an ontology representation. In general, the ‘closed-world’ semantics of UML are more restrictive than the ‘open-world’ model of OWL and RDFS. As well, the UML meta-model does not support properties as first-class entities. Nevertheless, UML bears similarities to frame-based knowledge modelling systems, and the Object Modelling Group has developed the Ontology Definition Metamodel (ODM) as a mechanism for modelling UML-based ontology development.
- Developing an ontology representation of an INSPIRE UML model would need also to address the ‘import’ of already-existing information models developed as international standards by ISO’s Technical Committee 211 (e.g. ISO 19107 for spatial schemas, ISO 19108 for temporal schemas, ISO 19115 for geospatial metadata, etc.). It would require the development of ontologies for these ‘imported’ models as well. These are substantial tasks in their own right requiring considerable involvement of the wider standards community.
- In order to provide conventional GML¹⁸ as a specific representation of INSPIRE geospatial data under a linked-data server (through content negotiation), further work is required on developing open-source implementations of the INSPIRE web services (i.e. the OGC ‘Web Feature Service’).

Notably, there have been a few such ontologies, albeit unofficial and generally incomplete, emerging from the INSPIRE and other related communities. For instance, the W3C Semantic Sensor Network Incubator Group (SSN XG)¹⁹ developed an ontology based in part on the ISO 19156 ‘Observations and Measurements’ conceptual model. And the UK Environment Agency has developed a linked-data representation of Bathing Water Quality including ‘sampling points’²⁰ motivated by the INSPIRE ‘Environmental Monitoring Facility’ theme. The OGC GeoSPARQL working group is developing a SPARQL extension to include spatial query predicates²¹ (touches, disjoint, overlaps, contains, etc.). In addition, the draft specification includes a number of OWL class definitions for geometry, topology, and geospatial ‘features’.

¹⁸ Geography Markup Language - <http://www.opengeospatial.org/standards/gml>

¹⁹ W3C Semantic Sensor Network Incubator Group - <http://www.w3.org/2005/Incubator/ssn/>

²⁰ <http://location.data.gov.uk/def/ef/SamplingPoint/SamplingPoint>

²¹ The so-called Egenhofer relations.

4 Existing Linked-data Servers

We assessed the suitability of a number of existing open-source linked-data servers for publishing geospatial datasets in accordance with the Cabinet Office’s guidelines discussed above. These servers included D2R²², Virtuoso²³, Triplify²⁴, SquirrelRDF²⁵ and Pubby²⁶. In general, these existing products enable exposing RDF views of datasets residing in relational databases through customisable HTTP URLs and querying them using SPARQL. In all cases, the desired mappings between an RDF schema and a relation database schema are specified in some form of mapping file(s) written in RDF/Turtle based languages with varying levels of syntactical and conceptual complexity. Additionally, some of these servers, such as D2R and Virtuoso support automated generation of the mapping files based on the schema of the relational database specified.

We also reviewed a number of geospatial linked-data services that are based on some of the aforementioned linked-data servers. This included the UK Environmental Agency’s linked-data server²⁷ that implements the Cabinet Office’s guidelines, and the “GeoLinked Data” service developed by the Ontology Engineering Group (OEG)²⁸ for publishing environmental data held by the National Geographic Institute of Spain.

Generally, most of the existing linked-data servers and services reviewed provide complete (and in some cases, complex) solutions for publishing linked-data. However, the major drawback of these solutions is the limited scope for their functional extensibility. For instance, adding a non-relational data source, such as a Web Feature Service for rendering GML representation of a geospatial resource, to any of these servers would likely require substantial re-implementation of its underlying architecture. Such a task, while achievable as most of these servers are open-source products, may not be practical within the related scope of work.

Providing alternative non-RDF representations (e.g. XML) of a linked-data resource is recommended in the linked-data principles. And considering that RDF is not a prevalent format for encoding geospatial datasets, there is added value for a geospatial linked-data server to have the capability to provide both non-RDF (such as GML) and RDF representations of a resource. Notably, the provision of GML-encoded environmental data is also not supported by either of the two geospatial linked-data services mentioned above.

To this end, we concluded that there would be merit in developing an open-source linked-data server for publishing geospatial datasets with the key features of the

²² <http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/>

²³ <http://virtuoso.openlinksw.com/linked-data/>

²⁴ <http://triplify.org/>

²⁵ <http://jena.sourceforge.net/SquirrelRDF/>

²⁶ <http://www4.wiwiss.fu-berlin.de/pubby/>

²⁷ <http://data.gov.uk/linked-data>

²⁸ <http://geo.linkeddata.es/web/guest/home>

aforementioned servers but crucially with appropriate extensibility points for adding new functionality as needed.

5 Methodology and Implementation

5.1 GeoTOD Linked-data Server Framework

To address the issues of the existing linked-data servers, in the GeoTOD project we developed a highly-extensible framework for a linked-data server, namely the GeoTOD Linked Data Server (GLS), which implements a set of Linked Resource interface specifications compliant with the Cabinet Office guidelines for the publication of geospatial data.

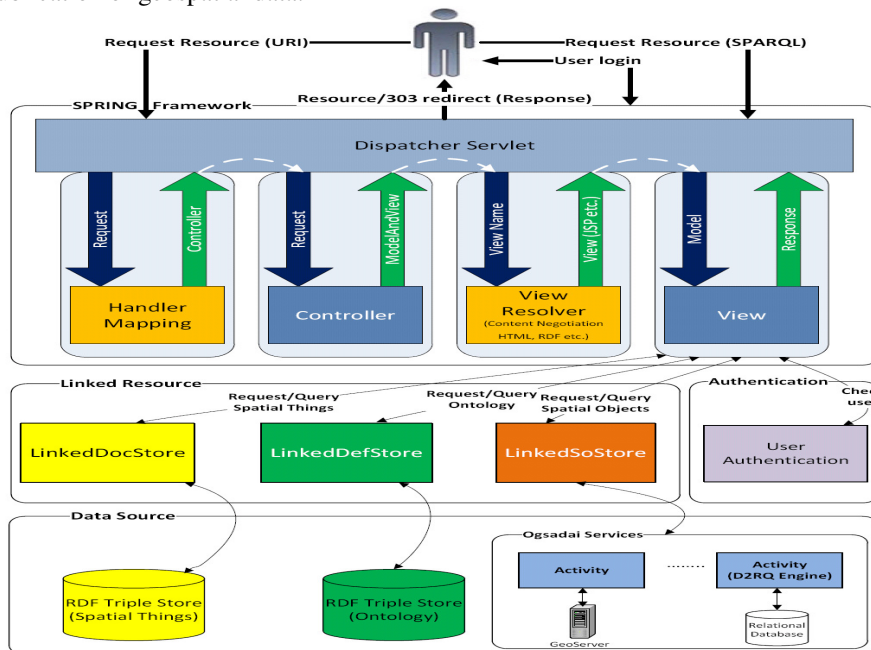


Fig. 3. An architectural view of the GLS Framework

The GLS architecture (Fig. 3) follows the Spring Model-View-Controller (MVC)²⁹-based Java EE³⁰ framework with four different layers of components: Spring MVC framework, Linked Resource, Authentication and Data Source. Of particular note is the Linked Resource layer, which integrates within the GLS framework three Linked Stores (*LinkedDocStore*, *LinkedSOSStore* and

²⁹ The Spring MVC Framework - <http://static.springsource.org/spring/docs/3.0.x/spring-framework-reference/html/mvc.html>

³⁰ Java Enterprise Edition 6- <http://www.oracle.com/technetwork/java/javaee/tech/index.html>

LinkedDefStore) representing the three resource types (Spatial Thing, Spatial Object and Ontology respectively) specified in the Cabinet Office’s guidelines (Section 2.2).

In general, the Spring MVC layer of the GLS framework receives requests related to any of the three Linked Stores as either HTTP URIs or SPARQL queries, and determines the appropriate response to be sent to the client. Formulation of the response to a client’s request mainly involves identification of and communication with an appropriate implementation of the Linked Store in question. The mode of communication between the Spring MVC layer and the Linked Resource layer depends on the type and output format (e.g. HTML, RDF etc.) of the resource requested.

In addition, the GLS Spring Framework handles authentication of users with administration privileges with the help of the “Authentication” layer. User authentication in GLS is required to perform administration related functions (e.g. adding, removing Spatial Objects etc.) in the “Linked Resource” layer.

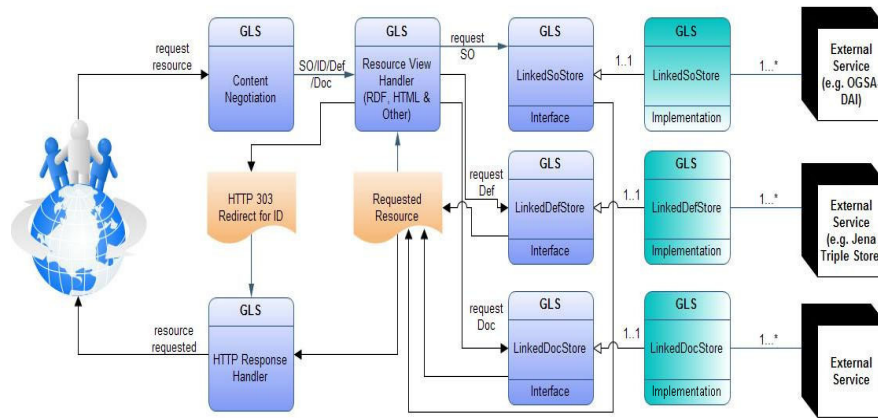


Fig. 4. Integration of the “Linked Stores” within the GLS Framework

As illustrated in Fig. 3 and Fig. 4, depending on the implementation, a GLS Linked Store can act as an interface to any type of service or data store within the GLS Data Source layer. For example, a GLS Linked Store could be implemented as an interface to a single service responsible for serving up linked resources in various supported output formats; or a set of aggregated services, where each service is responsible for producing a specific representation of a linked resource. In other words, the GLS can be integrated with any concrete implementations of Linked Stores conforming to the Linked Resource interface specifications. Thus, the GLS framework enables the publication of both existing and new geospatial data sources as linked-data.

5.2 Prototype Implementation

We implemented a prototype of the GLS framework for a demonstration dataset (described later in Section 6) with the *LinkedSoStore* implemented using OGSA-

DAI³¹ - an open source framework for distributed data management. OGSA-DAI provides the *LinkedSoStore* with an uniform interface to access and integrate third-party heterogeneous relational as well as web data sources (Fig. 3). We extended OGSA-DAI to support RDF resources using the D2RQ platform³² (the mapping engine behind the D2R server). Relational data sources are transformed into virtual RDF graphs using a mapping file, which describes the relation between an ontology and a relational data model. D2RQ also provides RDF/SPARQL access mechanisms to support different types of linked-data query on the legacy geospatial data resources. With this extension to OGSA-DAI, we are able to exploit different types of third party data services and convert their output into linked data representations using configurable OGSA-DAI workflows. With OGSA-DAI open framework, new data services can simply be wrapped and deployed into these workflow via software configuration. The *LinkedDocStore* and *LinkedDefStore* were implemented as interfaces to two native RDF triple stores (Fig. 3). The prototype GeoTOD Linked data server is available at: <http://tiger.dl.ac.uk:8080/geotodls/index.htm>

5.3 RDFS Generator

To support the provision of ontologies via RDF, we also developed a simple *UML to RDFS conversion* tool for converting UML-based conceptual models of a domain, such as INSPIRE thematic data specifications, to an RDF schema (ontology).

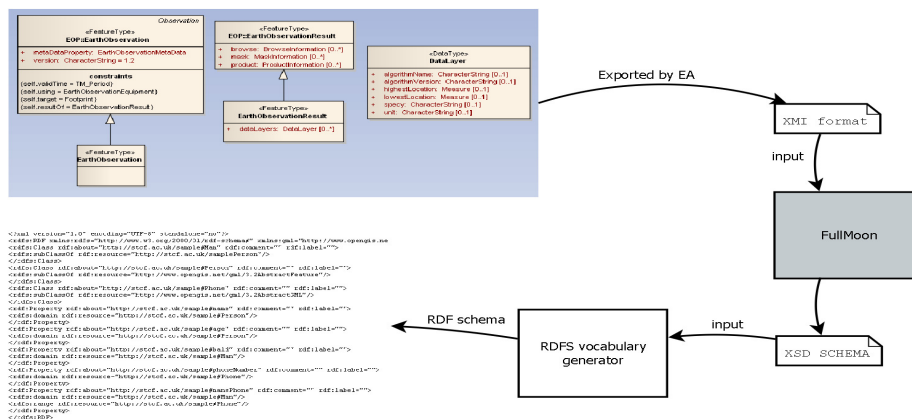


Fig. 5. Generation of RDFS vocabulary from UML Models created using Enterprise Architect (EA)³³

In general, this tool allows generating RDFS vocabulary from an XML Schema transformation of an INSPIRE UML model (Fig. 5). It is mainly designed to support the GML application schemas (XML schemas) generated by FullMoon³⁴ - a widely

³¹ OGSA-DAI - <http://www.ogsadai.org.uk/>

³² D2RQ - <http://www4.wiwiss.fu-berlin.de/bizer/d2rq/>

³³ Enterprise Architect - <http://www.sparxsystems.com/>

³⁴ FullMoon - <https://projects.arcs.org.au/trac/fullmoon/wiki/FullMoon>

adopted tool for generating (GML-based) XML Schema representations of the ISO 19000 series (adopted by INSPIRE) UML models. (It proved easier to generate RDFS from the XSD representation, rather than directly from UML.)

The underlying algorithm for performing the XML Schema-to-RDFS conversion is tightly coupled to the naming convention of XML type definitions and element declarations. However, these conventions arise from UML-to-XML encoding rules specified in ISO 19136 ('Geography Markup Language', Annex E), from which the underlying UML classes may directly be inferred. The RDFS Generator is available at: <http://tiger.dl.ac.uk:8080/rdfsgenerator>.

6 Demonstration Datasets

We have tested our prototype implementation of the GLS framework using the Ordnance Survey's 'Strategi'³⁵ dataset in order to demonstrate the transformation of an existing resource into linked-data form. The Strategi data is relevant to the UK, containing a view of the whole UK including natural and man-made features. This dataset has been made freely available online under the Ordnance Survey's OpenData™ initiative in support of the UK government's 'data transparency agenda'.

For this, we have used an ontology auto-generated from relevant INSPIRE conceptual models (i.e. the 'Hydrography' and 'Transport Networks' themes), using the RDFS Generator described above. Additionally, it was necessary to convert the Strategi data from the original ESRI Shapefiles³⁶ to relational data format (i.e. SQL) using a freely-available tool, namely shp2pgsql³⁷, and then store it in a PostgreSQL³⁸ database. As well as following the UK URI guidelines for spatial data, our prototype provided several representations of Spatial Objects through HTTP content negotiation – RDF, HTML, and GML. The latter was provided through the Geoserver³⁹ open-source Web Feature Service application.

7 Conclusions

The most significant outcome of the work presented here is a customisable linked-data framework that is aligned with the UK Cabinet Office's draft guidelines on applying linked-data in the geospatial context. In developing this framework, we have identified and attempted to articulate and address a number of issues and hidden assumptions with these guidelines. In addition, we have learned key lessons that should be considered by other members of the geospatial linked-data community. Foremost amongst these is the requirement to more fully develop mechanisms for mapping geospatial conceptual information models (normally formulated in UML) to

³⁵ <http://www.ordnancesurvey.co.uk/oswebsite/products/strategi/index.html>

³⁶ <http://en.wikipedia.org/wiki/Shapefile>

³⁷ <http://postgis.refractory.net/docs/ch04.html>

³⁸ <http://www.postgresql.org/>

³⁹ <http://www.geoserver.org>

RDF schemas and ontologies. Further, our implementation of the demonstrator provides an optimal solution combining both the strengths of OGSA-DAI for implementing database-to-linked data transformation together with a linked-data server that can be customised to support other both existing and new data stores and data formats.

On the whole, the work presented should benefit those organisations looking to deploy their geospatial data assets as linked-data among. While a number of industry players are developing commercial tools for linked-data, the availability of a conformant open-source framework will provide substantial benefit both to organisations wishing to publish their geospatial data as linked-data, and to the linked-data community (and Cabinet Office CTO Council itself) in developing best practice in this new field. Notably, the GeoTOD framework was used in a recently completed high-profile project, namely ACRID⁴⁰ that has developed a linked-data approach to publishing complex scientific workflows associated with climate research datasets held by the Climatic Research Unit of the University of East Anglia.

However, in order to fully exploit the work described here, further engagement with key players in both the linked-data and geospatial communities will be required, especially those involved with the UK Location Programme and INSPIRE. Furthermore, the initial development and proof-of-concept presented in this paper is only a small part of the effort required to develop hardened software – significant extra development resource would be required to take the project outcomes the next step to a fully tested, efficient, and reliable software product. Future work will need to address these issues.

Acknowledgments. We are grateful to Dr Brian Matthews (Group Leader, Scientific Information Group, e-Science Centre, STFC) for his advice and feedback. The work presented in this paper was funded by the OMII-UK (formerly the 'Open Middleware Infrastructure Institute UK').

References

1. Berners-Lee, T.: “Putting Government Data Online”, W3C (2009), <http://www.w3.org/DesignIssues/GovData.html>
2. Bizer, C., T. Heath and T. Berners-Lee: “Linked Data – The Story So Far”, International Journal on Semantic Web and Information Systems, 5(3), 1-22 (2009). <http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf>
3. UK Cabinet Office: “Designing URI Sets for Location”, v1.0 (2011). http://location.defra.gov.uk/wp-content/uploads/2011/09/Designing_URI_Sets_for_Location-V1.0.pdf
4. UK Cabinet Office: “Designing URI Sets for the UK Public Sector”, v1.0 (2009). <http://www.cabinetoffice.gov.uk/sites/default/files/resources/designing-URI-sets-uk-public-sector.pdf>
5. Berners-Lee, T.: “Linked Data – Design Issues”, W3C, (2009). <http://www.w3.org/DesignIssues/LinkedData.html>

⁴⁰ Advanced Climate Research Infrastructure for Data (ACRID) - <http://www.cru.uea.ac.uk/cru/projects/acrid/>

Semantic access to INSPIRE

How to publish and query advanced GML data

Sven Tschirner¹, Ansgar Scherp², and Steffen Staab²

¹ Federal institute of hydrology, Koblenz, Germany
tschirner@bafg.de

² Institute for Computer Science
University of Koblenz-Landau, Koblenz, Germany
<surname>@uni-koblenz.de

Abstract. The INSPIRE Directive establishes a pan-European "Spatial Data Infrastructure" (SDI) to make available multiple thematic datasets from the EU member states through stable Geo Web-Services. Parallel to this ongoing procedure, the Semantic Web has technologically fostered the Linked Data initiative which builds up huge repositories of freely collected data for public access. Querying both data categories within distributed searches looks promising. To tackle the associated prerequisites, this paper presents firstly a general approach to translate sophisticated INSPIRE GML data models into Semantic Web OWL ontologies. This is done according to Linked Data principles while preserving selective INSPIRE structural information as annotations. Secondly, a feasible conversion of the Semantic Web query language SPARQL to its Geo Web counterpart "OGC Filter Encoding" is proposed. The language mapping is required for a semantic wrapper over remote INSPIRE Download Services acting as a SPARQL-endpoint and bridging the gap between both worlds.

Keywords: INSPIRE, Linked Open Data, Semantic Web, SPARQL, Geo Web, GML, geospatial data, semantic enablement

1 Introduction

The INSPIRE Directive (2007/2/EC)³ obliges national authorities of the EU-member states to contribute their spatial data according to over 30 harmonized themes (e.g. Hydrography, Protected Sites or Elevation), make them accessible and described via standardized Geo Web-Services. These datasets are considered to be up-to-date, quite reliably, EU-wide and mostly free available forming a very impressive data source for multi-thematic information retrieval.

Because of its free data usage and distributed service-architecture INSPIRE does have a lot in common with the Linked Open Data initiative (LOD). Combining both data worlds while starting federated searches over INSPIRE and LOD-datasets looks attractive as repositories differ in thematic coverage and

³ <http://inspire.jrc.ec.europa.eu/>

data capture conditions. As a scenario let us consider a teacher planning a school trip which should be filled with leisure activities as well as some kind of nature exploration. He find accomodation and leisure facilities through Geo Linked Data⁴ and spatially overlay them with INSPIRE themes Land Use or Protected Sites. Finally the teacher may prepare himself with reliable background information referenced by INSPIRE data and disseminated by public authorities enriching the world of free data. The same governmental controlled Protected Sites data could be combined with the LOD project GeoSpecies⁵ to intersect INSPIRE protection goals with species populations from GeoSpecies at the same time.

To leverage such scenarios, the existing technological discrepancies between INSPIRE and LOD have to be overcome which are mainly due to different knowledge representation and web service interfaces. The goal is a feasible embedding of INSPIRE data in the Semantic Web. INSPIRE itself is based on the Geo Web technologies - which are ISO and OGC⁶ standardizations for Geo Web-services and -transfer formats. So INSPIRE applies the ISO/OGC-approach of modeling physical things, so-called "features", in the Geography Markup Language (GML)⁷. GML is a XML-derivative and GML data models are written in XML-Schema Language (XSD). Conventional GML data models define simple and flat XML-documents having features with only few attributes and thus could be easily queried and transformed to Semantic Web RDF/OWL-triples. With regard to INSPIRE we are faced with advanced GML data models which disclose features with many dependant complex elements and a heavily nested, verbose XML-tree structure. That's why transforming GML to OWL is not straight-forward and target OWL-models in which to transform INSPIRE data must be well-thought-of for not generating triples without proper content.

Former work [3] [13] have already introduced Semantic Web-queries which are translated in order to request OGC data access services - specified as OGC WFS⁸. The conversion of GML-results is done on-the-fly which is a reasonable way to avoid duplicated storage of GML and RDF/OWL-instances and facilitates the access to up-to-date information. But these approaches are concerned either with flat GML-models and so encompass a query translation which is much simpler than INSPIRE requires where XPath-expressions for filter processes became inevitably. Or they don't take into account that the content of the GML→OWL transformation must be refined in order to fulfill Linked Data principles with resolvable URIs and cross-references.

This paper presents a feasible way to perform SPARQL queries on INSPIRE which contains two main achievements. At first we propose a general approach for deriving INSPIRE ontologies from the INSPIRE UML/GML data models in order to define the target models for SPARQL querying. We also suggest common modeling aspects and refinements for Semantic Web-representations

⁴ <http://linkedgeodata.org/About>

⁵ <http://about.geospecies.org/>

⁶ Open Geospatial Consortium, see <http://www.opengeospatial.org/>

⁷ <http://www.opengeospatial.org/standards/gml>

⁸ <http://www.opengeospatial.org/standards/wfs>

due to Linked Data principles. Secondly we outline a SPARQL-endpoint - which is configured with these INSPIRE ontologies and acting as a proxy over WFS-services which are the main realizations for INSPIRE Download Services. Therefore we specify a viable translation from SPARQL to the WFS-query language "OGC Filter Encoding (FE)"⁹ and tackle the prerequisite of references between the INSPIRE ontology concepts and the INSPIRE GML data structure.

2 Towards INSPIRE Linked Data - Basic Architecture

To fulfill the overall goal for federated queries on INSPIRE and Linked Data, the INSPIRE data must be transformed into a Semantic Web representation as Semantic Web-formats are appropriate for integration tasks and SPARQL used to query different repositories in particular. Two key tasks are identified: 1. we need to create common target ontologies for INSPIRE themes and 2. provide query capabilities using SPARQL. INSPIRE UML- and derived GML-data models are regarded as community harmonizations and thus have the same characteristics as Semantic Web domain ontologies. Hence each INSPIRE theme is syntactically transferred to one domain ontology (here called "INSPIRE theme ontology") while preserving INSPIRE element names. Common concepts from different themes, e.g. ISO metadata elements and spatial representations, are outsourced into core ontologies and imported into the theme ontologies if needed.

The reasons for explicit INSPIRE ontologies are:

- they serve as target models for a proper GML→OWL conversion set-up
- they store GML structural information as annotations needed for querying
- they facilitate references to INSPIRE background information
- they prepare the basis for alignment with Semantic Web upper ontologies

Querying capabilities help to perform requests and filter the desired instances from the requested repository. A smart way to support SPARQL queries over INSPIRE data can be accomplished with a proxy application, wrapping remote or local INSPIRE Download Services and acting as a SPARQL-endpoint; see Fig. 1. The architecture has multiple advantages. For instance, no data replication and additional storage facilities are needed, hence queries always access up-to-date datasets. Furthermore, INSPIRE data authorities are not bothered with extra configuration efforts required for semantic WFS-profiles.

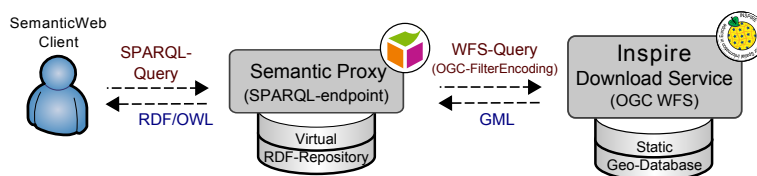


Fig. 1. Overall architecture - Proxy wrapping INSPIRE-Download Services

⁹ <http://www.opengeospatial.org/standards/filter>

Opposed to alternative solutions which comprise one single GML→OWL-transfer into a static INSPIRE triple-store building query facilities on top, the proxy solution has to cope with a virtual repository. That means that all resulting information is temporary and has to be combined with all side effects. The biggest challenge is the query mapping from SPARQL as a powerful general query language to the less powerful language OGC Filter Encoding. The latter is supported by WFS-services and focuses on domain-optimized filters.

We first introduce our proposed query strategy and more details will follow in Sec. 4. Fig. 2 shows the overall query procedure. A SPARQL query is received (1) and converted to SPARQL algebraic expression¹⁰ (2). The algebraic expression is assessed for INSPIRE concepts (3) for which XPath-expressions are concatenated to address the GML-element paths needed within Filter Encoding-query operators. Thereafter the translation SPARQL →Filter Encoding takes place (4) and the WFS GetFeature-request is send to all the WFS-services which serve the requested feature types (5). On return the GML-results are transformed to OWL-instances composing the virtual repository (6). This virtual repository is then queried with the former SPARQL-query from the second step (7) and finally SPARQL-results are returned to the user (8).

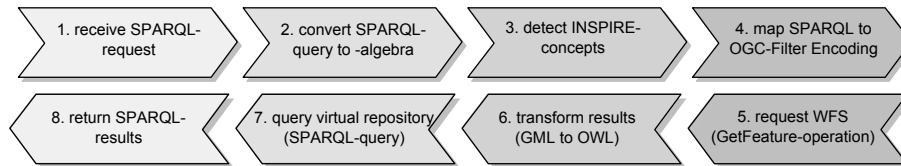


Fig. 2. SPARQL-query handling in eight steps

The resulting INSPIRE ontologies are designed to allow custom queries, for instance to filter Protected Sites and GeoSpecies data. Therefore we outline significant query types which are regarded during ontology modeling (see Sec. 3.2) in order to keep ontological structures simple and adequate for querying:

- distinguish instances by its classification attributes, e.g. "which protected sites are classified with designation = 'UNESCO World Heritage'?"
- do spatial reasoning, e.g. "which species habitats crosses protected site x?"
- filter temporally, e.g. "which sites originates from era x, between dates y/z?"
- assess measures, e.g. "which sites suit a hike, sites greater than x km²?"

This paper continues in Sec. 3 with a description of engineering and modeling aspects of the proposed INSPIRE ontologies. Sec. 4 is about the concept of a semantic query layer and a prototype mentioned in Sec. 5. Then, Sec. 6 outlines related work and Sec. 7 closes with conclusions and possible advancements.

¹⁰ <http://www.w3.org/TR/rdf-sparql-query/>; see Sec. 12.4 "SPARQL Algebra"

3 Deriving INSPIRE ontologies

3.1 Conversion rules: INSPIRE GML to RDF/OWL

Before examining ontological details, we have to consider the basic conversion rules from the original GML INSPIRE data models (written in XSD) to the targeted INSPIRE OWL ontologies. For generic XSD→OWL transformations several approaches [1] [2] and operational applications exist. Accordingly conversion rules are defined to transform either XML- or XSD-documents to OWL, converting e.g. every `xsd:complexType` to `owl:Class`. These rules are helpful for deriving INSPIRE ontologies but don't regard GML or INSPIRE specifics.

GML and RDF/OWL - common characteristics Having a closer look at GML itself which is akin to RDF/OWL [11], as GML-Version 1 had once an explicit RDF-encoding. Since then GML carries the "Object-Property-pattern". This means that XML-elements at odd numbered levels of the DOM-hierarchy represent GML-objects and ones at even level represent GML-properties. Hence one may compare GML-objects to RDF-resources and GML-properties to RDF-predicates. Another similarity is the cross-referencing, which is often used in INSPIRE for cross-thematic linking, and the identification of GML-objects. These basic mappings between GML and RDF are 1. GML-attribute for cross-referencing `xlink:href` which equals RDF-attribute `rdf:resource` and 2. GML-object identifier attribute `gml:identifier` which equals RDF-attribute `rdf:about`. While the Object-Property-pattern is crucial for engineering the theme ontology, the common linking and identification is used during OWL-instance generation.

INSPIRE UML-notations - leading to main conversion rules The UML-models disclose the GML Object-Property-pattern and give further orientation for ontology engineering with UML-class stereotypes and UML-associations as cross-references. The prominent UML-stereotypes are `featureType` (= a WFS record type), `dataType` (= a complex type dependant of `featureType`), `odelist` and `enumeration` (both key-value list-types, open for new values or a closed list respectively) and `union` (semantically equivalent to the XSD union type).

Given these modeling hints, we can list our few main rules which imply that derived OWL-concepts are named after corresponding INSPIRE element types:

- every INSPIRE UML-class except stereotype `union` is converted to an `owl:Class`. Subtypes of stereotype `union` are modeled each as one `owl:Class`
- every value of `odelist` or `enumeration` is converted to an `owl:Individual` typed with the corresponding enumeration/codelist `owl:Class`
- every UML-attribute corresponding to a GML-property is converted to an `owl:ObjectProperty` or `owl:DatatypeProperty`. If multiple, equally-named GML-properties lead to both OWL property types we name the `owl:DatatypeProperty` with suffix `_dataValue` to be conform to OWL-DL
- every UML-association is converted to an `owl:ObjectProperty`

3.2 Modeling aspects of frequent elements

Besides these general rules there have to be ontological refinements for frequently used element types which may leverage usual query types listed in the Sec. 2. This section will examine these aspects in more detail. Furthermore we present our considerations about OWL-instance identification and infrastructural opportunities with INSPIRE reference data.

Classifications with codelist- and enumeration-types In INSPIRE data models every sixth element is of such a type which are best suited to differ and filter instances. We model every codelist-type as one `owl:Class` and every codelist-value as one `owl:Individual`. Enumeration-types are equally modeled. However there is a constraint `owl:oneOf` to their enumeration values.

Modeling this way allows us to investigate every codelist-value of a certain codelist, e.g. `ont_ps:ProtectionClassificationValue` with SPARQL triple: `?x rdf:type ont_ps:ProtectionClassificationValue` and opens the possibility to annotate codelist-values with standard-annotations, e.g. `rdfs:label`. Codelist-value individuals are simply named after their codelist: e.g. the codelist `<http://inspire.ex.org/ProtectionClassificationValue>` has a value `<http://inspire.ex.org/ProtectionClassificationValue/archaeological>`. Hence no naming conflicts arise if two different codelists might have the same value names.

Geometries Concerning geometry handling, there is a current specification development called GeoSPARQL [10]. This state-of-the-art approach includes definitions for a unique geo-vocabulary for which SPARQL extended-functions for extensive spatial reasoning are defined, including topological and other geometrical functions. We extensively make use of this approach to serialize GML geometries resulting from WFS-queries. The proposed SPARQL proxy should provide corresponding GeoSPARQL-spatial filter functions.

Temporal values The conversion of INSPIRE temporal values to RDF/OWL is simple, since GML makes use of XSD-datatypes `xsd:date` or `xsd:dateTime`, too. Mostly resulting RDF typed literals are appropriate. Only in such cases where temporal values are coherent (e.g. start/end-times of a duration) complex RDF-resources have to be derived to retain the coherence and enwrap two or more typed literals for the actual temporal information (e.g. start/end-times).

Measure values Measure values, e.g. the GML-types `Area`, `Length` or `Velocity`, consist of numeric values and corresponding units of measurement (UOM). The encoding in RDF could be a `rdf:PlainLiteral`, e.g. "200 km2", or a typed literal, such as "200" `^^http://www.ex.org/units/km2`. The first option merges the two pieces of information within one textual representation, so that both would have to be discerned during analysis. The drawback of the typed literal

would be the definition of additional RDF-datatypes for maybe tens of UOM which should be avoided for interoperability reasons. Our design choice is therefore based on the "Measure Units Ontology"¹¹ and shown in Fig. 3 (we use namespaces 'ont_ps' for the theme ontology "ProtectedSites" and 'base' for INSPIRE basic types). The two pieces of information are bundled by a separate `base:Measure` instance. For numeric values we can simply adopt the XSD-types used in GML (`xsd:integer`, `xsd:decimal` etc.) and treat the UOM as an individual RDF-resource. Thus, we are able to filter measure values due to their UOM not only in SPARQL filters but even within SPARQL basic graph pattern.

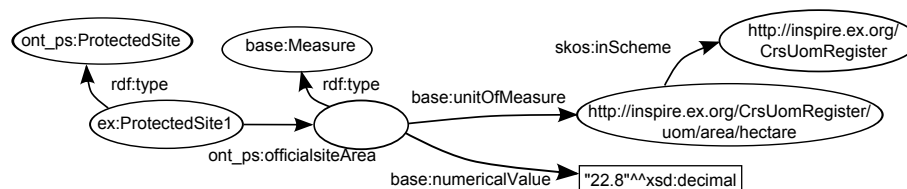


Fig. 3. Using measure values

Registries for reference data Like UOM, other kind of reference data exist, e.g. coordinate reference systems (CRS), languages-codes, country-codes or the INSPIRE codelists/ enumerations. Language-codes are trivial and equally treated in both worlds, INSPIRE and Semantic Web, by using either ISO 639-1/2 or RFC 3066. Other reference data should be explained and collected somewhere online, accessible via resolvable URIs. It makes sense to manage them centrally and in a unique fashion. The OGC has started to make SKOS-concepts¹² of specification elements and some CRS descriptions accessible online¹³. The main technical INSPIRE supporter, the JRC¹⁴ provides SKOS-concepts with the terms of the European GEMET-thesauri and the INSPIRE feature concept dictionary.¹⁵ Such infrastructural measurements will definitely facilitate a Semantic Web-enablement of INSPIRE but also could have positive side-effects for other domains, providing harmonized lists reused e.g. in folksonomies.

Identification As an important Linked Data principle, identifiers for RDF-resources should be kept unique and resolvable via HTTP-URIs. All INSPIRE features are identified with unique INSPIRE-IDs which are used here for individual URIs of OWL-instances. An INSPIRE-ID consists of 1.) a namespace

¹¹ http://forge.morfeo-project.org/wiki_en/index.php/Units_of_measurement_ontology

¹² SKOS: Simple Knowledge Organization System; used for semantically modeling taxonomies; <http://www.w3.org/TR/skos-primer/>

¹³ <http://www.opengis.net/def/>

¹⁴ EU Joint Research Centre, <http://www.jrc.ec.europa.eu/>

¹⁵ <https://semanticlab.jrc.ec.europa.eu/>

including details about the data-authority or -product, 2.) a `localId` which is an object identifier unique in the scope of the `namespace` and 3.) a `versionId` for an optional object versioning. These identifying attributes are proposed to be included in an HTTP-URI with additional format information for Linked Data content negotiation, distinguishing the return-format as e.g. `page` or `data`.

So the proposed URI-template looks like:

http://inspire.ex.org/{format}/{namespace}/{localId}/{versionId}

and an example would be:

http://inspire.ex.org/page/NL.KAD.AU.GEM/4507/V1.0

4 Semantic query layer for INSPIRE Download Services

In this section we show at first how to translate from SPARQL to the WFS-query language "OGC Filter Encoding" and then tackle the prerequisite of references between the INSPIRE ontology concepts and the INSPIRE GML data structure.

4.1 SPARQL conversion to OGC Filter Encoding

We map from a given SPARQL algebraic expression into the target language of the OGC Filter Encoding. More precisely the target language is actually a combination of a) WFS-Query addressing the feature-type to be returned and b) OGC-Filter Encoding (FE) providing spatial, comparison and logical filter operators. From now on, we only mention OGC-Filter Encoding in lieu of both.

SPARQL as well as the XML-encoded FE are declarative languages but they operate on different information units: SPARQL combines triples while FE filters complex GML-features as WFS-records. Given the proposed INSPIRE ontologies one GML-feature and its dependant GML-elements are usually transformed to more than one triple, so FE actually filters at a coarser level than SPARQL does. Besides, there are SPARQL-functions which have no real FE-correspondents, e.g. `isBlank(a)`, `langMatches(a,b)` or `regex(a,b)` (FE only supports a comparator with wildcards, not regular expressions). Furthermore SPARQL-solution modifiers could not be translated, either. With this in mind we conclude that SPARQL is more powerful (even though FE provides more domain-specific filters, e.g. spatial ones). A translation to FE may lose filter information and may be too restrictive at the same time. A viable solution is a two-step query process. In the first step, an overly coarse query is forwarded to INSPIRE Download Services returning a superset of intended query results. In the second step, the precise SPARQL query is re-executed on the returned result set in order to yield the intended SPARQL results.

Some SPARQL filter-functions and value-comparing graphpattern are mapped to FE comparison or spatial operators (e.g. SPARQL-function `sameTerm(A,B)` to FE-operator `<PropertyIsEqualTo>`). In cases where SPARQL applies path-comparing pattern (e.g. `?feature ont_ps:siteName ?siteName`), we propose the mapping to a combination of the two FE-operators `<Not><PropertyIsNull>` to assure that the resulting GML contains only features with these subelements

(here the info `siteName`) for further SPARQL-analysis afterwards. The table 1 shows the mapping of the main SPARQL-operators to their FE-pendants.

Table 1. Mapping SPARQL- to FE-operators

BGP(triple-pattern)	translate every path/value comparisons into appropriate FE-operators and combine them with the logical operator <AND>
Join(P1,P2)	like BGP-handling using both graph pattern P1 and P2
LeftJoin(P1,P2,F)	like BGP-handling, but only for mandatory branch P1, optional branch P2 and Filter F are ignored
Filter(F,P)	like BGP-handling using both filter F and graph pattern P
Union(P1,P2)	translate every comparisons under pattern P1 into appropriate FE-operators, enwrap them with logical <AND>, do the same for P2 and enwrap both <AND>-operators with logical <OR>

This simple conversion example transforms SPARQL-path comparisons and one temporal filter:

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ont_ps: <http://inspire.ex.org/ProtectedSites/>

Select ?feature ?name ?beginLifespan
Where{
  ?feature a ont_ps:ProtectedSite .
  ?feature ont_ps:siteName ?name .
  ?feature ont_ps:beginLifespanVersion ?beginLifespan .
  FILTER( ?beginLifespan > "2009-01-01T12:00:00"^^xsd:dateTime )
}
```

to its corresponding WFS-GetFeature request (abbreviated):

```
...
<wfs:Query xmlns:ps-f="urn:x-inspire:specification:gmlas:
ProtectedSitesFull:3.0" typeName="ps-f:ProtectedSite">
  <ogc:Filter>
    <ogc:And>
      <ogc:Not>
        <ogc:PropertyIsNull>
          <ogc:PropertyName>ps-f:beginLifespanVersion</ogc:PropertyName>
        </ogc:PropertyIsNull>
      </ogc:Not>
      <ogc:PropertyIsGreaterThan>
        <ogc:PropertyName>ps-f:beginLifespanVersion</ogc:PropertyName>
        <ogc:Literal>2009-01-01T12:00:00</ogc:Literal>
      </ogc:PropertyIsGreaterThan>
      <ogc:Not>
        <ogc:PropertyIsNull>
          <ogc:PropertyName>ps:siteName/gn:GeographicalName/gn:spelling/
            gn:SpellingOfName/gn:text</ogc:PropertyName>
        </ogc:PropertyIsNull>
      </ogc:Not>
    </ogc:And>
  </ogc:Filter>
</wfs:Query>
...
```

4.2 Annotations for GML-element references

In order to provide a SPARQL-endpoint which acts as a wrapper around WFS-services the SPARQL-endpoint must map INSPIRE ontologies to GML data models. When the SPARQL-endpoint receives a query, INSPIRE OWL-concepts have to be detected by their URIs (see Fig. 2, third step) and resolved to corresponding GML-element paths in order to compose a WFS-requests with OGC Filter Encoding filters (fourth step). For mapping OWL-concepts to GML-elements we annotate INSPIRE OWL-concepts with relative XPath-expressions for indicating their corresponding GML-element paths. We restrict the usage of XPath-version 1.0 to unique element references which means we avoid wild-cards for node-tests ('*') or predicate-filtering ('@*').

To this end we introduce instances of `owl:AnnotationProperty`:

- Annotation property `cf:xmlnamespace` is used at ontology-level to indicate which GML XML-namespaces are used in the particular INSPIRE ontology, e.g. `<http://inspire.ex.org/ProtectedSites> cf:xmlnamespace "xmlns:ps=\"urn:x-inspire:specification:gmlas:ProtectedSites:3.0\""`. Given these annotations, abbreviations of GML XML-namespaces with XML-prefixes are unique in the scope of the ontology
- Annotation property `cf:xmlname` is used at concept-level stating which OWL-class corresponds to which GML-object element, declared as a qualified XML-name, e.g. `ont_ps:ProtectedSite cf:xmlname "ps:ProtectedSite"` or `ont_ps:ResponsibleAgency cf:xmlname "ps-f:ResponsibleAgency"`
- Annotation property `cf:xpath` is used at concept-level, stating which OWL-property corresponds to which GML-property, declared as a relative XPath-expression (only XPath forward axes), e.g. `ont_ps:isManagedBy cf:xpath "ps-f:ProtectedSite/ps-f:isManagedBy/ps-f:ResponsibleAgency"`

With the `cf:xpath`-annotation, we adopt the triple-relationship "subject-predicate-object" to the GML-model where in the given example above the GML-object element `ps-f:ProtectedSite` stands for the subject, GML-property `ps-f:isManagedBy` for the predicate, `ps-f:ResponsibleAgency` for the object. Figure 4 shows an example with annotations explained in detail subsequently:

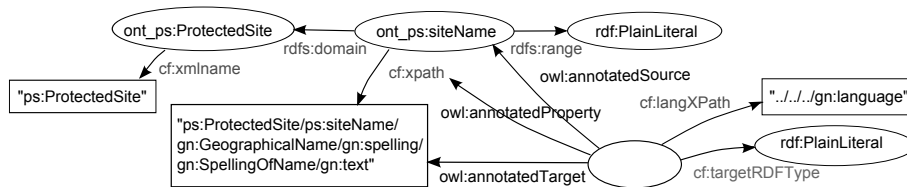


Fig. 4. XPath-expressions annotated to ontology concepts

In Fig. 4, the predicate `ont_ps:siteName` has a complex `cf:xpath`-annotation. This XPath-expression let us extract XML-information nested deeply within the

XML-tree so that we can ignore verbose GML element hierarchies (e.g. ISO-metadata elements) which are not needed in triple form. The other annotations `cf:langXPath` and `cf:targetRDFType` work as transformation hints and are realized as reification annotations so that they are unique for every `cf:xpath`. The reason is that there might be the need to assign multiple `cf:xpath` to one OWL-predicate if different INSPIRE GML-properties are equally named (distinguished only by their superordinated GML-object type) and so are derived both to one OWL-predicate. Or if INSPIRE GML data models allow alternative element storage, so multiple GML element paths each as one `cf:xpath` must be applied. The additional annotation `cf:langXPath` indicates where to find language-information about a text element directing there with a relative XPath-expression related to `cf:xpath`. The annotation `cf:targetRDFType` declares which RDF-datatype should be used for GML→OWL-transformation. Given the configuration example in Fig. 4, a transformation of this GML-snippet:

```
<ps:ProtectedSite>
  ...
  <ps:siteName>
    <gn:GeographicalName>
      <gn:language>deu</gn:language>
      ...
      <gn:spelling>
        <gn:SpellingOfName>
          <gn:text>Kleinkinzig- und Rötenbachtal</gn:text>
          <gn:script xsi:nil="true" />
        </gn:SpellingOfName>
      </gn:spelling>
    </gn:GeographicalName>
  </ps:siteName>
```

leads to the simple statement:

```
ex:ProtectedSite1 ps:siteName "Kleinkinzig- und Rötenbachtal"@de
```

Other GML-information which is likewise related to another GML-element as the language-code could be handled in the same way. A good example is the UOM-information related to its measure-value, so there could be an annotation `cf:uomXPath` and so on. In order to finally prepare WFS-querying, we need at least one more annotation (i.e. `cf:entityType`) for OWL-classes indicating which corresponding GML-object element is expected to serve as a WFS-feature type. In this case, the OWL-class has the annotation `cf:entityType "featureType"`.

5 Prototype implementation

The INSPIRE test platform consists of the INSPIRE-enabled WFS "Deegree3 inspireNode"¹⁶ and representative test data. We have created two ontologies for the INSPIRE themes "Protected Sites" and "Administrative Units"¹⁷. These are tested with national Slovak protected sites (419 features, GML-size: 11 MB) and administrative units from the Dutch Kataster (443 features, GML-size: 23 MB).

¹⁶ WFS-version 1.1.0; <http://www.deegree.org/>

¹⁷ soon available under: <http://inspire.west.uni-koblenz.de:8080/ontologies>

In the INSPIRE implementation process Download Services are scheduled to be operational not until the end of 2012, so for now there is only test data available like the data we have used so far ¹⁸.

The Proxy-application is based on the Sesame framework¹⁹ acting as an inference-layer using the "Sesame" Sail-API. Sesame was chosen due to its well-structured API and fast conversion of SPARQL-query to -algebra. Internally, experimental GeoSPARQL-filters as well as programmetically generated WFS-GetFeature requests are realized with the Deegree3-API. The WFS-results are transformed with the fast non-extracting, XPath-capable XML-parser "VTD"²⁰. For example the parsing of all protected sites elements is finished within 6 seconds, which also includes all further transforming into RDF-triples. Testing federated queries is done with the Sesame-based APIs "NetworkedGraphs" and "DistributedSAIL" from the University of Koblenz-Landau [12].

6 Related Work

At least since the OGC Geospatial Semantic Web Interoperability Experiment in 2006 [7], the Geo Web community is concerned with the integration of Semantic Web-technologies within OGC Geo Web-services (OWS) and vice versa. Some attempts are constrained to a semantic reasoning support of OWS-communication [5], not publishing the Geo Web-data in the Semantic Web. Other approaches, like the active OGC SensorWeb domain, provide measure and sensor data through LOD-endpoints [9] [8]. They even interlink OGC sensor data with other LOD-repositories [6]. Otherwise, without a SPARQL-endpoint, query capabilities are neglected which we suggest to be more important than RDF-browsing and necessary to foster many Semantic Web use-cases. Some projects tackle SPARQL- or DL-queries translated to WFS-queries [3] [13]. We take a further step and query WFSs which operate on sophisticated, heavily nested INSPIRE GML models. The resulting INSPIRE Linked Data is created regarding INSPIRE-similarities to LOD concerning syntax, identification and referencing [11] and based on ontologies (schema-level) which is not quite usual for LOD-data [4].

7 Conclusions and Outlook

This paper presents an approach to enable federated queries over INSPIRE and LOD. We identify two key tasks, INSPIRE ontologies and support for querying, and propose how to solve them. Our solution has been tested with a prototype and turned out to be viable and efficient while establishing a semantic query layer for arbitrary WFS-services not constrained to INSPIRE data publishing.

Open issues are investigations in linking this INSPIRE- with other LOD-data, e.g. the data about European directives from the "Reporting Obligations

¹⁸ <http://inspire.jrc.ec.europa.eu/index.cfm/pageid/44>

¹⁹ <http://www.openrdf.org/>

²⁰ <http://vtd-xml.sourceforge.net/>

Database”²¹, or the full integration of spatial reasoning as soon as the specification GeoSPARQL will become stable. One big challenge will be a coordinated Semantic Web-infrastructure for INSPIRE reference data and management facilities of INSPIRE ontologies.

The Semantic Web is about handling information resources directly so we focus on the WFS-service and GML-encoded geodata itself, not the metadata layer consisting of OGC Catalogue Services and ISO-metadata. But INSPIRE also builds up a vivid metadata layer so that a combination of metadata and geodata forming one harmonized distributed Semantic Web-graph will certainly be a matter for future work.

References

1. Bedini, I., Gardarin, G. & Nguyen, B. [2008]. Deriving ontologies from XML schema. in Proc. of EDA 2008. Vol. B-4. pp. 3-17.
2. Bohring, H. & Auer, S. [2005]. Mapping XML to OWL ontologies. in Leipziger Informatik-Tage. pp. 147-156.
3. Gomes Jr, L. C. & Medeiros, C. B. [2007]. Ecologically-aware queries for biodiversity research. in IX Brazilian Symp. on GeoInformatics. pp. 73-84.
4. Jain, P., Hitzler, P., Yeh, P., Verma, K. & Sheth, A. [2010]. Linked data is merely more data. in H. H. D. Brickley, V. K. Chaudhri & D. McGuinness, eds, *Linked Data Meets Artificial Intelligence*. AAAI Press. Menlo Park, CA. pp. 82-86.
5. Janowicz, K., Schade, S., Bröring, A., Keßler, C., Maué, P. & Stasch, C. [2010]. Semantic enablement for spatial data infrastructures. in *Trans. in GIS*. Vol. 14(2).
6. Keßler, C. and Janowicz, K. [2010]. Linking sensor data - Why, to What, and How?, in Proc. of the 3rd Int'l workshop on SSN10, CEUR-WS, vol. 668.
7. Lieberman, J. [2006]. Geospatial Semantic Web Interoperability Experiment Report. OGC Inc. OGC 06-002r1.
8. Page, K., Roure, D. D., Martinez, K., Sadler, J. & Kit, O. [2009]. Linked sensor data: RESTfully serving RDF and GML. in *Second Int'l Workshop on SSN2009*.
9. Patni, H., Henson, C. & Sheth, A. [2010]. Linked sensor data. in *IEEE*, ed., 2010 Int'l Symp. on Collaborative Technologies and Systems. pp. 362-370.
10. Perry, M. & Herring, J. [2011]. OGC GeoSPARQL - A geographic query language for RDF data. OGC Inc. OGC 11-052r3. Url: http://portal.opengeospatial.org/files/?artifact_id=44722
11. Schade, S., Cox, S., Panho, M., Santos, M. & Pundt, H. [2010]. Linked data in SDI or how GML is not about trees. in Proc. of the 13th AGILE Int'l Conf. on Geographic Information Science - Geospatial Thinking.
12. Schenk, S., Staab, S. [2008]. Networked Graphs: a Declarative Mechanism for SPARQL Rules, SPARQL Views and RDF Data Integration on the Web. In: Proc. Int'l. WWW Conf., New York, NY, USA, ACM, pp. 585-594.
13. Zhao, T., Zhang, C., Wei, M. & Peng, Z. [2008]. Ontology-Based geospatial data query and integration. in T. Cova, H. Miller, K. Beard, A. Frank & M. Goodchild, eds, *Geographic Information Science*. Vol. 5266 of LNCS. Springer Berlin. Heidelberg.

²¹ <http://rod.eionet.europa.eu/void.rdf>

An Architecture for Semantic Analysis in Geospatial Dynamics

Jan Oliver Wallgrün¹ and Mehul Bhatt²

¹ Spatial Information Science and Engineering, University of Maine, USA

² Spatial Cognition Research Center (SFB/TR 8), University of Bremen, Germany
 {wallgruen, bhatt}@informatik.uni-bremen.de

Abstract. We present the conceptual and operational overview and architecture of a framework for semantic – high-level, qualitative – reasoning about dynamic geospatial phenomena. The framework is based on advances in the areas of geospatial semantics, qualitative spatio-temporal representation and reasoning, and reasoning about actions and change. We present the main operational modules, namely the modules for data qualification and consistency, qualitative spatial data integration and conflict resolution, and high-level explanatory analysis.

Keywords: geographic information systems; spatio-temporal dynamics; events and objects in GIS; geospatial analysis

1 Introduction

Geographic information systems (GIS) and geospatial web applications are confronted with massive quantities of micro and macro-level spatio-temporal data consisting of precise measurements pertaining to environmental features, aerial imagery, and more recently, sensor network databases that store real-time information about natural and artificial processes and phenomena. In many applications multiple such data sets need to be combined on the fly in order to provide an adequate basis for high-level spatio-temporal analysis. Within next-generation GIS systems, the fundamental information theoretic modalities are envisioned to undergo radical transformations: high-level ontological entities such as *objects*, *events*, *actions* and *processes* and the capability to model and reason about these are expected to be a native feature of next-generation GIS [27]. Indeed, one of the crucial developmental goals in GIS systems of the future is a fundamental paradigmatic shift in the underlying ‘*spatial informatics*’ of these systems.

The spatial information theoretic challenges underlying the development of high-level analytical capability in dynamic GIS consist of fundamental representational and computational problems pertaining to: the semantics of spatial occurrences, practical abduction in GIS, problems of data qualification and consistency, and spatial data integration and conflict resolution. Research in the area of geospatial semantics, taxonomies of geospatial events and processes, and basic

ontological research into the nature of processes in a specific geospatial context has garnered specific interest from several quarters [3, 9, 19, 21, 22, 23, 30, 33]. Research has mainly been spurred by the realization that purely snapshot-based temporal GIS do not provide for an adequate basis for analyzing spatial events and processes and performing spatio-temporal reasoning. Event-based and object-level reasoning at the spatial level can serve as a basis of explanatory analyses within a GIS [13, 18, 26, 32]. Advances in formal methods in the areas of qualitative spatio-temporal representation and reasoning [11], reasoning about actions and change, and spatio-temporal dynamics [4, 8] provide interesting new perspectives for the development of the foundational spatial informatics underlying next-generation GIS systems.

Building on these existing foundations from the GIS and AI communities, we propose an overarching formal framework, and its corresponding conceptual architecture, for high-level qualitative modeling and analysis for the domain of geospatial dynamics. The input is assumed to consist of data sets from several data sources and the framework encompasses modules for different aspects such as *qualification*, *spatial consistency*, *data merging and integration*, and *explanatory reasoning* within a logical setting.

We give a brief overview of the proposed architecture in the next section and then describe and discuss the different components in detail in the following sections. In doing so, we address basic representational and computational challenges within the formal theory of space, events, actions and change.

2 Geospatial Analytics: A Formal Framework

In the following, we propose and explain a formal framework and its corresponding conceptual architecture for high-level qualitative modeling and (explanatory) analysis for the domain of geospatial dynamics.

Fig. 1 illustrates the architecture with its different modules, which we explain in detail in Sections (3–5). The main aspects of the proposed architecture are the following: The input consists of data sets from several data sources such as remote sensing, spatial databases, sensors etc. These data sets are then processed to derive qualitative spatial observations associated with specific time points to be handed over to the actual analytical reasoning component. This preprocessing is done by a module responsible for partitioning the input data into time points and merging data associated with the same time point including the resolution of spatial conflicts between the different data sources and wrt. given spatial integrity constraints. This module is supported by other modules for performing qualification and spatial consistency checking. The pre-processed temporally-ordered observations constitute configurational and narrative descriptions and serve as the input to the reasoning component, which embeds in the capability to perform explanatory reasoning. The knowledge derived by the reasoning component for a particular domain under consideration can be utilized by ex-

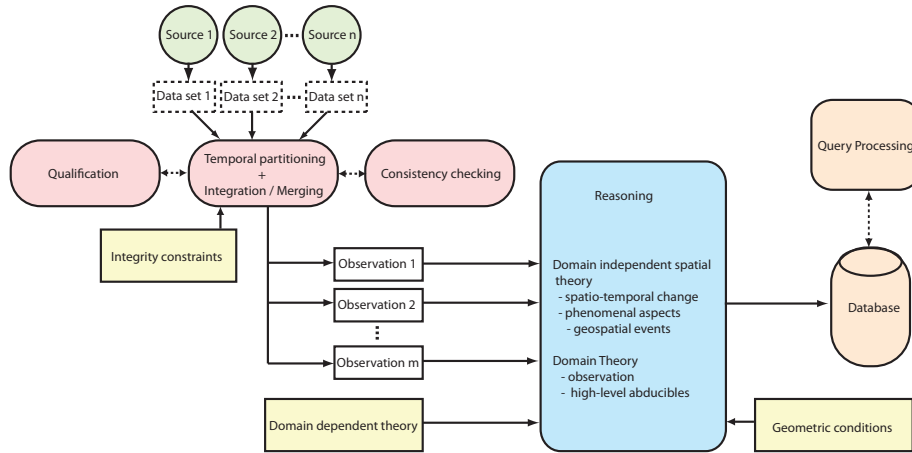


Fig. 1: Conceptual architecture for high-level modeling and explanatory analysis.

ternal services (e.g., query-based services) and application systems that directly interface with humans (e.g., experts, decision-makers).

3 Qualification and Spatial Consistency

Logical frameworks for performing explanation with spatial information generally require that the input information is consistent, meaning that the combined input data is compliant with the underlying logical spatial theory. However, in the geographic domain, the input data often stems from multiple sources, for instance from different sensors, remote sensing data, map data, etc., and the data itself is afflicted by measurement errors and uncertainty. Hence, the geo-referenced quantitative input data about spatial objects needs to be pre-processed in order to be used to perform explanation on a level of qualitative spatial relations. This preprocessing involves the temporal partitioning of the input data into an ordered sequence of time points and the formulation of consistent qualitative descriptions called *observations* for each time point. Crucial sub-components involved in the generation of these descriptions are modules for translating geo-referenced quantitative data into relations from several qualitative spatial models dealing with different aspects of space, a process referred to as *qualification*, and for checking the consistency of the combined information. Both modules are utilized by the main preprocessing module responsible for qualitative integration including the resolution of contradictions as explained further in the next section.

The qualification procedure needs to consider all data that concerns the same moment in time and compute relations for all n -tuples of objects where n corresponds to the arity of the relations in the given qualitative model (e.g., binary topological relations such as *contained* or *disjoint*, or cardinal directions rela-



Fig. 2: Information from four different sources which is inconsistent when combined.

tions such as *north-of*). If uncertainty of quantitative information is explicitly represented this needs to be taken into account and may lead to disjunctions on the qualitative level.

Due to the mentioned measurement errors and uncertainty of the quantitative input data, the qualitative descriptions resulting from qualification for particular moments in time may contain contradictions or violate integrity constraints stemming from background knowledge about the application domain. Fig. 2 illustrates the case of a spatial inconsistency on the level of topological relations when combining the information from four different sources (all concerning the same time period): From combining the fact that objects *c* and *d* (e.g., two climate phenomena) are reported to *overlap* by one source (a) with the reported relations *c* is completely *contained in a* (b) and *d* is completely *contained in b* (c) (let us say *a* and *b* are two neighbored states) it follows that the two states *a* and *b* would need to overlap as well. This contradicts the information from the fourth source (d) which could for instance be a spatial databases containing state boundaries (or alternatively be given in the form of a general integrity constraint).

As a result of the possibility of inconsistent input information occurring in geographic applications, frameworks for explanation and spatio-temporal analysis need the ability to at least detect these inconsistencies in order to exclude the contradicting information or, as a more appropriate approach, resolve the contradictions in a suitable way. Deciding consistency of a set of qualitative spatial relations has been studied as one of the fundamental reasoning tasks in qualitative spatial representation and reasoning [11]. The complexity of deciding consistency varies significantly over the different existing qualitative calculi. For most common qualitative calculi such as the Region Connection Calculus (RCC-8) [29], the consistency can be decided in cubic time when the input description is a scenario which means it does not contain disjunction. This is achieved by the path consistency or algebraic closure method [24]. For general descriptions including disjunctions a more costly backtracking search has to be performed.

Integrity constraints have been investigated in the (spatial) database literature [10, 16]. As the example above shows, in a geographic context, integrity rules often come in the form of qualitative spatial relations that have to be satisfied by certain types of spatial entities. These kinds of spatial integrity constraints can be dealt with by employing terminological reasoning to determine whether a certain integrity rule has to be applied to a given tuple of objects and feeding

the resulting constraints into a standard qualitative consistency checker together with the qualitative relations coming from the input data.

4 Spatial Data Integration and Conflict Resolution

When conflicts arise during the integration of spatial data, it is desirable to not only detect the inconsistencies but also resolve conflicts in a reasonable manner to still be able to exploit all provided information in the actual logical reasoning approach for explanation and analysis. Methods for data integration and conflict resolution have for instance been studied under the term information fusion [20]. Symbolic information fusion is concerned with the revision of logical theories under the presence of new evidence. Different information fusion settings have led to the formulation of different rationality criteria that corresponding computational approaches should satisfy such as the AGM postulates for belief change [1]. Such computational solutions often consist of merging operators that compute a consistent model that is most similar to the inconsistent input data. In distance-based merging approaches this notion of similarity is described using a distance measure between models. This idea has been applied to qualitative spatial representations [12, 14] using the notion of conceptual neighborhood [15, 17] to measure distance in terms of the number of neighborhood changes that need to be performed to get from inconsistent qualitative descriptions to consistent ones.

Fig. 3 shows an example from the domain of urban dynamics that illustrates the role of integration with conflict resolution as well as qualification and consistency checking. Let us assume that we have spatial data from different sources: Source 1 provides information about different land use zones including parks, residential zones, industrial zones, which are derived analyzing aerial images. Source 2 provides information about natural reservoirs, that is about parks and mangroves, stemming from a spatial database. Let us furthermore assume that the land use types are defined in a mutually exclusive way such that two different zones cannot overlap. We now follow the procedure for integrating this information sketched in Alg. 1 that takes a set of observations \mathcal{O} , one for each source, containing object identifiers with associated geometries and a set \mathcal{IC} of integrity constraints. Fig. 3(a) illustrates part of the combined information from all sources for a particular time point t . Source 1 and source 2 both contain geo-referenced polygons for a park but this information does not match. The first step now is to qualify the geometric data from source 1 and 2 which results in the qualitative constraint network Q .³ Using RCC-8 this network looks as shown in Fig. 3(b) (p and p' represent the different geometries for the same park object). If network Q is consistent and compliant with the integrity constraints, the result can directly be handed over as an observation to the reasoning module. However, as also shown in Fig. 3(b) this is not the case as integrity constraints

³ Alternatively, information for each data set could be qualified separately resulting in several constraint networks that have to be combined by a suitable merging operator.

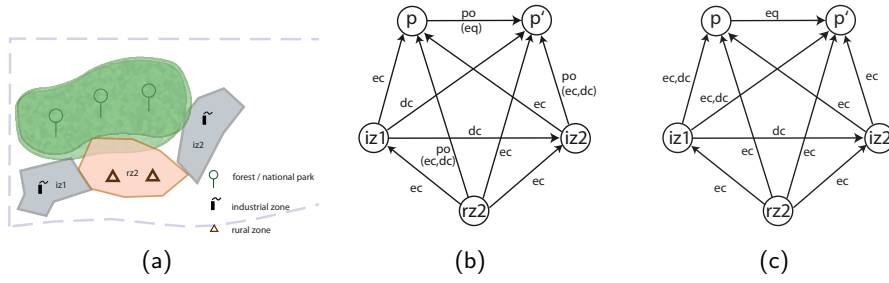


Fig. 3: Qualification of the combined geometric information (a) together with the integrity constraints results in an inconsistent qualitative model (b). The consistent model after resolving the conflicts (c).

are violated in three places indicated by listing possible relations following from the integrity constraint in brackets below the original relation. The relation between p and p' should be eq simply because it is known that both represent the same object. The relation between rz_2 and p should be either ec or dc because of our integrity constraint, and the same holds for the relation between p' and iz_2 . Therefore, the qualitative conflict resolution component needs to be called to find a qualitative representation that is as close as possible to the network from Fig. 3(b) but is overall consistent.

To achieve the conflict resolution, a resolution operator Λ based on the idea of distance-based merging operators for qualitative spatial representations [12, 14] is applied to Q . Our resolution operator Λ is based on a distance measure $d(s, s')$ between two scenarios over the same set of objects. It is computed by simply summing up the distance of two base relations in the conceptual neighborhood graph of the involved calculus given by $d_B(C_{ij}, C'_{ij})$ over all corresponding constraints C_{ij}, C'_{ij} in the input scenarios:

$$d(s, s') = \sum_{1 \leq i < j \leq m} d_B(C_{ij}, C'_{ij}) \quad (1)$$

The resolved network $\Lambda(Q)$ is then constructed by taking the union of those scenarios that are consistent, compliant with the integrity constraints and have a minimal distance to Q according to $d(s, s')$ ⁴:

$$\Lambda(Q) = \bigcup_{s \in S(Q)} s \quad (2)$$

with

$$S(Q) = \{s \in \llbracket QCN \rrbracket \mid \forall s' \in \llbracket QCN \rrbracket : d(s', Q) \geq d(s, Q)\} \quad (3)$$

and $\llbracket QCN \rrbracket$ standing for the set of all scenarios that are consistent and compliant with the integrity constraints. Following the approach described in [14], $\Lambda(Q)$

⁴ Taking the union here means we build a new network by taking the union of all corresponding constraints.

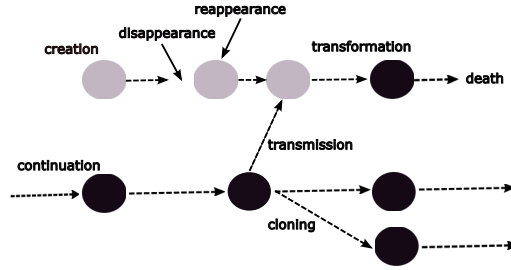


Fig. 4: Object Change History, Source: [32]

can be computed by incrementally relaxing the constraints until at least one consistent scenario has been found. This is illustrated in Alg. 2 where we assume that the function $\text{relax}(Q, i)$ returns the set of scenarios s which have a distance $d(s', Q) = i$ to Q .

The result of applying the resolution operator to the network from Fig. 3(b) is shown in Fig. 3(c): Both violations of integrity constraints have been resolved by assuming that instead of 'overlap' the correct relation is 'externally connected'. Interestingly, the resulting consistent qualitatively model contains two disjunctions basically saying that the relation between the park and iz_1 is either ec or dc. This is a consequence of the fact that both qualitative models are equally close to the input model such that it is not possible to decide between the two hypotheses.

Algorithm 1:
Qualify+Merge($\mathcal{O}, \mathcal{IC}$)

```

 $Q \leftarrow \text{qualify}(\mathcal{O})$ 
if  $\neg \text{consistent}(Q, \mathcal{IC})$  then
   $Q \leftarrow \Lambda(Q, \mathcal{IC})$ 
end if
return  $Q$ 

```

Algorithm 2: $\Lambda(Q, \mathcal{IC})$

```

 $i \leftarrow 0, N \leftarrow \emptyset$ 
while  $N = \emptyset$  do
   $R \leftarrow \text{relax}(Q, i)$ 
  for  $r \in R$  do
    if  $\text{consistent}(r, \mathcal{IC})$  then  $N \leftarrow N \cup r$ 
  end if
  end for
   $i \leftarrow i + 1$ 
end while
return  $N$ 

```

5 Analyses with Events and Objects

Our objective for the high-level reasoning module is to provide the functionality to enable reasoning about spatio-temporal narratives consisting of events and processes at the geographic scale. We do not attempt an elaborate ontological characterization of events and processes, a topic of research that has been addressed in-depth in the state-of-the-art (see Section 1). For the purposes of this paper, we utilize a minimal, yet rich, conceptual model consisting of a range of events such that it may be used to qualitatively ground metric geospatial

datasets consisting of spatial and temporal footprints of human and natural phenomena at the geographic scale.

From an ontological viewpoint, spatial occurrences may be defined at two levels: (O1) *domain-independent*, and (O2) *domain-dependent*:

O1. Domain Independent Spatial Occurrences These occurrences are those that may be semantically characterized within a general theory of space and spatial change. These may be grounded with respect to either a qualitative theory, or an elaborate typology of geospatial events. Distinctions as per (A–B) are identifiable:

A. Spatial Changes at a Qualitative Level

In so far as a general qualitative theory of spatial change is concerned, there is only one type of occurrence, viz - a transition from one qualitative state (relation) to another as (possibly) governed by the continuity constraints of the relation space. At this level, the only identifiable notion of an occurrence is that of a qualitative spatial transition that the primitive objects in the theory undergo. At the level of a spatial theory, it is meaningless to ascribe a certain spatial transition as being an event or action; such distinctions demand a slightly higher level of abstraction. For example, the transition of an object (o_1) from being *disconnected* to another object (o_2) to being a *tangential* – part of it could either coarsely represent the volitional movement of a person into a room or the motion of a ball. Whereas the former is an action performed by an agent, the latter is a deterministic event that will necessarily occur in normal circumstances. Our standpoint here is that such distinctions can only be made in a domain specific manner; as such, the classification of occurrences into actions and events will only apply at the level of the domain with the general spatial theory dealing only with one type of occurrence, namely primitive spatial transitions that are definable in it.

B. Typology of Events and Patterns

At the domain independent level, the explanation may encompass behaviours such as *emergence, growth & shrinkage, disappearance, spread, stability* etc, in addition to the sequential/parallel composition of the behavioural primitives aforementioned, e.g., *emergence* followed by *growth, spread / movement, stability* and *disappearance* during a time-interval. Certain kinds of typological elements, e.g., *growth* and *shrinkage*, may even be directly associated with spatial changes at the qualitative level in (A).

Appearance of new objects and disappearance of existing ones, either abruptly or explicitly formulated in the domain theory, is also characteristic of non-trivial dynamic (geo)spatial systems. Within event-based GIS, appearance and disappearance events are regarded to be an important typological element for the modeling of dynamic geospatial processes [9, 32]. For instance, Claramunt and Thériault [9] identify the basic processes used to define a set of low-order spatio-

temporal events which, among other things, include appearance and disappearance events as fundamental. Similarly, toward event-based models of dynamic geographic phenomena, Worboys [32] suggests the use of the appearance and disappearance events at least in so far as single object behaviours are concerned (see Fig. 4). Appearance, disappearance and re-appearances are also connected to the issue of object identity maintenance in GIS [3, 22].

O2. Domain-Specific Spatial Occurrences At a domain-dependent level, behaviour patterns may characterize high-level processes, environmental / natural and human activities such as *deforestation*, *urbanization*, *land-use transformations* etc. These are domain-specific occurrences that induce a transformation on the underlying spatial structures being modeled. Basically, these are domain specific events or actions that have (explicitly) identifiable occurrence criteria and effects that can be defined in terms of qualitative spatial changes, and the fundamental typology of spatial changes. For instance, in the example in Fig. 5, we can clearly see that region *a* has continued to *shrink* over a three-decade period, followed by a *split*, and eventually *disappearing* in the year 1990.

The following general notion of a ‘*spatial occurrence*’ is identifiable [6]:

‘Spatial occurrences are events or actions with explicitly specifiable occurrence criteria and/or pre-conditions respectively and effects that may be identified in terms of a domain independent taxonomy of spatial change that is native to a general qualitative spatial theory’.

As an example, consider an event that will *cause* a region to *split* or make it *grow / shrink*. Likewise, an aggregate cluster of geospatial entities (e.g., in wildlife biology domain) may *move* and change its orientation with respect to other geospatial entities. Thinking in agent terms, a spatial action by the *collective / aggregate* entity, e.g., *turn south-east*, will have the effect of changing the orientation of the cluster in relation to other entities. In certain situations, there may not be a clearly identifiable set of domain-specific occurrences with explicitly known occurrence criteria or effects that are definable in terms of a typology of spatial change, e.g., cluster of alcohol-related crime abruptly appearing and disappearing at a certain time. However, even in such situations, an analysis of the domain-independent events and inter-event relationships may lead to an understanding of spatio-temporal relationships and help with practical hypothesis generation [2].

Explanatory Reasoning in GIS: A Case for Practical Abduction

Explanatory reasoning requires the ability to perform abduction with spatio-temporal information. In the context of formal spatio-temporal calculi, and logics of action and change, this translates to the ability to provide *scenario and narrative completion* abilities at a high-level of abstraction.

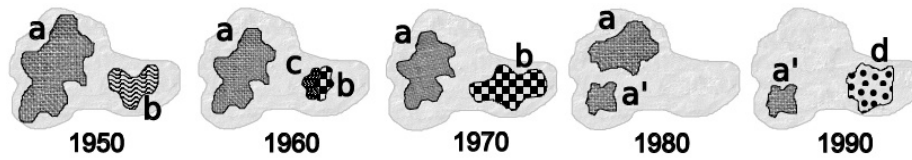


Fig. 5: Abduction in GIS

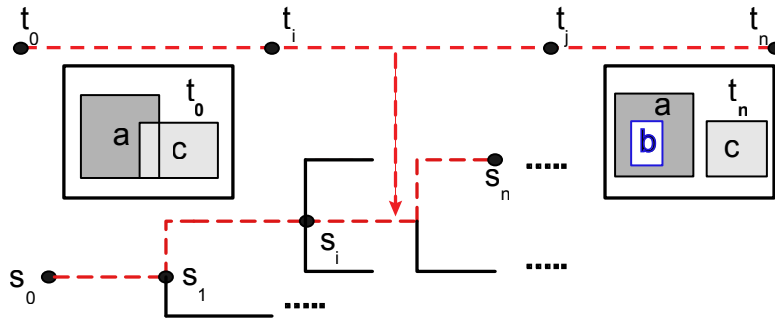


Fig. 6: Branching / Hypothetical Situation Space. Source [4]

Consider the GIS domain depicted in Fig. 5, and the basic conceptual understanding of spatial occurrences described in (O1–O2; Section 5). At a domain-independent level, the scene may be described using topological and qualitative size relationships. Consequently, the only changes that are identifiable at the level of the spatial theory are *shrinkage*, *splitting*, and eventual *disappearance* – this is because a domain-independent spatial theory may only include a generic typology (*appearance*, *disappearance*, *growth*, *shrinkage*, *deformation*, *splitting*, *merging* etc) of spatial change. However, at a domain-specific level, these changes could characterize a specific event (or process) such as *deforestation*. The hypotheses or explanations that are generated during an explanation process should necessarily consist of the domain-level occurrences in addition to the underlying (associated) spatial changes (as per the generic typology) that are identifiable. Intuitively, the derived explanations more or less take the form of existential statements such as: “Between time-points t_i and t_i , the process of deforestation is abducible as one potential hypothesis”. Derived hypotheses / explanations that involve both domain-dependent and as well their corresponding domain-independent typological elements are referred to as being ‘adequate’ from the viewpoint of explanatory analysis for a domain. At both the domain-independent as well as dependent levels, abduction requires the fundamental capability to interpolate missing information, and understand partially available narratives that describe the execution of high-level real or abstract processes. In the following, we present an intuitive overview of the scenario and narrative completion process.

Scenario and Narrative Completion Explanation, in general, is regarded as a converse operation to temporal projection essentially involving reasoning from

effects to causes, i.e., reasoning about the past [31]. Logical abduction is one inference pattern that can be used to realize explanation in the spatio-temporal domain [5, 6].

Explanation problems demand the inclusion of a narrative description, which is essentially a distinguished course of actual events about which we may have incomplete information [25, 28]. Narrative descriptions are typically available as *observations* from the real / imagined execution of a system or process. Since narratives inherently pertain to actual observations, i.e., they are *temporalized*, the objective is often to assimilate / explain them with respect to an underlying process model and an approach to derive explanations.

Given the set of observations resulting from the preprocessing which constitutes a partial narrative of the evolution of a system in terms of high-level spatio-temporal data, scenario and narrative completion corresponds to the ability to derive completions that bridge the narrative by interpolating the missing spatial and action / event information in a manner that is consistent with domain-specific and domain-independent rules / dynamics. Consider the illustration in Fig. 6 for a branching / hypothetical situation space that characterizes the complete evolution of a system [5]. In Fig. 6 – the situation-based history $\langle s_0, s_1, \dots, s_n \rangle$ represents one path, corresponding to an actual time-line $\langle t_0, t_1, \dots, t_n \rangle$, within the overall branching-tree structured situation space. Given incomplete narrative descriptions, e.g., corresponding to only some ordered time-points in terms of high-level spatial (e.g., topological, orientation) and occurrence information, the objective of causal explanation is to derive one or more paths from the branching situation space, that could best-fit the available narrative information [6]. Of course, the completions that bridge the narrative by interpolating the missing spatial and action/event information have to be consistent with domain-specific and domain-independent rules/dynamics.

6 Conclusion

In our research, we are addressing a broad question: what constitutes the (core) spatial informatics underlying (specific kinds) of analytical capabilities within a range of dynamic geospatial domains [7]. In continuation with the overarching agenda described in [7], this paper has demonstrated the fundamental challenges and presented solutions thereof encompassing aspects such as *spatial consistency*, *data merging and integration*, and *practical geospatial abduction* within a logical setting. Whereas independently implemented modules for these respective components have been developed in our projects at the Spatial Cognition Research Center, the main thrust of our ongoing work in the current context is to fully implement the integrated framework / architecture described in this paper.

Acknowledgements. This research has partially been financed by the Deutsche Forschungsgemeinschaft under grants SFB/TR 8 Spatial Cognition and IRTG GRK 1498 Semantic Integration of Geospatial Information. This article builds on, and complements a short paper & poster at the COSIT 2011 conference [7].

References

- [1] C. E. Alchourron, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985.
- [2] A. Beller. Spatio/temporal events in a GIS. In *Proceedings of GIS/LIS*, pages 766–775. ASPRS/ACSM, 1991.
- [3] B. Bennett. Physical objects, identity and vagueness. In D. Fensel, F. Giunchiglia, D. L. McGuinness, and M.-A. Williams, editors, *KR*, pages 395–408. Morgan Kaufmann, 2002.
- [4] M. Bhatt. Reasoning about space, actions and change: A paradigm for applications of spatial reasoning. In *Qualitative Spatial Representation and Reasoning: Trends and Future Directions*. IGI Global, USA, 2010.
- [5] M. Bhatt and G. Flanagan. Spatio-temporal abduction for scenario and narrative completion. In M. Bhatt, H. Guesgen, and S. Hazarika, editors, *Proceedings of the International Workshop on Spatio-Temporal Dynamics, co-located with the European Conference on Artificial Intelligence (ECAI-10)*, pages 31–36. ECAI Workshop Proceedings., and SFB/TR 8 Spatial Cognition Report Series, August 2010. URL <http://www.cosy.informatik.uni-bremen.de/events/ecai10/>.
- [6] M. Bhatt and S. Loke. Modelling dynamic spatial systems in the situation calculus. *Spatial Cognition and Computation*, 8(1):86–130, 2008.
- [7] M. Bhatt and J. O. Wallgruen. Analytical intelligence for geospatial dynamics. In *Proceedings of COSIT 2011: Conference on Spatial Information Theory*, 2011.
- [8] M. Bhatt, H. Guesgen, S. Woelfl, and S. Hazarika. Qualitative Spatial and Temporal Reasoning: Emerging Applications, Trends and Directions. *Journal of Spatial Cognition and Computation*, 11(1), 2011.
- [9] C. Claramunt and M. Thériault. Managing time in GIS: An event-oriented approach. In J. Clifford and A. Tuzhilin, editors, *Recent Advances on Temporal Databases*, pages 23–42. Springer, 1995.
- [10] S. Cockcroft. A taxonomy of spatial data integrity constraints. *GeoInformatica*, 1:327–343, 1997.
- [11] A. G. Cohn and J. Renz. Qualitative spatial reasoning. In F. van Harmelen, V. Lifschitz, and B. Porter, editors, *Handbook of Knowledge Representation*. Elsevier, 2007.
- [12] J.-F. Condotta, S. Kaci, and N. Schwind. A framework for merging qualitative constraints networks. In D. Wilson and H. C. Lane, editors, *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference, May 15-17, 2008, Coconut Grove, Florida, USA*, pages 586–591. AAAI Press, 2008.
- [13] H. Couclelis. The abduction of geographic information science: Transporting spatial reasoning to the realm of purpose and design. In K. S. Hornsby, C. Claramunt, M. Denis, and G. Ligozat, editors, *COSIT*, volume 5756 of *Lecture Notes in Computer Science*, pages 342–356. Springer, 2009.
- [14] F. Dylla and J. O. Wallgrün. Qualitative spatial reasoning with conceptual neighborhoods for agent control. *Journal of Intelligent and Robotic Systems*, 48(1): 55–78, 2007.
- [15] M. J. Egenhofer and K. K. Al-Taha. Reasoning about gradual changes of topological relationships. In *Proceedings of the International Conference GIS - From Space to Territory: Theories and Methods of Spatio-Temporal Reasoning on Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, pages 196–219, London, UK, 1992. Springer-Verlag.

- [16] R. Fagin and M. Y. Vardi. The theory of data dependencies - an overview. In J. Paredaens, editor, *ICALP*, volume 172 of *Lecture Notes in Computer Science*, pages 1–22. Springer, 1984.
- [17] C. Freksa. Conceptual neighborhood and its role in temporal and spatial reasoning. In M. Singh and L. Travé-Massuyès, editors, *Decision Support Systems and Qualitative Reasoning*, pages 181 – 187. 1991.
- [18] A. Galton and J. Hood. Qualitative interpolation for environmental knowledge representation. In *ECAI*, pages 1017–1018, 2004.
- [19] A. Galton and R. Mizoguchi. The water falls but the waterfall does not fall: New perspectives on objects, processes and events. *Applied Ontology*, 4(2):71–107, 2009.
- [20] É. Grégoire and S. Konieczny. Logic-based approaches to information fusion. *Information Fusion*, 7(1):4–18, 2006.
- [21] P. Grenon and B. Smith. Snap and span: Towards dynamic spatial ontology. *Spatial Cognition and Computation*, 4(1):69–104, 2004.
- [22] K. Hornsby and M. J. Egenhofer. Identity-based change: A foundation for spatio-temporal knowledge representation. *International Journal of Geographical Information Science*, 14(3):207–224, 2000.
- [23] K. S. Hornsby and S. J. Cole. Modeling moving geospatial objects from an event-based perspective. *T. GIS*, 11(4):555–573, 2007.
- [24] A. Mackworth. Consistency in networks of relations. *Artificial Intelligence*, 8(1):99–118, 1977.
- [25] R. Miller and M. Shanahan. Narratives in the situation calculus. *J. Log. Comput.*, 4(5):513–530, 1994.
- [26] G. D. Mondo, J. G. Stell, C. Claramunt, and R. Thibaud. A graph model for spatio-temporal evolution. *J. UCS*, 16(11):1452–1477, 2010.
- [27] NIMA. National Imagery and Mapping Agency, The Big Idea Framework, 2000.
- [28] J. Pinto. Occurrences and narratives as constraints in the branching structure of the situation calculus. *J. Log. Comput.*, 8(6):777–808, 1998.
- [29] D. A. Randell, Z. Cui, and A. Cohn. A spatial logic based on regions and connection. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference*, pages 165–176. Morgan Kaufmann, 1992.
- [30] A. Renolen. Modelling the real world: Conceptual modelling in spatiotemporal information system design. *Transactions in GIS*, 4:23–42(20), January 2000.
- [31] M. Shanahan. Prediction is deduction but explanation is abduction. In *IJCAI*, pages 1055–1060, 1989.
- [32] M. F. Worboys. Event-oriented approaches to geographic phenomena. *International Journal of Geographical Information Science*, 19(1):1–28, 2005.
- [33] M. F. Worboys and K. Hornsby. From objects to events: Gem, the geospatial event model. In M. J. Egenhofer, C. Freksa, and H. J. Miller, editors, *GIScience*, volume 3234 of *Lecture Notes in Computer Science*, pages 327–344. Springer, 2004.

Making *close to* Suitable for Web Search

A Comparison of Two Approaches

Iris Helming¹, Abraham Bernstein², Rolf Grütter¹, and Stephan Vock³

¹ Swiss Federal Research Institute WSL, Birmensdorf, Switzerland {iris.helming,
rolf.gruetter}@wsl.ch

² University of Zurich, Department of Informatics Zurich, Switzerland
bernstein@ifi.uzh.ch

³ stephan.vock@gmail.com

Abstract. In this paper we compare two approaches to model the vague german spatial relation *in der Nähe von* (English: "close to") to enable its usage in (semantic) web searches. A user wants, for example, to find all relevant documents regarding parks or forestal landscapes *close to* a city. The problem is that there are no clear metric distance limits for possibly matching places because they are only restricted via the vague natural language expression. And since human perception does not work only in distances we can't handle the queries simply with metric distances. Our first approach models the meaning of these expressions in description logics using relations of the Region Connection Calculus. A formalism has been developed to find all instances that are potentially perceived as *close to*. The second approach deals with the idea that everything that can be reached in a reasonable amount of time with a given means of transport (e.g. car) is potentially perceived as *close*. This approach uses route calculations with a route planner. The first approach has already been evaluated. The second is still under development. But we can already show a correlation between what people consider as *close to* and time needed to get there.

Keywords: Vague Spatial Relation, Local Expression, Region Connection Calculus (RCC), Route Planning, Reachability.

1 Introduction

Sometimes we want to search for places on the web and restrict the search results to a specific area. But we don't have an exact distance restriction in mind, we just want to look for something that is *close* or *not close to*, *a bit off* and so on. How can we make this restriction understandable to a search engine? So that future users could simply apply these expressions as keywords without further thinking about "translating" them?

Google⁴ already delivers results for queries that include *near*. But these results show that it's not really taken care of the meaning of the preposition: if

⁴ <http://www.google.ch>

you are looking for a "hotel in Zurich" for example, it returns also hotels which are in the area around Zurich. On the other hand, if disliking living in a city centre but wanting to get there quickly, you could search for "hotels near Zurich". The result will show you hotels near Zurich but also some which are located in the city centre. Also, these mechanisms can't scale with regards to the reference place (i.e. the place to which the first one is supposed to be near). So the area for searching doesn't have a bigger size if the reference place is bigger. With our knowledge representation approach (cf. [2]) such scaling is performed through the type of the administrative region of the reference place — the granularity of found nearby places is decreasing if the referent is situated on a more fine-grained scale, such as a village, or increasing if the referent is on a larger scale, such as a city. In this paper we will present a novel approach for decoding "nearness", which deals with statistical methods. We compare it to our previous knowledge engineering approach. With the new approach scaling works through the chosen means of transport — things that are near while driving a car may not be near while walking. Using this approach one could, in a future implementation, present the user the best results of *nearby* places for his traveling context.

Both approaches are meant to map the vague concepts of spatial relations that occur in natural language onto concrete geographical regions or places.

2 Related Work

Schokaert, De Cock and Kerre [8] suggest augmenting the structured information available to a local search service, such as Google Maps, with information extracted from the web. They show how nearness information in natural language and information about the surrounding neighborhood of a place can be translated into fuzzy restrictions and how such fuzzy restrictions can be used to estimate the location of a place with an unknown address. The vast amount of data addressed by the authors, together with the kinds of examples they provide, suggest that their approach is targeted on mass searches. In our case, the resources on the web, which could possibly be used to augment the searches, are sometimes scarce. Our second approach also is a statistical one, but we are using context information -traveling time- to give the best matching results for a special purpose.

Yao and Thill [9] also follow a statistical approach to handle vague natural language expressions of distance. Different to us, they directly transform their results to distances. We are avoiding this since human perception of vague spatial expressions doesn't work on metrical distances.

Also their discussion of general problems when dealing with vague expressions for distances is of interest for our approach. They are highlighting the importance of the context when a person has to judge if a place is near another place. Among others they name transport mode as an influence factor for the perception of nearness, e.g. Is city A close to city B? Yes for airplane, no for car. With the statistical approach we make use of this influence factor and show how contextual information in terms of means of transport can be modelled.

Mata [5] presents an approach to geographic information retrieval integrating topological, geographical and conceptual matching. For topological matching topological relations are extracted from overlaying data layers; for geographical matching constraints are obtained from dictionaries; for conceptual matching a geographic ontology is used. A constraint defines two geographic objects (points or polygons) as near provided they are connected by a third object (an arc, e.g., a road), the length of which is less than a given distance. Different from the approaches we are comparing, a metric distance measure thus is a necessary condition for nearness, although not a sufficient. However, the framework seems general enough to be aligned with that presented in section 3.

3 Knowledge Representation Approach

In this section we briefly summarize the important aspects of the knowledge representation approach, which we presented in [2]. For modelling nearness with this first approach we use information of the administrative structure of Switzerland, which can be obtained easily. It is freely available as a download for many countries. An administrative region like a district is responsible for administrative tasks like providing schools, medical supply, organizing elections and so on. Often borders for such regions are grown where also natural barriers like big streams exist. It has been shown before that the partitioning of a country into smaller parts has influence on human perception of space. Maki [3] for example showed that the affiliation to a state plays an important role in human perception of locations. In an experiment, subjects had to decide about the location of two cities regarding their orientation east-west. If the cities in question belong to different states, the reaction times were significantly shorter than with cities which belong to the same state. Human beings are able to judge faster about entities on a continuum if they can make use of category information. Among others, Carbon and Leder [1] showed that the membership to different political systems, structures or hierarchies influences the estimation of distance between two cities. They used estimation tasks for distances between cities east and west of the former border inside of Germany. Compared to pairs inside the same part of the former republic, distances were overestimated if the cities in question belonged to different parts.

Topologies of regions can be relatively easy obtained via modern geographic information systems (GIS) or spatial databases. Types of administrative regions and topological relations between them provide us with the possibility to reason on these regions as polygons. Randell, Cui and Cohn [7] developed a set of spatial relations in first-order logic, the Region Connection Calculus (RCC), which we use for it. How this works is described in the next section. With this approach we provide a qualitative method for qualitative search queries. As Mark and Egenhofer already conclude metric is not the most important parameter of the semantics of most spatial natural-language expressions⁵.

⁵ "The topological relations come out as the strongest discriminators approximately 22 times stronger than all metric parameters combined which confirms the under-

3.1 Methods for the Knowledge Representation Approach

Knowledge is organized in a sample OWL DL Knowledge Base KB, consisting of a TBox T and an ABox A. Partitions of regions are represented in T, partially ordered in a way that each element of a partition is a subset of an element of the next upper level of partitions. Each partition is typed and the concepts for typing are mutually disjoint, so that an individual can only be of one type. So assume you have the partitions $C(x_i)_{i \in I}$ and $D(y_k)_{k \in K}$ of the same region, their types are C and D . $C(x_i)_{i \in I}$ is understood as more fine-grained than $D(y_k)_{k \in K}$ if each element of $C(x_i)_{i \in I}$ is a subset of an element of $D(y_k)_{k \in K}$. For instance, $District(y_k)_{i \in I}$ is partitioned by elements of $Community(x_i)_{i \in I}$ and both are partitions of a canton. PartOf relations are kept functional, which means that regions are only asserted as part of the next upper region but not as part of the region above the next upper hierarchical step. In the ABox A one finds assertions like *partOf(Dietlikon, District-Buelach)*, stating that the individual Dietlikon, which is of the type community, belongs to Bülach, which is a district of type. Further all individuals are asserted as different from each other. The Region Connection Calculus can be used to represent spatial relations in first order logic. There are different sets available (e.g. RCC-3, RCC-5, RCC-8). For our purpose, we are using RCC-8, which means, we are using a RCC set that differentiates 8 relations. You can see the 8 relations in Figure 1.

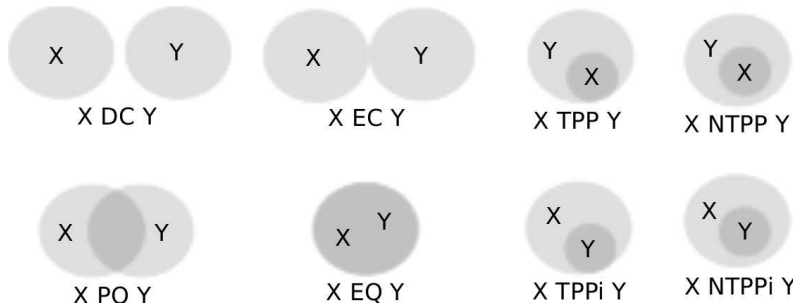


Fig. 1: The Region Connection Calculus with 8 relations

Altogether these relations form a jointly exhaustive and pairwise disjoint set. RCC relations can be interpreted temporal and spatial. Within the spatial interpretation, regions are considered as sets of points. According to that two regions which are connected to each other have at least one point in common. Rules are formulated in a subset of the Semantic Web Rule Language (SWRL)[6]. Our Rule Base is relatively small, in it, existing relations of the Region Connection Calculus (RCC) are used as basis to form Composition Rules. For example there are rules for the additional relation *close to* (cf. [2]).

lying assumption that topology is more critical for the semantics of spatial relations than metric” (Mark and Egenhofer 1994, p. 227 [4]).

From RCC-8, we are using the relations part of P (TPP, NTPP, TPPi, NTPPi), partially overlaps PO and externally connected EC to form the basic relation *close to*. Disjunctions of RCC relations in the bodies of composition rules are represented by auxiliary roles, such as {P, PO} subsuming the roles partOf and partiallyOverlaps. This allows composition rules that are expressed as (non-disjunctive) Horn rules (see equation 1).

3.2 The added Relation *CLOSETO*

To the set of RCC-8 relations we added a composed relation *CLOSETO*. A location x is close to y , stated as $CL(x,y)$. The following equation shows the actual *CLOSETO* rule:

$$\forall x \forall y \forall z [CL_{ap}(y, x) \wedge z \{P, PO\} y \rightarrow CL(z, x)] \quad (1)$$

It is read as region z is close to region x if region y is *a priori* close to x and z is part of or partially overlaps y . This rule makes up the basic building block of this approach. In addition to the basic rule the knowledge representation approach also includes the notion of "a priori"-closeness, which is derived by a second rule (cf. [2]). This second rule enables us to include functional micro regions additionally to the administrative regions into reasoning. These micro regions, consisting of mountane and space planning regions, were introduced to analyze the behavior of commuters. Since these micro regions are also related to traveling their addition seems to be useful for the comparison of the two approaches. For more details please refer to [2].

3.3 Results of the Knowledge Representation Approach

In previous papers it has already been shown that this approach works for the part of Switzerland that is covered by the sample ontology. Also an evaluation has been performed using the search engine "GoForIt"⁶, which provides general search and directory search as shown in [2]. For 170 communities two different kinds of queries were performed. Firstly plain queries, such as "communities close to Dietlikon" and afterwards a query which contained a concatenation of all the communities which have been inferred as "close to", such as "Nürens Dorf OR Dübendorf OR Rümliang OR Wallisellen OR Kloten OR Wangen-Brüttisellen OR Bassersdorf OR Opfikon" (all communities inferred as close to Dietlikon). Finally, the results showed that recall was significantly higher for the rewritten query.

4 Statistical Approach

We will represent now the second approach which is based on the idea that people speak of places as *close to* if they can reach them quickly. Imagine you

⁶ <http://www.goforit.com/>

plan a picnic in a forest nearby and because you have lots of food to take with you you want to go there by car. Then you will speak of a forest you can reach in a reasonable amount of time as *close to*. In terms of metric distance this place could be farer away from your location than another one. But the other place doesn't appear as being *close* to you because it would take longer to go there by car. Sometimes the occurrences of *close to* will differ from one means of transport to another. If you are using a car you can reach things in greater distance easier than while you are walking. On the other hand sometimes you find paths through woods which you can take while you are walking but not if you are driving a car.

4.1 Methods for the Statistical Approach

To gain language data in an appropriate amount the german newspaper text corpora of the Institut für Deutsche Sprache (IDS, Mannheim⁷) were used. Altogether it has a size of 5.3 million tokens. German is used as object language because the application of the approach should start in the German speaking part of Switzerland⁸. The keyword string *in der Nähe von* (i.e. "close to") was looked up in the corpus. Because of the great ambiguity of toponyms and since there are not yet good filters for toponyms available items were annotated manually for the two *close* places. Then all identified place names were looked up in a gazetteer to obtain the coordinates. Geonames⁹ and Swissnames¹⁰ were used for this. The additional inclusion of Swissnames aimed at getting rich data of Switzerland so that we could guarantee the comparability to the first approach which is only implemented for a part of Switzerland yet. Then the coordinates of the identified places were fed into a route planner. The routing API of cloudmade¹¹, which is based on OpenStreetMap¹², served best for this purposes. If the place name was ambiguous, the nearest match was chosen. The routes were calculated for trips by car, bike and walking. To get counter examples also hits for "nicht in der Nähe von" (*not close to*), "weitab von" (*further away from*) and other expressions for counterparts of *close to* were annotated. With these there were some difficulties since they are - except the *not close to* - not direct antonyms to *close to* and often they were used as a subjective statement of how far something is away with regard to some topic. Part of the instances, for example *weit entfernt von* ("far away from"), were often used to neutrally express distance in combination with a metric distance measure. An example would be:

⁷ <http://www.ids-mannheim.de/>

⁸ In other countries other amounts of traveling time maybe felt as *near*. For example, this amount of traveling time for Switzerland may be around 12 minutes. But for larger countries with only few cities this may even be 2 hours. In the future, one could calculate country specific traveling values.

⁹ <http://www.geonames.org/>

¹⁰ <http://www.swisstopo.admin.ch/internet/swisstopo/en/home/products/landscape/toponymy.html>

¹¹ <http://cloudmade.com/>

¹² <http://www.openstreetmap.org/>

10km weit entfernt von ("10km far away from"). And mostly, if places are not near each other, this fact is not explicitly mentioned.

Further there was a technical problem. Sentences like *Frankfurt, weitab von Asien* ("Frankfurt, far away from Asia") occurred. One can imagine, that it would not be difficult to calculate a route from Frankfurt to anywhere, but how to manifest the endpoint for that route in a whole continent like Asia? Therefore such routes could not be calculated. Nevertheless, there are some negative examples and calculated trips by car, bike and walking for them as well. Analogous to the "near-matches", when a name was ambiguous in geonames, we picked the match with farrest distance between the two places.

5 Preliminary Evaluation

The new approach is intended to model *closeness* via reachability with different means of transport (cf. section 4). Reachability is lowered by path obstacles.

5.1 Qualitative Illustration

See Figure 2 for an example where a mountain ridge would force you to make a detour of 3 times the length of the direct path when traveling by car. Because of this detour you would not say that Arosa and Davos, the places in question, are close to each other. The statistical approach can take care of path obstacles like mountains and lakes and missing direct connections between neighbored suburbs and so on. (For the knowledge representation approach these two places would be close to each other when applying the basic rule, because the community of Arosa borders the district of Davos. They would not be close to each other when applying the rule which has been extended for micro regions. Here we have an example where the inclusion of micro regions underlines the reachability.)

5.2 Quantitative Evaluation

For all the 345 pairs of places which occurred in the corpus connected with *in der Nähe von* (English: "close to"), we calculated the route for going by car, going by bike and walking. We collected traveling time and traveling distance. The same holds for the 30 pairs of places which are connected by *nicht in der Nähe von* ("not close to") and the above mentioned (cf. 4.1) synonyms. Then the true distance was also calculated for every pair.

Some data points are quite far away from the rest. A manual check of the 10 sentences in the corpus to the most far away data points showed that there have been mismatches to geonames, because for one of the place names there was only another item in geonames which did not match the place actually meant. Therefore SPSS¹³ was used to draw a histogram of traveling time by car only

¹³ The SPSS software can be obtained under <http://www-01.ibm.com/software/analytics/spss/>

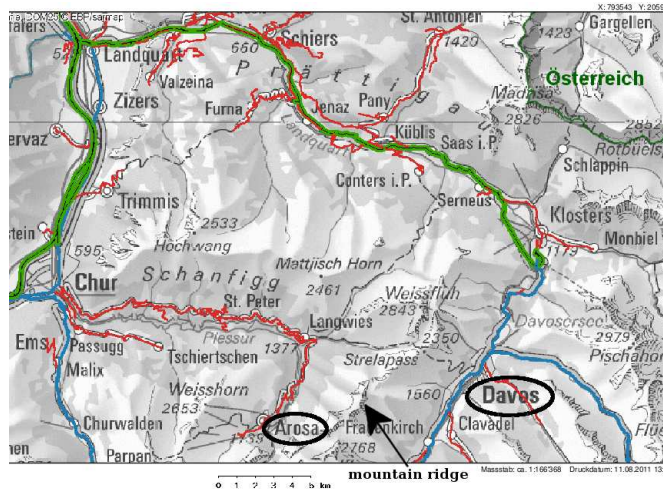


Fig.2: Arosa and Davosa are departed via a mountain ridge. Since there is no direct street over the mountain, a car has to take all the way around the mountains along Chur, Landquart, Küblis and Klosters. Source of Map: Kantonale Verwaltung Graubünden, GIS-Kompetenzzentrum (<http://mapserver1.gr.ch/kantonalesstrassennetz/kantonalesstrassennetz.phtml>)

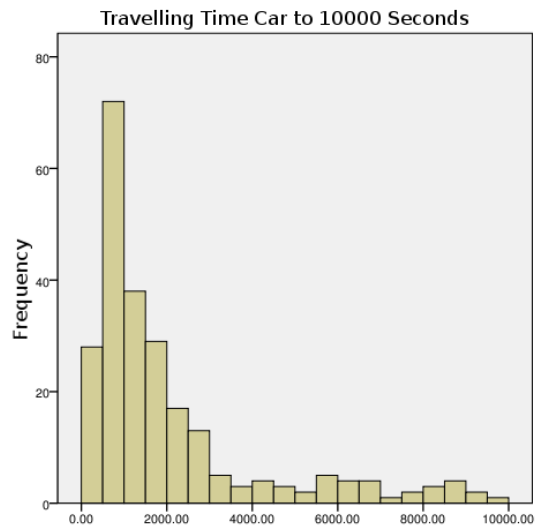


Fig.3: histogram with 20 equal-distance classes for traveling time up to 10,000 seconds

up to 10,000 seconds. You can see it in Figure 3. It has 20 equal distance classes which have a size of 500 seconds each. This histogram still includes 240 hits. It shows a right-skewed distribution: in the first class we only have 28 occurrences. With 71 occurrences, most points are found in the second class 500 to 1,000 seconds. Afterwards the amount of datapoints declines for the next two classes like normal distributed data. That there are relatively few occurrences in the first class let us conclude that things which are connected to each other are not mentioned as being *close to* each other.

		Number of total Occurrences in Corpus	Knowledge Representation Occurrences (Switzerland only)		Travelling Time by Car in Seconds				
			basic Rule	including Micro regions	Min.	Max.	Median	Average	Standard deviation
All	close to	345	-	-	129	706788	2166	31222	103053
	not close to	30	-	-	9720	605711	34713	166289	215039
Close to <= 10000 Seconds		240	-	-	129	9859	1200	2072	2160
Switzer- Land	close to	33	28	21	265	2382	729	728	485
	not close to	2	2	2	3375	7526	4304	4708	2163

Fig. 4: Overview of Results in Numbers

A table with the results of the comparison can be seen in Figure 4. Like in the histogram, we chose to display the results for the 240 cases up to 10,000 seconds traveling time as well. For example the maximum of time needed for a route to a place would be over 8 days (706788 seconds). As already mentioned this is because of mismatched place names with the gazetteer items. For the same reason average (8.67 hours; 31,222 seconds) and standard deviation (28.63 hours; 103,053 seconds) show high values. The shortest trip to a *closeby* place takes 2,15 minutes (129 seconds). For the 240 cases up to 10,000 seconds of traveling time we have better results. The maximum value is 2,74 hours (9,859 seconds). Average is 34,53 minutes (2,072 seconds). The standard deviation 36 minutes (2,160 seconds) is still high. A reason for this could be that also other means of transport have an influence here. So for example, one feels as *close to* a place where one has a direct flight connection to. This could explain that the histogram declines in waves, the second peak could for example make up the places with direct flight connections. And since the second peak is much smaller, we could conclude that this is because it is much more usual to go by car than to go by plane. The minimum traveling time of the *not close to* matches is slightly below 10,000 seconds. This may back up that our decision to have a closer look at occurrences below 10,000 seconds. With the *not close to* matches it is quite natural that standard deviation is high, since we also used differnt synonyms for *not close to*. Also the range for things which are felt as *not close to* may be very widespread.

Results for pairs of places in Switzerland are listed again to compare them to the knowledge representation approach which is only applicable to a part of Switzerland right now. Since the knowledge representation approach deals with administrative regions but not with villages we mapped every occurring village to its community for the comparison. For Switzerland we only have 2 negative matches, too few to say much about them. What can be seen by the positive matches is the smaller scale: the range is between 4.42 minutes (265 seconds) and 39.7 minutes (2,382 seconds) and the average for traveling time is only 12.13 minutes (728 seconds). We can conclude from that, that when both places of a *close to* relation are situated in a relatively small country like Switzerland, also the distances between *closeby* places are small.

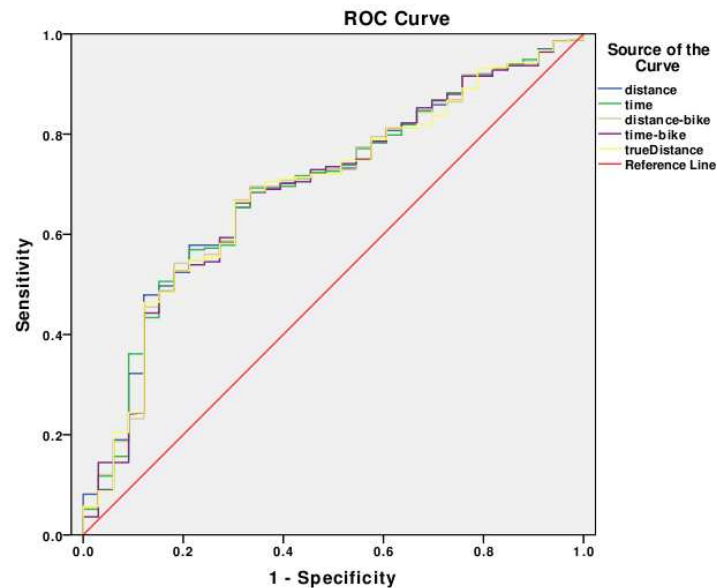


Fig. 5: Sensitivity and Specificity of time (car), distance (car), time-bike, distance-bike and true distance. It can be seen that time (car) and time-bike are slightly better predictors than true distance.

A logistic regression calculation using SPSS with nearness as dependent variable (1 for near pairs; 0 for not-near pairs) showed that distance, traveling distance and time are all correlating to nearness. But still we have too few values for pairs which are *not close to* to do valuable prediction. Also with SPSS sensitivity and specificity were calculated under the assumption that traveling time predicts what is felt as *close to*. Results are shown in the roc-curve in Figure 5. As you can see, the correlation of time was a bit stronger but with the data base available now it shows not significance.

5.3 Comparison with the Knowledge Representation Approach

The comparison with the knowledge representation approach shows, that places that occurred in the corpus as *close to* each other are also found to a great amount by the knowledge representation approach. The best fit was gained with application of the basic rule: 28 of 33 matches were found, which makes 84.85 %. For the micro region extension there have been 21 of 33 matches, which makes up 63.64 %. So the extension with the micro regions is more restrictive and the basic rule found many of the *closeby* places. Maybe in the future one could use additional regions, like the micro regions, to limit results of the knowledge representation approach for a typing structure of regions that is related to the context of searching.

The evaluation is not finished yet, we still have to gain more corpus data for places which are not near. But we already have established the process to get final results for our approach.

6 Discussion

		Knowledge Representation	Statistical
Amount and Type Of Input	+	Satisfied with little data	After establishment: Only route information needed
	-	Needs very special data: Polygons from topologies	Not suitable where no route can be calculated Establishment needs much data
Variability of Feature Types For Places	+	Extendable via extension of ontology	Applicable wherever there is a route between two places
	-	Only for entities in ontology (no bakeries, cinemas, villages for now; but extendable)	Data needed as points, e.g. no route for "Frankfurt far away from Asia"
Context For Usage	+	Not as restricted as Statistical Extendedable for further types of regions (e.g. inhabitation density areas)	Best matches for context of travelling with given means of transport
Output	+		Adjustable to vagueness
	-	Binary: either <i>close to</i> or <i>not close to</i>	

Fig. 6: Overview of Advantages and Disadvantages of both Approaches

The table in Figure 6 shows an overview of the advantages and disadvantages discussed in this section. The knowledge representation approach will be more precise wherever there is only little data about transport connections like streets available. But information about hierarchical structures of regions, information about topology and the type of a region is needed. It is good for modelling all kinds of landscapes (e.g. swamps, mountains) as polygons and reason on them. The statistical approach needs much data for setting up critical times for things which are *closeby*. Once after the approach started working with good predictions it will be applicable wherever you have route information, but not if the route cannot be calculated.

close to-calculations with the knowledge representation approach can be done with all types of places that are specified in the ontology and only these. If one wants to calculate *closeby* cinemas or bakeries, the ontology has to be extended with such types of places and entities which are of these types. An example for this is used in the discussion of the table in Figure 4: some places are villages and only the types of community, district and canton are available. So a mapping of the village to the community to calculate the *close to*-factor for the community in which it was situated was performed.

The statistical approach is applicable whenever there is a known path between places. It works with point data, not polygons. For cities, states and so on there is always used a point which lies within. Depending on where you are in the state/ city the approach is more or less accurate. For that reason we have not been able to calculate a route from Frankfurt to Asia (cf. the *Frankfurt weitab von Asien* ("Frankfurt far away from Asia") example from the *not close to* part of the corpus, section 4.1). Nevertheless, under normal circumstances it is not very likely that such a route is needed.

While the Knowledge Representation uses administrative regions, the statistical approach uses the context of traveling. Traveling is important for perception of things *closeby* but also the hierarchical administrative structure of a country has some influence (cf. section 2). The Knowledge Representation method will always lead to clearcut judgements *close to* or *not close to*. But since we are dealing with natural language data which is often quite vague and has many influence factors, in one context a place may be seen as *close to* the reference place whereas in another it is not. The statistical approach can also say something about the "shaded" areas, it can give the degree to which something is near.

7 Conclusion and Outlook

We have shown that it is possible to model human language concepts of spatial relations via description logical expressions using administrative regions as background knowledge or via reachability by different means of transport. They both have advantages and disadvantages. As already mentioned we have to extend the evaluation for the statistical approach. When this is done with satisfying results, we want to embed more possibilities to model humans perception of vague natural language expressions for spatial relations, for example *bei* (English: "next to"), *zwischen* (English: "in between"), etc. We will do this for the two presented approaches and maybe for others which we will develop in the future. Ontologies are providing good background knowledge for such additional models. There are many influence factors for the perception of spatial relation, one could build up a user-friendly system which first evaluates the most important models for the users needs and then computes the best-matching results.

Bibliography

- [1] Carbon, C.C., Leder, H.: The wall inside the brain: Overestimation of distances crossing the former iron curtain. In: *Psychonomic Bulletin & Review*, p. 746750. No. 12 (4) (2005)
- [2] Grütter, R., Helming, I., Bernstein, A., Speich, S.: Rewriting queries for web searches that use local expressions. In: Bassiliades, N., Governatori, G., Paschke, A. (eds.) *Rule-Based Reasoning, Programming and Applications; Proceedings of 5th International Symposium, RuleML 2011-Europe*. pp. 345–359. Springer, Heidelberg (2011)
- [3] Maki, R.: Categorization and distance effects with spatial linear orders. *Journal of Experimental Psychology: Human Learning and Memory* 7 (1), 1532 (1981)
- [4] Mark, D., Egenhofer, M.: Modeling spatial relations between lines and regions: Combining formal mathematical models and human subjects testing. Tech. Rep. 94-1, National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA (1994)
- [5] Mata, F.: Geographic information retrieval by topological, geographical, and conceptual matching. In: F., I.F., A., R.M., S., L. (eds.) *Proceedings of the Second International Conference on GeoSpatial Semantics (GeoS 2007)*. p. 98113. Springer, Lecture Notes in Computer Science No 4853, Berlin (2007)
- [6] Motik, B., Horrocks, I., Rosati, R., Sattler, U.: Can owl and logic programming live together happily ever after? In: Cruz, I.e.a. (ed.) *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*. LNCS, vol. 4273, pp. 501–514. Springer, Heidelberg (2006)
- [7] Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connections. In: Nebel, B., Rich, C., Swartout, W. (eds.) *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning (KR92)*. pp. 165–176. Kaufmann, San Mateo, CA Morgan (1992)
- [8] Schockaert, S., De Cock, M., Kerre, E.: Location approximation for local search services using natural language hints. *International Journal of Geographic Information Science* 22 (3), 315–336 (2008)
- [9] Yaoh, X., Thill, J.C.: How far is too far? - a statistical approach to context-contingent proximity modeling. *Transactions in GIS* 9 (2), 157–178 (2005)

Finding spatial equivalences across multiple RDF datasets

Juan Martín Salas¹ and Andreas Harth²

¹ FRLP, Universidad Tecnológica Nacional, Argentina juan.salas@ieee.org

² Institute AIFB, Karlsruhe Institute of Technology (KIT), Germany harth@kit.edu

Abstract. The importance of geospatial information is being reflected on the growing amount of spatial datasets on the Semantic Web. However, the high variability of the data presents challenges for integration. In this paper, we address the problem of finding spatial equivalences between geospatial RDF datasets. First, we present mappings between our NeoGeo vocabulary and the vocabularies used by some well-known spatial RDF datasets. Second, we describe a method to find spatially co-located features across spatial RDF datasets. To find equivalences, we rely on analyzing the Hausdorff distance distribution in the compared datasets, with the objective of finding a sensible criterion that aids the recognition of equivalent regions.

1 Introduction

Geospatial data is ubiquitous in information management, supporting scientific, industrial or just everyday activities. The relevance of geospatial data is reflected by the growing amount of geospatial datasets on the Web.

A feature is an abstraction of a real world phenomenon (e.g. a building, a mountain or an administrative region). A geographic feature is a feature associated with a location relative to the Earth, which is usually represented by a certain geometric shape (e.g. a point, a curve or a polygon). Features that are spatially co-located (i.e. share the same location) are not necessarily always the same. However, finding spatially co-located regions is a powerful measure of similarity between features.

Factors like rounding effects, different scales and different formats, present a challenge when attempting to elicit equivalences between geospatial resources. We define a method for obtaining a criterion that best fits the differences between the datasets merged.

This work was successfully applied to integrate our RDF representations of the NUTS nomenclature of the European Union ³ and of the GADM project ⁴ to other datasets describing spatial information on the Semantic Web, and also between each other.

Our contributions are as follows:

- Representation and modeling of datasets: we survey the representation of existing geospatial datasets, and distill an integration vocabulary which covers the

³ <http://nuts.geovocab.org/>

⁴ <http://gadm.geovocab.org/>

core set of classes and properties in existing data. We also integrate existing vocabularies and publish two geospatial datasets (Section 2).

- Integration and mapping of multiple datasets: we develop an algorithm for finding equivalences for geometric shapes across multiple datasets (Section 3).
- Evaluation of the presented approach: we conduct experiments in which we evaluate the accuracy of the results (Section 4).

We discuss the related work in Section 5. Finally, we identify areas for future work and conclude in Section 6.

2 Representing geospatial data on the web

2.1 Analyzed datasets

We start with providing a brief summary of the analyzed datasets.

- UN FAO Geopolitical Ontology⁵: The Food and Agriculture Organization of the United Nations (FAO) is a specialized agency of the UN. The UN FAO Geopolitical Ontology provides the FAO and its associated partners with a master reference for geopolitical information.
- OS OpenData⁶ [8]: The Ordnance Survey (OS) is the national mapping agency for Great Britain. OS has released a number of its products as Linked Data.
- GeoLinkedData.es [2, 3]: The initiative provides geospatial information about the national territory of Spain. The information provided as RDF at their website is gathered from different national sources. However, the integration process is based on string matching.
- LinkedGeoData.org [22]: The project provides data from OpenStreetMap⁷ as Linked Data.
- GeoNames.org: GeoNames is a geographical database that covers all countries and provides Linked Data under a Creative Commons attribution license.
- Uberblic.org: Uberblic provides an integration service that includes data from GeoNames, Wikipedia, MusicBrainz, Freebase, Last.fm and Foursquare.
- RAMON NUTS⁸: The Nomenclature of Units for Territorial Statistics (NUTS) is a geocoding standard for referencing the subdivisions of countries for statistical purposes developed by the European Union and published as Linked Data.
- DBpedia.org: The community effort extracts structured information from Wikipedia for publication as Linked Data.
- NeoGeo: We provide an integration vocabulary, described in more detail in Sections 2.4 and 2.5.

⁵ <http://www.fao.org/countryprofiles/geoinfo/geopolitical/resource/>

⁶ <http://data.ordnancesurvey.co.uk/>

⁷ <http://openstreetmap.org/>

⁸ <http://rdfdata.eionet.europa.eu/ramon/nuts2008/>

2.2 Representing location

The analyzed spatial datasets represent the location of features in different ways. We identified four main kinds of representation: point, bounding box, points in lists, points using a single property and literals. Geometric shapes are not only described using different vocabularies, but also these vocabularies are based on different structures, which increases the difficulty of working with GeoData across datasets.

- **Point** Location of objects is merely represented by a geographic point. The most common vocabulary to do so is W3C Geo[24], sometimes complemented with a GeorSS representation [21], such is the case of the UK Ordnance Survey, even if GeorSS is not a proper RDF vocabulary but an XML-Schema. In some cases, neither W3C Geo nor GeorSS is used, but an own vocabulary, as is the case of the Uberblic Ontology, which uses its own "latitude", "longitude" and "altitude" predicates.
- **Bounding box** The location is represented by two points or four line segments forming a georeferenced rectangle (on cylindrical projections). This is the case of the FAO Geopolitical Ontology, which uses four predicates (hasMinLongitude, hasMinLatitude, hasMaxLongitude, hasMaxLatitude) to represent a rectangle. The rectangle is represented by line segments, which should be tangential to the region at some point.
- **Points in lists** The geometric shape of a region is represented by a collection of points, each being described as a single RDF node. The whole collection of points is then linked together using either an RDF Collection or an RDF Container. LinkedGeoData.org represents geometric shapes by using a "hasNodes" object property, which links to a rdf:Seq container. The rdf:Seq container describes the nodes of a shape, which are represented using the W3C Geo Vocabulary.
- **Points using a single property** In the GeoLinkedData.es ontology, rivers are represented by a group of "Curva" (curve) RDF resources (similar to a GML LineString). "Curva" resources use a single "formadoPor" object property to link each of their nodes, which at the same time contain the WGS-84 coordinates (represented with the W3C Geo Ontology) and an "orden" (order) value property, defining the position of each node within the geometric shape.
- **Literals** Both Ordnance Survey and GeoLinkedData.es (for rivers only) ontologies include a predicate allowing to include a GML representation of the geometric data, which is coded in RDF as a literal. A "geometry:extent" property links a feature to its geometric representation.

2.3 Representing spatial relations

A spatial relation states the location of an object in relation to another. We created a set of vocabulary mappings to the NeoGeo vocabulary using the `rdfs:subPropertyOf` predicate. Table 1 shows which predicates are used in each dataset to describe spatial relations.

Dataset	Disjoint	Touches	Overlaps	Within	Contains	Equals	Nearby
UN FAO		hasBorder- With		isInGroup			
Ordnance Survey	disjoint	touches	partially- Overlaps	within	contains	equals	
GeoLinkedData.es				forma- ParteDe	formado- Por		
LinkedGeoData.org							
GeoNames.org		neighbour / neigh- bour- ingFea- tures		parent- Feature	children- Features		nearby / nearby- Features
Uberblic.org		adjoining- location		containing- location			
RAMON NUTS				partOf			
DBpedia.org				locatedIn- Area			
NeoGeo	DC	EC	PO	PP	PC	EQ	

Table 1: Equivalent properties for spatial relations across multiple vocabularies.

2.4 NeoGeo ontologies

Given the lack of a standardized spatial vocabulary, we developed our own set of spatial ontologies, which we call NeoGeo⁹. We manually created a set of mappings between our vocabularies and the vocabularies used by some acknowledged spatial datasets like the Ordnance Survey and LinkedGeoData.org. Both the GADM and NUTS datasets use the NeoGeo vocabularies.

The Geometry Vocabulary¹⁰ is an RDF vocabulary for the description of geo-referenced geometric shapes. It is based on the Core Profile of the Spatial Schema [12] and the General Feature Model [11]. Hopefully, the lack of a standardized RDF vocabulary in this domain will probably be addressed by GeoSPARQL[16] shortly. For experimentation reasons, the Geometry vocabulary allows to encode geometric shapes in a representation fully based on RDF or as a WKT representation [10] embedded in an XMLLiteral.

The Spatial Ontology¹¹ provides a vocabulary for the representation of the spatial relations used in the Region Connection Calculus (RCC) [19]. It also provides monotonic reasoning by mapping most of the semantics of RCC into OWL.

2.5 NeoGeo datasets

We provide two datasets, NUTS and GADM, containing geospatial information as Linked Data.

⁹ <http://geovocab.org/>

¹⁰ <http://geovocab.org/geometry>

¹¹ <http://geovocab.org/spatial>

The Nomenclature of Territorial Units for Statistics (NUTS) is a classification defined by the Eurostat office of the European Union. It is intended to divide the administrative regions of the European Union, in a way that the resulting regions are demographically equivalent.

The RDF representation of the NUTS nomenclature contains a 1:60,000,000 geospatial representation of the NUTS statistical units mapped to RDF. The resources representing NUTS regions in our dataset include (among others) links to resources in DBpedia, FAO Geopolitical Ontology, GeoNames, Ordnance Survey and GeoLinkedData.es.

The Global Administrative Areas (GADM)¹² is a project seeking to become a collaborative effort on building a spatial database containing information about all of the administrative regions in the world. GADM aims to provide high resolution mappings for all administrative areas in the world, along with additional information about them. The latest version of GADM (0.9) maps 226.439 administrative areas. The information can be downloaded at their website in the following formats: Shapefile, ESRI geodatabase, RData and KMZ.

Given the value of the GADM project, we have created an RDF representation of the information contained in the original GADM project, which we seek to enrich with additional capabilities like the materialization of spatial relations, mappings to other RDF datasets and SPARQL querying support.

3 Instance mappings

The alignment of the vocabularies is only the first step for the integration of the datasets. The second step in the process is to find matchings between the features.

We can classify the features into three general categories, in relation on how their location is represented in the datasets. First are the resources that present no quantitative spatial information at all. Second, the features that approximate their location by using only a single point. And finally, features which present rich information about their location (i.e. include a description of their geometric shape).

3.1 Resources with no quantitative geospatial information

Resources which include no quantitative information about their location can be integrated by relying either on text matching [14], or object property matching [23] [4]. These techniques are also suitable for the disambiguation of spatially obtained mappings [9]. This kind of resources is not the topic of this work.

3.2 Resources with poor quantitative geospatial information

Sometimes the location of a feature is approximated by using a single point (e.g. using the W3C Geo vocabulary) instead of representing its actual extent (i.e. the geometric shape). Examples of this kind of representation are DBpedia, LinkedGeoData.org, GeoNames and LinkedGeoData.es.

¹² <http://gadm.org/>

This kind of representation can lead to false assertions while performing comparisons against a spatial index if these features are not especially considered. For example, DBpedia uses the W3C Geo Vocabulary to describe the latitude and longitude coordinates of features as points. The resource for Germany in DBpedia <http://dbpedia.org/resource/Germany> is spatially represented by a point with latitude 52.516666 and longitude 13.383333. If we intended to obtain the containment relations for such resource by comparing it with a spatial index, the result would be that Germany is part of Berlin, which is false. Therefore, even though these relations can be obtained using the coordinates represented in DBpedia, first it is necessary to ensure that the process will not return such false statements. The false matches can be avoided, for example, by filtering the features that will be compared by its class, in a way that ensures that the feature will be properly contained in the features that it will be compared to (e.g. cities in provinces or restaurants in cities).

3.3 Resources with rich quantitative geospatial information

Resources that include an accurate description of their extent as a geometric shape can be compared using this information. We will focus on obtaining links between spatially co-located regions (we use the `spatial:EQ` predicate). Whether `owl:sameAs` links can be deduced from the obtained links depends on the modeling of the datasets (e.g. the class the resource belongs to).

To perform the comparison, we will adopt a Plate Carrée projection for both of the compared datasets. Being this projection equirectangular, we can treat latitude and longitude coordinates as if they were cartesian. Therefore, the units will be presented in centesimal degrees.

The benefit of using an equirectangular projection is that it simplifies the calculations by avoiding local reprojections (e.g. to UTM), and also allows to use a global spatial index, improving the performance of the process. In our approach it is not important if the projection distorts the real size or the actual geometric shapes on the surface of the Earth, as long as the geometric data is equally distorted for all datasets.

Due to a series of factors like rounding effects and different scales, there is no guarantee that both geometric shapes will be vertex by vertex identical. Figure 1 exemplifies these differences by showing the boundaries for Luxembourg as they are represented in the GADM and NUTS datasets.

An effective method of determining how similar two geometric shapes are is to compute the Hausdorff distance between them. The Hausdorff distance is the "maximum distance of a set to the nearest point in the other set" [20]. More formally, given two sets of points $A = \{a_1, a_2, \dots, a_n\}$ and $B = \{b_1, b_2, \dots, b_n\}$, the Hausdorff distance is defined by:

$$d_H(A, B) = \max(\{\arg \max_{a \in A} \arg \min_{b \in B} d(a, b), \arg \max_{b \in B} \arg \min_{a \in A} d(a, b)\})$$

It can be deduced from the formula that in the particular case of calculating the Hausdorff distance between points, the Hausdorff distance matches the Euclidean distance $d(a, b)$.

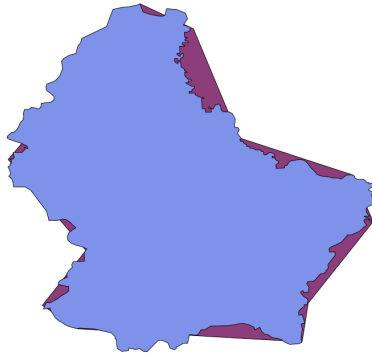


Fig. 1: Incongruency of the geometric data (GADM: blue, NUTS: violet) due to differences in resolution.

Figure 2 shows the values of correct and wrong guesses for similar regions in both datasets. In order to better appreciate the variability of the values, only small areas are plotted in the chart. From the figure it can be deduced that smaller regions (e.g. boroughs) require greater precision than larger regions (e.g. countries), in order to differentiate them from each other. Therefore, the Hausdorff distance margins allowed for regions which are suspected to be spatially co-located must be different, depending on the area size of the regions being compared.

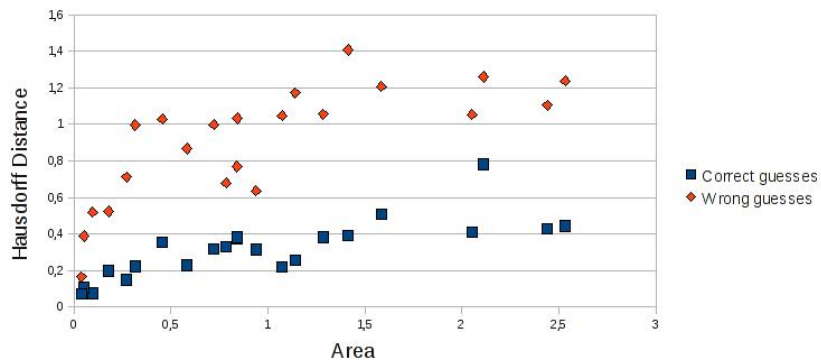


Fig. 2: Values for correct and wrong guesses for similar regions in NUTS and GADM.

To address this issue, it is desirable to obtain a function for a Hausdorff distance threshold for a given area size. In order to do this, first we calculate the midpoint between the lowest and second lowest Hausdorff distance values, for a representative set of features in both datasets. Afterwards we perform a quadratic

regression from the midpoint Hausdorff distance values. This produces a formula that allows to determine the maximum Hausdorff distance allowed between two regions, in order to consider them similar. The resulting function has the following form:

$$MaxHDist(x) = A \cdot x^2 + B \cdot x + C$$

Where MaxHDist is the maximum Hausdorff distance allowed between two regions in order for them to be considered similar. The x variable is the area of the region. The quadratic function gives more precision for small regions while allowing a greater margin for large regions. The A,B and C constants are tunable parameters for the integration procedure.

A yet unresolved issue of using a quadratic function is that the samples must include values for the approximated maximum area for which the integration will be performed. This is because the values of the function will tend to decrease after reaching a maximum value. We are performing experiments with logarithmic functions to solve this issue.

Table 2a shows sample execution times and Hausdorff distance values between features in the NUTS and GADM datasets.

Region Name	NUTS Id	Area	Hausdorff Distance	Time (ms)
Finland	FI	62.2835	1.3996	30353
Iceland	IS	19.3357	0.4163	567
Croatia	HR	6.2139	1.1374	7830
Schleswig-Holstein	DEF	2.1126	0.7281	1870
Karlsruhe	DE12	0.8433	0.1062	35
Seine-Saint-Denis	FR106	0.0358	0.0812	1

(a) for the original geometric shapes

NUTS Id	Hausdorff Distance	Time (ms)	NUTS Id	Hausdorff Distance	Time (ms)
FI	1.3483	2504	FI	1.3483	2257
IS	0.4613	66	IS	0.4863	49
HR	1.1366	1108	HR	1.1366	1053
DEF	0.7257	296	DEF	0.7801	278
DE12	0.1906	13	DE12	0.3762	14
FR106	0.0716	2	FR106	0.0716	2

(b) simplified with a separation of 0.2 degrees (c) simplified with a separation of 0.5 degrees

Table 2: Sample Hausdorff distance values and execution times

Calculating the Hausdorff distance between the original geometric data is quite expensive, especially for large regions. In order to increase the performance of the process, as an optional step, we chose to simplify the geometric shapes using the

Ramer-Douglas-Peucker algorithm [18] [5], prior to the calculation of the Hausdorff distances.

The Ramer-Douglas-Peucker algorithm starts by considering a line segment between the first and last points of the line. Then, it finds the furthest point from the line segment between the first and last points of the line. If the point found is closer than a predefined distance ε to the line segment, all other points that were not chosen to be used in the solution can be discarded. If the point furthest from the line segment is greater than ε , then the point is used in the solution. The algorithm then calls itself recursively with the found point and the last point as parameters.

Tables 2b and 2c show the Hausdorff distance between the NUTS regions and their matching GADM region, as well as execution times for different levels of simplification. As it can be seen, execution times are dramatically reduced, especially for large regions.

A further refinement of the process is to calculate the simplification distance for the Ramer-Douglas-Peucker algorithm depending on the Hausdorff distance threshold and therefore of the area of the regions. This is based on the same principle applied for the Hausdorff distance, where small areas require greater precision than large areas.

Given two spatial datasets A and B, the algorithm can be summarized as Algorithm 1.

4 Experiments

4.1 Implementation

We implemented the algorithm presented in Section 3.3 using the PostGIS 1.5.2 extension running on PostgreSQL 8.4.8. The computer is running on Ubuntu 10.04 on an Intel SU7300 processor with 4GB DDR3 RAM.

PostGIS includes the `ST_HausdorffDistance` function, which implements an approximation to the original algorithm. This approximation can be thought of as the "Discrete Hausdorff distance", which is the Hausdorff distance restricted to discrete points for one of the geometric shapes. If more precision is needed, the function receives also an optional "densityFrac" parameter which performs a segment densification before computing the discrete Hausdorff distance.

Since we are not concerned about the actual Hausdorff distance values, but just use it as a measure to determine if two regions are similar enough to be considered spatially co-located, this approximation is sufficient.

For the simplification of geometric data we will use the `ST_SimplifyPreserveTopology` function included in PostGIS. This function is a refined version of `ST_Simplify`, which is based on the Ramer-Douglas-Peucker algorithm [18] [5].

The query used with PostGIS to find regions which are supposed to be spatially co-located is very similar to the one presented below. To avoid having to perform the same calculations repeatedly, the values of the maximum Hausdorff distance function are cached into the "max_hausdorff_dist" column. The "geometry" column in both tables belongs to the "Geometry" datatype provided by PostGIS.

input : Datasets A, B
output: Equivalent regions from A and B
Convert the compared resources to a shared coordinate reference system.
Project the data into an equirectangular projection.
Obtain a representative set of regions in dataset *A* which intersect regions in dataset *B* and have a maximum arbitrary Hausdorff distance between each other.
foreach *region a of a representative set of regions in dataset A do*
 Get the minimum Hausdorff distance to a region in dataset *B*.
 Get the second minimum Hausdorff distance to a region in dataset *B*.
 Calculate the midpoint between the minimum and second minimum Hausdorff distances.
end
Perform a regression on the midpoints between the Hausdorff distances to calculate the Hausdorff threshold function.
foreach *region a in A do*
 foreach *region b in B do*
 if *a intersects b then*
 Calculate the Hausdorff distance between *a* and *b*.
 if *Hausdorff distance between a and b is lower than the threshold for the area of a then*
 a and b can be considered as spatially co-located.
 end
 else
 a and b cannot be considered as spatially co-located.
 end
 end
end
end

Algorithm 1: Matching algorithm

```
SELECT g.gadm_level, g.gadm_id, n.nuts_id
FROM nuts n INNER JOIN gadm g ON (n.geometry && g.geometry)
WHERE
    n.shape_area BETWEEN (g.shape_area*0.9) AND (g.shape_area*1.1)
    AND ST_HausdorffDistance(ST_SimplifyPreserveTopology(n.geometry, 0.5),
ST_SimplifyPreserveTopology(g.geometry,0.5)) < g.max_hausdorff_dist;
```

Basically this query selects the identifiers for the GADM region (level and id), and for the NUTS region (id). The && operator matches an intersection between the bounding boxes of the of the regions. Since similar regions will also have a similar area size, the first condition in the "where" clause filters regions that have a similar area size with an error of 10%. The second condition checks if the discrete Hausdorff distance between the simplified geometric shapes is within the limit calculated by the function presented in Section 3.3.

4.2 Evaluation

We can analyze the effectiveness of the method by looking at the results of the process of finding spatial equivalences between the NUTS and GADM datasets.

The NUTS dataset codes the geometric shapes fully in RDF using the NeoGeo vocabulary, and the coordinate system used is WGS-84. The data is retrieved by using a Construct SPARQL query and then converted into WKT using XSLT.

After retrieving the geometric data, it is merged with the GADM dataset using the method presented in Section 3.3.

Not all NUTS regions are expected to match a GADM region, since many NUTS regions represent parts or aggregations of administrative boundaries. Also a GADM administrative region in a certain level should be able to match different NUTS regions in different levels, and vice-versa.

From the existing 1,671 NUTS regions of the 2008 nomenclature that were included in the comparison, the algorithm detected 965 matches, from which 13 were false positives, as Table 3 shows.

NUTS Region	Incorrect guess	GADM Id	GADM Level	Area	Hausdorff Distance
UKM34	East Renfrewshire	14084	2	0.0214	0.1862
FR106	Val-De-Marne	13799	2	0.0334	0.1644
BE321	Soignies	2691	2	0.0654	0.3521
BE353	Thuin	2692	2	0.1188	0.2834
CH061	Aargau	531	1	0.1672	0.3653
LT	Latvija	136	0	9.5204	2.5098
LI	Appenzell Innerrhoden	533	1	0.0205	0.2783
UKM28	North Lanarkshire	14095	2	0.0689	0.3478
BE331	Lige	2696	2	0.1013	0.335
BE353	Thuin	2692	2	0.1188	0.2834
CH061	Aargau	531	1	0.1672	0.3653
SE3	Norge	168	0	60.585	7.8658
BE321	Soignies	2691	2	0.0654	0.3521

Table 3: False positives resulting on the application of the method

These false positives are due to the fact that the threshold is set too high for very small and very large areas. It is desirable to produce a larger gradient for small areas and a smaller Hausdorff distance threshold for large areas. This is still a matter for further research.

5 Related work

The problem of aligning spatial datasets is not new in the Semantic Web community and much work has been put on finding sensible solutions both at T-Box and A-Box level.

Ontology alignment is a heavily researched topic. Proposed solutions have been based on the terminological [15], structural [7], semantic [6] and extensional [17] aspects of the aligned ontologies. The last two works consider the alignment of spatial T-Boxes in particular.

Algorithms have also been proposed for feature matching across spatial datasets. [1] presents a series of algorithms for the integration of features, for which the location is approximated by a single point. These approaches have the problems exposed in Section 3.2 and therefore, are not suitable for all cases. [13] considers feature matching as an assignment problem based on a minimization of the Hausdorff distance between the geometric shapes. However, being this a case of Linear Programming, the method can only be applied for all the geometric shapes of both datasets at the same time, making it more difficult to integrate to live crawling.

6 Conclusion

We have presented a generic method that can be used to map multiple spatial datasets. We also showed its functioning by describing the integration between our two spatial datasets and analyzed its results.

Although the method has been used successfully to align the GADM and NUTS datasets, the false positive rate can still be improved when analyzing regions covering a wide spectrum of area sizes. However, the presented method has proven to be usable in Semantic Web applications.

Since the first experiments showed promising results, we are developing a tool that automates the whole mapping process. We are also further refining the algorithm to improve its precision and performance.

Acknowledgements

The authors acknowledge the support of the European Commission's Seventh Framework Programme FP7/2007-2013 (PlanetData, Grant 257641).

References

1. C. Beeri, Y. Kanza, E. Safra, and Y. Sagiv. Object fusion in geographic information systems. In *Proceedings of the thirtieth international conference on very large data bases. Volume 30*, pages 816–827. VLDB Endowment, 2004.
2. L.M.V. Blázquez, B. Villazón-Terrazas, V. Saquicela, A. de León, O. Corcho, and A. Gómez-Pérez. Geolinked data and inspire through an application case. *ACM SIGSPATIAL GIS*, pages 446–449, 2010.
3. A. de León, V. Saquicela, L.M. Vilches, B. Villazón-Terrazas, F. Priyatna, and O. Corcho. Geographical Linked Data: a Spanish use case. In *Proceedings of the 6th International Conference on Semantic Systems*, pages 1–3. ACM, 2010.
4. X. Dong, A. Halevy, and J. Madhavan. Reference reconciliation in complex information spaces. In *Proceedings of the 2005 ACM SIGMOD international conference on management of data*, pages 85–96. ACM, 2005.
5. D.H. Douglas and T.K. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica*, 10(2):112–122, 1973.
6. M. Dube and M. Egenhofer. Establishing similarity across multi-granular topological-relation ontologies. *Quality of Context*, pages 98–108, 2009.

7. J. Euzenat. An API for ontology alignment. *The Semantic Web-ISWC 2004*, pages 698–712, 2004.
8. J. Goodwin, C. Dolbear, and G. Hart. Geographical Linked Data: The administrative geography of Great Britain on the Semantic Web. *Transactions in GIS*, 12:19–30, 2008.
9. F. Hakimpour, B. Aleman-Meza, M. Perry, and A. Sheth. Spatiotemporal-thematic data processing for the Semantic Web. *The Geospatial Web*, pages 79–89, 2007.
10. J.R. Herring. OpenGIS® Implementation Specification for Geographic Information - Simple Feature Access - Part 1: Common Architecture. *Open Geospatial Consortium*, 2006.
11. International Organization for Standardization. ISO 19101. Geographic information Reference model, 2002.
12. International Organization for Standardization. ISO 19137. Geographic information Core profile of the spatial schema, 2007.
13. L. Li and M.F. Goodchild. Optimized feature matching in conflation. In *Geographic Information Science: 6th International Conference, GIScience 2010, Zurich, Switzerland, September 14-17, 2010. Proceedings*, 2010.
14. X. Li, P. Morie, and D. Roth. Semantic integration in text: From ambiguous names to identifiable entities. *AI magazine*, 26(1):45, 2005.
15. S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *Proceedings of the IEEE CS International Conference on Data Engineering*, pages 117–140. IEEE, 2002.
16. Open Geospatial Consortium Inc. GeoSPARQL - A geographic query language for RDF data, 2011.
17. Rahul Parundekar, Craig A. Knoblock, and José Luis Ambite. Aligning Ontologies of Geospatial Linked Data. In *Proceedings of the Workshop on Linked Spatiotemporal Data*, 2010.
18. U. Ramer. An iterative procedure for the polygonal approximation of plane curves. *Computer Graphics and Image Processing*, 1(3):244–256, 1972.
19. D.A. Randell, Z. Cui, and A.G. Cohn. A spatial logic based on regions and connection. *KR*, 92:165–176, 1992.
20. Günter Rote. Computing the minimum Hausdorff distance between two point sets on a line under translation. *Information Processing Letters*, 38:123–127, 1991.
21. Raj Singh, Ron Lake, Josh Liberman, Mikel Maron, and Carl Reed. An Introduction to GeoRSS: A Standards Based Approach for Geo-enabling RSS feeds. 2006.
22. Claus Stadler, Jens Lehmann, Konrad Höffner, and Sören Auer. LinkedGeoData: A Core for a Web of Spatial Open Data, 2011.
23. S. Tejada, C. Knoblock, and S. Minton. Learning object identification rules for information integration. *Information Systems*, 26(8):607–633, 2001.
24. World Wide Web Consortium (W3C) - Semantic Web Interest Group. W3C Geo Vocabulary. <http://www.w3.org/2003/01/geo/>, 2003.

EquatorNLP: Pattern-based Information Extraction for Disaster Response

Lars Döhling and Ulf Leser

Humboldt-Universität zu Berlin,
Department of Computer Science,
Unter den Linden 6, 10099 Berlin, Germany
{doehling,leser}@informatik.hu-berlin.de

Abstract. One of the most severe problems in early phases of disaster response is the lack of information about the current situation. Such information is indispensable for planning and monitoring rescue operations, but hardly available due to the breakdown of information channels and normal message routes. However, during recent disasters in developed countries, such as the flooding of New Orleans or the earthquake in New Zealand, a wealth of detailed information was posted by affected persons in media, such as Flickr, Twitter, or personal blogs. Finding and extracting such information may provide valuable clues for organizing aid, but currently requires humans to constantly read and analyze these messages. In this work, we report on a study for extracting such facts automatically by using a combination of deep natural language processing and advanced machine learning. Specially, we present an approach that learns patterns in dependency representations of sentences to find textually described facts about human fatalities. Our method achieves a F1 measure of 66.7% on a manually annotated corpus of 109 news articles about earthquake effects, demonstrating the general efficacy of our approach.

Keywords: Information Extraction, Dependency Graph, Earthquake, Disaster Response, Named Entity Recognition, Relationship Extraction

1 Introduction

After disastrous events like earthquakes, decision makers require precise and timely information about the current situation for planning and monitoring rescue operations effectively. During the last years, the Internet has become a major source for such information, in particular, if no acquaintance is available on-site. For earthquake events, many key information like the number affected are published on the Internet. This includes structured information provided by earthquake agencies (e.g. GEOFON¹ or USGS²) as well as textual updates published

¹ <http://geofon.gfz-potsdam.de/geofon>

² <http://earthquake.usgs.gov/earthquakes>

by news agencies or recently by Internet users themselves, called user-generated content (e.g. Twitter or personal blogs). Given the example sentence “The death toll in an earthquake in south-west China is now at least 32, with 467 injuries, state media says.”³, one can identify several text snippets expressing presumably demanded facts. It contains trigger words like “death toll” or “earthquake”, figures like “32” or “467” as well as temporal (“now”) or spatial (“south-west China”) attributes. Furthermore, these token or token sequences – subsequently called entities – are semantically connected to each other, forming so called relationships; “death toll” is related to “32” and “at least” whereas “467” refers to “injuries”. Moreover, both are associated with “earthquake” and “China”. Obviously, texts offer valuable information for decision making but require accurate analysis, which is still a manual and therefore time-consuming, expensive task. Hence, automating this analysis will aid humans to accomplish rescue operations successfully.

As a first step towards automatic textual analysis, we report on extracting facts from news articles, describing human impacts from earthquakes. To model these impacts, we define a 5-ary relationship, whose complexity imposes several challenges for extraction by

- consisting of more than two entities,
- allowing incomplete tuples and
- potentially spanning multiple sentences.

For extracting this relationship, we apply deep natural language processing combined with graph-based synthesis techniques. More specifically, we match patterns in sentence-based dependency graphs to compose a graphical model representing semantic connections between entities and examine this for connected subgraphs. Our evaluation demonstrates the general efficacy of our proposed method stack – called EquatorNLP⁴ – by achieving 66.7% F1 measure [23] on a novel, manually created news corpus.

1.1 Related Work

Due to the increasing amount of information available in a textual form (e.g. PubMed or Wikipedia), assisting humans by automatically analyzing these texts has become an important research topic in the last decade. Information extraction (IE) studies the problem of extracting structured information from unstructured text. Typically, this involves recognizing entities (named entity recognition, NER) and relationships between them (relationship extraction, RE).

Different methods have been proposed for NER, e.g. dictionary-based, rule-based or machine learning [20,29]. Hybrids like the one applied in this study usually perform best [28]. The achievable F1 measure highly depends on the concrete domain and ranges up to 95% [10,19,14,32]. To the best of our knowledge,

³ <http://news.bbc.co.uk/2/hi/asia-pacific/7591152.stm>

⁴ *EarthQUake dAta collecTOR* [8] with *Natural Language Processing*

this study is the first about IE in the earthquake domain, hence no quantitative results are available yet.

Regarding RE, co-occurrence forms an intuitive approach [15]. Beside that, pattern matching [30] and machine learning [21] have been adopted as well. As in EquatorNLP, these methods recently utilize deep natural language processing like dependency parsing [12]. Little is known about extracting complex n -ary relationships like the one examined in this paper, since most research has focused on binary relationships. Inspired by the promising results in [27], we transferred their subgraph-based idea into our domain (see section 2.4). In general, RE is regarded as being more difficult than NER, resulting in lower F1 measures, ranging from 40% [16] to 80% [12].

2 Materials and Methods

2.1 What we extract: Definition of the 5-ary Relationship

To model earthquake damages, our examined relationship consists of five different entity types, including several subtypes. Note that the concatenated parenthesized letters will subsequently be used as abbreviations.

- (O)bject: Describes the victims, e.g. “people” or “students”.
- (Q)uantity: Describes the number of victims and consists of the four subtypes
 - (c)ardinal: “12”, “ten”, “no”, “a”, “1.3 million”
 - (o)rdinal: “second”, “10th”
 - (v)ague: “many”, “hundreds”, “some”
 - (r)esidue: “everybody”
- (M)odifier: Refers to a quantity and modifies its value, e.g. “at least”, “about” or “more than”.
- (I)ndicator: Describes the type of damage and consists of six subtypes
 - (k)illed: “killed”, “death toll”, “died”
 - (i)njured: “injured”
 - (t)rapped: “trapped”
 - (m)issing: “missing”
 - (h)omeless: “homeless”
 - (a)ffected: “affected”
- (N)egation: Infrequently required to correctly describe a damage, e.g. “not”.

Given this definition, the previous example “The death toll in an earthquake in south-west China is now at least 32, with 467 injuries, state media says.” contains five entities: “death toll” (Ik), “at least” (M), “32” (Qc), “467” (Qc) and “injuries” (Ii). Note that entities may span multiple token – called multi-token entities. Together, these entities form two [N, M, Q, O, S] relationship tuples: [—, “at least”, “32”, —, “death toll”] and [—, —, “467”, —, “injuries”]. We define that not all entity slots have to be filled to form a valid tuple, indicated by —. However, we postulate two constraints concerning incomplete relationship instances: (i) An entity I is mandatory and (ii) an entity Q is mandatory, if an entity M is set.

2.2 Corpus

To train and later test our proposed machine-learning-based extraction methods, we required an annotated set of documents – called corpus – as a gold standard. As to the best of our knowledge, currently no appropriate corpus exists for our purpose, we created a new one. Our corpus consists of 109 English articles about earthquakes and their aftermath: 24 from BBC News⁵, 2 from Equator [8], 41 from Wikipedia⁶ and 42 from Yahoo! News⁷. They were randomly selected from a collection of documents retrieved from these four sources in spring 2010.

From each article, we extracted the text including the headline and annotated it manually according to the relationship definition given above. We removed cross-sentence (28) and unary (4) instances from the corpus, since our relationship extraction methods operate on the sentence level and are unsuitable for unary tuples (see section 2.4). Finally, we partitioned this altered corpus into a training ($\frac{2}{3}$) and an evaluation set ($\frac{1}{3}$) by stratified random sampling on the sentence level. Table 1 presents the resulting distribution of the relationship tuples in the different partitions.

Table 1. Data set statistics; Note that the Gold Standard values differ from the sum of training and evaluation set, owing to the removal of unary and cross-sentence instances.

Number of [...]		Training	Evaluation	Gold Standard
Sentences		1,964	986	2,950
containing a relationship instance		276	145	486
Token		39,856	20,796	60,652
Relationship instance		382	190	604
per type, defined by I sybtype	k	273	135	439
	i	56	24	80
	t	15	7	23
	m	19	11	30
	h	17	10	27
	a	2	3	5
per size, defined by filled entity slots	1	0	0	4
	2	152	69	245
	3	156	76	236
	4	74	45	119
	5	0	0	0

2.3 Named Entity Recognition

A prerequisite for relationship extraction is the detection of target entities in the text. For this task, we used a regular expression (Qc only) in combination with a

⁵ <http://news.bbc.co.uk>

⁶ http://en.wikipedia.org/wiki/Historical_earthquakes,
http://en.wikipedia.org/wiki/List_of_20th_century_earthquakes,
http://en.wikipedia.org/wiki/List_of_21st_century_earthquakes

⁷ <http://news.yahoo.com/science/earthquakes>

dictionary (all other types), both derived from the training data. As each token sequence can only be assigned to at most one entity type, the question emerged how to disambiguate competing matches. we applied the following plausible order of precedence: The regular expression matches prior to the dictionary, longer token sequences match prior to shorter (“as high as” M versus “high” Qv) and finally the most frequent type found for this token sequence in the training data. Overall, the dictionary extracted from the training data set contained 218 entries with an average length of 1.78 token.

2.4 Relationship Extraction

After recognizing the entities, the next step is to extract the actual relationship instances. Our proposed method consisted of two steps:

1. Discovering pairs of entities by pattern matching in dependency graphs.
2. Synthesizing complex instances from maximal cliques in entity graphs build from these entity pairs.

Dividing the extraction process into these two steps enabled us to apply well-known extraction methods for binary relationships. Furthermore, we gained more training instances, reducing the sparse data problem [22] existing for the complete relationship.

Matching Dependency Patterns Dependency models are syntactical models expressing the hierarchical dependencies between the words of a sentence. Those dependencies may be visualized as a directed, labeled graph whose root is the verb. Figure 1 depicts the running example in the Stanford Dependencies representation [25]. The arrows indicate the dependency direction from regent to dependent and are labeled with the dependency type.

These models offers a direct access to sentence structures [23] and have the potential to reveal relations between words apart more easily than regular expressions [12,7] (e.g. between “death toll” and “32” in the example). Therefore, examining patterns between entities in dependency graphs has been a successful approach in modern relationship extraction. For our work, we selected the shortest paths between two entities as patterns [4].

We applied the Stanford converter [24] to compute the dependency graph for each sentence, which requires constituent parses [23,1] as input (another syntactical model). Those parses were generated by the Charniak PCFG parser [5] in combination with the Charniak-Johnson Max-Ent reranking parser [6], using McClosky’s self-trained models [26].

During training, we extracted all shortest paths between entities and stored them in a pattern catalog. To abstract from the actual token of an entity, we joined all parting token vertices in advance into one entity vertex. Concurrently, we replaced each entity vertex in the pattern by its type to mask the actual value. Figure 2 illustrates the transformed example graph and the extracted patterns. Since dependency graphs can contain cycles, there might exist more than one

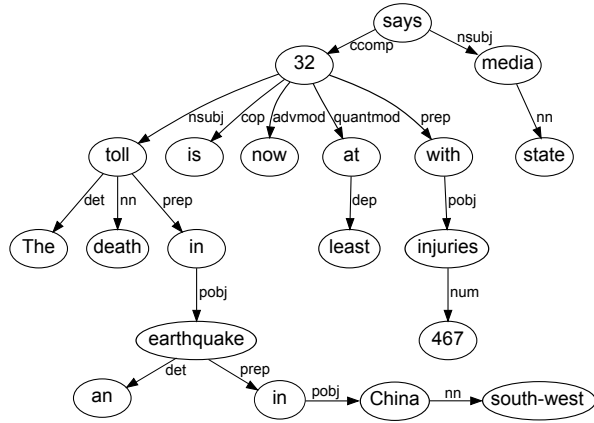


Fig. 1. Dependency graph of the sentence “The death toll in an earthquake in south-west China is now at least 32, with 467 injuries, state media says.”

shortest path between two entities with – of course – equal length. Hence, a relationship instance consisting of k entities will produce at least $\binom{k}{2}$ patterns. Overall, the catalog extracted from the training data set contained 396 unique patterns with an average length of 2.83 edges.

During extraction, we applied this catalog to create links between two entities in accordingly transformed dependency graphs, resulting in entity graphs (see Figure 3 for an example).

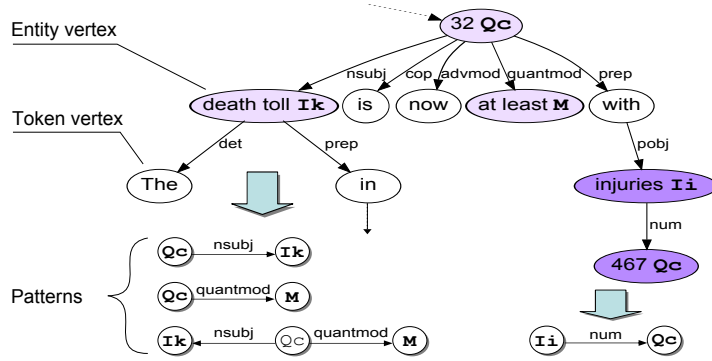


Fig. 2. Pattern extraction in the transformed example dependency graph (truncated)

Baseline To determine whether deep linguistic parsing like the dependency model is beneficial for relation extraction or not, we also use a co-occurrence-based classifier as a baseline for recognizing entity pairs. For each entity e , all

closest (in terms of token distance) entities within sentence scope having a different type than e are postulated as being linked to e . For example, this would imply a (false) connection between “32” and “injuries” in the running example, as for “32” the distance to “injuries” is less than to “death toll”.

Synthesizing Relationship Instances After detecting pairs of entities, the final step is to synthesize relationship instances from them. To address this, we identified maximal cliques in the entity graphs [27] which are consistent to our relationship definition.

Consider the entity graph in Figure 3 as one possible outcome of the previous pair-recognizing step when applied to the running example. To form relationship instances, we combined all those entities which are directly connected among each other in the entity graph. Such a set of vertices is called a clique. In Figure 3, all cliques of size two or greater are marked by eclipses (C_0 to C_5). Among these, we considered only those cliques that are consistent to our relationship definition (C_0 , C_2 and C_4). Furthermore, we ignored cliques which are contained in others (C_2). All such non-redundant cliques are called maximal. In our example, only C_0 and C_4 comply with the requirements ‘maximal’ and ‘relationship-definition-consistent’ and would in this case form the final output of the complete information extraction pipeline.

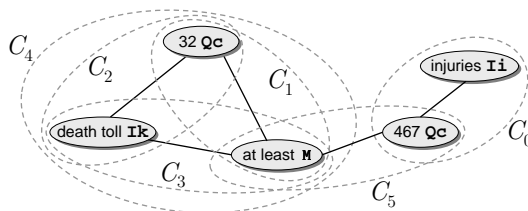


Fig. 3. An entity graph for the example sentence

3 Evaluation and Results

Based on the training data set, we derived optimal extraction pipeline configurations and tested them on the evaluation set. Before presenting our findings, we will explain the evaluation measures used and the underlying configuration parameters.

3.1 Evaluation Measures

To measure the performance of our pipeline, we determined precision (P), recall (R) and F1 measure [23] for all three extraction steps: recognizing entities (NER), extracting entity pairs (BinRE) and synthesizing relationship instances (RE).

Each measure is based on the concept of 'true positive'. We applied a strict evaluation schema, therefore considering a reported entity as a true positive, if and only if both the type and the token agreed with the gold standard [20]. Propagated to the relationship level, an instance was considered a true positive if and only if all participating entities were true positives and the instance had equal size.

3.2 Pipeline Configuration Parameters

Both NER and dependency-based BinRE are requiring well-defined matching criteria. On the entity level, we applied character-based equality. This could be relaxed by case insensitivity (IgnoreCase4NER) or stemming [31,17,23] (UseStem4NER). On the dependency level, we chose between different dependency schemata [25] (DependencySchema). Furthermore, we altered the token vertex matching by case insensitivity (IgnoreCase4RE), stemming (UseStem4RE) or using Part-Of-Speech tags [11,23] (UsePOS4RE). For matching entity vertices, we additionally ignored the subtype (IgnoreEntitySubtype). Moreover, matching pattern edges was modified by ignoring their direction (IgnoreDepDirection) and their label (IgnoreDepType). Given these parameters, Table 2 lists the configurations for maximal precision, recall and F1, estimated from stratified 10-fold cross-validation [13] on the training data.

Table 2. Optimal matching configurations; active: +, inactive: -

Parameter	P _{max}	R _{max}	F1 _{max}	OracleNER F1 _{max}
IgnoreCase4NER	+	+	+	
UseStem4NER	-	+	+	
DependencySchema	CollapsedTree	CCprocessed	Collapsed	CCprocessed
IgnoreCase4RE	-	-	-	-
UseStem4RE	+	-	+	-
UsePOS4RE	-	+	-	+
IgnoreEntitySubtype	+	+	+	+
IgnoreDepDirection	-	+	-	+
IgnoreDepType	-	+	-	+

3.3 Results

Based on the previously deduced pipeline parameters, we evaluated our proposed methods on the evaluation data set. The results for each pipeline step are shown in Table 3. While our approach achieved a surprisingly high recall for entity recognition (93.8%), the corresponding precision was quite low (22.7%). Further analysis revealed that the majority of false positives were produced by the regular expression matching each number in the text (e. g. year or monetary amount).

On the entity pair level, our proposed dependency pattern matching significantly outperformed the baseline in terms of precision (73.0% versus 29.0%). Considering recall, the relation was inverted (74.3% versus 87.2%), resulting in a

Table 3. Evaluation results for different pipeline setups, supplemented by bootstrapped 95% BC_α confidence intervals [9]

Pipeline Setup	NER			BinRE			RE		
	P	R	F1	P	R	F1	P	R	F1
Baseline	.227	.938	.366	.290	.872	.436	.260	.637	.369
	+0.033	+0.018	+0.042	+0.039	+0.035	+0.044	+0.038	+0.068	+0.046
	-.032	-.022	-.043	-.038	-.044	-.046	-.035	-.076	-.045
Dependency P_{\max}	.219	.942	.355	.748	.727	.737	.783	.568	.659
	+0.032	+0.018	+0.042	+0.052	+0.059	+0.046	+0.062	+0.073	+0.061
	-.031	-.022	-.042	-.061	-.067	-.052	-.077	-.078	-.069
R_{\max}	.207	.948	.339	.553	.836	.666	.403	.711	.514
	+0.031	+0.017	+0.041	+0.046	+0.045	+0.039	+0.052	+0.066	+0.051
	-.029	-.022	-.041	-.049	-.059	-.042	-.051	-.078	-.053
$F1_{\max}$.207	.948	.339	.730	.743	.736	.767	.589	.667
	+0.031	+0.017	+0.041	+0.053	+0.058	+0.045	+0.062	+0.072	+0.058
	-.029	-.022	-.041	-.061	-.066	-.050	-.076	-.079	-.069
OracleNER & Baseline				.867	.984	.922	.743	.884	.808
				+0.040	+0.012	+0.026	+0.072	+0.049	+0.060
				-.056	-.028	-.043	-.093	-.077	-.086
& Dependency $F1_{\max}$.930	.906	.918	.813	.826	.820
				+0.031	+0.036	+0.026	+0.070	+0.059	+0.054
				-.057	-.053	-.037	-.124	-.079	-.083

significantly higher F1 measure for the former (73.6% versus 43.6%). Obviously, matching dependency patterns is more insusceptible to low-precision NER than co-occurrence-based classification.

The same tendencies were observed for relationship instances with a significantly better F1 measure of 66.7% versus 36.9%. Additional examination showed that, for both methods, the reported overall precision and recall scores were roughly consistent across instance types (\mathbf{k} , $\mathbf{i} \dots$) and sizes (2, 3...).

Due to EquatorNLP’s pipeline architecture, the observed BinRE and RE performances were certainly biased by the preceding NER step. To quantify the effect of error propagation and therefore disclosing their ‘true’ capabilities, we also tested a perfect NER (OracleNER in Table 2 and 3). Although our results confirmed the global trends for the distribution of precision, recall among the two BinRE methods, their absolute difference in F1 measure were nearly eliminated (82.0% versus 80.8%).

4 Conclusions and Future Work

In this paper, we demonstrated that matching dependency patterns combined with detecting maximal cliques is a promising approach for extracting human

impacts from earthquake reports. Our evaluation on a manual annotated corpus resulted in a maximal F1 measure of 66.7%, outperforming a co-occurrence-based approach significantly. We also showed that our proposed extraction pipeline provides P / R adaptability. Additional experiments with oracle NER imply that under this setting, co-occurrence-based extraction provides competitive results, particularly with regard to its significantly lower computational runtime [2].

Note that the computed recall and F1 measures are slightly biased, since our proposed extraction pipeline operates only on the sentence level and does not cover unary relationship instances. As these instances form approximately 5% of all tuples in news articles (see section 2.2), this might be acceptable for particular applications.

As stated before, our evaluation was focused exclusively on domain specific texts. We cannot expect the same performance for unfiltered texts. The application on 113 general news articles yielded a precision of only 3.2%. This result is less surprising if one takes a closer look at Figure 1, showing that the domain trigger “earthquake” is not part of any shortest path between entities. In fact, only 1 out of all 396 extracted patterns contains a trigger word. Certainly, incorporating semantic knowledge for filtering texts would increase precision. On the other hand, this nonspecificity might be considered as an advantage, suggesting that our pipeline is applicable to other types of disasters.

To finally set the achieved F1 measure of 66.7% in context to a prospective human performance, we assessed this by calculating the inter annotator agreement (IAA) [3] for two independent annotations. The score of 70.3% for strict agreement on relationship instances for 30 articles indicates at least a cardinal task complexity. Great caution should be exercised in comparing these two values directly, since they belong to different dimensions: the former measures validity, while the latter measures objectivity.

4.1 Future Work

Given these results and our conclusions, we identified several challenges for future research. Obviously, we require domain specific texts as pipeline input, proposing text classification as a preprocessing step. As decision makers are interested in information about specific events, we plan to extend our relationship and its extraction to temporal and spatial attributes. Furthermore, we intend to apply high-precision machine learning techniques like condition random fields [18] for NER, hopefully increasing RE recall without losing precision by enabling less strict pattern matching criteria. Finally, we intend to explore user-generated content like on Twitter as a novel information source.

Acknowledgements We kindly thank Sebastian Arzt and Tim Rocktäschel for contributing the IAA annotations; furthermore Samira Jaeger for providing valuable feedback.

References

1. Bies, A., Ferguson, M., Katz, K., MacIntyre, R., Tredinnick, V., Kim, G., Marcinkiewicz, M.A., Schasberger, B.: Bracketing Guidelines for Treebank II Style, Penn Treebank Project (1995)
2. Bjerne, J., Ginter, F., Pyysalo, S., Tsujii, J., Salakoski, T.: Complex event extraction at pubmed scale. *Bioinformatics* 26(12), 382–390 (June 2010)
3. Brants, T.: Inter-annotator agreement for a german newspaper corpus. In: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000) (2000)
4. Bunescu, R., Mooney, R.: A shortest path dependency kernel for relation extraction. In: HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 724–731. Association for Computational Linguistics, Morristown, NJ, USA (2005)
5. Charniak, E.: A maximum-entropy-inspired parser. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. pp. 132–139. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2000)
6. Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and maxent discriminative reranking. In: ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 173–180. Association for Computational Linguistics, Morristown, NJ, USA (2005)
7. Clegg, A., Shepherd, A.: Benchmarking natural-language parsers for biological applications using dependency graphs. *BMC Bioinformatics* 8(1), 24+ (2007)
8. Döhling, L., Woith, H., Fahland, D., Leser, U.: Equator: Faster decision making for geoscientists. In: Proceeding of Workshop on IT support for rescue teams 2011 (2011), (to appear)
9. Efron, B., Tibshirani, R.J.: *An Introduction to the Bootstrap*. Chapman & Hall/CRC (1993)
10. Farmakiotou, D., Karkaletsis, V., Koutsias, J., Sigletos, G., Spyropoulos, C.D., Stamatopoulos, P.: Rule-based named entity recognition for greek financial texts. In: In Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX 2000). pp. 75–78 (2000)
11. Francis, W.N., Kucera, H.: *Brown Corpus Manual* (1979)
12. Fundel, K., Küffner, R., Zimmer, R.: RelEx – Relation extraction using dependency parse trees. *Bioinformatics* 23(3), 365–371 (2007)
13. Han, J., Kamber, M.: *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2nd ed. edn. (4 2006)
14. Isozaki, H., Kazawa, H.: Efficient support vector classifiers for named entity recognition. In: Proceedings of the 19th international conference on Computational linguistics. pp. 1–7. Association for Computational Linguistics, Morristown, NJ, USA (2002)
15. Jenssen, T.K., greid, A.L., Komorowski, J., Hovig, E.: A literature network of human genes for high-throughput analysis of gene expression. *Nature genetics* 28(1), 21–28 (5 2001)
16. Kawtrakul, A., Yingsaree, C., Andrès, F.: A framework of nlp based information tracking and related knowledge organizing with topic maps. In: NLDB. pp. 272–283 (2007)
17. Kraaij, W., Pohlmann, R.: Viewing stemming as recall enhancement. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 40–48. SIGIR '96, ACM, New York, NY, USA (1996)

18. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289. Morgan Kaufmann (2001)
19. Leaman, R., Gonzalez, G.: Banner: an executable survey of advances in biomedical named entity recognition. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing pp. 652–663 (2008)
20. Leser, U., Hakenberg, J.: What makes a gene name? named entity recognition in the biomedical literature. *Briefings in Bioinformatics* 6(4), 357–369 (2005)
21. Li, J., Zhang, Z., Li, X., Hsinchun, C.: Kernel-based learning for biomedical relation extraction. *Journal of the American Society for Information Science and Technology* 59(5), 756–769 (2008)
22. Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press (7 2008)
23. Manning, C., Schütze, H.: *Foundations of Statistical Natural Language Processing*. The MIT Press, 2nd printing w. corrections edn. (6 2000)
24. Marneffe, M., MacCartney, B., Manning, C.: Generating typed dependency parses from phrase structure parses. In: Proceedings of LREC-06. pp. 449–454 (2006)
25. de Marneffe, M.C., Manning, C.D.: *Stanford typed dependencies manual*, revised in february 2010 edn. (2008)
26. McClosky, D., Charniak, E., Johnson, M.: Effective self-training for parsing. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. pp. 152–159. Association for Computational Linguistics, Morristown, NJ, USA (2006)
27. McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y., White, P.: Simple algorithms for complex relation extraction with applications to biomedical ie. In: ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 491–498. Association for Computational Linguistics, Morristown, NJ, USA (2005)
28. Mikheev, A., Moens, M., Grover, C.: Named entity recognition without gazetteers. In: Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. pp. 1–8. Association for Computational Linguistics, Morristown, NJ, USA (1999)
29. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1), 3–26 (January 2007)
30. Pietschmann, S.: Relationship extraction by frequent patterns in dependency graphs (in German). Diplom thesis, HU-Berlin (September 2009)
31. Porter, M.F.: An algorithm for suffix stripping. *Program* 14(3), 130–137 (July 1980)
32. Zhou, G., Su, J.: Named entity recognition using an hmm-based chunk tagger. In: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 473–480. Association for Computational Linguistics, Morristown, NJ, USA (2002)

Associating Relevant Photos to Georeferenced Textual Documents through Rank Aggregation

Rui Candeias and Bruno Martins
rui.candeias,bruno.g.martins@ist.utl.pt

Instituto Superior Técnico, INESC-ID
Av. Professor Cavaco Silva, 2744-016 Porto Salvo, Portugal

Abstract. The automatic association of illustrative photos to paragraphs of text is a challenging cross-media retrieval problem with many practical applications. In this paper we propose novel methods to associate photos to textual documents. The proposed methods are based on the recognition and disambiguation of location names in the texts, using them to query Flickr for candidate photos. The best photos are selected with basis on their popularity, on their proximity, on temporal cohesion and on the similarity between the photo's textual descriptions and the text of the document. We specifically tested different rank aggregation approaches to select the most relevant photos. A method that uses the *CombMNZ* algorithm to combine textual similarity, geographic proximity and temporal cohesion obtained the best results.

1 Introduction

The automatic association of illustrative photos to paragraphs of text is a challenging cross-media retrieval problem with many practical applications. For instance the Zemanta¹ blog enrichment extension is a commercial application capable of suggesting photos from Flickr to blog posts. Another example concerns with textual documents describing travel experiences, usually called *travelogues*, which can give interesting information in the context of planning a trip. Today, there are several websites where these documents are shared and the use of web information for travel planning has also increased. However, the use of the travelogues by themselves is very restrictive. It is our conviction that the visualization of photos associated with specific parts from the travelogue, like common scenarios and points of interest, may lead to a better usage of travelogues.

Despite the huge number of high quality photos in websites like Flickr², these photos are currently not being properly explored in cross-media retrieval applications. In this paper, we propose methods to automatically associate photos, published on Flickr, to textual documents. These methods are based on mining geographic information from textual documents, using a free web service to

This work was partially supported by the Fundação para a Ciência e a Tecnologia (FCT), through project grant PTDC/EIA-EIA/109840/2009 (SInteliGIS)

¹ <http://www.zemanta.com/>

² <http://www.flickr.com>

recognize and disambiguate location names and points of interest mentioned in the documents. The places recognized in the documents are then used to query Flickr for related photos. Finally, the best photos are selected with basis on their popularity and on the similarity between their information (e.g., textual, geographical and temporal metadata) and the information from the document (e.g., textual contents, recognized places and temporal metadata).

The rest of this paper is organized as follows: Section 2 presents the main concepts and related works. Section 3 describes the proposed methods, detailing the mining of geographic information contained in texts and the selection of the best photos, based on their popularity and similarity. Section 4 describes how a system, containing the proposed methods, was implemented. It also presents the results of an initial evaluation experiment. Finally, Section 5 presents our conclusions and points guidelines towards future work.

2 Related Work

Problems related with the treatment of geographic references in textual documents have been widely studied in *Geographic Information Retrieval* [1,11,15,16]. Using this information requires the recognition of place names in the texts (i.e., delimiting the text tokens referencing locations) and the disambiguation of those place names in order to know their real location in the surface of the Earth (i.e., give unique identifiers, typically geospatial coordinates, to the location names that were found). The main challenges in both tasks are related with the ambiguity of natural language. Amitay et al. characterized those ambiguity problems according to two types, namely geo/non-geo and geo/geo [1]. Geo/non-geo ambiguity occurs when location names have a non-geographic meaning (e.g., Turkey, the country or the bird). Geo/geo ambiguity refers to distinct locations with the same name (e.g. London in England and London in Ontario).

Leidner studied different approaches for the recognition and disambiguation of geographic references in documents [11]. Most of the studied methods resolve places references by matching expressions from the texts against dictionaries of location names, and use disambiguation heuristics like default senses (e.g., the most important referenced location is chosen, estimated by the population size) or the spatial minimality (e.g., the disambiguation must minimize the polygon that covers all the geographic references contained in the document). Recently, Martins et al. studied the usage of machine learning approaches in the recognition and disambiguation of geographic references, using Hidden Markov Models in the recognition task, and regression models with features corresponding to the heuristics surveyed by Leidner, in the disambiguation task [15]. Other recent works focused on recognition and disambiguation problems that are particularly complex, involving the processing of texts where geographic references are very ambiguous and with a low granularity (e.g., mountaineering texts mention tracks and specific regions in mountains), and where it is important to distinguish between the location names pertinent to route descriptions and those that are pertinent to the description of panoramas [16].

Currently, there are many commercial products for recognizing and disambiguating place references in text. An example is the Yahoo! Placemaker³ web service, which was used in this work and is better described in Section 3.1.

Previous works have also studied the usage of Flickr as a *Geographic Information Retrieval* information source [4]. The information stored in this service revealed itself to be useful for many applications, due to the direct links between geospatial coordinates (i.e., the coordinates of the places where the photos were taken, either given by cameras with GPS capabilities or by the authors), dates (i.e., the moments when the photos were taken) and text descriptions that are semantically rich (i.e., descriptions and *tags* associated to photos).

In particular, Lu et al. addressed the automatic association of photos, published on Flickr, to Chinese travelogues [14], with basis on a probabilistic topic model detailed on a previous work [8], which is an extension of the Probabilistic Latent Semantic Indexing (pLSA) method [9]. The main idea in the work by Lu et al. is similar to the basis of our work, as the authors tested different methods for the selection of photos, obtained by querying Flickr’s search engine with the location names recognized in the texts. The probabilistic topic model is used by the authors to avoid the gap between the vocabulary used in the documents and the textual descriptions used in photos, modeling photos and/or documents as probabilistic distributions over words. The authors tested four different approaches for the selection of relevant photos, namely (i) a baseline approach based on a simple word-to-word matching with the words from the travelogue texts and the tags that represent the photos (ii) a mechanism based on a probabilistic model created with the travelogue texts (iii) a mechanism based on a probabilistic model created with tags that represent the photos, and (iv) a mechanism based on a probabilistic model using the texts and the tags, which obtained the best results. In our work, we approached the problem in a slightly different way, by querying Flickr with the geospatial information associated with the places recognized in the documents.

In terms of previous works related to the area of cross-media retrieval, Deschacht and Moens presented an approach that tries to find the best picture of a person or an object, stored in a database of photos, using the captions associated to each picture [5]. The authors built appearance models (i.e., language models that represent the text captions from images), to capture persons or objects that are featured in an image. Two types of entity-based appearance models were tested, namely an appearance model based on the visualness (i.e., the degree to which an entity is perceived visually), and another appearance model based on the salience (i.e., the importance of an entity in a text). As baseline approaches, the authors built two simpler appearance models, namely (i) a bag-of-words (BOW) model based on the words of the image captions, and (ii) a bag-of-nouns (BON) model based on the nouns and proper nouns contained in the image captions. From a dataset composed of several image-caption pairs, the authors created two different sets of images annotated with the entities, namely (i) an easy dataset composed of images with one entity, and (ii)

³ <http://developer.yahoo.com/geo/placemaker/>

a difficult dataset composed of images with three or more entities. The results showed that when the dataset was queried with only one entity, the method using the appearance model based on the visualness achieved the best results. On the other hand, when the query was composed of two entities, the method using the bag-of-words had better results.

3 Automatic Association of Photos to Texts

The proposed method for the automatic association of photos to textual documents is essentially based on a pipeline of three stages, which involves (i) recognizing and disambiguating location names and points of interest referenced in documents, (ii) collecting candidate photos through Flickr's API⁴, and (iii) selecting the best photos with basis on their importance and on their similarity (e.g., textual, geographical and temporal) towards the document. In this section we describe the three steps in detail.

3.1 Mining Geographic Information in Documents

In this work, we used the Yahoo! Placemaker web service in order to extract locations and specific points of interest from texts. Placemaker can identify and disambiguate places mentioned in textual documents. The service takes as input a textual document with the information to be processed, and returns an XML document that lists the referenced locations. For each location found in the input document, the service returns also its position in the text, the complete expression that was recognized as the location, the type of location (e.g., country, city, suburb, point of interest, etc.), an unique identifier in the locations database used by the service (i.e., the Where On Earth Identifier - WOEID - used by Yahoo! GeoPlanet⁵), and the coordinates of the centroid that is associated to the location (i.e., the gravity center of the minimum rectangle that covers its geographic area). Also, for each document taken as input, the service returns the bounding box corresponding to the document (i.e., the minimum rectangle that covers all its geographic locations).

3.2 Collecting and Selecting Relevant Photos

The main challenge in collecting and selecting photos relevant to a segment of text is related to the semantic gap between the photo metadata and the text, as well as the noise present in the documents and in the descriptions of the photos. For instance, in the case of travelogues, and despite the fact that these documents have a uniform structure, their authors frequently mention information related to transportation and accommodation, and not only descriptions of the most interesting locations. For example, if the text of a travelogue mentions an airport

⁴ <http://www.flickr.com/services/api/>

⁵ <http://developer.yahoo.com/geo/geoplanet/>

or the city where the trip ends, while describing the arrival, one can select photos related to these locations, which are not important for illustrating the most interesting contents of the document. We have that travelogues frequently mention locations that are only slightly relevant, and so it is very important to distinguish between relevant and irrelevant locations.

Other challenges in collecting and selecting relevant photos are related with the fact that photos published in Flickr are frequently associated to tags or textual descriptions irrelevant to their visual contents (e.g., tags are usually identical among different photos uploaded by the same person, at the same time), and also the vocabulary used in Flickr can be very different from the vocabulary used in textual documents.

Having these limitations in mind, we tested different approaches for the selection of relevant photos, combining different sources of evidence for estimating the relevance of the photos. These approaches are as follows:

- T1: Selection based on textual similarity:** We compute the textual similarity between the tags plus the title of the photos, and the text of the document. Specifically, we compute the cosine measure between the textual descriptions of the photos (i.e., joining tags and title) and the textual document, using the Term Frequency \times Inverse Document Frequency (TF-IDF) method to weight terms in the feature vectors. The idea behind this method is that, if a photo has textual descriptions more similar to the text of a document, then it can be considered as a good photo to be associated to the document.
- T2: Selection based on textual similarity and geographical proximity:** We combined the textual similarity from T1 with the similarity, based on the geospatial coordinates, between the locations recognized in the document and the locations where photos were taken. The geographical similarity is computed according to the formula $\frac{1}{(1+d)}$, where d is the great-circle distance between the two locations. Because multiple locations can be recognized in the document, we computed the maximum and the average similarity towards each photo. The idea behind this method is that a photo that was taken near a location recognized in the document can be considered as a good photo to be associated to the document.
- T3: Selection based on textual similarity, geographical proximity and temporal cohesion:** We combine the method from T2 with the temporal distance, in semesters, between the publication date of the document and the moment when a photo was taken. Similarly to what is done in method T2, the temporal similarity is computed according to the formula $\frac{1}{(1+t)}$, where t is the number of semesters separating the photo from the document. The idea behind this method is that a photo taken in a moment close to the date when the document was written can often be considered as a good photo to be associated to the document.
- T4: Selection based on textual similarity, geographical proximity, temporal cohesion and photo interestingness:** We combine the method T3 with other information related to the interestingness of the photos (e.g., the number of comments and the number of times other users considered the

photo as a favorite). In this case, if a photo was taken in a location inside the bounding box of the document (i.e., the bounding box that contains all locations), then the number of comments and the number of times a photo was marked as favorite are considered as features, and otherwise these features assume the value of minus one. The idea behind this method is that a photo that was taken near the locations recognized in the document, and that is considered an interesting photo due to the number of comments and the number of times users marked it as a favorite, can be considered a good photo to be associated to the document.

The above combination approaches were based on the usage of rank aggregation schemes to combine the multiple features. Specifically, two approaches were considered, namely the *CombSUM* and the *CombMNZ* methods originally proposed by Fox and Shaw [7]. Both *CombSUM* and *CombMNZ* use normalized sums when combining the different features. To perform the normalization, we applied the min-max normalization procedure to the scores of the individual features, which is given by Equation 1.

$$V_{normalized} = \frac{V - min}{max - min} \quad (1)$$

The *CombSUM* score of a photo p , for a given document D , is the the sum of the normalized scores received by the photo in each of the k individual rankings, and is given by Equation 2.

$$CombSUM(p, D) = \sum_{j=1}^k score_j(p, D) \quad (2)$$

Similarly, the *CombMNZ* score of a photo p for a given document D is defined by Equation 3, where r_e is the number of non-zero similarities.

$$CombMNZ(t, P) = CombSUM(t, P) \times r_e \quad (3)$$

For measuring the similarity between the textual description of the photos and the text of the document, in all the above methods, stopwords were first removed. To calculate the cosine measure between the photos textual descriptions and the document, using the Term Frequency \times Inverse Document Frequency (TF-IDF) method, we considered tags to be more important to describe the photo, followed by the title. Thus, we applied different weights for the different types of textual descriptions, weighting the tags as twice more important.

4 Validation Experiments

We implemented a prototype system based on the techniques described in the previous section, using the Qizx⁶ XQuery engine as an execution environment.

⁶ <http://www.xmlmind.com/qizx/>

This XQuery engine supports the latest version of the standard, together with the XQuery Full Text extension to perform full-text search with the cosine measure and TF/IDF vectors, in collections of XML documents.

In order to validate the proposed methods, we created a corpus of 450 photos downloaded from Flickr, with geographical information and a sufficiently large textual description (i.e., more than 100 words and containing location names or points of interest). We used expressions frequently used in travelogues, such as *monument*, *vacation*, *trip* or *castle* to filter the photos collected from Flickr. The collected photos were taken in a point contained in the bounding box corresponding to the geospatial footprint of one of the world's most visited cities⁷. Also, the considered photos were taken in a date from 2000-01-01 to 2010-05-01. For each photo, the number of comments and the number of times it was considered as favorite by other users were also collected.

In order to conduct the experiments, we needed a collection of documents with relevance judgments for photos, i.e., a correct relevant photo associated to the document. This collection was not already available and creating a collection of travelogue documents, illustrated with Flickr photos that had been manually selected by human experts, would be extremely time consuming, also implying some knowledge about the locations described in the documents. This collection is not already available, and creating a collection of photos from Flickr selected and associated by experts to travelogues would be extremely time consuming, and would imply a certain knowledge of the city to where the travel was made.

The photo descriptions from Flickr, with the above characteristics, are fairly good examples of documents with relevance judgments, because the owner considered the photo as a relevant example to be associated to the large textual description. So, for the purpose of our experiments, we considered the textual descriptions as representations of textual documents having the same characteristics as travelogues, and the photos from which the textual descriptions were taken as the relevant photos that should be automatically associated.

The prototype system, implementing different configurations for the proposed method, was then used to process the documents, associating them to relevant photos. The configurations used are described in Section 3.2.

With the results for each document, and considering all four possible configurations with the two voting schemes, we used the `trec.eval` evaluation tool to evaluate the matchings between photos and documents. Figure 1 presents the results obtained in terms of Precision at position 1 (Precision@1), and in terms of the Reciprocal Rank, in the all the considered cities. The horizontal lines represent the mean value of Reciprocal Rank, in red, and the mean value of Precision@1, in blue, for all the considered cities and when using the best configuration. In all the charts, the bar in red, full colored, represents the value of Reciprocal Rank, and the bar in blue, with a shaded color, represents the value for the metric of Precision@1.

The graphics show that method T3 using the *CombMNZ* approach (i.e., T3-MNZ) outperforms method T1 in all the cities. These results suggest that the

⁷ <http://en.wikipedia.org/wiki/Tourism>

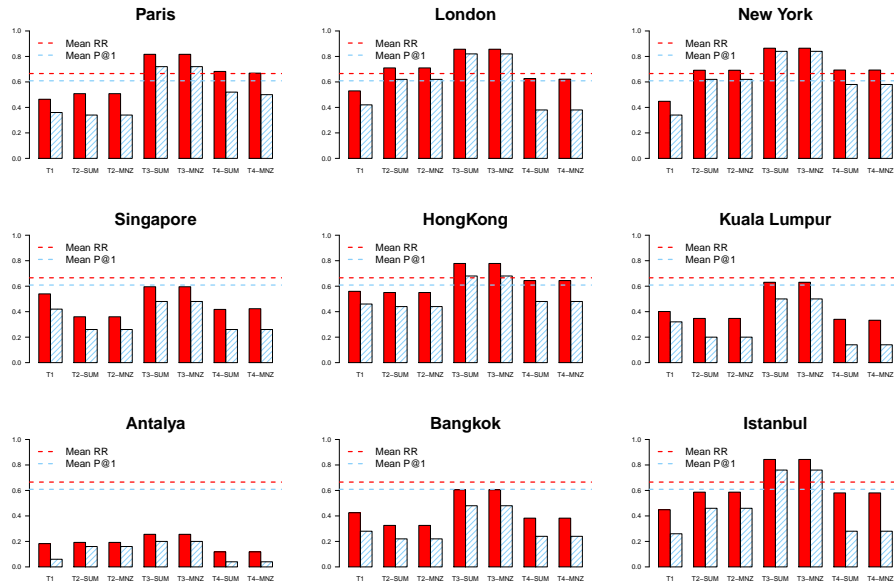


Fig. 1. Reciprocal Rank and Precision@1 for each method and city.

usage of multiple features (e.g., geographical proximity and temporal cohesion) combined with the textual similarity is better than the usage of the textual similarity alone. Also, methods using *CombMNZ* as the rank aggregation approach have similar results to the methods using *CombSUM*.

It is also interesting to notice that the values in the cities of Paris, London and New York are higher, although the dataset contained an equal number of photos for each city (i.e., 50 photos). In these cities, all the combination methods using *CombMNZ* outperform method T1. These results suggest a higher precision of Placemaker in the recognition and disambiguation of the location names mentioned in the descriptions for those cities, although it should be noticed that textual similarity alone also presents good results in these cities.

Figure 2 illustrates the obtained results for two example textual descriptions, presenting the top-3 most relevant photos as returned by the best performing method, together with their tags in Flickr.

Figure 3 presents the number of documents, in the collection, containing each possible number of words, and the number of documents mentioning different numbers of places. In the collection, there is a higher number of documents with 100 to 200 words. Also, the number of recognized places is frequently low, with most of the documents containing 1 to 5 places.

Figure 4 illustrates the relationships existing between the values of Precision@1 and Reciprocal Rank, with the number of words and the number of places, when considering the combination method that had the best results, i.e.,







<p>The Louvre Pyramid is a large glass and metal pyramid, surrounded by three smaller pyramids, in the main courtyard of the Louvre Palace in Paris. The large pyramid serves as the main entrance to the Louvre Museum. Completed in 1989, it has become a landmark for the city of Paris. The construction of the pyramid triggered considerable controversy because many people felt that the futuristic edifice looked quite out of place in front of the Louvre Museum with its classical architecture.</p>			
	<p>id=3756841917</p>	<p>id=3784201917</p>	<p>id=3418929006</p>
	<p>louvre paris pyramid palace europe vacation night reflections summer museum</p>	<p>paris france pompidou frontpage longexposure centrepompidou longexposure nikon guidomusch parijs</p>	<p>paris france colour seine boat filters lee leefilters sunset light</p>
<p>Water, without which, we would not be on Earth. This was captured to show that Kuala Lumpur is a blend of old buildings as well as new. This fountain sits on one end of the famous Selangor Club field used for Merdeka Day Celebrations. Reggie Wan of Singapore and I were here on this bright and hot day. Couldn't get out fast enough away from this tourist spot! Actually, all the photos showcased here were pretty nice and I really couldn't decide which one to be the main picture. I chose this one because it was more artistic (the fountain is dark, the tall building medium grayish and the minaret is white surrounded by blue)!</p>			
	<p>id=294945309</p>	<p>id=4194402499</p>	<p>id=327607369</p>
	<p>water fountain buildings sky tourist aqua cityscape reflections jalanraja fujifilm</p>	<p>kuala lumpur petronas explore digital blending dynamic nikon range malaysia</p>	<p>building kualalumpur malaysia soe southeastasia reggiewan asia top20travelpix klcc mosque</p>

Fig. 2. The top three most relevant photos returned for two example documents.

T3 using *CombMNZ*. These results suggest that a higher number of words does not improve the results, neither in terms of Precision@1 or Reciprocal Rank. The higher value of Reciprocal Rank and Precision@1 in documents with 1200 to 1300 words can be explained by the corresponding small number of documents (i.e., only 2 documents). It is also interesting to notice that the values for Precision@1 and for the Reciprocal Rank seem to improve when more than one place is referenced in the document.

5 Conclusions and Future Work

In this paper, we have described novel methods for the automatic association of photos to textual documents. The described methods are based on a pipeline of three steps, in which geographic references are first extracted from documents,

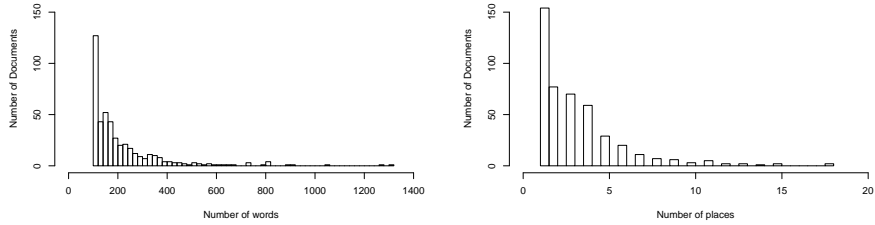


Fig. 3. Histograms with the number of words and the number of places.

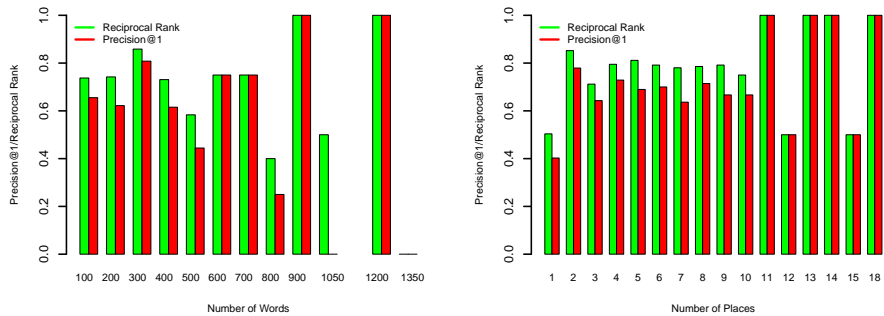


Fig. 4. Variations in the values of Precision@1 and Reciprocal Rank, in terms of the number of words and the number of places referenced in the documents.

then photos matching the geographic references are collected, using Flickr’s API, and finally the best photos are selected with basis on their similarity and relevance. Different methods to select relevant photos were compared and a method based on the combination of textual similarity, geographic proximity and temporal cohesion, using the *CombMNZ* rank aggregation method for performing the combination, obtained the best results.

Despite the good results from our initial experiments, there are also many challenges to future work. From our point of view, the major challenge lies in improving the evaluation protocol. The validation of the proposed methods should be made through a collection of static photos, with relevance judgments clearly established by humans. The Content-based Photo Image Retrieval (CoPhIR) collection, described in [3] and built from 106 million photos from Flickr, could be a starting point for building such a test collection. Another idea is to experiment the proposed methods in a collection not related to the domain of travelogues. For instance, the dataset with news texts from BBC which was described by Feng

and Lapata [6], containing approximately 3400 entries and where each entry is composed by a news document illustrated with a image that contains a textual caption, could also be used to as a starting point to build a better test collection to evaluate our method. This corpus contains near 3400 entries, where each entry is composed by a news document, a news image related with the document and its caption. Also, besides the usage of the cosine similarity to measure the textual similarity between photos and documents, it would be interesting to use different methods, for instance based on probabilistic topic models such as the Latent Dirichlet Allocation (LDA) model [2].

It would also be interesting to experiment with supervised learning methods for combining the different relevance estimators. Several supervised learning to rank methods [13,12], recently proposed in the information retrieval community to address the problem of ranking search engine results, could be used to develop models that can sort photos based on their relevance, considering different sources of evidence (i.e., several similarity and importance metrics). Recent works in the area of information retrieval have also described several advanced unsupervised learning to rank methods, capable of outperforming the *CombSUM* and *CombMNZ* approaches. This is currently a very hot topic of research and, for future work, we would for instance like to experiment with the ULARA algorithm, which was recently proposed by Klementiev et al. [10].

References

1. E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, 2004.
2. D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.
3. P. Bolettieri, A. Esuli, F. Falchi, C. Lucchese, R. Perego, T. Piccioli, and F. Rabitti. Cophir : A test collection for content-based image retrieval. Technical report, Institute of Information Science and Technologies, National Reasearch, Pisa, Italy, 2009.
4. D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World Wide Web*, 2009.
5. K. Deschacht and M. Moens. Finding the best picture: Cross-media retrieval of content. In *Proceedings of the 30th European Conference on Information Retrieval*, 2008.
6. Y. Feng and M. Lapata. Automatic image annotation using auxiliary text information. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2008.
7. E. A. Fox and J. A. Shaw. Combination of multiple searches. In *Proceedings of the 2nd Text Retrieval Conference*, 1994.
8. Q. Hao, R. Cai, X. Wang, J. Yang, Y. Pang, and L. Zhang. Generating location overviews with images and tags by mining user-generated travelogues. In *Proceedings of the 17th ACM international Conference on Multimedia*, 2009.

9. T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR conference on Research and development in information retrieval*, 1999.
10. A. Klementiev, D. Roth, K. Small, and I. Titov. Unsupervised rank aggregation with domain-specific expertise. In *Proceedings of the 21st International Joint Conference on Artificial intelligence*, 2009.
11. J. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh, 2007.
12. H. Li. *Learning to Rank for Information Retrieval and Natural Language Processing*. Morgan & Claypool Publishers, 2011.
13. T. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3), 2009.
14. X. Lu, Y. Pang, Q. Hao, and L. Zhang. Visualizing textual travelogue with location-relevant images. In *Proceedings of the 2009 international Workshop on Location Based Social Networks*, 2009.
15. B. Martins, I. Anastácio, and P. Calado. A machine learning approach for resolving place references in text. In *Proceedings of the 13th AGILE International Conference on Geographic Information Science*, 2010.
16. M. Piotrowski, S. Liubli, and M. Volk. Towards mapping of alpine route descriptions. In *Proceedings of the 6th ACM Workshop on Geographic information Retrieval*, 2010.