International Journal of Geographical Information Science
Vol. 00, No. 00, Month 2014, 1–23

RESEARCH ARTICLE

Thematic Signatures for Cleansing and Enriching Place-Related Linked Data

Benjamin Adams[a*] and Krzysztof Janowicz[b]
[a]*Centre for eResearch, Department of Computer Science, The University of Auckland, New Zealand*
[b]*Department of Geography, University of California, Santa Barbara, USA*
(*Received 00 Month 200x; final version received 00 Month 200x*)

There has been significant progress transforming semi-structured data about places into knowledge graphs that can be used in a wide variety of geographic information systems such as digital gazetteers or geographic information retrieval systems. For instance, in addition to information about events, actors, and objects, DBpedia contains data about hundreds of thousands of places from Wikipedia and publishes it as Linked Data. Repositories that store data about places are among the most interlinked hubs on the Linked Data cloud. However, most content about places resides in unstructured natural language text, and therefore it is not captured in these knowledge graphs. Instead, place representations are limited to facts such as their population counts, geographic locations, and relations to other entities, e.g., headquarters of companies or historical figures. In this paper we present a novel method to enrich the information stored about places in knowledge graphs using *thematic signatures* that are derived from unstructured text through the process of topic modeling. As proof-of-concept, we demonstrate that this enables the automatic categorization of articles into place types defined in the DBpedia ontology (e.g., *mountain*), and also provides a mechanism to infer relationships between place types that are not captured in existing ontologies. This method can also be used to uncover miscategorized places, which is a common problem arising from the automatic *lifting* of unstructured and semi-structured data.

1. Introduction

The complexity and ambiguity of *place* is well-documented in the field of geography, and thus has posed a riddle for those interested in creating large-scale computational representations of knowledge about places for information systems (Entrikin 1991, Massey 1991, Cresswell 2004, Winter *et al.* 2009). The path of least resistance for building such

*Corresponding author. Email: b.adams@auckland.ac.nz

ISSN: 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/1365881YYxxxxxxxx
http://www.informaworld.com

knowledge bases is to draw upon properties of places that are already recorded in a semi-structured manner on the Web, e.g., from tabular data (Auer *et al.* 2009). However, a tremendous amount of place knowledge remains obscured from formal computational representation, because it exists in formats designed for human, not machine, consumption. Unstructured data about places, such as crowdsourced natural language narratives and descriptions, offer rich and varied potential sources for building place representations because they provide thematic contextualizations of places that are not designed for a specific domain and its applications (Adams and McKenzie 2013, Vasardani *et al.* 2013). Having access to these different contexts of places is important because the identity of a place is fluid and depends upon social factors. Likewise the applications designed to utilize place knowledge bases will be similarly context-dependent. While the value of unstructured data about places has been recognized, methods to translate these unstructured data into computational representations that can interoperate with existing place models are largely missing. In this paper we propose a novel methodology that uses representations of latent properties built from machine learning techniques and combines them with top-down semantic engineering. By doing so we provide an effective means to integrate unstructured knowledge with the structures in existing place knowledge bases. We demonstrate our methods by utilizing DBpedia and its source Wikipedia.

In recent years substantial progress has been made on extracting structured facts from textual sources and building knowledge bases for structured querying (Medelyan *et al.* 2009). For example, researchers developing DBpedia and YAGO have extracted millions of facts from Wikipedia and other sources and published this information as Linked Data (Suchanek *et al.* 2007, Bizer *et al.* 2009). Although DBpedia has been successful at extracting place knowledge from structured tables and categories in Wikipedia, the majority of the semantic content in Wikipedia resides in the actual natural language text of the articles. Recently, Gangemi (2013) provided an overview and comparison of Semantic Web tools for knowledge extraction. A large number of these extracted facts are about places, e.g., types of places, point locations, relations to other places, and additional attributes. Typically, however, these tools focus on the extraction of relations and entities as well as learning ontologies. In contrast, we are interested in understanding the place types (i.e., classes) and taxonomies used by these tools, most of which rely on DBpedia/Wikipedia categories without questioning them. We investigate how stable these place types really are and whether we can automatically discover miscategorized place instances. To do so, we propose methods to derive *thematic signatures* for the instances and types in the DBpedia ontology from the natural language text in the Wikipedia pages associated with those entities.

Thematic signatures consist of characteristic *bands* that enable one to classify observations of geographic features into their types based on thematic (non-spatiotemporal) attributes. For the method presented in this article, the bands are characteristic mixtures of topics found in natural language articles about individual places. These signatures allow us to investigate regularities among and differences between places. For example, the Wikipedia article about *Napa Valley* consists largely of terms related to *wine* topics. Other articles about wine regions will share those topics. In contrast, a Wikipedia article about a different kind of place, e.g., the city of *Reykjavik*, will not. In addition, because thematic signatures are derived bottom-up from written descriptions of places they provide access to attributes that are important to people but that are often not quantitatively expressed in tabular place data (Edwardes and Purves 2007). In particular, this text will often refer to activities common to a place as well as historical events. Since the proposed thematic signatures are probabilistically defined, they also provide a

mechanism to model the vagueness of places in terms of their attributive model. Modeling this additional dimension of place vagueness complements existing work on using data to derive spatial representations of places (Montello *et al.* 2003, Jones *et al.* 2008).

Formal place ontologies restrict the interpretation of the meaning of named places and their types (e.g., cities and towns) to foster semantically interoperability between a multitude of data sources and applications. However, as with most formal ontology engineering, the place ontologies that have been created for DBpedia, schema.org, and similar efforts, reflect the particular choices of the developers who created them, the specific tools and methods that they used, and the (semi) structured data and taxonomies that were available to those developers. Therefore, existing place ontologies may be of limited applicability to many of their potential users. Similarly, such top-down categorization always introduce specific *ontological commitments* and biases. Thus, it is worth studying how well these ontologies represent the stored data. For example, one branch of an ontology may be very detailed in terms of the introduced subclasses, while another branch may remain on a more general level. Similarly, place types that may be semantically similar from the perspective of the stored instances, may be far away in the taxonomy provided by the ontology, and thus hinder semantics-based information retrieval. In contrast to place ontologies, the contents of the place articles in Wikipedia, a massively crowdsourced resource that has been created by a community of thousands of contributors, are the antithesis of one viewpoint or conceptual organization. It stands to reason that the categorizations that we want to use to describe the semantics of this content also be similarly broadly constructed, because this will lead to a robust categorization that is more applicable to the users. Toward this end, we advocate an observation-driven approach to representing the semantics of places (Janowicz 2012). This is not to say that bottom-up approaches do not reflect thematic biases either in terms of the distribution of themes in the source data or in terms of the choices of specific parameterisations in the statistical techniques employed. But they provide a useful counterpoint to the top-down approach, one where the biases and model used to uncover patterns are arguably more transparent and open to iterative improvement.[1]

The research contributions of the presented work are as follows:

- We explore how the thematic signatures for place *instances* (of which there are more than 500,000 associated English Wikipedia pages) can be used to identify thematic signatures for place *types*.
- We showcase how the semantic signatures can be applied to find anomalous thematic content for a place type instance, thereby allowing to clean big data sources that cannot be curated by hand.
- We demonstrate how to infer new subsumption relationships between place types and thus how ontologies can be semi-automatically enriched from a data-driven perspective.
- Finally, we point to new ways of disambiguating semantically dissimilar subclasses that are currently lumped in a single class in DBpedia.

For purposes of demonstration we will focus our evaluation on geographic knowledge in DBpedia, but the techniques are equally applicable to other Linked Spatiotemporal Data repositories extracted from heterogeneous, crowdsourced data.

The organization of the remainder of this article is as follows. Section 2 describes background on topic modeling and semantic signatures. Section 3 introduces our methodology

[1] A full discussion on the iterative process of building, testing, and improving statistical models for pattern discovery in big data is beyond the scope of this article and we suggest readers who are interested in this topic to read Blei (2014).

to generate thematic semantic signatures for DBpedia places and place types from natural language text. We also demonstrate their effectiveness for Linked Data cleansing. In Section 4 we present additional use cases for these signatures by giving concrete examples that showcase their abilities over a range of application areas. Finally, in Section 5 we conclude and present directions for future work.

2. Background Materials and Related Work

2.1. *Places Types in DBpedia*

DBpedia is a continually evolving knowledge base of entities that have been extracted from structured and semi-structured information in Wikipedia (Lehmann *et al.* 2014). Currently it is the most interlinked hub on the Linked Data cloud, GeoNames being second.[1] DBpedia also defines an ontology that describes a classification scheme for the entities. One part of this ontology defines subclasses of *Place*. Figure 1 shows this subsumption hierarchy of places. The DBpedia knowledge base and ontology are often revised; we use version 3.6 for our research. Most revisions add data and enrich some parts of the ontology, while the key classes (including the subclasses of *Place*) are more stable. Finally, it is worth mentioning that the DBpedia ontology aims at classifying *topics*. Consequently, the distinction between classes and individuals differs from many typical ontologies. Therefore, OWL2 DL *punning* is often used to type entities using DBpedia. While the USGS Topographic Ontology[2] is richer in terms of the defined classes, both ontologies use a *surface semantics* approach, i.e., the ontologies merely consist of subclass relations together with domain and range restrictions for the introduced binary relations. Consequently, it is difficult (and often impossible) to understand and distinguish the classes from a formal perspective. For instance, why are all bodies of water (including canals) categorized as *natural places* in the DBpedia ontology? Is the global distinction into cities and towns supported by data despite the highly *local* character of these classes (Janowicz 2012)? The proposed data-driven thematic signatures approach will allow us to answer such questions despite the lack of a deeper axiomatization of the DBpedia ontology. The goal is not so much to create a *more formal* ontology through a bottom-up approach, but rather to interrogate the designed ontology to see if the unstructured data supports it and if not provide one or more alternatives that do.

2.2. *Topic Modeling*

Latent Dirichlet allocation (LDA) is an increasingly popular unsupervised learning technique used to identify the latent structure of topics in a corpus of natural language text (Blei *et al.* 2003). Each document in the corpus is modeled as a mixture of topics, which themselves are multinomial distributions over terms. LDA is a generative model because it describes how the words in a set of documents are generated by a random process, given the distribution of topics over each document and the distribution of words over each topic. LDA inference is the task of finding the topics that best fit the observed data given this model. Since the topics learned via LDA inference are semantically coherent collections of terms, it is an effective way of reducing the dimensionality of the term space

[1] See `http://lod-cloud.net/versions/2014-08-30/lod-cloud_colored.png` for a visualization of the LD cloud.
[2] http://cegis.usgs.gov/ontology.html

- Place
  - Architectural structure
    - Building
      - Historic building
      - Hospital
      - Hotel
      - Lighthouse
      - Museum
      - Restaurant
      - Shopping mall
      - Stadium
      - Theatre
    - Infrastructure
      - Airport
      - Launch pad
      - Power station
      - Route of transportation
        - Bridge
        - Public transit system
        - Railway line
        - Road
        - Road junction
        - Tunnel
          - Railway tunnel
          - Road tunnel
          - Waterway tunnel
      - Station
    - Park
  - ...
- Place *(continued)*
  - Historic place
  - Monument
  - Natural place
    - Body of water
      - Lake
      - Stream
        - Canal
        - River
    - Cave
    - Lunar crater
    - Mountain
    - Mountain range
    - Valley
  - Populated place
    - Administrative region
    - Atoll
    - Continent
    - Country
    - Island
    - Municipality
      - City
      - Town
      - Village
  - Protected area
  - Ski area
  - Wine region
  - World heritage site

Figure 1. Subsumption hierarchy of place types as defined in the DBpedia ontology.

for a corpus in such a way that the dimensions are also interpretable. To give a visual impression, Figure 2 illustrates the kinds of latent topics LDA inference identifies from the Wikipedia corpus. This word cloud representation shows the top-20 terms of 3 topic, where the size of the term is based on their relative probability weight in the topic.

(a) Building topic (b) Army topic (c) Greek topic

Figure 2. Three topics from a set of LDA topics trained using a corpus of Wikipedia place articles.

Once LDA inference is performed, the distribution of topics for documents can be used to compare their similarity. Two common approaches are the Kullback-Leibler (KL) divergence and the Jensen Shannon (JS) divergence (Steyvers and Griffiths 2007). KL divergence (Equation 1) is an asymmetric measure of the divergence of one multinomial distribution, P, from another, Q. JS divergence (Equation 2) is a symmetric measure

derived from the KL divergences of both distributions from an average distribution.

$$KL(P \parallel Q) = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)} \tag{1}$$

$$JS(P \parallel Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M), \text{ where } M = \frac{1}{2}(P + Q). \tag{2}$$

2.3. *Semantic Signatures*

Semantic signatures (Janowicz 2012) are an analogy to spectral signatures used in remote sensing where geographic features on the earth's surface are recognized via their unique reflection and absorption patterns in different bands, i.e., wavelengths, of electromagnetic energy. Studying the reflection values within a particular band, such as the visible light, is often sufficient to distinguish between land use types. For instance, brick buildings can be easily distinguished from paving concrete. In other cases, however, the reflection patterns become only distinguishable in the near infrared or thermal band. Consequently, multiple bands have to be sensed. A typical example are conifers and deciduous trees as they are difficult to distinguish by merely comparing the visible band.

In analogy, semantic signatures can be defined via several bands by using spatial analysis, data mining, machine learning techniques, identity criteria, and so forth. The signatures are not derived from wavelengths but other, often derived, observations. In other words, semantic signatures apply concepts and methods from remote sensing to other types of data, e.g., from social media. In previous work *spatial* signatures have been computed from volunteered geographic information sources such as OpenStreetMap (Mülligann *et al.* 2011). Along the same lines, *temporal* signatures have been derived from Location-based Social Networks (Ye *et al.* 2011). Together with the thematic bands introduced in this work, the spatial and temporal bands can either be considered separately or regarded as bands that jointly form a more complex signature. Figure 3 illustrates the formal relationships between geographic features, signatures, and bands. So far, semantic signatures have been used for data cleansing, conflation, and for the prediction of labels for unnamed geographic locations. they form the ground level for the observation-driven ontology engineering (ODOE) methodology (Janowicz 2012).

3. Calculating Thematic Signatures for Data Cleansing

In this section we outline a method for deriving thematic bands of places and their types from analyzing unstructured data in Wikipedia and demonstrate how these signatures can be used for data cleansing.

3.1. *Creating Thematic Signatures from Articles*

The following enumeration outlines the thematic semantic signature extraction and analysis steps that will be discuss in more detail in the remainder of this article. These steps constitute the preprocessing, training, and extraction of the thematic bands. Shannon entropy is computed as intermediate step for both the instance and type level data to understand the LDA topic variability. Intuitively, place types that are dominated by a

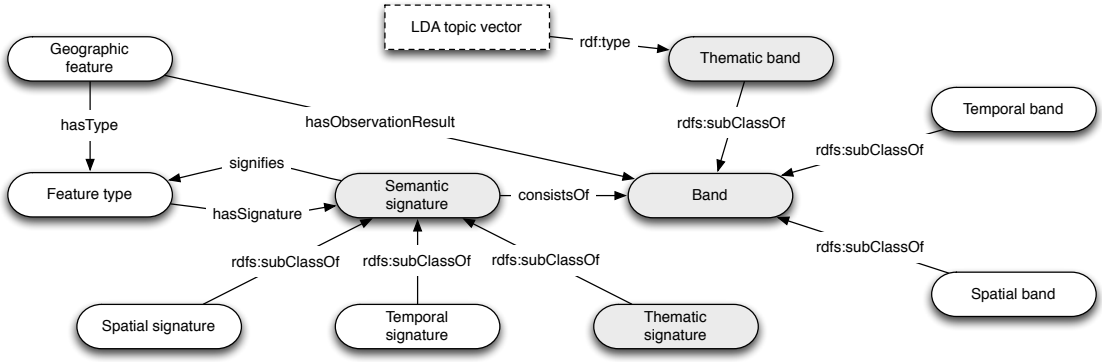


Figure 3. This diagram illustrates the ontological relationships between geographic features, feature types, signatures, and bands. A set of bands is a signature only if it is sufficient to classify observations into their feature types (i.e., the *signifies* relation). Spatial, temporal, and thematic signatures are semantic signatures that consist only of spatial, temporal, and thematic bands, respectively. In addition, complex semantic signatures are possible that combine spatial, temporal, and thematic bands. In this article we focus on thematic signatures that are composed of thematic bands defined using LDA on geographic feature descriptions. See also (Mülligann *et al.* 2011) and (Ye *et al.* 2011) for applications of spatial and temporal signatures.

few topics, e.g., *lunar crater*, will have stronger discriminatory power and will be placed deeper in a place type hierarchy than types that are represented by a wide variety of equally representative topics, e.g., *historic place*.

(1) Identify all place resources in DBpedia.[1] Let M be the number of place resources.
(2) Match the DBpedia resource identifier to page titles; get the Wikipedia article texts and prepare for input into LDA.
(3) Train LDA on all *place articles* (i.e., >500,000 Wikipedia articles categorized by subtypes of the place category).
    - This results in a topic distribution vector for each place article.
    - Compute information entropy for the topic vector to study how many or few topics characterize the article.
(4) Identify articles for each of the 67 Place subclasses in the DBpedia ontology.
(5) Create a new super-document for each place type that combines all the text from all articles tagged to the place type in question.
(6) Use the existing trained LDA model (from step 3) to infer a topic distribution for each of these *place type* super-documents.
    - This results in a topic distribution vector for each place **type**.
    - The entropy of the topic vector indicates how many or few topics characterize the place type.

We leverage the DBpedia knowledge base to find articles that are about specific feature types (Bizer *et al.* 2009). The DBpedia community has done a significant amount of data extraction by matching Wikipedia templates to well-formed classes in an ontology including several *place* subclasses. Similar to spatial and temporal bands, we use the semantic signature approach to characterize feature types with thematic signatures derived from LDA inference. Each feature type thereby corresponds to a vector in an n-dimensional space where n is the count of LDA topics and the probability values for each topic form the vector components. In our initial experiments we set n to 50, 300, and 1000 (follow-

[1]We used Factforge, a popular SPARQL endpoint for the web of data. `http://factforge.net/`

ing recommendations in (Griffiths and Steyvers 2004)). Changing this parameter does change the granularity (in terms of numbers of topics) at which places are compared. For a specific application scenario, we anticipate that analysts and ontology engineers would generate multiple runs of LDA with different numbers of topics and examine the results of each. Alternately, one can substitute a more sophisticated model based on Dirichlet processes, which not only train for topics but also optimize the number of topics based on the statistics of the corpus (Teh *et al.* 2006). All results shown subsequently are based on 300 topics. As the probabilities per topic differ for each feature type, we use these semantic signatures to uncover latent semantic structures, which allow us to differentiate between feature types.

Using the following SPARQL query we can find all the DBpedia resources associated with a place class in the DBpedia ontology:

```
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?c WHERE ?c rdf:type dbpedia-owl:ClassName
```

Next an LDA model is trained on the English Wikipedia corpus (or subset as appropriate) to uncover a set of latent topics that will be used to create the semantic signatures. Each article in the corpus is thereby described as a probability distribution over the topics (i.e., its thematic semantic signature).

The DBpedia place resources that are returned from the above SPARQL query are matched to original articles from the latest English Wikipedia database dump. To generate a thematic signature for a place class we aggregate the text from all articles for the class into a single document. Once we have this new aggregated document we can infer a topic distribution for it from the existing model using the Markov-Chain Monte Carlo method (Wallach *et al.* 2009). An alternative method is to average over the topic probability vectors for all the associated articles, but that has some disadvantages that the proposed method improves on. First, averaging over documents makes all documents weighted equally, whereas combining the text allows the length of an article impose a relative weight. Second, averaging will tend to create a flatter topic distribution the more articles are associated with the class. By using the topic model to infer the topic distribution, this flattening effect is mitigated because the distribution of topics for a document is determined by the optimized Dirichlet distribution for the data.

This distribution of topics for all the aggregated text of a class can be interpreted as the distribution of topics that would exist for the article of a prototypical instance of the class. We can use this semantic signature for the feature type to infer more about the semantics of instances in DBpedia than currently exist in the knowledge base. A *feature type signature* for an instance can be generated by calculating the similarity (i.e., KL divergence) of the topic distribution for the instance's associated article to the semantic signature for every class. By doing so, we get a feature type profile of an instance that goes well beyond simple classification and one that points to the discovery of relations based on similarities to multiple types through the semantic content of the article. As outlined above, arriving at a data-driven deeper understanding of the feature types is important as DBpedia only provides a flat subclass hierarchy.

3.1.1. Thematic Entropy of Feature Types

While some thematic bands for feature types contain high weights for diagnostic LDA topics, e.g., in the case of *city* and *museum*, others, such as *hospital*, do not. The diagnosticity of topics and, hence, bands, is quantified by measuring the entropy of the

Type	**Entropy**	**Type**	**Entropy**
City	4.8	Museum	4.79
ProtectedArea	4.62	Hotel	4.39
Lake	4.54	Restaurant	4.11
Mountain	4.41	Hospital	3.53

Table 1. Entropy values of selected meso-scale and micro-scale feature types.

probability vector (see Equation 3).

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{3}$$

Topics with a high entropy can be considered more geo-indicative, i.e., local; bands with a high entropy value have high weights for such topics. Table 1 shows entropy values for selected meso-scale feature types (left side) as well as those on micro-scale (right side). The entropy of a feature type can best be interpreted as a measure of how many different topics contribute to the characteristic make-up of the feature type. For example, *hospital* features have low entropy, which means that there are a few dominant topics that are found on average in articles about hospitals. Similarly, on average *city* articles will tend to have more different topics contributing to the text.

3.2. *Using Bands for Data Cleansing*

The following analysis steps introduce different views on the signatures, which can be used for data cleansing.

(1) Calculate an N x N matrix of JS divergence values between the N place types.
(2) Input the divergence matrix into multidimensional scaling (MDS) and hierarchical clustering algorithms to identify how place types are similar with respect to the topics mined from Wikipedia place articles.
(3) Calculate an N x M matrix of KL divergence values between N place type topic vectors and M place article vectors. For each place type, T, sort the articles by divergence value and bin the results into two groups:
   - Articles tagged as type T in DBpedia.
   - Articles not tagged as type T.
(4) Perform kernel density smoothing for both bins for each type in order to characterize how similar the instances and non-instances of place types are to the *prototypical* place type instance.
(5) Examine the instances for each type that have greatest KL divergence from the place type topic vector, and evaluate if they are misclassified or under-classified instances.

3.2.1. Understanding the Semantics of Place Types Using Divergence Curves

To evaluate their predictive capability we computed the KL divergences between all place *type* signatures and the semantic signatures for every *place* article in the corpus. For each place type the articles are binned into two groups: articles that are classified for that place type and those that are not. To create a smoothed curve we computed a kernel density estimation over the divergence values for both groups and plotted them on the same chart (Bowman and Azzalini 1997). Table 2 shows the resulting curves for

44 place types in the DBpedia ontology.

By comparing the divergence curves for place instances that are classified as a certain place type to those that are not, we can evaluate how characteristic (or prototypical) the semantic signature of the type is. In the charts shown in Table 2, the blue curves represent divergences of places instantiating the type, and the red dashed curve represents other place instances not classified as this type. For a place type signature that is *prototypical* one would expect the blue curve to reside primarily to the left (i.e., lower divergence / higher similarity) of the red curve. For many cases this is the result we see. For example, *hospital*, *bridge*, *lighthouse*, *cave* all exhibit this effect. In contrast, *administrative region*, *city*, and *world heritage site* all show significant overlap in the blue and red curves, which indicates that there is no single *prototypical* signature of themes for this type.[1] There are a few possible explanations for this. First, there is great variety in the types of topics used to describe these instances in the Wikipedia articles. *World heritage site* is exemplary of this as each world heritage site is chosen for its unique, defining characteristics and this is reflected in a great range of topics in the texts. Second, there are many other-classified instances that are very similar thematically to the articles of some types. This effect is seen most strongly when examining *administrative region* and *city* place types, which have instances that overlap greatly thematically. Finally, it is possible that feature types that have partonomic relationships with many other feature types (e.g., hotels and theaters are parts of cities) will exhibit thematic bands in common with many other feature types, and hence would be less differentiable instances of other types based on signature divergence. This is where the analogy between semantic signatures and spectral signatures comes into play. For some of the studied place types, the thematic band does not offer the discriminatory power to differentiate them effectively from other place types. In such cases additional thematic bands have to be mined in addition to temporal and spatial bands discussed in previous work.

The visual evidence of the divergence curves provides a quick means to evaluate the quality and consistency of categories, and they can be further evaluated quantitatively. In Table 3 the place types are listed in descending order based on the Earth Mover's Distance (EMD) between the two curves (same and different type instances). The EMD was popularized for comparing the similarity of grayscale images, but it can be used as a metric for comparing any probability distribution (Rubner *et al.* 2000). The magnitude of the KL divergence values (and thus the EMD measures) are dependent upon the number of topics used during LDA training (in this case 300). However, the important point is their relative values across types, which will not change greatly for different numbers of topics. The EMD results indicate a promising approach for automatically identifying types that are strongly defined semantically with high EMD (such as *lunar crater* and *launch pad*) and those that should be broken up or which are ambiguous with other types, indicated by low EMD (such as *city*, *town*, *village*, and *administrative region*). In addition to EMD, Table 3 also shows the measure of kurtosis and skewness for each curve. The vast majority of the curves are leptokurtic. Most of the curves for instances of the same type are right skewed, whereas most of the curves for instances of different types are left skewed. There tends to be an inverse relationship between skewness of the different-type divergence curve and the EMD. This is not surprising if low EMD indicates a semantically ambiguous type, as there will be a long tail of differently typed instances that will be similar to the prototype.

If the curve has two or more distinct peaks in the density distribution, then it points

[1]Though, there may be multiple, disconnected prototypes (Rosch and Mervis 1975).

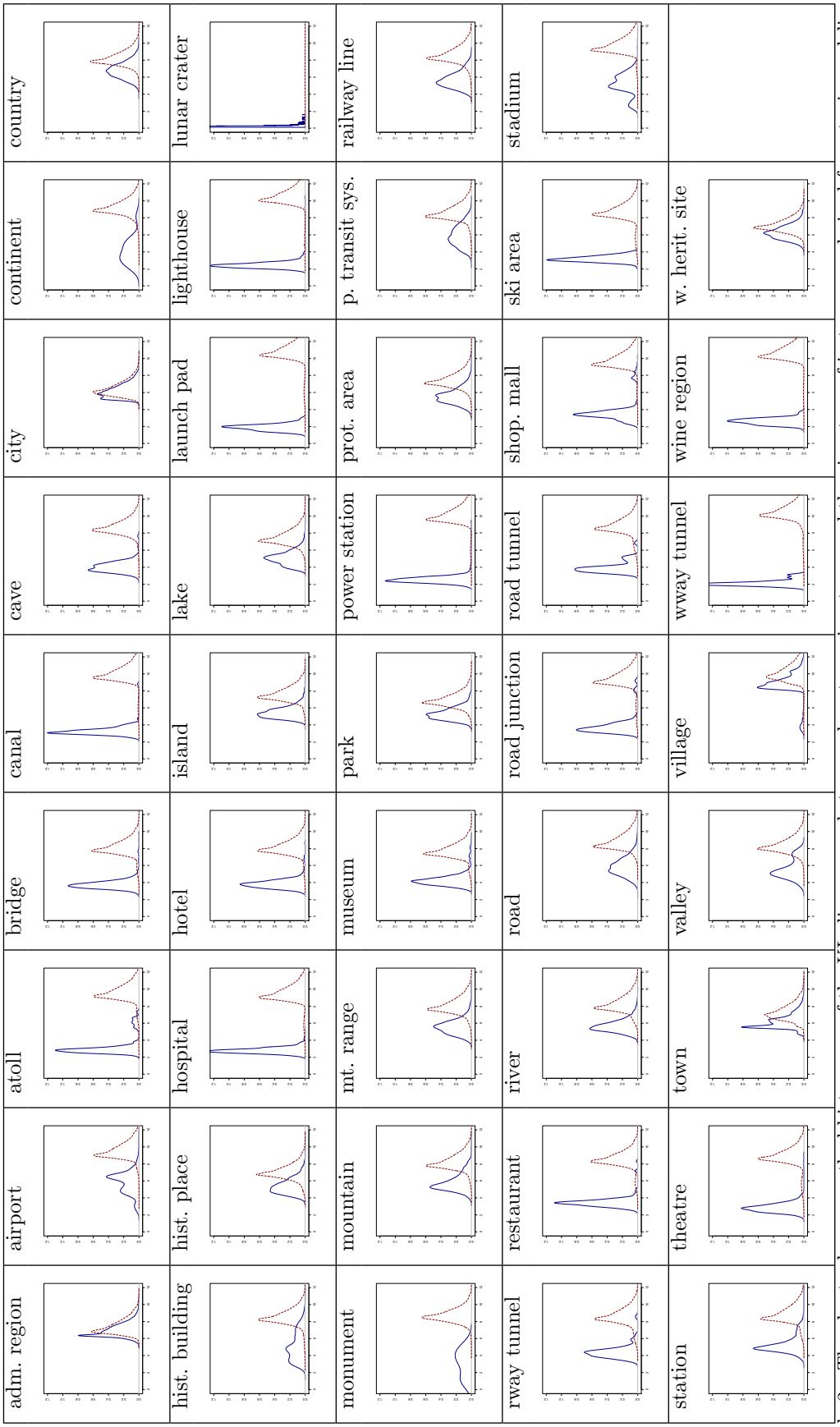


Table 2. The charts show smoothed histograms of the KL divergences between place type signatures and the signature of instances as derived from topic modeling on associated articles (x-axis: KL divergence, y-axis: density [area under the curve is 1.0]). The solid blue curve shows divergences of instances classified as the type, and the dashed red curve shows divergences of other place instances. Types where thematic signatures are predictive of the type are those charts where the blue curve lies mostly to the left of the red curve.

Type	**EMD**	**Kurtosis (s)**	**Kurtosis (d)**	**Skew (s)**	**Skew (d)**
Lunar crater	15.27	77.62	17.15	7.42	-2.02
Launch pad	9.15	0.02	6.27	0.36	-1.64
Wine region	8.36	0.07	8.28	0.32	-1.47
Waterway tunnel	8.31	0.59	5.02	1.14	-1.03
Lighthouse	8.24	91.67	8.37	7.33	-1.19
Power station	7.67	68.70	6.41	5.71	-1.45
Hospital	6.86	3.52	4.55	1.07	-1.16
Canal	6.80	47.34	5.93	5.30	-1.26
Atoll	6.44	2.79	3.39	2.02	-0.62
Shopping mall	6.28	5.78	4.21	2.48	-0.64
Theatre	6.02	27.69	3.25	3.54	-1.28
Monument	5.60	-2.33	1.08	-0.15	0.42
Ski area	5.59	-0.13	3.08	0.31	-1.12
Road junction	5.49	9.25	3.88	3.04	-1.24
Continent	5.21	0.29	1.60	0.91	-0.01
Restaurant	5.12	23.49	2.50	3.96	-0.74
Cave	4.91	6.63	1.70	1.66	0.15
Road tunnel	4.88	3.52	1.57	1.68	-0.52
Stadium	4.52	0.45	2.09	-0.22	0.02
Bridge	4.41	19.54	1.20	3.26	-0.29
Hotel	4.38	12.65	2.38	2.55	-0.34
Railway tunnel	4.26	8.87	1.20	2.40	-0.34
Airport	3.78	0.46	1.12	-0.19	0.46
Station	3.48	1.66	1.17	1.19	0.14
Museum	3.46	7.91	1.54	2.41	-0.23
Historic building	3.38	-0.59	2.76	0.48	-0.37
Railway line	3.27	0.14	1.18	0.59	0.21
Valley	3.20	-0.61	0.94	0.59	0.47
Road	3.02	0.80	1.04	0.60	0.75
Public transit system	2.85	-0.51	1.29	0.24	0.28
Lake	2.66	1.13	0.75	0.67	0.51
River	2.57	3.06	0.75	1.15	0.37
Mountain range	2.56	1.57	0.94	0.73	0.31
Mountain	2.47	0.70	0.94	0.86	0.38
Island	2.26	2.79	0.61	1.16	0.42
Protected area	2.01	1.01	0.82	0.65	0.53
Park	1.88	1.01	0.82	0.81	0.48
Historic place	1.70	1.34	1.05	0.87	0.26
Country	1.63	0.04	0.65	0.27	0.58
Village	1.42	4.31	4.48	-1.81	-0.34
Town	1.18	1.14	0.52	0.96	0.66
World heritage site	0.97	0.37	0.56	0.50	0.62
Administrative region	0.55	0.88	0.48	0.87	0.87
City	0.41	0.08	0.65	0.71	0.88

Table 3. Earth mover's distance (EMD), kurtosis, and skewness measures for the divergence curves of each type. (s) stands for the divergence curve of instances of the same type, and (d) stands for instances of different types. EMD can be seen as an index of semantic homogeneity.

toward the possibility of two or more disjoint subclasses of instances for the type. This effect is seen most clearly in the *village* type where the divergences peak at two clearly discrete values. It indicates that there are two thematic classes of *villages* that diverge in different respects from the *village* type signature. By examining the articles that exhibit these differing divergence values, an ontology engineer can then update the ontology as needed, e.g., by introducing subclasses. The distinction between *city* and *town* is another interesting case as their class definitions are highly local (Janowicz 2012). For instance, by law *city* and *town* are defined to be equivalent in the state of California,

while (for historic reasons) there is just one town in Pennsylvania. In contrast, Utah has a population count based schema to differentiate types of municipalities. As Wikipedia and DBpedia have no mechanism to handle these local differences, both classes cannot be efficiently distinguished. To give a concrete example, Stuttgart, which is the capital of the state of Baden-Württemberg in Germany, is categorized as *town* in DBpedia. This is most likely because it falls into the *Cities in Baden-Württemberg* and the *University towns in Germany* Wikipedia categories.

The thematic signatures based divergence curves do not only provide a visual way to uncover such cases but also provides guidelines how to address them. For instance, in the village case, new categories may be introduced, while the city/town case demonstrates a scenario in which existing ontological distinctions are not well supported by the data.

3.2.2. Examining the Long Tail to Find Anomalies

Place instances that fall in the long tail are particularly interesting because they indicate highly divergent cases. This information can be used to either identify potentially misclassified instances or instances that are underspecified if only identified by one place type. For big and noisy datasets such as the DBpedia, methods to find such outliers are essential. Table 4 shows the top outlying instances for selected place types discovered using the propose signatures. All these examples shown place information that could be greatly enriched or repaired with respect to their categorization in the DBpedia ontology. One clear example is that resources like *India* are misclassified as *city* because they are related to other instances through the `is dbpedia-owl:city of` relation. Another is the example of *Catholic High School, Singapore*, which is not categorized as a *country* in DBpedia, but inferred to a *country* type through an sameAs relation in a linked data repository. This lack of consistent discovery from different query endpoints has been identified as a key research problem in Linked Data (Buil-Aranda *et al.* 2013).

Obviously, these long tail anomalies also contribute to the text used to derive thematic signatures for the types, but this technique relies on the assumption that the vast majority of instances will be correctly classified and thus these outliers will have a small effect on the overall type signature.

3.2.3. Evaluation

In order to evaluate whether highly divergent instances indicative a need for data cleansing we took the ten-most divergent and ten-least divergent instances for each of the 44 leaf subclasses of the Place class and manually evaluated the classification. In all cases for the least divergent instances no misclassified instances were identified based on human interpretation of their semantics. However, for the most divergent instances several errors in the classification were identified. The number of errors varied greatly by feature type. Some types, such as *launch pad* and *lunar crater* did not have any errors, indicating clarity in the ontology engineering process for determining class membership. Some classes only have a few instances in total, making human validation easier and requiring less reliance on automatic methods. On the other hand, for most feature types several errors were observed in the most-divergent cases. All ten instances were identified as potential errors for the feature types *airport*, *bridge*, and *country*. Nine out of ten were errors for the type *city*, with the only 'correct' assignment being the Greek historical city-state of Sparta. Eight out of ten were identified as errors for the types *mountain* and *mountain range*. These results can directly be used to clean Linked Data as well as point to possible new types such as *city-state*.

The results point to systematic errors in the assignment of feature types in existing Linked Data. For example, seven of the ten most divergent places for the *airport* type are

Place type	N	Most divergent http://dbpedia.org/resource/*	KL div.
Administrative region	24027	Greater Manchester	11.08
		West Berlin	10.99
Airport	4296	Mukachevo	11.63
		Lutsk	11.40
Atoll	182	Scarborough Shoal	7.38
		Serranilla Bank	7.22
Bridge	1639	Reading and Leeds Festivals	9.75
		Valens Aqueduct	8.99
Canal	175	Bergse Maas	8.81
		Canal Bank Road	4.73
Cave	134	Kharosa	7.66
		Liang Bua	6.40
City	57276	Chiapas	10.06
		India	9.78
Continent	14	Australia (continent)	8.14
		Europa	5.67
Country	2347	Chesterfield Islands	10.17
		Spanish Florida	10.08
Historic building	2380	Dublin	11.10
		White Horse Temple	10.20
Historic place	1801	Transylvania	10.64
		Gaza Strip	10.50
Hospital	1165	The Madras Medical Mission	5.32
		Ignace Deen Hospital	4.44
Hotel	458	Bodysgallen Hall	8.63
		Miskin Manor	7.12
Island	2900	Java_(programming_language)	10.36
		List_of_districts_and_neighborhoods_of_Los Angeles	9.83
Lake	3025	Adriatic Sea	9.95
		Loch Alsh	8.89
Launch pad	42	Broglio Space Centre	2.95
		Cape Canaveral Air Force Station Launch Complex 34	2.57
Lighthouse	765	Kirby Cove Camp	10.42
		Flåvær	9.22
Lunar crater	1464	Shackleton (crater)	1.59
		Cabeus (crater)	1.28
Monument	3	Sigismund Bell	4.75
		Illinois Freedom Bell	3.38
Mountain	3722	Vatican City	11.26
		West Hills, Los Angeles	10.30
Mountain range	959	Appalachia	10.38
		Tourism in India	9.63
Museum	2020	Habitation at Port-Royal	9.54
		Hollywood Walk of Fame	9.15
Park	1035	Banias	8.85
		Tel Hazor	8.12
Power station	753	Okutataragi Pumped Storage Power Station	9.31
		LaColle Falls Hydroelectric Dam	8.19
Protected area	3503	Pennsylvania Avenue National Historic Site	10.49
		Port Chicago disaster	10.41
Public transit system	1722	West Norfolk Junction Railway	9.81
		Kolkata Suburban Railway	9.52
Railway line	1269	Orange Line (Jaipur Metro)	8.94
		Huddersfield Line	8.77
Railway tunnel	62	Market Street Subway	8.57
		Castle Shannon Incline	6.41
Restaurant	199	Leo's_Tavern	8.33
		Sheen_Falls_Lodge	7.23
River	3997	Copper	11.13
		Leaf	10.76
Road	9747	Wall_Street	12.41
		Potsdamer_Platz	11.86
Road junction	54	Oxford Circus	9.18
		Edappally Junction	8.24
Road tunnel	51	Gaviota Tunnel	6.85
		Northbrae Tunnel	5.31
Shopping mall	1451	Thrissur KSRTC Bus Station	10.19
		Shaktan Thampuran Private Bus Stand, Thrissur	10.09
Ski area	311	Falls Creek, Victoria	4.14
		Mount Ashland Ski Area	4.08
Stadium	2892	Wales Millennium Centre	11.21
		Melbourne Convention and Exhibition Centre	10.58
Station	7769	Charlotte, North Carolina	11.93
		Maplewood, New Jersey	11.52
Theatre	586	Delamere Park	9.02
		Center for Fine Art Photography	8.38
Town	19216	Chiapas	11.08
		Vermont	10.60
Valley	70	Lockyer Valley	8.05
		Fergana Valley	8.02
Village	5386	Missouri	13.96
		Avilla, Missouri	12.59
Waterway tunnel	16	Pennsylvania Canal Tunnel	3.11
		Greywell Tunnel	2.79
Wine region	140	Shawnee Hills AVA	3.78
		Connecticut wine	3.69
World heritage site	1197	Everglades	9.67
		Glacier National Park (U.S.)	9.05

Table 4. This table shows the most outlying instances of select place types based on KL divergence of thematic band of instances to their respective classes. Instance classifications can either be specified in DBpedia or inferred through sameAs and equivalentclass relations to classes in other Wikipedia data repositories, such as YAGO. The results show the potential of the proposed thematic signatures for knowledge base enrichment and cleaning.

cities in Ukraine–perhaps indicating a problem in the algorithms used to automatically lift semantics from the Ukranian language version of Wikipedia. Likewise, five of the ten most divergent instances of *island* type are in fact *peninsulas*. In other cases, the results point to judgments of semantic interpretation by the ontology engineers that might be questionable depending on the context. For example, three of the ten most divergent places for the *bridge* type are aqueducts built by the Romans. Although aqueducts share some structural similarity to bridges, in most cases they will not be satisfactory results for Linked Data queries for bridges. Likewise, many articles tagged as *mountains* are relatively low-lying hills in England such as Box Hill, Surrey, which is only 224 meters high. It is well known that the definitions of geographic feature types are highly location-context dependent. Finally, other feature types conflate semantics in a way that limits their usability. For example, Great Eastern Railway, a former British railway *company* is conflated with the feature type *public transit system*, which is a sub-class of *Place*.

3.2.4. Characterizing Multiple Feature Types Instantiated by an Article

To illustrate the usage of the thematic semantic signatures to further understand the semantics of place instances, Figure 4 shows how they can be applied to characterize the *multiple* feature types associated with a single Wikipedia article. In the DBpedia dataset, Santa Barbara, CA[1] is classified as a *City* and Cape Norman[2] in Newfoundland is not classified as any type. Using the LDA topic mixtures for the Wikipedia articles of these two places, we can automatically calculate the similarity between those mixtures and the thematic bands for different feature types. Figure 4 shows that the text in the Cape Norman article has a high similarity to the *lighthouse* type, and, in fact, it has a prominent lighthouse as described in the text. The Santa Barbara article, which is much longer, is most similar to the *city* type but it also has a similarity to some other feature types, including *historic place*, *park*, *museum*, and *hotel*. These results can be used to enrich the semantic representation of places, e.g., to enable better thematic search in semantically-enabled gazetteers. As these examples demonstrate, the signatures can also be used to enrich DBpedia by assigning types to places that have not yet been categorized.

4. Additional Use Cases for Thematic Bands and Signatures

In this section we outline additional results of applying the LDA-derived thematic signatures. To demonstrate their strengths as well as their wide applicability, we use them for the following tasks and give concrete examples: **1)** discovering bottom-up representations of feature types from unstructured data; **2)** inferring new class subsumption hierarchies from the signatures; and **3)** mapping locations associated with specific feature type-specific topics.

4.1. *Signature-based Similarity of Feature Types*

Figure 5 shows a two dimensional visualization of the similarities of 44 selected thematic bands for feature types on different scales ranging from natural, meso-scale landscape features, e.g., mountains, to man-made, micro-scale features such as restaurants. The metric Multi Dimensional Scaling (MDS) shows the semantic similarity between feature

[1] `en.wikipedia.org/wiki/Santa_Barbara,_California`
[2] `en.wikipedia.org/wiki/Cape_Norman`

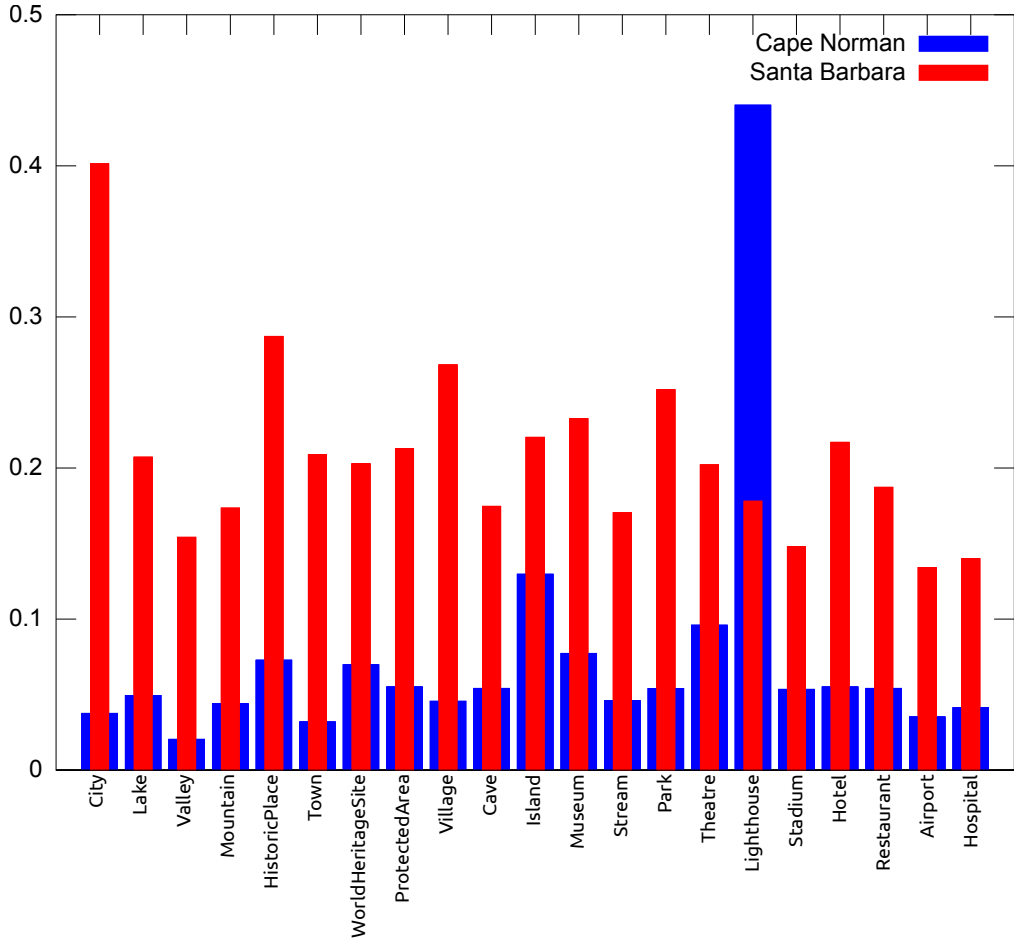


Figure 4. Feature type characterization for the Santa Barbara, CA and Cape Norman Wikipedia articles. The y-axis shows 1 / JS divergence between the topic mixture and thematic band for selected feature types. Higher values indicate more similarity between the type signature and the thematic bands for the sample articles.

types based on their thematic bands. For instance, *city* and *town* are closer than *city* to *mountain*. Such inter-type similarities are not only relevant for location estimation but also for semantics-based geographic information retrieval. Through the application of weights on specific topics these similarity measures can also be personalized to match a user's interests (Adams and Raubal 2014).

4.2. *Inferring new Subsumption Hierarchies*

The semantic signatures for place types provide an alternative approach to organize the knowledge in Wikipedia based on the natural language content of the articles. Given a set of (largely) mutually exclusive classes in the DBpedia ontology (e.g., all leaf subclasses of Place), one can use the similarities of the semantic signatures of the classes to automatically infer class relationships in a hierarchy using hierarchical clustering (Hastie *et al.* 2009). This bottom-up approach organizes the classes into a tree hierarchy based on the semantic content of the articles rather than imposing ontological distinctions top-down. Since these signatures come from what people are writing about place instances, it can also uncover relevant properties for categorization that an ontology designer may

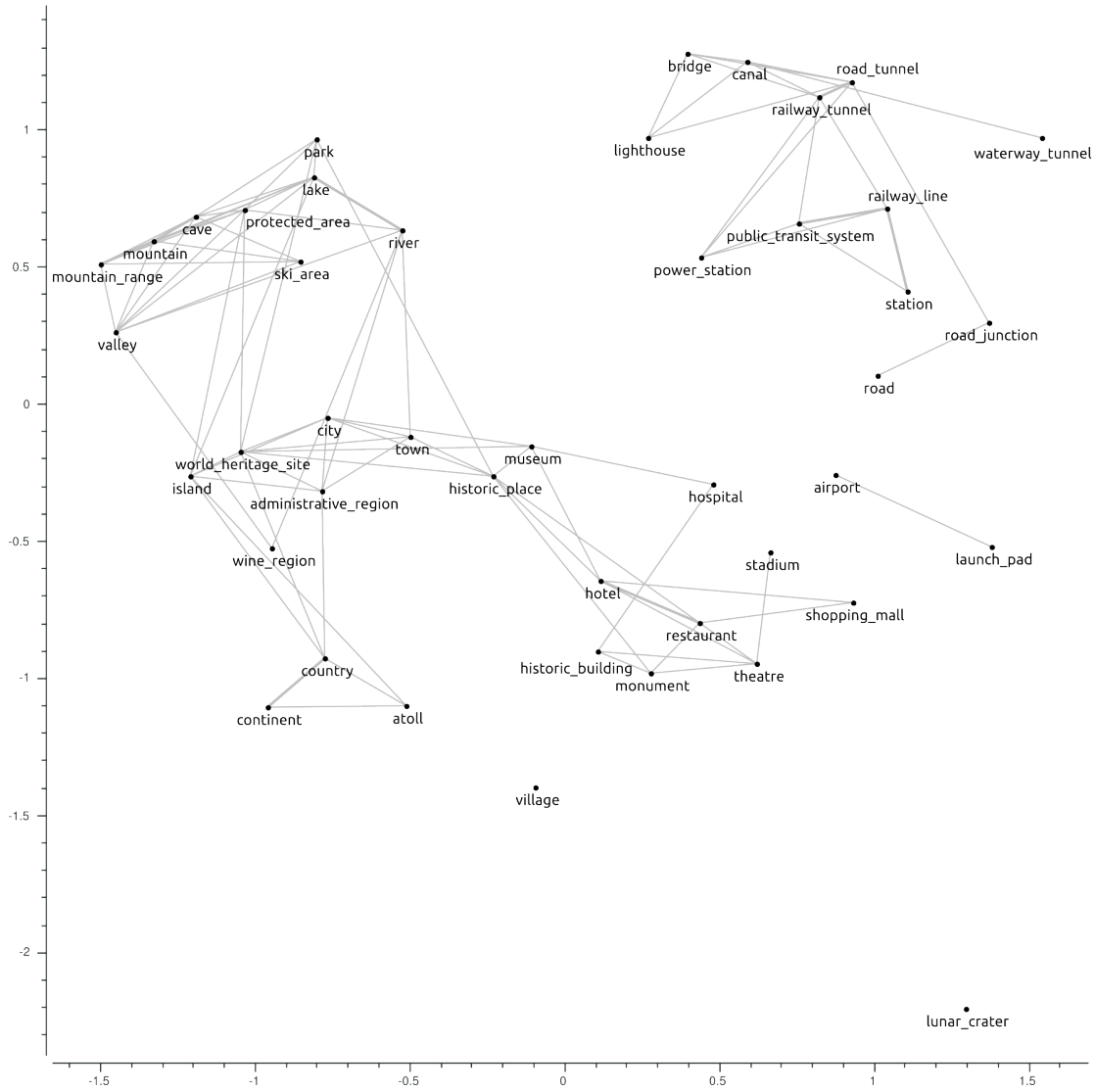


Figure 5. Metric MDS of 44 selected feature types based on their thematic bands; Kruskal's stress = 0.179 (Kruskal and Wish 1978). The similarity measure used in this case is the JS divergence as shown in Equation 2. The segments between the points indicate the most similar pairs.

not consider and which might be very different than the hierarchy currently used by the DBpedia ontology.

Figure 6 shows a sample result of hierarchical clustering of all the *place* leaf subclasses in the DBpedia ontology for which there are instances. The distance metric used in this example is the JS divergence between the semantic signatures for the two types. We use only leaf classes from the DBpedia ontology, because for classes higher up in the hierarchy there will be overlap in the articles used to generate the semantic signatures. This overlap would be reflected in their similarities and the result would be an artificial pressure to match DBpedia's hierarchy. As part of an interactive process (combining bottom-up and top-down approaches), an ontology designer can label the new superclasses based on their background knowledge. Alternately, the bands that are most shared between sibling classes (and particularly in this case the most probable words for the LDA topics) can provide useful suggestions for class labeling. In this case, methods to automatically label

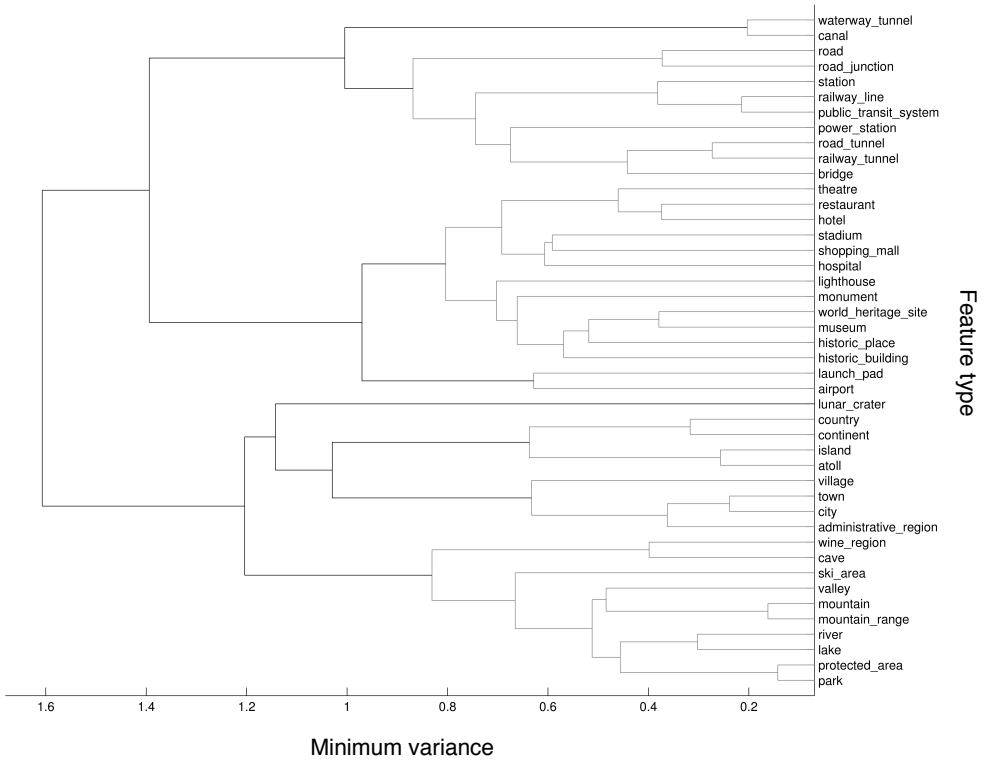
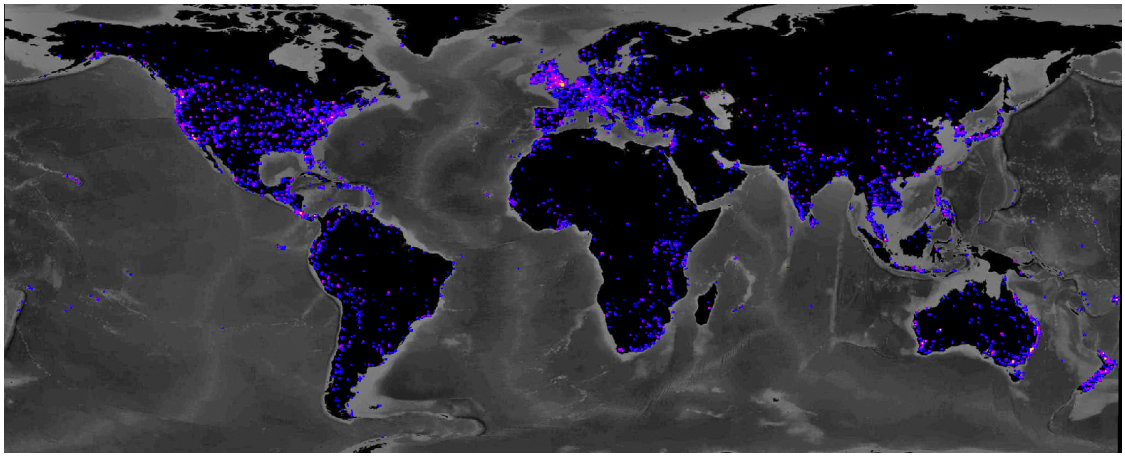


Figure 6. Hierarchical clustering result on a set of 44 leaf subclasses of Place in the DBpedia ontology using Ward's minimum variance method on the pairwise JS divergences (Ward 1963). The semantic signatures in this example were derived from a 50-topic model.

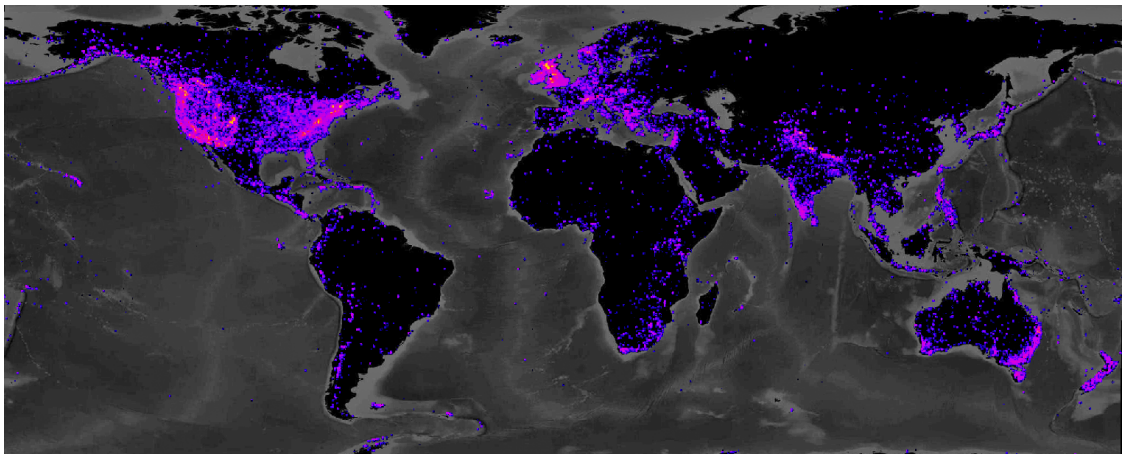
topics can be applied (Lau *et al.* 2011).

Some immediate differences can be seen between the organization of classes following our method and the DBpedia ontology. In DBpedia there is a rigid division of classes between *architectural structures*, *natural places*, *populated places*, and miscellaneous classes that are not subsumed under these three. This distinction between natural and artificial geographic features reflects an ontological perspective that has been widely adopted in geography (Mark *et al.* 2001). However, a more nuanced view that acknowledges that natural and man-made systems are interlinked has been a foundational perspective in socio-ecological research (Berkes *et al.* 2003). For example, the courses of many rivers have been engineered by humans to facilitate transportation networks that afford specific types of human activity, and thus are not purely natural features in any recognizable way. In contrast to the DBpedia classification, using a semantic signature approach, classes are organized by themes that are most prevalently associated with the feature types in the text. For example, based on semantic signatures *waterway tunnel* and *canal* are very closely related whereas in DBpedia they fall in completely different parts of the taxonomy. Likewise, our method uncovers similarities based on the historical aspects of *museum*, *world heritage site*, *historic place*, and *historic building* instances. In addition, *cities*, *towns*, and *administrative regions* are grouped together, separate from a group with *country*, *continent*, *atoll*, and *island*. This effect may reflect differences in how geographic features are conceptualized at differing scales (Montello 1993).

As the presented signature-based approach is derived from the content of Wikipedia,



(a) City band from Wikipedia



(b) Mountain band from Wikipedia

Figure 7. The thematic signature for the types *city* and *mountain* mapped on the Earth's surface.

one can conclude that parts of the DBpedia ontology introduce distinctions that are not reflected /supported by the data.

4.3. *Mapping Thematic Signatures*

In Adams and Janowicz (2012) a method was developed for mapping a mixture of LDA derived topics on the globe. A grid is created over the Earth for each LDA topic and a value for that topic is assigned to each grid cell using the following steps: **1)** Each document is binned into the appropriate grid cell based on the geo-reference location associated with the article. **2)** The average probability of the topic for all intersecting documents (based on their geo-reference) is calculated and the value is assigned to the cell. **3)** To smooth the result and generate regions associated with topics, a spatial kernel density operation is run on the topic raster to generate a surface over the Earth for the topic. The default bandwidth of the kernels is set to twice the grid square width allowing for region generation when high topic values are found in adjacent grids. Once a surface for each topic has been created, the mixture of topics (i.e., the LDA distribution) for a place type is rendered by performing a map algebra weighted sum of all the topic surfaces

associated with the type, where the weights are the probabilities of the type topics.

Applying this method, Figure 7 shows the spatial distribution of the thematic signatures for the *city* and *mountain* types. For instance, the Rocky Mountains, Appalachian Mountains, Alps, Himalayas, the Great Dividing Range, Eastern and Western Ghats, and the Swartberg range can be clearly seen in figure 7b. This is a remarkable result given that, in this example, all geographic names have been pruned prior to training the topics. The Andes are visible but less present, while the Pennines in England are over-represented. This result should not be misinterpreted as the Andes being less *mountainish*. It is caused by a higher number of articles related to the Pennines and their more detailed textual description per unit of space. This reflects the biases inherent in different language versions of Wikipedia, in this case an Anglo-American bias in the English Wikipedia, which has been well-documented (Graham *et al.* 2014). People tend to write more frequently about certain places in the Pennies than about those in the Andes; hence, the first mentioned are more likely to be referred in future texts. This effect is not critical for the bands as they are used in combination with the topics-based region estimations. Consequently, if a text can be geo-located to regions in South America and has a high probability of describing mountains, those regions can be further reduced to the Andes (given that their geographic footprint overlaps with those from the regions).

While figure 7a and 7b differ significantly in terms of regions and especially their weights, e.g., due to the low city density in the Himalayas, their overall coverage is similar. This is caused by the small number of Wikipedia articles covering features in Central Africa, North Asia, and rural parts of Brazil. We compared this distribution with a study on geo-locating Flickr images performed by Hayes and Efros and found them to be almost identical; see (Hays and Efros 2008, fig. 2).

5. Conclusion and Future Work

In this paper we proposed methods that exploit unstructured data about places and their types to discover bottom-up semantics to complement existing work on place semantics on the Semantic Web and the Linked Data cloud. As such, this work is an example of observation-driven ontology engineering from crowdsourced volunteered geographic information. We demonstrated that by using text mining and clustering on place articles we can derive thematic signatures that characterize places based on the content of these articles, not merely the (semi-)structured data that are associated with them. For our results we focused on Wikipedia and the related DBpedia ontology as source material, but the methods are widely applicable to other kinds of unstructured place descriptions, e.g., travelogues, newspaper articles, historical documents, and georeferenced social media.

To demonstrate the suitability of semantic signatures as well as their added value when combined with classical top-down ontology engineering, we presented multiple application areas by giving concrete examples. The signatures enable us to visually inspect alternative conceptual hierarchies using MDS and hierarchical clustering. We demonstrated how they provide insight into the relationships and different qualities of geographic feature types, such as how broadly or narrowly they are described as groups in Wikipedia. Additionally, we pointed out how our proposed approach can guide ontology engineers, e.g., by uncovering which top-down classes are not reflected in the data. Finally, signatures provide valuable information on outlier instances of places returned from SPARQL queries, which point to categorization errors in DBpedia and derived sources (e.g., FactForge). Thus the signatures can contribute to cleaning and enriching Linked Data about places.

While these methods show promise, we focused on one approach to deriving thematic semantic signatures, namely an using LDA. Future work will include comparing this approach against other methods of building semantic signatures of places from text, such as neural networks (Hinton and Salakhutdinov 2006). Developing more techniques to integrate these semantic signature representations with more traditional top-down knowledge representations, such as Semantic Web ontologies is also important. Furthermore, the signature approach points to the possibility of creating a universe of representations via different methods, which leads to a need to better evaluate different models against each other for specific applications (Gahegan and Adams 2014).

References

Adams, B. and Janowicz, K., 2012. On the Geo-Indicativeness of Non-Georeferenced Text.. *In*: J.G. Breslin, N.B. Ellison, J.G. Shanahan and Z. Tufekci, eds. *ICWSM* The AAAI Press, 375–378.

Adams, B. and McKenzie, G., 2013. Inferring thematic places from spatially referenced natural language descriptions. *In*: D. Sui, S. Elwood and M. Goodchild, eds. *Crowdsourcing Geographic Knowledge*. Springer, 201–221.

Adams, B. and Raubal, M., 2014. Identifying salient topics for personalized place similarity. *In*: S. Winter and C. Rizos, eds. *Research@Locate'14, Canberra, Australia, 07-09 April 2014*. CEUR-WS, 1–12.

Auer, S., Lehmann, J., and Hellmann, S., 2009. LinkedGeoData: Adding a spatial dimension to the web of data. *In*: A. Bernstein, D.R. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta and K. Thirunarayan, eds. *The Semantic Web-ISWC 2009.*, Vol. 5823 of *Lecture Notes in Computer Science* Springer, 731–746.

Berkes, F., Colding, J., and Folke, C., 2003. *Navigating social-ecological systems: building resilience for complexity and change*. Cambridge University Press.

Bizer, C., *et al.*, 2009. DBpedia - A Crystallization Point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7 (3), 154–165.

Blei, D.M., 2014. Build, compute, critique, repeat: data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1, 203–232.

Blei, D.M., Ng, A.Y., and Jordan, M.I., 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Bowman, A. and Azzalini, A., 1997. *Applied smoothing techniques for data analysis*. No. 18 Oxford statistical science series Oxford: Clarendon Press.

Buil-Aranda, C., *et al.*, 2013. SPARQL Web-Querying Infrastructure: Ready for Action?. *In*: H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J.X. Parreira, L. Aroyo, N. Noy, C. Welty and K. Janowicz, eds. *International Semantic Web Conference–ISWC 2013*, Vol. 8219 of *Lecture Notes in Computer Science* Springer, 277–293.

Cresswell, T., 2004. *Place: a short introduction*. Blackwell Publishing Ltd.

Edwardes, A.J. and Purves, R.S., 2007. A theoretical grounding for semantic descriptions of place. *In*: J.M. Ware and G.E. Taylor, eds. *Web and Wireless Geographical Information Systems.*, Vol. 4857 of *Lecture Notes in Computer Science* Springer, 106–120.

Entrikin, N., 1991. *The Betweenness of Place: Toward a Geography of Modernity*. Johns Hopkins University Press.

Gahegan, M. and Adams, B., 2014. Re-Envisioning Data Description Using Peirce's Pragmatics. *In*: M. Duckham, E. Pebesma, K. Stewart and A. Frank, eds. *Geographic*

Information Science., Vol. 8728 of *Lecture Notes in Computer Science* Springer International Publishing, 142–158.

Gangemi, A., 2013. A Comparison of Knowledge Extraction Tools for the Semantic Web. *In*: P. Cimiano, Ó. Corcho, V. Presutti, L. Hollink and S. Rudolph, eds. *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*, Vol. 7882 of *Lecture Notes in Computer Science* Springer, 351–366.

Graham, M., *et al.*, 2014. Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty. *Annals of the Association of American Geographers*, 104 (4), 746–764.

Griffiths, T.L. and Steyvers, M., 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 (Suppl. 1), 5228–5235.

Hastie, T., Tibshirani, R., and Friedman, J., 2009. *The Elements of Statistical Learning*. New York, NY: Springer.

Hays, J. and Efros, A.A., 2008. im2gps: estimating geographic information from a single image. *In*: *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* IEEE Computer Society.

Hinton, G.E. and Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. *Science*, 313 (5786), 504–507.

Janowicz, K., 2012. Observation-Driven Geo-Ontology Engineering.. *T. GIS*, 16 (3), 351–374.

Jones, C.B., *et al.*, 2008. Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, 22 (10), 1045–1065.

Kruskal, J. and Wish, M., 1978. *Multidimensional Scaling*. Sage Publications.

Lau, J.H., *et al.*, 2011. Automatic Labelling of Topic Models. *In*: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, Portland, Oregon Association for Computational Linguistics, 1536–1545.

Lehmann, J., *et al.*, 2014. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*.

Mark, D.M., Skupin, A., and Smith, B., 2001. Features, Objects, and Other Things: Ontological Distinctions in the Geographic Domain.. *In*: D.R. Montello, ed. *COSIT*, Vol. 2205 of *Lecture Notes in Computer Science* Springer, 489–502.

Massey, D., 1991. A global sense of place. *Marxism today*, 35 (6), 24–29.

Medelyan, O., *et al.*, 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67 (9), 716–754.

Montello, D.R., 1993. Scale and Multiple Psychologies of Space. *In*: A.U. Frank and I. Campari, eds. *Spatial Information Theory: A Theoretical Basis for GIS*, Vol. 716 of *Lecture Notes in Computer Science* Springer, 312–321.

Montello, D.R., *et al.*, 2003. Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, 3 (2-3), 185–204.

Mülligann, C., *et al.*, 2011. Analyzing Spatial-Semantic Interaction of Points of Interest in Volunteered Geographic Information. *In*: *COSIT 2011*, Vol. 6899 of *LNCS* Springer, 350–370.

Rosch, E. and Mervis, C.B., 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7 (4), 573–605.

Rubner, Y., Tomasi, C., and Guibas, L.J., 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40 (2), 99–121.

Steyvers, M. and Griffiths, T., 2007. Probabilistic Topic Models. *In*: T. Landauer, D. Mcnamara, S. Dennis and W. Kintsch, eds. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, 424–440.

Suchanek, F.M., Kasneci, G., and Weikum, G., 2007. YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. *In*: *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, Banff, Alberta, Canada New York, NY, USA: ACM, 697–706.

Teh, Y.W., *et al.*, 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101 (476), 1566–1581.

Vasardani, M., *et al.*, 2013. From descriptions to depictions: A conceptual framework. *In*: T. Tenbrink, J. Stell, A. Galton and Z. Wood, eds. *Spatial Information Theory.*, Vol. 8116 of *Lecture Notes in Computer Science* Springer, 299–319.

Wallach, H.M., *et al.*, 2009. Evaluation methods for topic models.. *In*: A.P. Danyluk, L. Bottou and M.L. Littman, eds. *ICML*, Vol. 382 of *ACM International Conference Proceeding Series* ACM, p. 139.

Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58 (301), 236–244.

Winter, S., Kuhn, W., and Krüger, A., 2009. Guest Editorial: Does Place Have a Place in Geographic Information Science?. *Spatial Cognition and Computation*, 9, 171–173.

Ye, M., *et al.*, 2011. What you are is When you are: The Temporal Dimension of Feature Types in Location-based Social Networks.. *In*: *ACM SIGSPATIAL 2011* ACM, 102–111.