TagNet: Using Soft Semantics to Search the Web

Geert Vanderhulst and Lieven Trappeniers
Alcatel-Lucent Bell Labs,
Copernicuslaan 50, 2018 Antwerpen, Belgium

Keywords: Tagging, Ambiguity, Soft Semantics.

Abstract: Semantic annotations are key to efficiently retrieve resources on the Web. On the one hand, ontologies underlying Web resources give rise to linked data. On the other hand, tagging has become increasingly popular to bring order in data across Web applications and social networks. While taxonomies and folksonomies serve the same purpose (i.e. classification), there is a large gap in semantics between uncontrolled keywords used for tagging and hierarchical concepts found in a taxonomy. In this paper we introduce sematags as light-weight 'soft semantics' which aim to bridge the gap between 'no semantics' and 'hard semantics'. Sematags define aliases (synonyms) and isas (hypernyms) which overcome the typical issues conventional tags cope with such as ambiguity. Furthermore, we present TagNet, a framework that extracts sematags from lexicons and existing knowledge bases and exploits them to annotate, link and retrieve resources on the Web. We evaluate how soft semantics can be used to semi-automatically map tagged photos in Flickr on DBpedia concepts and vice versa.

1 INTRODUCTION

The basic premise of the Semantic Web is to represent knowledge in a meaningful way so that computers can function more effectively by being able to distinguish different meanings of data. This is achieved by describing data using languages with a logical entailment such as OWL and RDF. We refer to this approach using the term *hard semantics* since linked data is typically mapped on ontological resources by domain experts or agents. Besides, we witness an emerge of folksonomies to create order in a rapidly expanding Web of data (Vander Wal, 2007; Specia and Motta, 2007). The increasing popularity of tags as a flat space of keywords is both visible on websites such as Youtube, Flickr, Del.icio.us and across social networks (e.g. hashtags on Twitter). However, there are still a number of limitations of the current state of technology as identified in (Garcia-Castro et al., 2009): (i) tag ambiguity, (ii) missing links between multiple synonyms, spelling variants, or morphological variants, and (iii) variation in the level of granularity and specificity of the tags used caused by differences in the domain expertise of agents. These issues are due to the fact that tags typically have *no semantics* associated.

In this paper we present TagNet, a framework that eases the task of annotating and searching for resources on the Web using *soft semantics*. Rather than defining hard links to ontological concepts, we add additional detail to a tag to remove ambiguity and facilitate automatic derivation of links to existing knowledge bases. In TagNet, tags (i.e. plain text keywords) are annotated in two dimensions: each tag (i.e. *sematag*) defines *aliases* and *isas* as illustrated in figure 1. Aliases are keywords that can be used as a synonym for a given tag (e.g. synonyms, acronyms, etc) and isas are keywords that generalize the meaning of the tag (e.g. sport is an isa for tennis). The combination of aliases and isas helps us to understand and express the meaning of a tag using additional tags, similar to the approach taken in (Garcia-Castro et al., 2009). We distinguish between aliases and isas because it matches well with the detail of information contained in dictionaries (e.g. WordNet (Fellbaum, 1998)) and ontologies (e.g. the DBpedia (Auer et al., 2008) ontology) which are suitable sources to extract tags from as outlined in section 2. For instance, the WordNet lexicon expresses linguistic relations between words such as synonyms (aliases) and hypernyms (isas). Also in ontologies, there is a notion of similar concepts (aliases) – expressed in OWL using constructs such as `owl:sameAs` (instance level) and `owl:equivalentClass` (class level) – and 'isa' relations contained in an ontology which directly map on the isas of a tag in TagNet. TagNet advances the state of the art by exploiting soft semantics both in the annotation process of arbitrary resources on the

Vanderhulst G. and Trappeniers L..
TagNet: Using Soft Semantics to Search the Web.
DOI: 10.5220/0004325001790188
In *Proceedings of the 9th International Conference on Web Information Systems and Technologies* (WEBIST-2013), pages 179-188
ISBN: 978-989-8565-54-9

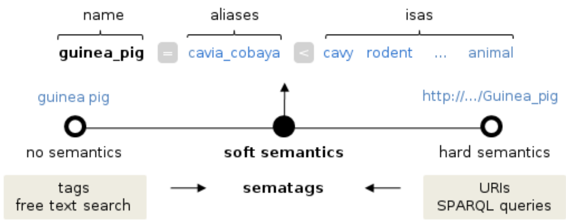

Figure 1: A sematag consists of a name, *aliases* and *isas* which we label as soft semantics. Its purpose is to link tag-based systems (no semantics) with semantic knowledge bases (hard semantics).

Web and during search operations. To solve ambiguity for end-users, we explain the different senses of a keyword using the isas and aliases of tags matching the keyword (section 3) which is also useful to refine search queries (section 4). A unique advantage of sematags over hard URIs is their scalability to tag-based systems combined with the ability to map them on linked data, as illustrated in section 5 by means of a case study.

2 TAGGING VOCABULARIES

To stimulate reuse and sharing of existing sematags and relieve users of specifying aliases and isas manually, TagNet relies on vocabularies from which tags are extracted and suggested to users. Basically, a tag (i.e. entered keyword) is only valid if it can be found in a vocabulary compatible with TagNet. However, since this constraint would prohibit free tagging, we relax it by requiring that the name of the tag can be freely chosen, as long as it is annotated with at least one isa that appears in a controlled vocabulary. Hence, a user can annotate a picture of a pet using the sematag `mickey` (the name of the pet), provided that it is enriched with e.g. a known `dog` isa tag.

We consider WordNet and DBpedia as main, controlled vocabularies for TagNet thanks to their wide coverage of contemporary terms. A lexicon such as WordNet contains a rich set of words that are part of the English language, but it still lacks several concepts such as names of places, people, companies, television shows, etc. In addition to language-specific words, we also want to include commonly accepted terms that are introduced by a community of users. The DBpedia vocabulary extracts tags out of content that was published on Wikipedia. Examples of tag names that are supported by this vocabulary are `google`, `san francisco`, `madonna`, etc. Whilst Wikipedia covers a large set of generally accepted terms, it still excludes concepts that only matter to specific users such as names of family members or pets and highly specialized terms related to a particular domain (e.g. medicines). To share such specialized tags, custom user- or domain-specific vocabularies can be integrated in TagNet as discussed in section 6. Figure 2 illustrates the main task of a vocabulary: taking a keyword as input, a vocabulary outputs one or more tags that attribute a meaning to the keyword using isas and aliases. Note that sematags do not contain direct references to concepts defined in the vocabulary's underlying knowledge base. For example, a tag extracted from the WordNet lexicon does not store a reference to a WordNet synset, nor does a DBpedia tag contain a link to a Wikipedia page. We opted for such a loose coupling because it allows us to describe and interpret all tags equally and independent of the semantics underlying a vocabulary. However, decoupling tags and vocabularies also introduces a new level of ambiguity between similar tags extracted from different vocabularies. For instance, the term `dog` is found both in WordNet and the DBpedia vocabulary with slightly different semantics. To overcome this ambiguity, we add a label to each tag that identifies the vocabulary from where the tag originates. In the next sections we outline how tags are extracted from WordNet and DBpedia.

Figure 2: A vocabulary takes a keyword as input and outputs one or more sematags.

2.1 WordNet Vocabulary

In WordNet, words are organized in synsets: sets of words with a similar meaning (i.e. synonyms). For each synset, hypernyms can be looked up (e.g. animal is a hypernym of dog). Hence, for each word in a synset, a sematag can be composed as follows:

1. the name of the tag is the name of the word;
2. the aliases of the sematag correspond to the names of all other words in the synset;
3. the isas of the sematag relate to the names of all direct hypernyms of the synset including hypernyms of hypernyms.

A keyword that is looked up in WordNet can appear in multiple synsets if it has several meanings and thus gives rise to multiple sematags, one for each sense. To improve the results, we filter out generic hypernyms such as `living thing`, `object`, `entity` and `whole` as they do not contribute to the differentiation of senses. The tag depicted in figure 1 is an example of a sematag produced by the WordNet vocabulary.

The WordNet vocabulary is further subdivided in the following subvocabularies: nouns, verbs, adjectives and adverbs. This allows users and agents to quickly distinguish between senses, knowing the lexical role of a keyword.

2.2 DBpedia Vocabulary

The DBpedia ontology organizes Wikipedia concepts in a structured hierarchy currently covering about 320 classes and 1650 different properties. There is an opportunity to generate sematags out of DBpedia classes (e.g. http://dbpedia.org/ontology/Person), instances (e.g http://dbpedia.org/resource/Semantic_Web) and properties (e.g http://dbpedia.org/ontology/birthDate). In this section, we will only elaborate on classes and instances, yet properties are briefly discussed in section 8. To understand how sematags are extracted from DBpedia, we will first introduce the notion of redirects and disambiguates in the DBpedia ontology:

Redirects. The `wikiPageRedirects` property maps a resource on another resource. An example is the resource `Cow` which does not have its own page on Wikipedia, yet is redirected to the resource `Cattle`. Hence we can interpret `cow` as an alias for `cattle` and vice versa. Several redirects are also defined to support different descriptions of the same resource. For instance, the resource `Winston_Churchill` is also referred to as `Sir_Winston` and `Prime_Minister_Churchill`.

Disambiguates. The `wikiPageDisambiguates` property maps a virtual resource on a collection of relevant resources that could be intended by a particular term. An example is the `Bird_(disambiguation)` resource which links to o.a. `Bird` (animal), `Birds,_Illinois` (community in the USA) and `The_Birds_(film)` (a Hitchcock movie). These resources can be considered as distinct senses of the term `bird`.

For a given keyword, the DBpedia vocabulary will lookup resources that match the keyword, also following redirects. If a match results in a disambiguating resource, each linked resource is also added to the temporary result set. Next, for each resource aliases are collected including the name of the resource if distinct from the keyword. This is achieved by asking for all resources for which a redirect exists to the current resource. The isas of a DBpedia sematag are populated by analyzing the classes to which a resource belongs (indicated by the `rdf:type` relation). Whilst DBpedia concepts are also mapped on other ontologies such as the YAGO knowledge base (Suchanek et al., 2007), we currently require isas to be part of the DBpedia ontology. The reason for this is the level of detail provided by YAGO classes (e.g. `FilmsBasedOnShortFiction`, `1960sHorrorFilms`) as compared to DBpedia classes (e.g. `Film`). Too much detail compromises the general applicability of a sematag. Some examples of tags extracted from DBpedia resources are listed in figure 2. In addition, keywords are directly matched with classes in the DBpedia ontology. If a matching class is found, a sematag is created as follows:

1. the keyword that serves as name for the tag is substituted by an asterisk, indicating that the tag name does not matter;
2. no aliases are added (no `owl:equivalentClass` relations are defined in the DBpedia ontology);
3. the class name is included as isa, as well as any parent classes.

The last sematag depicted in figure 2 is extracted from the `Bird` class in the DBpedia ontology.

3 EXPLAINING TAGS USING TAGS

When a resource is annotated using TagNet, a tag keyword is looked up in available vocabularies. If the keyword is found and multiple senses are detected, the user is requested to select the proper meaning in the context of the resource. However, disambiguating between different meanings of a keyword is not always a trivial task. This is largely due to the fact that several words have multiple senses, many of which we do not use in daily life or even are aware of. For instance, according to WordNet the noun `dog` has seven senses of which most are less commonly used, e.g. 'informal term for a man' and 'metal supports for logs in a fireplace'. Similar, when looking up a keyword in DBpedia, several concepts with the same name yet a different meaning are typically returned. These often include unexpected results because the search term also corresponds to the name of an (infamous) music album, place or alike. A straight-forward approach to let users distinguish between multiple meanings is to present them with a list of explanations as lined up above, and let them pick the intended one. However, this approach postulates some issues preventing quick disambiguation. First, it clearly takes time for a user to read all sense descriptions of a word as to identify the intended sense. Descriptions are often verbose and/or expressed using scientific terms, making

Figure 3: Ambiguous senses of a keyword are explained to users by means of the isas of matching sematags. A dialog visualizes the results of a tag lookup in available vocabularies and allows users to pick the intended sense.

it hard to grasp what is meant exactly (e.g. describing a dog as 'a member of the genus Canis' is still confusing). Secondly, several senses only marginally differ in semantics from each other. This level of detail is redundant for most tagging purposes and causes uncertainty when trying to select the proper sense. To increase the efficiency of perceiving a tag's senses, we present a filtered set of the isas of a tag – instead of sense descriptions – arranged by the (sub)vocabulary they originate from as illustrated in figure 3. These isas are fast to read and hence help to quickly differentiate between senses. Sematags are further organized by their aliases – e.g. `dog`, `utah_prairie_dog`, etc – and can be picked on class level (only isas are included) as well as instance level (aliases are included that map on a unique resource, similar to a URI).

The reason we opted for isas as the primary means to distinguish between tags with a similar name is because we learned that (i) tags extracted from WordNet and DBpedia generally contain more indicative isas than alias and (ii) broader terms seem more helpful than similar terms to understand the semantics of a tag. However, additional user experiments are needed to validate this claim which is based on our own experience. Moreover, to improve the understandability of tags explaining tags, it might be useful to incorporate statistics about the popularity of words to decide which tags are best suited to explain the semantics of tags. Another option is to give up some semantic detail in favor of a simplified tagging experience; a proper balance is needed. A coarser filter could for instance group the WordNet tags with isas `unpleasant_woman`, `chap` and `villain` into a single tag with isas `person` and `organism`. Similar, we could prune the DBpedia tags and only display key terms such as `animal`, `person`, `song`, `album`, `place` and `band`.

4 TagNet AS SEARCH TOOL

In this section we elaborate on the role of sematags to facilitate search operations in a repository populated by resources. We represent a resource by a URI, a name (label), a description and an optional image (thumbnail). Resources are annotated with sematags which are extracted from vocabularies as explained in section 2. The extra information contained in sematags is exploited when retrieving resources. Search terms are not only compared to a tag's name, but are also matched with its aliases and isas such that searching for `animal` will also yield resources that are tagged as `bird` or `dog`. TagNet implements a meta-search algorithm that accepts a mix of sematags and keywords – encoded as sematags with no isas and aliases – to find resources in connected repositories.

A sematag t_1 matches a sematag t_2 if and only if:

1. t_1 and t_2 originate from the same vocabulary, and;
2. t_1 has the same name as t_2 or an alias exists in t_2 with the same name as t_1 or an isa exists in t_1 with the same name as t_2[1], and;
3. all isas contained in t_1 also exist in t_2.

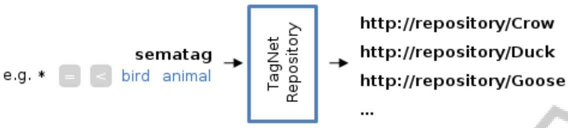

Figure 4: A repository takes sematags as input and outputs resources matching a search query composed from the input.

A search query tunnelled through TagNet can be refined dynamically. Initially, search keywords are passed to TagNet which are matched with the name and tags of resources in a target repository. If the results are considered too many and/or too diverse, the search results are narrowed down by indicating the actual meaning of one or more keywords. To this end, sematags representing the various senses of a keyword are looked up in available vocabularies and presented to the user in a dialog as depicted in figure 3. Finally, selected sematags are sent along with remaining keywords (that were not disambiguated) to a target repository and resources are returned. Hence sematags help users to resolve ambiguity *at search time* and refine a search query.

In the next two sections we discuss how sematags can be used to search for resources in WordNet and DBpedia repositories. Note that the knowledge base underlying WordNet and DBpedia is used both for tag extraction (using vocabularies) and retrieval of resources (using repositories). Sematags originating from the WordNet vocabulary relate to synsets which can be considered as annotated WordNet resources. Similar, it makes sense to query a DBpedia repository using DBpedia sematags because resources in this repository are already (virtually) annotated with DBpedia sematags.

4.1 WordNet Repository

In the WordNet repository, each synset is identified by a URI that is composed of an identifier such as `dog-0` with `dog` being the name of the synset and `0` corresponding to its sense number. Searching for synsets using sematags is achieved by looking up all synsets matching the keyword of the sematag and filtering out the results by comparing aliases and isas. Figure 5 shows that the ability to search through WordNet via sematags is useful for finding resources in a knowledge base that is mapped on WordNet such as SUMO (Niles and Pease, 2001), OpenCyc (Matuszek et al., 2006), DBpedia, etc. Sematags originating from WordNet can be translated into synset URIs which can then be used to query for resources that are linked to particular synsets.

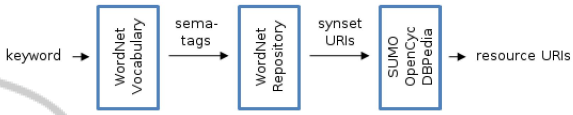

Figure 5: Facts about resources can be inferred from WordNet sematags by first translating the tags into synset URIs and then using these URIs to locate resources in a knowledge base mapped on WordNet.

4.2 DBpedia Repository

In the DBpedia repository, Wikipedia content is seen as a collection of resources that are virtually annotated with sematags. Given a sematag, a search algorithm can look for resources annotated with a matching sematag. Although these sematags do not really exist, we can assume they do according to the following rules based on the tag extraction method outlined in section 2.2:

- each resource is annotated with a sematag having the same name as the resource itself;
- the aliases of the sematag are derived from 'redirects' and 'disambiguates' pointing to the resource;
- the isas of the sematag relate to the class hierarchy of the resource in the DBpedia ontology.

The following SPARQL query collects resources that match this scheme:

```
SELECT DISTINCT ?r ?l ?c ?t
WHERE {
  ?r rdfs:label ?l; dbo:abstract ?c.
  ?r a <$tag.isa1>; a <$tag.isa2>.
  FILTER (bif:contains(?l,'"$tag.name" or
                         "$tag.alias1"')).
  FILTER ((langMatches(lang(?l), "en")) &&
          (langMatches(lang(?c), "en"))).
  OPTIONAL { ?r dbo:thumbnail ?t }
}
```

This query does not include resources connected through `dbo:wikiPageRedirects` or `dbo:wikiPageDisambiguates` properties. We include these resources using separate queries to keep the queries simple and performant.

[1]Note that wildcards are allowed in tag names. A * matches any tag name.

5 EVALUATION

To evaluate the effectiveness of soft semantics to link tagged resources with a semantic knowledge base, we tested how tags used in Web services such as Del.icio.us, Flickr and Youtube can be dynamically mapped on DBpedia or WordNet resources and vice versa. With such links in place, we can infer facts about a photo or video using its tags and involve tagged resources in semantic queries.In a first step, we explored how additional tag detail can be introduced in Flickr. Next, we investigated which steps should be traversed to unambiguously link a collection of popular tags to related DBpedia or WordNet resources.

5.1 Introducing Sematags in Flickr

In Flickr, photos are classified by means of user-generated (ambiguous) keywords. Rather than substituting these tags for URIs of semantic resources – wich are incompatible with a tag-based system like Flickr – we aim to upgrade these tags to sematags and hence remove ambiguity and facilitate mappings to resources through TagNet. However, this means that sematags need to be stored in Flickr, when annotating a photo. We thus need a way to seamlessly inject isas and aliases in a legacy tagging system without breaking its core functionalities (e.g. search functionality). To this end, we consider two approaches:

1. sematags are encoded in a string notation such as `name|alias1|isa1;isa2` (e.g. `atlantis||db:space_shuttle`) and added as a single tag to a link, or;
2. sematags are flattened into an array of tags composed of the name of the tag and its isas (e.g. `name`, `isa1`, `isa2`) and added as distinct tags to a resource.

The former approach is compatible with Flickr (and a.o. also Del.icio.us) and results in a number of benefits: i) free text searches in Flickr now also range over the synonyms and hypermyns of a tag, ii) search queries can be passed through TagNet, semantically refined and forwarded to Flickr using Flickr's open API, iii) unlike hard links, sematags can easily be understood by humans and machines and iv) sematags can be mapped on linked data and vice versa. However, we acknowledge that a custom encoding of sematags is not recognized by existing systems, resulting in poor textual representations of sematags. Sematags could be rendered in a more visually appealing way by hiding aliases and isas by default and depicting those when hovering over a tag or clicking on it. Furthermore, if sematags are not natively supported, internal free text searches will also match vocabulary labels prepended to tags such as `db:` which is not desirable.

Table 1: Scores attributed to the search results of 142 random tags in TagNet using its DBpedia vocabulary.

Sc	Description of score	Tags
A	The first hit matches a corresponding DBpedia resource.	110
B	The results contain a match, but not in the first hit.	10
C	None of the results correspond to a match.	14
D	No results were found.	8

The latter approach gives up on aliases and loses information about relationships between tags. In a flattened array of multiple sematags, it is unclear which tags are actually isas and to which tag they belong. Hence it is impossible to unambiguously map multiple flattened sematags on semantic resources. Yet, matching a sematag with a set of flattened sematags (i.e. the other way round) will only yield false positives in rare cases if tags are compared as follows. A sematag t matches an array of tags T if and only if:

1. T contains t or T contains an alias that exists in t;
2. T contains every isa that exists in t.

A situation where false positives are possible occurs if a sematag t_2 introduces keywords in T that compromise the semantics of a flattened sematag t_1. For instance, if a link is tagged using a keyword `dog` in the sense of an animal and another tag introduces the keyword `person`, then searching for a dog in the sense of a person would incorrectly return the resource. Sematags with wildcards in their name (see figure 2) should also be avoided here. However, the main drawback of this approach is the lack of aliases which are needed by a machine to distinguish between resources annotated with the same set of isas.

5.2 From Tagged Photo to Linked Photo

We used a *public beta release of TagNet*[2] and the all time most popular tags on Flickr[3] as a starting point for our study. These comprise 142 tags, related to the broad domain of photography, and are listed in figure 6. We looked up each tag in TagNet using DBpedia as primary repository and assigned a score based on the relevance of the search results which were limited to 20 results per query for usability reasons. The results of this step are summarized in table 1

[2]http://sematags.belllabs.be/
[3]http://www.flickr.com/photos/tags/

animals, architecture, art, asia, australia, autumn, baby, **band**$_C$, barcelona, beach, berlin, **bike**$_C$, bird, birds, birthday, black, **blackandwhite**$_D$, blue, **bw**$_C$, california, canada, **canon**$_B$, car, cat, chicago, china, christmas, **church**$_B$, city, clouds, color, concert, dance, day, **de**$_C$, dog, england, europe, fall, family, fashion, festival, film, florida, flower, flowers, food, football, france, **friends**$_B$, fun, garden, geotagged, germany, girl, graffiti, green, halloween, hawaii, holiday, house, india, **instagramapp**$_D$, iphone, **iphoneography**$_D$, island, italia, italy, japan, **kids**$_C$, **la**$_C$, lake, landscape, light, **live**$_B$, london, love, **macro**$_B$, **me**$_D$, mexico, **model**$_B$, museum, music, nature, **new**$_C$, newyork, **newyorkcity**$_D$, night, nikon, nyc, ocean, **old**$_C$, paris, park, party, people, photo, photography, photos, portrait, **raw**$_B$, red, river, **rock**$_B$, **san**$_C$, **sanfrancisco**$_D$, scotland, sea, seattle, **show**$_C$, sky, snow, spain, **spring**$_C$, **square**$_B$, **squareformat**$_D$, street, summer, sun, sunset, taiwan, texas, thailand, tokyo, travel, tree, trees, **trip**$_C$, uk, **unitedstates**$_D$, **urban**$_B$, usa, vacation, **vintage**$_C$, **washington**$_C$, water, wedding, white, winter, woman, yellow, zoo

Figure 6: By default, 110 out of 142 popular Flickr tags (77.5%) are mapped correctly on a valid DBpedia resource through TagNet (*A* score). Tags that need additional attention to resolve ambiguity are marked in bold and labeled with a *B*, *C* or *D* score (see table 1).

and marked on figure 6. It shows that 110 tags received an *A* score, meaning that they were correctly mapped on a corresponding DBpedia resource in the first hit. For example, the tag `fall` is resolved into http://dbpedia.org/resource/Autumn and `nyc` maps on http://dbpedia.org/resource/New_York_City. To denote the connection with DBpedia, we added the `db` prefix to tags with an *A* score such that TagNet knows which repository to use to map the tag on a URI.

For tags with a *B* score, additional detail should be added to overcome ambiguity. For instance, `canon` is in the first place known by DBpedia as a city in Georgia, a priest, a list of topics related to Dutch history, etc, while in the context of photography it refers to a company specialized in the manufacturing of imaging and optical products. From the related DBpedia resource (http://dbpedia.org/resource/Canon_(company), two isas can be extracted (`company` and `organisation`) which give rise to a sematag `db:canon||company,organisation` – encoded in a Flickr-compatible format as discussed in section 5.1 – that uniquely identifies the resource in TagNet. Similar, `friends` is recognized by default as a sitcom while the resource http://dbpedia.org/resource/Friendship is actually the best match for this tag's meaning. The `Friendship` resource has no specific isas, but by including its name as an alias to a `friend` tag (i.e. `friend|db:friendship`), TagNet can distinguish between the different senses. As such, each DBpedia resource can be described unambiguously by a human-understandable sematag that can be dereferenced to a URI via TagNet and vice versa. Note that we prefer to augment a tag with isas (if available) over aliases derived from a resource's label since these specific aliases often tend to be spelling variants of the tag name or informally refer to its isas. For instance, an alias of `canon` in the sense of the Japanese multinational would be `canon_(company)`.

Tags with a *C* or *D* score need extra attention. No resources with a matching name exist in DBpedia or they are not in line with the meaning of the tag. This leaves us with two options: i) lookup the tag in a secondary repository or ii) replace the tag by a similar tag or add *A*-rated aliases or isas to the tag. By relying on WordNet as secondary repository, seven more tags (`band`, `bike`, `kids`, `new`, `old`, `show`, `washington`) were attributed an *A* or *B* score and thus could be upgraded to sematags with a `wn:` prefix (e.g. `wn:bike`).

To clarify the semantics of the remaining 15 tags, we have to find at least one meaningful isa or alias for each tag. For instance, the tag `me` and `iphoneography` can be annotated with `db:person` and `db:blog` isas respectively. Tags like `newyorkcity` and `sanfrancisco` need an alias that is spelled differently (e.g. `db:nyc`, `db:san_francisco`) or substituted by this alias, while the tag `instagramapp` can be understood using an `instagram` alias.

Tags like `blackandwhite` (and by extension also typical Twitter hashtags such as `#savetheplanet`) are more difficult to map on linked data as they denote a very specific property, a state of mind or expression which is hard to describe using formal semantics. In summary, we showed that 127 out of 142 random tags (89.4%) could be mapped with minimal effort on known concepts using DBpedia as primary and WordNet as secondary vocabulary. *By systematically enriching a tag with additional tags, a (sema)tag becomes an alternate notation for a URI that scales better to tag-based systems like Flickr, as it is human readable and supports free text queries (including synonym and hypernym matching).*

6 ARCHITECTURE AND IMPLEMENTATION

TagNet is developed as a Java Web application, using Servlet technology in the back-end and AJAX technology in the front-end. It offers a REST API and has an open-ended design such that custom vocabularies and repositories can easily be plugged in by implementing a `Vocabulary` and `Repository` interface respectively. An overview of the architecture is presented in figure 7. The WordNet vocabulary and repository make use of the WordNet 3.0 database files[4] while the DBpedia vocabulary and repository rely on the online DBpedia Virtuoso SPARQL endpoint[5]. The dependency on the DBpedia SPARQL

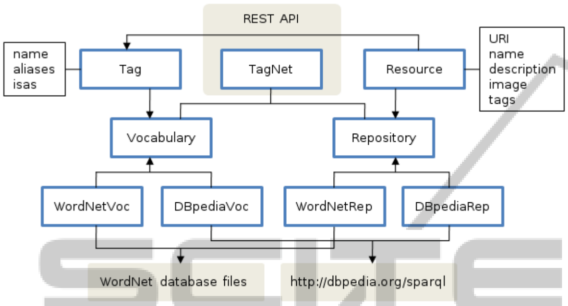

Figure 7: TagNet architecture overview.

query engine is a bottleneck in the current beta implementation. On the one hand, it allows TagNet to run on light-weight servers with limited memory available, but on the other hand we rely on live data which might not always be available in time. Each request for sematags or resources is translated into SPARQL queries which are directed to the online DBpedia SPARQL endpoint. Although a lot of effort was spent in optimizing these queries, we experienced huge differences in their processing time which is probably due to a variable load of the Virtuoso SPARQL engine over time. While the execution of a query can be considered relatively fast at one moment, the same query might time out moments later. To bypass these performance issues, we replicated the DBpedia database on a local server such that network latencies and processing delays caused by high loads were avoided. Furthermore, we could cache sematags that were looked up in the DBpedia vocabulary and pre-generate a repository of annotated resource URIs. This would dramatically speed up the matching of tags since additional queries are only needed to collect the data associated with matching resource URIs.

7 RELATED WORK

Previous work that has been done in the area of tagging is quite diverse. For instance, models have been proposed to represent relationships between agents, resources and tags and augment user-contributed data (Newman, 2005; Gruber, 2007); frameworks were proposed to add meaning to tags (Passant and Laublet, 2008; Garcia-Castro et al., 2009; Hepp, 2010); sharing and reuse of social tagging data has been studied (Golder and Huberman, 2006; Kim et al., 2008) as well as recommendation algorithms (Song et al., 2008; Araujo et al., 2010; Sigurbjörnsson and van Zwol, 2008).

In the remainder of this section, we elaborate on the works with the closest match to TagNet and discuss how they differ or match. MOAT (Passant and Laublet, 2008) extends an ontology designed for tagging (Newman, 2005) and aims to enrich free tags (i.e any user-defined keywords) with additional meaning. Similar to TagNet, MOAT looks up the global meaning of keywords in a controlled vocabulary and allows users to select the appropriate meaning, or define a new meaning by referring to a Web resource (e.g. a DBpedia resource). Unlike tags in MOAT which are stored externally, sematags can be injected in real-world tagging systems and mapped on knowledge bases through TagNet.

Another approach to add meaning to tags is presented in Tags4Tags (Garcia-Castro et al., 2009) where the underlying meaning of tag can be revealed by means of another tag. In this work, the typical meta-model in which a Web resource maintains one or more `hasTag` relations with tag literals is expanded with typed relationships between a pair of tags. The ideas postulated in Tags4Tags were reused in HyperTwitter (Hepp, 2010). Using so-called 'tripletweets', tag equivalence (e.g. `#webist13 = #webist2013`, tag specializations (e.g. `#tennis subtag #sports`) and predefined relations between tags (e.g. `#munich >translation #muenchen`) can be expressed. This is completely in line with the vision of TagNet: a Twitter vocabulary can process tripletweets and generate sematags out of them. Moreover, sematags could be incorporated in HyperTwitter to express that the hashtag `#webist13` is a subtag of a sematag `webist`.

To cope with large datasets and relieve users from manual tagging steps, recommendation algorithms were proposed that can (semi-automatically) generate annotations from Web pages (Song et al., 2008; Araujo et al., 2010) and images (Sigurbjörnsson and van Zwol, 2008). Additional research is needed to investigate how sematags could be generated from arbitrary Web resources, i.e. how the correct sense of a keyword could be derived from the current context.

In (Weller, 2007), Weller compares ontologies and folksonomies and suggests that they can be seen as the two ends of a scale of documentation languages

[4]http://wordnet.princeton.edu/wordnet/download/
[5]http://dbpedia.org/sparql/

ranging from unstructured to highly formalised systems. Rather than seeing them as rivals, they can be considered as elements in a toolbox which can be used together to support concrete applications. In this work, we showed how soft semantics blur the distance between folksonomies (no semantics) and ontologies (hard semantics) and help to complete each other.

8 CONCLUSIONS AND OUTLOOK

The key contributions of TagNet are twofold. First, TagNet introduces sematags which annotate regular keywords with isas and aliases, hence solving typical tag-related issues such as dealing with ambiguity, spelling variants and variations in the specificity of tags. Unlike other approaches, sematags do not include hard links to Web resources but rather contain a minimal set of information – extracted from pluggable vocabularies – that is used to lookup related resources in a repository. This loose coupling guarantees that folksonomies remain folksonomies (using richer tags) yet unambiguous links to concepts in formal knowledge bases can still be retrieved. By supporting WordNet and DBpedia as default vocabularies, we cover a wide range of contemporary meaningful tags. Second, TagNet serves as an extensible metasearch engine. We illustrated how TagNet is used to search through DBpedia using sematags and explained how other repositories can be supported. We also indicated how sematags can be scaled to support legacy tagging systems and give rise to enriched folksonomies. A beta version of TagNet is available on http://sematags.belllabs.be/.

In future work, we want to further explore and validate the effectiveness of using tags to explain the different senses of a keyword to users. Another interesting path to explore is the use of extended 'facets' (categories and subcategories to which resources belong) to narrow down search results. Sematags support basic faceted search by default as isas classify resources in categories. To better align with existing faceted search engines, we could indicate how many resources match a sematag while refining a search operation using the dialog depicted in figure 3. In (Ben-Yitzhak et al.,), Ben-Yitzhak et al. also explained the importance of gaining insight in the data behind facets which is far richer than just knowing the quantities of resources that belong to each facet. We see an opportunity to include information about the properties of a resource in (intermediate) search results and (refined) search queries. For instance, we can also annotate properties of resources using sematags. Searching for e.g. 'birthplace artist' in DBpedia with 'birthplace' and 'artist' both being resolved to sematags – the former matching a property, the latter matching resources – would result in a list of instances of the DBpedia `Artist` class for which a `birthPlace` property is defined (which is also included in the search results).

REFERENCES

Araujo, S., Houben, G.-J., and Schwabe, D. (2010). Linkator: Enriching Web Pages by Automatically Adding Dereferenceable Semantic Annotations. In *10th International Conference on Web Engineering (ICWE'10)*, pages 355–369. Springer-Verlag.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2008). DBpedia: A Nucleus for a Web of Open Data. In *6th International Semantic Web Conference (ISWC'07)*, pages 722–735.

Ben-Yitzhak, O., Golbandi, N., Har'El, N., Lempel, R., Neumann, A., Ofek-Koifman, S., Sheinwald, D., Shekita, E., Sznajder, B., and Yogev, S. Beyond Basic Faceted Search. In *International Conference on Web Search and Web Data Mining (WSDM'08)*.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.

Garcia-Castro, L. J., Hepp, M., and Garcia, A. (2009). Tags4Tags: Using Tagging to Consolidate Tags. In *20th International Conference on Database and Expert Systems Applications (DEXA'09)*, pages 619–628. Springer-Verlag.

Golder, S. and Huberman, B. A. (2006). The Structure of Collaborative Tagging Systems. *Journal of Information Science*, 32(2):198–208.

Gruber, T. (2007). Collective Knowledge Systems: Where the Social Web meets the Semantic Web. *Web Semantics Science Services and Agents on the World Wide Web*, 6(1):4–13.

Hepp, M. (2010). HyperTwitter: Collaborative Knowledge Engineering via Twitter Messages. In *17th International Conference on Knowledge Engineering and Management by the Masses (EKAW'10)*, pages 451–461. Springer-Verlag.

Kim, H.-L., Breslin, J., Yang, S.-K., and Kim, H.-G. (2008). Social Semantic Cloud of Tag: Semantic Model for Social Tagging. *Agent and Multi-Agent Systems: Technologies and Applications*, pages 83–92.

Matuszek, C., Cabral, J., Witbrock, M., and DeOliveira, J. (2006). An Introduction to the Syntax and Content of Cyc. In *AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, pages 44–49.

Newman, R. (2005). Tag Ontology Design. http://www.holygoat.co.uk/projects/tags.

Niles, I. and Pease, A. (2001). Towards a Standard Upper Ontology. In *International Conference on Formal Ontology in Information Systems (FOIS'01)*, pages 2–9. ACM.

Passant, A. and Laublet, P. (2008). Meaning Of A Tag: A Collaborative Approach to Bridge the Gap between Tagging and Linked Data. In *WWW'08 workshops: Linked Data on the Web (LDOW'08)*.

Sigurbjörnsson, B. and van Zwol, R. (2008). Flickr Tag Recommendation Based on Collective Knowledge. In *17th International Conference on World Wide Web*, pages 327–336. ACM.

Song, Y., Zhuang, Z., Li, H., Zhao, Q., Li, J., Lee, W.-C., and Giles, C. L. (2008). Real-time Automatic Tag Recommendation. In *31st International Conference on Research and Development in Information Retrieval (SIGIR'08)*, pages 515–522. ACM.

Specia, L. and Motta, E. (2007). Integrating Folksonomies with the Semantic Web. In *4th European Semantic Web Conference (ESWC'07)*, pages 624–639. Springer.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). YAGO: A Core of Semantic Knowledge. In *16th International World Wide Web Conference (WWW'07)*, pages 697–706. ACM Press.

Vander Wal, T. (2007). Folksonomy Coinage and Definition. http://vanderwal.net/folksonomy.html.

Weller, K. (2007). Folksonomies and Ontologies: Two New Players in Indexing and Knowledge Representation. In *Applying Web 2.0. Innovation, Impact and Implementation*, pages 108–115.