RENDER
FP7-ICT-2009-5
Contract no.: 257790
www.render-project.eu

RENDER

Deliverable D1.1.3

Final Collection of Data

Editor:	Mariana Damova (OT)
Authors:	Mariana Damova (OT), Delia Rusu (JSI)
Deliverable nature:	Report (R)
Dissemination level: (Confidentiality)	Public (PU)
Contractual delivery date:	31 December 2012
Actual delivery date:	30 December 2012
Suggested readers:	Data integrators, users of RENDER data layer
Version:	1
Total number of pages:	28
Keywords:	FactForge, reason-able view, diversity management, reasoning, RENDER data layer, RENDER data infrastructure, basic data layer, secondary data layer, KDO, PROTON, DBpedia, Freebase, Geonames, SIOC, DC, DMOZ, Schema.org, nested repositories

Full Project Title:	RENDER – Reflecting Knowledge Diversity
Short Project Title:	RENDER
Number and Title of Work package:	WP1 – Data collection and management
Document Title:	D1.1.3 - Final collection of data
Editor (Name, Affiliation)	Mariana Damova (OT)
Work package Leader (Name, affiliation)	Maurice Grinberg, Ontotext
Estimation of PM spent on the deliverable:	11

Executive Summary

This deliverable reports about the datasets that are part of the final collection datasets for the RENDER data layer. The RENDER data layer consists of a basic data layer, and several secondary data layers, which are domain specific, and serve the RENDER use cases, e.g. Telefonica, Google and Drupal. The datasets of the RENDER basic data layer represent a segment of the LOD cloud, comprising some of the most popular LOD datasets. The criteria for their selection and the versions that are included in it are described in greater detail. Except for the datasets, RENDER basic data layer includes several general purpose, very popular ontologies, and the reference knowledge stack, e.g. the unification ontologies layer, mainly PROTON, which allows to efficiently manage and access the data. The datasets of the secondary RENDER data layers are RDF data derived from the JSI Enrycher service, which has processed Google cluster news documents, and Twitter statuses. These secondary layers employ the developed within RENDER project OWLIM feature - nested repositories. Both the RENDER basic data layer and the RENDER secondary data layers are resolved as reason-able views, e.g. the explicit statements of the single datasets are augmented with inferred knowledge, based on OWL-Horst and FactForge specific inference rules, which include rules for consistency checks while loading the data. The PROTON ontology is extended with disjoint statements that allow for reducing the number of wrongly inferred statements. This deliverable shows the final data collection of datasets, data management methods and implementation approaches that will be part of the RENDER final data integration.

List of Authors

Organisation	Author
JSI	Delia Rusu
Ontotext	Mariana Damova

Table of Contents

Executive Summary ... 3
List of Authors... 4
Table of ContentsList of Figures ... 5
List of Figures... 6
Abbreviations... 7
1 Introduction ... 8
2 RENDER Basic Data Layer ... 10
2.1 Datasets ... 11
2.2 Reference Knowledge Stack... 12
2.3 Ontologies... 13
2.4 Inference Rules ... 14
2.5 SPARQL Endpoint of RENDER Basic Data Layer... 15
3 Diversity through Inferred Knowledge... 18
4 Render Secondary Data Layers... 20
4.1 News... 20
4.2 Tweets... 21
5 Sentiment Analysis... 23
6 Statistics ... 24
7 Conclusion... 25
References... 26

List of Figures

Figure 1 Results for the SPARQL query “Airports near London” 16
Figure 2 Visualization one of the results from the query “Airports near London” 16
Figure 3 Visualization two of the results from the query “Airports near London” 16
Figure 4 Visualization three of the results from the query “Airports near London”..................... 17
Figure 5 Mode “Exploration” showing the available explicit and implicit knowledge about Copenhagen 18
Figure 6 Mode “Exploration” showing the available implicit knowledge about Northern Europe 18
Figure 7 Results of the SPARQL query “Airports near London” with PROTON predicates only..................... 19
Figure 8 Unexpected inferred knowledge, the city of Missouri as a subject of a science fiction author 19
Figure 9 Secondary RENDER data layer, nested into the Basic one 20
Figure 10 RDF representation for Google news documents 21
Figure 11 RDF representation for Tweets statuses 22

Abbreviations

KDO	Knowledge Discovery Ontology
LOD	Linked Open Data
RDF	Resource Description Framework
TID	Telefónica Investigación y Desarrollo
UI	User Interface
URI	Unique Resource Identifier
OWL	Web Ontology Language

1 Introduction

Semantic Technologies have been introduced to meet the information management needs of 21st century. They are based on standards for data modeling and representation, and rely on data storage software, called triple stores or semantic repositories. Semantic technologies allow for an unprecedented ease of integration of heterogeneous data sources. When comparing with the RDBMs paradigm, there is nothing that can be expressed with semantic technologies that cannot be expressed in relational models. However, the difference between these two technologies is in the cost of production and subsequent maintenance and also in the efficient hardware utilization.

Unlike relational models, RDF, the basic data format of semantic technologies, is schema agnostic. Both data and schemata are represented in one and the same RDF format. It is as easy to add data as to change their schema. What is required is to add RDF triples into the semantic repository. This makes the cost of maintenance of semantic repositories much lower over time than the cost of maintenance of RDBMs. Additionally, only the available information is being stored in the triple stores, e.g. there is no waste of valuable space from the need of complying with specification of the relational tables by leaving empty cells.

Another feature of the semantic models is that they allow inference, e.g. the number of the explicitly introduced in the database facts can be up to 1/3 or even less than the actually available facts for querying, as new knowledge gets created based on the models that allow for the generation of new implicit facts out of the explicit ones.

Third, semantic technologies allow very easy interlinking between data from different sources, which enables retrieval of information from different datasets with a single query.

RENDER data infrastructure supports diversity in several ways. It lies inside OWLIM [4], the most scalable semantic repository, which has been the backend backbone of the first in the world real-life semantic application, BBC World Soccer Championship 2010 website, and BBC London 2012 Olympics website. It is capable of standing millions of queries a day, sent and executed concurrently. Its feature, nested repositories, developed within RENDER project, extends the technological infrastructure of RENDER providing with a mechanism of flexible sharing of a common pool of large amounts of data for different purposes and mixing them up with any number of other independent specific data pools.

RENDER data infrastructure is a reason-able view of the web of data. It contains 9 of the most popular LOD datasets with inference according to OWL-Horst [48] optimized performed on them. These data sets cover general knowledge (DBpedia [14], Freebase [20], New York Times [33]), geographical information (Geonames [21]), linguistic information (Wordnet [51], Lingvoj [30], Lexvo [29]), music information (Musicbrainz [32]), statistical information (CIA Factbook [8]) and are interlinked. This means that querying information about one of the datasets returns relevant information from all datasets. It also provides with a reference layer of an upper-level light-weight ontology (PROTON [36]) that allows for an easy and efficient access to the pull of heterogeneous data. The reference layer is being augmented with mappings to Schema.org [41]. So, the RENDER data infrastructure can be queried via the reference layer, or via the schemata of its datasets. It also contains several popular schemata, like SKOS [42], Dublin Core [17], FOAF [19] and the like. RENDER data infrastructure integrates two sets of data (news and tweets) produced by processing unstructured texts via the mechanism “nested repositories”, and includes the developed within RENDER project KDO (Knowledge diversity ontology).

This integration allows for formulating queries related both to the news or tweets and to the general knowledge in the RENDER data infrastructure layer. This is impossible without the integration. For instance, the news dataset contains information about documents, their topics, and their sentiments, where the topics refer to entities which are extensively characterized in the data of the basic RENDER data infrastructure layer. So, if the news layer has identified that a given news is about Barack Obama, who is a Person, the basic data infrastructure layer will provide information about Barack Obama’s date of birth, political party affiliation, job position, place of birth, spouse, etc. This will allow to formulate queries not about particular entities, but about more general characteristics of these entities, and link them with diversity information, based on the knowledge diversity ontology, for instance, “give me very negative documents about presidents of the USA” and inference over sentiments. The same holds for the tweets corpus. We have information about the tweets, where queries can be formulated that help conveniently combine rich factual information with diversity information. For instance, “positive tweets about Movistar and Artists born in Germany”.

So, RENDER data infrastructure gives much wider possibilities for querying and especially quickly updating the use case scenarios in which it is used. For instance, Telefonica use case is originally set up to take information from the tweets, e.g. “positive tweets from Germany in May 2012”. After the first prototype is shown and evaluated by their Marketing Department, and requirements analysis is being performed with respect to what other information will be interesting to have visualizations for, it will be easy to extend the demonstrator functionality, as the data is already available.

The RENDER data infrastructure, based on OWLIM, is capable of providing interesting and sometimes counter-intuitively diverse related entities, for instance “United Nations”, a topic of a news document, selects the following related entities: 2005 World Summit, Human immunodeficiency virus, Act of Free Choice, Activities of the Holy See within the United Nations system, Afghanistan.

This deliverable presents the final data collection for RENDER data infrastructure. It describes the version of RENDER data infrastructure developed and updated this year so that RENDER use cases can make use of it. We distinguish between RENDER basic data layer, which includes FactForge (cf. Section 2), and RENDER secondary data layer, which is a set of nested repositories storing RDF, processed by JSI Enrycher, and interlinking it with entities from RENDER basic data layer (cf. Section 4).

The final RENDER data infrastructure will contain the latest versions of the datasets, DBpedia 3.8, Freebase, October 2012, Geonames 2.0.2., Wordnet 3.0, Musicbrainz, december 2012. It will have two nested repositories to contain news and tweets data respectively. It will provide with the API to access and use the data programmatically, and GUIs to allow humans to access and navigate the data. A cloud solution is not necessary as the response times of OWLIM are satisfactory. Loading Wikidata data into OWLIM is another instance of data collection that will be carried out in RENDER project. Wikidata data model supports diversity, and will provide a data pool with collectively user generated knowledge in 200 languages. The expected size of the data is 4-10 mln entities with about a thousand triples per entity.

2 RENDER Basic Data Layer

The RENDER basic data layer is a reason-able view, based on FactForge (http://factforge.net) the public service developed and maintained by Ontotext. It is a compound dataset of some of the most popular datasets from the Linked Open Data Cloud [31].

The reason-able views are based on the following insights:

- The grouping of the selected datasets and ontologies in a compound dataset, which becomes a single body of knowledge – integrated dataset – with respect to reasoning and data evaluation
- Loading of the compound dataset in a single semantic repository in order to make query evaluation and reasoning practically feasible. It can be considered as an index which caches parts of the Linking Open Data (LOD) cloud and provides access to the datasets included as Web search engines index WWW pages and facilitate their usage
- The performance of inference with respect to tractable OWL dialects. Given all public results, only OWL-Horst-like languages seem to be suitable for reasoning with data in the range of billions of statements

Complying with all these aspects makes reasoning with the Linking Open Data (LOD) feasible. This is because a basic level of consistency of the data is being guaranteed, along with a guaranteed service availability because the compound dataset is loaded into a single repository. This allows for the easier exploration and querying of unseen data and ensures a lower cost of entry.

The constitution of reason-able views obeys special selection criteria for the datasets, for example, the datasets must allow inference and deliver meaningful results under the semantics determined for the view. Further, it is necessary that the datasets are easy to define and isolate, for example, they must be clearly distinguishable from other datasets. In many cases, additional manipulations on the datasets like cleanups are required. The datasets must allow easy and cheap cleanup manipulations that can be performed on them in an automated and semi-automated fashion. Finally, the datasets must be more or less static and able to function in predictable way, opposite to database wrappers, which implement complex mappings to be reusable in unplanned contexts, such as Web-based applications or federated systems, where RDF is generated in answer to retrieval requests.

In other words reason-able views are built according to the following design principles:
(a) All the datasets in the view represent linked data;
(b) Single reasonability criteria is imposed on all datasets;
(c) Each dataset is connected to at least one of the others.

FactForge, which emerges as RENDER basic data layer, is a collection of vast amounts of heterogeneous general purpose data, selected from the central datasets of the LOD cloud. FactForge is designed for two purposes:

- Serve as a useful index and entry point to the LOD cloud
- Present a good use case for large-scale reasoning and data integration

The final data collection of RENDER is intended to produce a new version of FactForge, FactForge 3.0, which will include the most recent versions of the main datasets. The datasets are described in section 2.1. Except from the datasets, the final data collection also comprises several popular ontologies, a new version of the Reference Knowledge Stack, synchronized with the newest version of the datasets and their schemata. The Reference Knowledge Stack is extended with one more ontology, e.g. schema.org. Finally, the set of inference rules that complement the standard OWL-Horst optimized rule set are revised and new rules are added to it.

FactForge is the biggest and most heterogeneous body of factual knowledge on which inference has been performed.

2.1 Datasets

The datasets of RENDER basic data layer, FactForge, are organized in three categories:

- General knowledge
- Linguistic knowledge
- Domain specific knowledge
    - Geography
    - Music

The general knowledge datasets of RENDER basic data layer, FactForge, are:

- **DBPedia 3.8**

DBPedia 3.8 is based on Wikipedia dumps from late May/early June 2012. DBPedia ontology has 359 classes (40 more than DBPedia 3.7). The localization covers 111 languages. The English version of DBpedia 3.8 describes 3.77 million things, out of which 2.35 million are classified in the ontology. This includes persons, places, creative works, organizations, species, diseases, etc. The full DBpedia has labels and abstracts for 10.3 million unique things in 111 different languages, 8.0 million links to images and 24.4 million HTML links to external web pages. The overall dataset has 1.89 billion RDF triples, where 400 mln come from the English Wikipedia, 1.46 billion come from other language editions, and about 27 million are data links into external RDF datasets.

- **Freebase**

Freebase (http://freebase.com) is a large collaborative knowledge base, an online collection of structured data harvested from many sources, including individual wiki contribution. Freebase contains data from Wikipedia [50], Chemoz, NNDB [34], MusicBrainz and individually contributed data from its users. It has about 20 million topics or entities and no defined ontology. The entities described in this knowledge base are in structured predicate names, which reflect a hidden class hierarchy. Freebase has an overall of 19632 predicates with a structure of the predicate name in which the left most word denotes the subject domain of the property; the middle word denotes a class which is the domain of the property denoted by the last right most word, e.g.

government.legislative_session.date_ended
celebrities.romantic_relationship.end_date

The topics and entities are described in two parallel models. Most of the topics are associated with one or more types (such as people, places, books, films, etc) and are assigned additional properties. These types and properties and related concepts are referred to as Schema by Freebase authors.
The latest RDF data dump of Freebase is from 04.12.2012. Its compressed size is of 7.5 GB, and uncompressed size is of 49.5 GB. It contains approximately 736 million triples.

- **CIA Factbook**

The CIA Factbook provides information about 267 world entities. It containts facts about history, people, government, economy, geography, communications, transportation, etc., including maps. The RDF data dump that is being used for FactForge and RENDER basic data layer is from 2012.

- **New York Times**

In 2010 New York Times published a dataset of manually tagged entities of people (4,978), organizations (1,489), locations (1,910) as linked open data. This dataset is also part of FactForge and RENDER basic data layer.

The linguistic knowledge datasets of RENDER basic data layer, FactForge, are:

- **Wordnet 3.0**

Wordnet 3.0 is the last version of the world thesaurus of linguistic knowledge comprising a vast amount of interlinked in synonym sets and hierarchical structure of hyponyms and hypernyms. It contains 155,287 words organized in 117,659 synsets for a total of 206,941 word-sense pairs; in compressed form, it is about 12 megabytes in size. Wordnet 3.0 with its RDF version dates from 2006. A newer version Wordnet 3.1 from 2011 is only available for online use. So, FactForge and RENDER basic data layer include Wordnet 3.0.

- **Lingvoj**

Lingvoj means languages in Esperanto. It is a knowledge base about the world languages. Lingvoj has been part of FactForge since its beginning. Lingvoj contains 22442 triples and has links to DBpedia, UMBEL. Its last version is from 2009, which is included in FactForge and RENDER basic data layer. This year its URIs have been redirected to another data base of human languages, Lexvo.

- **Lexvo**

Lexvo is another knowledge base that includes information about world languages. It comprises 7000 human languages. Its URIs are written with scripts like Latin, Cyrillic, Korean, etc. Its scope is broader than the scope of Lingvoj, providing information about languages, words, characters, and other human language-related entities. It shows that all knowledge in the world can be represented as links to and from linguistic objects by considering semantic relationships between multilingual labels. The URIs of Lexvo are dereferenceable and highly interconnected with other resources on the Web. Its latest version is from March 2012. This version is selected to be part of FactForge and RENDER basic data layer.

The domain specific knowledge datasets of RENDER basic layer, FactForge, are:

- **Geonames**

Geonames is a geographical knowledge base, containing over 8 million placenames, their geolocations and other geographical and demographic information. Its last RDF version is from October 2012. Previous load of Geonames into FactForge and RENDER basic data layer, pointed to flaws in Geonames model, which made it unadapted to the open world WWW. The new version of Geonames comes as a result of discussions between Ontotext and Geonames authors. The last RDF version of Geonames is part of the final data collection of FactForge and RENDER basic data layer.

- **Musicbrainz**

Musicbrainz is an initiative which collects information about music, musical artists, styles, etc. of any kind in a community maintained open source encyclopaedia. Its last RDF version is from December 2012. Musicbrainz is a new addition of FactForge and RENDER basic data layer datasets. It has been loaded successfully into FactForge in its version from June 2012. Musicbrainz URIs are very carefully elaborated. In many cases the first results from SPARQL queries about people, locations over FactForge are from Musicbrainz. The reason for this are under investigation.

2.2 Reference Knowledge Stack

The concept of a Reference Knowledge Stack has been introduced in the RENDER project. It refers to defining a unification ontology, which is linked to the ontologies of FactForge datasets, and serves as a common entry point to FactForge's data, making querying, accessing, managing and navigating this wealth of data more efficient and manageable.

The unification ontology of FactForge, and RENDER basic data layer is an upper-level ontology PROTON, mapped to the schemata of DBpedia, Geonames and Freebase covering diverging conceptualizations, structural differences, missing concepts [10], [11]. This reference layer makes loading of the LOD ontologies unnecessary, optimizing the reasoning processes and allows for quick and seamless data integration of new datasets with the entire LOD segment of FactForge. It also ensures better interfacing with other components via SPARQL as the queries are more compact and easy to formulate, response times are faster, because of less joins that are employed, and a wealth of inferred knowledge across the datasets, which allows for real journey of knowledge discovery, and navigation from different stand points, as triples with differently named semantically identical properties are retrieved by means of a single PROTON predicate, the inference brings valuable new insight as new knowledge is derived from all datasets in combination.

The PROTON ontology has been developed according to OntoClean principles [22] and generalizing over SUMO [46], DOLCE [16], Cyc models [9], has been used as the central unifying ontology for FactForge, or RENDER basic data layer. The effects of using it for querying are explained in section 3 below. Its official release PROTON 3.0 covers the schemata of DBpedia 3.7, Freebase, October 2011, Geonames 2.0.1. PROTON was moved to a new namespace, e.g. http://www.ontotext.com/proton/.

PROTON 3.0 contains close to 550 classes and about 150 properties. The schema level mappings allow to query the entire FactForge by using only the concepts of PROTON. The idea of the reference ontology has been exploited by approaches for federated querying over LOD. PROTON has been used in experiments using automated ontology matching to return federated results from multiple SPARQL end points, while formulating the initial query with PROTON predicates only [25].

As final data collection for RENDER PROTON ontology is aligned with the schemata of DBpedia 3.8, Freebase from October 2012, Geonames 2.0.2 and will be part of the Final data integration layer of RENDER.

Except from PROTON, the final data collection of RENDER basic data layer includes 78 Schema.org [41] mappings to PROTON. This extends the Reference Knowledge Stack with one reference ontology, and allows to formulate SPARQL queries using Schema.org predicates only. Schema.org and these mappings will be part of the RENDER final data integration.

2.3 Ontologies

Except for the datasets, described in section 2.1 and PROTON, described in section 2.2, FactForge and RENDER basic data layer includes several popular schemata. These are:

- **DCMI Metadata Terms (Dublin Core - DC)**

Dublin core is a popular metadata initiative providing a model adapted to represent library information. It has 15 properties. Dublin core has been very widely used by libraries and publishers. Dublin Core is part of a larger initiative DCMI, which aims at including these 15 properties into a larger system of vocabularies allowing to model richer library environments.

- **SKOS (Simple Knowledge Organization System)**

SKOS provides a standard way to produce knowledge systems in RDF, by providing the standards for building thesauri, classification schemes, taxonomies, subject heading sections, etc., all these connections to the framework of the Semantic Web.

- **RSS**

RSS stands for really simple syndication. It is a web content syndication format. It complies with W3C recommendations of XML representation. It includes information about the authorship and the date of publication of a given news feed.

- **FOAF**

FOAF, Friend of a Friend, is a model allowing to describe person data and relationships in machine readable form, abiding the principles of the Semantic Web. Its last version of 2010 is part of RENDER basic data layer.

- **Freebase ontology**

Ontotext is developing a Freebase OWL ontology, based on the Freebase schema, described in section 2.1. This ontology will allow to produce inferred statements out of Freebase entries that are produced as a result of OWL reasoning.

2.4 Inference Rules

The inferencing strategy in OWLIM, the repository of RENDER basic data layer, is one of materialization based on R-Entailment as defined by ter Horst [48], where Datalog [13] like rules with inequality constraints operate directly on a single ternary relation that represents all triples. In addition, free variables in rule heads are treated as blank nodes. Materialization involves computing all the entailed statements at load time. While this introduces additional reasoning cost when loading statements into a repository, the desirable consequence is that query evaluation can proceed extremely quickly. Several standard rule sets are included in all editions of OWLIM and these include:

empty – no inference;

rdfs[1] – RDFS semantics using rule entailment, but without data-type reasoning, i.e. without the literal generalization and related rules;

owl-horst – equivalent to pD*, again without data-type reasoning;

owl-max – RDFS and OWL-Lite (that can be captured in rules);

owl2-ql – a fragment of OWL2 Full based on DL-Lite$_R$, a variant of DL-Lite that does not require the unique name assumption;

owl2-rl[2] – the OWL2 RL profile, a fragment of OWL2 Full amenable to implementation on rule-engines, but without data-type reasoning.

In addition to the standard semantics, user-defined rule-sets can be used. In this case the user provides the full pathname to a custom rule file that contains definitions of axiomatic triples, rules and consistency checks. For ease of use, the rule files for the standard rule-sets are included in the distribution and users can modify or extend these for their specific purposes.

Consistency checks are used to ensure that the data model is in a consistent state and are applied whenever an update transaction is committed, for example to ensure that `owl:Nothing` has no members or that no pair of individuals have both `owl:sameAs` and `owl:differentFrom` relationships.

During loading, all inferred statements are materialized, except those generated as a result of the semantics of `owl:sameAs`. OWLIM-SE uses special data structures to maintain equivalence classes and uses the URI of the first asserted resource in each equivalence class in the statement indices. This allows for the correct expansion of results during query-answering while keeping the index sizes manageable. This technique has the further advantage that it can be switched off during query answering in order to limit the number of 'duplicate' results.

Except for the inference rules that serve the reference knowledge stack and allow to cover the structural mismatches in the conceptualizations of different FactForge ontologies, which count about 40, the final data collection of RENDER includes new inference rules, specifically crafted to provide consistency checks based on disjoint class definitions, which allow to capture conceptual contradictions, as for example, capturing that it is impossible for one entity to be both of type City and of type Book.

They are of the following form:

```
X owl:sameAs Y
X rdf:type W
Y rdf:type Z
W owl:disjoint Z
-----------------
```

[1] http://www.w3.org/TR/rdf-schema/
[2] http://www.w3.org/TR/owl2-profiles/


```
X ptop:error Y
```

This approach is the beginning of the creation of LOD curation methodology over explicit and implicit statements in an open world reasoning context.

2.5 SPARQL Endpoint of RENDER Basic Data Layer

A new version of FactForge and RENDER basic data layer, were released in 2012, loaded in OWLIM-SE 5.0. RENDER basic data layer (http://render.ontotext.com) counts 1.7 billion explicit statements; 15 billion retrievable statements available after inference and owl:sameAs expansion, which include 1.4 billion inferred statements. Full materialization with respect to OWL-Horst optimized [48] is performed at the time of data loading. The OWL-Horst optimized inference forward chaining rules are part of the distribution of OWLIM-SE. They allow for defining constraints, rules to handle types, transitive relations. The standard OWLIM-SE OWL-Horst optimized rule set has been augmented with additional rules enriching the datasets with instances which cover structural differences in their schemata. The additional rules cover mappings of LOD schemata to the reference layer ontologies of the Reference Knowledge Stack. The datasets combined in FactForge are: DBPedia, Freebase, Geonames, CIA World Factbook, Lingvoj, MusicBrainz, WordNet, New York Times. Only the unification ontology is loaded into the repository. The ontologies of LOD datasets are not, as they would only generate additional redundant statements, and make the inferencing process unnecessary complex.

FactForge provides several methods to explore the combined dataset that exploits some of the advanced features of OWLIM-SE. Firstly, 'RDF Search and Explore' keywords with a real-time auto-suggest feature ordered by 'RDF Rank' (similar to Google's Page Rank). Secondly, a SPARQL page allows users to write their own queries with clickable options to add each of the known namespaces. Lastly, a graphical search facility called 'RelFinder' [38] that discovers paths between selected nodes. This is a computationally intensive activity and the results are displayed and updated dynamically during each iteration. The resulting graph can be reshaped by the user with simple click and drag operations.

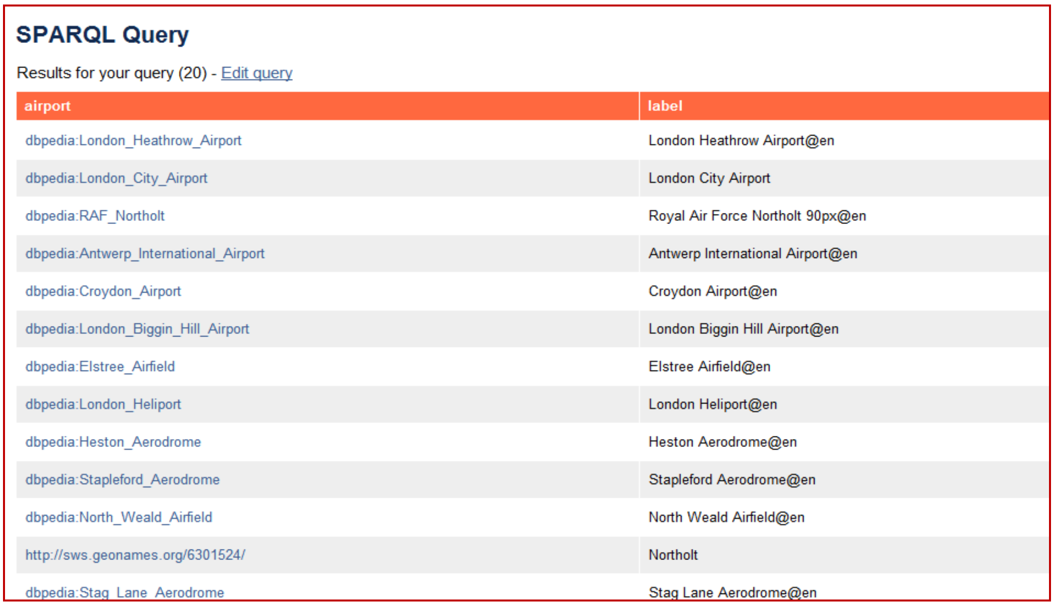
The current version of RENDER basic data layer at http://render.ontotext.com/sparql supports SPARQL 1.1 [44], the latest and most powerful edition of SPARQL language. This allows to make use of all new SPARQL 1.1 and OWLIM 5.0 features when querying. These are the most prominent feature examples:

- GROUP BY, min
  - Minimal and maximal population counts of European countries
- Federated Query between FactForge and LinkedLifeData
  - Drugs that cure the disease from which died Alexandre Graham Bell
- Literal index over dates
  - World governors in office between 1980 and 2005
- Literal index over digits
  - European countries with population above 20 MLN
- Geospatial index
  - Show the distance from London of airports located at most 50 miles away from it

The exemplary queries are provided at the SPARQL end point and can be executed at any time.

The SPARQL end point allows to connect to visualization services, and produced different kinds of visualizations of the query results. Experiments using Google visualization service sgvizler

(http://code.google.com/p/sgvizler/wiki/DesigningQueries), allowed to produce maps of the airports near London, c.f. figures 1-4 below;



SPARQL Query

Results for your query (20) - Edit query

airport	label
dbpedia:London_Heathrow_Airport	London Heathrow Airport@en
dbpedia:London_City_Airport	London City Airport
dbpedia:RAF_Northolt	Royal Air Force Northolt 90px@en
dbpedia:Antwerp_International_Airport	Antwerp International Airport@en
dbpedia:Croydon_Airport	Croydon Airport@en
dbpedia:London_Biggin_Hill_Airport	London Biggin Hill Airport@en
dbpedia:Elstree_Airfield	Elstree Airfield@en
dbpedia:London_Heliport	London Heliport@en
dbpedia:Heston_Aerodrome	Heston Aerodrome@en
dbpedia:Stapleford_Aerodrome	Stapleford Aerodrome@en
dbpedia:North_Weald_Airfield	North Weald Airfield@en
http://sws.geonames.org/6301524/	Northolt
dbpedia:Stag_Lane_Aerodrome	Stag Lane Aerodrome@en

Figure 1 Results for the SPARQL query “Airports near London”

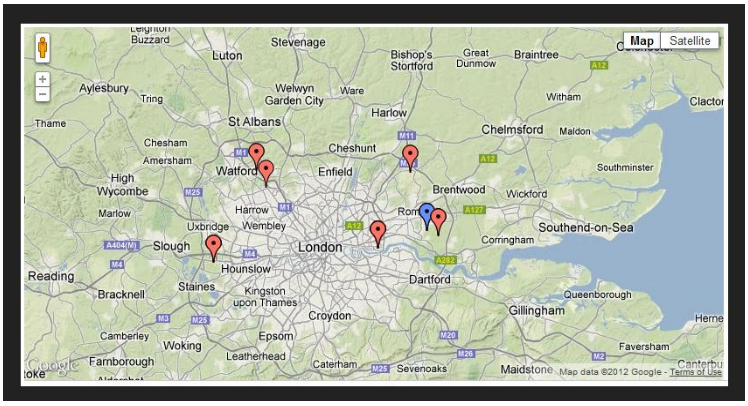


Figure 2 Visualization one of the results from the query “Airports near London”

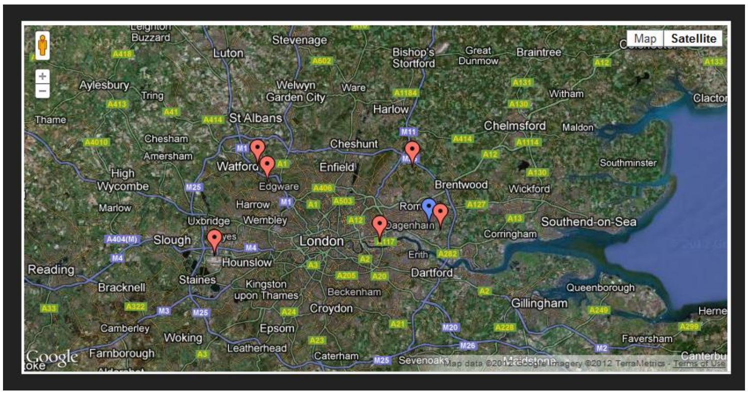


Figure 3 Visualization two of the results from the query “Airports near London”

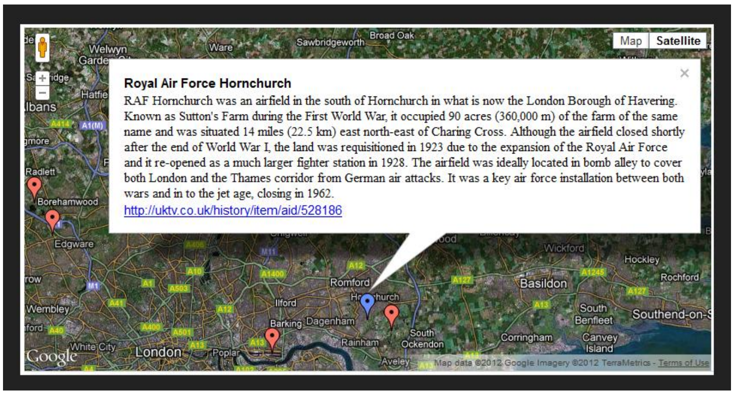


Figure 4 Visualization three of the results from the query “Airports near London”

- Geomaps, for example presenting a map based on GDP per capita in the European countries;
- Pie charts, column charts, bar charts, representing the European countries by population, etc.

3 Diversity through Inferred Knowledge

Using FactForge and RENDER basic data layer gives the opportunity to stumble upon examples of diverse information derived from inference over large interconnected LOD datasets with a reference unification ontology.

The exploration mode of FactForge allows to inspect the inferred knowledge in the query results. Figure 5 shows the results of a search about the Copenhagen, Denmark. The explicit statements are given in blue, whereas the implicit statements are given in red. We observe several things with respect to diversity in this example. Firstly, the transitive closure over the Geonames predicate geo-ont:parentFeature derives facts that Copenhagen is in Denmark and in Europe, expressed both with DBpedia and Geonames instance. Secondly, the PROTON predicate ptop:locatedIn however, has about 10 more implicit facts about the location of Copenhagen, e.g. that it is in Northern Europe, in Scandinavia, in a Nordic country, etc., which increases the variery of information available for querying. Thirdly, Dublin Core predicates give information about Geonames feature codes, which provides the freedom to access geographical information via generic predicates.

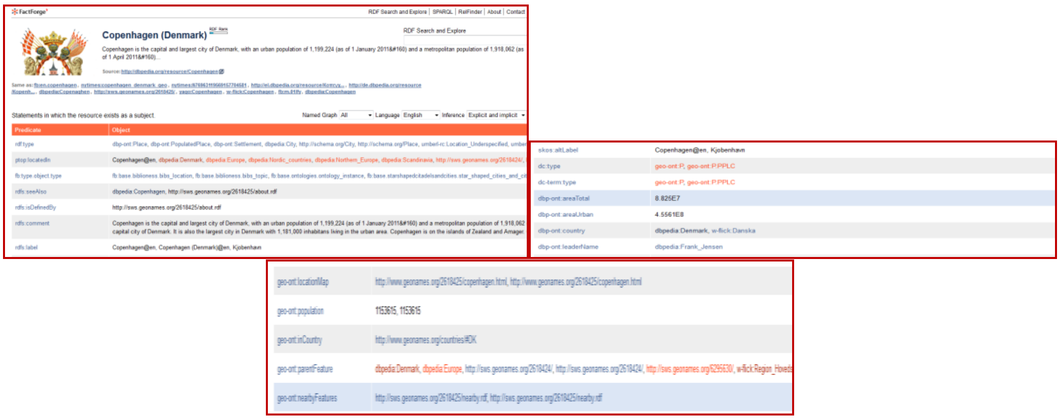


Figure 5 Mode “Exploration” showing the available explicit and implicit knowledge about Copenhagen

Further exploring North Europe, which we pick from the results of Copenhagen above, point us to several inferred facts with relevance to diversity, c.f. Figure 6.

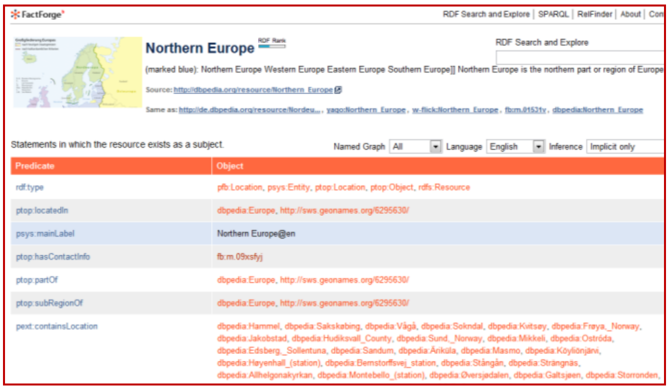
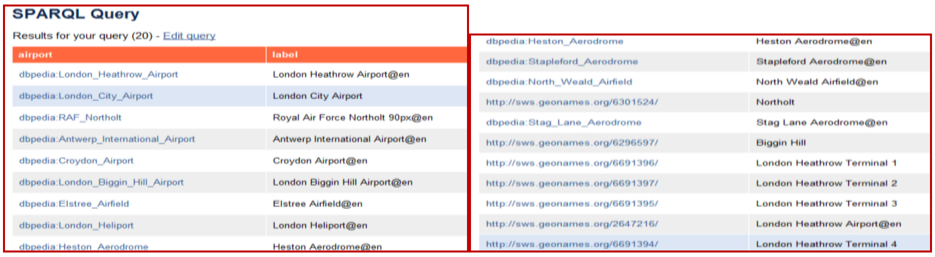


Figure 6 Mode “Exploration” showing the available implicit knowledge about Northern Europe

Two predicates refer to the location of Northern Europe in Europe, e.g. ptop:partOf, and ptop:subRegionOf. We observe an inferred reference to a Website of Northern Europe found in Freebase, and more than 50 locations that are inferred to be in Northern Europe.

PROTON mappings allow to query FactForge with PROTON predicates. This does not prevent however, to use the LOD datasets original predicates to query FactForge. Querying with PROTON though brings more results with additional, sometimes very useful information. For example, a query about airports of 50 miles away from London, brings 13 results when formulated with predicates from LOD, and 20 results when formulated with PROTON predicates. Interestingly, the additional 7 results come from Geonames, and point to airport Terminals, c.f. figure 7.



SPARQL Query

Results for your query (20) - Edit query

airport	label
dbpedia:London_Heathrow_Airport	London Heathrow Airport@en
dbpedia:London_City_Airport	London City Airport
dbpedia:RAF_Northolt	Royal Air Force Northolt 90px@en
dbpedia:Antwerp_International_Airport	Antwerp International Airport@en
dbpedia:Croydon_Airport	Croydon Airport@en
dbpedia:London_Biggin_Hill_Airport	London Biggin Hill Airport@en
dbpedia:Elstree_Airfield	Elstree Airfield@en
dbpedia:London_Heliport	London Heliport@en
dbpedia:Heston_Aerodrome	Heston Aerodrome@en
dbpedia:Heston_Aerodrome	Heston Aerodrome@en
dbpedia:Stapleford_Aerodrome	Stapleford Aerodrome@en
dbpedia:North_Weald_Airfield	North Weald Airfield@en
http://sws.geonames.org/6301524/	Northolt
dbpedia:Stag_Lane_Aerodrome	Stag Lane Aerodrome@en
http://sws.geonames.org/6296597/	Biggin Hill
http://sws.geonames.org/6691396/	London Heathrow Terminal 1
http://sws.geonames.org/6691397/	London Heathrow Terminal 2
http://sws.geonames.org/6691395/	London Heathrow Terminal 3
http://sws.geonames.org/2647216/	London Heathrow Airport@en
http://sws.geonames.org/6691394/	London Heathrow Terminal 4

Figure 7 Results of the SPARQL query "Airports near London" with PROTON predicates only

Finally, inferred knowledge can bring completely unexpected knowledge. For instance, the city of Missouri is inferred to be the subject in the work of the science fiction writer Joel Rosenberg, cf. figure 8.

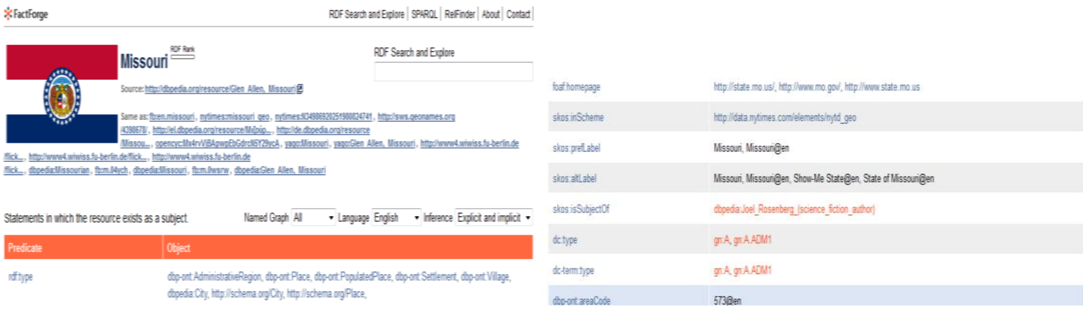


Figure 8 Unexpected inferred knowledge, the city of Missouri as a subject of a science fiction author

These examples present evidence that the inferred knowledge over LOD is a valuable resource of information, which calls for a new paradigm of data services, based on both explicit and implicit knowledge, not only on federating results over linked data physically dispersed over distant servers.

Semantic annotation and metadata enrichment are the core elements of linked data generation over LOD. They focus on facts and entity extraction and on assignment of an URI or respectively a list of related URIs following given rules and algorithms. Sentiment analysis and bias detection are typically disciplines strongly associated with language processing approaches. RENDER project advocates for a method of linking biased texts with LOD, which is achieved by providing RENDER secondary data layer, introduced in the next section. RENDER secondary data layer is implemented using a novel mechanism of RDF data management, called nested repositories, which was described in deliverable 1.3.1. This mechanism allows to share knowledge stored in multiple repositories both explicit and inferred knowledge.

4 Render Secondary Data Layers

The RENDER secondary data layer consist of two datasets, loaded in two repositories, nested into RENDER basic data layer, as shown on figure 9.

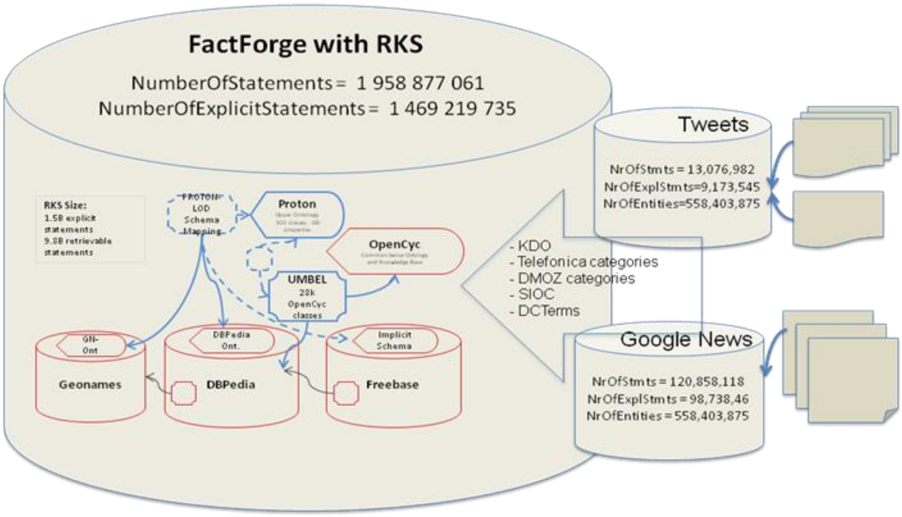


Figure 9 Secondary RENDER data layer, nested into the Basic one

The RENDER secondary data layer includes additional ontologies, e.g.

— **KDO (Knowledge Diversity Ontology)** [26]

KDO is developed within RENDER project and presents RENDERs view about diversity. It models statements, sentiments and opinions.

— **SIOC (Semantically Interlinked Online Communities)** [43]

SIOC is an ontology allowing to represent and integrate information on online communities.

Except for the additional ontologies, RENDER secondary data layer involves two categories of schemata, e.g.

— **DMOZ** [15]

DMOZ is a free dictionary that categorizes segments of the world to be used as media informational units, for instance, business, sports, computers, health and medicine, etc.

— **Telefonica categories** [47]

Telefonica categories are provided by Telefonica, a partner in RENDER project, and present company internal classification of their products and services.

4.1 News

The News dataset provided by JSI is a collection of news articles crawled from the Google News web site in the period of approximately two months. The collection contains about 23500 articles clustered (by Google) into stories 10-150 articles in size with median at 30. The articles are stripped of HTML markup and chrome (navigation, headers, footers etc.), then enriched with named entity detection and disambiguation algorithms and with full constituency parse trees. The total size of the enriched dataset is 1.5 GB (sqlite table, uncompressed). All the articles in this collection are in English, although they are gathered from publishers located all over the world. The average article is 550 words in length.

Google news articles were analyzed according to the following RDF model, developed by Ontotext, c.f. Figure 10 below. It includes diversity information, topic information, reference to an entity from the LOD cloud, document identification.

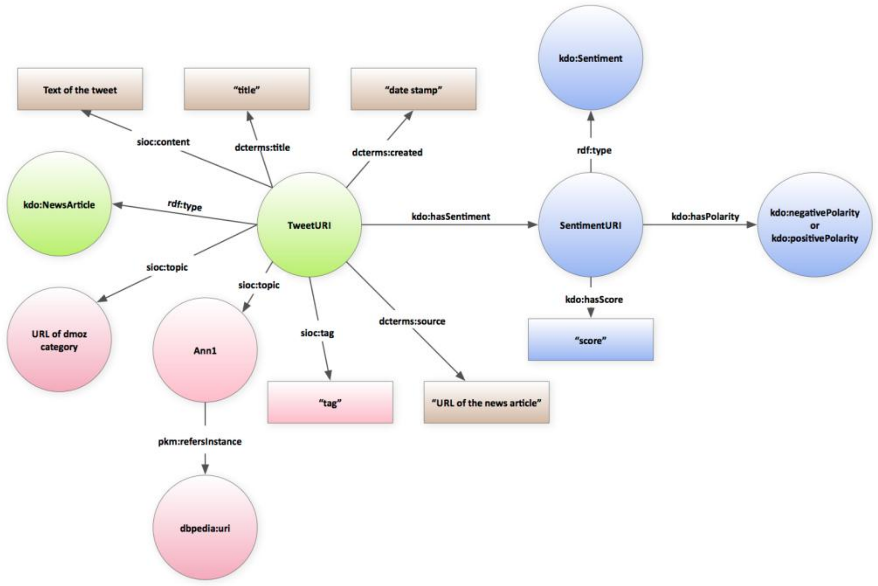


Figure 10 RDF representation for Google news documents

This nested repository is available at http://rendernews.ontotext.com.

One can query about Negative news that mention the United States, Positive news about Steve Jobs, Negative documents about people's occupations per country, Documents with positive bias towards given geographic regions and documents with negative bias towards given geographic regions, Which topics are regarded positively and which topics are regarded negatively by Forbes, Documents with positive or negative bias towards given geographic regions, industries and professions, Documents with positive or negative bias towards given geographic regions, Documents with positive or negative bias towards given professions, Documents with positive or negative bias towards given industries, Industries towards which a given document has a negative bias, Industries towards which a given document has a positive bias, Geographic regions towards which a given document has a negative bias, Geographic regions towards which a given document has a positive bias, Professions towards which a given document has a negative bias, Professions towards which a given document has a positive bias, and the like.

4.2 Tweets

The Twitter dataset collected by Telefonica is processed using the Diversity Mining Services [7] (Enrycher) and its RDF representation is stored in the OWLIM repositories.
The processed data comprises annotated tweets dated as follows:

- September 2010 [9 days]
- October 2010 [6 days]
- January 2011 [8 days]
- Monthly datasets starting April 2012

The size of a monthly unprocessed dataset is on average 5-6GB, smaller datasets having around 2GB. After processing and enriching, the resulting RDF dataset would be around 85GB uncompressed (given an input dataset of 4.5GB). Regarding processing time, approximately 90% is required for enriching the dataset (performing topic detection, entity extraction and resolution, sentiment analysis) and the remaining 10% is required for RDF export.

The RDF model of the tweets, developed by Ontotext, is presented in figure 11 below. It includes diversity information, topic information, reference to an entity from the LOD cloud, author account information, and tweet identification information.

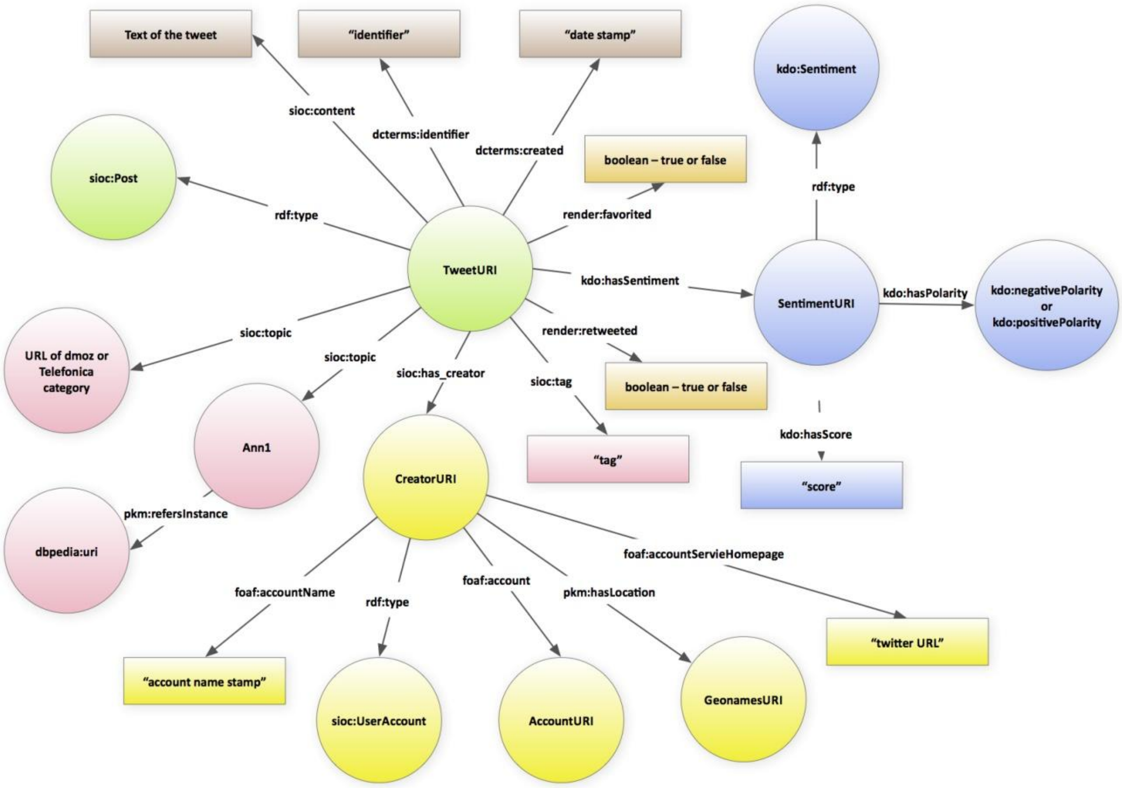


Figure 11 RDF representation for Tweets statuses

This nested repository is available at: http://rendertweets.ontotext.com

One can query about Topics of tweets by account 62855174 with negative bias, and topics of tweets with positive bias, PostivieTweets about Movistar and Politicians born in the United States, Negative tweets from Spain posted before August 2011, Accounts with positive bias and account with negative bias towards Western Europe, and the like.

5 Sentiment Analysis

For the sentiment analysis algorithms developed as part of the Opinion Mining Toolkit [7] JSI created the following datasets:

- Domain-specific sentiment lexicons for the telecommunications domain
- RenderEN. An English dataset containing 134 Twitter posts about a telecommunications provider (48 Positive, 84 Negative)
- RenderES. A Spanish dataset with 891 Twitter posts about a telecommunications provider (388 Positive, 445 Negative, 58 Objective)

In order to create the domain-specific lexicons we adapted the 4-step methodology described in [40] so as to generalize to other languages aside from English and provide sentiment information for words belonging to different parts of speech [6]. We have thus obtained two sentiment dictionaries for the telecommunications domain:

- English: 2000 adjectives, 1700 verbs and 8000 nouns
- Spanish: 650 adjectives, 2000 verbs and 4100 nouns

The RenderEN and RenderES datasets were used to evaluate the sentiment analysis algorithms. These datasets were obtained by manually annotating sample tweets from the Telefonica Twitter data collection, via the Interactive Modeling Tool [45].
The datasets are available at: aidemo.ijs.si/datasets/telcosenti.zip

6 Statistics

The current RENDER data infrastructure has the following size:

	News	Tweets	Base
NumberOfStatements	13,076,982	120,858,118	3,106,606,337
NumberOfExplicitStatements	9,173,545	98,738,468	185,938,3628
NumberOfEntities	558,403,875	558,403,875	558,403,875

7 Conclusion

Semantic annotation and metadata enrichment are the core elements of linked data generation over LOD. They focus on facts and entity extraction and on assignment of an URI or respectively a list of related URIs following given rules and algorithms. Sentiment analysis and bias detection are typically disciplines strongly associated with language processing approaches. RENDER project advocates for a method of linking biased texts with LOD to produce higher level of abstraction over diversity information, and allow generalizations over categories and sentiments. To achieve this a novel mechanism of RDF data management, called nested repositories was developed. It allows to share knowledge stored in multiple repositories both explicit and inferred knowledge. FactForge (http://factforge.net), and RENDER basic data layer (http://render.ontotext.com), reason-able views of the web of data, storing the most popular LOD datasets, and the biggest body of heterogeneous knowledge on which inference has been performed, is used as the central repository, which is nested in repositories storing RDF produced by processing news from Google news cluster and Twitter data with JSI Enrycher service. The bias of the texts is represented in RDF, via knowledge diversity ontology (KDO), and the recognized entities are assigned URIs from LOD datasets. The nested repositories mechanism allows inference across repositories, which results in ability to produce generalizations about statements and attitudes over higher levels of abstraction, for instance over categories of entities, and over sentiments. An implemented version of the described method stores RENDER secondary data layer at http://rendernews.ontotext.com and http://rendertweets.ontotext.com. It handles successfully queries about the attitudes towards CEOs, returning documents about Steve Jobbs and Steve Ballmer, or the countries towards which a certain author has a positive attitude, and the countries towards which the same author has a negative attitude, returning documents about cities and regions of these countries, and the like. It demonstrates the advantages of integrating structured data about sentiment and bias in RDF with general knowledge from LOD in reason-able views, because of the provided ability for a deeper analysis of bias, and representation of diversity. It also demonstrates the power of the concept of nested repositories as a flexible and very convenient innovative mechanism for semantic data management.

This deliverable presented the final data collection for RENDER data infrastructure. It discussed the advantages of using semantic web principles and standards, reason-able views and interlinking of diversity information with factual information allowing for unprecedented depth and richness of results produced as a function of reasoning, heterogeneous data interlinking, reference unification layers, sentiment and opinion modeling in RDF.

RENDER data infrastructure is being successfully used by RENDER use cases Telefonica, Google and Drupal.

The final data collection is the basis of the final data integration of RENDER data infrastructure.

References

1. *baseKB*. http://basekb.com/

2. Bergman, M. K. (2011). *Seeking a Semantic Web Sweet Spot. Blog post.* Retrieved from http://www.mkbergman.com/946/seeking-a-semantic-web-sweet-spot.

3. *Bing Google Yahoo* . (2011). Retrieved from Bing Google Yahoo unite to build the Web of Objects: http://www.bing.com/community/site_blogs/b/search/archive/2011/06/01/bing-google-and-yahoo-unite-to-build-the-web-of-objects.aspx?form=MFEHPG&publ=TWITTER&crea=TEXT_MFEHPG_SM0602_cc0602_TW006_1x.

4. Bishop, B., Kiryakov, A., Ognyanoff, D., Peikov, I., Tashev, Z., & Velkov, R. (2011). OWLIM: A family of scalable semantic repositories. (Hitzler, P., Ed.) *Semantic Web Journal. Special Issue: real-World Applications of OWL*.

5. Bishop, B., Kiryakov, A., Tashev, Z., Damova, M., Simov, K. (2012). OWLIM Reasoning over FactForge**.** In:Proceedings of *OWL Reasoner Evaluation Workshop* (ORE'2012), collocated with IJCAR 2012, Manchester, UK.

6. Bizau, A., Rusu, D., Mladenic. D. 2011. Expressing Opinion Diversity. In Proceedings of the 1st Intl. Workshop on Knowledge Diversity on the Web (DiversiWeb 2011), Hyderabad, India.

7. Caminero, J. (ed). 2012. Initial version of diversity information extensions for Telefónica tools. RENDER Deliverable D5.3.3.

8. *CIA The World Factbook*. (2011). Retrieved 2011, from https://www.cia.gov/library/publications/the-world-factbook/.

9. *Cyc*. (2012). http://www.cyc.com/.

10. Damova, M., Kiryakov, A., Grinberg, M., Bergman, M., Giasson, F., Simov, K.. Creation and Integration of Reference Ontologies for Efficient LOD Management. In: Semi-Automatic Ontology Development: Processes and Resources, IGI Global, USA, Armando Stellato and Maria Teresa Pazienza (Eds.) 2012.

11. Damova, M., Kiryakov, A., Simov, K.., Petrov, Sv. Mapping the Central LOD Ontologies to PROTON Upper-Level Ontology In Proceedings of Ontology Matching workshop, International Semantic Web Conference, Shanghai, China, November 2010.

12. Damova M., Simov K., Tashev Z., Kiryakov A. FactForge: Data Service or Diversity through Inferred Knowledge over LOD In Proceedings of AIMSA’2012 Varna, Bulgaria, September 2012.

13. *Datalog*. (2012) http://en.wikipedia.org/wiki/Datalog.

14. *DBpedia.* (2011). Retrieved from Structured information from Wikipedia: http://DBpedia.org.

15. *DMOZ.* (2012). http://www.deemoz.org/.

16. *DOLCE.* (2012). http://www.loa.istc.cnr.it/DOLCE.html.

17. *Dublin Core*. (2011). Retrieved 2011, from Dublin Core Metadata Element Set: http://dublincore.org/documents/dces/.

18. *FactForge*. (2011). Retrieved from A Reason-able View to the Web of Data: http://factforge.net, http://www.ontotext.com/factforge.

19. *FOAF*. http://www.foaf-project.org/

20. *Freebase*. (2011). Retrieved from http://www.freebase.com/.

21. *Geonames*. (2011). Retrieved 2011, from A geographical database: http://www.geonames.org.

22. Guarino, N., Welty C. (2002). “Evaluating Ontological Decisions with OntoClean.” Communications of the ACM, 45(2): 61-65 (2002).

23. Heim, P; Hellmann, S; Lehmann, J; Lohmann, S; Stegemann, T; (2009) RelFinder: Revealing Relationships in RDF Knowledge Bases. In Semantic Multimedia, volume 5887 of LNCS, pp. 182–187. Springer Berlin/Heidelberg, 2009.

24. Jain, P., Yeh, P. Z., Verma, K., Vasquez, R. G., Damova, M., Hitzler, P., & Sheth, A. P. (2011). Contextual Ontology Alignment of LOD with an Upper Ontology: A Case Study with Proton. In G. Antoniou (Ed.), *Proceedings of 8th Extended Semantic Web Conference.* Heraklion, Crete.

25. Joshi A.K., Jain P., Hitzler P., Yeh P.Z., Verna K., Sheth A.P., Damova M. Alignment-based Querying of Linked Open Data. In: Proceedings of Ontologies, DataBases, and Applications of Semantics (ODBASE) 2012, Rome, Italy, September 2012.

26. *KDO*. (2012) http://lov.okfn.org/dataset/lov/details/vocabulary_kdo.html.

27. Kiryakov, A., & Momtchev, V. (2009, June). Two Reason-able Views to the Web of Linked Data. *Paper presented at the Semantic Technology Conference* . San Jose, USA.

28. Kiryakov, A., Ognyanoff, D., Velkov, R., Tashev, Z., & Peikov, I. (2009). LDSR: Materialized Reason-able View to the Web of Linked Data. In R. H. Patel-Schneider (Ed.), *Proceedings of OWLED 2009* . Chantilly, USA.

29. *LEXVO.* http://www.lexvo.org/

30. *Lingvoj*. http://www.lingvoj.org/

31. *LOD*. http://linkeddata.org/

32. *MusicBrainz, community music metadatabase*. (2011). Retrieved 2011, from http://musicbrainz.org.

33. *New York Times Linked Open Data*. (2011). Retrieved 2011, from http://data.nytimes.com/.

34. *NNDB*. (2012). http://www.nndb.com/.

35. *OWLIM*. (2011). Retrieved from http://www.ontotext.com/owlim.

36. PROTON3.0. (2011). PROTON 3.0 Documentation. Sofia, Bulgaria. http://www.ontotext.com/proton-ontology.

37. Prud'hommeaux, E., Seaborne, A. (2008). *SPARQL Query Language for RDF, W3C Recommendation 15 January 2008.* http://www.w3.org/TR/rdf-sparql-query/.

38. *RelFinder*. http://www.visualdataweb.org/relfinder.php.

39. *RSS*. http://www.rssboard.org/rss-specification.

40. Rusu, D. (ed). 2012. Final version of the opinion mining toolkit. RENDER Deliverable D2.1.2.

41. *schema.org*. (2012). http://www.schema.org.

42. *Simple Knowledge Organization System*. (2011). Retrieved from http://www.w3.org/2004/02/skos/.

43. *SIOC*. (2012). http://sioc-project.org/.

44. *SPARQL 1.1*. (2012). http://www.w3.org/TR/sparql11-query/.

45. Stajner, T., Novalija, I., Mladenic, D. 2012. Informal sentiment analysis in multiple domains for English and Spanish. In Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2012) co-located with the 15th International Multiconference on Information Society.

46. *SUMO*. (2012). http://www.ontologyportal.org/.

47. Telefonica categories at RENDER Wiki

48. Ter Horst, H. J. Combining RDF and Part of OWL with Rules: Semantics, Decidability, Complexity . In Proceedings of The Semantic Web ISWC 2005, LNCS volume 3729 pp. 668–684. Springer Berlin / Heidelberg, 2005.

49. Upper Mapping and Binding Exchange Layer (UMBEL). (2011). Retrieved from http://www.umbel.org/.

50. *Wikipedia*. (2012). http://www.wikipedia.org/.

51. *Wordnet*. (2011). Retrieved from A lexical database for English: http://wordnet.princeton.edu/.

