# JRC-Names: Multilingual Entity Name variants and titles as Linked Data

Maud Ehrmann [a], Guillaume Jacquet [b,*] and Ralf Steinberger [b]

[a] *Swiss Federal Institute of Technology in Lausanne (EPFL)*
*Digital Humanities Laboratory - CDH, INN 116, Station 14, Lausanne, Switzerland*
*maud.ehrmann@epfl.ch*

[b] *European Commission - Joint Research Centre*
*Via Enrico Fermi 2749, 21027 Ispra, Italy*
*{guillaume.jacquet,ralf.steinberger}@jrc.ec.europa.eu*

Abstract

Since 2004 the European Commission's Joint Research Centre (JRC) has been analysing the online version of printed media in over twenty languages and has automatically recognised and compiled large amounts of named entities (persons and organisations) and their many name variants. The collected variants not only include standard spellings in various countries, languages and scripts, but also frequently found spelling mistakes or lesser used name forms, all occurring in real-life text (e.g. *Benjamin/Binyamin/Bibi/Benyamín/Biniamin/Беньямин/*بنيامين *Netanyahu/Netanjahu/Nétanyahou/Netahny/Нетаньяху/*نتنياهو). This entity name variant data, known as *JRC-Names*, has been available for public download since 2011. In this article, we report on our efforts to render JRC-Names as Linked Data (LD), using the lexicon model for ontologies *lemon*. Besides adhering to Semantic Web standards, this new release goes beyond the initial one in that it includes titles found next to the names, as well as date ranges when the titles and the name variants were found. It also establishes links towards existing datasets, such as *DBpedia* and *Talk-Of-Europe*. As multilingual linguistic linked dataset, JRC-Names can help bridge the gap between structured data and natural languages, thus supporting large-scale data integration, e.g. cross-lingual mapping, and web-based content processing, e.g. entity linking. JRC-Names is publicly available through the dataset catalogue of the European Union's *Open Data Portal*.

Keywords: multilingual semantic web, linguistic linked data, *lemon*, named entity, name variants

## 1. Introduction

Enhanced by Semantic Web technologies, the Linked Data publishing paradigm has become increasingly attractive in the recent years [4,30], giving rise to an ever-growing Web of Data[1]. The availability of such machine-readable, formally defined and interlinked data that can be used by computational agents bears the potential of a better and knowledgeable use of information and forms the basis of the Semantic Web vision [5]. Yet, even if the 'global giant graph' is under way, several challenges still need to be addressed before attaining web-scale data integration and full access to knowledge.

A crucial point relates to the natural language interfacing and processing capabilities of the Se-

[*] Corresponding author: guillaume.jacquet@jrc.ec.europa.eu

[1] As evidenced by the Linked Open Data (LOD) cloud: `http://lod-cloud.net/state/`.

1570-0844/0-1900/$27.50 © 0 – IOS Press and the authors. All rights reserved

mantic Web (SW). Indeed, if the Semantic Web is inherently language-independent [28], the question arises of how to mediate between, on the one hand, language-agnostic data representations and, on the other, language-based information needs and content. For that to happen, it is crucial to enrich structured data with linguistic information in several languages, and to enhance the Semantic Web infrastructure with language processing applications [9]. Overcoming the gap between the web of data and natural languages presents challenges and opportunities for both Semantic Web and Natural Language Processing (NLP), which stand here in a mutually beneficial relationship.

As regards the Semantic Web, such developments are key in several respects. Multilingual linguistic information can first support data integration. Given the growing trend towards the publication of non-English data sources and the risk of 'monolingual islands' of data that do not interoperate [27], cross-lingual mappings between datasets are necessary. In this context, the lexicalisation of data on a multilingual basis can be of great help [60]. Linguistic knowledge can also ease data access. Particularly, it can support the development of ontology-based Question-Answering systems in order to allow users to interact with data using their own language(s) [38,62]. Finally, even if data can be interlinked and accessed in several languages, the vast majority of content (i.e. the Web of Documents) remains unstructured. In order to facilitate information discovery and to further develop the scope of structured data, content needs to be marked-up with semantic metadata. This relies again on the availability of web-based linguistic information and technologies.

With respect to Natural Language Processing, adopting linked data principles for the distribution of linguistic resources can bring many benefits, including: resource interoperability, both at a structural and conceptual level; resource integration (via interlinking); and resource maintenance (via a rich ecosystem of technologies allowing, among other things, continuous updating) [14]. Based on such insights, members of the NLP and SW communities – in particular the Open Linguistics Working Group and the W3C Ontology-Lexica Community Group[2] – joined efforts for the definition of best practices [27] and the design of principled models for the representation of linguistic information [47,43]. This laid the foundation for the development of a Linguistic Linked Open Data cloud (LLOD) and provided a real impetus for the publication and the use of linguistic data collections on the Web[3]. Apart from an interoperable set of linguistic resources, NLP can additionally benefit from the plethora of semantic resources and knowledge bases (KB) available on the Semantic Web, e.g., as linked data. Finally, a web-based integration of NLP tools is foreseeable in the medium term. Some steps have already been taken in this direction with the definition of the *NLP Interchange Format* (NIF) [31], and further progress is being achieved through several international initiatives[4].

The task of *Entity Linking* (EL) is particularly representative of the symbiotic relationship between SW and NLP. It illustrates the evolution of information extraction from a document to a semantic-centric viewpoint [50,40] and is at the core of many knowledge extraction tools for the Semantic Web [25,17]. This task requires to align textual mentions of entities with a unique identifier in a knowledge base, typically Wikipedia or DBpedia [37]. Like in traditional named entity recognition, entities of interest are usually of type *person*, *organisation* and *geo-political*, although they can be extended to others. Many EL approaches have been developed [18,10,48,23], all of which acknowledge the lexical gap between KBs and textual content with, especially, the problem of entity surface form variation. Indeed, alternative spellings, abbreviations, aliases or other types of lexical variation make entity mention spotting and/or candidate selection difficult. When provided with extra surface forms, system performances increase, particularly with noisy texts [11] or specific domains [67]. There is thus a need for lexical information regarding entity names, especially across languages.

In this paper we present the release of a multilingual named entity resource for person and organisation names, namely *JRC-Names*, as linked data. The resource is freely available and comprises hun-

[2] `http://linguistics.okfn.org` and `www.w3.org/community/ontolex/`

[3] `http://linguistic-lod.org/`

[4] Such as the LIDER project (`lider-project.eu`) and the BPMLOD and LD4LT W3C community groups (`w3.org/community/{bpmlod|ld4lt}`)

dreds of thousands of entity names and their multilingual variants in over twenty languages, including across scripts. This is a follow-up of a first release [56], from which it differs in that (1) it is rendered as linked data using the **Le**xicon **M**odel for **On**tologies, and (2) it contains much more information, such as titles of persons and date ranges when title and name variants were found. Besides increasing the discoverability and reusability of the resource, the linked data release of *JRC-Names* can help better address the challenges of data integration and multilingual access, as well as support the SW to embrace the web of unstructured documents, e.g. through entity linking.

The remainder of the paper is organised as follows. In section 2 we introduce the JRC-Names resource; we briefly explain how it was produced (2.1), account for the quality of the resource (2.2) and specify what is included in the dataset (2.3). Next, we describe its conversion to linked data (section 3) and present its interconnections with other datasets (section 4). We then give accessibility details (section 5) and summarise known and potential usages (section 6); finally, after the discussion of related work (section 7), we conclude and consider future work (section 8).

## 2. JRC-Names

### 2.1. *Resource creation: Multilingual NER from the news*

*JRC-Names* is a by-product of the *Europe Media Monitor* (EMM) family of news analysis applications, which gathers and analyses up to 220,000 news articles per day fully automatically in about 70 different languages from up to 7,000 news sites (status January 2015; [55]). Once gathered, news texts enter a pipeline of different modules which cluster related news, link news clusters over time and across languages, and – for currently twenty-one languages – recognise direct speech quotations and perform named entity recognition (NER) and classification for the entity types *person* and *organisation.* Location names are also recognised, through a lookup procedure, and disambiguated via document-based heuristics.

NER is performed using a number of manually curated language-independent rules that make use of language-specific lists of titles and other words/phrases that are typically found next to names. As regards person names, these pattern words can be titles (*president*), professions or occupations (*tennis player*, *playboy*), references to countries, regions, ethnic or religious groups (*French*, *Bavarian*, *Berber*, *Muslim*), age expressions (*57-year-old*), verbal phrases (*deceased*) and more. Such phrases, which we generally refer to as *trigger words* because they include far more than only titles, can be further modified (*former*) or occur in combination (*57-year-old former British Prime Minister*). Trigger word lists are produced in a combination of machine learning and manual collection from online sources. Those found historically next to each name are stored in order to build up a frequency-ranked repository of common titles (and more) for each entity. Organisation name recognition is performed in a similar manner, i.e. it makes use of lists of typical organisation name parts (*organisation*, *club*, *international*, *bank*, etc.). However, it is relatively weakly developed in EMM and, due to a coarse entity type categorisation, other entity types are included such as *Belfast Agreement*, *Nobel Prize*, *Red Mosque* or *World War I.* We refer the reader to [54] for further details about the NER system.

Besides NER applied to multilingual news, *JRC-Names* is also the result of a name variant matching process. The NER tool identifies over 500 new name forms per day and, for each of them, the system shall determine whether it refers to a new entity or whether it is a spelling variant of an existing entity name. To this end, a language-independent name matching algorithm is applied, which computes a similarity measure (edit distance) between different name representations. These are obtained after several transformation steps including transliteration, normalisation and vowel removal to create consonant signatures. A newly identified name is merged with an existing one if their overall similarity is above an empirically defined threshold, and kept as separate entity otherwise. More advanced approaches for name similarity across scripts have been explored in [49].

It is important to clarify the concept of language with respect to names and their variants. We avoid talking about certain name variants as *being* in a certain language. Instead, we prefer to consider that a certain name variant *is more frequently found* in texts written in a certain language. The same variant may also be found in other languages, but probably with different distributions. For instance, *Michail Gorbatschow* is the most frequent spelling used in German news when referring to the former Soviet leader *Михаил Горбачев*, while *Mikhaïl Gorbatchev* is more frequent in French and this variant is also found in Portuguese texts. This relative frequency information is useful if the purpose is to generate an easy-to-read text in another language (e.g. during Machine Translation).

Finally, let us consider the question of morphological inflection. As other lexical units, proper names are morphologically inflected in many languages. Inflection mechanisms are numerous and heterogeneous, and they can be very difficult to handle when dealing with many languages. Some of the inflected forms found for the surname of the current US president are *Obamával* (Hungarian), *Obamę* (Polish) and *Obamas* (German). In order to avoid the storage of all inflected forms in the database (inefficient and untidy) while keeping the possibility to capture at least a large part of their occurrences in texts, EMM pre-generates the most common inflections for a subset of known name variants or it uses suffix replacement rules during the NER process. This mechanism allows to recognise a majority of name inflections in text and to return the base form for that name. Hence, morphological inflections of entity names are not meant to be part of *JRC-Names*. However, several of them have erroneously been missed as morphological variants and they have been categorised as variants of known names. This is rather an aesthetic issue because, from a practical point of view, their presence improves the lookup procedure of names in text.

Since 2004, the software has identified about 1.75 million different person and about 10,000 organisation names. In addition to these 'canonical' name forms, it contains about 390,000 additional lexical variants. The database grows by about 700 name forms (new names or variants of known names) per week.

### *2.2. Resource quality*

The JRC's software recognises entities in annotated gold standard NER corpora with an average Precision of 92,13% and a Recall of 50,33% for the nine languages De, En, Es, Hu, It, Nl, Pt, Ro and Tr. Precision is highest for English (96.83%) and lowest for Portuguese (83.41%). Recall is highest for Hungarian (73.89%) and lowest for Turkish (31.70%). The evaluation values of the real-life NER system are actually better than that because of the specific settings of JRC's system, which are geared towards (a) recognising each name at least once in a whole cluster of related news and (b) grounding each name to a real-life entity. When using the standard evaluation settings of the NER system by considering each individual mention of a name, Recall is thus low because the JRC's system ignores names consisting of only one name part (*Obama* alone could refer to either *Barack* or *Michelle Obama*). Furthermore, rather than only recognizing new names (which is the task in standard NER evaluation experiments), JRC's system will additionally look up the hundreds of thousands of known names it has repeatedly found in the past, boosting both recall and precision. Co-referencing name parts (e.g. *Obama*) and common nouns (e.g. *the US President*) with their full names is done independently further down in EMM's processing chain [52].

The result of EMM's automatic NER and variant merging process is subject to a (light-weight) human moderation process. Manual intervention is carried out daily (an average of maximally one hour), focusing on the most frequently mentioned names and on regular mistakes that affect large numbers of entities. The human moderator also has the possibility to mine - assisted by an automatic tool - name variants from cross-lingual Wikipedia links and to download entity images. This semi-automatic Wikipedia mining increases the number of languages for name variants beyond the ones covered by the NER system. Although extremely valuable, the manual verification mends only a small part of the data and *JRC-names* remains the product of an automated process and, as a consequence, contains noise. The main types of errors consist of non-entities (e.g. *Red Piano* or *French Doctor*), wrong name extents (e.g. *Even Obama*) and wrong entity type (e.g. *Merlin Biosciences* as a person). Additionally, it is possible that different entities have been merged into one and, conversely, that homonyms have the same identifier, as no disambiguation mechanism is in place. In order to keep most mistakes out of the *JRC-Names* distribution and also to stick to the more useful entities, only those entities whose frequencies are above a threshold are included in JRC-Names, as we shall see in the next section.

### *2.3. Content of the linked dataset*

A first version of *JRC-Names* has been released in 2011 in the form of a tab-separated text file, accompanied by a Java library for fast lookup. The named entity resource file corresponds to a subset of EMM's database, and it has since been available on JRC's website[5] where a daily update ensures the inclusion also of recent names. This initial version was subject of a coarse-grained transformation to RDF during the MLODE 2012 workshop[6], where participants collaboratively worked on bootstrapping the LLOD. The present linked data version of *JRC-Names* takes a leap forward from there in that it (1) encodes the data using a lexical data model, namely *lemon*, and (2) contains further types of data. The dataset is composed of the following:

(a) Person and organisation entity names. Those entities must have been found in at least five different news clusters (i.e. all mentions in all clustered articles of the same day count only as one)[7].
(b) Name variants. They must satisfy the threshold of having been found in at least 2 different news clusters.
(c) Trigger words. They correspond to titles and function names that have been found in news articles next to the person mentions (cf. section 2.1). Trigger words are included if they were found in at least five different news clusters.
(d) Time stamps. Each name variant or title is accompanied by two time stamps: the first insertion date into the database (when EMM first found this title), and the last update date. This information is useful to detect changing titles, e.g. when a person is mentioned with different positions.
(e) Frequency information. Each name variant has a news cluster frequency count.
(f) Prior probabilities. Name variants have monolingual and multilingual prior probabilities, which reflect how likely an entity is mentioned with a specific variant in a certain language, or across all languages[8].

For multilingual name variants harvested from Wikipedia, there is neither frequency nor time stamp information.

## 3. Multilingual entity names as Linked Data

The resource consists of lexical knowledge, *i.e.* name variants in multiple languages, about individuals, *i.e.* person and organisation entities. *Lemon* and other linguistic vocabularies (section 3.2), were used to render *JRC-Names* as linked data (sections 3.3 and 3.4).

### *3.1. The* lemon *model*

*lemon* is a model to represent linguistic information relative to ontologies in RDF. More specifically, it allows to specify the meaning of lexical units as well as to describe their constructions with respect to the vocabulary of an ontology. In line with the principle of *semantics by reference* [8,45], *lemon* maintains a clean separation between the lexical layer, which deals with the morphological and syntactic description of lexical entries (words or phrases), and the ontological layer, responsible for describing the meaning (or resolving the reference) of the lexical entries. The model builds on previous work for representing lexica and combines the strengths of LexInfo [15] and of the Linguistic Information Repository [42], both based on the Lexical Mark-up Framework [24]. The core of the *lemon* model[9] consists of the following elements:

– *Lexicon*, which collects lexical entries and is marked with a language,
– *Lexical entry*, which comprises all syntactic forms of an entry,

[5] `https://ec.europa.eu/jrc/en/language-technologies/jrc-names`

[6] Multilingual Linked Open Data for Enterprises: `http://sabre2012.infai.org/mlode`

[7] As mentioned in section 2.1, EMM groups related news articles into 'news clusters' and deals with each cluster as a meta-document. Frequency counts of named entities are relative to these clusters, and not to single news articles.

[8] The prior probability of a specific variant of an entity is calculated by dividing the frequency count of this variant by the sum of the frequency counts of other variants of the same entity in the same or across languages.

[9] `http://lemon-model.net/lemon#`

- *Lexical form*, which represents the surface realization of a lexical entry, usually in the form of a *written representation*,
- *Lexical sense*, which represents the usage of a lexical entry as a *reference* to an ontological entity.

The *lexical sense* acts, among others, as a 'glue'[10] between a lexical entry and an ontological entity and, as such, corresponds to the reification of the meaning of an entry [16]. *lemon* is linguistically agnostic and allows to use any vocabulary of linguistic categories. The model has already been used to represent various existing lexica [46,21,66,63,33,22,36] and proposals have been made for its extension [29,35,12]. Meeting the challenge of representing lexica and connecting them to ontologies is the current focus of the W3C OntoLex Community Group[11], which is actively working towards *lemon*'s final specification.

### *3.2. Other vocabularies*

Apart from *lemon*, which enables the representation of most *JRC-Names* data, other controlled vocabularies are used: LexInfo2 and OLiA, which provide linguistic categories and mapping between linguistic schemes, are used to specify linguistic categories and relation properties of name variants [15,13]; lexvo, which provides global IDs for language-related objects, is used to encode language information [19]; and the DBpedia ontology, which organises Wikipedia concepts, is used to encode entity types. As regards meta-data information, the VoID [1] and the DCTerms vocabularies are used. Finally, when no existing vocabulary could answer our needs, we defined our classes and properties in a dedicated vocabulary[12].

### *3.3. Representing entities and their multilingual name variants*

At the ontological level, JRC-Names entities are encoded as *dbo:Person* or *dbo:Organisation.* Each entity has a language-independent 'base name', i.e. the variant that was chosen to use for display purposes inside EMM. The choice was made according to the name being either the most frequently found variant in the news (across languages), or the variant found on Wikipedia, or a frequent Latin script version of a name originally written in another script. This base name is therefore not marked with a language (although it is typically a name form that is frequently found in English text) and is encoded as the *skos:prefLabel* of the RDF entity.

At the lexical level, entity name variants are encoded as *lemon:LexicalEntry*, the language of which is specified through *lemon* and lexvo language properties (ISO-639-1 and 3). These lexical entries are also defined as *olia:NamedEntity* and get further characterised with the lexinfo *properNoun* part-of-speech[13].

*JRC-Names* exhibits a relatively high degree of lexical variation. There are multiple scripts (e.g. Latin vs. Cyrillic *Barack Obama - Барка Обаму*), omission or addition of name parts (*Barack Hussein Obama Jr.*), inflected forms (*Barack Obamát*), typos (*Barrac Obama*), inversion of name parts (*Obama Barack*) and various other forms (e.g. *Barack O'Bama*). Because the collection of variants is based on string similarity, formally very different units such as diachronic variants or aliases (*Eric Blair*, alias *George Orwell*) do not exist in the resource (or if so, they were manually entered). Variant types, however, are not specified in *JRC-Names.* As a consequence, even if *lemon* offers the possibility to represent term variation at the level of surface form, word or sense [43,44], name variants are all *lemon:LexicalEntry* (i.e. words), although some could be conceived as different *lemon:Forms* of a variant. Accordingly, name variants of the same language (and of the same entity) are related through *lemon:lexicalVariant* relations.

The path from name variants to their referent is set via *lemon:LexicalSense.* As reification of the relation between a word and a concept (here an entity), a lexical sense can support the expression of information which is neither of lexical nor of ontological nature. *JRC-Names* associates contextual information to entity name variants, that is to say

[10] the expression is from [16].

[11] `http://www.w3.org/community/ontolex`

[12] URLs of all vocabularies are mentioned in Figure 1.

[13] Following the strict point of view that JRC name variants are nominal phrases composed of *proper nouns*, a more appropriate representation could be as *proper names.* Considering that the distinction proper noun/name is not universally applied and that the resource does not provide information about name composition (first-middle-last name, particle), we choose the *lexinfo:properNoun* property.

their news cluster frequency and the dates of their first insertion and last update in the database. Based on news cluster frequencies, we additionally compute monolingual and multilingual prior probabilities. This information is rendered as properties of name variant lexical senses. Such properties are circumstantial and do not qualify the linguistic usage but the incidence of the association of a given variant with a specific entity (how many times this name appears with this referent, when was the first and last time of this occurrence). This is the reason why we did not use the *lemon:context* property, which concentrates on pragmatics or discourse properties such as register or temporal and geographical usage constraints. With regards to proper names, such a context could for example specify the time span usage of *Byzantium* vs. *Constantinople* vs. *Istanbul*, or the register difference between *Michael Schumacher* and *Schumy.*

Lexical senses additionally allow the expression of translation relations between name variants in different languages referring to the same entity. Translation relations fall indeed within the domain of lexical sense, as they shall be stated between disambiguated names (the English lexical entry *London* will translate into the French *Londres* when referring to the city, into *London* when referring to the writer). These relations are represented through *lexinfo:translation* object properties, as there was no need to use a more principled way to do it [29].

### *3.4. Representing titles*

Besides name variants in multiple languages, the dataset also contains person entity ‘titles’. As detailed in section 2.1, titles correspond to the trigger words that helped recognise entities in texts and they consist of a heterogeneous set of nominal phrases referring to the function or the social status of a person. Titles are lexically defined as lexical entries and as *olia:TitleNoun*, a morphosyntactic category describing appropriately those items. They are marked with language, but their part-of-speech remain unspecified. Title lexical units refer through lexical senses to the *dbo:PersonFunction* class, in a kind of loose lexicalisation of this abstract concept.

Similarly as for name variants, frequency and time-stamp information are available. However, since these elements regard the relation between a title and a person entity and not the one between a title and its concept (*dbo:PersonFunction*), they cannot be stated on titles’ lexical senses. In other words, what is qualified here is not the linguistic relation between a word and its concept, but the factual one of a person entity having, or occurring with, its title(s). In order to correctly encode this information as well as to capture the person/title relation, we introduced a *jrc-model:Occurrence* class. It represents a specific occurrence of a title lexical sense and establishes the relation with a person entity via the *jrc-model:hasTitle* property. As expected, instances of *jrc-model:Occurrence* additionnally holds the frequency and time properties relative to a given person/title association.

Let us mention that in a more rigorous setting the occurrence of a title lexical sense (an instantiation of *jrc-model:Occurrence*) should point not to the person entity (*dbo:Person*) but to one of its name variants with which the title originally occurred. This information is however not available in the original database, where title expressions are directly associated with person entities.

A graphical representation of JRC-Names entity and lexical knowledge is given in Figure 1, with the example of the current President of the European Commission *Jean-Claude Juncker.* As it is not possible to represent all information, only a few items of each type of information are depicted.

## 4. Interlinking

JRC-Names introduces links towards two specialised datasets, *New York Times* and *Talk of Europe*, and a generic one, *DBpedia* [37]. The New York Times (NYT) initiated some years ago the linked data publication[14] of its news index, or subject headings, which includes data about people and organisations (among others). As of Talk of Europe, this project curates Linked Open Data about the European Parliament; the published dataset contains all plenary debates over a fifteen-year period (1999-2014), and biographical information about the members of parliament (MEP) [64]. Interlinks of type *owl:sameAs* are set from JRC persons towards person entities of both datasets, based on a label strict matching of non-

[14] `http://data.nytimes.com/`

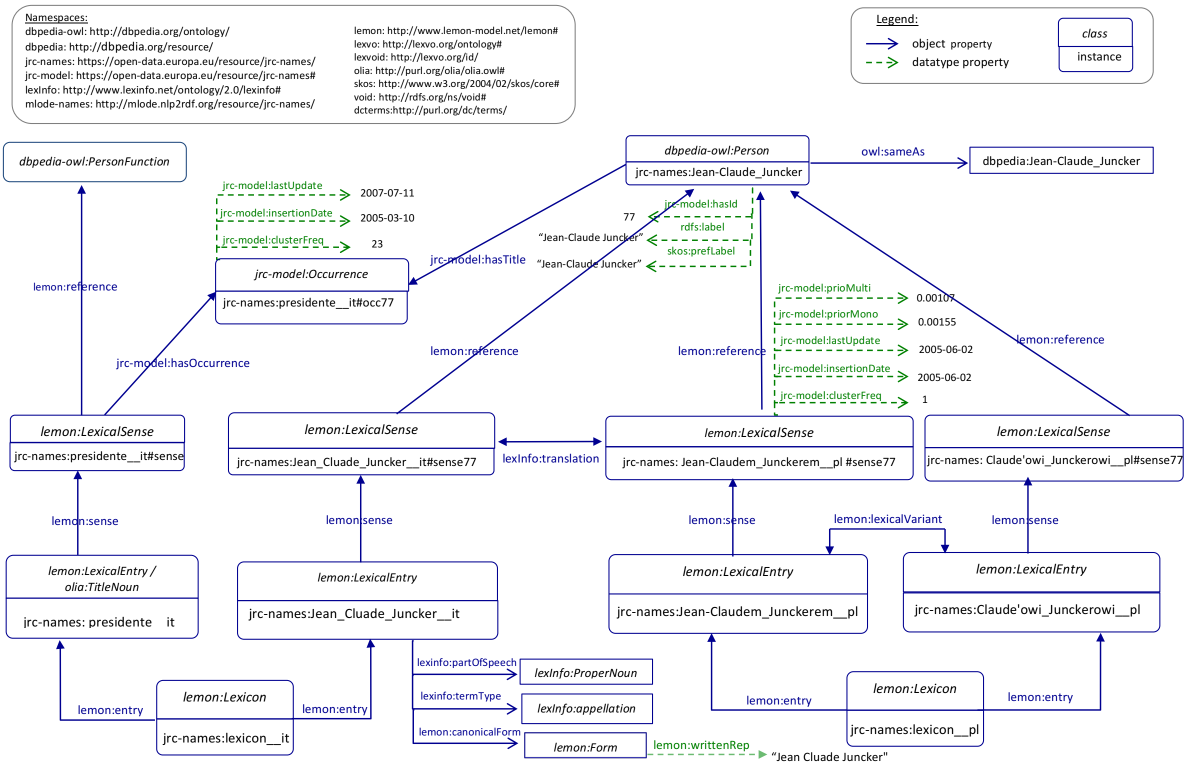

Figure 1. Graphical illustration of JRC-Names data representation, with the example of the entity *Jean-Claude Juncker*.

ambiguous entities. As indicated in Table 1, 2701 links are established towards NYT, 928 towards MEP.

DBpedia contains a great number of person entities with many properties in various languages. As briefly mentioned in the introduction, a well-known issue with knowledge bases is entity disambiguation. Although this was not the primary goal of the present work, we developed a light-weight strategy in order to link JRC-Names entities with their correct counterpart in DBpedia. Given a JRC source entity and its variants in all languages, the algorithm first looks for an exact match between the variants and the English *rdfs:label* of non-ambiguous person and organisation DBpedia entities. Next, if no match is found, ambiguous DBpedia candidates are selected (based on the variant surface forms) and if only one of these candidates is of the same type as the JRC source entity one, then the resources are interlinked. Finally, when there is more than one possible candidate (*i.e.* DBpedia entities having the same type and label than the JRC one), the set of English titles of the JRC entity is considered against a selection of English properties of DBpedia candidates (*dbo:office*, *purl:description* and *db-prop:title*), looking again for an exact match. Overall 95,437 links were created (cf. Table 1), 64,002 thanks to the first alternative, 31,340 thanks to the second and 95 to the third. We manually evaluated the correctness of 100 randomly selected links and obtained a Precision of 91%. Errors are mainly due to EMM mixing different persons, resulting into ambiguous entities difficult to link. The linking strategy could be improved in several ways, e.g. by exploiting multilingual features and making a joint use of the different DBpedia chapters.

Some interlinks are set at vocabulary level [34]. JRC's classes and properties being quite specific, only a few links could be set, mostly on NYT's vocabulary, with loose relationships (*rdfs:seeAlso*) from *jrc-model:clusterFreq*, *jrc-model:insertionDate*

| *Data* | |
|---|---|
| # Lexicons (total) | 170 |
| # Lexicons (with freq. metadata) | 21 |
| # Lexical Entries | 1,781,901 |
| # Lexical Senses | 1,781,901 |
| # Person entities | 331,242 |
| # Organisation entities | 7,391 |
| *Internal connectivity* | |
| # Lexical variants | 2,412,394 |
| # Translation relations | 32,564,928 |
| *External connectivity* | |
| # Talk of Europe | 928 |
| # New York Times | 2,706 |
| # DBpedia | 95,437 |
| Grand Total | 72,586,712 |

Table 1
Statistical profile of JRC-Names RDF dataset.

and *jrc-model:lastUpdate* towards New York Times *associated_article_count*, *first_use* and *latest_use* properties respectively. Finally, let us mention that backward links towards the MLODE dataset are set, based on JRC entity IDs.

## 5. Dataset features and Web access

The RDF version of JRC-Names features an overall number of 72,5 million triples. Table 1 gives further details on the statistical profile of the dataset. The majority of entities are persons, with 331,242 resources of this type against 7,391 of type *organisation.* Those entities are lexicalised through 1.7 million lexical entries, gathered into a total of 171 language-specific lexicons. It is worthwhile here recalling that NER is performed for 21 languages, and that data for other languages is added through Wikipedia mining. Next, there are about 2.4 million monolingual lexical variant relations, and 32 million translation relations. Finally, external connectivity is reasonably good, with a third of the entities being connected to either DBpedia, New York Times, or Talk of Europe.

Resource metadata are expressed using the VOID vocabulary; provided descriptions include general, access and structural metadata. Usage conditions are specified through the *dc-terms:license* property.

The JRC-Names linked dataset is served on the web via the EU *Open Data Portal* with: an RDF dump file[15], a public SPARQL endpoint[16] and dereferenceable URIs[17]. Occasional updates of the LOD version of JRC-Names are foreseen to maintain appropriate synchronisation with the database.

## 6. Known and Potential uses

JRC-Names has been used for a whole range of tasks. The major usage probably is the improvement of the recall of searches in databases (including audio-visual) and text collections (including the internet) [57,2] by expanding the initial user query by all name variants. Alternatively, name mentions in the search space can be normalised by replacing variants with a standard form. Search expansion is particularly important across scripts as even approximate matching techniques will not find foreign script variants of the searched name. Hands-on users of JRC-Names have either replaced the whole entity name by the set of its variants ‘George Bush’ (‘George Busch’, ‘George Buhs’, ‘Corc Uolker Buş’), or they have split all entities in JRC-Names to produce lists of variants for each name part, e.g. ‘Georgius’, ‘Georges’, ‘Georg’, ‘Džordž’, etc. for the English standard spelling of ‘George’. By doing this, the knowledge contained in the data collection can be applied to any names and not only to media VIPs. Another usage of JRC-Names relates to Machine Translation systems, which typically have problems translating proper names [6]. This challenge can be overcome by identifying and removing names before the translation process and by then reinserting the target language equivalent [61]. Also, lists of names in two different scripts are often used to learn transliteration rules, e.g. [49]. Collections of names and their variants have been used to train and/or improve Named Entity Recognition tools [7,20,65] or to disambiguate name mentions [2], but also,

[15] `https://open-data.europa.eu/en/data/dataset/jrc-names`
[16] `http://open-data.europa.eu/en/linked-data`
[17] `https://open-data.europa.eu/resource/jrc-names/`

more generally, to develop Language Technology tools for lesser-resourced languages [69,58]. The development of higher-level Language Technology tools has benefited from JRC-Names, such as coreference resolution [52] and cross-lingual linking of related documents in different languages [53]. Furthermore, JRC-Names has been used in higher-level sociological or political studies such as tracking researchers' mobility on the web [26] or preprocessing text for a subsequent political science study [3]. In principle, JRC-Names can also be useful as a component in Language Technology tools for opinion mining, summarisation, topic detection and tracking, and more.

The LOD version of JRC-Names contains more information and links to other LOD resources. This not only widens the application areas, but most of all it opens the way to a fully-automatic usage of the data. First, the machine-readable version of JRC-Names can be queried by agents[18] and the retrieved information can easily be integrated into NLP web services. Second, due to the list of spelling variants for each name, the LD resource allows establishing richer links between unstructured natural language texts and structured information (for e.g. entity linking), what is more at a multilingual level. Furthermore, the LD resource can support cross-lingual access to information with e.g. the automatic retrieval of entity information spread over several monolingual resources, as well as cross-lingual mapping between datasets, including accross scripts. Finally, interlinks towards other resources connect JRC-Names to the web of data, enabling further data enhancement at both content and linguistic levels: while interlinks towards *New York Times* and *Talk of Europe* datasets can support political studies with questions such as "How and when members of parliament or politicians where mentioned in news articles", links towards the DBpedia nucleus provide additional lexicalizations of DBpedia person entities and have the potential to facilitate integration with other named entity resources.

## 7. Related work

This section summarises previous efforts to compile multilingual lexical information about names, and considers named entity-related data on the LLOD.

Named entities, or proper names when limited to the core categories of *person*, *location* and *organisation*, represent an open word class which evolves endlessly. Dedicated resources or gazetteers are therefore not easy to acquire and require constant updates. In this context, the collaboratively built, semi-structured and multilingual Wikipedia resource appeared as a great relief, and several named entity dictionaries were built out of it [68,57,59]. Prolexbase [39], a manually produced multilingual ontology of proper names built up over many years, recently adopted a semi-automatic enrichment strategy based on Wikipedia [51]. All of these resources are the result of exploiting Wikipedia and, with the exception of [59] which makes use of LMF, they are not interoperable.

Many linguistic resources have been exposed as linked data recently. As for entities, they appear mainly in encyclopaedic dictionaries and knowledge bases, such as BabelNet [22], DBpedia [37] and YAGO [32], but some are present in lexical resources. In the latter case, resources such as WordNet RDF [33] or *lemon*UBY [21] do include entity names, but in a rather limited number and with little information about lexical variation. In the former, all entities derive from Wikipedia and are primarily the focus of encyclopaedic descriptions. At lexical level, Wikipedia is strong at providing cross-lingual and cross-script variants, but it contains only few spelling variants within the same language and it does not contain information on morphological variants. In contrast, *JRC-Names* is mostly built up by recognising name variants in real-life multilingual text. A dedicated resource has been compiled as part of DBpedia Spotlight [41], which consists of entity lexicalisations collected over the graph of labels, redirects and disambiguations of the KB. Anew, the range of name variants is bounded to Wikipedia data, while JRC-Names provides name occurrences of real-life texts. Overall, the picture that emerges is one of complementarity, where various datasets could provide different types of information about entities.

[18] Examples of queries are available at: `http://open-data.europa.eu/en/linked-data`.

## 8. Conclusion

We have presented the new release of the JRC-Names resource as linked data using *lemon*, a model for representing ontology lexica. This work is the continuation of previous efforts and is in line with the general effort of the European Commission to support multilingualism and language diversity. Compared with the initial release of JRC-Names in 2011, the current one is available as linked data and provides more information, namely person titles, occurrence time-stamps and frequency information. With name variants extracted from multilingual news, this resource complements those based on Wikipedia and contributes to the ongoing developments within the SW and NLP communities to support data access in several languages.

Future work could be manifold. At data level, it would be useful to further specify the variant types, to carry out a *lemon*-based publication of morphological generation rules, and to clean erroneously conflated entities (e.g. using titles). At web level, interlinking with other datasets (lexical, encyclopaedic or factual) could be expanded, as well as intralinking among titles.

### Acknowledgments

The *Europe Media Monitor* EMM is a multi-annual group effort involving many tasks and we would thus like to thank all past and present EMM team members for their help and dedication. We are also grateful to Valentina Fratto and her team from the EU Publication Office for the fruitful collaboration. Finally, we would also like to thank the members of the LIDER project, whose activities emphasise the need for and greatly support the development of linguistic linked data.

### References

[1] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. *Describing Linked Datasets with the VoID Vocabulary*, 2011. URL `http://www.w3.org/TR/void/`. W3C Interest Group Note (work in progress).

[2] C. Aliprandi, B. Tomas, and P. Sérgio. Language Processing and Linguistic Data in the CAPER Project. *Language Resources for Public Security Applications*, page 23, 2012.

[3] N. Andreas, W. Gregor, and H. Gerhard. Leipzig Corpus Miner-A Text Mining Infrastructure for Qualitative Data Analysis. In *Terminology and Knowledge Engineering 2014*, pages 10–p, Berlin, Germany, 2014.

[4] S. Auer, J. Lehmann, and A. N. Ngomo. Introduction to Linked Data and its Lifecycle on the Web. In *Reasoning Web. Semantic Technologies for the Web of Data*, pages 1–75. Springer, 2011.

[5] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.

[6] B. Bogdan and A. Hartley. Improving Machine Translation quality with Automatic Named Entity Recognition. In *Proceedings of the* $7^{th}$ *International EAMT workshop on MT and other Language Technology Tools*, Budapest, Hungary, 2003.

[7] S. Buchholz and A. van den Bosch. Integrating seed names and ngrams for a Name Entity list classifier. In *Proceedings of the* $2^{nd}$ *International Conference on Language Resources and Evaluation*, Athens, Greece, 2000.

[8] P. Buitelaar. Ontology-based Semantic Lexicons: Mapping between Terms and Object Descriptions. *Ontology and the Lexicon*, pages 212–223, 2010.

[9] P. Buitelaar, Key-Sun K.S. Choi, P. Cimiano, and E. Hovy. The Multilingual Semantic Web. Technical Report 12362, Report from the Dagstuhl Seminar, 2012.

[10] E. Charton, M. Gagnon, and B. Ozell. Automatic Semantic Web Annotation of Named Entities. In *Advances in Artificial Intelligence*, pages 74–85. Springer, 2011.

[11] E. Charton, M.J. Meurs, L. Jean-Louis, and M. Gagnon. Improving Entity Linking using Surface Form Refinement. In *Proceedings of the* $9^{th}$ *International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, 2014.

[12] C. Chavula and C. M. Keet. Is *lemon* Sufficient for Building Multilingual Ontologies for Bantu Languages? In *Proceedings of the* $11^{th}$ *OWL: Experiences and Directions Workshop*, volume 1265, pages 61–72. CEUR-WS, Oct 2014.

[13] C. Chiarcos. Ontologies of Linguistic Annotation: Survey and Perspectives. In *Proceedings of the* $8^{th}$ *International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 2012.

[14] C. Chiarcos, J. P. M$^{c}$Crae, P. Cimiano, and C. Fellbaum. Towards open data for linguistics: Linguistic linked data. In A. Oltramari, P. Vossen, L. Qin, and E. Hovy, editors, *New Trends of Research in Ontologies and Lexical Resources*, pages 7–25. Springer, 2013.

[15] P. Cimiano, P. Buitelaar, J. P. M$^{c}$Crae, and M. Sintek. LexInfo: A Declarative Model for the Lexicon-Ontology Interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51, March 2011 2011.

[16] P. Cimiano, J.P M$^{c}$Crae, P. Buitelaar, and E. Montiel-Ponsoda. On the Role of Senses in the Ontology-Lexicon. In *New trends of research in ontologies and Lexical resources*, pages 43–62. Springer Berlin Heidel-

berg, 2013.

[17] M. Cornolti, P. Ferragina, and M. Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the* $22^{nd}$ *International conference on World Wide Web*, pages 249–260. International World Wide Web Conferences Steering Committee, 2013.

[18] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the* $9^{th}$ *International Conference on Semantic Systems*, pages 121–124. ACM, 2013.

[19] G. de Melo. Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud. *Semantic Web Journal*, 7:1–5, 2013.

[20] L. Derczynski, D. Maynard, N. Aswani, and K. Bontcheva. Microblog-genre noise and impact on Semantic Annotation Accuracy. In *Proceedings of the* $24^{th}$ *ACM Conference on Hypertext and Social Media*, pages 21–30. ACM, 2013.

[21] J. Eckle-Kohler, J. P. M$^c$Crae, and C. Chiarcos. *lemon*Uby-a large, interlinked, syntactically-rich resource for ontologies. *Semantic Web Journal, Special issue on Multilingual Linked Open Data*, 2014.

[22] M. Ehrmann, F. Cecconi, D. Vannella, J. P. M$^c$Crae, P. Cimiano, and R. Navigli. Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In *Proceedings of the* $9^{th}$ *International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, May 2014.

[23] Peter Exner and Pierre Nugues. Entity extraction: From Unstructured Text to DBpedia RDF Triples. In *The Web of Linked Entities Workshop (WoLE 2012)*, 2012.

[24] G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, C. Soria, et al. Lexical Markup Framework (LMF). In *International Conference on Language Resources and Evaluation*, pages 233–236, 2006.

[25] A. Gangemi. A comparison of knowledge extraction tools for the Semantic Web. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, and S. Rudolph, editors, *The Semantic Web: Semantics and Big Data*. Springer, 2013.

[26] F. Jorge J García, P. Zweigenbaum, Z. Yue, and W. Turner. Tracking Researcher Mobility on the Web Using Snippet Semantic Analysis. In *Advances in Natural Language Processing*, pages 180–191. Springer, 2012.

[27] A. Gómez-Pérez, D. Vila-Suero, E. Montiel-Ponsoda, J. Gracia, and G. Aguado de Cea. Guidelines for Multilingual Linked Data. In *Proceedings of the* $3^{rd}$ *International Conference on Web Intelligence, Mining and Semantics*. ACM, 2013.

[28] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Goméz-Pérez, P. Buitelaar, and J. P. M$^c$Crae. Challenges for the Multilingual Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11:63–71, 2011.

[29] Jorge Gracia, Elena Montiel-Ponsoda, Daniel Vila-Suero, and Guadalupe Aguado-De-Cea. Enabling Language Resources to Expose Translations as Linked Data on the Web. In *Proceedings of the* $9^{th}$ *International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, may 2014.

[30] T. Heath and C. Bizer. Linked data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1): 1–136, 2011.

[31] S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. Integrating NLP using Linked Data. In *Proceedings of the* $12^{th}$ *International Semantic Web Conference*, pages 97–112, 2013.

[32] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61, 2013.

[33] P. Cimiano J. P. M$^c$Crae, C. Fellbaum. Publishing and Linking WordNet using lemon and RDF. In *Proceedings of the* $3^{rd}$ *Workshop on Linked Data in Linguistics*, 2014.

[34] K. Janowicz, P. Hitzler, B. Adams, D. Kolas, and C. Vardeman II. Five Stars of Linked Data Vocabulary Use. *Semantic Web*, 5(3):173–176, 2014.

[35] F. Khan, F. Frontini, R. Del Gratta, M. Monachini, and V. Quochi. Generative Lexicon Theory and Linguistic Linked Open Data. In *Proceedings of the* $6^{th}$ *International Conference on Generative Approaches to the Lexicon*, pages 62–69, 2013.

[36] Lars L. Borin, D. Dannélls, M. Forsberg, and J. P. M$^c$Crae. Representing Swedish Lexical Resources in RDF with lemon. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the* $13^{th}$ *International Semantic Web Conference*, 2014.

[37] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 2013.

[38] V. Lopez, V. Uren, M. Sabou, and E. Motta. Is Question Answering fit for the Semantic Web?: A survey. *Semantic Web*, 2(2):125–155, 2011.

[39] D. Maurel. Prolexbase: a multilingual relational lexical database of proper names. In *Proceedings of the* $6^{th}$ *International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008.

[40] P. McNamee and T. H. Dang. Overview of the TAC 2009 Knowledge Base Population Track. In *Text Analysis Conference (TAC)*, volume 17, pages 111–113, 2009.

[41] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the* $7^{th}$ *International Conference on Semantic Systems*, pages 1–8. ACM, 2011.

[42] E. Montiel-Ponsoda, G. Aguado de Cea, and A. Gómez Pérez. Modelling Multilinguality in Ontologies. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 67–70, 2008.

[43] E. Montiel-Ponsoda, J. Gracia del Río, G. Aguado de Cea, and A. Gómez-Pérez. Representing Translations on the Semantic Web. In *Proceedings of the* $2^{nd}$ *International Workshop on the Multilingual Semantic Web*, pages 30–42, 2011.

[44] E. Montiel-Ponsoda, J. P. M$^c$Crae, G. Aguado de Cea, and J. Garcia. Multilingual Variation in the context of Linked Data. In *Proceedings of the* $10^{th}$ *International Conference on Terminology and Artificial Intelligence*, pages 19–26, 2013.

[45] J. P. M$^c$Crae, D. Spohr, and P. Cimiano. Linking Lexical Resources and Ontologies on the Semantic Web with lemon. In *The Semantic Web: Research and Applications*, pages 245–259. Springer, 2011.

[46] J. P. M$^c$Crae, P. Cimiano, and E. Montiel-Ponsoda. Integrating Wordnet and Wiktionary with lemon. *Linked Data in Linguistics*, pages 25–34, 2012.

[47] J. P. M$^c$Crae, G. Aguado de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 2012.

[48] B. Pereira, N. Aggarwal, and P. Buitelaar. AELA: an Adaptive Entity Linking Approach. In *Proceedings of the* $22^{nd}$ *International conference on World Wide Web companion*, pages 87–88, 2013.

[49] B. Pouliquen. Similarity of Names Across Scripts: Edit Distance Using Learned Costs of N-Grams. In B. Nordström and A. Ranta, editors, *Advances in Natural Language Processing*, volume 5221 of *Lecture Notes in Computer Science*, pages 405–416. Springer Berlin Heidelberg, 2008.

[50] D. Rao, P. McNamee, and M. Dredze. Entity Linking: Finding Extracted Entities in a Knowledge Base. In *Multi-source, Multilingual Information Extraction and Summarization*, pages 93–115. Springer, 2013.

[51] A. Savary, L. Manicki, and M. Baron. Populating a multilingual ontology of proper names from open sources. *Journal of Language Modelling*, 1(2):189–225, 2013.

[52] J. Steinberger, J. Belyaeva, B. Crawley, L. Della-Rocca, M. Ebrahim, M. Ehrmann, M. Kabadjov, R. Steinberger, and E. Van der Goot. Highly Multilingual Coreference Resolution Exploiting a Mature Entity Repository. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 254–260, Hissar, Bulgaria, 2011.

[53] R. Steinberger. Multilingual and cross-lingual news analysis in the Europe Media Monitor (EMM). In M. Lupu ans E. Kanoulas and F. Loizides, editors, *Multidisciplinary Information Retrieval,* $6^{th}$ *Information Retrieval Facility Conference (IRFC'2013)*, volume 8201 of *Springer Lecture Notes in Computer Science*, pages 1–4. Springer, 2013.

[54] R. Steinberger and B. Pouliquen. Cross-lingual named entity recognition. *Lingvisticae Investigationes*, 30(1): 135–162, 2007.

[55] R. Steinberger, B. Pouliquen, and E. van der Goot. An introduction to the Europe Media Monitor family of applications. In *Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009)*, pages 1–8, Boston, USA, July 2009.

[56] R. Steinberger, B. Pouliquen, M. Kabadjov, and E. van der Goot. JRC-Names: A Freely Available, Highly Multilingual Named Entity Resource. In *Proceedings of the* $8^{th}$ *International Conference Recent Advances in Natural Language Processing (RANLP'2011)*, pages 104–110, Hissar, Bulgaria, September 2011.

[57] R. Stern and B. Sagot. Resources for Named Entity Recognition and Resolution in News Wires. In *Workshop on Resource and Evaluation for Entity Resolution and Entity Managment, collocated with the* $7^{th}$ *International Conference on Language Resources and Evaluation*, 2010.

[58] O .Täckström. *Predicting Linguistic Structure with Incomplete and Cross-Lingual Supervision*. PhD thesis, Uppsala University, 2013.

[59] A. Toral, S. Ferrández, M. Monachini, and R. Muñoz. Web 2.0, Language Resources and standards to automatically build a multilingual Named Entity lexicon. *Language Resources and Evaluation*, 46(3):383–419, 2012.

[60] C. Trojahn, B. Fu, O. Zamazal, and D. Ritze. State-of-the-Art in Multilingual and Cross-Lingual Ontology Matching. In P. Buitelaar and P. Cimiano, editors, *Towards the Multilingual Semantic Web*, pages 119–135. Springer Berlin Heidelberg, 2014.

[61] M. Turchi, M. Atkinson, A. Wilcox, B. Crawley, S. Bucci, R. Steinberger, and E. van der Goot. ONTS: "OPTIMA" News Translation System. In *Proceedings of the* $13^{th}$ *Conference of the European Chapter of the Association for Computational Linguistics*, pages 25—30, Avignon, France, 2012.

[62] C. Unger and P. Cimiano. Pythia: Compositional Meaning Construction for Ontology-Based Question Answering on the Semantic Web. In *Proceedings of the* $16^{th}$ *International Conference on Applications of Natural Language to Information Systems*, pages 153–160. Springer, 2011.

[63] C. Unger, J. P. M$^c$Crae, S. Walter, S. Winter, and P. Cimiano. A *lemon* lexicon for DBpedia. In S. Hellmann, A. Filipowska, C. Barriere, P. Mendes, and D. Kontokostas, editors, *Proc. of* $1^{st}$ *International Workshop on NLP and DBpedia*, Sydney, Australia, 2013.

[64] A.E. van Aggelen and L. Hollink. Plenary debates of the european parliament as linked open data, Website accessed in February 2015. URL `http://www.talkofeurope.eu/data/`.

[65] A. van den Bosch and T. Bogers. Memory-based Named Entity Recognition in Tweets. In *Proceedings of Making Sense of Microposts (MSM2013) Concept Extraction Challenge*, pages 40–43. Rio de Janeiro, Brazil, 2013.

[66] M. Villegas and N. Bel. PAROLE/SIMPLE 'Lemon' ontology and lexicons. *Semantic Web Journal*, 2013.

[67] A. Weichselbraun, D. Streiff, and A. Scharl. Linked Enterprise Data for Fine Grained Named Entity Linking and Web Intelligence. In *Proceedings of the* $4^{th}$ *International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, pages 13:1–13:11, 2014.

[68] W. Wentland, J. Knopp, C. Silberer, and M. Hartung. Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Translitera-

tion. In *Proceedings of the* $6^{th}$ *International Conference on Language Resources and Evaluation*, 2008.

[69] W. Zaghouani. Critical survey of the freely available Arabic corpora. In *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*, page 1, 2014.