Semantic Web 0 (2016) 1–11
IOS Press

The debates of the European Parliament as Linked Open Data

Editor(s): Natasha Noy, Stanford University, USA
Solicited review(s): Konrad Höffner, University of Leipzig, Germany; Alvaro Graves, Scale Free Labs LLC, USA; Adegboyega Ojo, National University of Ireland Galway, Ireland

Astrid van Aggelen [a,*], Laura Hollink [b], Max Kemman [c], Martijn Kleppe [d], and Henri Beunders [d]
[a] *Department of Computer Science, VU University, Amsterdam, The Netherlands*
E-mail: a.e.van.aggelen@gmail.com
[b] *Center for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands*
E-mail: l.hollink@cwi.nl
[c] *Department of History, University of Luxembourg, Luxembourg*
E-mail: max.kemman@uni.lu
[d] *Department of History, Erasmus University Rotterdam, The Netherlands*
E-mail: {kleppe, beunders}@eshcc.eur.nl

Abstract
The European Parliament represents the citizens of the member states of the European Union (EU). The accounts of its meetings and related documents are open data, promoting transparency and accountability, and are used as source data by researchers. However, the official portal of these documents provides limited search facilities. This paper presents *LinkedEP*, a Linked Open Data translation of the verbatim reports of the plenary meetings of the European Parliament. These data are integrated with a database of political affiliations of the Members of Parliament, and enriched with detected topics from the EU's topic hierarchy and links to four other Linked Open Datasets. The results of this work are available through a SPARQL endpoint and a user interface with extensive browse and search facilities. It is now possible to combine in one query the time and topic of the debate, the spoken words - in any available translation - and information about the speaker uttering these, such as affiliations to countries, parties and committees. This paper discusses the design and creation of the vocabulary, data and links, as well as known use of the data.

Keywords: Linked Open Data, European Parliament, open government data, RDF, data modeling, multilingual data

1. Introduction

The European Parliament (EP) is the only directly elected body of the European Union (EU), composed of the representatives of the member states. During the plenary meetings, it debates and votes upon the laws and budget of the EU. To residents of the European Union, access to the documents of the European Parliament is a formal right[1] in order to make informed votes and hold the Members of Parliament accountable.

From a scientific perspective, the proceedings of the EU parliament are a valuable source of data, in particular for studies in Political Science and Public Administration. For instance, Proksch and Slapin [14] relate the speeches held in the EP to the speakers' political ideology and country of representation. By virtue of their

*Corresponding author. E-mail: a.e.van.aggelen@gmail.com

[1] Regulation (EC) No 1049/2001 of the European Parliament and of the Council

multilingualism, the proceedings of the EP have further proven a valuable resource for studies into Natural Language Processing and Machine Translation [8,15].

The European Parliament publishes its proceedings as Open Data. A search portal[2] gives access to HTML pages of the speeches held in the plenary sittings. These so-called 'Verbatim Reports of Proceedings' or 'Comptes Rendus in Extenso' (henceforth referred to as 'proceedings') contain the word for word transcripts of each speaker's utterances as well as translations to languages of other member states. The search interface allows users to query by date, speaker, and words occurring in the title of the debate, but does not support search by textual content or by speaker profile.

This paper demonstrates how the EU proceedings can be published as Linked Open Data to support a wider range of queries. It provides an account of the choices made in the design of the data and vocabulary, especially with regard to multilingualism and speaker roles. The proceedings are linked to other Open Data on the Web, including a general-purpose encyclopedia to provide information about the Members of Parliament, a geographical knowledge base for the EU countries, and a topic hierarchy covering the activities of the EP. The resulting dataset, which is called *LinkedEP*, thus allows users to formulate queries of greater complexity and expressiveness than is currently supported, combining speech content with speaker and country information. In the seven months following its release, *LinkedEP* has been queried 7500 times on our servers.

The work presented here fits in a series of efforts to translate government data into the machine readable Semantic Web standard RDF. Some of these are realized by governments (e.g. the parliaments of Italy[3] and the United Kingdom[4]), others by civic parties (e.g. Votewatch[5], Open Congress[6]), or, like the current work, in academia (e.g. the projects Political Mashup[7], PoliMedia[8], Whattheysaid[9], and the Data-gov Wiki[10]). A Linked Open Data version of the European Parliament data can play a central role in these initiatives. Not only are the topics discussed in the EP relevant to all EU countries, the people and parties involved also play a role in national politics, making *LinkedEP* a potential hub in a Web of Linked Government Datasets. The multilingual nature of the EP facilitates making links to data in each language. As a first example of this, links from the proceedings of the EP to those of the Italian parliament are provided.

In the next section the source materials of the dataset are presented. Section 3 gives an overview of how we represented the data in RDF classes and properties, and the rationale behind the modeling choices. The links to other RDF sources are presented in Section 4. Section 5 describes the data portal and Section 6 demonstrates observed uptake of the data. In Section 7 we reflect on the quality of the dataset and on directions for future work.

2. Source data

The plenary meetings of the European Parliament are organised in four- or two-day *sessions*[11] in Strasbourg and Brussels taking place roughly each month. On a typical session day, a number of matters are debated, interspersed with votes, questions, administrative duties, and occasional statements. Each separate activity taking place in the plenary session is referred to as an *agenda item*. The debates typically have a few dozen speeches, in which the floor is given to the President of the EP, Members of Parliament, EU officials, and invited speakers.

The proceedings of the plenary meetings are published on the website of the European Parliament. Supplemented with an external database with background information about the parliamentary members, they form the basis of our dataset. The content of the two source corpora of the dataset is discussed below. Reports, vote statistics and other documents available on the EU website are beyond the scope of this endeavour.

2.1. Proceedings

The account of the plenary meetings in the proceedings includes the structure of the parliamentary events from the session up to the speech level, and the content of the speeches. The proceedings provide dates and ordering information, the titles of agenda items, and for each speech, the language in which it is spoken, the speaker name, the speaker's official numerical ID

[2] http://www.europarl.europa.eu/plenary/en/debates-video.html
[3] http://dati.camera.it/
[4] http://lda.data.parliament.uk/
[5] http://www.votewatch.eu
[6] https://www.opencongress.org
[7] http://politicalmashup.nl/
[8] http://polimedia.nl/
[9] http://whattheysaid.org.uk
[10] http://data-gov.tw.rpi.edu

[11] also called *part-sessions*

(when applicable), the spoken text, and additional annotations. These annotations serve many purposes, for instance, to quote the speaker when his words are difficult to translate, or to mark special events or circumstances, for instance when a speech is received with applause or is spoken on behalf of a party. Speeches are presented in the proceedings as single-actor events, except when multiple speakers speak out collectively, for instance in a collaborative statement. There may be speeches without text, for instance to indicate a non-verbal act, which may be clarified by an annotation.

The account of what is said in the plenary meetings is multilingual, and parallel proceedings are available for each of the EU languages. Members of Parliament have (limited) rights to request translations [13] of their speeches for the proceedings if these are not already provided.

2.2. Members of Parliament in ADEP

The publicly available online Automated Database of the European Parliament (henceforth referred to as *ADEP*) [6] provides the source for the background information on the Members of Parliament. For each Member is given, in comma-separated format: the official ID, the first and last name, birth date, country of representation, and partisan history. The latter includes affiliations to EU committees, EU parties (including role descriptions), and national parties. *ADEP* is linked to the EP data through the Members of Parliament' unique numerical identifier.

3. Data model

This chapter explains the modeling principles we followed, and discusses and visualises the various sections of the resulting schema, such as the structure of the plenary events, the textual information and translations, and the Members of Parliament and their roles. Finally, it elucidates the choice of URIs and noteworthy translations from the source data to the schema.

3.1. Modeling principles

The data and vocabulary of *LinkedEP* are designed to facilitate use, re-usability, and interoperability.

To promote uptake, querying the data should be as straightforward as possible. For this reason, the backbone of the model is a direct translation of the structure of the events in Parliament. Moreover, a number of properties are introduced that are redundant but enable shorter and less complex queries, or avoid the need for reasoning engines. While this increases the number of RDF statements, the modest size of the dataset allows us to prioritise ease of use over the price of data storage. Finally, intuitive names are chosen for properties and classes. Experts from the information services of the European Union were consulted about the vocabulary used in practice, leading us to adopt, for instance, the term *session* instead of *part-session*.

The vocabulary for *LinkedEP* accommodates reuse for other proceedings and political datasets, such as EP committee meetings, national parliament meetings, and other types of events that cannot be foreseen at this moment. For this reason we call it the `Linkedpolitics` vocabulary and adhere to a minimum of semantic commitment. Domain and range restrictions are avoided, just like cardinality restrictions and statements about disjointness and functional properties. With this approach we follow van Hage et al. [18]. Reference to the EU is avoided where it can restrict reuse - e.g. in the names of properties and in most class and instance URIs - and added where it is deemed necessary to distinguish resources. For example, instances of countries, roles and institutions are not marked as EU-specific, while speeches, sessions, session days, and agenda items do have a designated EU component in their URI.

To increase interoperability with other Linked Open Datasets, properties from widely used vocabularies are reused or linked to where possible, in particular FOAF[12] and Dublin Core[13].

3.2. Structure of the plenary sessions

The backbone of the model, depicted in Figure 1, consists of the hierarchical structure of the events in Parliament, with classes denoting the sessions, session days, agenda items and speeches. The `hasPart` relationship relates higher level events to their parts. Information about the chronological order of events is contained in dates and numbering.

Additionally, the model contains redundant predicates `isPartOf` and `hasSubsequent`. The former is the inverse of `hasPart` and enables users to click through to the higher-level event while browsing through speeches and debates through the interface, for instance to see the title and topic of the debate in which

[12] `http://www.foaf-project.org/`
[13] `http://dublincore.org/`

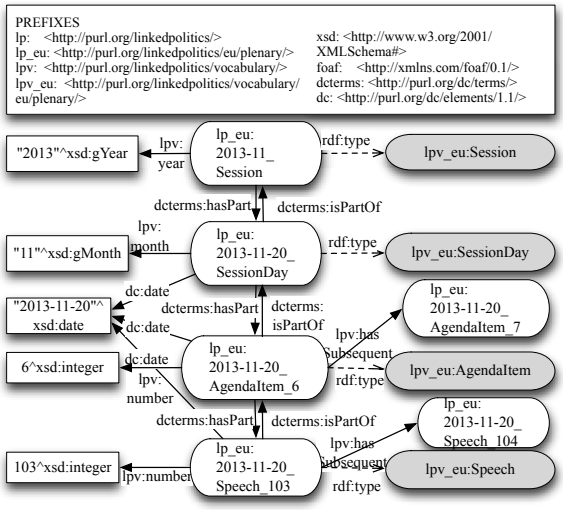

Figure 1. The exemplified backbone of the model, which expresses the hierarchy and order of the parliamentary events. The coloured boxes denote classes. The namespaces are clarified at the top.

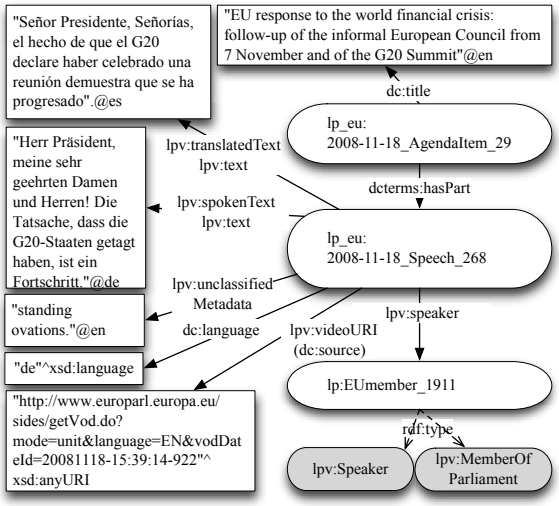

Figure 2. The content-level information in the model, exemplified. Parenthesized are the superproperties, where applicable. The coloured boxes denote classes.

a speech was held. The predicate `hasSubsequent` serves to request follow-up items for speeches and agenda items. While the dates and numbers suffice to find subsequent instances of each kind of event, this requires the use of query operators, which might be cumbersome to inexperienced SPARQL users.

3.3. Unclassified metadata of speeches

As illustrated in Figure 2, `Speech` instances are sometimes accompanied by miscellaneous annotations regarding the delivery of the speech, which we call `unclassifiedMetadata`. Examples include mentions of interruptions and applause, and role-statements such as "on behalf of PPE". All information in the speech that is presented on the EU website in italics is taken to be such meta-information.

3.4. Languages and translations of textual data

All textual data – titles of agenda items, speech transcripts, and unclassified metadata – are subject to translation. Each `Speech` instance has a `language` property to denote the language in which it was originally spoken. This facilitates queries for all speeches uttered in a certain language. Each speech instance has a `text` property for all available translations. These text literals are complemented with a language tag, so that speech texts in a particular language can easily be queried. Similarly, parallel language-annotated literals are available for `title` and `unclassifiedMetadata` literals.

The data model includes two auxiliary properties for speech contents: `translatedText` and `spokenText` (Figure 2). The words of the speech in the original language are pointed to by `spokenText` to facilitate users who are only interested in original transcripts; and the translated text by property `translatedText`. The triples generated from these properties eliminate the need for users to combine the spoken language and the transcription language in one query. `translatedText` and `spokenText` are subproperties of `text`, which retrieves transcripts regardless of their original language.

3.5. Speakers and Members of Parliament

The `speaker` property connects a speech to a speaker (Figure 2). All speech actors are assigned to class `Speaker`; if a numerical ID is provided in the online proceedings, the instance is additionally assigned to class `MemberOfParliament`. In that case, the URI is based on the ID number, while for non-MEP speakers the URI contains the full name provided in the online proceedings.

While non-MEP speaker instances have just a `name` property, the Members of Parliament are annotated with extensive information from *ADEP*, including a separate `givenName` and `familyName`. The date of

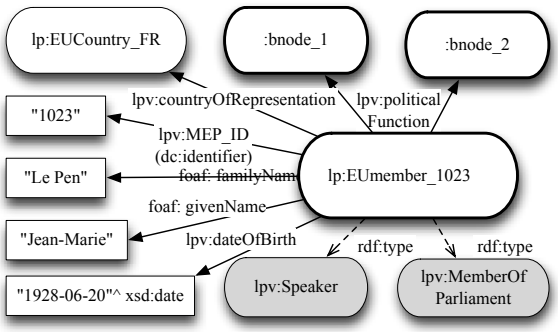

Figure 3. Example representation of a Member of Parliament. The coloured boxes denote classes and thick-edged boxes denote entities that reoccur in Figure 4.

birth and country of representation are also given, as well as political functions (Figure 3).

3.6. Political functions

Figure 4 shows how the political affiliations of MEPs are modeled, building on the example in Figure 3. A `PoliticalFunction` instance reflects one entry in *ADEP*. It connects a `Role` and a `Political-Institution` instance, and on- and offset literals of type `xsd:date`. The `PoliticalInstitution` class currently has subclasses `NationalParty`, `EUParty`, and `EUCommittee`. The `Role` class has about a dozen instances, denoting concepts such as *member* and *vice-chair*. The concept of *political function* is defined solely by its attributes, and no meaningful identifier could be assigned to `Political-Function` instances other than a concatenation of their property values. For this reason they are represented as blank nodes.

A supplementary relation, `spokenAs`, connects speeches and the momentary political affiliations of the speaker. This relation is derived from the speaker and date information of the speech and all political functions defined for the speaker. While the `politicalFunction` property is convenient for querying politicians and their functions, it does not accommodate searching for *speeches* by politicians in certain functions. For example, a query for speeches held by the chair of a given committee returns all speeches by MEPs who once had that role, even the ones spoken years after. Relating speeches to partisan interests directly, the `spokenAs` property frees the user from the burden of defining date restrictions on `politicalFunction` and running possibly expensive queries.

3.7. Dataset description and provenance

The content and provenance of the data and vocabulary are described using the `void`[14], `prov`[15] and `omv`[16] vocabularies. For the dataset as a whole, information is given about the content, the makers, the license, the URIs, and access. To allow for separate annotations, the dataset is split into several RDF graphs. For instance, the information about the structure of the events in the EP is separated from the textual information, which is stored in one graph per language. For each graph is given: a title, a description, the used source and a description of the generation process, the download link, and - for linksets - the source and target dataset. The metadata are collected in a single graph on the server and as a turtle file in the *well-known* directory[17], as recommended in the `void` specification[14].

3.8. URIs

The namespace `http://purl.org/linked-politics` forms the basis for all URIs, reflecting our aim to gather different political datasets under one umbrella. Schema URIs are marked by an additional component `vocabulary`. Some classes and instances, for instance the speeches, have additional components `eu` and `plenary` in their URIs. This is to distinguish them from possible equivalents at other levels of organisation, that would otherwise get the same URI. For example URIs, we refer to Section 3, in particular Figure 1, which declares the used namespaces.

3.9. Conversion process

The proceedings were extracted from the website of the European Parliament following the method of Gielissen and Marx [5], who proposed an XML Document Type Definition for parliamentary proceedings. The raw XML data were then translated into the *LinkedEP* RDF data model using a SWI Prolog toolkit for converting XML to RDF [18]. This package supports an initial, automatic conversion step, and provides a syntax to rewrite the resulting RDF resources into the desired schema.

[14] `https://www.w3.org/TR/void/`
[15] `https://www.w3.org/TR/prov-o/`
[16] `http://omv2.sourceforge.net/`
[17] `http://purl.org/linkedpolitics/.well-known/void`
[18] `http://semanticweb.cs.vu.nl/xmlrdf`

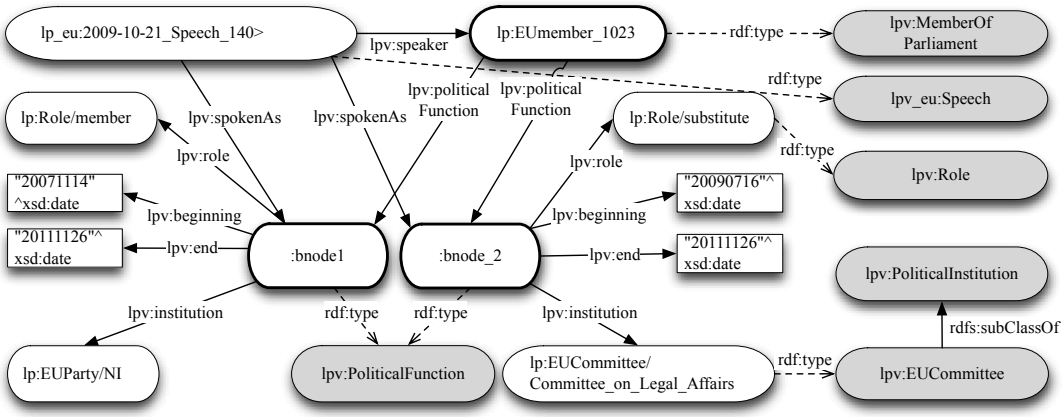

Figure 4. Example representation of the political functions of a Member of Parliament, as defined by a role, institution, and time span. The coloured boxes denote classes and thick-edged boxes denote entities that reoccur in Figure 3.

Textual properties such as paragraphs within speeches (separated by white space) were left out. Also, entities were introduced for Members of Parliament, which were denoted as literals in the raw XML data.

The value of the `dc:language` property for non-English speeches was extracted from the speech level annotations on the website of the EP. It was assumed that every (single) ISO-compliant, abbreviated, parenthesized language code found in italics at the beginning of a speech denotes a language. The language tag of the speech literals was derived from the display language selected in the EP web interface. This language was taken to apply to all text on the web page, including titles of agenda items and speech texts. This feature was exploited to automatically language-annotate the titles and speech texts while scraping the site in the selected language.

The auxiliary properties `spokenText` and `translatedText` were derived from the extracted language information. The `spokenAs` property was obtained by combining date and speaker information extracted from the EP website with political functions of the speaker from ADEP. All Prolog source code used for the conversion is available from GitHub[19].

4. Links to the LOD cloud

A start has been made with connecting *LinkedEP* to the LOD cloud with links to four external knowledge sources: two about politicians' backgrounds, one geographical database, and a topic taxonomy. Additionally, the dataset has been linked to from a third party source: the European Union Data Portal[20] provides 887 links between Member of Parliament instances in *LinkedEP* and its named entity resource JRC-Names [4], available through a SPARQL endpoint. For each linked source, an example is given in Figure 5.

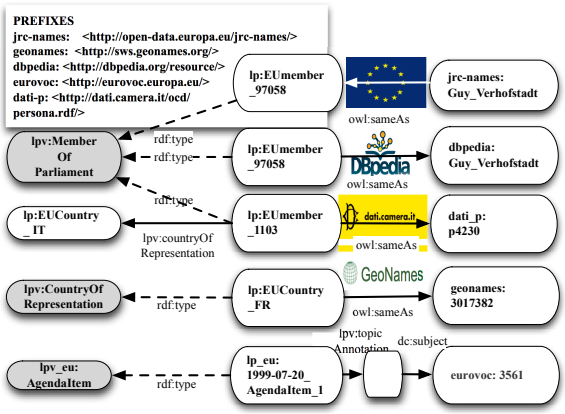

Figure 5. Examples of links to and from *LinkedEP*: inlinks from JRC-Names to *LinkedEP* MEPs, outlinks from *LinkedEP* MEPs to DBpedia and the Italian parliament, from *LinkedEP* countries to GeoNames, and from agenda items to the Eurovoc thesaurus.

The country entities in *LinkedEP* are connected to their counterparts in GeoNames[21], a geographical database. This connection brings in information that could be useful in debate analyses, such as the area, population, languages, neighbouring countries,

[19] https://github.com/aan680/LinkedEP

[20] https://open-data.europa.eu/en/data

[21] http://geonames.org

and territorial dependencies of the EU countries. These links are based on the two-letter ISO 3166 country codes, which are stored as the value of property `acronym` and are manually verified.

The Members of Parliament are linked to their entries in DBpedia[22], the RDF counterpart of Wikipedia. Besides structured biographical properties, some of which overlap with *ADEP*, DBpedia provides textual descriptions, references to the comprehensive Yago ontology, and a link to the corresponding entry in Wikipedia. First, candidate matches in DBpedia were generated based on a simple string matching process. Then, all links were verified and if necessary corrected by human judges. Out of 1258 generated matches, 75 needed correction or removal. This process resulted in 1226 links to the English language DBpedia on a total of 3115 MEPs in our dataset (1423 of whom were a Member of Parliament during the time covered in *LinkedEP*). The proportion of MEPs with a link into DBpedia can be increased when localised DBpedia chapters are included. To illustrate this, the matching process was repeated on the Polish DBpedia, which is among the most complete chapters on this topic. This resulted in an additional 186 links. *LinkedEP* provides only one link to a DBpedia resource per MEP, as DBpedia contains (albeit incomplete) `owl:sameAs` links between corresponding resources in the localised editions.

The politicians representing Italy are matched to the official RDF database[23] of the Italian parliament. This connection allows users to compare politicians' utterances in the European and the national setting. The cues taken for this mapping are the name and birth date of the politicians. Because of the modest number of Italian MEPs, the mapping results were manually checked for correctness and completeness.

Finally, the agenda items are related to a topic hierarchy. EuroVoc[24] is the EU's multilingual thesaurus, which captures all domains in which the European Parliament is active. It can be downloaded in RDF from the EU Linked Data Portal[25]. Eurovoc comes with special purpose classification software, JEX, machine-trained to label documents with one or more instances from the topic hierarchy [16]. Using this software on the collated English transcripts of speeches within an agenda item resulted in topic annotations for over 90 percent of all debates for which textual content is available. The topic annotations are represented in RDF using the ontology pattern for n-ary relations [12], to accomodate a confidence value of each annotation. The EuroVoc thesaurus is imported to the Linkedpolitics portal to enable users to query the hierarchical index and the multilingual labels within our server.

5. Access, scope and size of the data

The *LinkedEP* data are available under an open license[26] and can be accessed from data portal `http://purl.org/linkedpolitics`, providing several search, browse and access possibilities including a SPARQL endpoint.

The data portal runs on the Semantic Web server ClioPatria[27]. It displays summaries of each RDF graph, allowing users to browse through the classes and properties up to the instance level. A free-text search bar accommodates keyword queries. ClioPatria provides a SPARQL endpoint and query editor implementing most features of the latest SPARQL version, 1.1. Through an environment called SWISH, it supports querying using SWI Prolog, which features libraries for federated querying amongst other functionalities. The RDF graphs can be downloaded in Turtle and RDF/XML serialisations.

All URIs are dereferenceable and return an overview of the triples defined for the given resource. To guarantee their persistence, the domain `http://purl.org/linkedpolitics` is registered as a Persistent Uniform Resource Locator (PURL[28]), which currently redirects to a service hosted at VU University Amsterdam. This service hosts the latest version (currently version 1.0, released on 23 Oct 2105). It contains the proceedings from 20 July 1999 onward, i.e. the start of the fifth term, when the EP started publishing the proceedings in the current interface[29], and contains around 300K speeches, embedded in 22K agenda items and 1K session days, featuring 1.5K distinct Members of Parliament. We aim to update the repository yearly to include the latest debates of the

[22] `http://dbpedia.org`
[23] `dati.camera.it`
[24] `http://eurovoc.europa.eu`
[25] `http://open-data.europa.eu`

[26] CC0 1.0 Universal
[27] `http://cliopatria.swi-prolog.org/`
[28] `http://purl.org`
[29] `http://www.europarl.europa.eu/plenary/en/debates-video.html`. In the legacy interface `http://www.europarl.europa.eu/omk/omnsapir.so/calendar?LANGUE=EN&APP=CRE`, the debates date back to 15th of April 1996.

EP. These updates will not change the existing data or the data model, and are therefore not treated as new versions. In case of changes to previous data or to the data model, however, a new version will be served with the appropriate version number in the URIs; old versions remain available as data dumps. To ensure persistence of the data for decades to come, a stable version including the 5th, 6th and 7th EP has been deposited[30] at the institute for Data Archiving and Networking (DANS)[31].

6. Third party use

In the 29 weeks following its announcement, the homepage of *LinkedEP* was visited more than five thousand times and the dataset was queried through our service 7504 times, of which 3654 times in SWISH/SWI-Prolog and 3850 times in SPARQL. Manual inspection of the logs reveals that queries containing regular expressions are particularly prevalent, as well as queries with count operations. In total, 1648 out of the 3850 SPARQL queries in our logs include a regular expression. 1600 queries have a count operation, and 906 have both.

While query log analysis gives a good indication of the use of the data, it does not identify the information need or envisaged application behind the queries. In the remainder of this section, we will delve deeper into a selection of the logged queries. For each of these queries we have had contact with the user that ran the query to determine the underlying research questions and application scenarios. The interaction with users took place in the context of three week-long workshops that were organized by the authors[32].

Use case 1: A study into the role of higher education in the EP Birkholz [2] studies how higher education is proposed as a solution to various policy problems that are not in themselves related to higher education. She thereby considers the role of parliamentary committees, individual members of the EP, political parties, coalitions, and nation-states. Displayed below is a query that was used within this study to select speeches with the keyword 'education', and that returns their identifier, text, and the name of the EU party of the corresponding speaker.

```
SELECT ?speech ?text ?partyname
WHERE {
   ?speech lpv:text ?text.
   FILTER ( langMatches(lang(?text),"en") )
   FILTER regex( str(?text), "education" )
   ?speech lpv:spokenAs ?politicalFunction.
   ?politicalFunction lpv:institution ?party.
   ?party rdf:type lpv:EUParty.
   ?party lpv:acronym ?partyname.
}
```

Use case 1 is an example of a frequently observed usage pattern of selecting potentially relevant items for further close reading, a pattern that was also identified by Traub et al. [17] among users of digital historical archives. In the *LinkedEP* dataset, speeches are typically selected based on dates, the occurrence of a keyword or topic in the debate, and/or information about the speaker, such as country, party or committee membership. Other use cases that we observed that follow this pattern include a study of debates about data privacy and transparency, a study into the perspectives of the different parliamentary groups on the financial crisis, and an analysis of the use of emotionally charged words by MEPs.

Use case 2: A comparison of the discourse of political groups The discursive practices of MEPs affect public opinion on the issues debated in the EP. Nerghes et al. [11] explore speeches during the recent Eurozone financial crisis to expose discursive practices of the two largest political groups on either side of the left-right political ideology spectrum. Using a text mining tool, they semi-automatically code the English speech texts and carry out semantic network analysis. The query below retrieves the data for one of the selected parties.

```
SELECT DISTINCT ?date ?text
WHERE {
   ?sessionday dcterms:hasPart ?agendaItem.
   ?sessionday dc:date ?date.
   ?agendaItem dcterms:hasPart ?speech.
   ?speech lpv:text ?text.
   ?speech lpv:speaker ?speaker.
   ?speaker lpv:politicalFunction ?function.
   ?function lpv:institution ?party.
   ?party lpv:acronym ?partyname.
   FILTER regex(str(?partyname), "S&D").
   FILTER ( ?date >= "2009-08-01"^^xsd:date
     &&?date <= "2014-07-31"^^xsd:date )
   FILTER ( langMatches(lang(?text), "en") )
}
```

[30] http://dx.doi.org/10.17026/dans-2xg-umq8

[31] www.dans.knaw.nl/

[32] www.talkofeurope.eu/creative-camp-1/, www.talkofeurope.eu/creative-camp-2/,www.talkofeurope.eu/creative-camp-3/

In a similar study [3], the evolution of the discourse during the financial crisis is explored. The query below searches for speeches that contain mentions of financial or economic crisis (or crises), as is captured by a regular expression, and returns the counts by date. This query was repeated with various keywords to verify the occurrence of the targeted keywords and their correspondence with the economic crisis. This task corresponds to the usage pattern of investigating quantitative results over time as identified by Traub et al. [17].

```
SELECT DISTINCT ?year ?month
 (COUNT(DISTINCT ?speech) AS ?speechno)
WHERE {
    ?speech lpv:text ?text.
    FILTER ( langMatches(lang(?text),"en") )
    FILTER regex( str(?text),
      "financ*|econom*&&cris*s","i" )
    ?speech dc:date ?date.
}
GROUP BY ?date
```

In use case 2, relevant speeches are retrieved for consecutive offline processing by other tools. Other examples of this usage pattern were encountered: (1) for a visualization of (statistically) salient words used by MEPs per country and per month [9] (2) in a study about how the EP talks about rulings of the European Court of Justice [19], in which speeches that mention the court in combination with the word 'ruling' or 'case' were processed offline by a custom matching algorithm to link the speech to a specific court case in the EUR-Lex database.

Usage logs of Linked Data servers typically capture only part of the actual use of the data; downloading all RDF onto a local disk for further querying and processing is a common practice on the Semantic Web. Also, the usage of the links to the *LinkedEP* data provided by the European Union Data Portal cannot be tracked.

7. Quality

In the star system by Berners-Lee [1], *LinkedEP* is a five-star collection. The first three stars are credited for, respectively, the open license, the structured format, and the non-proprietariness of the latter. The use of URIs and the links to other data grant *LinkedEP* the fourth and fifth star.

Dataset quality Zaveri et al. [20] provide an inventory of indicators of the intrinsic quality of linked datasets. We have checked [33] the data for each of the described metrics for consistency (where applicable) and found no contradictions.

Vocabulary quality Janowicz et al. [7] propose quality indicators for vocabularies. Following their rating scheme, the `Linkedpolitics` vocabulary is worth four stars out of five: it is in machine-readable format (2 stars), it is linked to other vocabularies such as FOAF and Dublin Core (3 stars), and it is annotated with properties from the `void`, `prov` and `omv` vocabularies (4 stars). However, to gain the 5th star requires the vocabulary to be taken up by others. While the vocabulary presented here was designed for other events than the meetings of the European Parliament, to the best of our knowledge it has not yet been reused.

Known shortcomings and future work The translation service of the EP translates debates into a selection of other languages, depending on the topic and importance of the debate. In cases where a translation into a particular language is not available, the quality of the language tags of the speech literals in *LinkedEP* drops. This is due to the fact that *LinkedEP* is based on the website of the EP, where the same problem exists: speeches without translation in the selected language are displayed in their original language without warning. A start has been made to remedy this: all `spokenText` literals were processed with an off-the-shelf language identification tool [10] and had their language tag corrected with the detected language. Moreover, `translatedText` triples were removed for speeches that were not translated at all. However, some incorrect language tags remain.The exact quantification of the problem and the effects of the correction procedure remains future work.

There is considerable room for outreach to a wide range of other datasets, including the records of national parliaments and other open government data, encyclopedic sources such as the CIA Factbook, and news media archives. For instance, the EU parties, national parties and EU committees in our dataset can be linked to their entries in DBpedia or country-specific Open Datasets. The sources that are currently linked to were chosen either because of their low cost (e.g. country names are relatively unambiguous and therefore easy to match) or high gain (e.g. DBpedia's central

[33] source code (Prolog rules) available on GitHub

position in the LOD cloud means that it gives access to many other datasets). Future work includes expanding the links to more Open Datasets.

8. Conclusion

LinkedEP is an RDF translation of the verbatim proceedings of the plenary sessions of the European Parliament, supported by a newly introduced vocabulary, `Linkedpolitics`. With its links to various background datasets, *LinkedEP* supports rich user queries. To facilitate ease of use of the data, established vocabularies were re-used where possible; redundant properties were introduced to facilitate shorter queries; and source and provenance information were added to make the data self-evident.

Acknowledgments

This work was done in the Talk of Europe project, funded by Clarin ERIC and Clarin-NL. To generate the dataset, we gratefully used scripts from Political Mashup. Jan Wielemaker assisted us with ClioPatria and SWI-Prolog. The connection to the Italian Parliament data was made by Silvia Giannini during the Creative Camp organised by Talk of Europe. We thank Adina Nerghes, Jonathan Gray and Julie Birkholz for involving us in their research and writing queries together. Ingelise de Boer from the European Parliament Information Office taught us everything we needed to know about the workings of the European Parliament.

References

[1] Tim Berners-Lee. Linked Data, 2006. URL `http://www.w3.org/DesignIssues/LinkedData.html`. Accessed: 2014-11-28.

[2] Julie M. Birkholz. Network of higher education institutions: A social network approach to the study of governance arrangements, September 2015. Presented at the 28th Annual Conference of the Consortium of Higher Education Researchers (CHER 2015).

[3] Anastasia Deligiaouri. Analysis of the political discourse of European Parliament political groups during economic crisis (2008-2014). In *24th World Congress of Political Science, Istanbul, Turkey, July 2016*, 2016. URL `https://wc2016.ipsa.org/my-ipsa/events/istanbul2016/paper/analysis-political-discourse-european-parliament-political-groups-`. Submission under review.

[4] Maud Ehrmann, Guillaume Jacquet, and Ralf Steinberger. JRC-Names: Multilingual Entity Name variants and titles as Linked Data. *Semantic Web*, 2016. URL `http://www.semantic-web-journal.net/content/jrc-names-multilingual-entity-name-variants-and-titles-linked-data-1`. In press.

[5] Tim Gielissen and Maarten Marx. Exemelification of parliamentary debates. In Robin Aly, Claudia Hauff, Ida den Hamer, Djoerd Hiemstra, Theo Huibers, and Franciska de Jong, editors, *Proceedings of the 9th Dutch-Belgian Information Retrieval Workshop (DIR 2009)*, number 09-01 in Workshop Proceedings Series, pages 19–25, Enschede, February 2009. Centre for Telematics and Information Technology. URL `http://doc.utwente.nl/65379/`.

[6] Bjørn Høyland, Indraneel Sircar, and Simon Hix. Forum Section: An automated database of the European Parliament. *European Union Politics*, 10(1):143–152, 2009. doi:10.1177/1465116508099764.

[7] Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, and Charles Vardeman. Five stars of linked data vocabulary use. *Semantic Web*, 5(3):173–176, 2014. doi:10.3233/SW-140135.

[8] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

[9] Ilya Kuzovkin Konstantin Tretyakov and Aleksandr Tkatsenko. Talk of Europe: Significant words, 2015. URL `http://europe.all.my/`. Accessed: 2015-11-04.

[10] Marco Lui and Timothy Baldwin. langid.py: An off-the-shelf language identification tool. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the System Demonstrations, July 10, 2012, Jeju Island, Korea*, pages 25–30. Association for Computational Linguistics, 2012. URL `http://www.aclweb.org/anthology/P12-3005`.

[11] Adina Nerghes, P. Groenewegen, and I. Hellsten. Europe Talks: An analysis of discursive practices, their structural functions and the left-right political ideology spectrum in the European Parliament, July 2015. Presented at the International Sunbelt Social Networks Conference, Brighton, UK.

[12] Natasha Noy and Alan Rector, editors. *Defining N-ary Relations on the Semantic Web*. W3C Working Group Note, 12 April 2006. URL `https://www.w3.org/TR/swbp-n-aryRelations/`.

[13] The Bureau of the European Parliament. Code of conduct on multilingualism: Bureau decision of 16 June 2014, 2014. URL `http://www.europarl.europa.eu/pdf/multilinguisme/coc2014_en.pdf`.

[14] Sven-Oliver Proksch and Jonathan B. Slapin. Position taking in European Parliament speeches. *British Journal of Political Science*, 40:587–611, 7 2010. doi:10.1017/S0007123409990299.

[15] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, May 22-28, 2006, Genoa, Italy*. European Language Resources Association, 2006. URL `http://www.lrec-conf.org/proceedings/lrec2006/pdf/340_pdf.pdf`.

[16] Ralf Steinberger, Mohamed Ebrahim, and Marco Turchi. JRC EuroVoc Indexer JEX - A freely available multi-label categori-

sation tool. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 798–805. European Language Resources Association (ELRA), 2012. URL http://www.lrec-conf.org/proceedings/lrec2012/summaries/875.html.

[17] Myriam C. Traub, Jacco van Ossenbruggen, and Lynda Hardman. Impact analysis of OCR quality on research tasks in digital archives. In Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla, editors, *Research and Advanced Technology for Digital Libraries - 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015. Proceedings*, volume 9316 of *Lecture Notes in Computer Science*, pages 252–263. Springer, 2015. doi:10.1007/978-3-319-24592-8_19.

[18] Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. Design and use of the Simple Event Model (SEM). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):128–136, 2011. doi:10.1016/j.websem.2011.03.003.

[19] Karin van Leeuwen, Hilde Reiding, Bart Vredebregt, and Radboud Winkels. Chambers to chambers: ECJ rulings in European Parliamentary debate, 2015. URL http://www.talkofeurope.eu/creative-camp-3/abstracts/#Chambers. Accessed: 2015-11-06.

[20] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for Linked Data: A survey. *Semantic Web*, 7(1):63–93, 2016. doi:10.3233/SW-150175.