The debates of the European Parliament as Linked Open Data

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@vu.nl
[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. 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 linked to three other Linked Open Datasets. The resulting data of over 25 million triples are available through a user interface and a SPARQL endpoint, enabling queries about the monthly sessions of the European Parliament, the agenda of the debates, the spoken words and their translations into other EU languages, and information about the speakers such as affiliations to countries, parties and committees. The 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

This paper presents a Linked Open Data version of the proceedings of the European Parliament (EP). The 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 [7] relate the speeches held in the EP to the speakers' political ideology and country of representation. By virtue of their multilingualism, the proceedings of the EP have further proven a valuable resource for studies into Natural Language Processing and Machine Translation [6,8].

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

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

[2] `http://www.europarl.europa.eu/plenary/en/debates-video.html`

'Comptes Rendus in Extenso', which will simply be called 'proceedings' in the remainder of this paper, contain the verbatim transcripts of each speaker's utterances. Speakers are free to talk in any of the 24 official EU languages, and for parts of the proceedings translations to languages of other member states are available. The search interface allows users to query by date, speaker, and words occurring in the title of the debate.

This paper demonstrates how the EU proceedings can be published as Linked Open Data. 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: they are integrated with an online database containing background knowledge about the Members of Parliament, the politicians are linked to their entry in an online general-purpose encyclopedia, and the country names are connected to a rich and well-established geographical knowledge base. The resulting dataset, which is called *LinkedEP*, thus allows users to formulate queries that combine speech content with speaker and country information, and to formulate queries of greater complexity and expressiveness than is currently supported. In the seven months following its release, the *LinkedEP* data have been queried 7,500 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 it a potential hub in a Web of Linked Government Datasets. The multilingual nature of the EP facilitates the creation of 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-day *sessions*[11] in Strasbourg, taking place almost every month, and in two-day sessions, which are held in Brussels roughly every other month. On a typical session day, a number of matters are debated, interspersed with votes, questions and administrative duties, as well as occasional statements. Each separate activity taking place in the plenary session is referred to as an *agenda item*. An agenda item typically consists of a sequence of a few dozen speeches, with the President giving the introductory and the closing speech, where the floor is given to 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. Reports, vote statistics and other documents on the EU website are beyond the scope of this endeavour. The content of the two source corpora of the dataset is discussed below.

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 (when applicable), the spoken text, and additional annotations. These annotations denote special events or circumstances, for instance when a speech is re-

[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*

ceived with applause or is spoken on behalf of a party. Speeches are presented in the proceedings as single-actor events. Therefore, whenever a speaker is interrupted, a new speech starts. There may be speeches without text, for instance to indicate a non-verbal act, which is usually clarified by an annotation. Also, speeches can list multiple speakers, in case these behave as one actor, for instance when a collaborative statement is read out.

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 [1] of their speeches for the proceedings, if these are not planned to be provided.

2.2. Members of Parliament in ADEP

The publicly available online Automated Database of the European Parliament (henceforth referred to as *ADEP*) [4] 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, and national parties. In total, *ADEP* describes 1,813 politicians active between 1999 and 2014; this includes all the Members who spoke in Parliament between July 1999 and July 2014. *ADEP* is linked to the EU data through the ID of the Members of Parliament.

2.3. Scope and statistics

LinkedEP covers the complete fifth, sixth, and seventh term of the European Parliament. The materials range between 20 July 1999, when the EU started publishing the proceedings in the current interface [12] and July 2014. This 15-year collection of proceedings contains 304,500 speeches, embedded in 21,678 agenda items and 879 session days, featuring talking turns by 1,692 different Members of Parliament. *LinkedEP* is a close translation of the source data.

3. Data model

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

3.1. Modeling principles

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

To promote uptake by users unfamiliar with Semantic Web practices, querying the data should be as straightforward as possible. After all, the organisation of the EP is inherently complex. To this end, first, a number of properties are introduced which are redundant but enable shorter and less complex queries. This increases the number of RDF statements, but given the modest size of the dataset priority was given to ease of use over the price of data storage. Similarly, the model makes inferable information explicit. We have explicitly included all triples that could be generated at query time by a reasoning engine that supports OWL, such as inverse properties. Users are not assumed to have access to such reasoners.

Second, the account of reality is simplified where possible. Concretely, the parts of the plenary sessions are treated as events and archived documents at the same time. That is, one item is simultaneously assigned document-like properties, such as textual content, and event-like properties, such as speaker, or properties ambiguous in this respect, e.g. *has part*. This choice is supported by the nature of the source materials, since verbatim reports are a direct account of reality.

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* was designed to accommodate 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 that reason, we call it the `LinkedPolitics` vocabulary. First, this was done by adhering to a minimum of semantic commitment of the model. Domain and range restrictions are avoided, just like cardinality restrictions and state-

[12] `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.

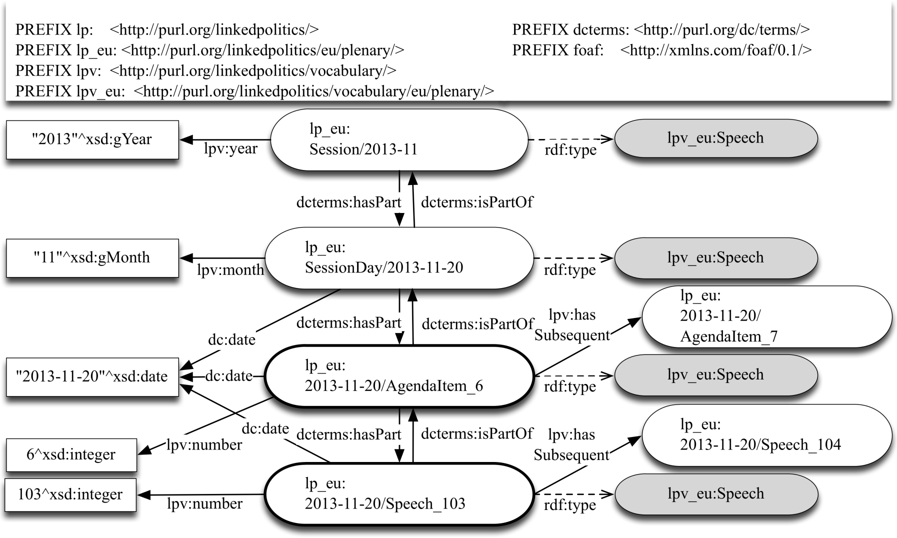

PREFIX lp: <http://purl.org/linkedpolitics/>
PREFIX lp_eu: <http://purl.org/linkedpolitics/eu/plenary/>
PREFIX lpv: <http://purl.org/linkedpolitics/vocabulary/>
PREFIX lpv_eu: <http://purl.org/linkedpolitics/vocabulary/eu/plenary/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
"2013"^xsd:gYear
lpv:year
lp_eu: Session/2013-11
rdf:type
lpv_eu:Speech
dcterms:hasPart
dcterms:isPartOf
"11"^xsd:gMonth
lpv:month
lp_eu: SessionDay/2013-11-20
rdf:type
lpv_eu:Speech
dc:date
dcterms:hasPart
dcterms:isPartOf
lpv:has Subsequent
lp_eu: 2013-11-20/ AgendaItem_7
"2013-11-20"^xsd:date
dc:date
lp_eu: 2013-11-20/AgendaItem_6
rdf:type
lpv_eu:Speech
6^xsd:integer
lpv:number
dc:date
dcterms:hasPart
dcterms:isPartOf
lpv:has Subsequent
lp_eu: 2013-11-20/Speech_104
103^xsd:integer
lpv:number
lp_eu: 2013-11-20/Speech_103
rdf:type
lpv_eu:Speech

Figure 1. The exemplified backbone of the model, which expresses the hierarchy and order of the parliamentary events. The coloured boxes denote classes and thick-edged boxes denote entities that reoccur in other figures. The namespaces are clarified at the top.

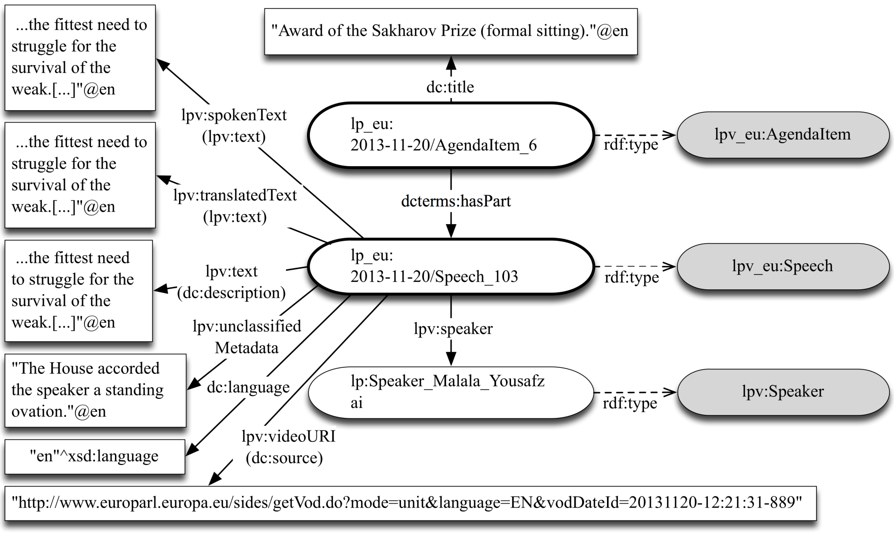

...the fittest need to struggle for the survival of the weak.[...]"@en
lpv:spokenText (lpv:text)
"Award of the Sakharov Prize (formal sitting)."@en
dc:title
lp_eu: 2013-11-20/AgendaItem_6
rdf:type
lpv_eu:AgendaItem
...the fittest need to struggle for the survival of the weak.[...]"@en
lpv:translatedText (lpv:text)
dcterms:hasPart
...the fittest need to struggle for the survival of the weak.[...]"@en
lpv:text (dc:description)
lp_eu: 2013-11-20/Speech_103
rdf:type
lpv_eu:Speech
lpv:unclassified Metadata
lpv:speaker
"The House accorded the speaker a standing ovation."@en
dc:language
lp:Speaker_Malala_Yousafz ai
rdf:type
lpv:Speaker
"en"^xsd:language
lpv:videoURI (dc:source)
"http://www.europarl.europa.eu/sides/getVod.do?mode=unit&language=EN&vodDateId=20131120-12:21:31-889"

Figure 2. The content-level information in the model, exemplified. Parenthesized are the superproperties, where applicable. The coloured boxes denote classes and thick-edged boxes denote entities that reoccur in other figures.

ments about disjointedness and functional properties. With this approach, we follow Van Hage et al. [11]. Second, reference to the EU is avoided where it can restrict reuse - e.g. in the names of the classes and properties, and in most instance URIs - and added where it is deemed necessary to distinguish resources. For instance, 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 Data sets, properties from widely used vocabularies

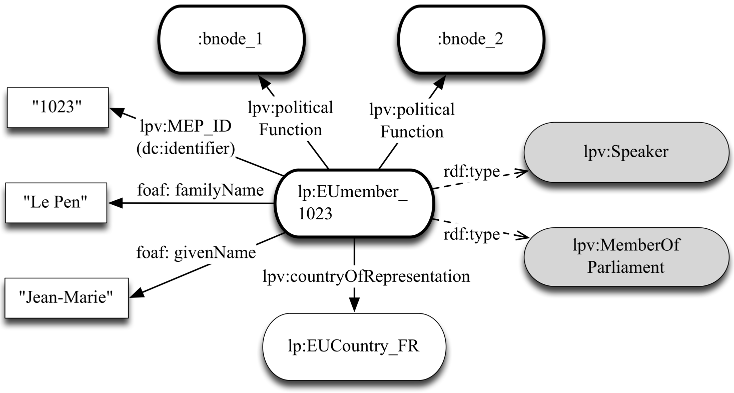


Figure 3. Example representation of a Member of Parliament. Parenthesized are the superproperties, where applicable. The coloured boxes denote classes and thick-edged boxes denote entities that reoccur in other figures.

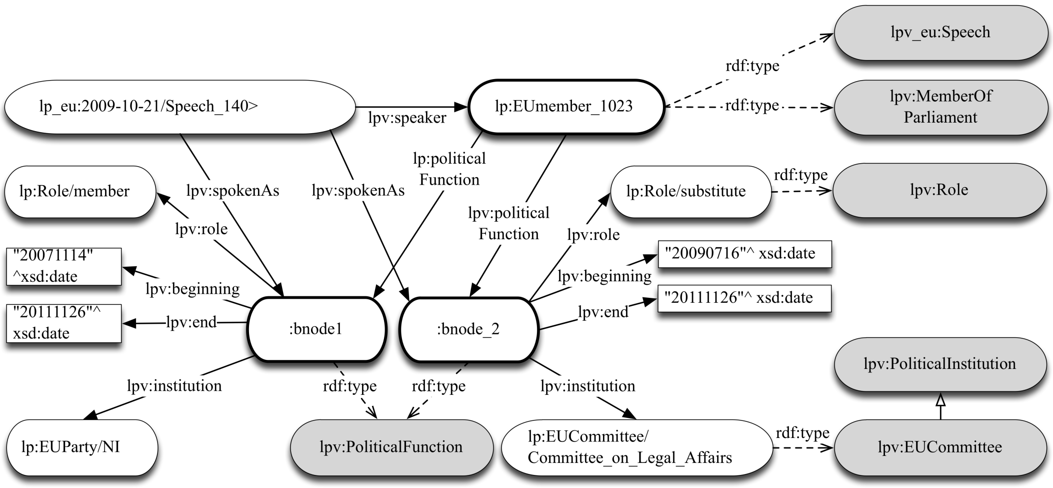


Figure 4. Example representation of the political functions of a Member of Parliament, as defined by a role, institution, and time span. The coloured boxes denote classes and thick-edged boxes denote entities that reoccur in other figures.

are reused or linked to where possible, in particular FOAF[13] for people and Dublin Core[14] for archival resources.

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 (monthly) sessions, session days, agenda items and speeches. The `hasPart` relationship relates higher level events to their parts, and the `isPartOf` relationship does the inverse.

Information about the relative ordering of items is contained in dates and numbering. While these features in theory suffice to find subsequent instances of each kind of event, they require the use of query operators, which might be cumbersome to inexperienced SPARQL users. As a supporting feature we therefore introduced the relation `hasSubsequent` to request follow-up items for speeches and agenda items.

3.3. *Contents of the plenary sessions*

As illustrated in Figure 2, agenda items and speeches are annotated with titles (`title`) and text (`text`),

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

respectively. Besides spoken text, Speech instances have meta-level annotations which are not assigned to any specific category. These unclassifiedMetadata literals are notes of interruptions and applause, as well as 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.

Section 3.4 describes how the model accommodates different translations for these properties, and Section 3.5 how speeches connect to their speaker.

3.4. Languages and translations

All textual data, i.e. titles, speech transcripts, and speech-level annotations, are subject to translation. There is a distinction between the language in which speeches are spoken and the language of display selected on the EP website. 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 it can be easily queried for speech texts in a particular language. Similarly, parallel language-annotated literals exist for unclassifiedMetadata and for the debate-level property title.

The model supports combining the spoken language and the transcription language. The words of the speech in the original language are pointed to by a specific property spokenText to facilitate users who are only interested in original transcripts; for the translated text property translatedText is introduced. Both 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 speakers 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 birth and country of representation are also given, as an xsd:date and a CountryOfRepresentation instance, as well as political functions. Figure 3 displays how the Members of Parliament are modeled.

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 is taken to represent one entry from *ADEP*. It is connected to Role and a PoliticalInstitution instances, 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 PoliticalFunction instances other than a concatenation of their property values. For this reason these instances are represented as blank nodes.

The politicalFunction property is convenient for querying politicians and their functions. However, it is cumbersome for querying for *speeches* by politicians in certain functions. For example, a user meaning to retrieve the speeches by the chair of a given committee might actually retrieve all speeches by people who have ever been chair of that committee, even if they were spoken years after they had that role, if they use the politicalFunction property without a date restriction. To free the user from the burden of defining date restrictions and running these possibly expensive queries, a direct relation, spokenAs, is added between speeches and the momentary political affiliations of the speaker.

3.7. Dataset description and provenance

The content and provenance of the data and vocabulary are described using the void, prov and omv vocabularies. To allow for fine-grained metadata, 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.

The dataset as a whole, as well as each separate graph, has a title and a description of the contents.

To document provenance, the source data for each graph are given, as well as a description of the generation process that underlies the graph. For instance, the graph with the structure of the plenary debates has the homepage of the EP proceedings as its source. The generation of this graph is ascribed to a process of crawling the EP website and translating it to RDF. Contact details of the makers of the dataset are included. For graphs that contain links to other data - which the `void` vocabulary calls *linksets* - the source and target dataset are given. To support access to and use of the data, the used vocabularies are listed and an example resource is provided for all classes. Download links are provided for all graphs, as well as the query endpoint and license information of the dataset as a whole. The metadata are collected in a single graph on the Cliopatria server (see Section 5) and as a turtle file in the *well-known* directory.

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.

4. Links to the LOD cloud

A start has been made with connecting our data to the LOD cloud with links to three external datasets. Additionally, as far as we know, the dataset has been linked to from another source by a third party: the European Union Data Portal[15] provides 887 links between Member of Parliament instances in *LinkedEP* and their named entity resource JRC-Names [10], available through their SPARQL endpoint. For each source connected to, an example is given in Figure 5.

The country entities in our dataset are connected to their counterparts in GeoNames[16], a geographical database. This connection brings in information that could be useful in debate analyses, such as the area, population, languages, neighbouring countries, 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`. Because this task concerned just a few dozen instances, the results were manually verified.

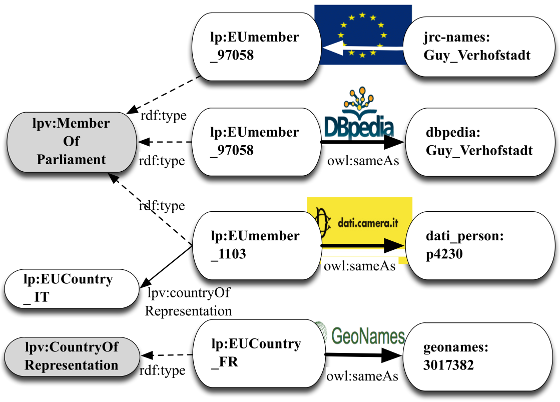


Figure 5. Examples of links to and from *LinkedEP*: inlinks from JRC-Names to *LinkedEP* MEPs, outlinks from *LinkedEP* MEPs to DBpedia and the Italian parliament, and from *LinkedEP* countries to GeoNames.

The Members of Parliament are linked to their entries in DBpedia[17], 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. Section 4.1 details the generation process and the quality of these links.

The politicians representing Italy are in addition matched to the official RDF database[18] 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 Members of Parliament, the mapping results are manually checked for correctness and completeness.

4.1. Method and evaluation

DBpedia contains records for a large part of the MEPs, with small differences between the DBpedia

[15] `https://open-data.europa.eu/en/data`
[16] `http://geonames.org`
[17] `http://dbpedia.org`
[18] `dati.camera.it`

chapters of different languages. DBpedia contains categories and classes related to our topic - such as `Member_of_the_European_Parliament` or `Politician` - but membership of these is too incomplete to be used as input for a matching approach. Instead, a simple bottom-up approach is taken to create links from our dataset to these records. For each MEP in *LinkedEP*, a potential match in DBPedia was stipulated of the form `dbpedia.org/resource/firstname_lastname`. To verify this hypothesized match, it was embedded as the subject of an `ASK` SPARQL query, and accepted if the query returned `true`. To increase the number of matches, not just the English but also the Polish DBpedia was included, which is among the most complete localised DBpedia editions on this topic. In case the Polish (`pl.dbpedia.org/resource/firstname_lastname`) and the English DBpedia both returned a hit, preference was given to the latter. We provide only one link to a DBpedia resource per MEP, as DBpedia contains (albeit incomplete) `owl:sameAs` links between corresponding resources in the localised editions.

For 1455 out of the 1813 Members of Parliament, matching DBpedia instances were found, 196 of which in the Polish DBpedia. On a set of 50 randomly chosen matches, 45 were judged to describe the intended MEP, a precision level of 90 % (± a margin of error of 8%). Of the 5 incorrect matches, 3 refered to a disambiguation page that did list the intended match. Of 50 random Members of Parliament for whom no match was found, only 11 (22 %) were found to have been rightfully overlooked, i.e. lacking a representation in the considered DBpedia sets, considering possible spelling variants. The remaining 39 cases appeared in at least one of the considered DBpedia versions but under a different URI. This was mainly caused by missing diacritic symbols and superfluous additional names in the hypothesized URIs. This small evaluation suggests that the DBpedia linkset has a high precision but a lower recall, which could be improved by a more elaborate use of name variants and inclusion of multiple DBpedia chapters.

5. Access

The described data 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[19]. 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 [20]), which currently redirects to a service hosted at VU University Amsterdam.

With users from a humanities background in mind, the full collection and the individual graphs are described in ISOcat data categories, conform to CLARIN standards [12], the result of which is published on the homepage in a CMDI [3] file.

6. Third party use

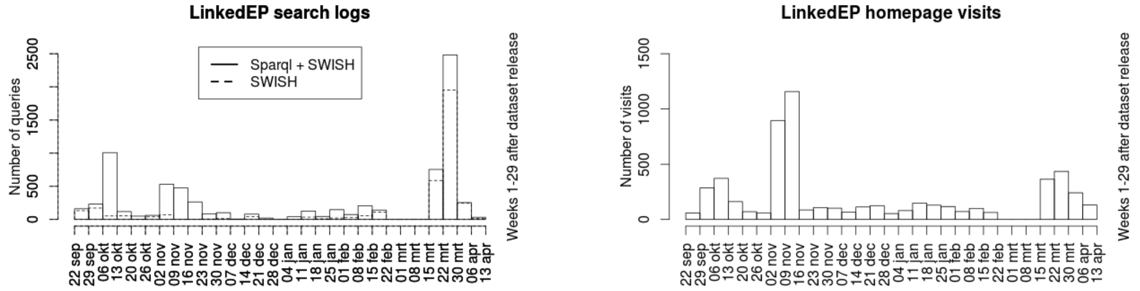
In the 29 weeks following its announcement, the homepage of *LinkedEP* has been visited over 5.5 thousand times and the dataset has been queried through our service about 7.5 thousand times, of which 3,654 times in SWISH/SWI-Prolog and 3,850 times in SPARQL (Figure 6). Manual inspection of the logs reveals that queries containing regular expressions are particularly prevalent, as well as queries with count operations. In total, 1,648 out of the 3,850 SPARQL queries in our logs include a regular expression. 1,600 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 sample of the logged queries.

On several occasions we have had direct contact with users of the *LinkedEP* dataset. Two peaks in query activity that can be seen in Figure 6 each correspond to

[19] `http://cliopatria.swi-prolog.org/`

[20] `http://purl.org`



(a) Number of queries performed, in SPARQL and SWISH/ SWI-Prolog

(b) Number of visits of the homepage

Figure 6. Usage of the dataset in the first 29 weeks since its release. For the period 22 Feb - 15 March, figures are missing due to site and data maintenance.

one-week workshops[21] that were organized by the authors in October 2014 and March 2015 to promote interdisciplinary research around the *LinkedEP* data; another increase in activity corresponds to a Digital Humanties course that was taught at VU University Amsterdam in November 2014, where students queried *LinkedEP*. We will use the queries of known users which match with the typical query patterns found in the logs to illustate the application of *LinkedEP*.

6.1. Example 1: Speaker characteristics and keyword mentions

Displayed below is an example of a query that selects speeches that contain a given keyword. The example searches for speeches with the term 'education' in their English transcript, and returns their identifier, text, and the name of the EU party of the corresponding speaker. This query was formulated to select speeches concerning educational matters in the EU debates, and to explore whether this topic has been addressed more by speakers of certain parties.

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

6.2. Example 2: Trends over time in keyword mentions

The query below requests statistics of the use of a keyword over time. It 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 used by social scientists who were examining how the financial crisis was discussed in the European Parliament. With this query they verified the occurrence of the targeted topic and its correspondence with the economic crises taking place in recent years.

```
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
```

The two given example queries are part of two frequently observed usage patterns of exploring the occurrance of a topic over time or across parties (or other organizational units such as commmittees or countries), and selecting potentially relevant speeches for further close reading. Other examples that we encountered include an exploration of debates about data privacy and transparency, and a study into the use of emotionally charged words by MEPs.

[21] www.talkofeurope.eu/creative-camp-1/ and www.talkofeurope.eu/creativecamp2/

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. During the workshops this occurred several times, among other things to compare word use across parties and countries, and the sentiments expressed across countries. Also, the usage of the links to the *LinkedEP* data provided by the European Union Data Portal cannot be tracked.

7. Discussion and future work

This paper describes the design, generation and use of *LinkedEP*, an RDF translation of the verbatim proceedings of the plenary sessions of the European Parliament, including links to four other datasets. 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. This section provides a discussion of the quality of the dataset and how it can be further improved and extended in future work.

One way to describe the quality of a Linked Dataset is the star system by Berners-Lee [2]. *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. There is considerable room for improvement of the data when it comes to the links to other data, as the EP data are potentially relevant for a wide range of content, including the records of national parliaments and other open government data, encyclopedic sources such as the CIA Factbook, news media archives, etc. The datasets 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 position in the LOD cloud means that it gives access to many other datasets). Future work includes expanding the links to related Open Datasets.

LinkedEP is a close translation of the underlying source data from the website of the EP, implying that the quality of the former depends on the quality of the latter. Fortunately, the online proceedings of the EP are checked by the speakers and are consequently of a high quality. One known flaw is that speeches for which the translation is missing are displayed on the website in their original language, without any warning. Diverging slightly from the aim of a close translation, a heuristic was used to correct these anomalies in the RDF version. However, we know that some incorrect language tags remain. A quantification of and solution to this problem remains future work.

Janowicz et al. [5] proposed quality indicators for vocabularies. Following their rating scheme, the vocabulary described in this paper 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).

The real value of a dataset lies in its uptake. According to the rating system by Janowicz et al., 5-star vocabularies are those that are used by others. While the vocabulary presented here was designed to be used also for other events than the meetings of the European Parliament, to the best of our knowledge this has not happened so far. The data, on the other hand, has been used by third parties. The European Union Data Portal provides links to the Member of Parliament instances of *LinkedEP*. In the seven months following its release, *LinkedEP* has been queried 7,500 times on our servers. An inspection of the server logs revealed that users often retrieve speeches based on the occurrence of one or multiple keywords that signal relevance to the user's topic of study. This usage pattern was confirmed by several use cases observed during two workshops in which researchers from several disciplines worked with *LinkedEP*. The current version of *LinkedEP* does not include any annotations of the content of the speeches. To better support the observed usage pattern, we plan to include topic annotations in future versions, initially using the JEX Eurovoc Indexer [9], a supervised named entity and topic detection tool created for the European Commission.

Acknowledgments

The material described in this paper is based on work carried out in the project Talk of Europe 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] Code of Conduct on Multilingualism Article 10.8. Bureau decision of 16 june 2014. URL `http://www.europarl.europa.eu/pdf/multilinguisme/coc2014_en.pdf`.

[2] Tim Berners-Lee. Linked data: Design issues. `http://www.w3.org/DesignIssues/LinkedData.html`. Accessed: 2014-11-28.

[3] Daan Broeder, Menzo Windhouwer, Dieter Van Uytvanck, Twan Goosen, and Thorsten Trippel. CMDI: a component metadata infrastructure. In *In proceedings of the LREC2012 workshop on Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, page 1, 2012.

[4] Bjørn Høyland, Indraneel Sircar, and Simon Hix. Forum section: an Automated Database of the European Parliament. *European Union Politics*, 10(1):143–152, 2009.

[5] Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, and Charles Vardeman II. Five stars of Linked Data vocabulary use. *Semantic Web*, 5(3):173–176, 2014.

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

[7] Sven-Oliver Proksch and Jonathan B Slapin. Position taking in European Parliament speeches. *British Journal of Political Science*, 40(03):587–611, 2010.

[8] 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. *arXiv preprint cs/0609058*, 2006.

[9] Ralf Steinberger, Mohamed Ebrahim, and Marco Turchi. JRC EuroVoc Indexer JEX - a freely available multi-label categorisation tool. *arXiv preprint arXiv:1309.5223*, 2013.

[10] Ralf Steinberger, Bruno Pouliquen, Mijail Kabadjov, and Erik Van der Goot. JRC-Names: a freely available, highly multilingual named entity resource. *arXiv preprint arXiv:1309.6162*, 2013.

[11] 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.

[12] Tamás Váradi, Steven Krauwer, Peter Wittenburg, Martin Wynne, and Kimmo Koskenniemi. CLARIN: Common language resources and technology infrastructure. In *LREC*, 2008.