POWER - Politics Ontology for Web Entity Retrieval

Silvio Moreira, David Batista, Paula Carvalho, Francisco M. Couto, and Mário J. Silva

University of Lisbon, Faculty of Sciences, Portugal

Abstract. POWER is an ontology of political processes. It is designed for tracking politicians, political organisations and elections, both in mainstream and social media. In social media, these entities (particularly humans) are frequently named by emergent abbreviations, non-standardized acronyms, nicknames, metaphoric expressions and neologisms. Politicians are also frequently mentioned in texts by their roles in the political scene, which may change rapidly over time. This paper describes how POWER was designed for tracking such complex and dynamic setting, with the purpose of making it a key resource to analytics applications mining the media

Key words: social systems, natural language, politics, ontology

1 Introduction

The intensive use of the web as a media channel and the democratisation of publishing tools is leading to a paradigm shift in media production and distribution. On one hand, the line between producers and consumers of news content is fading and, on the other hand, the abundance of content dictates the need for technologies that effectively monitor, gather, analyse and integrate news.

Politics is one the most relevant and prolific topics in the media, but representing and describing this domain pose different challenges on advanced Information Systems. In this context, names are mostly unambiguous, which may suggest that such types of entities would be adequately represented in simple knowledge bases, such as specialized dictionaries. However, political actors and their roles in the political scene change quickly over time, being crucial to create a resource capable of modelling this type of dynamic information. For example, *José Sócrates* is the incumbent Prime Minister of Portugal and the General-Secretary of the Portuguese Socialist Party, since 2005, but he was also the Environment Minister, an ergonym often used in texts to refer this politician. Politicians can also be mentioned by a multiplicity of other forms, which poses additional challenges to their modelling and recognition. For example, they are often mentioned by means of *nicknames* (e.g. *Pinócrates*, instead of Sócrates) and non-standardized *acronyms* (e.g. *MFL; FL*, instead of *Manuela Ferreira Leite*).

Identifying all these mentions in text and mapping them to a unique real-world referent requires up-to-date knowledge of the world and society, robustness to “noise” introduced by metaphorical mentions, neologisms, abbreviations and nicknames, and the capability of performing co-reference resolution.

In this paper, we present POWER, an ontology that formalises the dynamic domain knowledge defining the political landscape, i.e. the political actors (politicians and political organisations), their roles in the political scene, and the relationships and interactions that can be observed among these entities.

Currently, the ontology only covers the Portuguese political environment, but it was designed to describe the different electoral systems at multiple levels, from local to national and supranational . The ontology contains information about the possible forms of mention (birth name, media name(s), ergonym(s), acronym(s) and nickname(s)), enabling their recognition both in mainstream and social media.

The remainder of the paper is organized as follows. Section 2 describes related work. Section 3 presents the conceptual model, identifying the concepts and the relationships used to describe the political landscape. The development and deployment of the ontology are described in Section 4. Section 5 describes the approaches used to populate POWER and provides statistics on the created individuals. The paper concludes by highlighting the future research steps, namely concerning the ontology expansion, enrichment and alignment with other public datasets.

2 Related Work

The success of most natural language applications depends on the adequate recognition of named entities and their normalisation, i.e. identifying the different named-entity mentions in text and mapping them to a unique real-world referent.

Wikipedia has been promoting the creation of large-scale knowledge bases, which can support named entity recognition and normalisation tasks, such as DBpedia [1], YAGO [2], YAGO2 [3]. Typically, these knowledge resources describe a vast immensity of named entities, which are classified into relatively flexible semantic classes, and identify the possible relationships between them.

DBpedia describes more than 1.67 million entities (including 364,000 persons, 462,000 places and 148,000 organisations) and over 672 million RDF triples, extracted from Wikipedia. YAGO’s first version comprises more than 1.7 million entities and 15 million facts. These have been automatically extracted from the category system and from the infoboxes of Wikipedia, and have been combined with taxonomic relations from WordNet [4]. YAGO2 also uses GeoNames [5]. Currently, it describes more than 10 million entities (including 882,534 people, 240,047 organisations and 695,712 locations) and more than 80 million facts about these entities. In YAGO2, the basic triple model was extended to include time and location. The temporal and spatial data were derived from Wikipedia and assigned to semantically meaningful facts.

Freebase[6] is a collaboratively created database of general, structured information intended for public use. It differs from typical ontologies as it does not define a controlled vocabulary; instead, it follows a folksonomy approach allowing the users to create new types to express new relations and properties. This model allows the fast evolution of the knowledge base, but it raises questions about the quality and authority of the data.

Despite of being extremely valuable, these resources do not contain information about a huge number of media personalities, such as most of the Portuguese politicians at local level, mainly because these entities do not have (yet) an article in Wikipedia. Additionally, such type of resources do not systematically represent the multiplicity of forms that each human noun can take in text. As shown by Carvalho et al. [7], proper names only cover 36% of the mentions to human targets, in a collection of comments posted by the readers of a daily newspaper to a set of news articles covering the 2009 Portuguese parliament election debates.

The POWER ontology is being developed to support the recognition of media named entities, particularly politicians, appearing in different genres and types of text, ranging from well structured journalistic articles to highly unstructured and spontaneous opinionated text from user generated content in social media. The development of this ontology follows some of the vocabularies and principles used in GeoNet-PT [8], a geospatial ontology of Portugal, such as the separation between entities and their names.

3 Domain Conceptualisation and Scope

As in any other realm of knowledge, the political reality can be described from countless perspectives. Therefore, the first step in defining this knowledge base was delimiting its scope. The domain of the ontology is politics (using Portuguese post-1974 politics as case-study). In order to define the scope and validate it, we began the conceptualisation phase by posing questions that POWER should be able to answer, such as:

- *How is José Sócrates referred to in the media?*
- *Who is known as* 'Pinócrates' *?*
- *In which mandates has served Cavaco Silva?*
- *Who were the general secretaries in the government in 2000?*
- *Who were the head members of the main political parties in 1995?*
- *Who are the members of the list endorsed by the Socialist Party (Partido Socialista) for the last legislative elections?*

Then, keeping in mind the target and range of applications intended for the ontology we have defined ”political landscape” as a set of relationships between the concepts of:

- Politician (*Mário Soares, Cavaco Silva, Jerónimo de Sousa,...*)

- Political Association (*Partido Socialista, Partido Comunista Português, CDS-Partido Popular,...*)
- Political Institution (*Parliament, government, the president,...*)
- Election (*Legislative, Presidential, Regional,...*)
- Mandate (*Minister of defense in 2001, general secretary of foreign affairs in 2010, ...*)

With these concepts we are able to capture knowledge about: i) **political structure**, the institutions that represent the different political powers (legislative, executive and moderator) at any level of intervention (international, national or local) and the respective geographic scope; ii) **relationships between politicians, political organisations and institutions**, such as offices held in institutions and organisations (mandates), political affiliations and endorsements; iii) **elections**, the process by which politicians attain mandates in a political institution. Depending on the type of election, politicians can run individually for a specific mandate, like the mandate for president of the republic, or run in lists. Elections based on lists of runners define mandates either directly, which means that each list must have candidates for all the offices in the political institution, or with the notion of list proportional representation. In a list proportional representation system each list gets a number of mandates proportional to the number of received votes. Runners for mandates can have party endorsements or run independently and the results of an election determine one or several mandates. Some mandates can be attained by inherency or by appointment from the head of the executive.

One of the main motivations to build POWER was the necessity of a knowledge base to support the tasks of recognizing and resolving named entities in the media. As a result, the ontology model separates the entities (*PoliticalEntity*) from the way(s) they are mentioned (*EntityName*). These may include the full name, acronym or media name (the name by which an entity is normally mentioned in the media).

Figure 1a shows the proposed model represented as UML class diagram depicting political individuals and organisations and Figure 1b shows the additional concepts required for representing electoral processes. These concepts and relations are detailed in Appendix A.

4 Architecture and Deployment

This section details some of the key aspects of the ontology development, namely, the choice of vocabularies for the definition of terms (classes, relations and properties), data flow, deployment and linking of the dataset.

POWER adopts *OWL2* [9], *DCMI* [10] and *SKOS*[11] vocabularies and defines additional, domain specific, terms (see Figure 3 and Table 2). None of the terms has a range or domain defined, to promote their reusability. The definition of all the classes and properties, known as *T-Box* (terms), was generated using *Protegé 4.1.0* [12]. The instances, known as *A-Boxes* (assertions), are collected

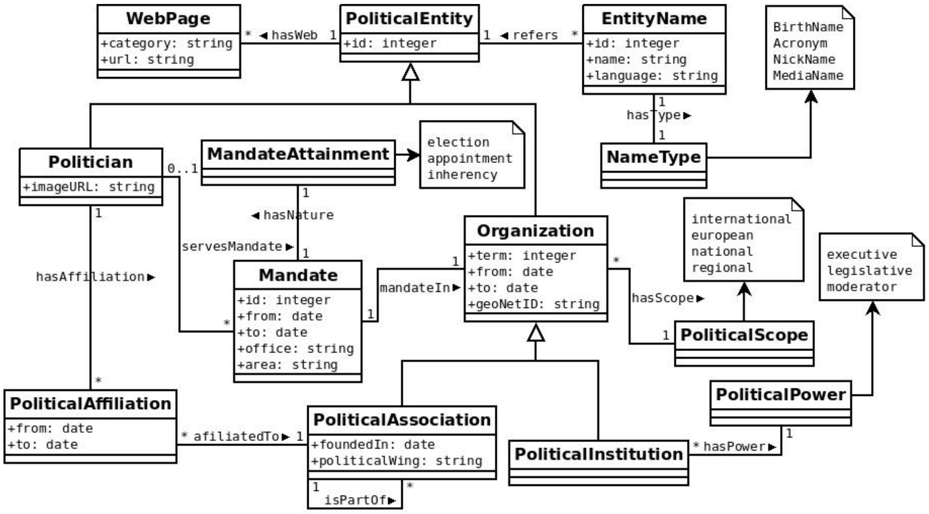

Fig. 1a. Conceptual Model

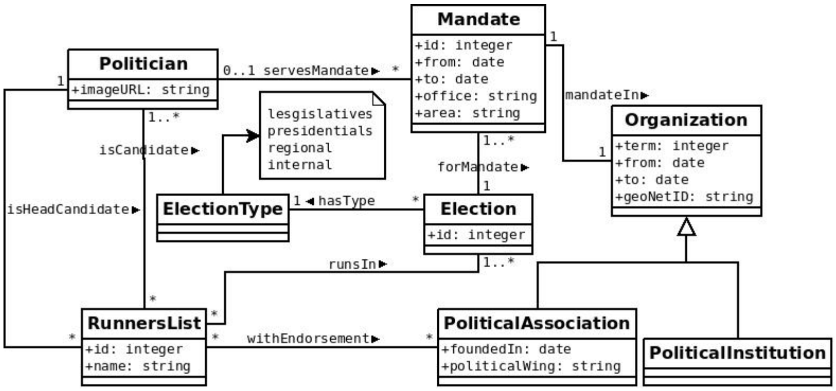

Fig. 1b. Conceptual Model (Elections)

using several information extraction tools, organised under a common framework providing reusable functions. The tools process data from selected databases and websites (see Section 5). The ontology is serialized in the RDF/XML format and deployed under a *Virtuoso* tripletstore [13] . The organisation of the ontology production software enables the enrichment and expansion of the ontology by simply adding new files to the triplestore. The dataset is available via a *SPARQL* endpoint and a web user interface is planned. This data flow is illustrated in Figure 2.

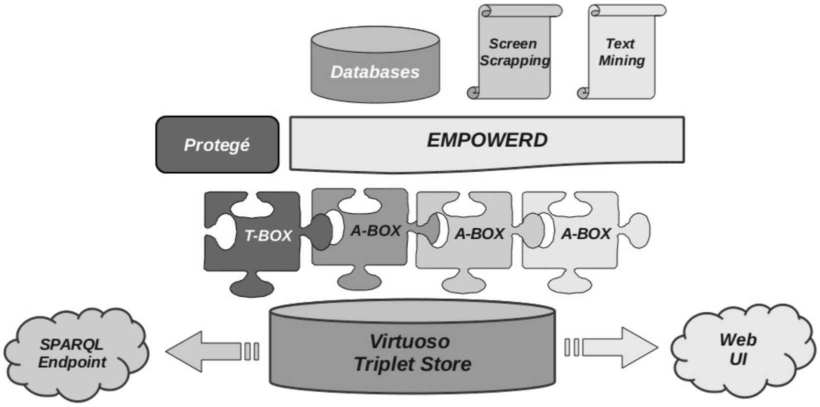

Fig. 2. Data Flow

4.1 Linked Data and Provenance

POWER is publicly available, following the linked data principles outlined in [14], allowing it to serve several purposes and applications. In the linked data vision of the web, a knowledge base is as valuable as the links it provides to additional data. POWER accomplishes this principle in several ways: i) providing links to homepages, Wikipedia articles or other relevant websites trough the relation `power:hasWeb` between political entities and their webpages (`power:WebPage`); ii) mapping the geographical scope of a political institution to a location in the GeoNet-PT ontology [1] using the property `power:geoNetID`; iii) using `skos:extactMatch`, `skos:closeMatch` and `owl:sameAs` relations for the enrichment of POWER individuals and connection to individuals in other datasets.

Another important aspect of the ontology design is the assertion of its data provenance [15]. Providing lineage information allows its consumers to ascertain the authority of the claimed facts and decide whether or not, and to what degree, they should be considered. The provenance metadata for each statement describes its *source*, *creator* (script, tool or person) and *creation date*. These annotations are implemented with terms from the *DCMI* vocabulary using RDF reification [16] (see Figure 4).

5 Populating POWER

For the population of POWER we have developed *EMPOWERD (Enrichment Manager for POWER Dataset)*, a framework that wraps the details of extracting information from various sources and generating RDF statements that assert

[1] This ontology also defines formal mappings to the Yahoo!Planet geographic dataset

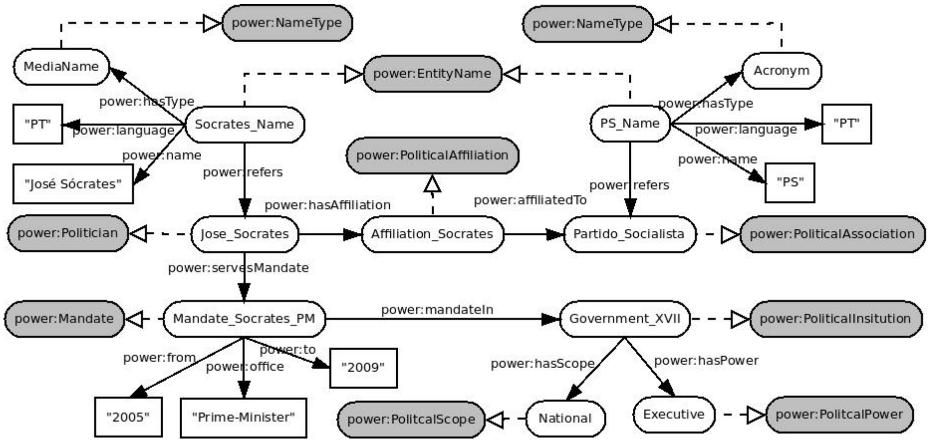

Fig. 3. Political Actors and Mandates

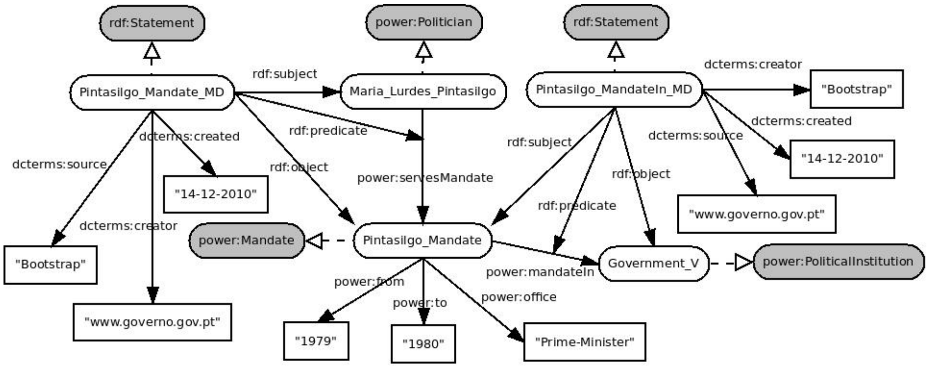

Fig. 4. Provenance Metadata (Reification)

knowledge about the political domain expressed using the POWER vocabulary. This process follows a two-step approach: i) a **bootstrap phase** that creates instances semi-automatically trough a set of scripts that harvest data from selected sources, and ii) an **enrichment phase**, which adds new individuals and their properties using text mining methods and other tools for the extraction of data from relevant sources.

5.1 Bootstrap

The bootstrap step consisted in the development of a set of scripts that *scrap* data from authoritative sources and generate statements about instances of each of the POWER classes. The generated code is then merged into a *bootstrap script* that deploys the ontology with an initial set of assertions.

We have bootstrapped POWER with data from the Portuguese Government [17] and National Elections Commite[18] websites. Using the EMPOWERD framework, we have generated statements about Portuguese parties, politicians, their affiliations and their mandates in the constitutional governments, parliament and presidency of the republic since 1976. Considering that these instances constitute the *backbone* of the ontology, against which new concepts and individuals will be attached in the future, we have selected highly authoritative sources for populating it in this phase. We also manually inspected each of the added instances as an additional normalisation and validation step. However, the obtained dataset at the end of this phase is not rich enough for our purpose. For instance, it does contain the full names of all elected politicians, but not their short names used in the (traditional) media or nicknames used in social media.

5.2 Enrichment

The EMPOWERD framework handles the enrichment of the ontology by providing methods to: i) create new individuals in POWER; ii) add new facts (properties or relations) to an existing individual. The extraction of relevant individual names, properties and facts is based on text mining tools.

Tools and scripts developed for the enrichment of POWER rely on this framework to generate new *A-Box* files augmenting the knowledge base with newly collected facts extracted after scanning the media. Individuals identified in different runs of distinct text mining tools may have similar or identical names, but having similar or identical names may not necessarily imply that they may refer to the same individual. We therefore handle them as distinct, but relate these individuals through SKOS mapping properties, such as `skos:exactMatch` and `skos:closeMatch`.

Initial enrichment of POWER with Portuguese politics personalities was done with the Voxx tool at Sapo, a Portuguese media portal [19]. Voxx provides web services for obtaining the list of mentioned personalities in the media (together with mentioning dates and frequencies). For each personality, Voxx also provides webservices for querying the list of ergonyms associated to each personality over time. We run an EMPOWERD script periodically that scans these personality names and matches them against the POWER personality names loaded in the bootstrap. We assert SKOS matches between POWER entities and Voxx-collected media names through a statistical learning model that takes as features the Jaccard similarity coefficient [20] between the POWER entity full name and the media name, and a set of common heuristics for inferring these associactions. Among others, matching first names, matching last names, two matching family names (common in Portugal).

5.3 Statistics

Following the approach presented in the previous sections we were able to deploy the *backbone* of POWER, with data extracted from authoritative sources, describing:

- 3590 Politicians
- 3043 terms of office in Political Institutions
- 74 Political Associations (18 of them are in fact coalitions)
- 5959 Mandates

This first version of the dataset is publicly available for download as RDF file and may be queried at a SPARQL endpoint [21].

In terms of evaluation, all the political entities were manually validated. We will be assessing the coverage and quality of the POWER ontology based on the performance of the the entity tracking tools using it.

6 Conclusions and Future Work

The work described in this article represents the inception phase of POWER, an ontology for the political domain tailored to aid in the tasks of named entity recognition and resolution. It represents the complexity and dynamic nature of relations between political agents (politicians, political associations and political institutions) over time. This knowledge base is specially useful to support entity tracking, expert finding and question-answer systems. It ensures the recognition of entities referred by several types of mention, such as birth name, acronym or ergonym. We believe it is a valuable resource and could be useful in other contexts like political science, sociology and academic research, thus it is published as a public resource following the guidelines of linked data and can be accessed via SPARQL endpoint [21].

To further expand the value and scope of this resource we will enhance it in two separate, but complementary, directions:

i) **enrichment**, using text mining tools to scan and extract facts from traditional media (newspapers and articles) and from informal and social media (blogs, microblogs, forums and social networks). Defining formal mappings to other public knowledge bases and datasets, such as DBPedia, YAGO, FOAF [22] and Freebase will also increment the knowledge base and allow the inference of new properties and relations of previously ”known” individuals.

ii) **expansion**, by refining and extending the presented model to accommodate other types of media personalities such as sportsmen, actors, influent businessmen or other relevant personalities of the society.

Following these lines of development we hope to achieve a rich dataset capable of describing common knowledge about our society, enabling reasoning and discovery of unknown patterns and relations among social organisations, personalities and events. The development of POWER takes place within the REACTION project [23], an iniative for developing a computational journalism [24] platform for automatic analysis of content (news, blogs, micro-blogs, comments) and implicit and explicit networks in social media.

7 Acknowledgments

We thank João Ramalho for his assistance in the development of the POWER software. This work was partially supported by FCT (Portuguese research funding agency) under grant UTA-Est/MAI/0006/2009 (REACTION project), and scholarships SFRH/BPD/45416/2008 and SFRH/BD/70478/2010. We also thank FCT for its LASIGE multi-annual support.

References

1. DBpedia, `http://wiki.dbpedia.org/About`
2. Suchanek F., Kasneci G., Weikum G.: YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In: Proceedings of the 16th international conference on World Wide Web, pp. 697-706 (2007)
3. Hoffart J., Suchanek F., Berberich K., Weiku G.: YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. Research Paper (2010)
4. WordNet: A lexical database for English, `http://wordnet.princeton.edu/`
5. Geonames Ontology, `http://www.geonames.org/ontology/documentation.html`
6. Freebase: An entity graph of people, places and things, `http://www.freebase.com`
7. Carvalho P., Sarmento L., Teixeira J., Silva M.: Liars and Saviors in a Sentiment Annotated Corpus of Comments to Political Debates. In: 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, June 19-24, 2011 (submitted) (2011)
8. Lopez-Pellicer F., Chaves M., Rodrigues C., Silva M.: Geographic Ontologies Production in GREASE-II Technical Report (2009)
9. OWL2 Vocabulary, `http://www.w3.org/TR/owl2-overview`
10. Dublin Core Metadata Iniative Vocabulary, `http://www.dublincore.org/documents/dcmi-terms/`
11. SKOS Vocabulary, `http://www.w3.org/TR/skos-reference/skos.html`
12. *Protegé* Ontology Editor, `http://protege.stanford.edu`
13. *Virtuoso* Triplet Store, `http://virtuoso.openlinksw.com`
14. Linked Data Design Principles, `http://www.w3.org/DesignIssues/LinkedData`
15. Buneman P., Khanna S., Tan W.: Data Provenance: Some Basic Issues. In: Foundations of Software Technology and Theoretical Computer Science (2000)
16. RDF Semantics, `http://www.w3.org/TR/rdf-mt`
17. Portuguese Government, `http://www.governo.gov.pt`
18. Portuguese Election Committee, `http://www.cne.pt/`
19. Voxx website, `http://voxx.sapo.pt/`
20. Tan P., Steinbach M., Kumar V.: Introduction to Data Mining. Addison-Wesley (2005)
21. POWER SPARQL Endpoint, `http://xldb.di.fc.ul.pt/wiki/POWER-PT_01_SPARQL_endpoint`
22. FOAF Vocabulary, `http://www.foaf-project.org`
23. Reaction Project, `http://xldb.di.fc.ul.pt/wiki/Reaction`
24. Hamilton J., Turner F.: Accountability Through Algorithm: Developing the Field of Computational Journalism. Report, Center for Advanced Study in the Behavioral Sciences at Stanford University and the DeWitt Wallace Center for Media and Democracy (2009)

A Summary of Classes, Relations and Terms

Table 1. POWER classes and relations

Classes	*EntityName*	implements the separation between names and concepts allowing the reference of a concept trough different types of mention
	PoliticalEntity	political individual (politician) or organization
	Politician	a politician
	Organization	an organization that can be a political association or institution. An organization has a political scope to represent organizations in different levels (international, national or local)
	PoliticalAssociation	a political association such as a political party or non-governmental organization
	WebPage	a political entity's web page
	PoliticalAfilliation	affiliation of an individual in a political association
	PoliticalInstitution	institutions that represent the political powers such as the parliament or the government
	Mandate	an office held by an individual in an organization for a period of time
	RunnersList	group of candidates that run together in an election. A group can have endorsements from political associations.
	Election	an election can be legislative, executive, regional or internal
Relations	*refers*	a political entity is referred by an entity name
	hasWeb	a political entity has a webpage
	hasType	an entity name has a type (acronym, birth name, ...)
	hasScope	a political organization has a scope (regional, national, ...)
	hasPower	a political institution has a type of political power (executive, legislative, ...)
	hasNature	the nature of a mandate attainment (election, appointment, ...)
	isPartOf	a political association can be part of coalition
	hasAffiliation	an individual has a political affiliation
	affiliatedTo	an individual is affiliated to a political association
	servesMandate	an individual serves a mandate
	mandateIn	the mandate is served in an political organization
	isCandidate	an individual is candidate in a list
	isHeadCandidate	an individual is head candidate of a list
	runsIn	a candidate list runs for election
	hasType	an election has a type (legislatives, presidentials, ...)
	withEndorsementOf	a candidate list has endorsement of political association
	forMandate	an election is for mandate(s)

Table 2. Power terms

Type	**Term**	**Specializes**
Class	power:EntityName	-
	power:NameType	-
	power:PoliticalEntity	-
	power:WebPage	-
	power:Politician	power:PoliticalEntity
	power:Organization	power:PoliticalEntity
	power:PoliticalAssociation	power:Organization
	power:PoliticalInstitution	power:Organization
	power:PoliticalAfilliation	-
	power:PoliticalScope	-
	power:PoliticalPower	-
	power:Mandate	-
	power:MandateAttainment	-
Object Property	power:refers	-
	power:hasType	-
	power:hasWeb	-
	power:hasScope	-
	power:hasPower	-
	power:isPartOf	-
	power:hasAffiliation	-
	power:affiliatedTo	-
	power:servesMandate	-
	power:mandateIn	-
	power:hasNature	-
	power:isCandidate	-
	power:isHeadCandidate	-
	power:forMandate	-
	power:withEndorsementOf	-
	power:runsIn	-
	dc:source	-
	dc:creator	-
	dc:created	-
	skos:closeMatch	-
	skos:exactMatch	-
DataType Property	power:language	-
	power:id	-
	power:category	-
	power:url	-
	power:term	-
	power:from	-
	power:to	-
	power:geoNetID	-
	power:foundedIn	-
	power:politicalWing	-
	power:office	-
	power:area	-
	power:imageURL	-