Development of an ontology to support information retrieval in the media industry

Ricardo Eito Brun, 12.09.2018

Context of the project

- A Spanish Communication/Media Company.
- This Company is using IR system based on OpenText BRS to index and search textual content and transcriptions from video/TV recordings.
- An improvement program was launched 2 years ago to:
  - Execute the automatic conversion of video/TV recordings to text.
  - Add SW solutions to support names' disambiguation.
  - Improve the IR user interface to support query expansion (adding terms related to those initially proposed by the user.
- For Query expansion, the thought on the possibility of using an ontology.
- Business objectives:
  - Empower end-users (no mediation)
  - Dedicate less time to the manual indexing of the content,
  - Documentalists can dedicate time to enrich the ontology instead of searching content on behalf of the users.

Context of the project

- Typical case:
  - End-user interested in getting multimedia content about what **agents** (e.g. politicians) said about a Specific **topic**.
  - The back-end ontology should provide:
    - The list of all the agents (persons, entities) who belong to that group AND
    - Terms that are related (similar, synonyms, etc.) to the one topic chosen by the user.
  - Then, the end-user can:
    - Launch the query (terms shall be combined with different Boolean operators).
    - Select some of the suggested items to add them to the query.
- The UI builds the query, which shall be directed to the target indexing tool (SOLR, ElasticSearch, etc.)

Context of the project

- The Project involves different tasks and activities.
- UC3M responsibilities are limited to:
  - Design the schema for the ontology, using W3C standards.
  - Collect an initial, representative data set for the different entities in the scope of the demonstrator.
  - Build a prototype to demonstrate the feasibility of the proposed approach, using Semantic Web standards and a semantic/RDF repository.
  - Support the IT department in the selection and testing of the tool.
  - Provide guidance for the future evolution and maintenance of the data sets.
- It is excluded:
  - Development / deployment of the operational SW solution.
  - Activities related to content-indexing.

Context of the project

Two important requirements:

- **Descoping of the IR tool and the conceptual design of the ontology.**
  - The ontologies shall be created using Semantic Web standard languages.
  - Ontologies should be reused with different IR tools used in the future.
  - The data set could be easily reused in other initiatives (not only query expansion).
- **Descoping of the data set and how it is displayed / edited.**
  - Although users proposed some visualization models (e.g. Visual Thesaurus), data should be independent of any specific visualization tool.
  - The data set should be easily maintained using any editing tools.

Methodology.Competence questions

- Project started with the identification of "competence questions"
- A group of 8 specialists in different areas: politics, sports, celebrities, etc.
- Example competence questions:
  - European leaders talking about immigration.
  - People involved in a judiciary case or cases of political corruption.
  - Politicians from a specific political party talking about Catalonia, migration, etc.
  - Celebrities criticising Donald Trump.
  - Spanish celebrities talking against the political measures of the government.
  - Spanish football players playing in the Premier League.
  - Athletes involved in doping scandals.

Methodology. Identification of main Entities

- The analysis of the competence questions led to the identification of two main areas / entities:
  - People and organizations / institutions.
  - Topics about which, those people have made declarations / talked about.
- The first set of entities supports the answers to the « first part » of the competence questions:
  - « *Politicians belonging to this political party talking about…* »
  - « Football players from Chelsea talking about…. »
  - « People imputed in the Gurtel case … »
  - « Relatives of *a specific politician…* »
- The second area (topics) groups set of related terms that could be combined to expand the queries about one topic.
- For these topics, the typical thesaurus structure could be used (RT, NT, BT)

Methodology.Data collection

- Several reliable data sources were provided by the journalists / documentation specialists.
- In most of the cases, data were taken from « official sources ».
- Data were collected / received:
  - In a structured format (Excel sheets)
  - Directly from web pages (screen scrapping) with semi-automated procedure.
  - DBPEDIA was used as a source of data, although some data were not reliable, and the data offered by the same type of entities was not consistent (e.g. Iturraspe vs Adúriz)
  - Spanish Open Data portal (lack of a well-defined organization)
- Problems with data collection:
  - Data are dynamic in nature and subject to change without notice (membership of local governments)
  - This was necessary to explore the possibility of reaching agreements with the main data providers

Methodology. Data conversion and integration

- As data were collected, they were converted into RDF following the rules in the defined schema.
- Data conversion also required the integration and linking of the different entities:
  - References between the list of geographical names and jurisdictions for local governments.
  - References between the list of geographical names and the cities where companies' headquarters are located.
  - References between the list of nationalities and the nationality of sportsmen.
- The use of different variants / alternative names and codes to the same location, nationality, etc., in the different data sources was one of the main issues.
- Cross-check tool was needed to ensure that object properties containing these references pointed to the right value in the "master data lists".

Ontolology Schema – Main classes

- Agents
  - Persons
    - Politicians
    - Sportsmen
      - Football players…
    - Public Administration Staff
  - Organizations
    - Companies
    - Political Parties
    - Governments (local, national, regional)
    - Sport entities
      - National Leagues
      - Sport clubs
- Jurisdiction
  - Countries
  - Regions
  - Cities / Towns
- Situations / Events
  - Sport competitions
  - Judiciary cases
    - Political corruption cases
    - Doping cases
  - Social acts (divorce, marriage, parties…)

Ontolology Schema – Relations

One key requirement of the schema is to keep and maintain data about the relationships between persons, and between persons and institutions.

- W3C Org Schema used as a reference:
  - **Relations between people and organisations**: member, memberOf, headOf, hasMembership / role.
  - **Relations between people**: relativeOf (and sub properties), *hasBusinessWith*.
  - **Relations between organizations**: unitOf, hasUnit.
  - **Additional relations**: worksFor, manages, sub properties of memberOf.
- How to represent a temporary relationship?
  - The Membership class establishes a link between one person and one organization for a specific time period, playing a specific role or function.
  - E.g.: one person belongs to a regional government / parliament for a specific period with a specific role.
  - In some cases it is not feasible to know the start and ending dates, so the relationship must be tagged as « finished » or « expired ».

Ontolology Schema – Relations

How to represent that an entity is part of another entity?

- *For example: if you belong to the regional branch of a political party, can we infer that you belong to that party?*
- *In this case, we opted to keep naming conventions similar to those used in Library Headings and Authority records.*
- *For example: for the Popular Party, different « parts » were recorded for its regional parties, e.g. Popular Party. Andalucía, linked together by a dedicated relation/object property.*

Ontolology Schema – name variants

- Another problem was how to name individuals.
  - URI creation was based on DBPEDIA conventions.

  Example:

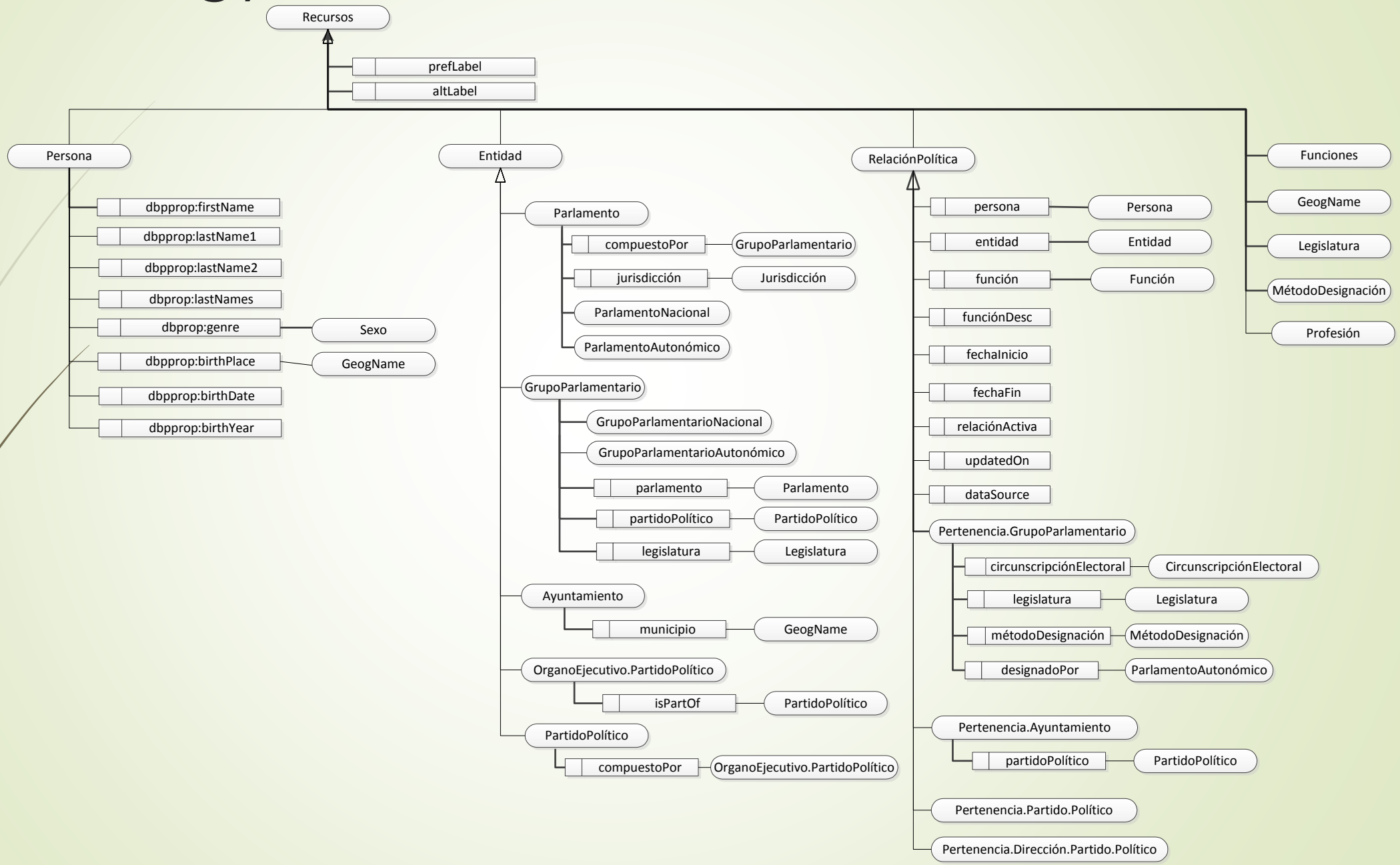
  http://www.uc3m.com/onto/resources/Mariano_Rajoy_Brey

  - To support the use of alternative forms of names when launching the queries, variants for names were recorded in the ontology:

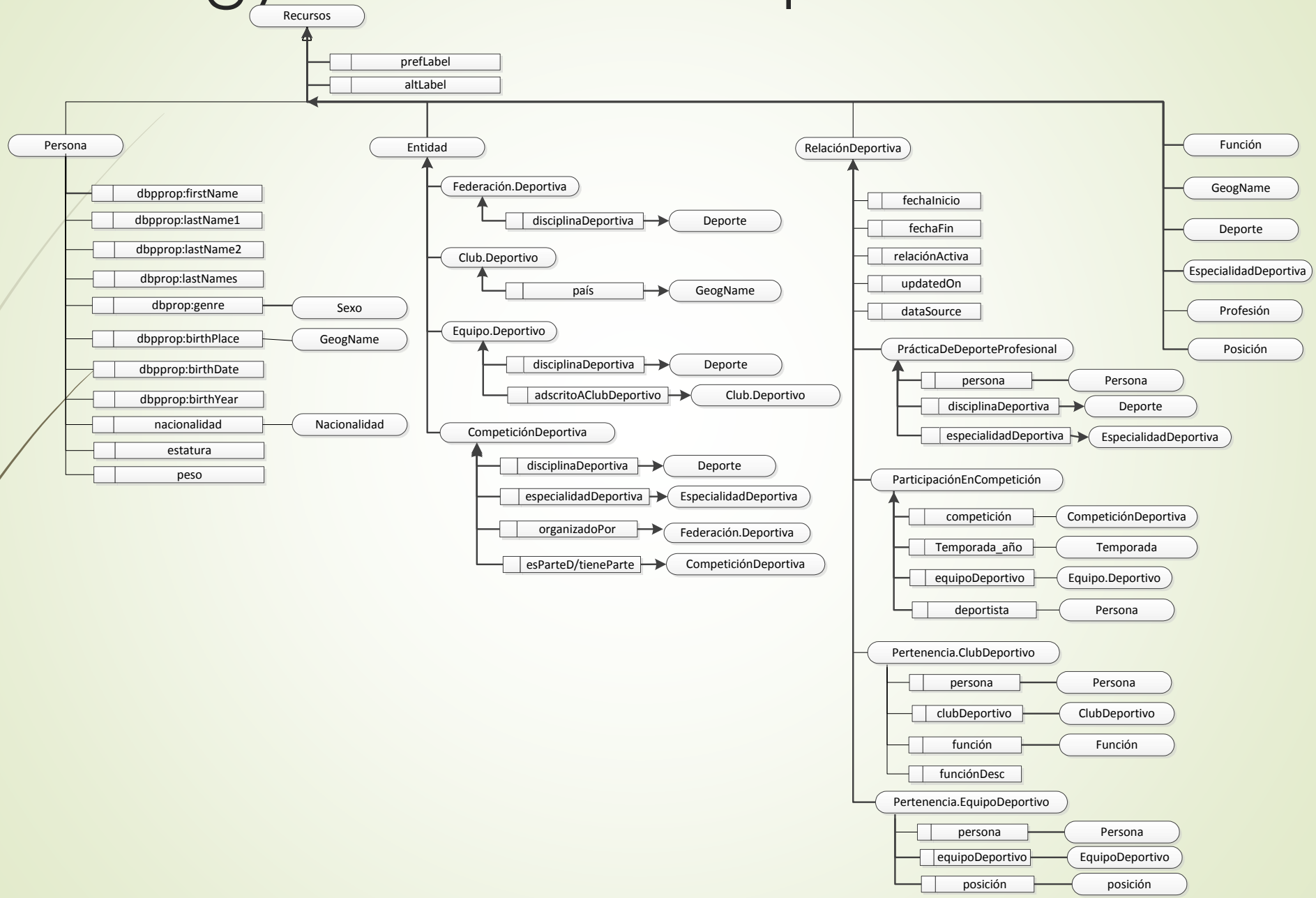
  (« mariano rajoy » OR rajoy OR « rajoy brey »)

  - Individuals have a *@prefLabel* propety and *@altLabel* properties to keep different, feasible names.
  - By default, basic form «First name plus last name » was combined with « First name plus last names ».
  - For institutions, preferred name was based on Library recommendations and Authority records.

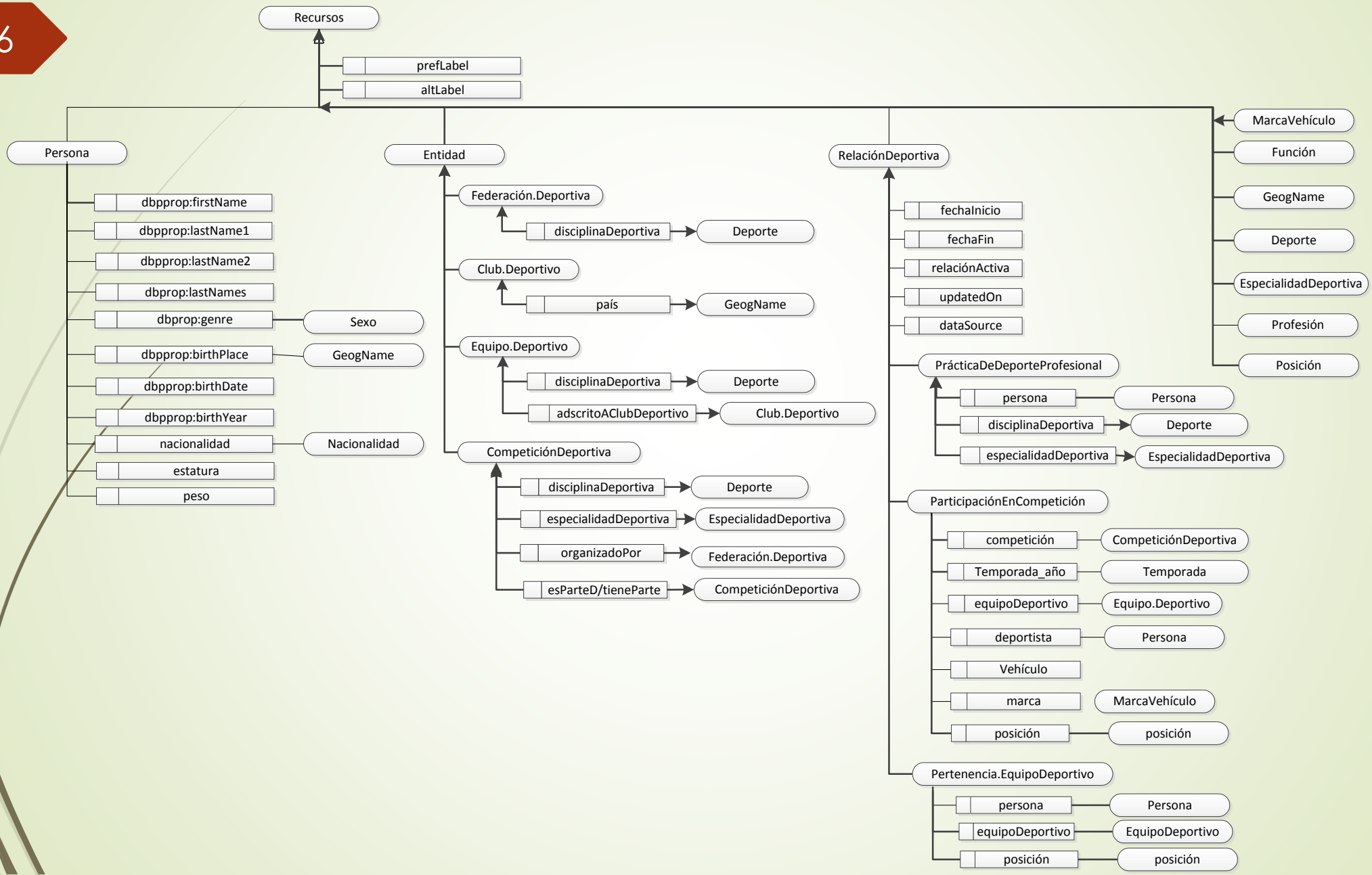
Ontology – Schema for Politics



Ontology – Schema for Sports



Ontology – Schema for Motor Sports



Ontolology Schema – Topics

- Topics are included in the ontology to cover the « second part of the *standard* competence questions ».

  *« Politicians of this party talking about **this topic** »*

- A set of thematic areas (families) was identified.
  - The closed set of terms belonging to these families were represented using SKOS tags.
  - In some cases, additional properties and classes were used to represent more specific relations.

Ontolology Schema – Topics

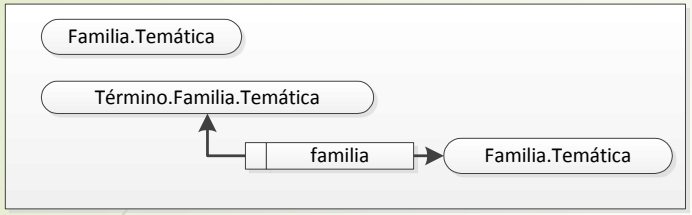
- INMIGRATION
  - MIGRATORY ROUTES
    - TE – Route of Haití
      - *Facet Destination*
      - TR – Bahamas
      - TR – Caiman Island
      - *Facet Means of transport*
      - TR – Ships
      - *Facet Ethnic group /Nationality*
      - TR - Dominicans
    - TE – Traditional Route
    - TE -- …..
    - TE -- Calais-Eurotunnel Route

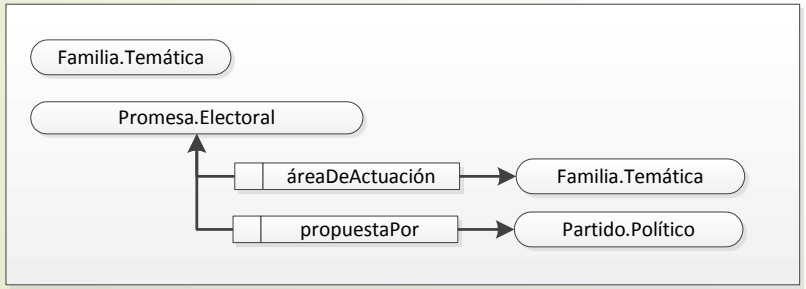
When the end-user selects one topic for the « right hand side » of the query, e.g. INMIGRATOIN, the related terms are proposed as candidates for query expansion.

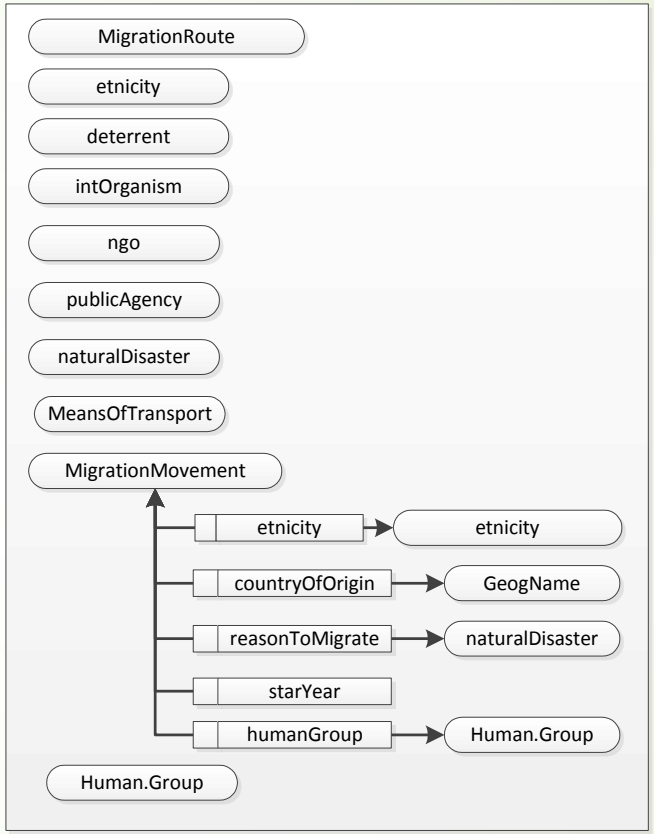
The selection of one of the terms within this « family » shall also add those related terms.

In this example, selecting « INMIGRATION » shall add the terms « MIGRATORY ROUTES », and all their NT terms; « Route of Haití » OR « (Bahamas OR Caiman Island) OR Dominicans ».

Ontology – Schema for Topic Families







Ontology - Data Sets

- Politics
  - Members of the Spanish governments (national, regional and local for cities with more than 50K inhabitants)
  - Deputies and representatives in the Spanish Parliament (National and regional)
  - Executive committees of Spanish political parties: PP, Ciudadanos, PSOE, Podemos, ERC, PNV, EH-Bildu, JuntsPerCatalunya, Compromís, UPN, ANOVA, BNG, En Marea, demòcrates, PDeCAT.
  - Full list of majors of Spanish town hall
  - Parties' electoral promises
- Sport:
  - Football teams and players of main competitions: Spanish Liga Santander, Liga 123, BundesLiga, Premier League, Portuguese Liga, Italian Calcio and French Ligue1.
  - Teams and players who participated in UEFA Champions League and Europa League 2016/17, 2017/18.
  - Basketball, handball teams and players (Spanish competitions)
  - Football Teams and players who participated in UEFA EURO and World Championship.
  - Teams and pilots who participate in Dakar, WCR rallies, Formula1, MOTOGP, Moto2, Moto3.
  - Participants in the last Olympic games.
  - Oher sports: list of golf players, boxers according to Specific rankings.

Ontology - Data Sets

- Business and Economy
  - Companies in the Spanish market Exchange.
  - CEOs and Executive directors of the companies above.
- Other data (Society)
  - Actors, actresses, film directors.
  - Courts, Judges and Prosecutors.
  - Celebrities.
  - Forbes list.
  - Members of royal families.
- Auxiliary
  - Geographical Names
  - Nationalities
  - Occupations
  - Roles and functions.
- Subject (topic families)
  - Inmigration
  - Political Independence movements.
  - Social events.
  - Violence against women
  - Terrorism

Ontology – Validation

- Ontology data are searched using SPARQL.
- A set of basic SPARQL queries were used to tests the design of the ontology with the sample data:
  - To which political parties belong the members of the « Junts pel Sí » group in the Parliament of Catalonia?
  - Who are the women, member of the Socialitst party, with a role in the local government of cities with more than 50K inhabitants?
  - Who are the representatives of the Popular party for the geographical areas of Madrid and Sevilla?
  - Etc.

Ontology – Validation

- Head of Regional Governemtns who belong to a political party.

```
SELECT ?name ?region
WHERE
{
  ?subject skos:prefLabel ?name .
  ?subject rdf:type <http://www.a3media.com/onto/Politic> .
  ?subject org:memberOf <http://www.a3media.com/onto/PP> .
  ?subject org:hasMembership ?member .
  ?member org:role <http://www.a3media.com/onto/Presidente_Gobierno_Autonómico> .
  ?member org:organization ?regionID .
  ?regionID skos:prefLabel ?region .
}
```

- People involved in a judiciary case:

```
SELECT ?name
WHERE
{
  ?subject skos:prefLabel ?name .
  ?subject rdf:type <http://www.a3media.com/onto/Person> .
  ?subject :isImputedIn* <http://www.a3media.com/onto/Gurtel> .
}
```

Ontology – Validation

- Relatives of people involved in corruption

```
SELECT ?imputado ?pareja
WHERE
{
  ?subject skos:prefLabel ?imputado .
  ?subject :isImputedIn* <http://www.a3media.com/onto/Gurtel> .
  ?relative :marriedWith ?subject .
  ?relative skos:prefLabel ?familiar .
}
```

- Responsibilities/roles of a politician:

```
SELECT ?funcionN ?entidad
WHERE
{
  <http://www.a3media.com/onto/Mariano_Rajoy_Brey> org:hasMembership ?cargo .
  ?cargo org:role ?funcion .
  ?funcion skos:prefLabel ?funcionN .
  ?cargo org:organization ?entidad .

}
```

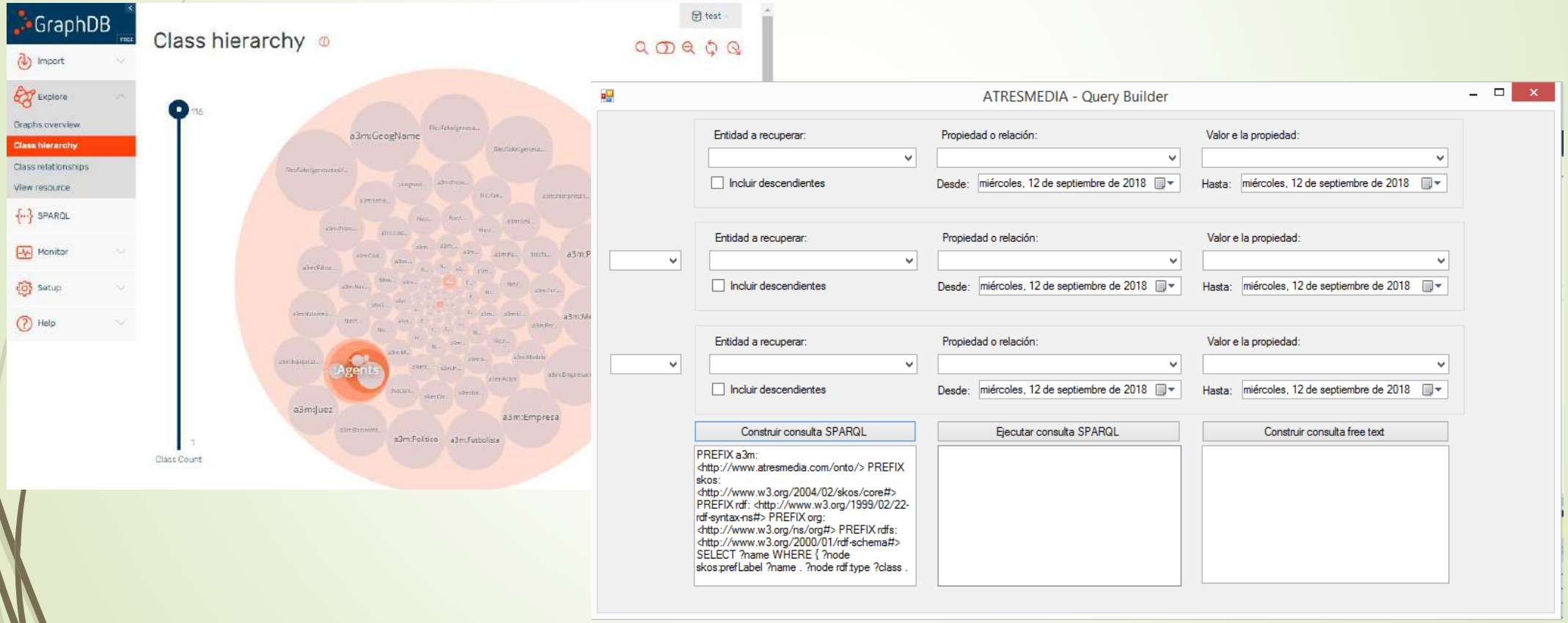
Ontology – Validation

- Political representatives who joined the Socialist group in the Parliament from January 2015.

```
SELECT ?diputado ?ini
WHERE
{
  ?subject skos:prefLabel ?diputado .
  ?subject org:hasMembership ?membership .
  ?membership org:organization
<http://www.a3media.com/onto/PSOE_._Grupo_Parlamentario> .
  ?membership org:dateIn ?ini .
FILTER (?ini > "2015-01-01"^^xsd:date)
}
```

Ontology – Validation

- RDF data files were loaded into a GraphDB database.
- A simple user interface built on Visual C# with dotNetRDF library.



Conclusions

- Approach considered valid and useful by journalists / researchers.
- The use of the ontology saves time and ensures exhaustiveness of the search terms.
- Better queries can be easily constructed, improving recall.
- Main difficulties are related to the maintenance of the ontology.
- Changes need a close monitoring to the selected data sources.
- Some data (e.g. people in judiciary and court cases) are sensitive and require double check before using them.