http://latc-project.eu

D2.1 Report on the Publication of Business-related Datasets

Project GA No.	FP7-256975
Project acronym	LATC
Start date of project	2010-09-01
Document due date	2010-09-30
Actual date of delivery	2010-09-30
Lead Partner	FUB
Reply to	Anja Jentzsch, mail@anjajentzsch.de
Document status	Final

Project GA No.	FP7-256975
Project acronym	LATC
Project full title	Linking Open Data Around The Clock
Dissemination level	PU
Number of pages	17
Task responsible	FUB
Other contributors	INFAI
Author(s)	Anja Jentzsch, Ulrich Zellbeck, Michael Martin
EC Project Officer	Stefano Bertolo
Keywords	Linked Data, Publication, FTS, EURES, CORDIS, EPSO

Table of Contents

1 Executive Summary

This deliverable reports on the publication of business-related produced by European Institutions as Linked Data on the Web. Besides the publication, the datasets were interlinked with corresponding datasets on the Web of Data. Procedures for keeping the data sources up-to-date in respect to original sources have been established and reported on.

2 Financial Transparency System of the European Commission (FTS)

2.1 Executive Summary

The Financial Transparency System of the European Commission (FTS)[1] is a set of information about beneficiaries, which were granted (or otherwise supported) by the commission.
The publication of the FTS as Linked Data was divided into the following steps, which will be covered by this document in more detail:

1. Analysis of original data set
2. Retrieval of original data set
3. Modeling the Linked Data version
4. Publishing FTS as Linked Data
5. Interlinking FTS with existing Linked Data sets

2.2 Analysis of the original dataset

The original dataset is published in HTML, which makes it possible for users to retrieve the information in a human readable way. Users are able to filter the information by:

- the name of the beneficiary
- the year of grant
- the country / territory / geographical zone / postal code
- the amount
- the action type
- the expense type
- the budget line name or number

The correct usage of the filter form leads to a result list containing the information about the beneficiaries. For most of the information listed there, support information (help) is available.

[1] http://ec.europa.eu/beneficiaries/fts/

2.3 Retrieval of the original dataset

In addition to the HTML publication of the FTS it is possible to download the dataset(s) in CSV and XML format. At the moment it is possible to download 4 different datasets containing the information according to the year of grant: 2007, 2008, 2009 and 2010.

In order to create an RDF version out of these datasets we developed a PHP script that parses the CSV versions of these datasets, and processes a set of rules to transform the data into a ntriple RDF notation.

2.4 Modeling the Linked Data version

To enable a more meaningful version of the created RDF, we developed the schema displayed in Figure 1. This schema (Terminology-Box) contains RDF resources (descriptions of classes and properties) used to type and aggregate RDF-resources of the Assertion-Box. All resources (A-Box and T-Box) located in the same model[2]. Schematic elements can be addressed by using the namespace of the T-Box[3] in combination of the local name of the element. Data resources are located in the A-Box namespace[4] .

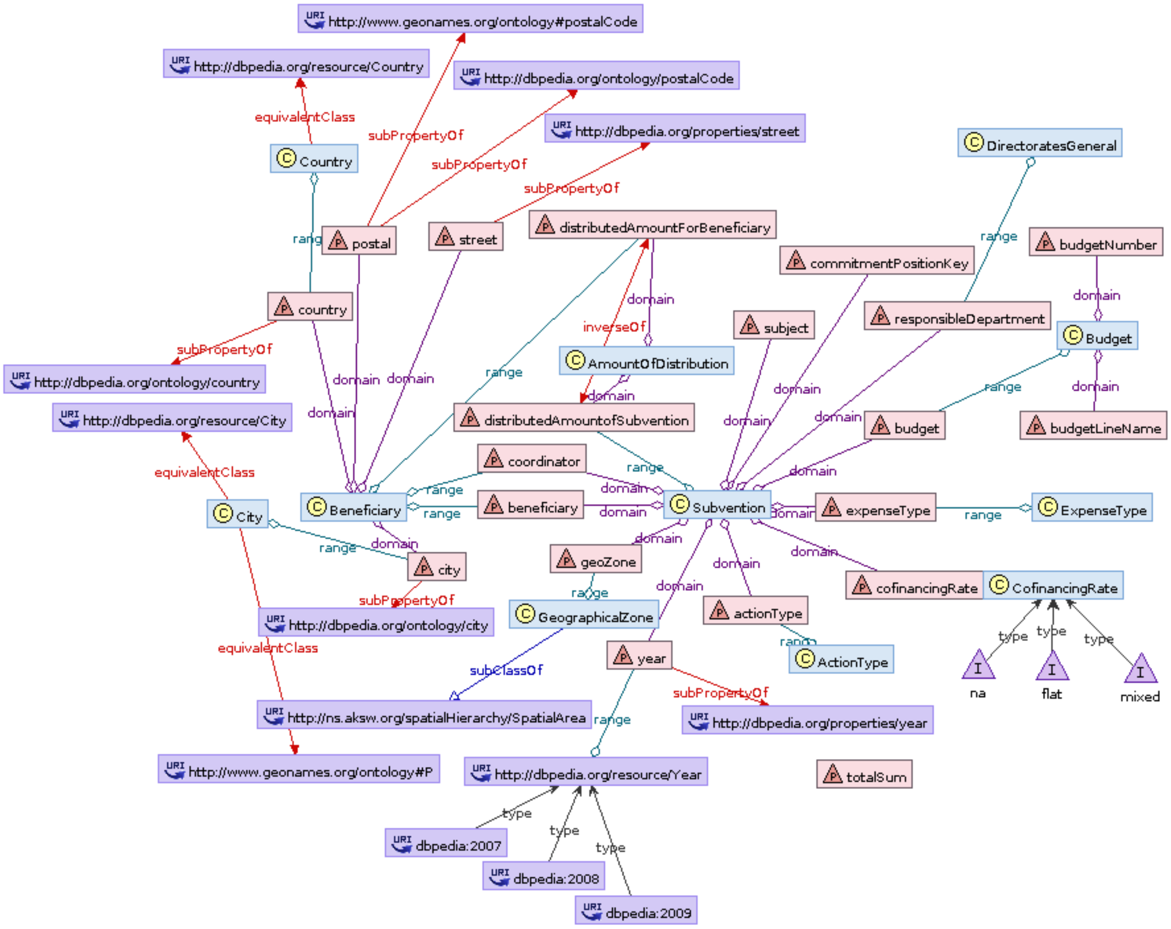

Figure 1: Schema of the FTS RDF version

[2] FTS model-namespace for the RDF-version: http://fintrans.publicdata.eu/ec/
[3] FTS namespace of the T-Box: http://fintrans.publicdata.eu/ec/ontology/
[4] FTS namespace of the A-Box: http://fintrans.publicdata.eu/ec/resource/

The major classes used to type resources are as follows:

- Subvention
- Beneficiary
- ExpenseType
- DirectoratesGeneral
- GeographicalZone
- CofinancingRate
- ActionType
- Budget
- AmountOfDistribution

Some of the resources are typed with classes not located in the FTS-namespace (described in the next section). For instance, classes used to type countries, cities and years are taken from the DBpedia ontology as displayed in Figure 1.

The major properties used to aggregate resources are as follows:

- **beneficiary** This is the name indicated by the beneficiary in the documents submitted to the Commission. Since this is a legal name it may vary from the name known to the general public. Moreover, certain organisations are registered in different countries (for example in the case of cooperation with national subsidiaries), possibly with variations in their name, each of which is a different beneficiary in the eyes of the Commission.
- **year** This is the year in which the amount awarded to the beneficiary has been booked in the budget in the form of a committed amount. The commitment comes after the awarding decision and implies booking in the Commission budget accounts the total amount awarded, which is then paid out on the basis of a legal agreement, usually in different tranches
- **country** The countries/territories are split between "EU" countries and "Other". Among the latter are a few territories which are in fact part of the EU (usually part of France or the UK), but are classified under the "other" heading simply because of the international (SWIFT) code used for financial transfers to relevant beneficiaries.
- **coordinator** In some cases the beneficiary is not the sole end-receiver of EU funds, but the project coordinator responsible for redistributing the funds among the participants. This is typical for research grants which are redistributed by a coordinating scientific institution to scientists and research centres working together on the same project.
- **responsibleDepartment** Grants and procurements are administered by Commission departments called "Directorates General" (DG) in charge of implementing policies and EU funded programmes. The name of the DG can help you identify a policy area. However, be aware that grants and procurements in certain

areas are managed by more than one DG (for example "research" grants which are handed out not only by DG Research but also DG Information Society, DG Enterprise, etc.) and that one DG may manage funds in more than one area. Please note that the names of departments may change over the years, as a result of internal reorganisation.

- **geoZone** For development aid grants paid by EuropeAid, this is the location of the action financed by the grant.
- **budget** The EU budget is structured around titles (two digits: TT), chapters (four digits: TT.CC), articles (six digits: TT.CC.AA) and posts (eight digits: TT.CC.AA.PP).
- **distributedAmountForBeneficiary** referes to an amount resource containing the amount information given to the concrete beneficiary as part of this subvention.

All explanations about these properties are interlinked with the help page[5] of the original publication page of the FTS. Some of the resources expressing cities and countries are interlinked with the help of the spatial hierarchy vocabulary[6] enabling the creation of geographical hierarchies.

The major properties used for textual descriptions about resources are as follows:

- **subject** The subject of a grant or procurement provides general information on the nature and purpose of the expenditure, when available in the system.
- **postal** This is the postal code indicated by the beneficiary in the documents submitted to the Commission.
- **budgetLineName** Budget line name
- **budgetNumber** Budget line number
- **totalSum** This is the amount of the budgetary commitment made in favour of the beneficiary. This is the maximum amount the beneficiary may receive, based on costs incurred. The total amount actually paid out may therefore be smaller.

2.5 Publishing FTS as Linked Data

The RDF version of that dataset was published as linked data. Therefor a virtual host was set up using the same domain as used for the namespace of the FTS RDF version[7]. To enable that SPARQL and Linked Data endpoint we used OntoWiki[8] as publishing and maintenance tool in combination with Virtuoso[9] as storage solution for the RDF model.

In addition to the SPARQL and Linked Data endpoints for machine usage it's also possible to browse the data with the HTML output of OntoWiki, which is more comfortable than the original user interface of FTS.

[5] http://ec.europa.eu/beneficiaries/fts/help_en.htm
[6] http://ns.aksw.org/spatialHierarchy/
[7] http://fintrans.publicdata.eu/
[8] http://ontowiki.net/
[9] http://virtuoso.openlinksw.com

2.6 Interlinking FTS with existing Linked Data sets

At the moment THE RDF-Version of FTS contain the datasets of 2007, 2008 and 2009. The dataset containing data of 2010 was newly published and is at the moment in transformation process.

Table 1 gives an overview of targeted data sets and types as well as the published link counts.

Target Dataset	Type	Link count
DBpedia	overall	199168
DBpedia	cities (instances)	42155
DBpedia	countries (instances)	42366
DBpedia	years (instances)	114636
DBpedia ontology	classes and properties	11

Table 1: Data sets FTS is linked to

2.7 Questions to be answered using the Linked Data version of FTS

There are several advantages of having FTS available as Linked Data, for instance comparing several indicators between EU countries. The interlinking of resources with FTS allows constructing complex queries. At the moment we interlink with spatial and temporal DBpedia resources. In the future, we will increase the number of links to other government datasets like CORDIS and EURES.

3 European Employment Services (EURES)

3.1 Executive Summary

European Employment Services (EURES)[10] is a cooperation network designed to facilitate the free movement of workers within the European Economic Area (Switzerland is also involved). It publishes Job vacancies in 31 European countries. Partners in the network include public employment services, trade union and employers' organisations. The network is coordinated by the European Commission.

The publication of EURES as Linked Data was divided into the following steps, which will be covered in more detail by this document:

1. Analysis of original data set
2. Retrieval of original data set
3. Modeling the Linked Data version
4. Publishing EURES as Linked Data

[10] http://eures.europa.eu/

5. Interlinking EURES with existing Linked Data sets
6. Questions to be answered using the Linked Data version of EURES

3.2 Analysis of the original dataset

EURES is available through its website as HTML only.
The job positions are contained in HTML tables of various variants. These will have to be aligned during the Linked Data publication process.

3.3 Retrieval of the original dataset

In order to publish EURES as Linked Data, the original data set publishers have been approached and asked for an API as well as a dump of EURES data in April 2011.
Since neither of these are available, the Linked Data version of EURES is scraped from the HTML on the EURES website. An initial crawl was done in May 2011. Further updates can be done by scraping new job position pages non-intrusively by scraping daily only the new pages.

3.4 Modeling the Linked Data version

When modeling the EURES ontology, existing ontologies like the Knowledge Nets[11] ontology have been analyzed and integrated.
The main concepts in the EURES data set are: jobs, organizations, contact persons, skills, languages, countries, regions, cities.
The EURES ontology contains vocabulary mappings to the following vocabularies: Knowledge Nets, LEXVO, and DBpedia.

3.5 Publishing EURES as Linked Data

EURES has been published as Linked Data using different tools.
D2R Server[12] is being used to hold a database version of the whole data set and to allow fast data integration. Using a D2RQ mapping for mapping the EURES database to RDF, the data set can be dumped as RDF as well as browsed.
In order to offer a high-performance SPARQL endpoint, the data is loaded into Fuseki[13].
The Silk Link Discovery Framework[14] is used to interlink EURES with existing Linked Data sets.
The publication process contained the following steps:
The EURES data is scraped using ScraperWiki[15] from HTML pages and then imported into a relational database. Using D2RQ, a mapping from the database schema to the EURES ontology was created. The dataset is dumped as RDF using D2R Server and loaded into the Fuseki endpoint. Several Silk link specifications has been defined which specify the link conditions for EURES to other Linked Data sets. The resulting links are imported into the database as well as the SPARQL endpoint.

[11] http://wissensnetze.ag-nbi.de/
[12] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/
[13] http://openjena.org/wiki/Fuseki
[14] http://www4.wiwiss.fu-berlin.de/bizer/silk/
[15] http://scraperwiki.com/

D2R Server allows for browsing parts of the Linked Data version of EURES:
http://www4.wiwiss.fu-berlin.de/eures/
The Linked Data version of EURES consists of 7,011,899 RDF triples.

3.6 Interlinking EURES with existing Linked Data sets

EURES has a huge thematic overlap with various other data sets. Amongst geographical data, it contains information on organizations.

Interlinking the EURES Linked Data set with other Linked Data sets allows for extensive querying (see also section 2.7).

So far 3,236 links have been created using either Silk or a manual mapping.

Table 2 gives an overview of targeted data sets and types as well as the published link counts.

Target Dataset	Type	Link count
DBpedia	cities	1,382
DBpedia	regions	326
DBpedia	countries	246
DBpedia	languages	184
DBpedia	currencies	8
LEXVO	languages	184
Eurostat FUB	regions	300
Eurostat Riese	regions	300
Eurostat Ontology Central	regions	300
Geonames		
CORDIS		
FTS		

Table 2: Data sets EURES is linked to

3.7 Questions to be answered using the Linked Data version of EURES

The advantages of having EURES available as Linked Data are widespread.

The EURES ontology ensures a clean and consistent view on the EURES data. Data is only published if it matches the schema requirements.

Furthermore, having EURES interlinked with other Linked Data sets facilities a rich search for different user groups. We will be focusing on use cases for employees in this section.

When searching for a suitable job, different factors might need to be taken into consideration.

1. Employees may want to compare outside work life standards for different jobs, e.g. weather or social structures. Geographical, touristic as well as economic information on the job region can be found on different data sets, including DBpedia, Geonames, LinkedGeoData etc.
2. Employees may want to compare salaries as well as contrast them with the existing life standard in the job region. Eurostat as well as DBpedia offer monthly labor costs as well as the GDP which can be used to compare salaries in EURES.
3. Statistical (CENSUS) data can be included in the job decision process by using DBpedia, US Census or OCS.
4. For families information on the quality of regional schools might be of interest. There will be an All Ireland's data set on these by the end of June 2011 which can be accessed through DBpedia.
5. The classification of language and education skills in EURES enables a fine grained search for suitable jobs. DBpedia can be used for descriptional texts on the different skills or levels.

4 Community Research and Development Information Service for Science, Research and Development (CORDIS)

4.1 Executive Summary

Community Research and Development Information Service for Science, Research and Development (CORDIS)[16] is the official source of the information of the EU's seventh framework proposal (FP7). Its database includes information of previous framework proposal as well as information about related companies and projects.
The publication of CORDIS as Linked Data was divided into the following steps, which will be covered in more detail by this document:

- Analysis of original data set
- Retrieval of original data set
- Modeling the Linked Data version
- Publishing CORDIS as Linked Data
- Interlinking CORDIS with existing Linked Data sets
- Questions to be answered using the Linked Data version of CORDIS

4.2 Analysis of the original dataset

The original dataset is published on the CORDIS website[17] in HTML. Access to the information is given by a search interface which is currently renewed and published as beta[18]. There, the user can search for:

[16] http://cordis.europa.eu/
[17] http://cordis.europa.eu/home_en.html
[18] http://cordis.europa.eu/newsearch/index.cfm?page=simpleSearch

- Programmes
- Subjects
- Document types
- Countries

First of all, the user has to enter a specific query which gives him a result list.
As a second step he can either choose one of the results or constrain the results by choosing one of the topics on the left.

4.3 Retrieval of the original dataset

The base of the dataset was given by a database dump from CORDIS which was sent to us on a CD-Rom. After creating a database schema the data was converted from Microsoft Access into MySQL with a PHP script. Around this base another PHP script updates the data using the RSS feed of CORDIS.

4.4 Modeling the Linked Data version

The CORDIS dataset is originally available in a relational database with. The database schema is depicted in Figure 2.

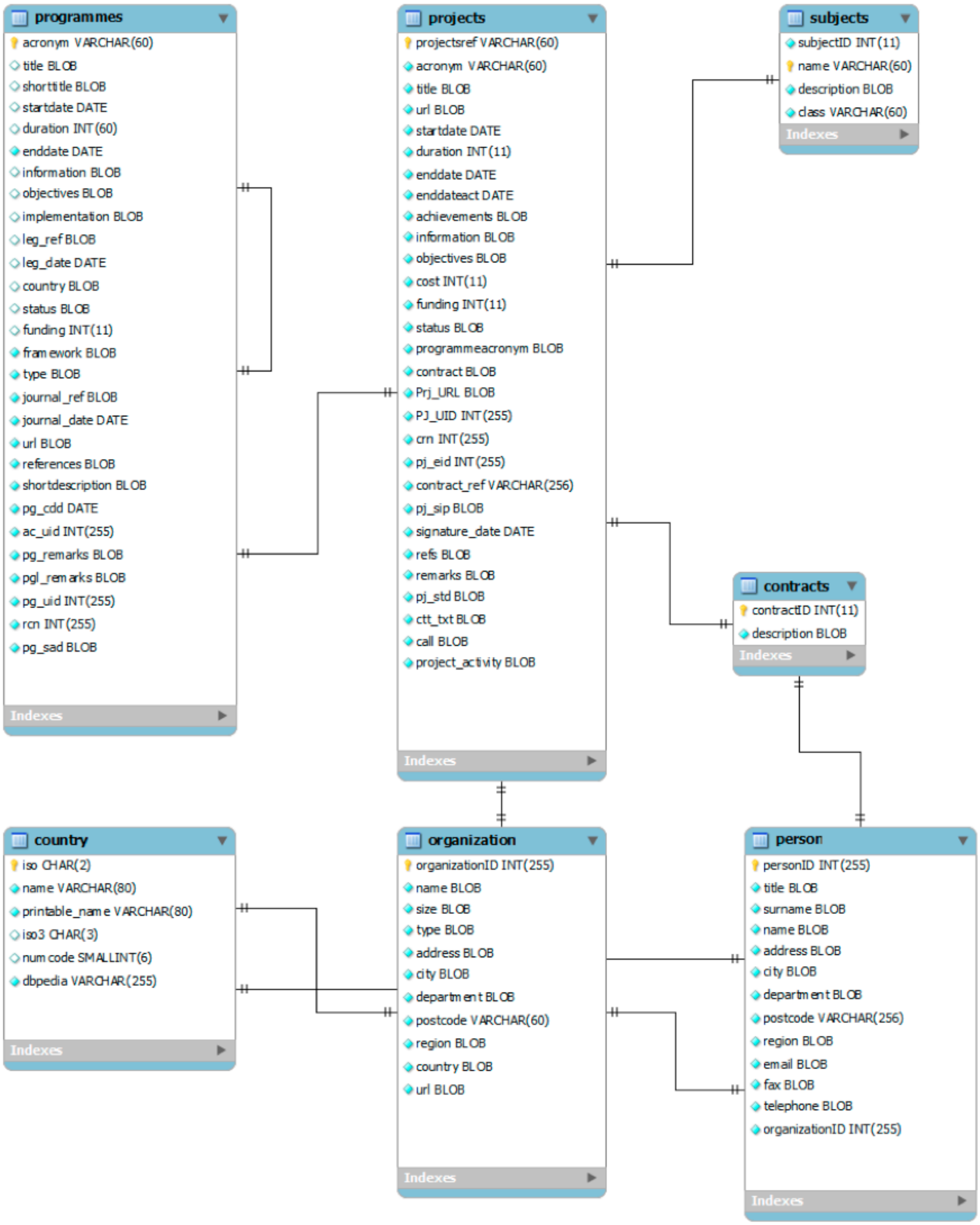

Figure 2: CORDIS database schema

The main classes are:

- projects
- programmes
- organization
- person
- country

The main focus of the dataset is on describing projects. A unique ID for a project is its record control number which is set by CORDIS.
The project has a project coordinator (of class person) who himself is related to an organization.
Other participants involved are listed with their company and, if available, with their department. For categorizing purpose at least one subject is assigned. The dataset also provides information about the start date, end date and internal categories like project ID and reference number.

As stated before, projects are connected to companies by their employees. As a result, companies do not have an unique address, as they are set to their employees which may work in different departments on different locations.

A project is also connected to its corresponding programme, i.e. Framework Programme 7 which itself is related to other programmes either as following / previous programme or as umbrella programme.

For modeling the language as Linked Data a new vocabulary has been created to describe CORIDS specific items (projects, programmes, employer and employee). Other vocabularies used are FOAF.

4.5 Publishing CORDIS as Linked Data

Various tools have been used to publish CORDIS as Linked Data.
As latest link on the chain of publication, a D2R Server[19] is used for publishing the relational database. D2R Server allows navigating the content in RDF as well as in HTML:

http://www4.wiwiss.fu-berlin.de/cordis/

Using a D2RQ mapping for mapping the CORDIS database to RDF, the data set can be dumped as RDF.
A SPARQL endpoint[20] is provided to query the dataset for users as well as for machines.
For generating links to other Data sets, Silk Framework[21] has been used. Several Silk link specifications have been created which specify the link conditions for CORDIS to other Linked Data sets. The resulting links are imported into the database as well as the SPARQL endpoint.

4.6 Interlinking CORDIS with existing Linked Data sets

Table 3 gives an overview of targeted data sets and types as well as the published link counts.

[19] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/
[20] http://www4.wiwiss.fu-berlin.de/cordis/sparql
[21] http://www4.wiwiss.fu-berlin.de/bizer/silk/

Target Dataset	Type	Link count
DBpedia	organization	283
DBpedia	project	31
DBpedia	country	239

Table 3: Data sets CORDIS is linked to

4.7 Questions to be answered using the Linked Data version of CORDIS

As Linked Data could be queried like a normal database, the following answers could be given:

- Which are the projects an organization is participant?
- Which projects have had lower costs than funding provided by the EU?
- Which projects have a funding greater than 1.000.000?
- Which projects do have more than 10 participants?
- Which partners cooperate more than x times with another?
    - What are the project details / reasons for that?
- Which projects do have participants from more than 5 different countries?
- How many projects does Programme "FP6-FOOD" have?

By interlinking to DBpedia:

- More information about the participating organization

By interlinking to FTS:

- Does the participating organization get any other beneficiaries from the European Union?

5 European Personnel Selection Office (EPSO)

5.1 Executive Summary

An investigation of the European Personnel Selection Office (EPSO)[22] website has shown that the information that could be extracted from the website is not informative enough. There is essential information missing, e.g. a detailed job description or the employing organization (except for temporary jobs).
Also, there are overall 500 jobs, which would result in a very small dataset.

[22] http://europa.eu/epso/

5.2 Analysis of the original dataset

EPSO is available through its website as HTML only.
The jobs are categorized in 3 categories:

5.2.1 Temporary jobs (60)

http://europa.eu/epso/apply/today/temporary_en.htm

These jobs are no calls but offer several jobs at EU institutions.
Information provided: Type, position, reference, grade, deadline, location, link to job proposition

5.2.2 Ongoing competitions (100)

e.g.: http://europa.eu/epso/apply/on_going_compet/adm/index_en.htm

These calls are currently on going.
Information on these jobs vary widely over the jobs but are overall very little. Most of the information is provided in PDFs (Official journals of the EU), which are hard to parse.
Information provided on the website: Competition number, publication date, closing date (number of applications, publication of test results, date of assessment center).
What's missing: More information on the job, at which institution is the job.

5.2.3 Closed competitions (330):

http://europa.eu/epso/success/archives/index_en.htm

Closed calls of competitions.
Very little information, as there is only information about: Competition No, Publication date, grade, title, journal link, dates.
More information is provided in PDFs (Official journals of the EU).

The official journals are provided by EUR-LEX. There might be an access to XML files, but it is charged. (See http://eur-lex.europa.eu/en/editorial/legal_notice.htm)