Designing semantic Application Programming Interfaces for open government data

Citation for published version:
Degbelo, A, Trilles, S, Kray, C, Bhattacharya, D, Schiestel, N, Wissing, J & Granell, C 2016, 'Designing semantic Application Programming Interfaces for open government data', *eJournal of eDemocracy and Open Government*, vol. 8, no. 2, pp. 21-58. https://doi.org/10.29379/jedem.v8i2.420

Digital Object Identifier (DOI):
10.29379/jedem.v8i2.420

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
eJournal of eDemocracy and Open Government

JeDEM 8(2): 21-58, 2016
ISSN 2075-9517
http://www.jedem.org

Designing Semantic Application Programming Interfaces for Open Government Data

Auriol Degbelo*, Sergio Trilles**, Christian Kray*, Devanjan Bhattacharya***, Nicholas Schiestel*, Jonas Wissing*, Carlos Granell**

**Institute for Geoinformatics, University of Münster,*
degbelo|c.kray|nicholas.schiestel|jonas.wissing@uni-muenster.de

***Institute of New Imaging Technologies, Universitat Jaume I, Spain*
strilles|carlos.granell@uni-muenster.de

****NOVA Information Management School (NOVAIMS), Universidade Nova de Lisboa, Portugal*
dbhattacharya@isegi.unl.pt

Abstract: Many countries currently maintain a national data catalog, which provides access to the available datasets – sometimes via an Application Programming Interface (API). These APIs play a crucial role in realizing the benefits of open data as they are the means by which data is discovered and accessed by applications that make use of it. This article proposes semantic APIs as a way of improving access to open data. A semantic API helps to retrieve datasets according to their type (e.g., sensor, climate, finance), and facilitates reasoning about and learning from data. The article examines categories of open datasets from 40 European open data catalogs to gather some insights into types of datasets which should be considered while building semantic APIs for open government data. The results show that the probability of inter-country agreement between open data catalogs is less than 30 percent, and that few categories stand out as candidates for a transnational semantic API. They stress the need for coordination - at the local, regional, and national level - between data providers of Germany, France, Spain, and the United Kingdom.

Keywords: open data, semantic API, user requirements

Acknowledgement: The authors gratefully acknowledge funding from the European Union through the GEO-C project (H2020-MSCA-ITN-2014, Grant Agreement Number 642332, http://www.geo-c.eu/). Carlos Granell has been funded by the Ramón y Cajal Programme (grant number RYC-2014-16913). Sergio Trilles has been funded by the postdoctoral programme Vali+d (GVA) (grant number APOSTD/2016/058). We also thank the two anonymous reviewers for their constructive comments.

1. Introduction

Various factors have contributed to the increased availability of Open Data, including national and international legislation, requests for transparency and hopes for enabling new services. Cities are a particular 'hotbed' for producing data and consuming it, for example, a lot of sensor data is produced in a smart city (see Hancke, Silva, & Hancke Jr., 2013; Lecue, Kotoulas, & Mac Aonghusa, 2012), and numerous apps have been developed that make use of city data (see Lee, Almirall, & Wareham, 2016). While many have pointed out the high potential for open data in terms of better participation, increased transparency, new services and better use of ressources (e.g., Fechner & Kray, 2014; Hartog, Mulder, Spée, Visser, & Gribnau, 2014; Janssen, Charalabidis, & Zuiderwijk, 2012; Masip-Bruin, Guang-Jie, Serral-Gracià, & Yannuzzi, 2013; Ojo, Curry, & Zeleti, 2015), there are still several challenges that need to be tackled. These include managing the vast amount (and high bandwidth) of data being produced (Chen, Mao and Liu, 2014), the heterogeneity of the data (Janssen, Charalabidis and Zuiderwijk, 2012; Masip-Bruin *et al.*, 2013; d'Aquin *et al.*, 2014), data quality and recency (Janssen, Charalabidis and Zuiderwijk, 2012; Masip-Bruin *et al.*, 2013), coordination mechanisms at the technical and political levels (Lee, Almirall and Wareham, 2016), as well as privacy issues (Janssen, Charalabidis and Zuiderwijk, 2012; Chen, Mao and Liu, 2014), to name but a few.

One important challenge when working with open data is to enable machines (or applications) to re-use it. This aspect is not considered in many definitions of open data. For example, a commonly used definition says that open[1] data is data which is freely available and shareable online, without charge (The World Wide Web Foundation, 2015). While this definition ensures that *humans* can access and inspect the data, it says little about the use by *machines*. Two key aspects must be considered while enabling open data reuse by machines: The provision of open data in a structured (also referred to as *machine-readable*) data format such as Comma Separated Values (CSV), Extensible Markup Language (XML) or Resource Description Framework (RDF); and the querying of the data by machines. The first aspect is addressed by the five stars deployment scheme for Linked Data introduced by Berners-Lee (Berners-Lee, 2006). Essentially, the greater the degree of structure of the data, the easier its further processing by machines. Regarding the second aspect, the common way to enable open data querying over the Web for machines is to provide an Application Programming Interface (API) that describes the functions an application can execute to access an open data repository, the parameters that are expected and the kind of results returned. APIs were mentioned as one of the 12 critical factors for success of open data initiatives in (Susha *et al.*, 2015). This notwithstanding, the recent edition of the Open Data Barometer pointed out that "*More elaborated APIs that facilitate access to data are still very rare among government data*" (The World Wide Web Foundation, 2015). This article aims at initiating a discussion on the required components of such APIs. More specifically, the work focuses on the types of data categories that more elaborated APIs should offer, a topic which has received little attention in the literature so far.

[1] The world 'open' is associated with many different connotations (for a recent discussion see Pomerantz & Peek, 2016). For the purposes of this article, the definition of open data as "data which is freely available and shareable online, without charge" (The World Wide Web Foundation, 2015) is adopted.

APIs form an essential component of the World Wide Web. For instance, ProgrammableWeb.com, the *"Web's defacto journal of the API economy"*[2] lists more than 15,000 Internet-based APIs as of June 30, 2016. One of the main motivations for building APIs is to improve programmer's productivity by enabling code reuse instead of code writing from scratch (see Stylos & Myers, 2007). Note that improving programmers' productivity and enabling data access to machines are, in the context of API development, two sides of the same coin. Machines or software agents use APIs to autonomously access data stored in external repositories, but it is the programmer who tells these machines or software agents which APIs to use, and how they can best access the data. Programming is an important aspect of API development, but it is not the most important of it. As Henning (2009) pointed out, *"an API is not about programming, data structures, or algorithms - an API is a user interface"*. Thus, designing an API essentially boils down to providing a useful interface by which machines can access resources (e.g., datasets). API design is equally providing a useful interface by which programmers can access resources for the purpose of application development. API design is thus, like any other design problem in the context of information sharing, best viewed as a "human-machine-human" conversation problem (see (Scheider and Kuhn, 2015) for a detailed discussion of the "human-machine-human" perspective on information sharing).

The salient peculiarity of API design is the *"stakes of getting the design right in the first place"* (Myers and Stylos, 2016). Since APIs are used by many applications after they have been developed, any change in their interface incurs thousands of broken apps, and therefore considerable loss of time and money[3]. Stylos and Myers (2007) list three different stakeholders of an API: **API designers** (whose goals are to maximize the adoption of an API and minimize its support costs), **API users** (willing to write error-free programs, and use APIs that many other programmers use), and **consumers of products built with the API**. The work in this article is mostly relevant to API designers and users (i.e., programmers). Citizens (as consumers of Apps built with these APIs) will benefit indirectly from semantic APIs for exiting government data.

RESTful APIs are one of the major types of APIs nowadays, covering about 60% of the API market[4]. In a nutshell, a RESTful API provides the opportunity to retrieve resources via the methods from the Hypertext Transfer Protocol (HTTP). Richardson and Amundsen (2013) recommend using the HTTP methods GET, POST, PUT, DELETE, and PATCH for Web API development. The first step of designing a RESTful API is, following (Richardson and Amundsen, 2013), to list *semantic descriptors*. Semantic descriptors are all the pieces of information that API users might want to get out of the API, or put into the API: they are the data items (also referred to as *informational resources*) that the API should return. Depending of the application scenario, these semantic descriptors can be grouped together and organized into hierarchies. An example of semantic descriptor for an API returning a list of books is 'books'. A RESTful API with a base URL,http://mylibrary.com could:

- Return all books in the library via GET http://mylibrary.com/books
- Return the book with the ISBN 1098-6596 via GET http://mylibrary.com/books/1098-6596

[2] See http://www.programmableweb.com/about (last accessed: June 30, 2016).

[3] For an anecdote showing undesirable consequences of minor changes in an API, see (Henning, 2009).

[4] As of June 30, 2016, ProgrammableWeb.com lists about 9,500 RESTful APIs.

- Add a new book with the ISBN 1098-6596 to the library catalog via
  POST http://mylibrary.com/books/addbook/1098-6596
- Delete this book from the catalog via DELETE http://mylibrary.com/books/1098-6596

The work reported in the next sections aims at providing an empirical basis for the choice of semantic descriptors for APIs for open government data. The research question motivating the work is: What are the recurrent *types of open datasets* relevant in an open government context? The assumption is that providing an answer to this question is key to the development of new application programming interfaces, which will ease data access to programmers and reduce barriers to data re-use. Imagine for example programmers accessing environmental datasets from two catalogs CAT1 (belonging to City 1) and CAT2 (belonging to City 2) using:

- GET http://cat1.city1.com/dataset/environment;
- GET http://cat2.city2.com/dataset/environment.

Such a situation would be a great improvement over the current state of affairs where datasets are accessed via cryptic items' identification numbers[5] (IDs). First, it is more user-friendly to interact with APIs which return items according to their types (e.g., climate, finance), rather than their IDs. APIs which return data items according to their types are termed *semantic APIs* in this paper. Becasue RESTful APIs require the definition of semantic descriptors, there is an appropriate architectural style for the technical implementation of semantic APIs. Second, naming schemes re-used consistently by many API designers will create an environment where programmers' learning struggles for the re-use of datasets from different open data catalogs in their application could be drastically reduced[6]. Third, semantic APIs contribute to greater transparency. As Michener and Bersch (2013) pointed out, transparency has two dimensions, namely visibility and inferability. Visibility means that the information is (i) reasonably complete and (ii) found with relative ease; inferability refers to the degree to which the information at hand can be used to draw accurate inference. A semantic API increases information visibility (i.e., it makes available the *types of data* which are used while building city applications).

[5] For instance, an example url to retrieve a table in Comma Separated Value (CSV) from a ckan catalog is http://giv-oct.uni-muenster.de:5000/api/action/datastore_search?resource_id=a774f073-ba31-44e3-8edb-ed0fca79c216&limit=5.

[6] It is a well-known fact that "*programmers at all levels, from novices to experts, repeatedly spend significant time learning new APIs*" (Myers and Stylos, 2016); and "*until […] standards are more universal, coders must write numerous interfaces for each city and maintain them individually*" (Lee, Almirall and Wareham, 2016). In the current context, a programmer willing to use datasets from two different catalogs would need to go through two learning phases to become familiar with the naming policies of their different APIs. Learning of the APIs' interfaces (and maintenance of the Apps built with these interfaces) may never be entirely removed, but it could be reduced to the strict minimum if similar naming policies were adopted while developing semantic APIs for current open data catalogs.

2. A Survey of Existing Categories for Open Government Data

Many open data catalogs classify the open datasets they provide according to categories (the term 'theme' is occasionally used to denote these categories). The goal of this section is to survey these categories across different European countries and extract some recurrent patterns (if any). 40 open data catalogs from four different countries - Germany, Spain, France and the UK - are assessed. These countries were chosen partly because the authors of this paper are native speakers of these languages. In addition, the four countries are currently among the top 10 European countries which are most-ready for open data initiatives according to the Open Data Barometer[7]. The steps followed in collecting this data were as follows:

- **Step 1**: go through the catalog and list the categories offered by each of them;
- **Step 2**: translate the categories in the target language. The target language used for this paper is English (all researchers understand it), but it is conceivable that other target languages (e.g., German, French, Spanish) could have been used for the same purpose;
- **Step 3**: harmonize the terms' translation across the four countries;
- **Step 4**: generate descriptive statistics about the dataset.

The data collection took place from June 20th to July 6th 2016. Steps 1 and 2 were carried out independently by the first, second and fourth author of the paper[8]. Conjunctions meaning 'AND' in the original languages were left out during the translation because semantic descriptors for APIs should be single words. Step 3 (harmonization) is necessary because of the possibility of translating certain terms differently in English. For instance, the terms 'labour' and 'job' were both present in the dataset after the researchers performed the initial translation. After the harmonization, only the term 'job' was kept in the dataset to facilitate the comparison of the results across countries. The choice of 'job' rather than 'labour' is, of course, a matter of personal preference, and does not influence the validity of the final conclusions. In addition, Step 3 was useful to prepare the dataset in a format useful for further processing. For example, 'urban planning' was transformed into 'urbanplanning' so that it is treated as a single word (and thus data category). Words with hyphens (e.g., E-Administration, procès-verbaux) were also converted into single words during this step. Step 3 was performed through discussions between the first, second, and fourth author. Finally, Step 4 consisted in counting the frequencies of the different data categories[9], and is meant to help answer two questions:

- What data categories *must* semantic API designers consider? The answer to this question is data categories which appear in *all* catalogs;

[7] The UK ranks 1st in Europe (1st worldwide), France ranks 2nd in Europe (2nd Worldwide), Germany ranks 7th in Europe (11th worldwide), and Spain ranks 8th in Europe (12th worldwide) according to (The World Wide Web Foundation, 2015). 'Readiness', in (The World Wide Web Foundation, 2015), means the degree of preparation for, as well as the policies in place to support open data initiatives.

[8] Only three were needed during these steps because the first author speaks French as native language, and German fluently. This researcher has also collected the data for Germany.

[9] The online tool Online-Utility.org (http://www.online-utility.org/, last accessed: July 5th, 2016) was used to count the frequencies of the different data categories.

- What data categories *could* semantic API designers consider? The answer to this question is all data categories obtained from our data collection (or put differently, data categories which appear *at least once* in the dataset). Table 1 presents all catalogs surveyed, their spatial granularities (i.e., whether they catalog datasets for a city, a region, or the whole country), as well as their URLs.

Table 1: Open Data Catalogs Surveyed

Catalog Name	**Granularity**	**URL**
GERMANY		
OffeneDaten.de	country	https://offenedaten.de/
GovData	country	https://www.govdata.de/web/guest/daten
Open Data Berlin	city	http://daten.berlin.de/
Open Data Köln	city	http://www.offenedaten-koeln.de/dataset
Open Data HRO	city	http://www.opendata-hro.de/group
Open Data München	city	https://www.opengov-muenchen.de/dataset
Open Data ULM	city	http://daten.ulm.de/datenkatalog/offene_daten
Open Government Data Portal Rheinland-Pfalz	region	http://daten.rlp.de/group
Open NRW	region	https://open.nrw/de/dat_kat
Transparenzportal Hamburg	city	http://transparenz.hamburg.de/
SPAIN		
Datos Abiertos JCYL	region	http://www.datosabiertos.jcyl.es/
Datos Abiertos Junta de Andalucía	region	http://www.juntadeandalucia.es/datosabiertos/portal.html
Datos Abiertos Madrid	city	http://datos.madrid.es/
Open data Ajuntament de Valencia	city	http://gobiernoabierto.valencia.es/
Open data Aragon	region	http://opendata.aragon.es/

OpenDataBCN	city	http://opendata.bcn.cat/opendata/
Open Data Euskadi	region	http://opendata.euskadi.eus/w79-home/eu/
Open data Gobierno de Canarias	region	opendata.gobiernodecanarias.org/
Open Data Navarra	region	www.gobiernoabierto.navarra.es/es/open-data
Portal Open Data Xunta de Galicia	region	abertos.xunta.gal/
FRANCE		
data.gouv.fr	country	http://www.data.gouv.fr/fr/datasets/
Data GrandLyon	region	http://data.grandlyon.com/
Montpellier Territoire Numérique	city	http://opendata.montpelliernumerique.fr/Les-donnees
Nantes Ouverture des Données	city	http://data.nantes.fr/
Open Data Nice Côte d'Azur	region	http://opendata.nicecotedazur.org/site/news
Open Data Bordeaux	city	http://opendata.bordeaux.fr/catalogue-des-donnees
Open PACA	region	http://opendata.regionpaca.fr/donnees.html?no_cache=1
ParisData	city	http://opendata.paris.fr/page/home/
Rennes métropole en accès libre	city	http://www.data.rennes-metropole.fr/les-donnees/catalogue/
Toulouse Métropole Data	city	https://data.toulouse-metropole.fr/page/home/
UNITED KINGDOM		
Birmingham DataFactory	city	https://data.birmingham.gov.uk/dataset
Bournemouth Data Stream	city	http://bournemouthdata.io/
Data.gov.uk	country	https://data.gov.uk/

Data- Liverpool City Council	city	http://liverpool.gov.uk/council/key-statistics-and-data/data/
Edinburgh Open Data Portal	city	http://edinburghopendata.info/
Leeds Data Mill	city	http://leedsdatamill.org/
London Datastore	region	http://data.london.gov.uk/
Open Data Bristol	city	https://opendata.bristol.gov.uk/
OpenDataNI	region	https://www.opendatani.gov.uk/
Sheffield City Council Open Data	city	https://data.sheffield.gov.uk/

Appendices A1, B1, C1, and D1 presents all catalogs' data categories as well as their translations into English. Figure 1 presents the categories' respective frequencies for Germany. The figure shows 48 distinct categories[10]. The figure shows also that five terms are used in all catalogs surveyed, namely: *culture, elections, education, sport,* and *economy*. Figure 2 shows example categories for Spanish open data catalogs. The word frequency count for Spanish open data catalogs yields 65 distinct categories. Contrary to the German case, none of the categories shown appear in all catalogs. The categories *culture, leisure, economy, education, health, environment, transport, tourism* and *employment* seem the most popular, with 8 of 10 of the catalogs surveyed proposing them for the access of open data. The French catalogs surveyed present 78 distinct categories, some of which are shown in Figure 3. *Culture* is both the most popular, and the only category which appears in all French catalogs surveyed. Figure 4 presents the different categories from the UK's catalogs, as well as their respective frequencies. Fifty nine distinct categories were obtained[11] as the result of the words frequencies count. The categories of *education* and *health* are the most popular among the UK catalogs with 9 occurrences each.

Figure 1: Categories of German Open Data Catalogs and their Frequencies

[10] The categories 'tax' and 'taxes' could have been merged into one single category, reducing this number to 47. However, both categories were kept distinct, because of the different spellings (i.e., 'Steuern', and 'Steuer') in the original German open data catalogs.

[11] The categories 'art' and 'arts' could have been merged into one single category. However, both categories were kept distinct, because of the different spellings (i.e., 'art', and 'arts') in the original open data catalogs from the UK.

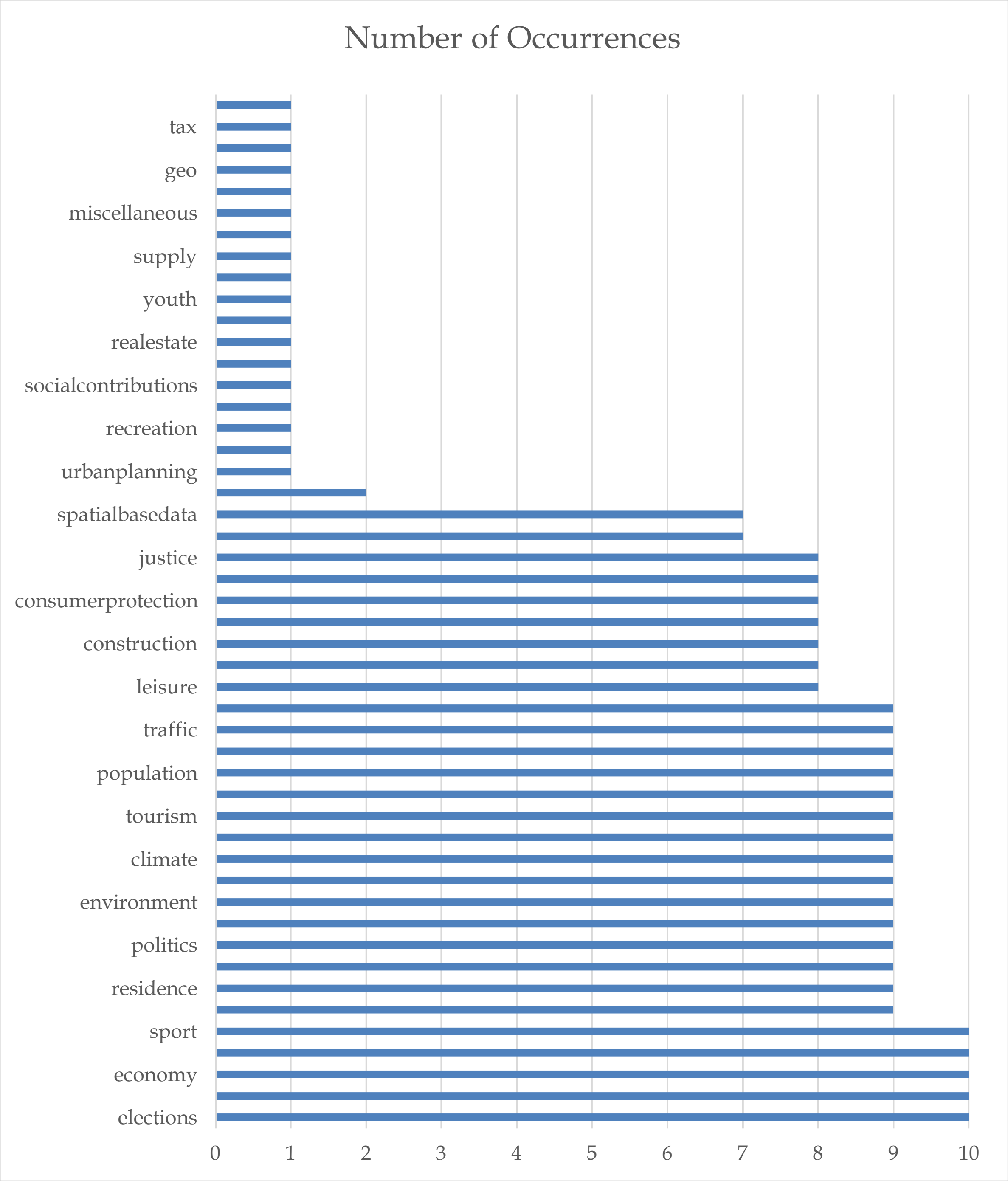

Number of Occurrences
tax
geo
miscellaneous
supply
youth
realestate
socialcontributions
recreation
urbanplanning
spatialbasedata
justice
consumerprotection
construction
leisure
traffic
population
tourism
climate
environment
politics
residence
sport
economy
elections
0
1
2
3
4
5
6
7
8
9
10

Figure 2: Categories of Spanish Open Data Catalogs and their Frequencies

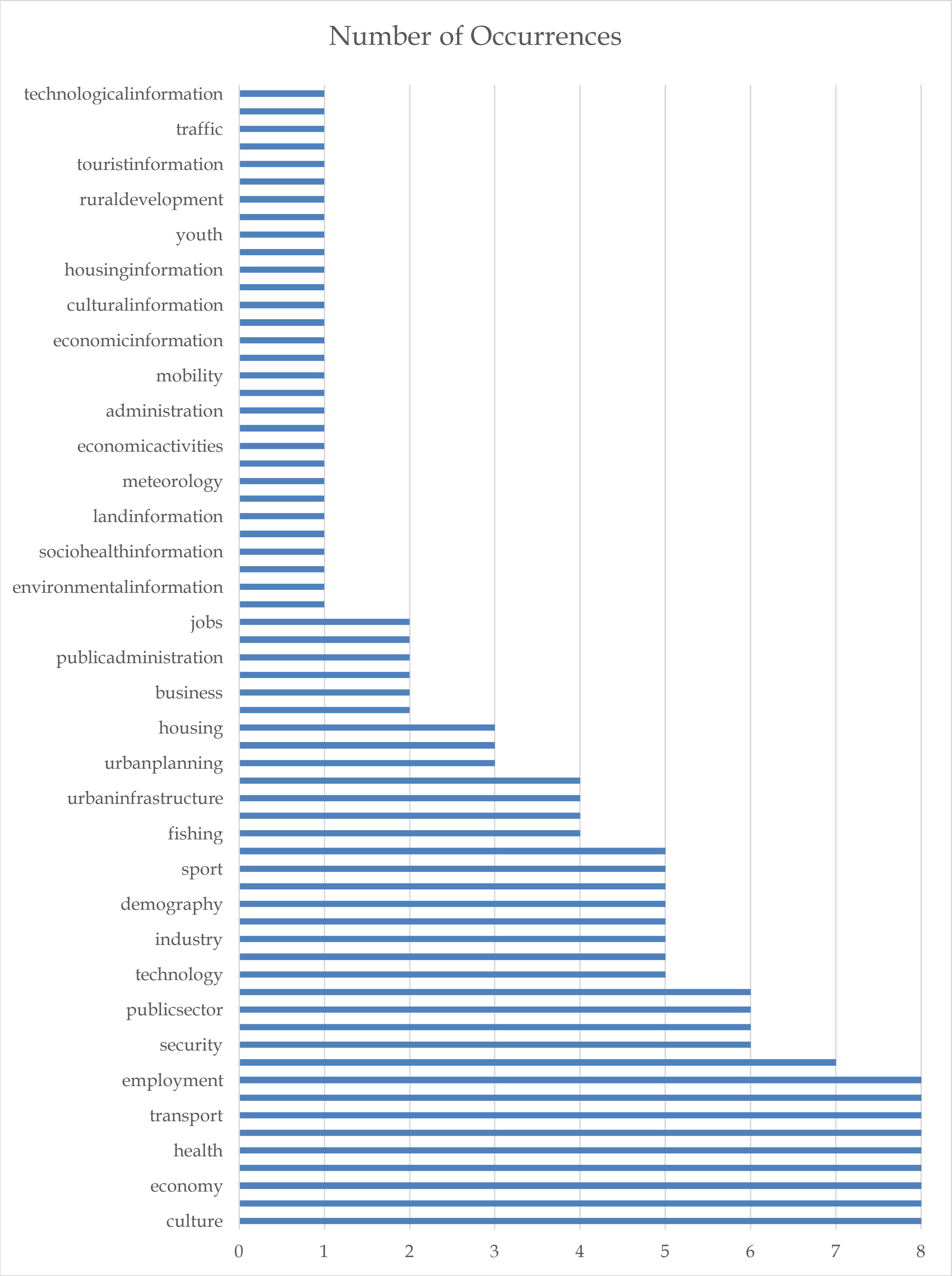


Figure 3: Categories of French Open Data Catalogs and their Frequencies

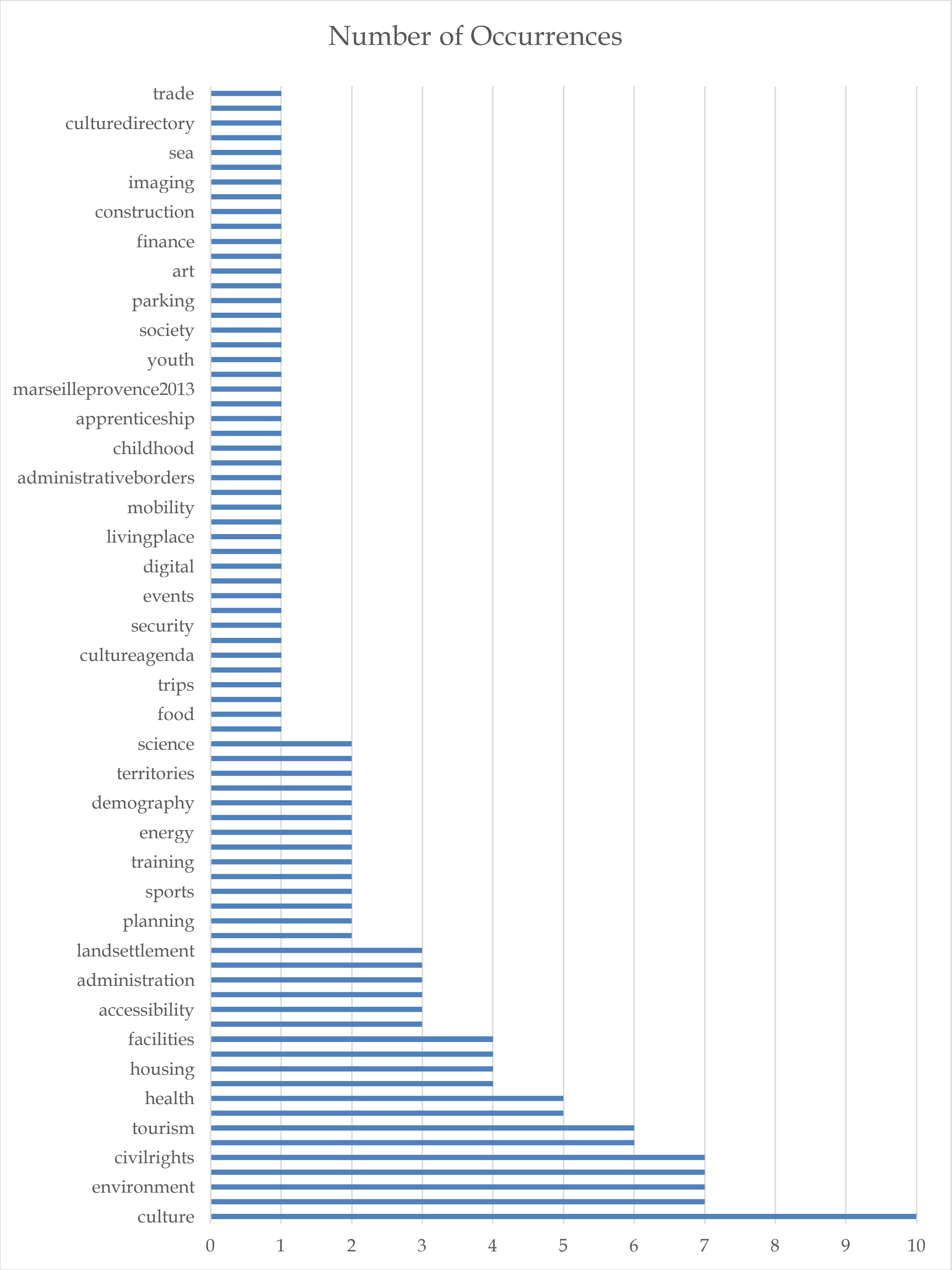
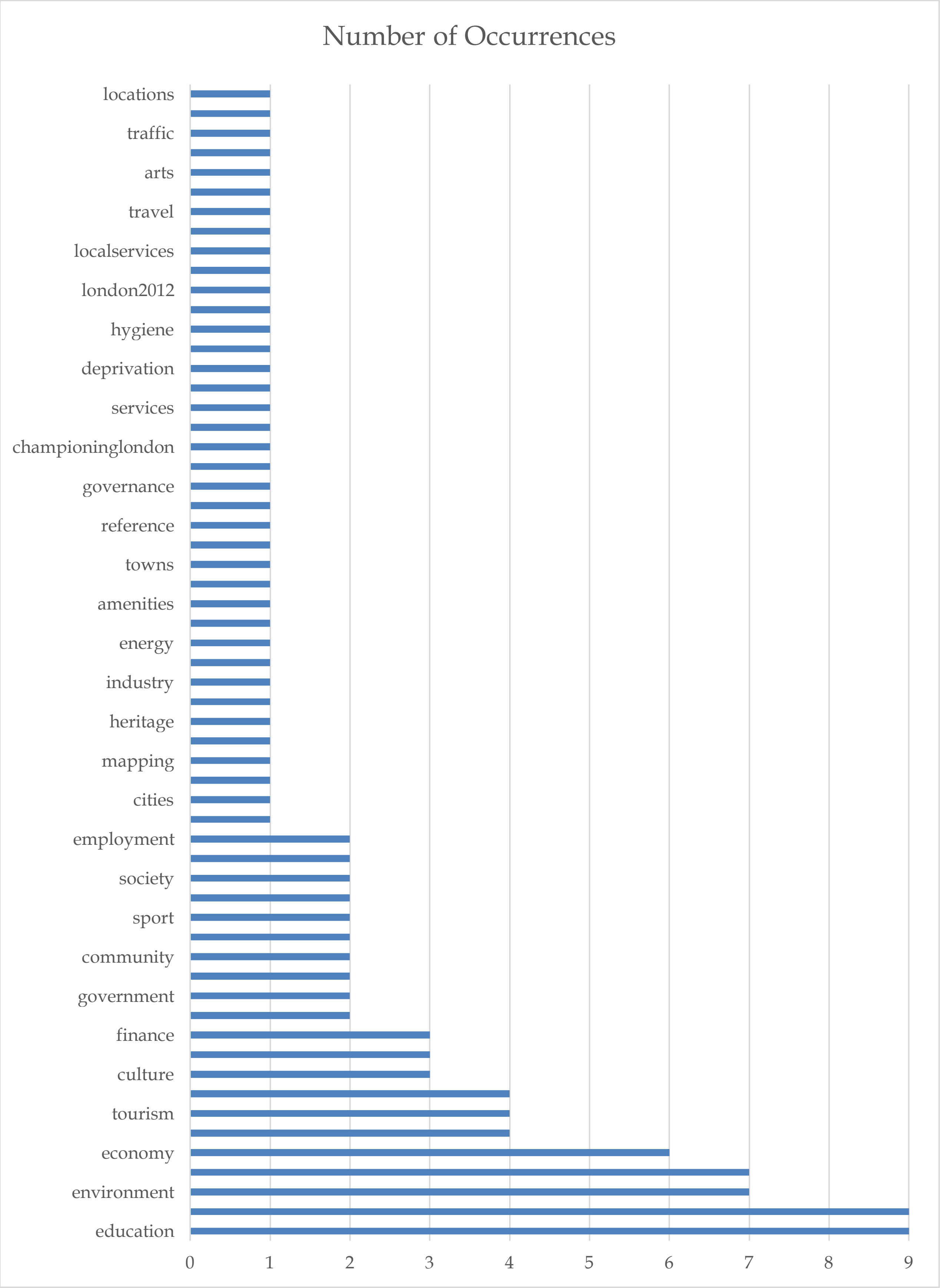


Figure 4: Categories of the UK's Open Data Catalogs as well as their Frequencies



3. Towards Semantic APIs for Open Government Data

Section 2 has surveyed 40 European open data catalogs and the different categories they provide for open data access. This process yielded 171 distinct words, which were used by providers of open datalogs to offer access to their data. This section discusses in detail how the categories gathered in the previous section can inform the choice of core categories for semantic APIs for cities. The section also briefly touches upon the different technical components useful to implement a semantic API for open government data.

3.1. Categories for a Semantic API for Open Government Data

The main motivation behind this work is to shed some light on the recurrent *types of open datasets* relevant in an open government context. Beyond informing the design of semantic APIs, the recurrent types of open datasets are an indicator of the *topics of interests in the respective countries*, the types of questions data publishers *assume* users will ask, and ultimately the types of questions citizens *can* ask. It is worth mentioning that some categories from the surveyed catalogs (appendices A1, B1, C1, and D1) could have been translated differently from how they were in this article. For example, 'Wohnen' was translated as 'residence', but could have also been translated as 'habitation' or 'home' (which are appropriate synonyms for the word retrieved from the Merriam-Webster dictionary[12]); 'Arbeit' was translated as 'jobs' ('Arbeitsmarkt' was translated as 'job market'), but the world 'labour' could have been used in lieu of 'jobs'; 'Stadtplanung' could have been translated as 'city planning' instead of 'urban planning'; and so on. This limitation is an inherent limitation of all studies which will endeavour to compare different categories offered by open data providers in Europe. The harmonization stage (Step 3, Section 2) has ensured that the translation remained consistent both within tables, and across countries. Two questions were mentioned in Section 2, namely: What data categories *must* semantic API designers consider? And what data categories *could* semantic API designers consider? The first question is referred to as Q1, and the second as Q2 in the rest of the paper. Both are now considered in turn.

3.1.1. Categories for National Semantic APIs

Appendices A2, B2, C2, and D2 present the inter-catalog agreements for the different catalogs surveyed in Germany, Spain, France, and the UK respectively. The values for inter-catalog agreement were computed using the Jaccard index, i.e., the size of the intersection of two sets divided by the size of their unions. The indices show some great differences between the countries examined: The average inter-catalog agreement for German catalogs is 0.70 (standard deviation: 0.21); this value drops to 0.34 (standard deviation: 0.31) for Spanish open data catalogs; the mean inter-catalog agreement for French open data catalogs is 0.18 (standard deviation: 0.08); and the average inter-catalog agreement for catalogs from the UK surveyed is 0.19 (standard deviation: 0.09)

[12] See http://www.merriam-webster.com/dictionary/residence (last accessed: July 1st, 2016).

[13]. The differences between the average inter-catalog agreements in the countries surveyed indicates that the level of harmonization of terms used in open data catalogs accross these countries is, at the moment, quite disparate.

The averages of inter-catalog agreements in each country were also computed by taking into account the spatial granularities (see appendices A2, B2, C2, and D2), and are as follows:

- Germany: 0.57 (city), no average at the regional level because there were only two catalogs available at this level, no average at the national level computed because there were only two catalogs in the datasets for the national level[14];
- Spain: 0.17 (city), 0.37 (region), no average at the national level computed because there is no catalog in the dataset for the national level;
- France: 0.16 (city), 0.22 (region), no average at the national level computed because there is only one catalog in the dataset for the national level;
- UK: 0.17 (city), no average at the regional level because there were only two catalogs available at this level, no average at the national level computed because there is only one catalog in the dataset for the national level.

These values lead to the following observations: Within each of the countries surveyed, the instances of open data catalogs for the national level are too few to draw some a meaningful conclusion; and the inter-catalog agreements at the city level and at the regional level are in general quite low. The former observation is not surprising because there may not be many institutions in a single country which can take an inventory of open government data across a whole country. The latter observation suggests that within each of the country, more effort - at both the local and the regional levels - is needed to harmonize the categories offered by open data providers.

Five terms appeared in all catalogs for open data in Germany surveyed: *culture, elections, education, sport*, and *economy*. Their very high rate of occurrence suggests that they seem inevitable in the German open data landscape, and that designers of semantic APIs for German cities should include them in their own APIs. Put differently, a possible answer to Q1 is *culture, elections, education, sport*, and *economy*. As regards data categories, API designers could consider as eligible categories (i.e., Q2), there are different ways of providing an answer:

- Include all categories *already used* by other open data catalog publishers (i.e., the 48 terms from Figure 1);
- Set a threshold T (0 ≤ T < 1) that categories to be included should surpass. T denotes here the frequency of appearance in existing open data catalogs. The choice of T will necessarily involve some degree of conventionality and arbitrariness, but a similar value of T across all open data catalogs makes transnational comparison possible. In the rest of this work, the illustrative value of T = 0.75 is chosen, that is, the answer to Q2 is limited to categories which appear in at least 75% of the surveyed catalogs. In the case of Germany, 27 categories fulfill this

[13] The Jaccard indices and the statistical values were calculated using two open source libraries, namely https://github.com/ecto/jaccard and https://github.com/simple-statistics/simple-statistics respectively.

[14] One needs at least three catalogs to get a meaningful value for the mean inter-catalog agreements.

requirement. These are: *culture, elections, education, sport, economy, population, transport, jobs, geography, household, traffic, health, environment, residence, science, climate, tourism, publicadministration, infrastructure, politics, justice, construction, taxes, consumerprotection, leisure, law,* and *geology.*

With respect to Spanish open data catalogs, no term seems so popular that it can be deemed as inevitable (Q1). However, the data collected suggests that designers of semantic APIs for Spanish cities could consider the following nine terms (threshold of appearance T = 0.75): *culture, transport, education, health, environment, leisure*, *tourism, economy* and *employment*. *Culture* is the only term appearing in all French open data catalogs surveyed. It imposes thus itself as a category of semantic APIs for French cities (Q1). However, (and contrary to Germany and Spain), no other term appears in at least 75% of the catalogs surveyed to be suggested as a possible answer to Q2. As to open data catalogs in the UK, no term appears in all the catalogs surveyed (Q1). Nevertheless, *health* and *transport* appear in at least 75% of the catalogs surveyed, and could be considered while designing semantic APIs for the UK's cities (Q2).

3.1.2. Categories for Bi-National Semantic APIs

What if semantic API designers want to provide APIs for a bi-national audience? Relevant categories for this task are the *intersection* of the two countries' sets of categories. With respect to Q1 (i.e, terms which appear in all catalogs surveyed), and Q2 (i.e., terms which appear in at least 75% of the catalogs surveyed), the following categories are possible answers:

- Germany-Spain: Q1 (no category found); Q2 (*culture, economy, education, health, environment, transport, tourism, leisure* and *sport*);
- Germany-France: Q1 (*culture*); Q2 (*culture, economy, environment, transport, education* and *tourism*);
- Germany-UK: Q1 (no category found); Q2 (*education, health, economy, environment* and *transport*);
- Spain-France: Q1 (no category found); Q2 (*culture, economy, environment* and *transport*);
- Spain-UK: Q1 (no category found); Q2 (*education, health, environment* and *transport*);
- France-UK: Q1 (no category found); Q2 (no category found).

Table 2 presents the values of the Jaccard indices between the different countries. The Jaccard indices show the inter-country agreement between the categories offered by the open data providers. This inter-country agreement is again quite low, oscillating between 0.14 and 0.25. Below are the values of the inter-catalog agreements for open data catalogs at the same spatial granularity. They indicate that this low inter-catalog agreement is more or less homogenously present at all levels (local, regional and national).

- Germany-Spain: 0.26 (city), 0.23 (region), no value computed because there is no Spanish catalog in the dataset for the national level;
- Germany-France: 0.20 (city), 0.18 (region), 0.23 (country);
- Germany-UK: 0.17 (city), 0.20 (region), 0.14 (country);
- Spain-France: 0.25 (city), 0.20 (region), no value computed because there is no Spanish catalog in the dataset for the national level;

- Spain-UK: 0.24 (city), 0.18 (region), no value computed because there is no Spanish catalog in the dataset for the national level;
- France-UK: 0.19 (city), 0.14 (region), 0.2 (country).

Table 2: Jaccard Indices Showing Inter-Country Agreement between Data Categories Offered by Open Data Providers

Country (A)	Jaccard Index (A, France)	Jaccard Index (A, Germany)	Jaccard Index (A, Spain)	Jaccard Index (A, UK)
France	1	0.17	0.25	0.17
Germany	0.17	1	0.24	0.14
Spain	0.25	0.24	1	0.18
UK	0.17	0.14	0.18	1

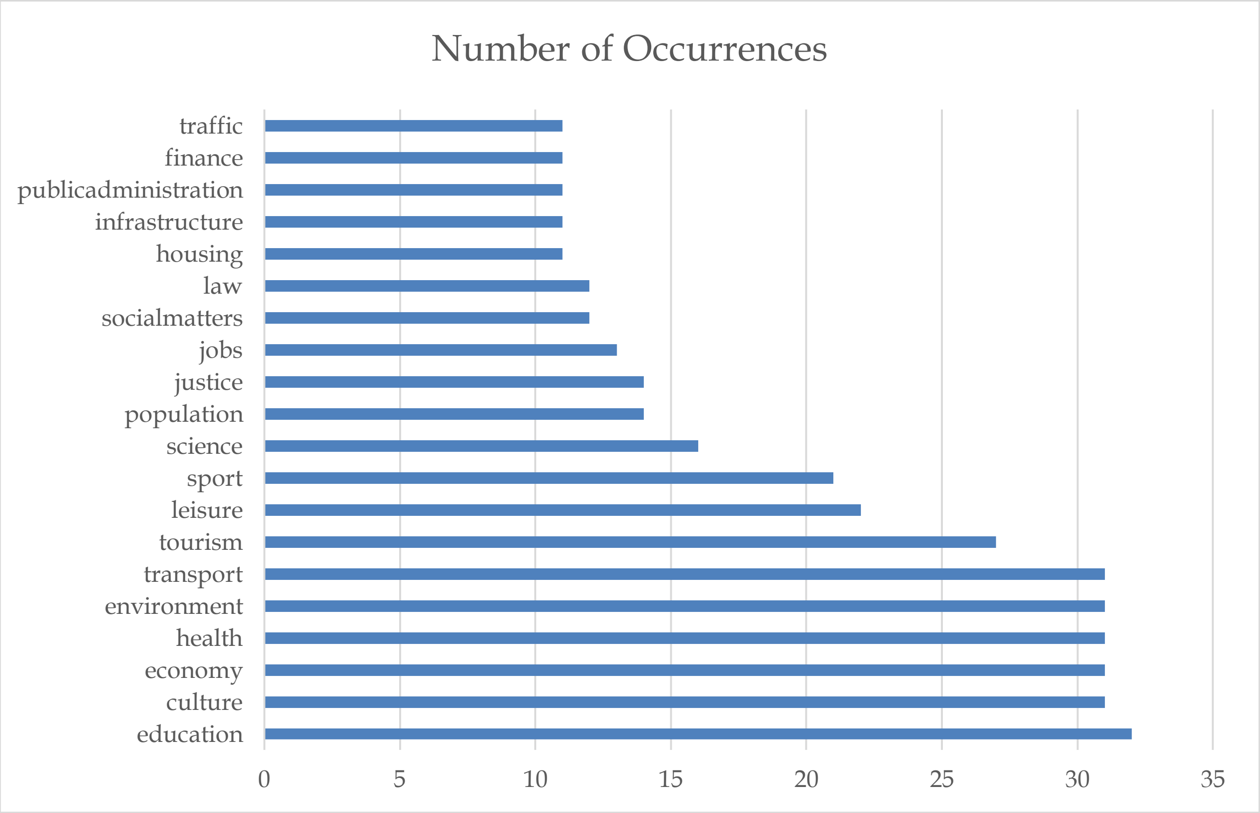
3.1.3. Categories for a Transnational Semantic API

There are various ways of obtaining a list of useful categories to consider while designing semantic APIs at a transnational level. One way is to adopt a minimalistic approach, i.e., only categories from the national level appearing in all countries should be considered. This approach yields no category for a transnational semantic API (i.e., no answer for Q1). A variant of the minimalist approach is to consider categories which appear in at least three of the four countries to be relevant. In that case, *culture*, *health* and *transport* would be good candidate categories for a transnational semantic API (Q1).

An alternative to the minimalist approach would be a maximalist approach, i.e., each of the category from the national level should be included at the transnational level. This results in the following list of 27 categories (mostly inherited from German open data catalogs) for a transnational semantic API: *culture, elections, education, sport, economy, population, transport, jobs, geography, household, traffic, health, environment, residence, science, climate, tourism, publicadministration, infrastructure, politics, justice, construction, taxes, consumerprotection, leisure, law,* and *geology*. These categories are also possible answers to Q2.

A third way of generating categories for a transnational semantic API is to compute descriptive statistics using all terms from all catalogs surveyed (appendices A1, B1, C1 and D1) altogether. Figure 5 presents the list of terms resulting from this approach (only the 20 terms which occurred the most are shown). If the threshold of appearance is set to 0.75 (i.e., T= 0.75), the candidate list of terms obtained (Q2) is *education, culture, economy, health, environment* and *transport*.

Figure 5: Candidate Categories for a Transnational Semantic API



3.1.4. Discussion

Sections 3.1.1 to 3.1.3 have examined possible answers to the two questions motivating this work: what data categories *must* semantic API designers consider? And what data categories *could* semantic API designers consider?. A couple of insights can be summarized from these sections. First, the sections illustrate that an empirical approach to generate terms for semantic APIs is applicable, and may be used to generate terms for semantic APIs at the local, regional and national levels. However, these sections also illustrate that the answers obtained are strongly dependent on the approach taken. Since any of the approach mentioned necessarily involves some degree of conventionality and arbitrariness, any empirical approach to generate categories for semantic APIs should make explicit what the underlying parameters (e.g., maximalist vs minimalist approach, theshold of appearance) are to facilitate traceability.

Second, previous sections have presented inter-catalog agreements from different perspectives. The following conclusions can be drawn based on the values obtained:

- The four countries examined are non-homogeneous with respect to level of harmonization of terms used in their open data catalogs;
- Within each of the countries, the inter-catalog agreements at the city level and at the regional level are in general quite low;
- There is also a low inter-catalog agreement between terms used in a country, and terms used in another (in average less than 30%). In other words, the probability that a category chosen

by a data provider in one European Country will also be chosen by a data provider in another European country is somewhere less than 30 percent.

The consistently low values obtained for inter-catalog agreements remind of the 'vocabulary problem', i.e., the low probability that two people use the same term to refer to a specific object in computer applications. The solution to this problem proposed in (Furnas et al., 1987) is unlimited aliasing, i.e., the provision of many alternative words to users so that they can get what they want from large and complex systems. Unlimited aliasing is not entirely suitable for the case of semantic API design, since designers can only choose one term as entry point for their data items. A possible solution to the problem (perhaps the only one?) is the *coordination of efforts* between different data providers. Coordination, in this case, would involve some commitment from different data providers (local, regional, national) to use a set of terms, with agreed upon definitions, in their catalogs. Finally, Susha et al. (2015) identified a set of 12 critical factors for the success of open data initiatives. One of these factors is to "Integrate metadata schemas and *federated controlled vocabularies for properly categorizing information*" (emphasis added). The Susha et al. (2015) study derived the factor from two workshops conducted with a number of experts. The consistently low values for inter-catalog agreements, obtained from an empirical survey of existing open data catalog categories, confirm the need to implement this success factor in current open government initiatives from a different perspective. Moreover, the low values for inter-catalog agreements suggest that *federated controlled vocabularies for properly categorizing information* is necessary both a local, regional and national level in the four European countries surveyed.

A couple of limitations of the study are also worth mentioning. One limitation is that the results are dependent on the quality of the translated terms in English. As mentioned in Section 2, some of the terms could have been translated differently yielding slightly different results. These effects were minimized through Step 3 and Step 4 of the method. In addition, since no information is available on how the categories were chosen by the open data provider (e.g., based on institutional mandate, or simply because it fits best their existing data, etc.), there are some limits to the explanatory power of this study (i.e., why some of the difference are observed).

3.2. Technical Components of a Semantic API

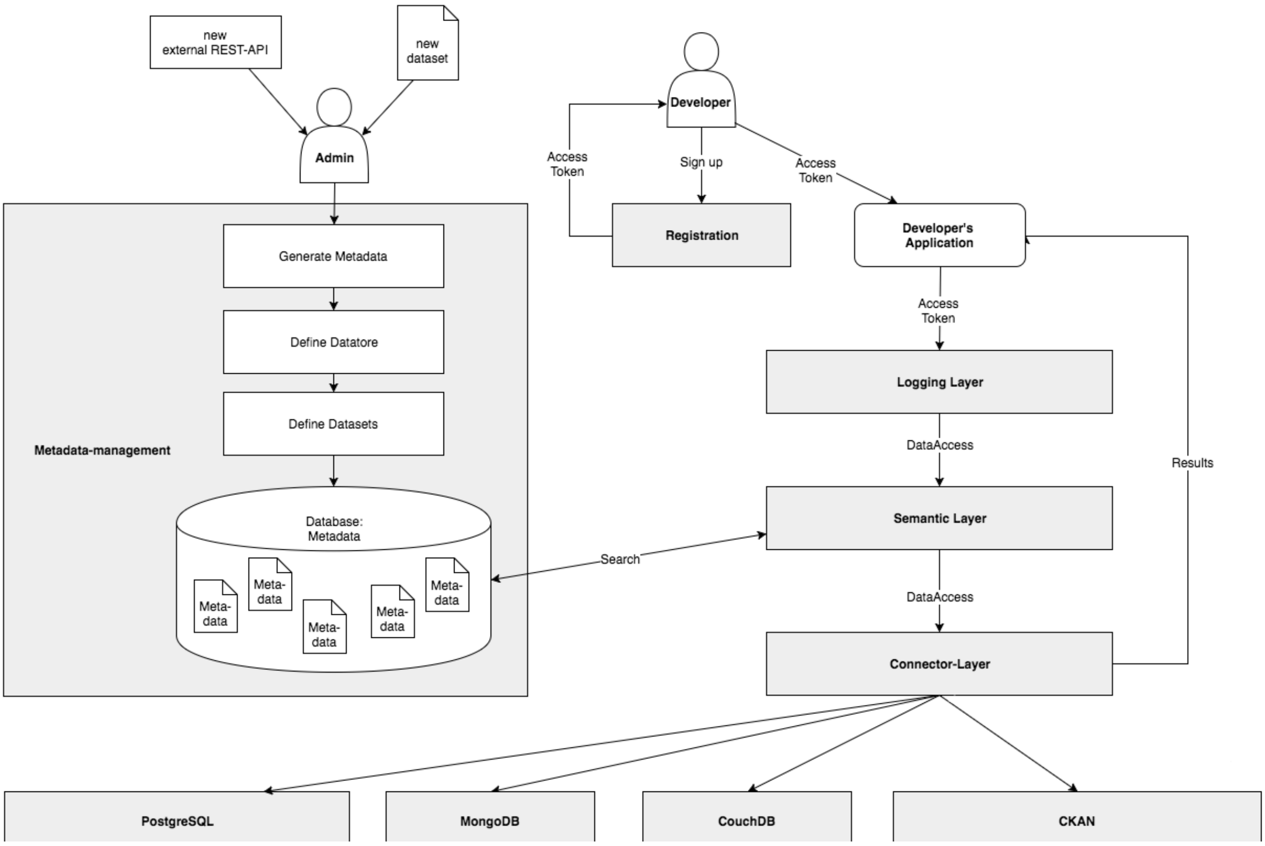
As mentioned in Section 1, one of the benefits of semantic APIs is to increase transparency. This happens because semantic APIs improve information visibility (i.e., they make available the *types of data* which are used while building applications with open government data). The implementation of semantic APIs necessitates some technical considerations which are discussed in this section. Six main components are needed to realize them: a metadata-management component, a registration component, a logging layer, a semantic layer, a connector, and the databases. All components introduced are illustrated in Figure 6. Their role is described below:

- **Registration component**: this component is helpful to register any developer who wants to use a semantic API to build city applications. After a successful registration, a developer receives an access-token[15] which will be used to identify the app making calls to the semantic API. The ability to know *who is making an API call* is one feature of semantic APIs;

- **Logging layer**: this component records all events related to the semantic API, for example, the ids of the applications which request a certain type of dataset (the applications can be automatically identified using via the access-token), the types of data requested (e.g., culture, health or transport), the number of applications accessing a certain data, and the frequency of API calls. The log files generated by the logging layer can be formatted using a well-known open format such as the common log format presented in (World Wide Web Consortium, 1995). The logging layer generates information about *what is happening* with the API, and *when*.
- **Metadata-management component**: the role of this component is to establish all mappings between the requests of the users, and the databases relevant to process these requests. It is the 'brain' of the semantic API because it stores all relevant conceptual relationships for the functioning of the API. This component is built and maintained by the API provider (e.g., an institution such as a city council). The metadata-management component would specify for example that 'health' in English is equivalent to 'Gesundheit' (in German) and that any request related to health is also a request about 'Gesundheit'. The metadata-management component can also specify hierarchical relationships between concepts (e.g., a request about health is a request about all items with a direct relation to the 'health' concept, and more specific health concepts such as 'health insurance' and 'preventive care').
- **Semantic layer**: this component provides a bridge between the request of the user (or software agent) and the metadata-management component; it retrieves the databases (and all concepts) to look for based on the user requests, and forwards this information to the connector layer.
- **Connector-layer**: The queries performed on the databases, as well as the query languages (e.g., SQL, SPARQL, interfaces of a RESTful API) needed to return data items are dependent on the user request and the database being queried. This layer stores therefore the different queries needed to retrieve specific datasets from the databases.
- **Databases**: they store the data which can be of any type: relational data (PostgreSQL), document-oriented data (MongoDB, CouchDB), data coming from CKAN-based platforms, graph-based data (stored in triple stores such as Parliament, Fuseki or Virtuoso), or even data coming from other semantic APIs.

[15] See https://msdn.microsoft.com/en-us/library/Aa374909.aspx (last accessed: July 7th, 2016) for a short introduction to access-tokens.

Figure 6: Generic components of a semantic API



The JavaScript development environment Node.js[16] is currently used to implement these components. The main reason for choosing Node.js is its portability; all Node.js applications (irrespective of their functionality) can be run using two commands, namely 'npm install' followed by 'npm start'. That is, any city council could install and run the API with relative ease (i.e., only two commands). Semantic APIs is an essential component for the realization of the vision of the Open City Toolkit described in (Degbelo *et al.*, 2016). Figure 7 illustrates one practical use of a semantic API, namely generate information about applications in a city which access some types of datasets, and types of datasets which are often requested in a city. The example on the figure as well as the documentation of the features of the API implemented can be accessed from https://github.com/geo-c/OCT-Core.

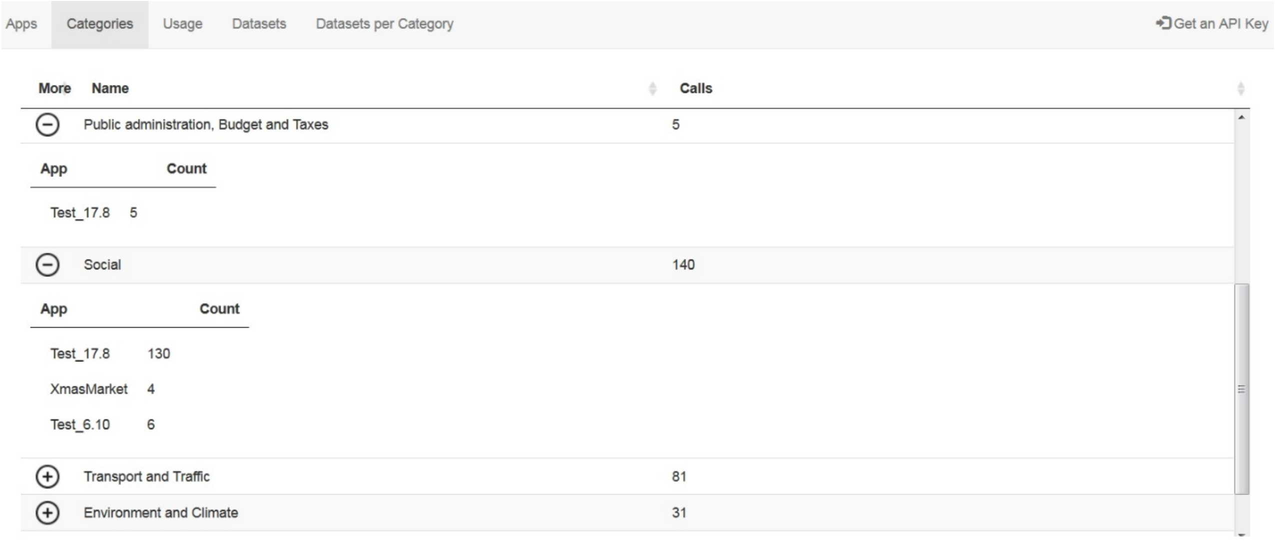
4. Related Work

To the best of the authors' knowledge, there has not been any attempt to classify the different categories of open data catalogs to inform the design of semantic APIs in the literature. Example classifications of city data appear in (Lecue, Kotoulas and Mac Aonghusa, 2012; Bischof et al., 2014), yet they were not generated based on an empirical consideration of data categories provided by current open data catalogs. Attard, Orlandi, Scerri, and Auer (2015) provided a systematic survey of

[16] See https://nodejs.org/en/ (last accessed: July 7th 2016) for further information about Node.js.

open (government) data initiatives but did not specifically look at categories offered by open data catalogs.

Figure 7: An example of practical use of semantic API - information can be produced about applications which access some types of datasets in a city. Datasets of type 'social' seem to be the most requested in this example; the app Test_17.8 re-uses dataset which are related to the categories 'Public administration, Budget and Taxes', and 'Social'.



More	Name	Calls
⊖	Public administration, Budget and Taxes	5

App	Count
Test_17.8	5

More	Name	Calls
⊖	Social	140

App	Count
Test_17.8	130
XmasMarket	4
Test_6.10	6

More	Name	Calls
⊕	Transport and Traffic	81
⊕	Environment and Climate	31

From a practical perspective, designers of semantic APIs could resort also to categories provided in (The World Wide Web Foundation, 2015), though these categories were not proposed to this end. The categories found in (The World Wide Web Foundation, 2015) are: *maps, land, statistics, budgets, spending, companies, legislation, transport, trade, heath, education, crime, environment, elections*, and *contracts*. Because these categories were already used to assess the progress of 92 countries with respect to their adoption of open data, they could also be considered while providing access to open data. Nevertheless there is no information as to the reason why these specific categories were chosen to perform the assessment in (The World Wide Web Foundation, 2015), and further iterations may find alternative classifications in this article.

A Spanish standard called UNE 178301:2015 (Aenor, 2015) was elaborated by a group of Spanish smart cities. UNE 178301:20 defines a set of indicators divided into five categories: political, organizational, technical, legal and economy. Only, one category of this standard appears in Figure 2, namely 'economy'. Also, this standard defines the metrics to quantify the level of the open data in Spanish cities. Another possible source of categories for the design of semantic APIs is the European Data Portal which "harvests the metadata of Public Sector Information available on public data portals across European countries"[17]. Categories for data access offered by this portal are: *agriculture, fisheries, forestry, foods, energy, regions, cities, transport, economy, finance, internationalissues, government, publicsector, justice, legalsystem, publicsafety, environment, education, culture, sport, health, population, society, science*, and *technology*. A good sign is that 11 of these 25 categories appear in

[17] See http://www.europeandataportal.eu/en/what-we-do (last accessed: July 8th, 2016).

Figure 5. Chances are that the team developing the European Data Portal has asked itself questions similar to those asked in this paper, though there is no information of how the categories were created.

The CitySDK Linked Data API[18] was developed to provide unified and direct access to "open" data, with an interface for writing data. It was designed to work closely with other open source projects such as OpenTripPlanner, OpenTripPlanner Analyst, Open311, GTFS, and OpenStreetMap, where one query about one object provides results from multiple datasets, annotated using semantic web technologies. CitySDK provides a web service offering integrated and direct access to open data from government, commercial and crowd sources identically. The web service is adopted by six European cities. The CitySDK Linked Data API makes data available by collecting data or web services from different sources, describing the data, linking the data to reference datasets when applicable (viz. Cadastre/OSM), offers the data as a unified service to other applications (API), also allowing the applications to annotate and enrich the data. Independent of file format, refresh rate or granularity open data is easily accessible for commercial use, research and software developers. The research in this paper considered the European cities' open data held by CitySDK platform and incorporated the strength of semantics links for those data sets to yield the result for the openness of the respective city/country.

The data API from Data.gov.uk is RESTful, and may be considered the most advanced implementation of semantic APIs in the current open data landscape. However only two categories are available at the time of this writing, namely *health* and *transport*[19]. Though the API offers the opportunity to retrieve datasets according to these categories, its documentation says nothing about registration and logging capabilities which are the pre-requisite for increased transparency as regards the use of data sources in a city context. The categories obtained in this work as well as the technical discussion in Section 3.2 provide a solid ground for making this API more sophisticated at the technical level, and adding new topics to it.

5. Conclusion

As the recent edition of the Open Data Barometer (The World Wide Web Foundation, 2015) has pointed out, open data is entering the mainstream, but the more elaborated APIs that facilitate access to data are still very rare among government data. This work has proposed that semantic APIs could be such 'elaborated APIs', and presented their technical components. The work also pointed out that the REST architectural style is an adequate paradigm for the implementation of semantic APIs. As semantic APIs rely on data categories to make data items available to both programmers and machines, this paper has looked into data categories relevant for semantic APIs designers in European countries. The article has surveyed 40 European data catalogues from four countries (France, Germany, Spain, and the United Kingdom) and observed the recurrent data categories offered by open data providers in these countries. The results show great disparities between the

[18] http://www.citysdk.eu/mobility/ (last accessed: November 11th, 2016).

[19] See https://data.gov.uk/data/api/ (last accessed: July 8th, 2016).

countries surveyed, but suggest that *culture*, *health* and *transport* would be good candidate categories for a transnational semantic API. The results also show that the probability of inter-country agreement between open data catalogs is less than 30 percent. This suggests that effort is needed with respect to coordination among countries so that semantic APIs built in one country have greater chances of adoption by other countries. Any study like the one presented in this paper is dependent upon the quality of the translation of terms between the languages (which is by definition never perfect) and the translator. This aspect puts some limits on the generalizability of the results. Nonetheless, the merit of this work has been to provide a set of categories based on the current practice to inform the design of semantic APIs. The results obtained stress the need for *federated controlled vocabularies for properly categorizing information* at both a local, regional and national level in the four European countries surveyed.

The data categories surveyed reflect a data provider perspective of the current open data landscape. That is, they give an indication of the types of datasets that open data providers assume citizens will look for. A useful complement to this study could look at the most requested data categories (e.g., number of downloads) of open data catalogs to get an understanding of what citizens actually often look for[20]. Another direction for future work would be a large-scale survey asking programmers across the four countries to assess the types of datasets they would retrieve in case they were provided with semantic APIs for their cities. Finally, future work could also, in addition to the aspects discussed in this paper, have a closer look at other usability factors (e.g., complexity, documentation, error handling: for a complete list, see Zibran, Eishita, & Roy, 2011) which will favor API adoption in the open data landscape.

References

Aenor (2015). UNE 178301:2015. Ciudades inteligentes. Datos abiertos http://www.aenor.es/aenor/normas/normas/fichanorma.asp?tipo=N&codigo=N0054318#.WBCU6vmLSUl

Attard, J., Orlandi, F., Scerri, S. and Auer, S. (2015) 'A systematic review of open government data initiatives', Government Information Quarterly, 32(4), pp. 399–418. doi: 10.1016/j.giq.2015.07.006.

Berners-Lee, T. (2006) 'Linked Data - Design issues (http://www.w3.org/DesignIssues/LinkedData.html; last accessed: November 07, 2016)'.

Bischof, S., Karapantelakis, A., Sheth, A., Mileo, A. and Barnaghi, P. (2014) 'Semantic modelling of smart city data', in W3C Workshop on the Web of Things Enablers and services for an open Web of Devices. Berlin, Germany, pp. 1–5. Available at: http://www.w3.org/2014/02/wot/papers/karapantelakis.pdf.

Chen, M., Mao, S. and Liu, Y. (2014) 'Big data: A survey', Mobile Networks and Applications, 19(2), pp. 171–209. doi: 10.1007/s11036-013-0489-0.

d'Aquin, M., Adamou, A., Daga, E., Liu, S., Thomas, K. and Motta, E. (2014) 'Dealing with diversity in a smart-city datahub', in Omitola, T., Breslin, J., and Barnaghi, P. (eds) Proceedings of the Fifth Workshop on Semantics for Smarter Cities (S4SC 2014). Riva del Garda, Italy: CEUR-WS.org, pp. 68–82.

[20] At the moment of this writing most of the catalogs surveyed in the paper do not provide this information.

Degbelo, A., Granell, C., Trilles, S., Bhattacharya, D., Casteleyn, S. and Kray, C. (2016) 'Opening up smart cities: citizen-centric challenges and opportunities from GIScience', ISPRS International Journal of Geo-Information, 5(2), p. 16. doi: 10.3390/ijgi5020016.

Fechner, T. and Kray, C. (2014) 'Georeferenced open data and augmented interactive geo-visualizations as catalysts for citizen engagement', eJournal of eDemocracy and Open Government, 6(1), pp. 14–35.

Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T. (1987) 'The vocabulary problem in human-system communication', Communications of the ACM, 30(11), pp. 964–971. doi: 10.1145/32206.32212.

Hancke, G. P., Silva, B. de C. e and Hancke Jr., G. P. (2013) 'The role of advanced sensing in smart cities', Sensors, 13(1), p. 393. doi: 10.3390/s130100393.

Hartog, M., Mulder, B., Spée, B., Visser, E. and Gribnau, A. (2014) 'Open data within governmental organisations', eJournal of eDemocracy and Open Government, 6(1), pp. 49–61.

Henning, M. (2009) 'API design matters', Communications of the ACM, 52(5), pp. 25–36. doi: 10.1145/1506409.1506424.

Janssen, M., Charalabidis, Y. and Zuiderwijk, A. (2012) 'Benefits, adoption barriers and myths of open data and open government', Information Systems Management, 29(4), pp. 258–268.

Lecue, F., Kotoulas, S. and Mac Aonghusa, P. (2012) 'Capturing the pulse of cities: A robust stream data reasoning approach', in Srivastava, B., Lecue, F., and Joshi, A. (eds) The AAAI 2012 Workshop on Semantic Cities. Toronto, Ontario, Canada: AAAI Press, pp. 9–14.

Lee, M., Almirall, E. and Wareham, J. (2016) 'Open data and civic apps: First-generation failures, second-generation improvements', Communications of the ACM, 59(1), pp. 82–89. doi: 10.1145/2756542.

Masip-Bruin, X., Guang-Jie, R., Serral-Gracià, R. and Yannuzzi, M. (2013) 'Unlocking the value of open data with a process-based information platform', in 15th IEEE Conference on Business Informatics (CBI 2013). Vienna, Austria: Institute of Electrical and Electronics Engineers (IEEE), pp. 331–337. doi: 10.1109/CBI.2013.54.

Michener, G. and Bersch, K. (2013) 'Identifying transparency', Information Polity, 18(3), pp. 233–242. doi: 10.3233/IP-130299.

Myers, B. A. and Stylos, J. (2016) 'Improving API usability', Communications of the ACM. ACM, 59(6), pp. 62–69. doi: 10.1145/2896587.

Ojo, A., Curry, E. and Zeleti, F. A. (2015) 'A tale of open data innovations in five smart cities', in 48th Hawaii International Conference on System Sciences (HICSS 2015). Kauai, Hawaii: IEEE, pp. 2326–2335.

Pomerantz, J. and Peek, R. (2016) 'Fifty shades of open', First Monday, 21(5). doi: 10.5210/fm.v21i5.6360.

Richardson, L. and Amundsen, M. (2013) RESTful web APIs, RESTful Web application programming interfaces. O'Reilly Media, Inc. doi: 10.1017/CBO9781107415324.004.

Scheider, S. and Kuhn, W. (2015) 'How to talk to each other via computers - Semantic interoperability as conceptual imitation', in Zenker, F. and Gärdenfors, P. (eds) Applications of Conceptual Spaces. Springer International Publishing, pp. 97–122. doi: 10.1007/978-3-319-15021-5_6.

Stylos, J. and Myers, B. (2007) 'Mapping the space of API design decisions', in Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC 2007). Coeur d'Alène, Idaho, USA: IEEE, pp. 50–57. doi: 10.1109/VLHCC.2007.36.

Susha, I., Zuiderwijk, A., Charalabidis, Y., Parycek, P. and Janssen, M. (2015) 'Critical factors for open data publication and use: A comparison of city-level, regional, and transnational cases', eJournal of eDemocracy and Open Government, 7(2), pp. 94–115.

The World Wide Web Foundation (2015) Open Data Barometer - Global Report. 3rd edn. The World Wide Web Foundation. Available at: http://opendatabarometer.org/doc/3rdEdition/ODB-3rdEdition-GlobalReport.pdf

World Wide Web Consortium (1995) Logging Control In W3C httpd. Available at: https://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format (Accessed: 7 July 2016).

Zibran, M. F., Eishita, F. Z. and Roy, C. K. (2011) 'Useful, but usable? Factors affecting the usability of APIs', in Pinzger, M., Poshyvanyk, D., and Buckley, J. (eds) 18th Working Conference on Reverse Engineering (WCRE 2011). Lero, Limerick, Ireland: IEEE Computer Society, pp. 151–155. doi: 10.1109/WCRE.2011.26.

About the Authors

Auriol Degbelo

Auriol Degbelo is postdoctoral researcher at the Institute for Geoinformatics, University of Münster, Germany. His research interests include ontology for geographic information processing and semantic integration of geospatial information.

Sergio Trilles

Sergio Trilles received his PhD in Integration of Geospatial Information from the Jaume I University in 2015. He had the opportunity to work four months as researcher in the Digital Earth and Reference Data Unit of the European Commission's Joint Research Centre (JRC). Currently, he is a postdoc researcher at the GEOTEC group.

Chris Kray

Chris Kray is a professor in Geoinformatics at the Institute for Geoinformatics (ifgi) at the University of Münster, Germany. His research interests include location-based services, smart cities and human-computer interaction, particularly interaction with spatial information. Chris is the scientific coordinator of the ITN "GEO-C: enabling open cities" at ifgi, where he works on realising transparency, accessibility and privacy protection in the context of smart cities.

Devanjan Bhattacharya

Devanjan Bhattacharya holds a PhD in Geomatics Engineering and his research interests are in applications of geoinformatics for societal challenges, geohazard management, smart cities, and spatial technologies. He is currently a post-doctoral manager of EU H2020 project GEO-C at NOVA IMS, Universidade Nova de Lisboa, Lisbon, Portugal.

Nicholas Schiestel

Nicholas Schiestel is a master student at the Institute for Geoinformatics, University of Münster. He has a bachelor's degree in Geoinformatics and he has worked since 2014 as a student assistant in the Situated

Computing and Interaction Lab, University of Münster. His interests lie in modern web technologies, the design of software architectures and the internet of things.

Jonas Wissing

Jonas Wissing is a student assistant at the Institute of Geoinformatics (ifgi) at the University of Münster, Germany. He is currently studying a Bachelor of Science in Geoinformatics and a Bachelor of Science in Information Systems at the University of Münster.

Carlos Granell

Carlos Granell currently holds a 5 year Ramón y Cajal post-doctoral fellowship at the UJI of Castellón, Spain. Before re-joining GEOTEC in 2014, he worked for 3 years as a post-doc in the Digital Earth and Reference Data Unit of the European Commission's Joint Research Centre (JRC), and was a post- and pre-doctoral researcher during the period 2003-2010 at the Universitat Jaume I of Castellón, from which he holds a Ph.D. (2006). His research interests lie in multi-disciplinary GIS, model web, and spatial analysis and visualization.

Appendix A1: 10 German Open Data Catalogs with their Respective Data Categories

Catalog name	Publisher	Data Categories (German)	Data Categories (English)
OffeneDaten.de	Open Knowledge Foundation Germany	Geographie, Geologie und Geobasisdaten, Bildung und Wissenschaft, Umwelt und Klima, Soziales, Infrastruktur, Bauen und Wohnen, Bevölkerung, Öffentliche Verwaltung, Haushalt und Steuern, Verbraucherschutz, Wirtschaft und Arbeit, Transport und Verkehr, Kultur, Freizeit, Sport und Tourismus, Gesundheit, Politik und Wahlen, Noch nicht kategorisiert, Gesetze und Justiz	geography, geology, spatialbasedata, education, science, environment, climate, infrastructure, construction, residence, population, publicadministration, household, taxes, consumerprotection, economy, jobs, transport, traffic, culture, leisure, sport, tourism, health, politics, elections, notyetcategorized, law, justice
GovData	Finanzbehörde Hamburg, Feschäfts- und Koordinierungsstelle GovData	Bevölkerung, Bildung und Wissenschaft, Geographie, Geologie und Geobasisdaten, Gesetze und Justiz, Gesundheit, Infrastruktur, Bauen und Wohnen, Kultur, Freizeit, Sport und Tourismus, Politik und Wahlen, Soziales, Transport und Verkehr, Umwelt und Klima, Verbraucherschutz, Öffentliche Verwaltung, Haushalt und Steuern, Wirtschaft und Arbeit	Population, education, science, geography, geology, spatialbasedata, law, justice, health, infrastructure, construction, residence, culture, leisure, sport, tourism, politics, elections, socialmatters, transport, traffic, environment, climate, consumerprotection, publicadministration, household, taxes, economy, jobs
Open Data Berlin	Senatsverwaltung für Wirtschaft, Technologie, und Forschung	Arbeitsmarkt, Bildung, Demographie, Geographie und Stadtplanung, Gesundheit, Jugend, Kunst und Kultur,	Jobmarket, education, demography, geography, urbanplanning, health, youth, art, culture,

		Öffenliche Verwaltung, Haushalt, und Steuern, Protokolle und Beschlüsse, Sonstiges, Sozialleistungen, Sport und Erholung, Tourismus, Umwelt und Klima, Ver- und Entsorgung, Verbraucherschutz, Verkehr, Wahlen, Wirtschaft, Wohnen und Immobilien	publicadministration, household, taxes, procèsverbaux, decrees, miscellaneous, sport, socialcontributions, recreation, tourism, environment, climate, supply, disposal, consumerprotection, traffic, elections, economy, residence, realestate
Open Data Köln	Stadt Köln	Geo, Bevölkerung, Politik und Wahlen, Transport und Verkehr, Umwelt und Klima, Verwaltung, Haushalt und Steuern, Kultur, Freizeit, Sport und Tourismus, Soziales, Bildung und Wissenschaft, Infrastruktur, Bauen und Wohnen, Gesundheit, Gesetzte und Justiz, Wirtschaft und Arbeit	Geo, population, politics, elections, transport, traffic, environment, climate, administration, household, taxes, culture, leisure, sport, tourism, socialmatters, education, science, infrastructure, construction, residence, health, law, justice, economy, jobs
Open Data HRO	Hansesdadt Rostock	Bevölkerung, Bildung und Wissenschaft, Geographie, Geologie und Geobasisdaten, Gesetze und Justiz, Gesundheit, Infrastruktur, Bauen und Wohnen, Kultur, Freizeit, Sport und Tourismus, Politik und Wahlen, Soziales, Transport und Verkehr, Umwelt und Klima, Verbraucherschutz, Öffentliche Verwaltung, Haushalt und Steuern, Wirtschaft und Arbeit	Population, education, science, geography, geology, spatialbasedata, law, justice, health, infrastructure, construction, residence, culture, leisure, sport, tourism, politics, elections, socialmatters, transport, traffic, environment, climate, consumerprotection, publicadministration, household, taxes, economy, jobs
Open Data München	Landeshauptstadt München	Bevölkerung, Wirtschaft und Arbeit, Geographie, Geologie und Geobasisdaten, Transport und Verkehr, Kultur, Freizeit, Sport und Tourismus, Soziales, Politik und Wahlen, Infrastruktur, Bauen und Wohnen, Bildung und Wissenschaft, Öffentliche Verwaltung, Haushalt und Steuern, Gesundheit	Population, economy, jobs, geography, geology, spatialbasedata, transport, traffic, culture, leisure, sport, tourism, socialmatters, politics, elections, infrastructure, construction, residence, education, science, publicadministration, household, taxes, health
Open Data ULM		Geographie, Geologie und Geobasisdaten, Bildung und Wissenschaft, Umwelt und Klima, Soziales, Infrastruktur, Bauen und Wohnen, Bevölkerung, Öffentliche Verwaltung, Haushalt und	geography, geology, spatialbasedata, education, science, environment, climate, infrastructure, construction, residence, population, publicadministration,

		Steuer, Verbraucherschutz, Wirtschaft und Arbeit, Transport und Verkehr, Kultur, Freizeit, Sport und Tourismus, Gesundheit, Politik und Wahl, Gesetze und Justiz	household, tax, consumerprotection, economy, jobs, transport, traffic, culture, leisure, sport, tourism, health, politics, elections, notyetcategorized, law, justice
Open Government Data Portal Rheinland-Pfalz	Ministerium des Innern, für Sport und Infrastruktur des Landes Rheinland-Pfalz	Bevölkerung, Bildung und Wissenschaft, GDI-RP, Geographie, Geologie und Geobasisdaten, Gesundheit, Infrastruktur, Bauen und Wohnen, Gesetze und Justiz, Kultur, Freizeit, Sport und Tourismus, Politik und Wahlen, Soziales, Transport und Verkehr, Umwelt und Klima, Verbraucherschutz, Öffentliche Verwaltung, Haushalt und Steuern, Wirtschaft und Arbeit	Population, education, science, GDIRP, geography, geology, spatialbasedata, health, infrastructure, construction, residence, law, justice, culture, leisure, sport, tourism, politics, elections, socialmatters, transport, traffic, environment, climate, consumerprotection, publicadministration, household, taxes, economy, jobs
Open NRW	Innenministerium NRW	Bevölkerung, Bildung und Wissenschaft, Geographie, Geologie, und Geobasisdaten, Gesetze und Justiz, Infrastruktur, Bauen und Wohnen, Kultur, Freizeit, Sport und Tourismus, Öffentliche Verwaltung, Haushalt und Steuern, Politik und Wahlen, Soziales, Transport und Verkehr, Umwelt und Klima, Verbraucherschutz, Wirtschaft und Arbeit	Population, education, science, geography, geology, spatialbasedata, law, justice, infrastructure, construction, residence, culture, leisure, sport, tourism, publicadministration, household, taxes, politics, elections, socialmatters, transport, traffic, environment, climate, consumerprotection, economy, jobs
Transparenzportal Hamburg	Freie und Hansestadt Hamburg	Bevölkerung, Bildung und Wissenschaft, Geographie, Geologie und Geodaten, Gesetze und Justiz, Gesundheit, Infrastruktur, Kultur und Sport, Politik und Wahlen, Soziales, Transport, Umwelt und Klima, Verbraucherschutz, Öffentliche Verwaltung, Wirtschaft und Arbeit	Population, education, science, geography, geology, geodata, law, justice, health, infrastructure, culture, sport, politics, elections, socialmatters, transport, environment, climate, consumerprotection, publicadministration, economy, jobs

Appendix A2: Inter-Catalog Agreement for the 10 German Catalogs Surveyed

	Offene Daten. de	GovD ata	Open Data Berlin	Open Data Köln	Open Data HRO	Open Data München	Open Data ULM	OGDP R	Open NRW	TH
OffeneDat en.de	1	0.9	0.38	0.72	0.93	0.77	0.93	0.90	0.9	0.65
GovData	0.9	1	0.39	0.74	0.97	0.79	0.84	0.93	0.93	0.67
Open Data Berlin	0.38	0.39	1	0.31	0.38	0.33	0.35	0.37	0.36	0.28
Open Data Köln	0.72	0.74	0.31	1	0.77	0.67	0.67	0.75	0.74	0.55
Open Data HRO	0.93	0.97	0.38	0.77	1	0.83	0.87	0.97	0.97	0.7
Open Data München	0.77	0.79	0.33	0.67	0.83	1	0.71	0.8	0.79	0.53
Open Data ULM	0.93	0.84	0.35	0.67	0.87	0.71	1	0.84	0.84	0.65
Open Data Rheinland-Pfalz (OGDPR)	0.90	0.93	0.37	0.75	0.97	0.8	0.84	1	0.93	0.68
Open NRW	0.9	0.93	0.36	0.74	0.97	0.79	0.84	0.93	1	0.67
Transparen zportal Hamburg (TH)	0.65	0.67	0.28	0.55	0.7	0.53	0.65	0.68	0.67	1

Mean (all)[21]: 0.70; Standard deviation: 0.21; Min: 0.28; Max: 0.97; Mode: 0.67

Mean (city): 0.57; Standard deviation: 0.19; Min: 0.28; Max: 0.87; Mode: 0.67

Mean (region)/Standard deviation/Min/Max/Mode: N/A

Mean (country)/Standard deviation/Min/Max/Mode: N/A

[21] The values in the table above are rounded to the second decimal place to ease readability, but the values for the descriptive statistics were computed based on non-rounded values of the jaccard indices.

Appendix B1: 10 Spanish Open Data Catalogs with their Respective Data Categories

Catalog name	Publisher	Data Categories (Spanish)	Data Categories (English)
Datos Abiertos JCYL	Junta de Castilla y León	Ciencia tecnología, comercio, cultura ocio, demografía, deporte, economía, educación, empleo, energía, hacienda, industria, legislación justicia, medio ambiente, medio rural pesca, salud, sector público, seguridad, sociedad bienestar, transporte, turismo, urbanismo infraestructuras, vivienda	Science, technology, commerce, leisure, culture, demography, sport, economy, education, employment, energy, finance, industry, law, justice, environment, ruralenvironment, fishing, health, publicsector, security, socialwelfare, transport, tourism, urbaninfrastructure, livingplace
Datos Abiertos Junta de Andalucía	Junta de Andalucía	Ciencia tecnología, comercio, cultura ocio, demografía, deporte, economía, educación, empleo, energía, hacienda, industria, legislación justicia, medio ambiente, medio rural pesca, salud, sector público, seguridad, sociedad bienestar, transporte, turismo, urbanismo infraestructuras, vivienda	Science, technology, commerce, leisure, culture, demography, sport, economy, education, employment, energy, finance, industry, law, justice, environment, ruralenvironment, fishing, health, publicsector, security, socialwelfare, transport, tourism, urbaninfrastructure, livingplace
Datos Abiertos Madrid	Ayuntamiento de Madrid	Ciencia tecnología, comercio, cultura ocio, demografía, deporte, economía, educación, empleo, energía, hacienda, industria, legislación justicia, medio ambiente, medio rural pesca, salud, sector público, seguridad, sociedad bienestar, transporte, turismo, urbanismo infraestructuras, vivienda	Science, technology, commerce, leisure, culture, demography, sport, economy, education, employment, energy, finance, industry, law, justice, environment, ruralenvironment, fishing, health, publicsector, security, socialwelfare, transport, tourism, urbaninfrastructure, livingplace
Open data Ajuntament de Valencia	Ayuntamiento de Valencia	Medio ambiente, sociedad y bienestar, transporte, urbanismo e infraestructuras, salud, turismo, cultura y ocio, sector público, comercio, economía, hacienda, ciencia y tecnología, educación, seguridad y vivienda	Environment, socialwelfare, transport, urbanplanning, infrastructure, health, tourism, culture, commerce, publicsector, trade, economy, finance, science, technology, education, security, housing
Open data Aragon	Gobierno de Aragón	Ciencia tecnología, comercio, cultura ocio, demografía, deporte, economía, educación, empleo, energía, hacienda, industria, legislación justicia, medio ambiente, medio rural pesca, salud, sector público, seguridad, sociedad bienestar,	Science, technology, commerce, leisure, culture, demography, sport, economy, education, employment, energy, finance, industry, law, justice, environment, ruralenvironment, fishing, health, publicsector, security, socialwelfare, transport,

		transporte, turismo, urbanismo infraestructuras, vivienda	tourism, urbaninfrastructure, livingplace
OpenDataBCN	Ajuntament de Barcelona	Territorio, población, Ciudad y servicios, Economía y empresa y Administración	Territory, population, city, services, economy, business, administration
Open Data Euskadi	Gobierno Vasco	Actividades económicas, Administración Pública, Asuntos Sociales, Cultura, Euskera, Educación, Medio Ambiente, Justicia, Meteorología, Ocio y Turismo, Salud, Seguridad e Interior, Transporte y movilidad, Trabajo y Empleo, Urbanismo y territorio, Vivienda	Economicactivities, publicadministration, socialmatters, culture, basquelanguage, education, environment, justice, meteorology, leisure, tourism, health, security, localgovernment, transport, mobility, jobs, employment, urbanplanning, territory, housing
Open data Gobierno de Canarias	Gobierno de Canarias	Sociedad y bienestar, Sector Público, Medio rural, Empleo, Demografía, Urbanismo e infraestructuras, Turismo, Educación, Salud, Economía, Transporte, Medio Ambiente, Hacienda, Cultura y ocio	Socialwelfare, ruralenvironment, publicsector, employment, demography, urbanism, infrastructure, tourism, education, health, economy, transport, environment, finance, culture, leisure
Open Data Navarra	Gobierno de Navarra	Administración electrónica, Administración pública, Ámbito local, Asuntos sociales, Deporte, Desarrollo rural, Economía y finanzas, Educación, Energía, Estadística, Formación, Industria, Justicia, Juventud, Medio ambiente, Salud, Territorio y urbanismo, Trabajo y Empleo, Tráfico, Transporte, Turismo, ocio y cultura, Vivienda	eAdministration, publicadministration, localgovernment, socialmatters, sport, ruraldevelopment, economy, finance, education, energy, statistics, training, industry, justice, youth, environment, health, territory, urbanplanning, jobs, employment, traffic, transport, tourism, leisure, culture, housing
Portal Open Data Xunta de Galicia	Xunta de Galicia	Información medioambiental, información geográfica, información turística, información cultural, deportiva y de ocio, información sobre transporte, información territorial y de vivienda, información administrativa y legal, información socio-sanitaria, información económica, empresarial y de empleo, información científico-tecnológica	Environmentalinformation, geographicinformation, touristinformation, culturalinformation, sports, leisure, transportinformation, landinformation, housinginformation, administrativeinformation, legalinformation, sociohealthinformation, economicinformation, business, employment, scientificinformation, technologicalinformation

Appendix B2: Inter-Catalog Agreement for the 10 Spanish Catalogs Surveyed

	D.A. JCYL	D.A. Andalucia	D.A. Madrid	Open Data Valencia	Open Data Aragon	Open Data BCN	Open Data Euskadi	Open Data Canarias	Open Data Navarra	Open Data Galicia
D.A. JCYL	1	1	1	0.47	1	0.03	0.27	0.5	0.36	0.05
D.A. Andalucia	1	1	1	0.47	1	0.03	0.27	0.5	0.36	0.05
D.A. Madrid	1	1	1	0.47	1	0.03	0.27	0.5	0.36	0.05
Open Data Valencia	0.47	0.47	0.47	1	0.47	0.04	0.3	0.48	0.29	0
Open Data Aragon	1	1	1	0.47	1	0.03	0.27	0.5	0.36	0.05
Open Data BCN	0.03	0.03	0.03	0.04	0.03	1	0.04	0.27	0.5	0.36
Open Data Euskadi	0.27	0.27	0.27	0.3	0.27	0.04	1	0.28	0.5	0.06
Open Data Canarias	0.5	0.5	0.5	0.48	0.5	0.27	0.28	1	0.30	0.06
Open Data Navarra	0.36	0.36	0.36	0.29	0.36	0.5	0.5	0.30	1	0.04
Open Data Galicia	0.05	0.05	0.05	0	0.05	0.36	0.06	0.06	0.04	1

Mean (all)[22]: 0.34; Standard deviation: 0.31; Min: 0; Max: 1; Mode: 1

Mean (city): 0.17; Standard deviation: 0.20; Min: 0.03; Max: 0.47; Mode: 0.03

Mean (region): 0.37; Standard deviation: 0.30; Min: 0.05; Max: 1; Mode: 0.05

Mean (country)/Standard deviation/Min/Max/Mode: N/A

[22] The values in the table above are rounded to the second decimal place to ease readability, but the values for the descriptive statistics were computed based on non-rounded values of the jaccard indices.

Appendix C1: 10 French Open Data Catalogs with their Respective Data Categories

Catalog name	**Publisher**	**Data Categories (French)**	**Data Categories (English)**
data.gouv.fr	Etalab	Agriculture et alimentation, culture, économie et emploi, éducation et recherche, international et Europe, Logement, développement durable et énergie, santé et social, société, territoires, transports, tourisme	Agriculture, food, culture, economy, jobs, education, science, international, Europe, housing, sustainabledevelopment, energy, health, socialmatters, society, territories, transport, tourism
Data GrandLyon	Métropole de Lyon	Transport, imagerie, citoyenneté, services, culture, localisation, limites administratives, économie, environnement, occupation du sol, urbanisme, équipements, accessibilité, démographie	Transport, imaging, civilrights, services, culture, localization, administrativeborders, economy, environment, landuse, urbanplanning, facilities, accessibility, demography
Montpellier Territoire Numérique	Ville de Montpellier	Environnement, patrimoine/tourisme, économie, urbanisme, arts & culture, numérique, équipements, localisation, santé, politique publique & démocratie, démographie, transport, éducation, vie associative, sports & loisirs, proximité, habitat & aménagement	Environment, heritage, tourism, economy, urbanplanning, art, culture, digital, facilities, localization, health, publicpolicy, democracy, demography, transport, education, communitylife, sports, leisure, proximity, accommodation, planning
Nantes Ouverture des Données	Ville de Nantes	Citoyenneté/Institution, mobilité, santé/social, culture/tourisme, territoires, éducation/formation, environnement, économie, urbanisme, logement, jeunesse	Civilrights, institutions, mobility, health, socialmatters, culture, tourism, territories, education, training, environment, economy, urbanplanning, housing, youth
Open Data Nice Côte d'Azur	Métropole Nice Côte d'Azur	Accessibilité, administration électronique, aménagement du territoire, citoyenneté, culture, économie, éducation, environnement, événementiel, loisirs, santé, sécurité, sport, tourisme, transport	Accessibility, eGovernment, landsettlement, civilrights, culture, economy, education, environment, events, leisure, health, security, sport, tourism, transport

Open Data Bordeaux	Mairie de Bordeaux	Cadre de vie, citoyenneté et administration, culture, sports et loisirs	Livingplace, civilrights, administration, culture, sports, leisure
Open PACA	Région Provence-Alpes-Côte d'Azur	Administration-marchés publics, agriculture, aménagement du territoire, citoyenneté-démocratie, culture-patrimoine, économie-emploi, éducation-recherche, environnement-énergie, équipement collectif, finances, fonds institutionnels, formation-apprentissage, information-TIC, international-Europe-Bassin méditerranéen, Marseille-Provence 2013, Mer-Littoral, réseau de distribution, santé-social-sport, secteur public, tourisme, transports, urbanisme	Administration, publiccontracts, agriculture, landsettlement, civilrights, democracy, culture, heritage, economy, jobs, education, science, environment, energy, publicfacilities, finances, institutionalfunds, training, apprenticeship, information, TIC, international, Europe, Mediterraneanbasin, MarseilleProvence2013, sea, coastline, distributionnetwork, health, socialmatters, sport, publicsector, tourism, transport, urbanplanning
ParisData	Mairie de Paris	Services, déplacements, urbanisme, citoyens, culture, environnement, administration, finances, commerces	Services, trips, urbanplanning, citizens, culture, environment, administration, finances, trade
Rennes métropole en accès libre	Service Innovation Numérique, Hôtel de Rennes Métropole	Accessibilité, citoyenneté, culture, culture:agenda, culture:annuaire, culture:statistiques, données budgétaires, environment, equipements, logement, référentiel géographique, sports et loisirs, stationnement, transports	Accessibility, civilrights, culture, cultureagenda, culturedirectory, culturestatistics, budget, geographicreferenceframe, facilities, housing, sport, leisure, parking, transport
Toulouse Métropole Data	Mairie de Toulouse	Citoyenneté, culture, transport, finance, statistiques, sport, aménagement du territoire, urbanisme, bâtiments, equipements, logement, environnement, enfance, patrimoine, services, aménagement, tourisme, economie	Civilrights, culture, transport, finance, statistics, sport, landsettlement, urbanplanning, construction, facilities, housing, environment, childhood, heritage, services, planning, tourism, economy

Appendix C2: Inter-Catalog Agreement for the 10 French Catalogs Surveyed

	data.gouv.fr	Data Grand Lyon	Montpellier T.N.	Nantes O.D.D.	Open Data Nice	Open Data Bordeaux	Open PACA	Paris Data	Rennes M.E.A.L	Toulouse M.D.
data.gouv.fr	1	0.10	0.18	0.32	0.22	0.04	0.32	0.04	0.10	0.16
Data Grand Lyon	0.10	1	0.29	0.21	0.26	0.11	0.14	0.21	0.22	0.33
Montpellier T.N.	0.18	0.29	1	0.23	0.28	0.12	0.21	0.10	0.13	0.29
Nantes O.D.D.	0.32	0.21	0.23	1	0.30	0.10	0.25	0.14	0.12	0.27
Open Data Nice	0.22	0.26	0.28	0.30	1	0.17	0.25	0.09	0.26	0.32
Open Data Bordeaux	0.04	0.11	0.12	0.10	0.17	1	0.08	0.15	0.17	0.09
Open PACA	0.32	0.14	0.21	0.25	0.25	0.08	1	0.13	0.09	0.23
Paris Data	0.04	0.21	0.10	0.14	0.09	0.15	0.13	1	0.05	0.17
Rennes M.E.A.L	0.10	0.22	0.13	0.12	0.26	0.17	0.09	0.05	1	0.23
Toulouse M.D.	0.16	0.33	0.29	0.27	0.32	0.09	0.23	0.17	0.23	1

Mean (all)[23]: 0.18; Standard deviation: 0.08; Min: 0.04; Max: 0.33; Mode: 0.09

Mean (city): 0.16; Standard deviation: 0.07; Min: 0.05; Max: 0.29; Mode: 0.05

Mean (region)/Standard deviation/Min/Max/Mode: N/A

Mean (country)/Standard deviation/Min/Max/Mode: N/A

[23] The values in the table above are rounded to the second decimal place to ease readability, but the values for the descriptive statistics were computed based on non-rounded values of the jaccard indices.

Appendix D1: 10 Open Data Catalogs from the UK with their Respective Data Categories

Catalog name	Publisher	Data Categories (Catalogue)	Data Categories (API)
Birmingham DataFactory	Birmingham City Council	Travel and transport, council business, your local area, locations, environment, education	Travel, transport, councilbusiness, yourlocalarea, locations, environment, education
Bournemouth Data Stream	Bournemouth Borough Council	Tourism and population, traffic and geography, amenities, services and buildings, health and hygiene, finance	Tourism, population, traffic, geography, amenities, services, buildings, health, hygiene, finance
Data.gov.uk	UK Government	Environment, towns & cities, mapping, government, society, health, government spending, education, business & economy, transport	Environment, towns, cities, mapping, government, society, health, governmentspending, education, business, economy, transport
Data- Liverpool City Council	Liverpool City Council	Economy, population, education and skills, health, deprivation, labour market, housing, crime	Economy, population, education, skills, health, deprivation, labourmarket, housing, crime
Edinburgh Open Data Portal	City Council Edinburgh	Environment, health, education, transport, tourism, leisure, community	Environment, health, education, transport, tourism, leisure, community
Leeds Data Mill	Leeds City Council	Local services, transport, education, housing, health, business and economy, art and culture, geospatial, licenses, tourism, sport, transparency	Localservices, transport, education, housing, health, business, economy, art, culture, geospatial, licenses, tourism, sport, transparency
London Datastore	Greater London Authority	Demographics, employment and skills, transparency, environment, housing, health, transport, business and economy, education, planning, crime and community safety, young people, sport, art and culture, championing london, london 2012	Demographics, employment, skills, transparency, environment, housing, health, transport, business, economy, education, planning, crime, communitysafety, youngpeople, sport, art, culture, championinglondon, london2012
Open Data Bristol	Bristol City Council	Community, education, energy, environment, finance, government, health, internet of things, land use, mobility, reference, safety	Community, education, energy, environment, finance, government, health, internetofthings, landuse, mobility, reference, safety

OpenDataNI	The Open Data Team	Property & land, population & society, transport, health, finance, environment & agriculture, economy, industry & employment, tourism, leisure, culture & arts, education	Property, land, population, society, transport, health, finance, environment, agriculture, economy, industry, employment, tourism, leisure, culture, arts, education
Sheffield City Council Open Data	Sheffield City Council	Economy, education, environment, governance, health, heritage, housing, population, transport	Economy, education, environment, governance, health, heritage, housing, population, transport

Appendix D2: Inter-Catalog agreement for the 10 Catalogs from the UK Surveyed

	Birmingham D.F	Bournemouth D.S.	Data.gov.uk	Data Liverpool	Edinburgh O.D.P.	Leeds Data Mill	London Datastore	Open Data Bristol	Open Data NI	Sheffield Open Data
Birmingham D.F	1	0	0.19	0.07	0.27	0.11	0.13	0.12	0.14	0.23
Bournemouth D.S.	0	1	0.05	0.12	0.13	0.09	0.03	0.1	0.17	0.12
Data.gov.uk	0.19	0.05	1	0.17	0.27	0.24	0.23	0.2	0.26	0.31
Data Liverpool	0.07	0.12	0.17	1	0.14	0.21	0.26	0.10	0.18	0.38
Edinburgh O.D.P.	0.27	0.13	0.27	0.14	1	0.24	0.17	0.27	0.33	0.33
Leeds Data Mill	0.11	0.09	0.24	0.21	0.24	1	0.42	0.08	0.24	0.28
London Datastore	0.13	0.03	0.23	0.26	0.17	0.42	1	0.10	0.23	0.26
Open Data Bristol	0.12	0.1	0.2	0.10	0.27	0.08	0.10	1	0.16	0.17
OpenData NI	0.14	0.17	0.26	0.18	0.33	0.24	0.23	0.16	1	0.3

Sheffield Open Data	0.23	0.12	0.31	0.38	0.33	0.28	0.26	0.17	0.3	1

Mean (all)[24]: 0.19; Standard deviation: 0.09; Min: 0; Max: 0.42; Mode: 0.12

Mean (city): 0.17; Standard deviation: 0.09; Min: 0; Max: 0.38; Mode: 0.12

Mean (region): 0.23; Standard deviation: 0; Min/Max/Mode: N/A

Mean (country)/Standard deviation/Min/Max/Mode: N/A

[24] The values in the table above are rounded to the second decimal place to ease readability, but the values for the descriptive statistics were computed based on non-rounded values of the jaccard indices.