Linked Open Data Technologies for Publication of Census Microdata

Gustavo Pabón, Claudio Gutiérrez, Javier D. Fernández, and Miguel A. Martínez-Prieto
Department of Computer Science, University of Chile, Avenida Blanco Encalada 2120, 837-4059 Santiago, Chile. E-mail: {gpabon,cgutierr,jafernan,mmartine}@dcc.uchile.cl

Javier D. Fernández and Miguel A. Martínez-Prieto
Dataweb Research, Department of Computer Science, University of Valladolid, Escuela Técnica Superior de Ingeniería Informática (E.T.S.I.I.), Campus Miguel Delibes, Paseo de Belén, número 15, ES-47011, Valladolid, Spain. E-mail: {jfergar,migumar2}@infor.uva.es

Censuses are one of the most relevant types of statistical data, allowing analyses of the population in terms of demography, economy, sociology, and culture. For fine-grained analysis, census agencies publish census microdata that consist of a sample of individual records of the census containing detailed anonymous individual information. Working with microdata from different censuses and doing comparative studies are currently difficult tasks due to the diversity of formats and granularities. In this article, we show that novel data processing techniques can be applied to make census microdata interoperable and easy to access and combine. In fact, we demonstrate how Linked Open Data principles, a set of techniques to publish and make connections of (semi-)structured data on the web, can be fruitfully applied to census microdata. We present a step-by-step process to achieve this goal and we study, in theory and practice, two real case studies: the 2001 Spanish census and a general framework for Integrated Public Use Microdata Series (IPUMS-I).

Census Data Dissemination and Open Data

Let us quote in full this reflection of the United Nations Statistics Division (2010):

> A census is not considered complete until the collected information is made available to users in the form suited to their needs. Recent decades have witnessed an increasing demand, by census data users, for a broad range of census products and services. Each census product, and its media of dissemination, offers respective advantages and limitations. In most instances, the various census products and services complement each other and can provide effective ways to reach out to a wide range of users in the public and private sectors. With the widespread use of microcomputers and growing access to the Internet, an increasing number of data users prefer to obtain census data in electronic media rather than in printed form. (p. 37)

One of the main challenges in this area is the computational support to fulfill the demands posed by these trends. A recent survey shows that most countries publish their census results in electronic media. The preferred media are (in this order): static website (95%), CD/DVD (80%), geographic information systems (GIS) web-based tools (59%), and online database (53%) (United Nations Statistics Division, 2010).

From a computational point of view, current web trends indicate that increasingly people consider that a piece of information is "available" or "public" only if it can be obtained online. This follows technological trends that indicate that intermediate "artifacts" of information (printed, CD/DVD, flash memory, etc.) are rapidly becoming obsolete. Our research work relies on this fact and examines the behavior of web technologies regarding publication of census microdata.

Microdata

Results of censuses are disseminated via tabulations of different topics and granularities and also, in many cases, via microdata samples. The United Nations Statistics Division (2010) defines *microdata* as

> the individual records that contain information collected in a census or survey on the characteristics of each person, household and housing unit. A census microdata sample allows researchers to know, simultaneously, all the personal characteristics of every individual in the sample. (p. 39)

Statistical organizations have diverse policies regarding the distribution of basic census information. Usually there is a standard distribution policy for data for the general public and specialized, more restricted case-by-case policies for

Received July 24, 2012; revised October 31, 2012; accepted November 1, 2012

 DOI: 10.1002/asi.22876

providing more fine-grained tabulations for specialized types of users, such as researchers, policy makers, teachers, and students.

The restrictions for not having detailed access to census data as a general policy can be grouped in two areas of concern. One is the economic costs of processing the data. Different specialized users hardly need the same views and the same granularity of data. Thus, elaborating tabulations that meet special requests (e.g., detailed tabulations, data on specific topics and subgroups, etc.) is costly and thus can only be made for a fee, usually on a cost recovery basis. The second consideration, closely related to the previous one, is that of privacy and legal regulations. Different areas and countries have diverse regulations on this matter. Censuses have sensitive information about people. Thus, this information can be released only under strict privacy and statistical constraints. For the case of microdata, only samples of the full records, assuring anonymity, are released.

There is a growing tendency to disseminate census microdata. Today 77 countries or areas (59% of a total of 131 in the study) disseminate census microdata, and Latin America is the region with the highest percentage of countries that disseminate microdata (United Nations Statistics Division, 2010). However, these microdata are currently presented in different formats, structures, and levels of detail that hinders comparative analysis (Esteve & Sobek, 2003) and do not stimulate interoperability. The social, economic, and scientific impacts of microdata would increase if they could be openly available in standard formats for multiple types of users, agencies, researchers (in a wide range of fields), and citizens. Besides, the use of common schemas would allow microdata to be harmonized across countries and time periods, enabling comparative analysis. Although there exist some particular projects to harmonize censuses of different countries, such as the Integrated Public Use Microdata Series International (IPUMS-I) (Hall, McCaa, & Thorvaldsen, 2000) and the North Atlantic Population Project,[1] they produce isolated datasets, noninteroperable between them and not linkable to other sources. Nevertheless, data interoperability between countries is a well-known, yet difficult to solve, problem in almost every area; harmonization of trade data (Trade and Investment Division, 2012), environmental information and services (Perego et al., 2012), geospacial data,[2] or data sharing policies fighting against transnational organized crime[3] are just a few examples of standardizing datasets across governments. One of the latest efforts is the Interoperability Solutions for European Public Administrations program (ISA),[4] aimed to foster interoperability between public administrations, yet highlighting that "legal compatibility, semantic interoperability, technical aspects of information systems, organisational cooperation and a favourable political climate are all necessary to make interoperable public services a reality." Other efforts, such as the Guardian's World Government Data website,[5] are aimed at providing horizontal access to government catalogs. However, the catalogs have different standards, limiting flexibility, reusability, and extensibility as it substantially increases the required effort for each new catalog addition (Maali, Cyganiak, & Peristeras, 2010). Semantic interoperability addresses this issue, extending the capabilities of applications on the basis of the collected and harmonized semantics-aware data structures (Perego et al., 2012).

Semantic Web Technologies Applied to Census Microdata

This article deals with the application of the latest semantic web technologies to the publication of census microdata in order to (a) use a semantic data model that formalizes the representation while hindering the different levels of detail (semistructured), (b) link census data with external sources, and (c) integrate censuses across countries.

Given the early stage of these technologies, their application to any particular scenario, as studied in this article, requires a prior study to establish methodologies and procedures for specific data modeling and processing. Although semantic web technologies have been applied to related problems, such as the publication of historical data (Meroño-Peñuela et al., 2012) or recent projects on census publication, such as CEDAR (Dutch census data in a web of global cultural and historic information),[6] there is no previous work, to the best of our knowledge, outlining the general steps required for publishing census microdata over the semantic web infrastructure. It is worth noting that our concern here is neither the statistical issues of what types of data should be published, nor the important legal and privacy restrictions that should apply. Acknowledging that several countries are already publishing census microdata, we propose to enhance their publication and consumption by means of semantic web technologies.

The core of our approach lies in how microdata are structured and represented for their publication. Censuses, like other kinds of statistical or historic information, can be seen as a number of entities, attributes of these entities, and relationships between entities. For instance, a census could describe an individual ("person" entity) who is 40 years old ("age" attribute) and has one son (relation to another "person" entity). This can be formalized and effectively modeled using semantic web technologies.

The choice of an adequate data model is essential to address the requirements of data processing of a specific area. In this article, we follow the standard model for publication of semantic data at web scale, namely, the standard resource description framework (RDF) (Malona & Miller, 2004). RDF has flexibility for handling semistructured information (entities with different levels of detail). In summary, RDF describes entities through triples (subject, predicate, value), in which the subject is the entity being described, the predicate is a property applied to it, and the value is the property value. For instance, (individual #1, age, 40) and (individual #1, has_son, individual #2) draw two RDF triples. This can be seen as a graph of knowledge in which entities and values are linked via labeled edges, that is, the predicates are the labels. How these predicates are formally defined constitutes the second main decision. In our case, an

DOI: 10.1002/asi

appropriate set of vocabularies must be provided to cover the diversity of different census semantics. Thus, the original plain representation of microdata is turned into a rich semantic graph in which relationships are undoubtedly understood by people with diverse backgrounds, but also by automated systems that can easily parse and consume the census knowledge for different purposes.

Open Publication of Semantic Census Microdata

Once the microdata are semantically described, how this representation is finally published on the web constitutes the second part of our approach. There is today a growing tendency toward open data and open access, but making the data available to increasing groups of people brings several challenges. Besides the legal and political issues, one of the most challenging is the issue of the formatting of data. However, our prior decisions guarantee microdata to be delivered in a standardized way, opening their publication to the World Wide Web (WWW) showcase. Nowadays, the project Linked Open Data[7] (LOD) is the reference for exposing semantic datasets in the WWW, and its application to government data has created a new field, Linked Open Government Data[8] (LOGD), pushing the interconnection, openness, modularity, and scalability of government data to a new dimension. Summarizing (a detailed review is given in the next section), LOD establishes a set of recommendations to publish the aforementioned RDF data, encouraging the establishment of connections between different data sources on the web (Bizer, Heath, Idehen, & Berners-Lee, 2008). For instance, a census dataset can include an RDF sentence (individual #1, born, geoInformation:Madrid), where geoInformation:Madrid is a link to another dataset (geoInformation) which expands the information about Madrid. This simple approach has lifted traditional hyperlinks to a new level: From a web of isolated pages we are entering the era of the "web of linked data."

The fulfillment of our approach shows how census microdata can take advantage of this web of linked data, but also how these microdata can also contribute to the globalized knowledge that it makes available. First, we describe a step-by-step process proving that it is feasible and advantageous to develop a general infrastructure to publish microdata censuses using semantic standards, and this process is then applied to two real-world case studies:

- The case of the 2001 Spanish Census is an example of how linked open census microdata can be published using a bottom-up procedure. That is, the semantic web technologies are directly applied to the model and the vocabularies originally agreed by the Spanish agencies. Microdata published following this procedure preserve the original agency guidelines, but enhances the final representation to take advantage of the semantic web possibilities.
- The IPUMS-I conversion goes a step further and provides a top-down procedure that allows microdata to be exposed, in a standardized way, for any potential publisher agency. This procedure enables census integration across countries and time periods because it regards the IPUMS-I proposal to develop a universal vocabulary that harmonizes, over the RDF data model, the entity properties and relations mostly used in censuses.

These practical results show that, although the harmonized approach (IPUMS-I case) is sufficiently universal, the LOD conversion also benefits a particular case such as the Spanish Census, and it is also a valid strategy for other types of microdata that are not covered by IPUMS-I (e.g., local censuses, sociological surveys).

As a whole, this work shows that there are easy and feasible solutions to implement conversion processes, encouraging the adoption of semantic web technologies for the publication of census microdata in particular, and open government data in general.

The rest of the article of the paper is organized as follows: The State of the Art section contextualizes historically census microdata publication and provides a brief review about Open Government and Open Data trends, establishing the path to LOGD and census publication. In the Linked Open Data Conversion of Census Microdata section, we detail the semantic web technologies and their relevance in the LOD principles, showing how census microdata can take advantage of their adoption. This background is then used for describing the specific steps involved in the conversion process from census microdata to LOD. The Case of the Spanish Census section reports our first case study. We model the specific Spanish Census scheme, develop its dedicated vocabulary, and build demo applications over the translated microdata. The RDF Publication of International Census Microdata section complements this practical experience from the perspective of harmonized data. We also set its data scheme and develop a vocabulary of integrated census microdata based on the IPUMS-I project. Last but not least, we provide tools for converting particular instantiations of the IPUMS-I to the principles of LOD. Finally, in Conclusions and Future Work we discuss our findings and delineate further developments.

State of the Art

Censuses have had an unquestionable importance throughout history. As stated by Anderson (1988), "the census is deeply embedded in American political life through myriad apportionment mechanisms; it is also a crucial maker for American history" (p. 2). This assertion applies not only to the United States but to almost all countries around the globe. The census information and its published results are defined by political, social, and technical reasons, but at the same time the census information is crucial to makers of world history.

In the following we focus on putting census microdata in context and gaining insights into LOD publication.

Historical Context of Census Microdata

To give a historical context of census microdata, we briefly present three different national cases of efforts to

DOI: 10.1002/asi

integrate and make publicly available census microdata. We chose the Argentinean Census because it pioneered historic census microdata publication (contents from 1869 and 1895 were recovered in 1967); the Chinese case because it is an interesting example of social and demographic politics in the most populous country in the world (microdata were published first in 1982); and the IPUMS-USA because it plants the seed of the IPUMS-I project for universal microdata harmonization (IPUMS-USA released microdata from the 1960 census in 1964). For each case we summarize (following the work of Hall et al. (2000), where the reader can find a more complete review) the practices developed, the problems encountered in the handling of such volume data, and how the technology (data storage, processing, and dissemination), and sociopolitical matters, in some interesting way, affected its production and consumption.

Argentina: The first national historical census microdata (Somoza & Lattes, 1967). In 1967 two Argentine demographers, Jorge L. Somoza and Alfredo E. Lattes, completed the first national computerized samples of historical census microdata. They used punch cards and worked from manuscripts of the original enumeration sheets of the Argentine censuses of 1869 and 1895, producing more than 200,000 cases at a cost of just USD $21,000. Unfortunately, the original tapes were lost and the punch cards were recycled long ago, leaving just the results and published work.

Besides this, we know now how Somoza and Lattes made this remarkable work. For 6 months 12 researchers selected the samples, transcribed, coded cases, checked, and rechecked the cases. The key-pushing process took only 3 months. After the key-pushing and verification process, four researchers spent an additional 6 months in data processing tasks: dealing with missing data, checking and rechecking for errors and omissions, computing new national tables, evaluating results, and drafting documentation.

The entire project was completed October 1, 1967. The final data of 1869 and 1895 censuses were freely published as ASCII text files, of three and four megabytes, respectively. Their compressed representation fits on a conventional floppy disk.

China: Censuses of 1982 and 1990 (Hall et al., 2000). China's history of public use census microdata is quite short, even though it has the largest population and a long history of censuses. It was just in the 1982 census that microdata began to be stored in machine-readable format and made available for public use. Chinese censuses are considered national efforts, mobilizing millions of participants.

The 1982 census records were processed by a central computer center using mini- and mainframe computers, while the 1990 census was largely processed with personal computers. At the first stages of the data analysis, the 1990 census used 3,690 PCs. These microdata have not been made public outside China, with very few exceptions of collaborative research with foreign institutions.

IPUMS-USA: Integrated Public Use Microdata Series for the United States (Hall et al., 2000). The U.S. Bureau of Census made available in 1964 the first public use census microdata sample of the 1960 census. This was the result of an effort to meet the needs of scholars who increasingly requested specialized tabulations not included in the published census volumes. It was an immediate success even though its density was quite small, 1-in-1,000 extract, yielding about 180,000 person-records. Such data analysis was not an easy task given the modest computer capacity available in 1964. The low density did not help with the analyses of small population groups, and for privacy matters the records do not include any geographic information. These problems were addressed by the 1970 public samples, and a new 1960 sample was produced with a density of 1-in-100 and with integrated variables that allows cross-analyses including both censuses. By the late 1970s, the public use samples for 1960 and 1970 had become essential tools of American social scientists. The Census Bureau also released public use samples from the 1940 to 1990 censuses, but unfortunately, each sample was coded entirely differently and had separate documentation.

In 1992, the Minnesota Population Center of the University of Minnesota started the IPUMS project (Integrated Public Use Microdata Series) to build a coherent national database of a high-precision individual-level census data describing the characteristics of the U.S. population between 1850 and 1990.

Preliminary samples exist for all census years with the exception of 1890, which was destroyed in a fire, and 1930, which was unavailable until 2002 for privacy restrictions. The sample density of the early censuses had to be variable depending on special-interest population subgroups to achieve representative sampling (i.e., Black and Indian population); it was achieved by oversampling strategies. A key feature of the samples constructed as part of the project is their harmonization (use of a common coding system for all variables) to guarantee integration on multiple census years. An important feature of the IPUMS is the allocation of missing, illegible, and inconsistent data in all the censuses from 1850 to 1920 using logical editing and "hot deck" probabilistic editing.

As a result of the success of the IPUMS project, a new project, called IPUMS International (IPUMS-I), started in 1999 as an effort to inventory, preserve, integrate, and disseminate census microdata around the world. According to the last activity report (Minnesota Population Center, 2011) there are 185 fully processed samples of 63 countries into the IPUMS-I database.

Linked Open Data Publication

The notion of open government can be traced back to the Enlightenment and its policies toward freedom of the press advocated by the American and French revolutions (Lathrop & Ruma, 2010). Nevertheless, freedom of information was an achievement of the last century, and many countries did not pass corresponding legislation until the early 21st century. For instance, the United States passed its Freedom

DOI: 10.1002/asi

of Information Act (FOIA) in 1966. This law allows previously unreleased information (and documents controlled by the U.S. government) to be fully or partially disclosed. The "OPEN Government Act" amended the FOIA in 2007, highlighting, among many other changes, the specific recognition of electronic media as a document source. An updated survey of the current Open Government status was made by Lathrop and Ruma (2010), who argue that

> open government is emerging as a new kind of public sector organization which opens its doors to the world; co-innovates with everyone, especially citizens; shares resources that were previously closely guarded; harnesses the power of mass collaboration; drives transparency throughout its operations; and behaves not as an isolated department or jurisdiction, but as something new a truly integrated and networked organization. (p. 16)

Currently, Open Data is one of the most important movements within Open Government. It basically relies on the idea that government data should be publicly available to enable its reuse by different stakeholders. It is worth noting that, in the internet era, the same openness that enables transparency also enables innovation (Lathrop & Ruma, 2010). This philosophy is not exclusive for government data, and this opening movement is also arising in scenarios related to libraries, culture, economics, or science (Uhlir & Schröder, 2007) among others.

Open data and open government movements have been gaining momentum thanks to initiatives such as the Open Data Foundation and the Open Knowledge Foundation.[9] One of the first actions of the newly elected President of the United States in 2009 was the Open Government Memorandum (Obama, 2009), which calls for more transparent, participatory, and collaborative government.

The mutual understanding between (linked) open data and open government initiatives began to be advocated, the creator of the web, Tim Berners-Lee, being among the first to propose such a program (Berners-Lee, 2009).

LOD is an initiative that proposes to use the WWW to connect related data that were not previously linked, or use the WWW technologies to lower the barriers to link data currently linked using other methods. Linked Data is also "about employing the Resource Description Framework (RDF) and the Hypertext Transfer Protocol (HTTP) to publish structured data on the Web and to connect data between different data sources, effectively allowing data in one data source to be linked to data in another data source" (Bizer et al., 2008, p. 1265).

Therefore, LOD uses the WWW infrastructure and leverages it for publishing and linking any type of data, overcoming the traditional uses of linkage restricted to documents and web pages. Nevertheless, the meaningful interlinking of this universe of data is challenging (Hausenblas & Karnstedt, 2010). Currently, it relies on the following four principles (Berners-Lee, 2006): (a) use URIs for naming resources; (b) use HTTP URIs so that people can look up those names; (c) provide useful information using standards, such as RDF and SPARQL (Prud'hommeaux & Seaborne, 2008), when someone looks up a URI; and (d) include links to other URIs so that they can discover other related resources on the web. Behind these technical recommendations underlies the idea of releasing data under open licenses, which permits the reuse and crossing of published data.

LOD principles of modularity, openness, and scalability fit with government's goals of efficiency, transparency, and service to citizens, particularly regarding information. That is why LOD appears as a natural interconnection bus to ease access to public-sector data (Bizer, 2009). In fact, the W3C eGovernment Interest Group and the Government Linked Data Working Group are starting to work closely with governments in highlighting the benefits of LOGD (Alonso, Novak, & Acar, 2009).

Governments around the world are beginning to understand this potential (Bizer, 2009). The U.S. government, through the Data.gov site, hosts several hundred thousand datasets offering citizens and developers the possibility of creating applications from these government data. Its main research is focused on semiautomatic or low-cost LOGD conversion (from CSV and other formats) and publication (Ding et al., 2010). A complete background of Data.gov, as well as the current and planned use of linked data for organizing its knowledge and vocabularies, is described in Hendler, Holm, Musialek, and Thomas (2012). The U.K. government is also making strong efforts in this direction. Currently there are thousands of datasets at data.gov.uk, with particular emphasis on Linked Data integration. The experience of deploying this public data catalog, its research challenges, and the lessons for governments and technical communities can be found in Shadbot et al. (2012). The U.S. and U.K. Data.gov experiences are among the most representative examples of LOGD, but this movement is growing progressively in numerous countries and states. By October 2012, 47 countries participated in the Open Government Partnership initiative[10] and 10 more are ready to join.

In this article, two strategies have been explored to make census microdata publicly available using LOD principles.

The first consists of modeling and translating directly the microdata into a web language such as RDF. An example of this approach—but for aggregated data—is the 2000 U.S. Census,[11] which is part of the LOD cloud. It comprises geographical data and detailed population statistics by region. Note that in this case the original source is aggregated data. Developing such conversion for microdata presents other types of challenges. The 2001 Spanish Census conversion to RDF (Fernández, Martínez-Prieto, & Gutiérrez, 2011) is a good example of employing microdata. These two cases are examples of bottom-up procedures.

The second strategy is to proceed top-down, that is, to get a general model that abstracts the common features of different censuses and use it to produce particular instances.

DOI: 10.1002/asi

One of the most comprehensive developments in this direction is the work of Integrated Public Use Microdata Series, International (IPUMS-I) (Hall et al., 2000). The goals of this project are to inventory machine-readable census microdata, to preserve census microdata identified as at risk, to create an integrated international census database with a harmonized system of concepts, and to disseminate integrated microdata via the web. To the best of our knowledge, there have been no efforts to integrate this information under Linked Data principles.

Linked Open Data Conversion of Census Microdata

As explained in the introduction, the application of these young technologies to a new problem deserves a study defining methodologies and procedures for specific data modeling and processing. We define, at the end of this section, our own step-by-step process for transforming raw census microdata into LOD. Nevertheless, we first analyze how the specific census microdata particularities impact the semantic web infrastructure.

Semantic Microdata Modeling

Although census microdata are gathered in a structured way and it is traditionally seen as a structured source of information, this view entails several disadvantages. Let us consider that a census describes entities (person, family nucleus, house, etc.), their attributes (age, studies, number of members in one family, rooms of a house, etc.), and relations between entities (mother of another person, main person in a family nucleus, owner of house, etc.). First, each type of entity is mainly related to different attributes, but repeated attributes can have a different meaning depending on the entity (e.g., "structure" of a family differs from "structure" of a house). Second, entities are described in different levels of detail and do not usually have a value for all the attributes (e.g., a census could gather the four previous types of jobs of every person, but not all the persons would have them). Finally, each entity can also have different types and numbers of relationships (e.g., a relationship between two houses is rare, and one person can be related to more than one house or just one). A structured model (such a relational scheme) could solve "somehow" some of these deficiencies, but others are clearly out of the scope of a structured view and results in a poor representation (e.g., having several NULL attributes in the description of a person). Thus, a flexible semistructured data model better fits these particularities of census microdata.

Although the need to capture the semantics by this data model was already outlined, this becomes a main requirement when harmonizing censuses between different periods or countries. In this case, together with format harmonization, the data model has to capture the very different levels of detail of the entities and relationships and to deal with a particular meaning of the attributes of each census sample.

In such a scenario the resource description framework (RDF) (Malona & Miller, 2004) is an ideal candidate because of its flexibility for representing the semantics of semistructured information with entities (resources) with different levels of detail. It is based on atomic triple units. An RDF triple is described as a statement of the form (subject, predicate, value) in which the subject is the entity (also called *resource*) being described, the predicate is a property applied to it, and the value is the concrete value. For instance, (registry #2, type, person) and (registry #2, gender, female) model that the second registry of our census represents a person and this person is a woman. Note that this type of semantic representation is supposed to avoid ambiguity in order to harmonize different collections and, when possible, to be processed automatically. Thus, "gender," "sex," "has_gender" are different possibilities that should be harmonized. To do so, RDF provides a clear semantic in which subjects and predicates are modeled with URIs (Uniform Resource Identifiers), so that if two entities use the same URIs we can assume that they convey the same meaning. This way, the previous example could be formally written as (http://example.org/registry#2, rdf:type, "Person") and (http://example.org/registry#2, foaf:gender, "female"), in which rdf: and foaf: are well-known vocabularies providing standard properties. As can be seen, the correct URI design for the entities, vocabulary selection, and vocabulary creation for the attributes (for those properties not found elsewhere), and the establishments of relationships, are the main cornerstones when converting census microdata to RDF.

A final, but no less important, observation refers to the scalability of this data model when it is applied for modeling any particular microdata collection. On one hand, when the data designer builds the vocabulary, it can freely model any kind of feature and relationship. Note that this URI-based identification enables each predicate in the vocabulary to be undoubtedly understood by simply de-referencing it. On the other hand, the labeled graph structure underlying the RDF model allows new semantics to be easily added in advance. Let us suppose, for instance, that a new property is recorded for the "person" entity. In this case, all values for this property are stored by adding edges (labeled with the URI identifying the new property) from the entity nodes to their corresponding values.

Semantic Microdata Publication

These important issues must be dealt with by considering the final dissemination target and the potential consumption of information. As explained, the project LOD is the reference for exposing open semantic datasets on the web.

Linked open census microdata refers to providing census microdata in RDF following the aforementioned Linked Data principles. The publication of linked open census microdata provides, therefore, several advantages: (a) The philosophy makes use of de-referenceable URIs for each resource so that the information is referenceable across the web; (b) the statistical variables are also URIs,

DOI: 10.1002/asi

hence the use of a common vocabulary for harmonized parts of a census is encouraged; (c) the data and their models can be easily extended by adding, deleting, or updating RDF triples (units of information); (d) the use of data standards provides access to a set of standard tools and the support of an emerging research community; and (e) census data can be potentially integrated with noncensus data published under the Linked Data philosophy.

The latest LOD cloud estimations[12] show that more than 30 billion RDF triples are being shared and increasingly linked. Anyone can publish data on the web of Linked Data, comprising large self-describing RDF datasets from fields such as bioinformatics, social networks, geographic locations, films, and so on. Thus, millions of entities are connected and the open philosophy assures that applications can discover new data at run-time by following links. Census data have a great integration potential with other datasets in the LOD. For instance, from an anonymous individual census record, external links can be followed to get geographic or wiki information on the place of birth (or current housing), wiki information of events happening in a period of time referred to in the census record, information about music, films, publications related to the data, and so on. It is worth mentioning that most of these datasets are already present in the cloud (such as Geonames, DBpedia, ACM, and DBLP publications, etc.), whereas others can potentially be available (we review the state of the art in LOGD publication in the previous section).

The Step-by-Step Process

The conversion of census microdata into LOGD is a process that can be generalized in a series of six steps leading to the adaptation of the information to the required form and format. In the next sections we provide two different case studies, instantiations, of these steps.

Corpus definition. Governments might provide several sources of information in the same microdata file, or, on the other hand, they might disseminate samplings in diverse ways and even formats. Thus, a first study must be carried out to narrow the data to be finally converted.

Data modeling. The exposed microdata tend to organize the information by records, that is, one record per person. However, within each record census information can be modeled in many ways; other entities could have their own associated information, such as country regions, buildings, or family nucleuses. The second phase of the data processing is to identify the schema of the data used by the corresponding administration, its study, and analysis.

Vocabulary definition. Data must be converted to a standard format, RDF. Then the identified entities and their properties (attributes) must be named with a unique URI. For instance, each anonymous person or building, and each property, such as *age* or *year of building*, should have a URI. Following the Linked Data principles, these URIs must be de-referenceable (accessible) via the HTTP protocol. It is worth mentioning that there exist two standard policies for entity and property naming: slash URIs and hash URIs approaches (Sauermann & Cyganiak, 2008). They establish a common scheme to be followed in the assignment of URIs.

Linked data links. The Linked Data fourth principle recommends making links elsewhere to connect the data to conform to a linked data web. Thus, potential out-links, from the data to well-known datasets, have to be identified. This process is particularly peculiar in the case of a census because of the special entities and properties used. Entities are anonymous and their IDs respect the original data; thus, they cannot be present in any other external dataset. The vocabulary is also domain-dependent, hence the commonalities must be identified in the literal objects. For instance, geographical terms are a good source of out-links.

Parsing and publication. The process of conversion ends with the original file parsing and final publication. The original file is parsed, (a) identifying the entities and properties of the defined data model, (b) converting them to RDF following the defined URI scheme, and (c) establishing the identified out-links. The original data may include some errors and inconsistencies that have to be solved in the parsing process.

Architecture and services for the citizen. The conversion of a census into RDF and their links to the web of data offer the basis for constructing rich applications by the administration and interested third parties. This opens a wide range of applications and a new window for disseminating results, which constitutes one of the main objectives of the census. Thus, this step is almost required in every conversion process, although it goes beyond the basic open data publication.

It is worth mentioning that these steps can be also seen as a particular instantiation of the guidelines provided in Villazón-Terrazas, Vilches, Corcho, and Gómez-Pérez (2011). In this work, the authors group some recommendations for general government Linked Data publishing, in the following five categories: (a) specification, including the analysis of the data sources and the URI design; (b) modeling, referred mainly to the vocabulary construction; (c) generation which includes parsing and linking; (d) publication; and (e) exploitation. As can be seen, we find some similarities and differences. Our first step specifically addresses corpus definition, whereas we share its analysis of data sources, renamed "data modeling" in our second step. This way, we pay special attention to the original data study because of the potential census microdata particularities. We merge the URI design (in their first step) together with its modeling to conform to our third step, called "vocabulary definition." The linking subprocess of their generation step is upgraded to our dedicated four step "Linked Data links" to draw attention to its importance in the case of a census. Finally, we share the main

DOI: 10.1002/asi

features of the latest two steps, although we focus exploitation specifically in "Architecture and Services for the Citizen."

The Case of the Spanish Census

Since its inception in 1768, the Spanish Census has considered the individual as the analysis unit. The latest published census, in 2001, was designed, primarily, to count the population, housing, and buildings. The 2001 Spanish Census was based on a classical census model, that is, an exhaustive query of the territory. The total process involved the efforts of more than 42,500 people. Thirteen million homes, which included a total population of 40,847,371, were visited in 3 months. The total budget amounted to 165,278,328€, of which 70% was addressed to contracting census agents.

The final data were published through (a) microdata files, (b) population figures, and (c) a basic query system. Microdata are ASCII files which sample the total data and represent each field with a fixed-length code. A conversion table describes each field (length/meaning) corresponding to a statistical variable. The main file (5% sampling) includes one register per anonymous individual. Population figures summarize the most important results. The data are also available through a basic query system which allows the user to construct predefined or new tables on a set of variables.

The publication and accessibility of such data present several technical drawbacks. An advanced user could construct only a limited set of queries conditioned by the query system interface. Third-party applications could use the microdata files, but the parsing is neither trivial nor standard. It is difficult to develop automatic tools that consume such data.

Additionally, from an economic and administrative point of view, there are also problems. The government is forced to implement internal processes to provide full access for local administrations. The common citizen obtains few direct visible benefits (e.g., a set of visual applications).

In the following we use the 2001 Spanish Census as a case study to show how these problems can be solved. The general six-step LOD conversion, presented in the previous section, is followed; in this case, the original census microdata are processed to obtain a high-level semantic model to be published as LOD.

Corpus Definition

In this case we focus on the microdata files of the 2001 Spanish Census concerning the individuals, that is, ASCII files with a 5% sampling of the population.

Data Modeling

We identify the statistical data model used by the administration. For this purpose, we enrich the online available documentation with the file parsing and the study of each field that describes an individual.

The resulting data model is shown in Figure 1. Entities are represented by circles, their statistical variables within rectangles, and the relations between entities through arrows. As shown, the given information is about people, homes, houses, family nucleuses, and buildings. The set of properties of the entities are well described in the documentation. These are the fields completed in the data-gathering phase of the census. People belong to a unique family nucleus and a unique home, although that home could house several families. A home is sited physically in a house belonging to a building. People can be related through *father, mother*, and *main person* relations, that is, the head of the family nucleus.

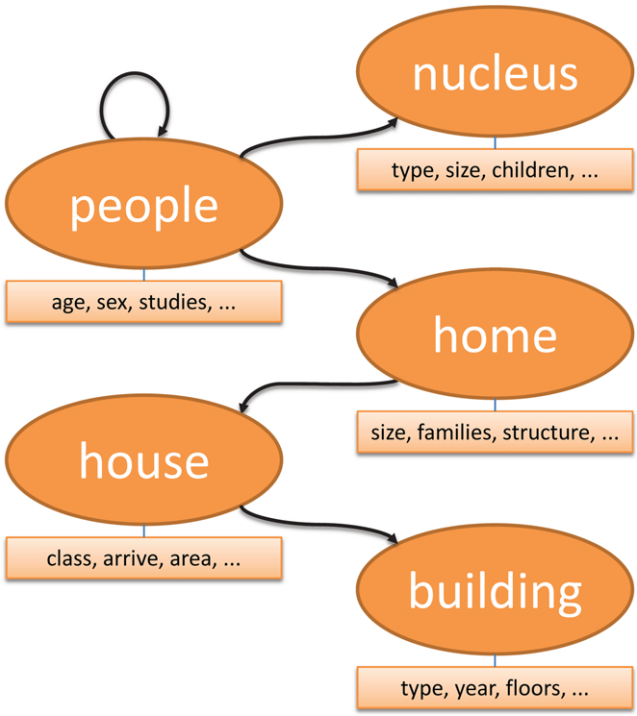

FIG. 1. Spanish Census data model. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Vocabulary Definition

The RDF model requires the use of URIs for entities and properties. We design a URI scheme with a common prefix http://dataweb.infor.uva.es/census2001/censo/2001. Then, entities are named by adding an appropriate ID to this common prefix; the nomenclature for the IDs of people, homes, houses, and nucleuses was taken from the identification of records within the original text file. In turn, the attributes of these entities must be converted to a standard URI representation. As stated, LOD encourages the reuse of vocabularies, but little effort has been made in census properties definition. Thus, in this case ad-hoc URI properties are created by adding the original census attribute name to the aforementioned prefix.[13]

Figure 2 shows a reduced example of census data. Two persons are present; an individual was born in Chile (predicate person#BORN) and links to his mother (predicate

DOI: 10.1002/asi

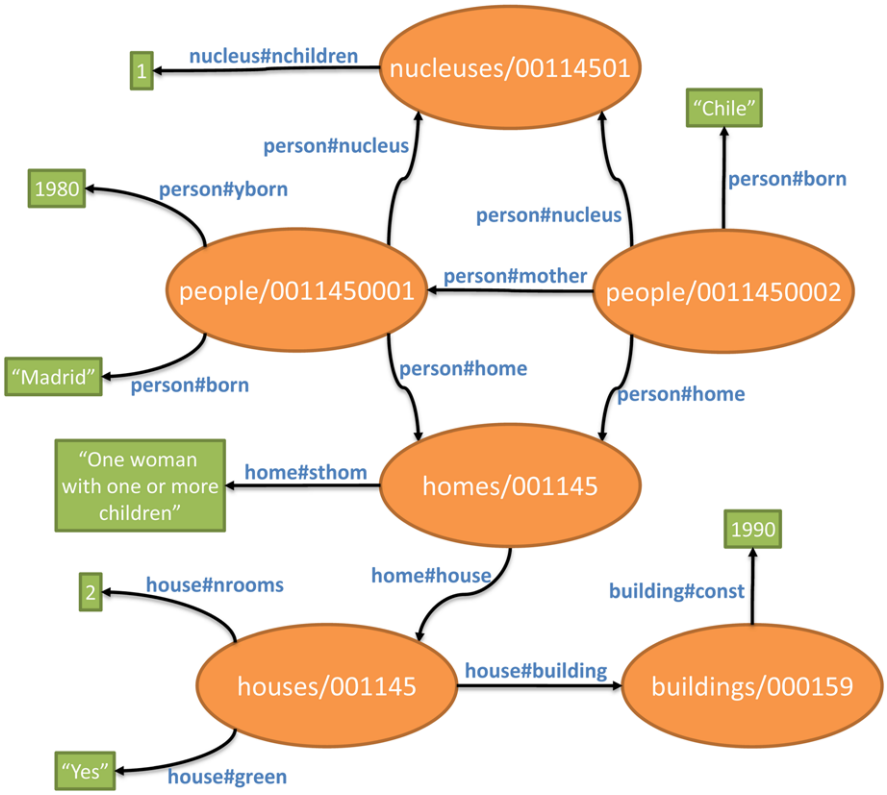

FIG. 2. RDF excerpt from the Spanish Census data (URIs have been summed up). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

person#MOTHER) who was born in Madrid in 1980. They belong to the same family nucleus (with one child, predicate nucleus#NCHILDREN) and the same home, which is structured as "One woman with one or more children" (predicate home#STHOM). The home is sited in a house with two rooms (predicate house#NROOMS) and the occupants of the house complain of not having parkland nearby (predicate house#GREEN). Finally, the house belongs to a building constructed in 1990 (predicate building#CONST).

As can be shown, the URI scheme for entities follows a slash URIs approach where the web server answers requests all these URIs with the URL of a document that represents the resource (Sauermann & Cyganiak, 2008). The URI scheme for properties follows a hash URIs approach in which the fields are separated from the rest of the URI by a hash symbol ("#"') (Sauermann & Cyganiak, 2008). A web server answers requests for all these URIs with the URL of the document describing the vocabulary.

It is worth mentioning that the building entities have been created explicitly in the conversion process and hence their sequential numbering. Although people filled in data about their building, the microdata files do not include identification for each building and thus the trace of common neighbors is impossible. We have created building entities which group a type of building with the same features, that is, two or more individuals agree on the properties of their building. Note that these commonalities are region-independent, thus the statistical secrecy is preserved while new valuable information is published for future studies.

Additional properties have also been added to link entities: (a) person#HOME links people and homes, (b) person#NUCLEUS links people and nucleuses, (c) home#HOUSE links homes and houses, and (d) house#BUILDING links houses and buildings. Person properties relating familial relationship have been redefined to link directly the URI of the related person, thus saving the repetition of data.

This conversion shows that the final data model is simple, intuitive, and expressive. It preserves the original conception of the census and its properties but adopts RDF and Linked Data principles as common standards in the web.

Linked Data Links

The accepted values for literal objects are well described in the original documentation and, as shown in the example of Figure 2, they are usually strings that are hardly repeated in other contexts except for the countries, regions, and town names. These are the candidates for external linkage.

Geographic concepts are referred through GeoNames.[14] We create new URIs for countries, regions, and town names, which act as the properties of the entities. These concepts link to GeoNames entities (owl:sameAs predicate) whose URI is constructed with a unique GeoNames Identifier: the

DOI: 10.1002/asi

geoNameId. The naming convention for geographic URIs follows the pattern small geo/<code>, in which <code> consists of the original code for the country or the region, or the string concatenation of region and town code for the towns.

Parsing and Publication

The accepted values for the original fields were encoded with fixed length integers. In parsing, we substitute each integer by the corresponding value. In turn, the GeoNames linkage is performed over the geographical concepts. The final RDF file is built in N3 format. The total process reports these statistics: 2,039,274 persons, 118,255,550 triples, and 147 predicates. The size of the resultant dataset exceeds 15 GB.

Architecture and Services for the Citizen

The new publication scheme allows us to construct rich applications. As an example, a website[15] was built on top of the Spanish Census converted information. RDF information is included in an RDF store providing a configurable SPARQL endpoint to run queries to the census data. Thus, applications can be built on top of the RDF store and its SPARQL interface or over other existing applications and demos. This allows users to have online access to the data, the endpoint, and the applications. We chose Virtuoso 6.1.2[16] to implement the RDF store, which is built on a relational database.

An example of a real SPARQL user query running on the system is presented below.

```
PREFIX ppl:<http://dataweb.infor.uva.
es/census2001/censo/2001/persona#>
SELECT distinct(?age), count(?a) AS
?es, count(?b) AS ?ex
FROM     <http://dataweb.infor.uva.es/
census2001/censo/2001>
WHERE{
  {
  ?a ppl:NACI ?born .
  ?born rdfs:label ?country FILTER
  (?country="ESPANA").
  ?a ppl:EDAD ?age .
  }
  UNION
  {
  ?b ppl:NACI ?born2 .
  ?born2 rdfs:label ?country2 FILTER
  (?country2!="ESPANA").
  ?b ppl:EDAD ?age .
  }
}
ORDER BY ASC (?age)
```

The query returns the number of Spanish and foreign people by age in the census data. It makes use of normative SPARQL syntax and some Virtuoso features from the relational databases such as the variable renaming AS and the count function.

RDF Publication of International Census Microdata

As shown, the previous case successfully addresses the open publication challenges of census microdata and their direct use by third parties. There are other types of uses that could benefit from a more general approach, particularly the integration of several censuses across countries and time periods that have different statistical variables. In this case, a top-down approach would be more appropriate, that is, to get a general model that can be instantiated to a particular case. In this section, based on international census standards, we develop a general high-level vocabulary for census microdata built using Linked Data principles for publishing data on the web. In this case, specific data can be instantiated selecting the desired countries, time-periods, and statistical variables.

The Integrated Public Use Microdata Series (IPUMS-I) (Hall et al., 2000), is one of the biggest international projects on census microdata, whose goals are (a) to provide readable census microdata; (b) to preserve census microdatasets identified as at risk; (c) to create an integrated international census database with a harmonized system of concepts, variables, and codes, incorporating both historical and contemporary microdata of individuals, households, and dwellings; and (d) to disseminate integrated microdata via the Internet. IPUMS-I currently has available 185 samples encompassing 62 countries with a total of 397 million person records harmonized (McCaa, Ruggles, Sobek, & Thomas, 2011). We will concentrate on goals (c) and (d) using the LOD Conversion defined in the Linked Open Data Conversion of Census Microdata section.

Corpus Definition

IPUMS-I has recollected and integrated a vast set of census data from its international partners, the local statistics offices, to shape its census data. These data include some metadata about each sample, for example, census and sample characteristics, local office identification, and variable categorization, such as its type (household, demographic, education, etc.) and its domain. From this vast census data integrated by IPUMS-I, we only focus on the census and variable information that can be linked to external RDF datasets.

Data Modeling

The next step is to identify the relations between elements to build a common microdata model. Figure 3 shows a very general representation of the main parts of our model. This model tries to represent, in RDF format, the IPUMS-I relations but mainly focuses on creating external links.

DOI: 10.1002/asi

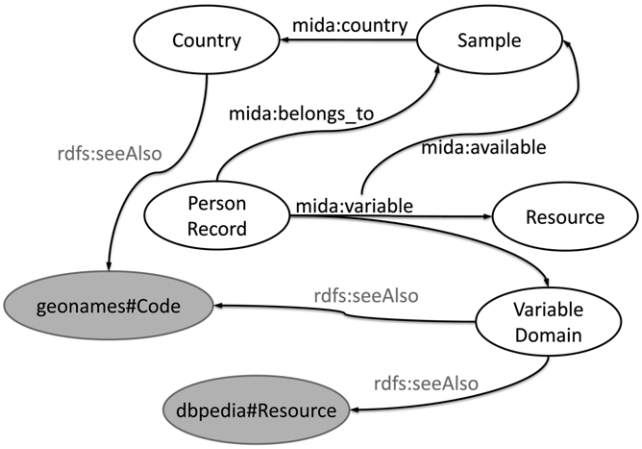

FIG. 3. MIDA vocabulary overview.

Unlike the Spanish Census case (including several entities), the only recognized entities in IPUMS-I are Person Records, Sample, and Countries. Each record groups the information of a given anonymous person, providing its fields as a set of variables. These variables and their description are given then in the vocabulary definition.

Vocabulary Definition

In this step, we focus on creating a general RDF vocabulary for the attributes in the IPUMS-I corpus (referred to as "variables"). Note that the effort of alignment (harmonization) of the attributes between censuses and periods has already been done by IPUMS-I, hence we take advantage of it to provide a general semantic vocabulary. The resultant set of URI properties is the microdata vocabulary[17] (MIDA). It is a direct RDF conversion of the variable metadata of the 185 samples encompassing 62 countries integrated to date, published on the IPUMS-I web page (Hall et al., 2000). The MIDA vocabulary has a total of 162.068 RDF triples.

Figure 3 shows the main elements of the MIDA vocabulary. A Person Record belongs to a Sample and has a set of variables. Two types of variables are present: (a) discrete variables that only take a set of possible values, represented as a Variable Domain in the figure, and (b) continuous variables whose values are not predefined. We represent this latter by pointing to a Resource in Figure 3.

It is worth mentioning that, because of the integration of different censuses, not all variables are available on all samples. This is represented by the mida:available predicate in Figure 3. Note also that Sample has several more properties not included in this figure, such as census title, census agency, population universe, enumeration forms, field of work period, or sample design.

The MIDA vocabulary respects the IPUMS-I variable categorization; a variable could be integrated (available in more than one sample) or unharmonized (available in just one sample) and could be a household variable or a person variable. Household variables are, among others, geography, economic, utility, or amenity variables. Person variables could be demographic, fertility-and-mortality, ethnicity-and-language, labor, income, migration, or disability variables.

Linked Data Links

We include external links to Geonames, which are represented as geonames#Code entities, for the country of each sample and those Variable Domains which indeed have a geographic location. We also include links to DBpedia,[18] represented as dbpedia#Resource for Variable Domains containing years. The aim is to point to the DBpedia resource describing important events occurring in a given year. These Linked Data out-links are defined in the vocabulary consumed by the tool.

As an example of the usefulness of the MIDA vocabulary we present a simple SPARQL query that asks for the common census years from Chile and Brazil and obtains the abstract DBpedia information from returned years. This dbpedia-owl:abstract information is a summary of international events that took place in such years.

```
PREFIX mida: <http://purl.org/mida/
0.1/mida.rdf#>
PREFIX rdf: <http://www.w3.org/1999/
02/22-rdf-syntax-ns#>
SELECT ?year ?abstract
WHERE {
  ?scl mida:country mida:Chile .
  ?sbz mida:country mida:Brazil .
  ?scl mida:sampleYear ?year .
  ?sbz mida:sampleYear ?year .
  ?year mida:dbpediaResource ?reso .
  ?reso dbpedia-
  owl:abstract ?abstract .
}
```

Parsing and Publication

For the generation of the RDF triples, we implemented a simple Java tool (midafc.jar[19]). The first step is the execution of a SPARQL query on the MIDA vocabulary to identify the samples and variables to be downloaded. For instance, we can select all common integrated person variables between two censuses of different countries and different years. The next step is to download the identified samples and variables from the IPUMS-I page, as usual. Two files have to be downloaded: (a) the codebook, in the Data Documentation Initiative (DDI) XML format, provides the description of the samples and variables included in (b) the IPUMS-I "dat" file, a plain file with variables of fixed length. Finally, the Java tool parses the files and returns the RDF data, including the predefined Linked Data out-links.

DOI: 10.1002/asi

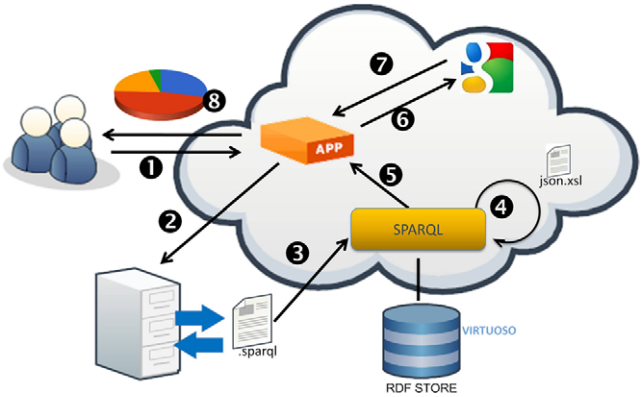

FIG. 4. Visualization flow in the Spanish Census site. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Architecture and Services for the Citizen

Generated RDF triples can be published using the same architecture and services for the citizen presented at the end of The Case of the Spanish Census section.

Conclusions and Future Work

This article shows that the latest semantic web technologies are suitable for formalizing the representation, publication, and exploitation of census microdata. The advantages of such an approach include the flexibility of the data that allows any user to adapt it to diverse applications for diverse purposes; the discovery of new relationships among the data; the accessibility of pieces and views of the data; the possibility of integrating them with other data; and last but not least, the possibility of providing simple services for complex querying, demos, and visualizations.

We have discussed the publication of census microdata under the principles of LOD, providing a step-by-step process. We instantiate this process in two case studies; an ad-hoc conversion of the 2001 Spanish Census and a generalization using the IPUMS-I framework. The most important conclusion here is that—independently of the approach chosen—the process of publishing LOD for census microdata is feasible, simple, and shows immediate advantages. In the case of a particular census, a specific schema is developed for each given census dataset. It gives flexibility to the modeling process and allows for including particular local features of each census data. The harmonized approach sacrifices flexibility (by forcing using a general model, like the IPUMS-I) at the cost of gaining interoperability: several countries and time periods census can be integrated and converted together.

We also developed visualization demos to show that the use of the semantic information is easy and extensible. The visualization flow is represented in Figure 4. It follows a similar approach to the data-gov demos in Ding et al. (2010). Once the user accesses the application, the request is redirected to the server hosting the appropriate SPARQL query, that is, each demo has a related query for extracting the matching information. The SPARQL endpoint runs the query and the result is converted to the JavaScript Object Notation[20] (JSON). These data are passed through the Google Visualization API,[21] which finally produces the visualization shown to the user, such as pie and bar charts or maps.

We think that these initial deployments will encourage administrations to provide the means for their members to deal with the new Open Government requirements and to incorporate these technologies into their publication processes. In addition, we would like to encourage researchers in other fields (law, history, government, etc.) to examine the potentialities and risks arising in initiatives such as LOGD. Although the main aim of LOD is developing methods for data structuring and publication to facilitate the discovery of related information, it opens the door for malicious individuals to harness it for fraudulent data-matching across collections in the LOD cloud. Thus, even though personal data-matching across databases is a privacy issue ruled by specific laws (e.g., United States,[22] United Kingdom,[23] or Australia[24]), more research is needed to address the application of these techniques to sensitive data. It is worth noting that the first step of our conversion process allows the released data to be carefully selected, enabling administrations to obfuscate or directly discard sensitive information.

We envision different approaches for future work. We are currently developing applications to take further advantage of the semantic data. In particular, a graphical query interface would fill the gap between census practitioners and formal languages for querying semantic data such as SPARQL. We are also refining the conversion process, by semiautomatizing the external link generation, for example, to automatically recommend links to DBpedia resources or any dataset in the LOD cloud. We also expect that these methodologies will become widespread and, hopefully, more census microdata will be openly available in standards formats. Then much work could be done in data linking, reasoning over the semantic data, and automatic consumption (summarization of important information, snippets in web engines, etc.).

Acknowledgments

This work was partially funded by MICINN (TIN2009-14009-C02-02), Fondecyt 1110287, and Fondecyt 1-110066. The third author was funded by Erasmus Mundus, the Regional Government of Castilla y León (Spain), and the European Social Fund. The fourth author was funded by the University of Valladolid: programme of Mobility Grants for Researchers (2012).

Endnotes

1. http://www.nappdata.org/napp
2. http://www.opengeospatial.org/standards
3. http://www.unodc.org/unodc/en/organized-crime/law-enforcement.html

DOI: 10.1002/asi

4. http://ec.europa.eu/isa/
5. http://www.guardian.co.uk/world-government-data
6. http://ehumanities.nl/ceda_r/
7. http://linkeddata.org/
8. http://logd.tw.rpi.edu/
9. The Open Data Foundation (http://www.opendatafoundation.org/), the Open Knowledge Foundation (http://okfn.org/)
10. http://www.opengovpartnership.org
11. http://www.rdfabout.com/demo/census
12. http://www4.wiwiss.fu-berlin.de/lodcloud/
13. The Spanish Census field descriptions are available at ftp://www.ine.es/temas/censopv/cen01_ph/disenoreg_ph01.zip
14. http://www.geonames.org
15. http://dataweb.infor.uva.es/census2001
16. http://virtuoso.openlinksw.com
17. http://purl.org/mida/0.1/mida.rdf
18. http://dbpedia.org
19. http://purl.org/mida/0.1/
20. http://www.json.org
21. https://developers.google.com/chart/
22. http://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/final_guidance_pl100-503.pdf
23. http://www.audit-commission.gov.uk/SiteCollectionDocuments/Downloads/CodeDMPFinalJuly08.pdf
24. http://www.privacy.gov.au/materials/types/download/8688/6527

References

Alonso, J.M., Novak, K., & Acar, S. (2009). Improving access to government through better use of the web. W3C Interest Group Note 12 May 2009. Retrieved from http://www.w3.org/TR/egov-improving/

Anderson, M.J. (1988). The American census: A social history. New Haven: Yale University Press.

Berners-Lee, T. (2006). Linked data. W3C Design Issues. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T. (2009). Putting government data online. W3C Design Issues. Retrieved from http://www.w3.org/DesignIssues/GovData.html

Bizer, C. (2009). The emerging web of linked data. IEEE Intelligent Systems, 24, 87–92.

Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. (2008). Linked data on the web. In Proceedings of the 17th International Conference on World Wide Web (pp. 1265–1266).

Ding, L., DiFranzo, D., Graves, A., Michaelis, J., Li, X., McGuinness, D., & Hendler, J. (2010). TWC data-gov corpus: Incrementally generating linked government data from data.gov. In Proceedings of the 19th International Conference on World Wide Web (WWW) (pp. 1383–1386).

Esteve, A., & Sobek, M. (2003). Challenges and methods of international census harmonization. Historical Methods, 36(2), 37–41.

Fernández, J.D., Martínez-Prieto, M.A., & Gutiérrez, C. (2011). Publishing open statistical data: The Spanish census. In Proceedings of the 12th Annual International Conference on Digital Government Research (dg.O) (pp. 20–25).

Hall, P., McCaa, R., & Thorvaldsen, G. (2000). Handbook of international historical microdata for population research. Minneapolis, MN: The Minnesota Population Center.

Hausenblas, M., & Karnstedt, M. (2010). Understanding linked open data as a web-scale database. In Proceedings of the 2nd International Conference on Advances in Databases (DBKDA) (pp. 56–61).

Hendler, J., Holm, J., Musialek, C., & Thomas, G. (2012). US government linked open data: semantic.data.gov. IEEE Intelligent Systems, 27(3), 25–31.

Lathrop, D., & Ruma, L. (Eds.) (2010). Open government: Collaboration, transparency, and participation in practice. Sebastopol, CA: O'Reilly.

Maali, F., Cyganiak, R., & Peristeras, V. (2010). Enabling interoperability of government data catalogues. In Proceedings of the 9th IFIP WG 8.5 International Conference on Electronic Government (EGOV) (pp. 339–350).

Malona, F., & Miller, E. (2004). RDF primer. W3C Recommendation 10 February 2004. Retrieved from http://www.w3.org/TR/rdf-primer/

McCaa, R., Ruggles, S., Sobek, M.L., & Thomas, W. (2011). IPUMS-International: free, worldwide microdata access now for censuses of 62 countries–80 by 2015. In 58th ISI World Statistics Congress.

Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., . . . (2012). Semantic technologies for historical research: A survey. Semantic Web Journal [under review]. Retrieved from http://www.semantic-web-journal.net/sites/default/files/swj301.pdf

Minnesota Population Center. (2011). Integrated public use microdata series, international: Version 6.1 [machine-readable database]. Minneapolis, MN: University of Minnesota.

Obama, B. (2009). Transparency and open government. Memorandum for the Heads of Executive Departments and Agencies. Retrieved from http://www.whitehouse.gov/the_press_office/TransparencyandOpen Government

Perego, A., Fugazza, C., Vaccari, L., Lutz, M., Smits, P., Kanellopoulos, I., & Schade, S. (2012). Harmonization and interoperability of EU environmental information and services. Intelligent Systems IEEE, 27(3), pp. 33–39.

Prud'hommeaux, E., & Seaborne, A. (2008). SPARQL query language for RDF. W3C Recommendation 15 January 2008. Retrieved from http://www.w3.org/TR/rdf-sparql-query/

Sauermann, L., & Cyganiak, R. (2008). Cool URIs for the semantic web. W3C Interest Group Note. Retrieved from http://www.w3.org/TR/cooluris/

Shadbot, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., & Schraefel, M.C. (2012). Linked open government data: Lessons from ″Data.gov.uk. IEEE Intelligent Systems, 27(3), 16–24.

Somoza, J.L., & Lattes, A.E. (1967). Muestras de los dos primeros censos nacionales de población, 1869 y 1895. Buenos Aires, Argentina: Instituto Torcuato Di Tella, Centro de Investigaciones Sociales, Documento de Trabajo no 46.

Trade and Investment Division. (2012). Data harmonization and modelling guide for single window environment. Economic and Social Commission for Asia and the Pacific (ESCAP), ST/ESCAP/2619. Retrieved from http://www.unescap.org/tid/publication/tipub2619.pdf

Uhlir, P., & Schröder, P. (2007). Open data for global science. Data Science Journal, 6, 36–53.

United Nations Statistics Division. (2010). Report on the results of a survey on census methods used by countries in the 2010 census round. Working Paper: UNSD/DSSB/1. Retrieved from http://unstats.un.org/unsd/demographic/sources/census/2010_phc/docs/ReportOnSurvey For2010Census.pdf

Villazón-Terrazas, B., Vilches, L.M., Corcho, O., & Gómez-Pérez, A. (2011). Methodological guidelines for publishing government linked data. In D. Wood (Ed.), Linking government data (pp. 27–49). New York: Springer.

DOI: 10.1002/asi