Linked Open Data as the fuel for Smarter Cities

Mikel Emaldi, Oscar Peña, Jon Lázaro, Diego López-de-Ipiña

Deusto Institute of Technology - DeustoTech, University of Deusto
Avda. Universidades 24, 48007, Bilbao, Spain
{m.emaldi, oscar.pena, jlazaro, dipina}@deusto.es

Abstract. In the last decade big efforts have been carried out in order to move towards the Smart City concept, from both the academic and industrial points of view, encouraging researchers and data stakeholders to find new solutions on how to cope with the huge amount of generated data. Meanwhile, Open Data has arisen as a way to freely share contents to be consumed without restrictions from copyright, patents or other mechanisms of control. Nowadays Open Data is an achievable concept thanks to the World Wide Web, and has been re-defined for its application in different domains. Regarding public administrations, the concept of Open Government has found an ally in Open Data concepts, defending citizens' right to access data, documentation and proceedings of the governments.

We propose the use of Linked Open Data , a set of best practices to publish data on the Web recommended by the W3C, in a new data life cycle management model, allowing governments and individuals to handle better their data, easing its consumption by anybody, including both companies and third parties interested in the exploitation of the data, and citizens as end users receiving relevant curated information and reports about their city. In summary, Linked Open Data uses the previous Openness concepts to evolve from an infrastructure thought for humans, to an architecture for the automatic consumption of big amounts of data, providing relevant and high quality data to end users with low maintenance costs. Consequently, smart data can now be achievable in smart cities.

1 Introduction

In the last decade, cities have been sensorised, protocols are constantly refined to deal with the possibilities that new hardware offers, communication networks are offered in all flavours, and so on, generating lots of data that needs to be dealt with. Public administrations are rarely able to process all the data they generate in an efficient way, resulting in large amounts of data going un-analysed and limiting the benefits end-users could get from them.

Citizens are also being encouraged to adopt the role of linked open data providers. User-friendly Linked Open Data applications should allow citizens to easily contribute with new trustable data which can be linked to already existing published (more static generally) Linked Open Data provided by city councils. In addition, people-centric mobile sensing, empowered by the technology inside actual smartphones , should progress into continuous people-centric enriched Linked Data . Linked Open Data also encourages the linkage to other resources described formally through structured vocabularies, allowing the discovery of related information and the possibility to make inferences, resulting in higher quality data.

Data management is becoming one of the greatest challenges of the 21^{st} century. Regarding urban growth, experts predict that global urban population will double by the year 2050, meaning that nearly 70% of the whole planet's inhabitants will be living in a major town or city [1]. This prediction makes clear the need to deal with the huge amounts of data generated by cities, enabling the possibility to manage their resources in an efficient way. The *Smart Data* term has been coined to address data that makes itself understandable, extracting relevant information and insights from large data sources and presenting the conclusions as human-friendly visualisations.

The problems of managing data are moving to a new level. It is not only a matter of caring about *more* data, but how we can use it efficiently in our processes. It is about how we can deal with increasing volumes of data (from standalone databases to real *Big Data*) and integrate them to our advantage, making data useful and digestible in order to make better decisions.

In the last few years, the *Smart city* concept has been adopted by cities aware of their citizens' life quality, worried about the efficiency and trustworthiness of the services provided by governing entities and businesses. Smart data (understood as curated, high-quality and digestible data) can help cities promote to a *Smart City* status, analysing the generated data streams and providing useful information to their users: citizens, council managers, third parties, etc.

Although, efficient data lifecycle management processes need to be adopted as best practices, avoiding the provision of input data that can not be used to improve council's services, thus incrementing the noise around high quality data.

Our approach is based on an actual review of the state of the art regarding data lifecycle management, proposing our own model as a more refined approach to the existing ones. We also encourage the adoption of

Linked Open Data principles[1] to publish both the whole generated data and the processed data, in order to allow further research on the area by third parties and the development of new business models relying on public access data. Similar proposals regarding Linked Data are defended by [2].

[1] http://5stardata.info/

2 Background and definitions

As envisaged by Sir Tim Berners-Lee, the Web is moving from an interlinked documents space to a global information one where both documents and data are linked: The Semantic Web [3]. Related to this Semantic Web , Linked Data is a set of best practices to publish data on the Web in a machine-readable way, with a explicitly defined semantic meaning, linked to other datasets and allowed to be searched for [4]. In 2006, Sir Tim Berners-Lee described a set of principles to publish Linked Data on the Web:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information, using the standard (RDF, SPARQL)
4. Include links to other URIs, so that they can discover more things

Later, in 2010, he established a five-star rating system to encourage people and governments to publish high-quality Linked Open Data:

⋆	Available on the web (whatever format) but with an **open licence**, to be Open Data .
⋆ ⋆	Available as **machine-readable** structured data (e.g. excel instead of image scan of a table).
⋆ ⋆ ⋆	as (2) plus **non-proprietary format** (e.g. CSV instead of excel).
⋆ ⋆ ⋆ ⋆	All the above plus: Use **open standards** from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff.
⋆ ⋆ ⋆ ⋆ ⋆	All the above plus: **Link your data** to other people's data to provide context.

Over the years, enormous amount of applications based on Linked Data have been developed. Big companies like Google[2], Yahoo![3] or Facebook[4] have lately invested resources on deploying Semantic Web and Linked Data technologies. Several governments from countries like the United States of America or the United Kingdom have published a big amount of Open Data from different administrations in their Open Data portals. Linked Data is real and can be the key for data management in smart cities. Following these recommendations, data publishers can move towards a new data-powered space, in which data scientists and application developers can research on new uses for Linked Open Data.

[2] `http://www.google.com/insidesearch/features/search/knowledge.html`
[3] `http://semsearch.yahoo.com/`
[4] `http://ogp.me/`

3 Data Life Cycle

Through the literature, a broad variety of different definitions of data life cycle models can be found. Although they have been developed for different actuation domains, we describe here some of them which could be applied for generic data, independently of its original domain.

3.1 Data Documentation Initiative

The first model to be analysed is the model proposed by the Data Documentation Initiative (DDI). The DDI introduced a Combined Life Cycle Model for data managing [5]. As Figure 1 shows, this model has eight elements or steps which can be summarized as follows, according to [6]:

- **Study concept.** At this stage, apart from choosing the research question and the methodology to collect data, the processing and analysis stage of the needed data to answer the question is planned.
- **Data collection.** This model proposes different methods to collect data, like surveys, health records, statistics or Web-based collections.
- **Data processing.** At this stage, the collected data are processed to answer the proposed research question. The data may be recorded in both machine-readable and human-readable form.
- **Data archiving.** Both data and metadata should be archived to ensure long-term access to them, guaranteeing confidentiality.
- **Data distribution.** This stage involves the different ways in which data are distributed, as well as questions related to the terms of use of the used data or citation of the original sources.
- **Data discovery.** Data may be published in different manners, through publications, web-indexes, etc.
- **Data analysis.** Data can be used by others to achieve different goals.
- **Repurposing.** Data can be used outside of their original framework, restructuring or combining it to satisfy diverse purposes.

3.2 Australian National Data Service

In late 2007, the *Australian National Data Service* (ANDS) was founded with the objective of creating a national data management environment. ANDS established a set of verbs, denominated *Data Sharing Verbs*, that describe the entire life cycle of the data [7]:

- **Create.** *Create* (or *collect* for disciplines with an observational focus) is about the kinds of metadata that could be collected and the tools to fulfil this gathering task.
- **Store.** This *Data Sharing Verb* remarks the need for stable and web-accessible storage, taking care of the appropriate storing of data.
- **Describe.** The more information inside the storage, the more difficult its discovery, access and exploitation is. Annotating the data with the proper metadata solves this issue.

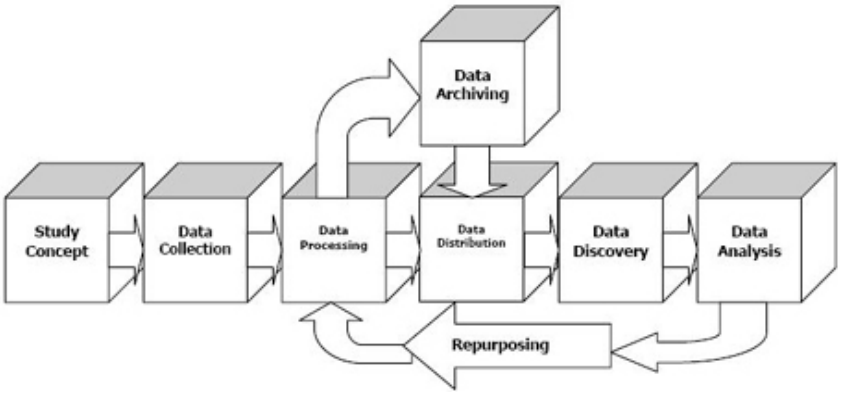

Fig. 1. Combined Life Cycle Model (ownership: DDI Alliance).

- **Identify.** The application of this verb implies the proper identification of each data resource, assigning a persistent identifier to each of them.
- **Register.** This step pertains to record the descriptions of the different data collections with one or more public catalogues.
- **Discover.** To improve data-reusing, ANDS suggests to enable different discovery services.
- **Access.** To guarantee the appropriate access to data, ANDS advises to provide a suitable search engine to retrieve these data. If data is not electronically available, ANDS recommends to provide contact details to get data in conventional formats.
- **Exploit.** *Exploit*, the final *Data Sharing Verb*, comprises the tools, methodologies and support actions to enable reutilisation of data.

3.3 Ecoinformatics data life cycle

Michener and Jones define in [8] the concept of "ecoinformatics": *a framework that enables scientists to generate new knowledge through innovative tools and approaches for discovering, managing, integrating, analysing, visualising and preserving relevant biological, environmental, and socioeconomic data and information.* To manage these data, the following data life cycle has been defined, as can be seen at Figure 2:

- **Plan.** This step involves the confection of a data management planning.
- **Collect.** This step considers both manual (hand-written data sheets) and automatic (sensor networks) data-gathering methods.
- **Assure.** Quality assurance and quality control (QA/QC), an issue addressed in previously mentioned models, is not taken into account. Michener and Jones proposal is based on developing methods to guarantee the integrity of data. Quality assurance can also include the definition of standards for formats, codes, measurement units, metadata, etc.

– **Describe.** As other data life cycle models, this model remarks the value of the metadata to answer questions about *who, when, where, how* and *why*.
– **Preserve.** Data preservation implies the storage of the data and metadata, ensuring that these data can be verified, replicated and actively curated over time.
– **Discover.** The authors describe the data discovering process as *one of the greatest challenges*, as many data are not immediately available because they are stored in individual laptops. The main challenges to publish the data in a proper way are related to the creation of catalogues and indexes, and about the implementation of the proper search engines.
– **Integrate.** Integrating data from different and heterogeneous sources can become a difficult task, as it requires *understanding methodological differences, transforming data into a common representation, and manually converting and recording data to compatible semantics before analysis can begin.*
– **Analyze.** As well as the importance of a clear analysis step, this models remarks the importance of documenting this analysis with sufficient detail to enable its reproduction in different research frameworks.

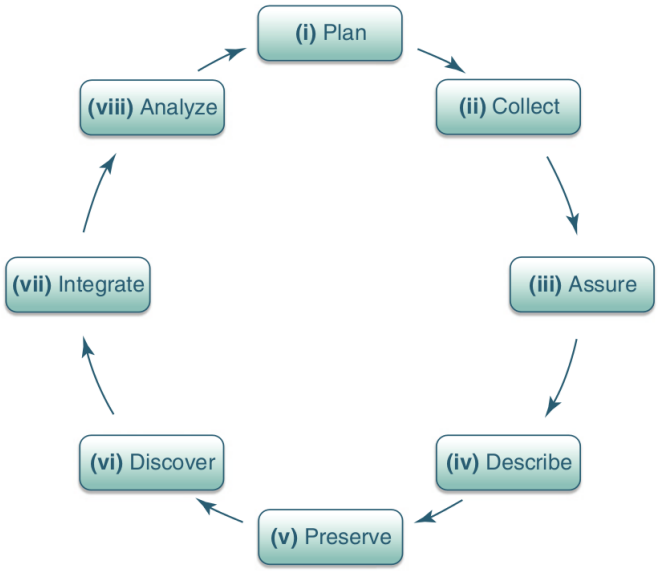

Fig. 2. Data life cycle in ecoinformatics. Taken from [8].

3.4 UK Data Archive

Another data life cycle model is the one proposed by *UK Data Archive*[5]. This model is oriented to help researchers publish their data in a manner that allows other researchers to continue their work independently. In Figure 3, the following stages can be observed:

- **Creating data.** Creating the data involves the design of the research question, planning how data are going to be managed and their sharing strategy. If we want to reuse existing data, we have to locate existing data and collect them. Whether data is new or existing, at this stage the metadata has to be created.
- **Processing data.** Like in other models, at this stage the data is translated, checked, validated and cleaned. In the case of confidential data, it needs to be "anonymized". The UK Data Archive recommends the creation of metadata at this stage too.
- **Analysing data.** At this stage data are interpreted and derived into visualisations or reports. In addition, the data are prepared for preservation, as mentioned in the following stage.
- **Preserving data.** To preserve data properly, they are migrated to the best format and stored in a suitable medium. In addition to the previously created metadata, the creating, processing, analysis and preserving processes are documented.
- **Giving access to data.** Once the data is stored, we have to distribute our data. Data distribution may involve controlling the access to them and establish a sharing license.
- **Re-using data.** At last, the data can be re-used enabling new research topics.

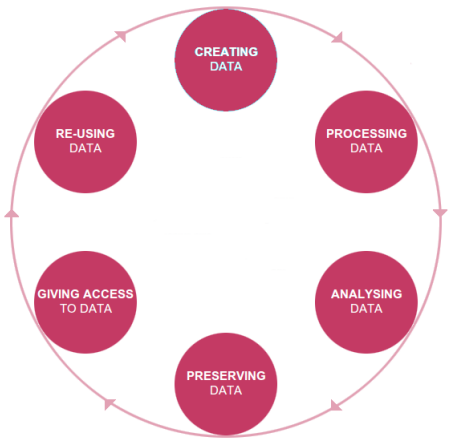

Fig. 3. Data life cycle proposed by UK Data Archive.

[5] http://www.data-archive.ac.uk/create-manage/life-cycle

3.5 The LOD2 Stack Data Life Cycle

The last analysed data life cycle has been developed under the LOD2[6] project. This project proposes a technological and methodological stack which supports the entire life cycle of Linked Data [9]. As Figure 4 shows, the proposed life cycle phases are the following:

Fig. 4. Linked Data life cycle from LOD2 stack.

- **Storage.** As RDF data presents more challenges than relational data, they propose the collaboration between known and new technologies, like columnstore technology, dynamic query optimization, adaptive caching of joins, optimized graph processing and cluster/-cloud scalability.
- **Authoring.** LOD2 provides provenance about data collected through distributed social, semantic collaboration and networking techniques.
- **Interlinking.** At this phase, LOD2 offers approaches to manage the links between different data sources.
- **Classification.** This stage deals with the transformation of raw data into Linked Data. This transformation implies the linkage and integration of data with upper level ontologies.
- **Quality.** Like other models, LOD2 develops techniques for assessing quality based on different metrics.

[6] `http://lod2.eu/`

– **Evolution/Repair.** At this stage, LOD2 deals with the dynamism of the data from the Web, managing changes and modifications over the data.
– **Search/Browsing/Exploration.** This stage is focused on offering Linked Data to final users through different search, browsing, exploration and visualisation techniques.

3.6 A common data life cycle for Smart Cities

Based on these data life cycle models, we propose a common data life cycle for managing data in Smart Cities . As can be seen in Figure 5, the different stages of the mentioned models have been aggregated, forming our proposed model.

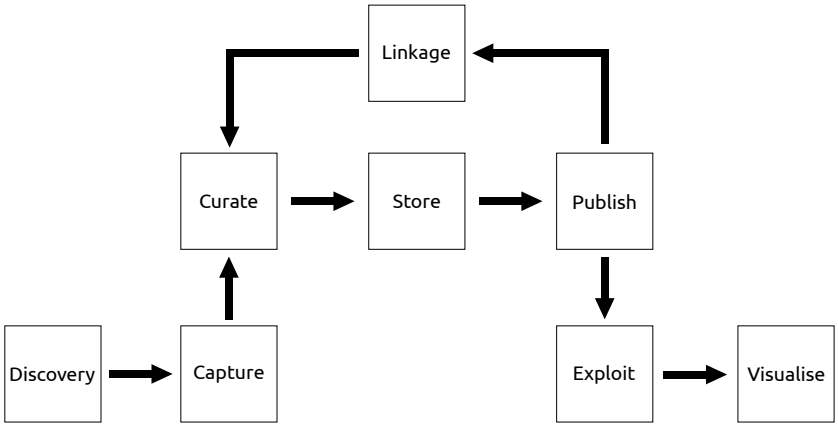

Fig. 5. Proposed model.

The different stages of this model, which are going to be explained widely in following sections, are:
– **Discovery:** The first step in our model consists of discovering where data can be taken from, identifying the available datasets which contain the necessary data to accomplish our task. Datasources can either be maintained by us, or by external entities, so the more metadata we can gather from datasets, the easier further steps will become.
– **Capture:** Once datasources are identified, data need to be collected. In a Smart City environment, there are a lot of alternatives to capture data, like sensors, data published by public administration, social networks , mobile sensing, user generated data or in more traditional ways like surveys.
– **Curate:** After the required data are captured, they are prepared to be stored and in need of proper methods to explore them. This processing involves the analysing, refining, cleaning, formatting and transformation of the data.

- **Store:** The storage of data is, probably, the most delicate action in the life cycle. Above the storage all the analysis tools are build, and is the "final endpoint" when someone requests our data. A suitable storage should have indexing, replication, distribution and backup features, among other services.
- **Publish:** Most of the previously mentioned models prioritize the analysing stage over the publication stage. In our model, we defend the opposite approach for a very simple reason: when you exploit your data before their publication, and using different processes as the rest of the people who is going to use them, you are not making enough emphasis on publishing these data correctly. Everybody has ever met a research paper or an application in which accessing the data was difficult, or, when once the data was collected it became totally incomprehensible. To avoid this issue, we propose to publish the data before their consumption, following the same process the rest of the people do.
- **Linkage:** Before consuming data, we suggest to search for links and relationships with other datasets found in the discovery step. Actual solutions do not allow the linkage with unknown datasets, but tools are developed to ease link discovery processes between two or more given datasources.
- **Exploit:** Once the data is published, we use the provided methods to make use of the data. This data consumption involves data mining, analytics or reasoning.
- **Visualise:** To understand data properly, designing suitable visualisations is essential to show correlations between data and the conclusions of the data analysis in a human-understandable way.

4 Identified challenges

Taking into consideration the large amounts of data present at Smart Cities , data management's complexity can be described in terms of:

- Volume
- Variety
- Veracity
- Velocity

This four variables can also be found in *Big Data*-related articles (also known as the *Big Data's Vs*) [10, 11], so it is not surprising at all that Smart Cities are going to deal with Big Data problems in the near future (if they are not dealing with them right now).

Data scientist need to take into account these variables, which could overlap in certain environments. Should this happen, each scenario will determine the most relevant factors of the process, generating unwelcome drawbacks on the other ones.

4.1 Volume

The high amount of data generated and used by cities nowadays needs to be properly analysed, processed, stored and eventually accessible. This means conventional IT structures need to evolve, enabling scalable storage technologies, distributed querying approaches and massively parallel processing algorithms and architectures.

There is a growing trend which defends that the minimum amount of data should be stored and analysed without significantly affecting the overall knowledge that could be extracted from the whole dataset. Based on *Pareto's Principle* (also known as the 80-20 rule), the idea is to focus on a 20% of the data to be able to extract up to the 80% of knowledge within it. Even being a solid research challenge in Big Data, there are occasions where we can not discard data from being stored or analysed (e.g., sensor data about monitoring a building can be used temporally, while patient monitoring data should be kept for historical records).

However, big amounts of data should not be seen as a drawback attached to Smart Cities. The larger the datasets, the better analysis algorithms can perform, so deeper insights and conclusions should be expected as an outcome. These could ease the decision making stage.

As management consultant Peter Drucker once said: *"If you can not measure it, you can not manage it"*, thus leaving no way to improve it either. This adage manifests that should you want to take care of some process, but you are not able to measure it or you can not access the data, you will not be able to manage that process. That being said, the higher amounts of data available, the greater the opportunities of obtaining useful knowledge will become.

4.2 Variety

Data is rarely found in a perfectly ordered and ready for processing format. Data scientists are used to work with diverse sources, which seldom

fall into neat relational structures: embedded sensor data, documents, media content, social data, etc. As can be seen at section 5.2 there are many different sources where data can come from in a smart city. Despite of presented data cycle can be applied for all kind of data, the different steps of the cycle have to be planned, avoiding the overloading of implemented system. For example, data from social media talking about an emergency situation may be prioritised over the rest of the data, to allow Emergency Response Teams (ERTs) react as soon as possible.
Moreover, different data sources can describe the same real world entities in such different ways, finding conflicting information, different datatypes, etc. Taking care of how data sources describe their contents will lead to an easier integration step, lowering development, analytics and maintenance costs over time.

4.3 Veracity

There is also an increasing concern on data trustworthiness. Different data sources can have meaningful differences in terms of quality, coverage, accuracy, timeliness and consistency of the provided data. In fact, [12] conclude that redundancy, consistency, correctness and copying between sources are the most recurrent issues we have to deal with when we try to find trustworthy information from a wide variety of heterogeneous sources.
As pointed out by [13], *data provenance is fundamental to understanding data quality* . They also highlight that established information storage systems may not be adequate to keep semantic sense of data.
Several efforts are trying to convert existing data in high quality data, providing an extra confidence layer in which data analysts can rely. In a previous research [14], we introduced a provenance data model to be used in user-generated Linked Data datasets, which follow W3C's PROV-O ontology[7]. Some other researches as [15] or [16] also provide some mechanisms to measure the quality and trust in Linked Data.

4.4 Velocity

Finally, we must assume that data generation is experiencing an exponential growth. That forces our IT structure to not only tackle with volume issues, but with high processing rates. A widely spread concept among data businesses is that sometimes you can not rely on five-minute-old data for your business logic.
That is why *streaming data* has moved from academic fields to industry to solve velocity problems. There are two main reasons to consider streaming processing:

- Sometimes, input data is too fast to store in their entirety without rocketing costs.
- If applications mandate immediate response to the data, batch processes are not suitable. Due to the rise of smartphone applications, this trend is increasingly becoming a common scenario.

[7] http://www.w3.org/TR/prov-o/

5 Linked Open Data as a viable approach

In the previous section, we identified some of the challenges Smart Cities will need to face in the following years. The data lifecycle model proposed at Figure 5 relies on Linked Open Data principles to try to solve these issues, reducing costs and enabling third parties to develop new business models on top of Linked Open Data.
Next we describe how Linked Open Data principles could help in the model's stages:

5.1 Discovery

Before starting any process related with data management, where that data can be found must be known. Identifying the data sources that can be queried is a fundamental first step in any data life-cycle. These data sources can be divided in two main groups: *a)* internal, when the team in charge of creating and maintaining the data is the same that makes use of it, or *b)* external, when data is provided by a third party.
The first scenario usually provides a good understanding of the data, as their generation and structure is designed by the same people who are going to use them. Whereas in real applications, its becoming more common to turn to external data sources to use in the business logic algorithms. Data scientists and developers make use of external datasets for analysing them, expecting to get new insights and create new opportunities from existing data. Luckily, some initiatives help greatly whilst searching for new Open Data sources.

- *The Datahub*[8] is a data management platform from the *Open Knowledge Foundation*, providing nearly 11,000 open datasets as of September 2013. The Datahub relies on *CKAN*[9], an open-source software tool for managing and publishing collections of data. The Datahub's datasets are openly accessible, but data formats can vary from CSV files to RDF , going through JSON, XML, etc.
- The *Linking Open Data Cloud* (LOD Cloud)[10] is an Open Data subset whose catalogues are available on the Web as Linked Data, containing links to other Linked Data sets. LOD Cloud is commonly referred as the biggest effort to bring together Linked Open Data initiatives, grouping 337 datasets as of September 2013. The central node of the LOD Cloud is DBpedia [17, 18], a crowd-sourced community effort to extract structured information from Wikipedia and make it available on the Web.
- Sindice [19] is a platform to build applications on top of semantically markup data on the Web, such us RDF, RDFa, Microformats or Microdata. The main difference is that Sindice does not keep the found documents, but the URL where semantic data can be found. This makes Sindice the closest approach to a traditional document search engine adapted for the Semantic Web.

[8] `http://datahub.io/`
[9] `http://ckan.org/`
[10] `http://lod-cloud.net/`

– Finally, Sig.ma [20] uses Sindice's search engine to construct a view on top of the discovered data on the Web in an integrated information space.

The projects described above can establish the basis to search for external data sources, on top of which further analysis and refinement processes can be built.

5.2 Capture

Data are undoubtedly the basis of Smart Cities: services offered to citizens, decisions offered to city rulers by Decision Support Systems, all of them work thanks to big amounts of data inputs. These data are captured from a wide variety of sources, like sensor networks installed along the city, social networks, publicly available government data or citizens/users who prosume data through their devices (they crowdsource data, or the devices themselves generate automatically sensing data). In most cases, these sources publish data in a wide set of heterogeneous formats, forcing data consumers to develop different connectors for each source. As can be seen in section 5.3, there are a lot of different and widely extended ontologies which can represent data acquired from sources found in Smart Cities, easing the capture, integration and publication of data from heterogeneous domains. In this section, different sources of data which can be found in Smart Cities are shown, while in section 5.3 the transformation process from their raw data to Linked Data is exposed.

Sensor Networks A sensor network is composed by low-cost, low-power, small sized and multifunctional sensor nodes which are densely deployed either inside the phenomenon or very close to it [21]. In a smart city, these sensor networks are used for a wide range of applications, from the simple analysis of air quality[11] to the complex representation of public transport services[12], through the sensors embedded in citizens smartphones. For example, the SmartSantander project envisions the deployment of 20,000 sensors in four European cities [22]. Nowadays, due to the existence of open-source and cheap hardware devices like Arduino[13] or Raspberry Pi[14], the amount of collaborative and social sensor networks is growing faster and faster. Furthermore, there are software platforms like Xively[15] or Linked Sensor Middleware [23], which allow users to share the captured data from their own sensor networks in an easy way.

Social Networks Since the adoption of the Web 2.0 paradigm [24], users have become more and more active when interacting with the Web.

[11] `http://helheim.deusto.es/bizkaisense/`
[12] `http://traintimes.org.uk/map/tube/`
[13] `http://www.arduino.cc/`
[14] `http://www.raspberrypi.org/`
[15] `https://xively.com`

The clearest example of this transformation of the Web can be found in social networks and the high growth of their users. For example, at the end of the second quarter of 2013, Facebook has almost 1.2 billion users[16], while at the end of 2012, Twitter reached more than 200 million monthly active users[17]. Although users of social networks generate a lot of data, it is hard to manipulate them because users write in a language not easily understood by machines. To solve this issue many authors have worked with different Natural Language Processing (NLP) techniques. For example, NLP and Named Entity Recognition (NER) systems [25] can be used to detect tweets which talk about some emergency situation like a car crash, an earthquake and so on; and to recognize different properties about the emergency situation like the place or the magnitude of this situation [26, 27]. Extracting data from relevant tweets could help emergency teams when planning their response to different types of situations as can be seen at [28–30].

Government Open Data Government Open Data has gained a lot of value in recent years, thanks to the proliferation of Open Data portals from different administrations of the entire World. In these portals, the governments publish relevant data for the citizens, in a heterogeneous set of formats like CSV, XML or RDF. Usually, data from these portals can be consumed by developers in an easy way thanks to the provided APIs, so there are a lot of applications developed over these data. As citizens are the most important part of Smart Cities, these applications make them an active part in the governance of the city.
To illustrate the importance of Government Open Data, in Table 1 some Open Data portals are shown.

Table 1. Open Data portals around the World.

Name	**Public Administration**	**No. of datasets (Sept. 2013)**	**API**
Data.gov	Government of USA	97,536	REST, SOAP, WMS
Data.gov.uk	Government of UK	10,114	REST
Data.gc.ca	Government of Canada	197,805	REST
Open Data Euskadi	Government of Basque Country	2,127	RSS, Java API, REST
Datos Abiertos de Zaragoza	Council of Zaragoza	112	SPARQL

[16] http://techcrunch.com/2013/07/24/facebook-growth-2/
[17] https://twitter.com/twitter/status/281051652235087872

Mobile sensing and user generated data. Another important data source in which Smart Cities can capture data are the citizens themselves. Citizens can interact with city in multiple ways: for example, in [14] an interactive 311 service is described. In this work, authors propose a model to share and validate, through provenance and reputation analysis, reports about city issues published by its citizens. In Urbanopoly [31] and Urbanmatch [32], different games are presented to tourist with the aim of gathering data and photographies from tourist points of interest in the city using their smartphones' cameras. In the same line, csxPOI [33] creates semantically annotated POIs through data gathered by citizens, allowing semi-automatic merging of duplicate POIs and removal of incorrect POIs.
As have been shown, in a Smart City a lot of data sources can be found, publishing an abundant stream of interesting data in a different and heterogeneous manner. In section 5.3, how to transform these data into standard formats is shown.

5.3 Curate

As can be seen at section 2, the Linked Data paradigm proposed the Resource Description Framework (RDF) as the best format to publish data and encourage the reuse of widely extended ontologies. In this section we explain **what** is an ontology, **which** are the most popular ontologies and **how** we can map previously captured raw data to a proper ontology. At the end of this section a set of best practices to construct suitable URIs for Linked Data are shown.
As defined by [34], an ontology is *a formal explicit description of concepts in a domain of discourse, properties of each concept describing various features and attributes of the concept, and restrictions on slots*. According to this definition, an ontology has *Classes* which represent the concept, *Properties* which represent different characteristics of *Classes* and *Restrictions* on the values of these properties and relationships among different *Classes*. An ontology allows modelling data avoiding most ambiguities originated when fusing data from different sources, stimulating the interoperability among different sources. As seen in section 5.2, data may come from a wide variety of sources in Smart Cities, whereby the ontologies seem to be a suitable option to model these data.
The following works use ontologies to model different data sources which can be found in a Smart City. In Bizkaisense project [35], diverse ontologies like Semantic Sensor Network ontology (SSN) [36], Semantic Web for Earth and Environmental Terminology (SWEET) [37] or Unified Code for Units of Measure ontology (UCUM)[18] are used to model raw data from air quality stations from the Basque Country. AEMET Linked Data project[19] has developed a network of ontologies composed by SSN ontology, OWL-Time ontology[20], wsg84_pos ontology[21], GeoBuddies on-

[18] `http://idi.fundacionctic.org/muo/ucum-instances.html`
[19] `http://aemet.linkeddata.es/models.html`
[20] `http://www.w3.org/TR/owl-time/`
[21] `http://www.w3.org/2003/01/geo/wgs84_pos`

tology network[22] and its own AEMET ontology, to describe measurements taken by meteorological stations from AEMET (Spanish National Weather Service). In [38] authors extend SSN ontology to model and publish as Linked Data the data stream generated by the sensors of an Android powered smartphone.

Another example of semantic modelling of infrastructures from a city can be found in LinkedQR [39]. LinkedQR is an application that eases the managing task of an art gallery allowing the elaboration of interactive tourism guides through third parties Linked Data and manual curation. LinkedQR uses MusicOntology [40] to describe the audioguides and Dublin Core [41], DBpedia Ontology and Yago [42] to describe other basic information.

LinkedStats project[23] takes data about waste generation and population of Biscay to develop a statistical analysis about the correlation between these two dimensions of the data. It models these statistical data through the RDF Data Cube Vocabulary [**?**], an ontology developed for modelling multi-dimensional data in RDF. At last, in [43] the authors show how Linked Data enables the integration of data from different sensor networks.

The mapping between raw data and ontologies, usually is made by applications created *ad-hoc* to each case; Bizkaisense, AEMET Linked Data and LinkedStats have their own Python scripts to generate proper RDF files from raw data. In the case of LinkedQR, it has a control panel where the manager can manually type data and map to a desired ontology. Instead, there are tools designed for transforming raw data into structured data. One of them is Open Refine[24] (formerly Google Refine). Open Refine is a web-tool which can apply different manipulations to data (facets, filters, splits, merges, etc.) and export data in different formats based on custom templates. Additionally, Google Refine's RDF Extension allows exporting data in RDF.

Another interesting tool is Virtuoso Sponger, a component of OpenLink Virtuoso[25] which generates Linked Data from different data sources, through a set of extractors called *Cartridges*. There are different Cartridges which support a wide variety of input formats (CSV, Google KML, xHTML, XML, etc.) and vendor specific Cartridges too (Amazon, Ebay, BestBuy, Discogs, etc.).

After modelling data, one of the most important concepts in Linked Data are the URIs or Unified Resource Identifiers. As shown in section 2, to publish a data resource as Linked Data it has to be identified by an HTTP URI which satisfies these conditions:

- An HTTP URI is unique and consistent.
- An HTTP URI can be accessed from everywhere by everyone.
- The URI and its hierarchy are auto-descriptive.

Designing valid URIs is a very important step into the publication of Linked Data: if you change the URIs of your data, all the incoming links

[22] `http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/ontologies/83-geobuddies-ontologies`

[23] `http://helheim.deusto.es/linkedstats/`

[24] `http://openrefine.org/`

[25] `http://virtuoso.openlinksw.com/`

from external sources are going to be broken. To avoid this issue there is a set of good practices, proposed in [44]:

- **Be on the web.** Return RDF for machines and HTML for humans through standard HTTP protocol.
- **Don not be ambiguous.** Use a URL to describe a document and a different URI to identify real-world objects. A URI can not stand for both document and real-world object.

To apply these good practices correctly, authors propose two solutions, which nowadays have been widely adopted by Linked Data community:

- **303 URIs.** Use `303 See Other` status code for redirecting to the proper RDF description of the real world object or to the HTML document. For example:
  - `http://helheim.deusto.es/hedatuz/resource/biblio/5112` - A URI identifying a bibliographic item.
  - `http://helheim.deusto.es/hedatuz/page/biblio/5112` - The HTML view of the item.
  - `http://helheim.deusto.es/hedatuz/data/biblio/5112` - A RDF document describing the item.
- **Hash URIs.** URIs contains a *fragment*, a special part separated by the symbol #. The client has to strip-off the fragment from the URI before requesting it to server. Server returns a RDF document in which the client has to search the fragment.

Tools implementing these techniques are described on section 5.5.

5.4 Store

Once data is mapped to a proper ontology and the RDF files are generated, is time to store them. Due to the big amount of data generated in a city, an appropriate storage has to:

- Support different input data formats.
- Manage and index big amounts of data properly.
- Execute queries over data in an efficient way.
- Offer different interfaces and APIs to users, allowing them to exploit data in a wide variety of formats.

Along this section, the first three points are discussed, while the fourth is discussed in section 5.5. Before a wide description of each analysed datastore is given, a brief description is presented in Table 2.
The first datastore to be reviewed is Virtuoso by Openlink, mentioned in section 5.3. Virtuoso is a hybrid server that manages SQL, XML and RDF in a single server. It is available in both Open Source and commercial versions. As seen in Table 2, it supports a wide variety of input and output formats. It has a SPARQL endpoint, accessible thus by web-interface as by HTTP GET requests, allowing to web-agents the access to data. Virtuoso supports the SPARQL UPDATE syntax, allowing the update of datastore through HTTP POST requests; and it provides connectors for different Java powered RDF engines, like Jena[26], Sesame[27]

[26] `http://jena.apache.org/`
[27] `http://www.openrdf.org/`

Table 2. Summary of characteristics of selected datastores.

Datastore	Input formats	API	Output formats
Virtuoso	RDF/XML, N3, Turtle, N-Triples, N-Quads, etc.	SPARQL, SPARQL UPDATE, Java (Jena, Sesame, Redland)	HTML, Spreadsheet, XML, RDF+JSON, JS, NTriples, RDF/XML, CSV
Stardog	NTRIPLES, RDF/XML, Turtle, TRIG, TRIX, N3, N-Quads	SPARQL (CLI, HTTP), Java, JS, Groovy, Spring, Ruby	N-Triples, RDF/XML, TURTLE, TRIG, TRIX, N3, N-Quads
Fuseki	RDF/XML, N-Triples, Turtle, etc.	SPARQL, SPARQL UPDATE, Java	JSON, XML, Text, CSV, TSV

or Redland[28]. Further, it supports some OWL properties for reasoning. According to the Berlin SPARQL Benchmark [45] Virtuoso 7 can load one billion triples in 27:11 minutes.

Another datastore which is becoming popular is Stardog[29]. Developed by Clark & Parsia, Stardog is a RDF database which supports SPARQL querying and OWL reasoning. It offers a Command Line Interface to manage the different databases (create, remove, add/remove data, SPARQL queries, etc.), while they can be queried through HTTP too. Furthermore it has its own query syntax which indexes the RDF literals; and its own Java library to manage databases from Java applications. It supports OWL 2 reasoning, supporting different OWL profiles[30] like QL, RL, EL or DL.

The last analysed datastore is Fuseki[31]. Part of Jena's framework, Fuseki (formerly known as Joseki) offers RDF data over HTTP, in a REST style. Fuseki implements W3C's SPARQL 1.1 Query, Update, Protocol and Graph Store HTTP Protocol. It has a web-panel to manage the datastore and can interact with the rest of the Jena components.

As can be seen at Table 2, all analysed datastores are similar in terms of input/output formats or offered APIs. But there are differences in other aspects, like security: Fuseki does not support users nor access roles. Another difference is the installation and executing complexity: while Fuseki and Stardog are launched as a Java JAR file, Virtuoso can be installed through Debian's package system and launched as a UNIX daemon. In the other hand, Virtuoso is more than a "simple" RDF store, Virtuoso is a relational database engine, an application server in which both preinstalled applications and our own applications can be launched, and much more. Furthermore, Virtuoso can be installed in a cluster formed by multiple servers.

[28] `http://librdf.org/`
[29] `http://stardog.com/`
[30] `http://www.w3.org/TR/owl2-profiles/`
[31] `http://jena.apache.org/documentation/serving_data/index.html`

Concluding this section, we can say that Fuseki can be used in light-weight installations, when hardware and data are limited; Stardog in more complex systems, due to its fast query execution times. Meanwhile, Virtuoso offers more services like Sponger (described in section 5.3) or Semantic Wiki, whereby, it can be suitable for environments which need more than the simple storage of RDF triples.

5.5 Publish

Publication stage is one of the most important stages in the life cycle of Linked Data in Smart Cities, because this stage determines how citizens or developers can acquire Linked Data to exploit (section 5.7) through different discovery methods (section 5.1). As we saw in section 5.4, the three proposed RDF stores include some publication API or SPARQL endpoint, but, sometimes the specifications of the system to be deployed need additional features at this publication stage.

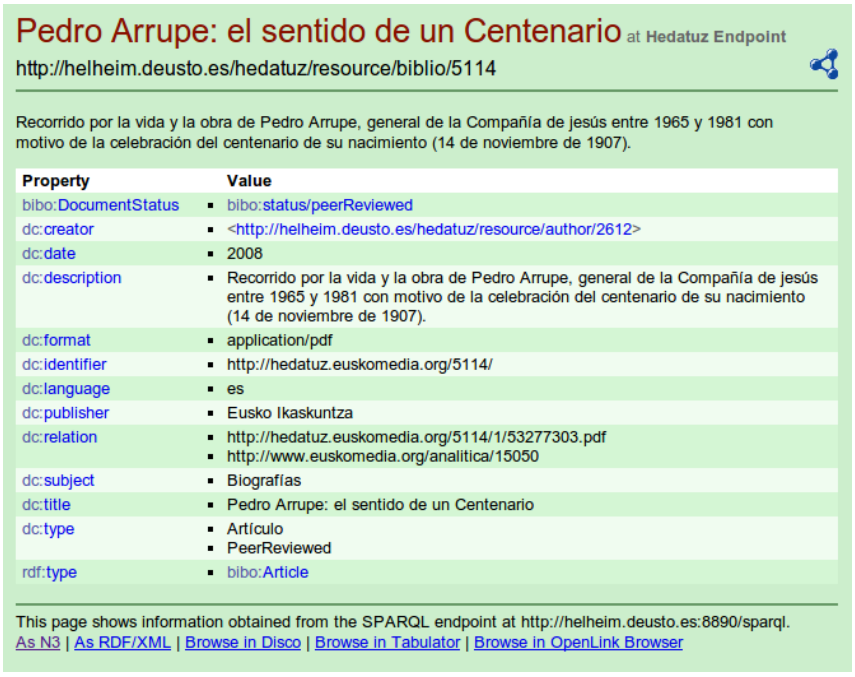

Fig. 6. Example of HTML visualisation of a resource from Hedatuz dataset by Pubby.

One of these features can be the `303 Redirection` explained at section 5.3. Although all datastores mentioned on section 5.4 offer a SPARQL endpoint to explore data, Linked Data paradigm demands resolvable HTTP URIs as resource identifiers. Fortunately, there are tools which

fulfil this demand. One of them, Pubby[32], adds Linked Data interfaces to SPARQL endpoint. Pubby queries the proper SPARQL endpoint to retrieve data related to a given URI and manages the `303 Redirection` mechanism. Depending on the `Accept` header of the HTTP request, Pubby redirects the client to the HTML view of data or to the RDF document describing the resource. Pubby can export data in RDF/XML, NTriples, N3 and Turtle. In Figure 6 an example of the HTML view of a resource is shown.
D2R Server [46] allows the publication of relational databases as Linked Data. At first D2R requires a mapping from the tables and columns of the database to selected ontologies using D2RQ Mapping Language. Once this mapping is done, D2R offers a SPARQL endpoint to query data and a Pubby powered interface.

```
# D2RQ Namespace
@prefix d2rq:        <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
# Namespace of the ontology
@prefix : <http://annotation.semanticweb.org/iswc/iswc.daml#> .

# Namespace of the mapping file; does not appear in mapped data
@prefix map: <file:///Users/d2r/example.ttl#> .

# Other namespaces
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

map:Database1 a d2rq:Database;
    d2rq:jdbcDSN "jdbc:mysql://localhost/iswc";
    d2rq:jdbcDriver "com.mysql.jdbc.Driver";
    d2rq:username "user";
    d2rq:password "password";
    .
# ------------------------------------------------
# CREATE TABLE Conferences (ConfID int, Name text, Location text);

map:Conference a d2rq:ClassMap;
    d2rq:dataStorage map:Database1.
    d2rq:class :Conference;
    d2rq:uriPattern "http://conferences.org/comp/confno@@Conferences.ConfID@@";
    .
map:eventTitle a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:Conference;
    d2rq:property :eventTitle;
    d2rq:column "Conferences.Name";
    d2rq:datatype xsd:string;
    .
map:location a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:Conference;
    d2rq:property :location;
    d2rq:column "Conferences.Location";
    d2rq:datatype xsd:string;
    .
```

Fig. 7. D2RQ mapping file. Example taken from http://d2rq.org/d2rq-language.

Besides the publication of the data itself, it is important to consider the publication of provenance information about it. In [47] the authors identify the publication of provenance data as one of the main factors that influence web content trust. At the time of publishing provenance infor-

[32] `http://wifo5-03.informatik.uni-mannheim.de/pubby/`

mation two approaches can be taken: the first is to publish basic metadata like when or who created and published the data; the second is to provide a more detailed description of where the data come from, including versioning information or the description of the data transformation workflow, for example. Some ontologies help us in the process of providing provenance descriptions of Linked Data. Basic provenance metadata can be provided using Dublin Core terms, like *dcterms*[33]*:contributor*, *dcterms:creator* or *dcterms:created*. Other vocabularies like the Provenance Vocabulary [48] or the Open Provenance Model (OPM) [49] provide ways to publish detailed provenance information like the mentioned before. The W3C has recently created the PROV Data Model [50], a new vocabulary for provenance interchange on the Web. This PROV Data Model is based on OPM and describes the entities, activities and people involved in the creation of a piece of data, allowing the consumer to evaluate the reliability of the data based on the their provenance information. Furthermore, PROV was deliberately kept extensible, allowing various extended concepts and custom attributes to be used. For example, the Uncertainty Provenance (UP) [51] set of attributes can be used to model the uncertainty of data, aggregated from heterogeneously divided trusted and untrusted sources, or with varying confidence.

Here we can find an example of how bio2rdf.org -an atlas of post-genomic data and one of the biggest datasets in the LOD Cloud- represent the provenance data of its datasets, using some of the mentioned ontologies in conjunction with VoID, a dataset description vocabulary:

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix void: <http://rdfs.org/ns/void#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix prov: <http://www.w3.org/ns/prov#> .

<http://bio2rdf.org/bio2rdf_dataset:bio2rdf-affymetrix-20121002>
    a void:Dataset ;
    dcterms:created "2012-10-02"^^xsd:date ;
    rdfs:label "affymetrix dataset by Bio2RDF on 2012-10-02" ;
    dcterms:creator <affymetrix> ;
    dcterms:publisher <http://bio2rdf.org> ;
    void:dataDump <datadump> ;
    prov:wasDerivedFrom
        <http://bio2rdf.org/bio2rdf_dataset:affymetrix> ;
    dcterms:rights "restricted-by-source-license" ,
        "attribution" , "use-share-modify" ;
    void:sparqlEndpoint <http://affymetrix.bio2rdf.org/sparql> .
```

In the process of publishing data from Smart Cities as Linked Data, new ontologies are going to be created to model the particularities of each city. The creators of these ontologies have to publish a suitable documentation, allowing the proper reuse of them. A tool for publishing ontologies and their documentation is Neologism[34]. Neologism shows ontologies in a human-readable manner, representing class, subclass and property relationships through diagrams.

[33] http://purl.org/dc/terms/
[34] http://neologism.deri.ie/

5.6 Linkage

Connecting existing data with other available resources is a major challenge for easing data integration. Due to its interlinked nature, Linked Data provides a perfect base to connect the data present in a given dataset.
The linkage stage starts a loop on the model after the publishing step, establishing relationships between existing data and external datasets, in order to provide links to new information stores.
Different frameworks have been developed to deal with class and properties matching. The basis of these frameworks is to provide data discovery features through links to external entities related to the items used in the analysis.
The *Silk - Link Discovery Framework* [52] offers a flexible tool for discovering links between entities within different Web data sources. Silk makes use of *Silk - Link Specification Language* (Silk-LSL), a declarative language which lets data publishers specify which RDF link types should be discovered providing two related datasets, and the conditions under data items must fulfil to be interlinked. Silk framework's architecture is depicted in Figure 8. As an example, a script in Silk-LSL can be written to match cities between *DBpedia* ontology's *City* or *PopulatedPlace* classes, and *GeoName*'s feature class *gn:P*. As constraints, string similarity metrics can be used to match city names, and take into consideration cities' bounding boxes (i.e. the margins projected on a map) to check overlaps.

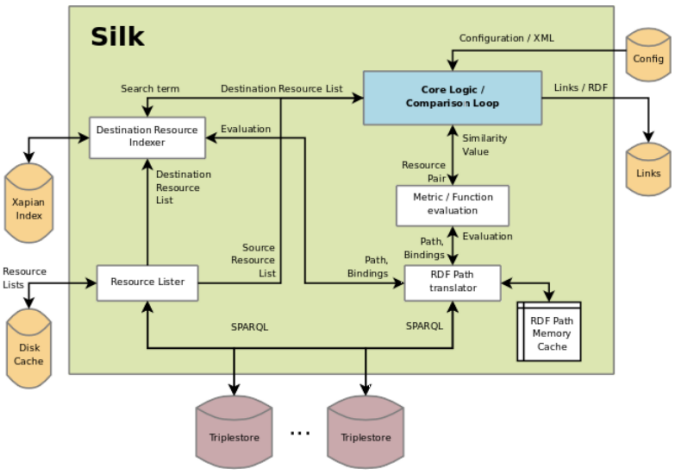

Fig. 8. Silk framework architecture

With a similar approach *LIMES* (LInk discovery framework for MEtric Spaces) [53] can be used for the discovery of links between Linked Data knowledge bases, focusing on a time-efficient approach especially when

working with large-scale matching tasks. LIMES relies on *triangle inequality* mathematical principles for distance calculations, which reduce the number of comparisons necessary to complete a mapping by several orders of magnitude. This approach helps detecting the pairs that will not fulfil the requirements in an early stage, thus avoiding spending time in more time-consuming processing. The architecture followed by LIMES framework is depicted in Figure 9.

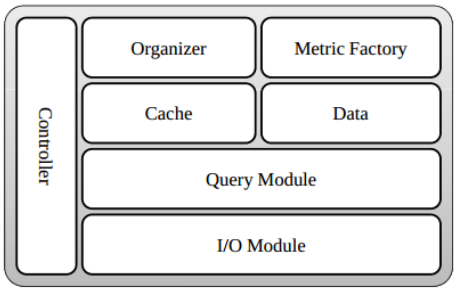

Fig. 9. LIMES framework architecture

5.7 Exploit

At this stage, the focus is located on exploiting data for business-logic processes, should they involve data mining algorithms, analytics, reasoning, etc.

Whereas complex processing algorithms can be used independently of the dataset format, Linked Open Data can greatly help at reasoning purposes. Linked Open Data describes entities using ontologies, semantic constraints and restriction rules (belonging, domain, range, etc.) which favor the inference of new information from the existing one. Thanks to those semantics present in Linked Data, algorithms are not fed with raw values (numbers, strings...), but with semantically meaningful information (height in cm, world countries, company names...), thus resulting in higher quality outputs and making algorithms more error aware (i.e., if a given algorithm is in charge of mapping the layout of a mountainous region, and founds the height of one of the mountains to be *3.45*, it is possible to detect the conversion has failed at some point, as height datatype was expected to be given in *meters*).

As seen in sections 5.2 and 5.2, sensors and social networks are common data input resources, generating huge amounts of data streamed in real time. The work done in [54] comprises a set of best practices to publish and link stream data to be part of the Semantic Web.

However, when it comes to exploiting Linked Data streams, SPARQL can find its limits [55]. Stream-querying languages such as CQELS's[35]

[35] `https://code.google.com/p/cqels/`

language (an extension of the declarative SPARQL 1.1 language using the EBNF notation) can greatly help in the task. CQELS [56] (Continuous Query Evaluation over Linked Stream) is a native and adaptive query processor for unified query processing over Linked Stream Data and Linked Data developed at DERI Galway.
Initially, a query pattern is added to represent window operators on RDF Stream:

```
GraphPatternNotTriples ::= GroupOrUnionGraphPattern |
OptionalGraphPattern | MinusGraphPattern | GraphGraphPattern |
*StreamGraphPattern* | ServiceGraphPattern | Filter | Bind
```

Assuming that each stream has an IRI as identification, the *StreamGraphPattern* is defined as follows:

```
StreamGraphPattern ::= 'STREAM' '['Window']' VarOrIRIref
'{'TriplesTemplate'}'
Window ::= Rangle|Triple|'NOW'|'ALL'
Range ::= 'RANGE' Duration ('SLIDE' Duration |'TUMBLING')?
Triple ::= 'TRIPLES' INTEGER
Duration ::= (INTEGER 'd'|'h'|'m'|'s'|'ms'|'ns')+
```

An example query could be:

```
PREFIX lv: <http://deri.org/floorplan/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?locName FROM NAMED <http://deri.org/floorplan/>
WHERE {
        STREAM <http://deri.org/streams/rfid> [NOW]
        {?person lv:detectedAt ?loc}
        {?person foaf:name "AUTHORNAME"^^xsd:string }

        GRAPH <http://deri.org/floorplan/>
        {?loc lv:name ?locName}
}
```

Eventually, both batch and streaming results can be consumed through *REST* services by web or mobile applications, or serve as input for more processing algorithms until they are finally presented to end-users.

5.8 Visualise

In order to make meaning from data, humans have developed a great ability to understand visual representations. The main objective of data visualisation is to communicate information in a clean and effective way through graphical means. It is also suggested that visualisation should also encourage users engagement and attention.
The *"A picture is worth a thousand words"* saying reflects the power images and graphics have when expressing information, and can condense big datasets into a couple of representative, powerful images.
As Linked Data is based on subject-predicate-object triples, graphs are a natural way to represent triple stores, where subject and object nodes are inter-connected through predicate links. When further analysis is applied on triples, a diverse variety of representations can be chosen to show processed information: charts, infographics, flows, etc.[57]

Browser-side visualisation technologies such as d3.js[36] (by Michael Bostock) and Raphaël[37] are JavaScript-based libraries to allow the visual representation of data on modern web browsers, allowing anybody with a minimum internet connection try to understand data patterns in a graphical form.
For developers not familiar with visualisation techniques, some investigations are trying to enable the automatic generation of graphical representations of Linked Data query results. The *LDVM* (Linked Data Visualization Model) [58] is proposed as a model to rapidly create visualisations of RDF data. *LODVisualization*[38] is an implemented prototype which supports the LDVM.
Visualbox[39] is a simplified edition of LODSPeaKr[40] focused on allowing people create visualisations using Linked Data. In Figure 10, a SPARQL query to retrieve the number of Nobel laureates in Physics by country is displayed in interactive pie and bar charts.

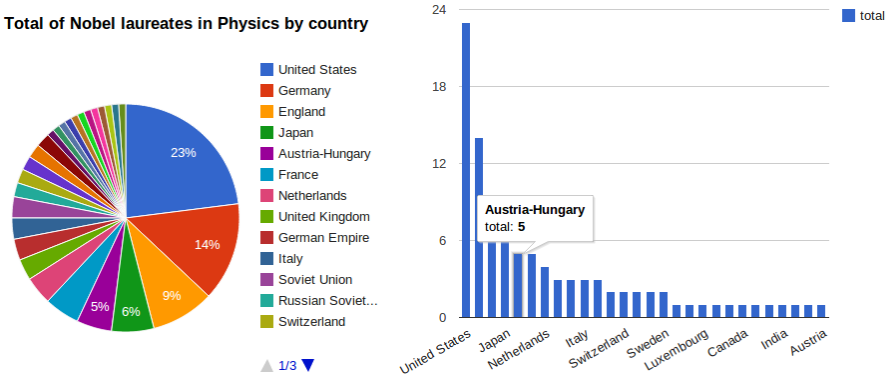

Fig. 10. Visualbox graph example (ownership: Alvaro Graves).

Independently of how visualisations are generated, they provide a perfect solution to present high-quality, refined data to end-users.

[36] http://d3js.org/
[37] http://raphaeljs.com/
[38] http://lodvisualisation.appspot.com/
[39] http://alangrafu.github.io/visualbox/
[40] http://lodspeakr.org/

6 Conclusions

In this chapter we have proposed Linked Data as a suitable paradigm to manage the entire data life cycle in Smart Cities. As can be seen, along this chapter we expose a set of guidelines for public or private managers which want to contribute with data from their administration or enterprise into a Smart City, bringing closer existing tools and exposing practical knowledge acquired by authors while working with Linked Data technologies. The proposed data life cycle for Smart Cities covers the entire travelling path of data inside a Smart City, and the mentioned tools and technologies fulfil all the needed tasks to go forward on this path.

But Linked Open Data is not all about technology. The *Open* term of Linked Open Data is about the awareness of public (and private) administrations to provide citizens with all the data which belong to them, making the governance process more transparent; the awareness of developers to discover the gold behind data and the awareness of fully informed citizens participating on decision making processes: Smart Cities, smart business and Smart Citizens.

Urban Linked Data applications also empower citizens' role of first level data providers. Thanks to smartphones, each citizen is equipped with a full set of sensors which are able to measure the city's pulse at every moment: traffic status, speed of each vehicle to identify how they are moving, reporting of roadworks or malfunctioning public systems, and so forth. Citizens are moving from data consumers to data *prosumers*, an aspect data scientists and application developers can benefit from to provide new services for Smart Cities.

7 Acknowledgments

This work has been supported by research project grants Future Internet II (IE11-316) and SmarTUR (IE12-343) granted by the Basque Government and ADAPTA (IPT-2011-0949-430000) by the Spanish Government. Mikel Emaldi and Jon Lázaro are grateful to University of Deusto for their PhD grants. Oscar Peña holds a PhD grant from the Basque Government.

8 Acronyms and terms

CSV Comma Separated Values.
RDF Resource Description Framework.
JSON JavaScript Object Notation.
XML eXtensible Markup Language.
LOD Linked Open Data.
RDFa Resource Description Framework in Attributes.
NLP Natural Language Processing.
NER Named Entity Recognition.
API Application Programming Interface.
REST REpresentational State Transfer.
SOAP Simple Object Access Protocol.
WMS Web Map Service.
RSS Really Simple Syndication.
SPARQL SPARQL Protocol and RDF Query Language.
POI Point Of Interest.
URI Uniform Resource Identifier.
KML Keyhole Markup Language.
HTML HyperText Markup Language.
HTTP HyperText Transfer Protocol.
N3 Notation3.
JS JavaScript.
TSV Tab Separated Values.
SQL Structured Query Language.
OWL Web Ontology Language.
W3C World Wide Web Consortium.

References

1. World Health Organization: Urbanization and health. Bull World Health Organ **88** (2010) 245–246
2. Bizer, C., Boncz, P., Brodie, M.L., Erling, O.: The meaningful use of big data: four perspectives – four challenges. SIGMOD Rec. **40**(4) (January 2012) 56–60
3. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific american **284**(5) (2001) 28–37
4. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. International Journal on Semantic Web and Information Systems (IJSWIS) **5**(3) (2009) 1–22
5. Initiative, D.D.: Overview of the DDI version 3.0 conceptual model (April 2008)
6. Ball, A.: Review of data management lifecycle models. (2012)
7. Burton, A., Treloar, A.: Designing for discovery and re-use: the 'ANDS data sharing verbs' approach to service decomposition. International Journal of Digital Curation **4**(3) (2009) 44–56
8. Michener, W.K., Jones, M.B.: Ecoinformatics: supporting ecology as a data-intensive science. Trends in ecology & evolution **27**(2) (February 2012) 85–93
9. Auer, S., Bühmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., Lehmann, J., Martin, M., Mendes, P.N., Van Nuffelen, B., Stadler, C., Tramp, S., Williams, H.: Managing the life-cycle of linked data with the LOD2 stack. In: The Semantic Web–ISWC 2012, Springer (2012) 1–16
10. deRoos, D., Eaton, C., Lapis, G., Zikopoulos, P., Deutsch, T.: Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media (2011)
11. Russom, P.: Big data analytics. TDWI Best Practices Report, Fourth Quarter (2011)
12. Li, X., Dong, X.L., Lyons, K., Meng, W., Srivastava, D.: Truth finding on the deep web: is the problem solved? In: Proceedings of the 39th International Conference on Very Large Data Bases. PVLDB'13, VLDB Endowment (2013) 97–108
13. Buneman, P., Davidson, S.B.: Data provenance–the foundation of data quality (2013)
14. Emaldi, M., Pena, O., Lázaro, J., López-de-Ipiña, D., Vanhecke, S., Mannens, E.: To trust, or not to trust: Highlighting the need for data provenance in mobile apps for smart cities. In: Proceedings of the 3rd International Workshop on Information Management for Mobile Applications. (2013) 68–71
15. Hartig, O., Zhao, J.: Using web data provenance for quality assessment. In: Proceedings of the International Workshop on Semantic Web and Provenance Management, Washington DC, USA. (2009)
16. Bizer, C., Cyganiak, R.: Quality-driven information filtering using the WIQA policy framework. Web Semantics: Science, Services and Agents on the World Wide Web **7**(1) (2009) 1–10
17. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The semantic web. Springer (2007) 722–735

18. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia-a crystallization point for the web of data. Web Semantics: Science, Services and Agents on the World Wide Web **7**(3) (2009) 154–165
19. Tummarello, G., Delbru, R., Oren, E.: Sindice. com: Weaving the open linked data. In: The Semantic Web. Springer (2007) 552–565
20. Tummarello, G., Cyganiak, R., Catasta, M., Danielczyk, S., Delbru, R., Decker, S.: Sig. ma: Live views on the web of data. Web Semantics: Science, Services and Agents on the World Wide Web **8**(4) (2010) 355–364
21. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks. Communications magazine, IEEE **40**(8) (2002) 102–114
22. Sanchez, L., Galache, J.A., Gutierrez, V., Hernandez, J., Bernat, J., Gluhak, A., Garcia, T.: SmartSantander: the meeting point between future internet research and experimentation and the smart cities. In: Future Network & Mobile Summit (FutureNetw), 2011. (2011) 1–8
23. Le-Phuoc, D., Quoc, H.N.M., Parreira, J.X., Hauswirth, M.: The linked sensor middleware–connecting the real world and the semantic web. Proceedings of the Semantic Web Challenge (2011)
24. O'reilly, T.: What is web 2.0: Design patterns and business models for the next generation of software. Communications & strategies (1) (2007) 17
25. Maynard, D., Tablan, V., Ursu, C., Cunningham, H., Wilks, Y.: Named entity recognition from diverse text types. In: Recent Advances in Natural Language Processing 2001 Conference. (2001) 257–274
26. Sixto, J., Pena, O., Klein, B., López-de-Ipiña, D.: Enable tweet-geolocation and don't drive ERTs crazy! improving situational awareness using twitter. In: SMERST 2013: Social Media and Semantic Technologies in Emergency Response. Volume 1., Coventry, UK (2013) 27–31
27. Martins, B., Anastácio, I., Calado, P.: A machine learning approach for resolving place references in text. In: Geospatial Thinking. Springer (2010) 221–236
28. Abel, F., Hauff, C., Houben, G.J., Stronkman, R., Tao, K.: Twitcident: fighting fire with information from social web streams. In: Proceedings of the 21st International Conference Companion on World Wide Web. (2012) 305–308
29. : Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, author = Vieweg, Sarah and Hughes, Amanda L and Starbird, Kate and Palen, Leysia, year = 2010, pages = 1079–1088
30. Hughes, A.L., Palen, L.: Twitter adoption and use in mass convergence and emergency events. International Journal of Emergency Management **6**(3) (2009) 248–260
31. Celino, I., Cerizza, D., Contessa, S., Corubolo, M., Dell'Aglio, D., Valle, E.D., Fumeo, S.: Urbanopoly – a social and location-based

game with a purpose to crowdsource your urban data. In: Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust. SOCIALCOM-PASSAT '12, Washington, DC, USA, IEEE Computer Society (2012) 910–913
32. Celino, I., Contessa, S., Corubolo, M., Dell'Aglio, D., Valle, E.D., Fumeo, S., Krüger, T.: UrbanMatch - linking and improving smart cities data. In Bizer, C., Heath, T., Berners-Lee, T., Hausenblas, M., eds.: Linked Data on the Web. Volume 937 of CEUR Workshop Proceedings., CEUR-WS (2012)
33. Braun, M., Scherp, A., Staab, S.: Collaborative semantic points of interests. In: The Semantic Web: Research and Applications. Springer (2010) 365–369
34. Noy, N.F., McGuinness, D.L.: Ontology development 101: A guide to creating your first ontology. Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880 (2001)
35. Emaldi, M., Lázaro, J., Aguilera, U., Peña, O., López-de-Ipiña, D.: Short paper: Semantic annotations for sensor open data. In: Proceedings of the 5th International Workshop on Semantic Sensor Networks, SSN12. (2012) 115–120
36. Lefort, L., Henson, C., Taylor, K., Barnaghi, P., Compton, M., Corcho, O., Garcia-Castro, R., Graybeal, J., Herzog, A., Janowicz, K.: Semantic sensor network XG final report. W3C Incubator Group Report (2011)
37. Raskin, R.G., Pan, M.J.: Knowledge representation in the semantic web for earth and environmental terminology (SWEET). Computers & Geosciences **31**(9) (2005) 1119–1125
38. d'Aquin, M., Nikolov, A., Motta, E.: Enabling lightweight semantic sensor networks on android devices. In: The 4th International Workshop on Semantic Sensor Networks 2011 (SSN2011). (October/Autumn 2011)
39. Emaldi, M., Lázaro, J., Laiseca, X., López-de-Ipiña, D.: LinkedQR: improving tourism experience through linked data and QR codes. In: Ubiquitous Computing and Ambient Intelligence. Springer (2012) 371–378
40. Raimond, Y., Abdallah, S., Sandler, M., Giasson, F.: The music ontology. In: ISMIR 2007: 8th International Conference on Music Information Retrieval, Vienna, Austria (September 2007) 417–422
41. Weibel, S., Kunze, J., Lagoze, C., Wolf, M.: Dublin core metadata for resource discovery. Internet Engineering Task Force RFC **2413** (1998) 222
42. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide Web, ACM (2007) 697–706
43. Stasch, C., Schade, S., Llaves, A., Janowicz, K., Bröring145, A.: Aggregating linked sensor data. SEMANTIC SENSOR NETWORKS (2011) 46
44. Ayers, A., Völkel, M.: Cool uris for the semantic web. Woking Draft. W3C (2008)

45. Bizer, C., Schultz, A.: The berlin sparql benchmark. International Journal on Semantic Web and Information Systems (IJSWIS) **5**(2) (2009) 1–24
46. Bizer, C., Cyganiak, R.: D2r server-publishing relational databases on the semantic web. In: Proceedings of the 5th International Semantic Web Conference. (2006) 26
47. Gil, Y., Artz, D.: Towards content trust of web resources. Web Semantics: Science, Services and Agents on the World Wide Web **5**(4) (2007) 227–239
48. Hartig, O.: Provenance information in the web of data. In: Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009. (2009)
49. Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., Van den Bussche, J.: The open provenance model core specification (v1.1). Future Generation Computer Systems **27**(6) (2011) 743 – 756
50. Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., Miles, S., Myers, J., Sahoo, S., Tilmes, C.: PROV-DM: The PROV data model (2013)
51. De Nies, T., Coppens, S., Mannens, E., Van de Walle, R.: Modeling uncertain provenance and provenance of uncertainty in W3C PROV. In: Proceedings of the 22nd international conference on World Wide Web companion, Rio de Janeiro, Brazil (2013) 167–168
52. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk-a link discovery framework for the web of data. In: Proceedings of the International Semantic Web Conference 2010 Posters & Demonstrations Track, Citeseer (2009)
53. Ngomo, A.C.N., Auer, S.: Limes: a time-efficient approach for large-scale link discovery on the web of data. In: Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three, AAAI Press (2011) 2312–2317
54. Sequeda, J., Corcho, O., Taylor, K., Ayyagari, A., Roure, D.D.: Linked stream data: A position paper. In: Proceedings of the 2nd International Workshop on Semantic Sensor Networks (SSN09) at ISWC 2009. Volume 522., CEUR Workshop Proceedings (November 2009) 148–157
55. Della Valle, E., Ceri, S., van Harmelen, F., Fensel, D.: It's a streaming world! reasoning upon rapidly changing information. Intelligent Systems, IEEE **24**(6) (2009) 83–89
56. Le-Phuoc, D., Dao-Tran, M., Parreira, J.X., Hauswirth, M.: A native and adaptive approach for unified processing of linked streams and linked data. In: The Semantic Web–ISWC 2011. Springer (2011) 370–388
57. Khan, M., Khan, S.S.: Data and information visualization methods, and interactive mechanisms: A survey. International Journal of Computer Applications **34**(1) (2011) 1–14
58. Brunetti, J.M., Auer, S., García, R.: The linked data visualization model. In: International Semantic Web Conference (Posters & Demos). (2012)