Linked Open Data as the fuel for Smarter Cities

No Author Given

No Institute Given

Abstract. In the last decade big efforts have been carried out in order to move towards the Smart City concept, both from the academic and industrial points of view. Cities have been sensorized, protocols are constantly refined to deal with the possibilities that new hardware offers, communication networks are offered in all flavours... and meanwhile researchers and data stakeholders try to figure out how to cope with the huge amount of generated data.

Open Data has arisen in the last years as a way to share data to be consumed freely 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. The data generated by a city council is highly difficult to manage in an efficient way, due to the facts described earlier, usually resulting in large amounts of data not being analysed and affecting negatively the end-users of the information.

We propose Linked Open Data, a set of best practices to publish data on the Web proposed by W3C, as a way of publishing data among governments, easing their consumption by anybody, 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.

We also advocate the role of citizens as linked open data providers. User-friendly Linked Data apps should enable citizens to easily contribute with new trustable data that can be linked to already existing published (more static generally) Linked Data provided by councils. In addition, people-centric mobile sensing, enabled by ubiquitously wearing smartphones, should progress into continous 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.

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. Smart data can now be achievable in smart cities.

1 Introduction

2 Data Life Cycle

Throughout the literature, a 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 we think that can be applied for generic data, independently of its original domain.

2.1 Data Documentation Initiative

The first model to be analysed is the model proposed by Data Documentation Initiative (DDI). The DDI introduced a Combined Life Cycle Model for data managing [1]. As Figure 1 shows, this model has eight elements or steps which can be summarized as follows, according to [2]:

- **Study concept.** At this stage, apart from choosing the research question and the methodology for collecting the data, also plans the processing and analysis of data to answer the question.
- **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 is processed to answer the proposed research question. The data may be recorded both machine-readable and human-readable form.
- **Data archiving.** Both data and metadata should be archived to ensure long-term access to them, ensuring the confidentiality.
- **Data distribution.** This stage involves the different ways which the data are distributed and questions related to terms of use of the data or citation of the original sources.
- **Data discovery.** The data may be published in different manners, through publications, web-indexes, etc.
- **Data analysis.** The data can be used by others to achieve different goals.
- **Repurposing.** The data can be used outside of their original framework, restructuring them or combining with different data.

2.2 Australian National Data Service

In late 2007 the Australian National Data Service (ANDS) was founded with the target 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 [3]:

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

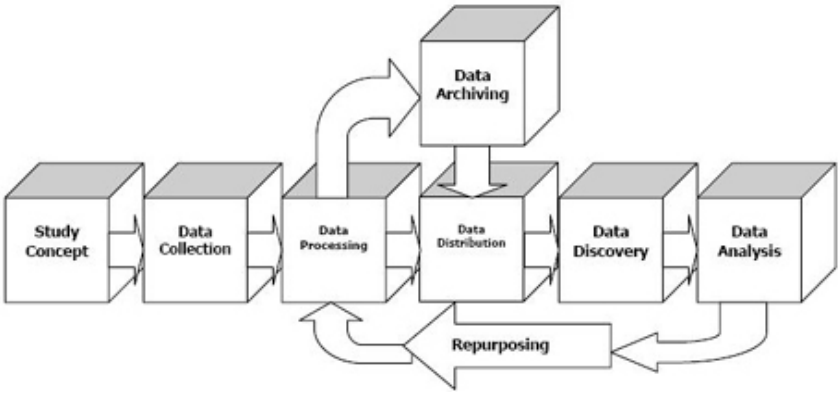

Fig. 1. Combined Life Cycle Model (ownership: DDI Alliance).

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

2.3 Ecoinformatics data life cycle

Michener and Jones define in [4] the concept of "ecoinformatics": *a framework that enables scientists to generate new knowledge through innovative tools and approaches for discovering, managing, integrating, analysing, visualizing 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 consider both manual (hand-written data sheets) and automatic (sensor networks) data-collection methods.
- **Assure.** Quality assurance and quality control (QA/QC), an issue which in previously mentioned models is not taken into account, refers to developing methods to guarantee the integrity of the data. Quality assurance also can consists of defining 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 to 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 discover as *one of the greatest challenges*, as many data are not immediately available because there are stored in individual laptops. The main challenges to publish the data in a proper way are related about 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 into a difficult task because 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, this models remarks the importance of documenting this analysis with sufficient detail to enable its reproduction in different research frameworks.

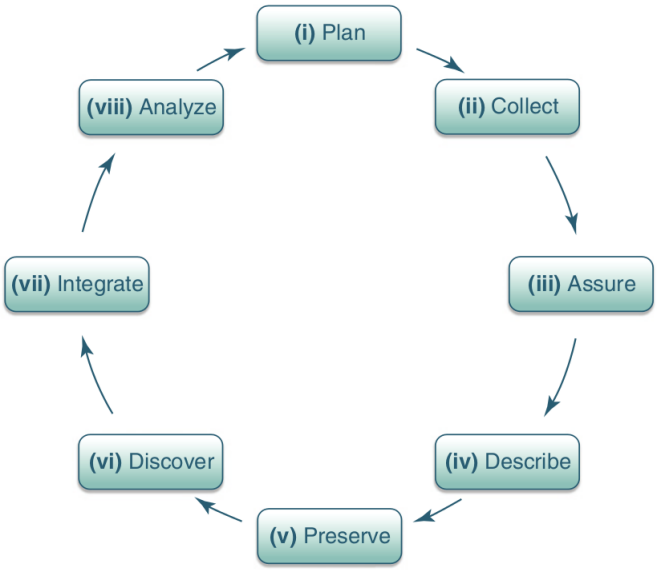

Fig. 2. Data life cycle in ecoinformatics. Taken from [4].

2.4 UK Data Archive

The last analyzed data life cycle model, is the one proposed by UK Data Archive[1]. This model is oriented to help researchers to publish their data in a manner which allows other researchers continuing their work independently. In Figure 3, we can observe the following stages:

- **Creating data.** Creating the data involves the design of the research question, planning data management and their sharing. If we want to reuse existing data, we have to locate existing data and collect these data. As well as the data is new or is an existing data, 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 the confidential data, they have to be "anonymized". The UK Data Archive recommends to create the metadata at this stage too.
- **Analysing data.** At this stage the data are interpreted and derived into visualizations or reports. In addition, the data are prepared for preservation, as can be seen at following stage.
- **Preserving data.** To preserve the 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 process are documented.
- **Giving access to data.** Once the data is stored, we have to distribute our data. This distribution of the data 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.

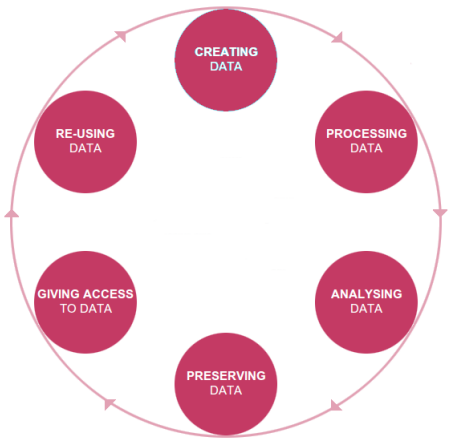

Fig. 3. Data life cycle proposed by UK Data Archive.

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

2.5 A common data life cycle for smart cities

Based on these data life cycle models, we proposed a common data life cycle for managing data into smart cities. As can be seen at Figure 4, the different stages of mentioned models have been aggregated, forming our proposed model.

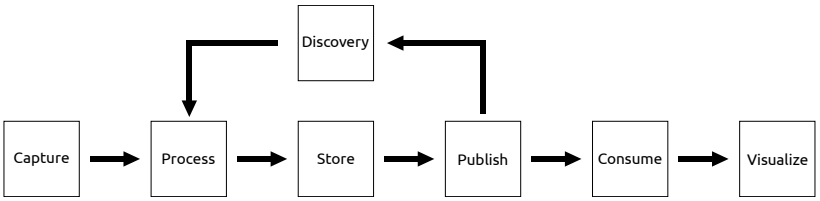

Fig. 4. Proposed model.

The different stages of this model, which are going to be explained widely in following sections, are:

- **Capture.** The first step of our model consists on capturing the data. In a smart city there are a lot of alternatives to capture data, like sensors, data published by public administration, social networks or in more traditional way like surveys.
- **Process.** Once the required data is captured, they are prepared to store, allowing proper methods to explore them. This processing involves the analyzing, refining, cleaning, formatting and transformation of the data. A this stage we also suggest to search links and relationships with other data.
- **Store.** The storage of data is, probably, the most delicate action in the life cycle. Above the storage are build all the analysis tools and is the "final endpoint" when someone requests our data. A suitable storage should have indexing, replication, distribution and backup capabilities, among other things.
- **Publish.** Most of previously mentioned models prioritize the analyzing stage over the publication stage. In our model, we propose the opposite for a very simple reason: when you consume your data before the publication of them, and through different way which the rest of the people is going to consume them, you do not emphasize enough on publishing these data correctly. Everybody has ever met a research paper or an application in which accessing the data was difficult, or, once we have the data, these data are incomprehensible. To avoid this issue, we propose to publish the data before consuming them, and consume them through the same way as the rest of the people does.
- **Consume.** Once the data is published, we use the provided methods to consume the data. This data consumption involves the data mining, analytics or reasoning.

- **Visualize.** To understand the data properly, doing suitable visualizations is essential to show correlation between data or the conclusions of the data analysis in a human-readable way.
- **Discovery.** This step is about discovering data from external sources, or about our data being discovered by others. This step enables a new execution of the entire life cycle, avoiding the capturing stage, since we are using third parties data. Otherwise, the data life of different datasets can elapse concurrently, being merged at processing stage.

3 Open Linked data as a viable approach

References

1. Initiative, D.D.: Overview of the DDI version 3.0 conceptual model (April 2008)
2. Ball, A.: Review of data management lifecycle models. (February 2012)
3. 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
4. Michener, W.H., Jones, M.B.: Ecoinformatics: supporting ecology as a data-intensive science. (2012)