International Journal of Advances in Computer Science & Its Applications
Volume 6 : Issue 1 *[ISSN 2250-3765]*
Publication Date : 18 April, 2016

Towards an approach to measure how open data is reused

Visualization of open data metrics

[Jose Vicente Carcel, Jose-Norberto Mazón,
Llorenç Vaquer, Jose Zubcoff and Irene Garrigós][1]

***Abstract*—Data economy drives us into a world in which reusing data has an important economical impact. However, data consumer has no enough mechanisms for developing apps that reuse data. A developer is willing to know which is the impact of the data to decide which is consumed within the envisioned smartphone application.**

One of the best ways of measuring impact comes from the research arena, in which metrics based on citations are provided (such as bibliometrics and altmetrics) to measure the impact of authors, articles and journals. Accordingly, to measure how open data is reused it must be cited. However, data citation is not a common practice and it is quite difficult to obtain metrics of reused data and apps. This paper describes first step of an approach to (i) extract information about used data in app catalogs with scraping and other techniques to compose a metadata database, (ii) compute some metrics to measure the impact of open data sets, and (iii) visualize results to help data consumers in the process of selecting the most impact data for reusing and develop apps.

***Keywords*—open data, bibliometrics, altmetrics, SOA**

I. Introduction

Nowadays, governments and other institutions are opening their data to society for the sake of transparency and accountability. Undoubtedly, open data has a social impact but also an economical one where data and technology work together for creating new added-value services through applications. Importantly, to improve the process of selecting adequate data, more information is required. Actually, if data consumers were concerned about the real impact of reusing one or other data sets, they could make informed decisions about which data sets would be selected for reusing to create a successful application.

Measuring impact has been profusely studied in other research areas, such as library science [1,2]. Interestingly, there are two types of science metrics that can be useful for data consumers, namely:

- Bibliometrics: they are part of scientometrics, a body of knowledge that applies mathematical and statistical methods to the entire scientific literature and authors who produce it, with the aim of studying and analyzing scientific activity. Impact of authors, papers or journals can be measured by bibliometrics. For example, impact factor measures the impact of a paper by its citations and it provides a paper magnitude which can be useful to compare papers or journals.
- Altmetrics: they are metrics based on social networks that are proposed as alternative to bibliometrics. E.g., there are altmetrics for measuring the well-known impact factor, used for scientific journals, and the citation indexes of person, as the h-index.

Our hypothesis is that these strategies for measuring scientific impact can also be used for measuring impact of open data, and how it is reused. The rationale behind it is that open data has, by definition, a license in order to state reusing conditions [3]. One of them consists on acknowledging the source of data, i.e., citing the source of the reused data. To sum up, citation of data must be a common practice since: (i) data license forces to cite source of reused data, (ii) data citation allows everybody to easily reuse, (iii) data citation enables to develop tools for measuring the impact of data and (iv) an open data structure that recognises and rewards data producers can be settled.

Therefore, data consumers should cite any data they use in their applications, projects or papers. Unfortunately, this is not carried out in the right way to be understood by machines in order to easily automate processes to measure data metrics.

In conclusion, to encourage open data reuse, open data consumers need the data counterparts of bibliometrics and altmetrics. In this way, their impact can be visualized and informed decisions can be made for selecting appropriate data to reuse and developing useful and demanding apps. A good starting point is consider the well-known metrics coming from the scientific arena. In this position paper, we present a first attempt to provide a visualization approach that shows these kinds of metrics about data.

II. Current State of (Open) Data Citation

As previously explained, citation of data promotes reusing open data among data consumers. However, data

Jose-Vicente Carcel, Jose-Norberto Mazón, Llorenç Vaquer, Jose Zubcoff[1] and Irene Garrigós
WaKe Research Group, University of Alicante
Spain

[1] This research is supported by Open.MinD project (GV/2014/098) funded by Conselleria de Educación, Cultura y Deporte de la Generalitat Valenciana.

citation is not a common practice nowadays, and one may find citations in some documents or papers but it is not readable by machines.

The five star model proposed in [4] is widely accepted to determine the quality of data regarding their potential for reusing. This model establishes five levels labeled with stars according to their fitness for reutilization:

1 star. Data are available on the Web, independently of the format (e.g., pdf file).

2 stars. Data are published on the Web in a machine-readable structured format (e.g. using Excel format instead of a scanned image of a table).

3 stars. Data are published in a non-proprietary format (e.g. CSV instead of Excel).

4 stars. Data are identified by URIs (e.g. by using W3C standards such as RDF) in such a way that they are easily and persistently accessible.

5 stars. Data are linked with other data in such a way that they are contextualized.

There are several data quality criteria that have been studied in Batini et al (2009)[5] and they can be applied to open data as well, however, there are some citation issues and other data quality criteria that must be considered with special emphasis on reutilization of data, i.e., data availability, suitability, relevance, and so on. In this sense, Oviedo et al (2013)[6] specified a set of measures to consider in an open data initiative beyond the 5-star schema.

However, citation of data will allow to measure the impact of the data, and the tracking of this measure. Without data citation, developers and stakeholders are limited to measure the impact of the data reuse.

Based on our own experience, citation of data in mobile apps [7,8] records should be one of the best practice to help data consumers in selecting open data and also to help developers to create useful tools for data consumers but, unfortunately applications do not cite data that reuse (or, at least, not in a right way).

To go further into the details, there exist two sources of applications that reuse open data:

- Applications stores like "App Store" (https://itunes.apple.com/es/genre/ios/) or "Google Play" (https://play.google.com/store). These types of app stores do not have a specific metadata fields for citing reused data and you can rarely find references to reused data in their descriptions.
- Applications catalogs in open data portals: These catalogs only contain applications which reuse open data from their own data catalogs and sometimes they have metadata fields in the app record for specifying reused data, although, normally, in wrong way for be readable directly by machine.

Bearing all the previous considerations in mind, data consumer does not have tools to obtain more information (e.g. metrics such as bibliometrics or altmetrics) about datasets to visualize the impact of reusing data and develop appropriate applications or stakeholders to measure the return of investment.

III. Case Study: Citation in Spanish Open Data Portals

At the University of Alicante we have developed the open data portal (http://datos.ua.es) and now we are trying to provide data consumers with different tools for reusing data, being one of them the aim of this paper .

After detecting this specific need (metrics about open data), we have taken Spanish open data portals and their application catalogs as a case study, focusing on open data citation in applications.

We have studied a large number of these open data portals and we have tried to answer these questions: Do they have an application catalog? Do they cite reused data? Do they have a metadata field for reused data?

Our general conclusions are: (i) there are open data portals that do not have an specific apps section because they are in a lower level of development, (ii) most portals with app section do not reference used data and do not has a specific metadata field for used data, (iii) the Spanish government open data portal (http://datos.gob.es) is the only site that has, in most applications, a metadata field for used data and it cites used data or catalog. Then, the open data portal of Spanish government is the site that has more citations of used data in applications. In addition, most of Spanish open data initiatives federate their catalogs and applications to it, although they have their own portal. That is why we can take datos.gob.es as a reference site for our purpose.

IV. An Architecture for Measuring How Open Data is Reused

We propose a visualization approach for data consumers that shows metrics about data and applications. The envisioned metrics are (i) bibliometric-like (ranking of most used data by applications and map of areas where datasets have been used), or (ii) altmetrics-like (number of mentions on tweets, number of retweets of leading users and number of likes in Facebook). These metrics can help data consumers and developers to develop an application. For example, it can help them to visualize what kind of application and/or data has more social impact and which area is more appropriate to develop an app.

In order to realize our approach, we have developed the architecture shown in Fig. 1: (i) An integration layer to extract, by scraping and other techniques, information about applications and used data from the datos.gob.es, (ii) this extracted information about apps and his used data is indexed into the metadata database, (iii) this database contains homogenized metadata such a name of app, used data, area, app type, store link, record link and (iv) there is a data model and a data access layer to communicate with the database and methods to calculate aforementioned metrics. On the top, there are services to provide metrics about data and applications for the visualization front-end app.

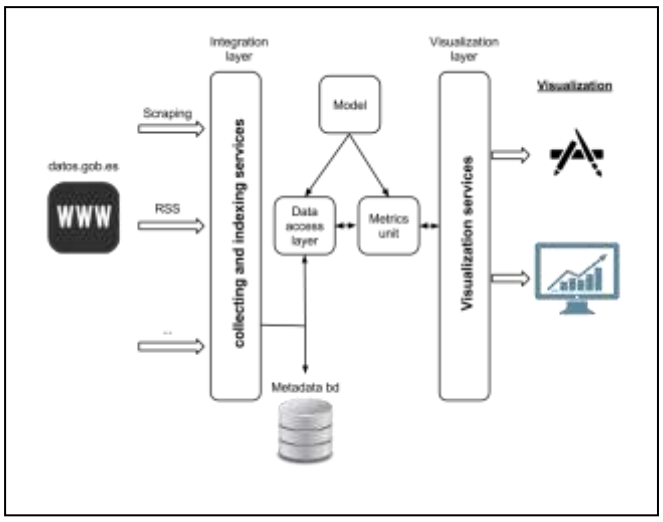

Figure 1. Architecture for open data reuse metrics and visualization.

v. Conclusions

In this work we propose an approach to use metrics for open data and applications to allow developers and stakeholder to measure the relevance of a data source or an app. We also propose an architecture for visualize open data reuse metrics that works also for applications. From our approach, users can visualize the impact of a source taking better decisions by selecting appropriate data to reuse and developing useful and demanding apps. Future work will be to provide a mechanism to collect and index open data and apps from other sources improving the use of metrics.

References

[1] R. Roemer, R. Borchardt (2012) From bibliometrics to altmetrics. College and Research Libraries News vol. 73, no 10 596-600

[2] D. Torres, A. Cabezas, E. Jiméenez (2014) Altmetrics: new indicators for scientific com- munication in Web 2.0. arXiv:1306.6595

[3] P. Miller, R. Styles, T. Health (2008) Open Data Commons, a License for Open Data. LDOW.

[4] C. Bizer, T. Heath, T. Berners-Lee (2009): Linked Data - The Story So Far. Int. J. Semantic Web Inf. Syst. 5(3): 1-22.

[5] C. Batini, C. Cappiello, C. Francalanci, A. Maurino. (2009). Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41(3).

[6] E. Oviedo, J.N. Mazón, J. Zubcoff (2013): Towards a data quality model for open data portals. CLEI 2013: 1-8.

[7] DataCite. Making your data more accesible and more useful. Retrieved October 2014 from https://www.datacite.org/services/cite-your-data.html

[8] A. Ball, M. Duke (2012). 'Data Citation and Linking'. DCC Briefing Papers. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/briefing-papers/

[Data citation is not a common practice, then, it is quite difficult to obtain metrics of reused data and apps.]