Towards Ecosystems based on Open Data as a Service

Position Paper

Kiev Santos Gama, Bernadette Farias Lóscio
Centro de Informática – Universidade Federal de Pernambuco (UFPE) Recife, Brazil
{kiev, bfl}@cin.ufpe.br

Keywords: Open Data, Software Ecosystem, Business Ecosystem, Data as a service.

Abstract: Despite several efforts in contests throughout the world that encourage local communities to develop applications based on government Open Data, the solutions resulting from such initiatives do not have longevity, lacking maintenance and rapidly falling into disuse. This is due mainly to the lack of investment or even a model for monetizing the use of such applications. Therefore, it is necessary to develop a model that fosters the value chain for Open Data aiming an economically self-sustained ecosystem. Such ecosystem should promote new businesses through the creation of systems and applications focused on citizens. This article discusses the creation of software ecosystems for services and applications underpinned by a platform based on Open Data as a Service.

1 INTRODUCTION

Recently, a global trend is being observed on the release of data access in public administration, so far restricted to government organizations that hold these data. Increasingly, public institutions are planning and implementing strategies for open government data, in order to increase transparency and efficiency of government as well as increasing citizen participation (Huijboom & Van den Broek 2011). The data in question range from high-level administrative data to city infrastructure data. While the former focuses on data such as public expenditure, tax revenues and similar information, the former consists of diverse domains such as rainfall indices; data about schools, demographics, sewage, transportation, water and electricity; public spaces geopositioning data; waste collection points, and so forth.

The United States and the United Kingdom pioneered the efforts around Open Data, releasing their portals in 2009 (Hogge 2010). Many countries followed the global trend on Open Data, and nowadays, Open Data portals can be found in many places. The main idea behind Open Data consists of promoting the transparency and the empowerment of the citizens through the free access to government data. Though, besides of promoting transparency and enabling the empowerment of citizens, Open Data may also enable the creation of significant economic value. Recently, a research developed by the McKinsey Global Institute shows that Open Data could help to generate more than $3 trillion in value every year in different domains of the global economy (McKinsey 2013).

Townsend (2013) points out competitions as a way to stimulate the usage of Open Data in various cities of many countries – USA (Washington D.C., New York, San Francisco, Portland), Canada (Edmonton), Netherlands (Amsterdam), Ireland (Dublin), to cite a few. The format of these contests usually ranges from medium duration (a few months) to short duration (*hackathons*, that only last a few days). Most applications in such scenarios end up being quickly abandoned. In the long run, very few applications resulting from such contests are actually scalable and sustainable (in an economical perspective).

As discussed in (Goldstein & Dyson 2013), just providing an Open Data access is not enough to unlock significant amounts of economic value. In order to achieve its full economic potential, an Open Data initiative must be part of a whole ecosystem, which is mainly composed by Open Data providers and Open Data consumers that collaborate to promote the initiative. An Open Data ecosystem should foster new businesses through the creation of systems and applications focused on improving productivity of existing companies or government

institutions as well as improving the well being of citizens. Equally important, an IT-based platform must also be provided to promote the interactions between the actors of the Open Data ecosystem. Furthermore, this platform should enable the creation in large scale of innovative products and services through the use of standard formats and common tools.

The effective creation of an Open Data ecosystem along with an IT platform to enable the communication between data providers and data consumers are fundamental issues for the success of an Open Data initiative. In this paper, we are interested on those issues and we propose a promising approach that can facilitate the provisioning of Open Data, allowing a flexible and easy way to reuse it. In a world where the IT industry started to design Cloud Computing systems that deliver everything as a service (Banerjee et al 2011), this article specifically discusses the creation of ecosystems underpinned by a platform based on the concept of Open Data as a Service. We also argue that the use of such solution contributes significantly to the generation of the expected economic value of the Open Data. The remainder of this paper is organized as follows: Section 2 provides some background on Open Data and its challenges, followed by a brief overview on data as a service and business ecosystems; Section 3 introduces our perspective concerning the Open Data ecosystem; Section 4 presents our view of a platform for Open Data as a Service; and, finally, Section 5 concludes this paper.

2 BACKGROUND

2.1 Open Data

Nowadays, Open Data portals can be found in most developed countries (e.g, United Kingdom, France, Germany) as well as in developing economies (e.g., Brazil, Russia, India). Open Data is typically presented through raw bulks of data, in machine-readable format, that can be used by citizens and institutions as they wish (e.g., development of systems and applications, data analysis using software tools, etc).

Data consolidation and creation of these portals that provide masses of raw data already follow precepts dictated by the open government initiatives. An open format is considered independent of data technology platforms; it is based on open standards and tools, and made available to the public without restrictions (Dietriech et al 2009). The information portals on the Internet providing that data, often make it available through files of various formats (XML, CSV, JSON), sometimes accompanied by metadata describing them.

This strategy of publishing data enables third parties to build applications that use these data sources to correlate them, analyze them and extract information relevant to the objective of each use. However, access to such data sources is typically done manually. Those who wish to use such data need to access the portal, copy the bulks of data, then extract what is relevant and finally integrate the data into their applications which must access local databases with full or partial copies of the acquired Open Data.

2.2 Open Data Challenges

Despite of the several Open Data portals that have been released in the last few years, substantial challenges need to be addressed when designing and planning a new Open Data initiative. Identification of relevant data, integration of distributed data, the lack of metadata and standards, and the quality of the released data are good examples of challenges to be faced when starting an Open Data initiative. Moreover, once the project is running, it will be necessary to deal with new issues, including, for example, stale data as well as problems related with the scalability, availability and reliability of the adopted Open Data solutions. Most of the existing Open Data initiatives use solutions like CKAN, Socrata and Data Hub, which focus basically on providing platforms for just storing and providing access to Open Data. In general, those solutions may be classified as data-centric catalogues and do not offer appropriate tools to address the previously mentioned limitations.

According to the McKinsey report (McKinsey 2013) some high level key issues have to be addressed in order to realize the full Open Data value potential, including: i) Investment in technology to provide suitable tools for collecting and sharing data, and to perform data analyses to uncover important insights; ii) Development of standards to facilitate the integration of data distributed in multiple data sources; iii) Releasing of metadata to make Open Data more usable; and iv) Creation of an Open Data marketplace to provide clear channels for sharing data and to build a community with group norms and rules.

In this context, new Open Data solutions are required to deal with the limitations of existing approaches and to help to unlock the expected Open Data economic value. Considering the increasing

interest of the IT industry on service-centric approaches and the design of Cloud Computing systems that deliver everything as a service (Banerjee et al 2011), we argue that those solutions should deliver Open Data as a Service on top of a Cloud Computing infrastructure. Specifically, an Open Data solution must offer:

- Services that can be easily combined for constructing applications, including data publication services and data analysis services;
- An infrastructure that can provide scalability, availability and reliability on the available data services;
- APIs for easily allowing to crowdsense data that can replace or complement stale Open Data;
- Transparently linking and integration of data from different sources.

2.3 Data as a Service

Software services are a promising approach that can facilitate the provisioning of Open Data, allowing a flexible and easy way to reuse it. Service-Oriented Computing (SOC) (Papazoglou 2003) offers an approach to build applications through services as building blocks. Services consist of self-descriptive elements that are platform-independent and hosted in a distributed environment. A service could consist in any type of functionality, whether simple or complex. Service-oriented architectures (SOA) act as a logical model to represent how services should be delivered and consumed through service contracts. The applications delivered as services over the Internet have long been referred to as Software as a Service (SaaS), while the datacenter hardware and software delivering such services is what is called a Cloud (Amburst et al 2010).

Although the Open Data handbook (Dietriech et al 2009) suggests the use of software services to provide access to data only in cases where they change frequently, using this approach to also access data that do not necessarily change in a regular basis would provide several advantages. Besides the ability to reuse existing functionality, the construction of applications would be facilitated by the usage of services. There would be no need to copy the entire mass of relevant data (e.g. mappings of schools, streets, census data). The typical scenario for third party application is using only a small fraction of that data, which would be available through services.

The Web has evolved to become a programmable Web, where a new generation of applications is based on *mashups* (Yu & Woodard 2009). They consist on applications that are created by integrating data from different sources. The availability of data as service is one of the enablers for an open programmable Web (Maximilien, Ranabahu & Gomadam 2008). Others have discussed the approach to provide data as services. The concept of Information-as-a-Service (IaaS) discussed by Dan and colleagues (Dan et al., 2007), is referred by Truong and Dustdar (Truong and Dustdar, 2009) as Data as a Service (DaaS). Typically, it is available in the form of infrastructure that extracts data from other sources and provide services in "slices" of smaller data, under a specific context (e.g., stock market data). There are basically two categories of service delivery in that case: (1) access to data based on other data sources, and (2) access to data held in own infrastructure provided, allowing, in addition to the consultations, the addition, deletion and modification of information.

2.3 Business ecosystems

The concept of business ecosystem was introduced by James F. Moore in the early 1990s (Moore, 1993). Moore defines a business ecosystem as: "*An economic community supported by a foundation of interacting organizations and individuals: the organisms of the business world. This economic community produces goods and services of value to customers, who are themselves members of the ecosystem. The member organizations also include suppliers, lead producers, competitors, and other stakeholders. Over time, they coevolve their capabilities and roles, and tend to align themselves with the directions set by one or more central companies"*

Two important issues of this definition are: the collaborative nature of a business ecosystem and the diversity of actors participating in the ecosystem. The first one is related with the idea that member organisations or parts of the ecosystem evolve in alignment, instead of the traditional view where companies are part of a single industry and try to evolve individually. Moore (1993) argues that in a business ecosystem, companies co-evolve capabilities around a new innovation: they work cooperatively and competitively to support new products, satisfy costumer needs, and eventually incorporate the next round of innovations. Other important issue is that a business ecosystem should include not only partners and subcontractors but also complementors, competitors, customers, and

potential collaborator companies, as well as public bodies, local incubators, investors, and even research institutes and universities (Moore 1993).

Based on those perspectives, in the next section we present our definition for an Open Data ecosystem.

3 OPEN DATA ECOSYSTEM

Recently, Open Data ecosystems have been the subject of a lot of discussions. The McKinsey report on Open Data, for example, highlights the need of a vibrant ecosystem in order to transform Open Data into valuable tools (McKinsey 2009). However, to the best of our knowledge, the concept has not been formally defined. In this section, we present some of our preliminary ideas concerning the definition of an Open Data ecosystem. In our proposal, we consider the two issues discussed in the previous section: the collaborative nature and the diversity of actors participating in a business ecosystem. We also argue that in order to promote the effective collaboration among the actors and to allow the quickly development of new products and services an IT-based platform is required.

Figure 1 shows our perspective for an Open Data ecosystem. The main actors of our model are: government, application developers, small and medium enterprises (SMEs), startups, civil society, universities, funding agencies and investors. Each one of them would play at least one of the following roles: *data provider*, *data consumer, data aggregator* and *data sponsor*.

The government interacts with all the other actors and it can play different roles. Its main role is as a *data provider*, i.e., it is responsible for providing free access to the governmental data. On the other hand, the government is also a consumer of solutions developed by companies or single application developers. Finally, the government promotes the Open Data initiative by providing communication channels between the other actors as well as encouraging the development of Open Data products. Universities are mainly responsible for promoting the innovation and the development of specialized knowledge. They act like *aggregators* bringing other actors together to form a shared understanding and to promote the development of standard formats and platforms. The civil society, together with the government, is mainly responsible for identifying new demands for Open Data access, acting like an *indirect data provider*. Moreover, the civil society is also a *data consumer*, once it consumes data and information produced by market solutions. Application developers, startups and SMEs compose the market, whose main role is as a *data consumer*. The market is the responsible for the development of Open Data based products and, similar to universities, it should also promote the innovation and the development of Open Data solutions in large scale. Funding agencies and investors can be seen as Open Data *sponsors*. They are responsible for promoting the Open Data initiative through both public funding programs and private investments.

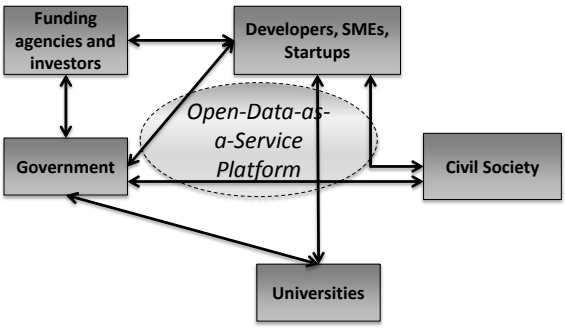

Figure 1. Open Data ecosystem

Other fundamental component of our proposal is the Open-Data-as-Service platform, which allows the interactions among a subset of actors in the ecosystem. In our context, this platform allows them to play their different roles as data providers, data consumers and data aggregators, by means of services and applications built on top of those services. The platform would promote the integration of these different actors, allowing the development of products and services based on Open Data that can provide practical benefits to users (e.g., civic apps). Therefore, a marketplace for services and applications can be built around the platform, allowing monetizing the solutions and thus generating economic value from Open Data.

Considering the important role that the software platform plays in the ecosystem, our proposal can be considered as a Software ecosystem, which is a special type of business ecosystems (Jansen & Cusumano 2013). In a Software ecosystem, the relationships among actors *"are frequently underpinned by a common technological platform or market and operate through the exchange of information, resources and artifacts"* (Jansen et al 2009).

4 A PLATFORM FOR OPEN DATA AS A SERVICE

Currently, no data-as-a-service approach focuses on providing Open Data. From a service-oriented perspective, what one may find is limited to Open Data portals following the recommendation from the World Wide Web Consortium (W3C) (Bennet & Harvey 2009), that suggests exposing Open Data through application programming interfaces (API) based on Web Services (REST or SOAP). In our perspective, Open Data should not only be exposed as services, but all building blocks of the platform can be built based on services. The platform would provide core services and allow developers to consume such services for: integrating Open Data providers with the platform; building their own services (e.g. aggregating value based on Open Data) for offering them in the same platform; constructing applications using Open Data services. Figure 2 illustrates the different service layers that comprise our vision of such platform, as detailed next in bottom-up order:

Virtualization Services: This is the basic Cloud Computing infrastructure that will provide virtualized services such as elastic storage, thus ensuring availability, reliability and scalability;

Linked Open Data Publication Services: This service layer is responsible for services involving mediation and integration of data, as well as services for storing ontologies (i.e., ontologies catalogs), metadata, matching and merging ontologies, data transformation (e.g., XML to RDF) and data publication. Platform users (developers) can plug their own services in this layer for extracting data from their data sources, which can either be data originated from conventional databases or crowdsensed information (e.g., obtained from end users through the usage of Apps);

***Query Services*:** this category of service is responsible for providing transparent access to distinct data sources, with services responsible for query rewriting (e.g.: REST to SPARQL) and query routing.

Analysis and Visualization Services: they offer support for discovering useful information through the use of services that encompass traditional data analysis techniques like data mining and statistical techniques, as well as visualization techniques like line graphs, pie and bar charts, interactive maps and infographics. These services are responsible of turning raw data into something useful. The idea is to provide high-level services that allow the most inexperienced users to quickly start performing sophisticated analysis and produce great visualizations from many different types of data.

Service Registry: As illustrated in Figure 2, this layer communicates with all other layers, being the central registry that maps the services offered in the platform. This is the key point of interaction with service and application developers. In this layer, a catalog service can group service descriptions by category and allows for searching and viewing what services are currently available. During development time, a user may query the platform to select one or more desired services and sees their description and details, or he can develop solutions (services or applications) that programmatically select the best available services according to specified criteria (e.g., Quality of Service) at runtime.

***Applications*:** They are built through mashups or compositions of different services provided by the platform. Applications use Open Data with different purposes, including: to promote the transparency, to improve accountability, to help people dealing with mobility and transportation problems and to encourage citizens to get involved in civic and community activities.

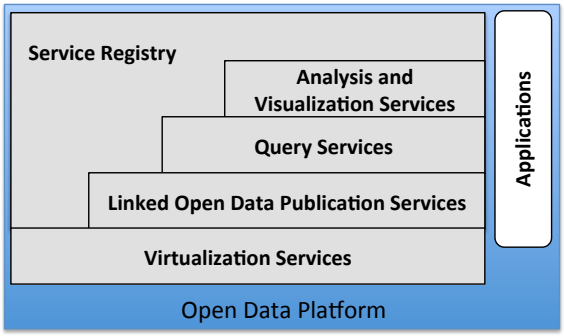

Figure 2. Different service layers of an Open Data platform

This platform can be offered as a public cloud that could be partially maintained by government and funding agencies, to foster the value chain around Open Data. It would rely on the core services provided with the platform and also allow its extension through the publication of services from third parties (i.e., an ecosystem of solutions), which would enable the creation of a marketplace and strengthen the sense of community around Open Data solutions. The proposed platform addresses the challenges enumerated in section 2.2, by providing services for collecting and sharing data, and to perform data analysis. The platform would also rely both on standards for integrating different data sources and on metadata for facilitating data publication. As stated, marketplace around services and applications can be established. Custom added-value services developed by SMEs or startups can charge for access, as well as applications for niche

markets. For instance, a developer can use consume two distinct Open Data services offered in the platform: one providing data from public transportation, and another one providing real time traffic data. These services can be composed and with additional functionality the developer can provide an optimized routing service from which he can charge and monetize from. In another scenario, a government agency may want to use a data transformation service to upload raw XML data and convert it into RDF according to the ontology used by the platform.

5 CONCLUSIONS

Nowadays many countries are following a global trend where public institutions are planning and implementing strategies for open government data. As a way to encourage the use of such data, local competitions for developing applications are being organized, but the resulting applications are quickly abandoned and, in the long run, very few of them can monetize. Such evidence supports the argument that just providing access to data is not enough to unlock the potential of extracting significant economic value from the usage of Open Data. In order to achieve such goal, many challenges have to be addressed, such as investing in technology to deal with Open Data, and creating a marketplace for sharing data. Since the IT industry is evolving to systems that deliver everything as a service, we support the idea that Open Data can take advantage of such approach. Services can help to address such challenges and be used as the building blocks on the construction of a platform for Open Data as a Service. The platform would underpin an ecosystem that involves many actors interacting and playing different roles. These interactions would be done through data services and mashups of such services that would generate other services (e.g., service compositions, added-value services) and end-user applications that could be charged for their use, thus fostering the generation of business around Open Data.

REFERENCES

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M., 2010. A view of cloud computing. Communications of the ACM, 53(4), 50-58.

Banerjee, P., Friedrich, R., Bash, C., Goldsack, P., Huberman, B. A., Manley, J. & Veitch, A., 2011. Everything as a service: Powering the new information economy. Computer, 44(3), 36-43.

Bennett, D., & Harvey, A., 2009. Publishing open government data. W3C Working Draft. Available from: [http://www.w3.org/TR/gov-data].

Dan, A., Johnson, R., & Arsanjani, A., 2007. Information as a service: Modeling and realization. In Systems Development in SOA Environments, 2007. SDSOA'07: ICSE Workshops 2007. International Workshop on (pp. 2-2). IEEE.

Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, P., Tait, J., & Zijlstra, T., 2009. Open Data handbook.. Available from: <http://opendatahandbook.org>. [18 January 2014]

Goldstein, B. & Dyson, L., 2013. Beyond Transparency: : Open Data and the Future of Civic Innovation, Code For America Press.

Hogge, B., 2010. Open Data study. a report commissioned by the Transparency and Accountability Initiative, http://www.transparency-initiative.org/wp-content/uploads/2011/05/open_data_study_final1.pdf

Huijboom, N., & Van den Broek, T., 2011. Open Data: an international comparison of strategies, European Journal of ePractice, No. 12, March/April 2011.

Jansen, S., Finkelstein, A., & Brinkkemper, S., 2009, May). A sense of community: A research agenda for software ecosystems. In Software Engineering-Companion Volume, 2009. ICSE-Companion 2009. 31st International Conference on (pp. 187-190). IEEE.

Jansen, S., & Cusumano, M. A., 2013. Defining software ecosystems: a survey of software platforms and business network governance. Software Ecosystems: Analyzing and Managing Business Networks in the Software Industry, 13.

Maximilien, E. M., Ranabahu, A., & Gomadam, K., 2008. An online platform for web apis and service mashups. Internet Computing, IEEE, 12(5), 32-43.

McKinsey Global Institute, 2013. Open Data: Unlocking innovation and performance with liquid information. Available from: <http://www.mckinsey.com>. [17 January 2014]

Moore, J. F., 1993. Predators and prey: a new ecology of competition. Harvard business review, 71(3), 75-86.

Papazoglou, M. P., 2003 Service-Oriented Computing: Concepts, Characteristics and Directions, 4th International Conference on Web Information Systems Engineering (WISE'03), Rome, Italy, 2003.

Townsend, A., 2013. Smart cities: Big data, civic hackers, and the quest for a new utopia. WW Norton & Company 2013 p. 199-212.

Truong, H.L., & Dustdar, S., 2009. On analyzing and specifying concerns for data as a service. Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific. IEEE, 2009.

Yu, S., & Woodard, C. J., 2009. Innovation in the programmable web: Characterizing the mashup ecosystem. In Service-Oriented Computing–ICSOC 2008 Workshops (pp. 136-147). Springer Berlin Heidelberg.