2012 IEEE 36th International Conference on Computer Software and Applications Workshops

Identification and Utilization of Components for a linked Open Data Platform

Evanela Lapi, Nikolay Tcholtchev, Louay Bassbouss,
Florian Marienfeld, and Ina Schieferdecker
Fraunhofer FOKUS
Berlin, Germany
firstname.lastname@fokus.fraunhofer.de

***Abstract*—Open Data (OD) is an emerging trend that aims to facilitate the freedom and reuse of information. Therefore, tools, applications and platforms are required that enable the publishing and consumption of data. In this paper, we present our experience from the integration of components that should constitute an OD platform. The proposed solution is able to store datasets as linked data, to catalogue the datasets, and provides a portal with features for supporting community activities. Further, we exemplify the utilization of the proposed platform, by describing a touristic city guide web mashup that uses published Open Data. This application consumes data in machine readable format over APIs provided by the OD platform.**

I. Introduction

The idea of Open Data is increasingly gaining importance. Open Data can be seen as closely related to the notion of Open Source Software that allows for the free distribution and reuse of software artifacts and corresponding code. An insightful definition of Open Data is provided by the Open Knowledge Foundation which defines it as *"A piece of content or data is open if anyone is free to use, reuse, and redistribute it subject only, at most, to the requirement to attribute and share-alike."* [1].

In order to utilize this momentum, tools and platforms are required to facilitate the economical and social benefits of Open Data. That is, a platform is required that acts as a logically centralized point for publishing and sharing data, thereby announcing it as Open Data. This platform would enable the smooth access to information for the community. That aspect would provide ground for a breakthrough in the relationship between citizens and business on one hand, and the public agencies on the other hand [2]. It is expected that citizens get involved in the Open Data publishing process by requesting, voting, rating, and commenting datasets within the emerging platform. Furthermore, such a platform enables the development of innovative services and applications (e.g. mobile apps) that have the potential to improve citizens' life.

In this paper, we present our considerations and experience on the components that should constitute such a platform. To maximize the value of Open Data, the proposed solution is able to: 1) store datasets as linked data, 2) store the belonging metadata in a dedicated data catalogue, and 3) provide an easily extensible data portal with features for supporting community activities. Further, we exemplify the utilization of the proposed platform, by describing a touristic city guide web mashup that uses Open Data hosted on the platform.

The rest of this paper is organized as follows: Section 2 gives an overview of the platform. Section 3 presents the data catalogue component. The following section 4 focuses on the need for linked data store. Section 5 describes the utilization of the proposed solution. Finally, section 6 concludes this paper and outlines future research directions.

II. The Open Data Platform

Social media phenomena emerging in the course of Web 2.0 are driving Open Data around the world. The main goal of Open Data initiatives is to make public sector data and corresponding resources available by the *Open Data Principles* [3]. These principles facilitate easy access to data that encourages users to reuse, mashup, and share available information through innovative applications and services.

The Open Data (OD) platform is an integrated solution for public agencies and private institutions that want to publish data, which is stored in a non-proprietary format and should be directly accessible via open protocols. That way the data can be easily reused and consumed by citizens and business. The primary benefit brought by the OD platform is that it offers a point of consolidation, minimizing the efforts of developers and consumers for searching, accessing, downloading and sharing datasets. Consumers have a consolidated view on all of the datasets that have been catalogued, and can navigate, identify, access and use data of interest. In addition, the OD platform is not only restricted to handling metadata, but it also allows for storing and keeping datasets which in turn can be used for applications and services. Open Data can be stored within the proposed platform in a linked format (e.g. RDF). That way the quality of the data and its usability for application and services is maximized given the powerful expressiveness of linked data [4]. Formats such as RDF and OWL can be used to store data and to describe relationships between datasets. Developers can consume data over an open API and hook to the linked data storing component in order to access the belonging linked data resources.

Figure 1 illustrates the central role of the OD platform within the whole Open Data life cycle, i.e., Identify-Publish-Discover-Enrich-Consume. It starts with data owners (e.g. civil servants) identifying and preparing relevant raw data in

978-0-7695-4758-9/12 $26.00 © 2012 IEEE
DOI 10.1109/COMPSACW.2012.30

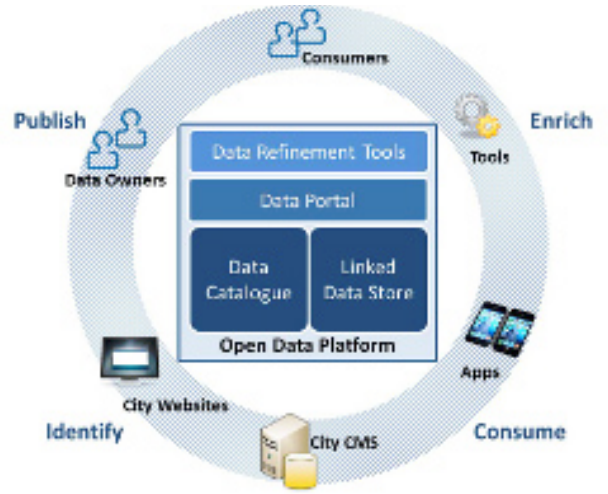

Fig. 1. An overview of the OD process and OD platform components

various non-proprietary formats following the corresponding internal processes. In the next step, the selected datasets are published. Hence, the corresponding metadata is stored within the open data catalogue. The metadata includes aspects such as description of the data, its purpose, maintainer, licensing, etc.

Once the datasets are published, citizens and business can discover them by using portal features such as searching, filtering and RSS feed notifications. Data Enrichment represents an optional task that is meant to increase the quality and correspondingly the potential for datasets to be utilized in an application. To enable machine-readability as well as seamless processing and data interpretation, datasets can be transformed into semantically richer formats, which explicitly express the context and the relation between data.

Finally, data consumption closes the life cycle of the Open Data process. On one hand citizens, businesses and civil society may use apps via Web and mobile devices. On the other hand, they also generate new data, which is further published, discovered, enriched and consumed.

III. Open Data Catalogue: Handling Metadata

The key idea of an open data catalog is to provide data seekers/consumers with a "one-stop-shop" user experience. In a typical scenario, public sector information is widely dispersed over a range of web sites of governmental entities. Thus, the main purpose of a data catalog is to gather all metadata at a central place. The metadata mainly includes attributes such as the dataset's name, description and the URL of the actual resources i.e., files or service end points. The second important function is complementary to the gathering of information. It is to enable the federation of different data catalogues, such as catalogues operating on different administration levels, or topic specific ones.

Following requirements for the catalogue software result from these use cases: 1) metadata should be stored and handled based on a well defined syntax and semantics, i.e. a documented schema, 2) the catalogue software must offer both a user interface and a widely accepted application programming interface for access by other software like applications and data portals.

In terms of metadata semantics, the most important initiative that a data catalog should accommodate is the DCat vocabulary [5]. It is a set of concepts for data catalogs and was declared W3C Editor's Draft on February 9th, 2012 [6]. DCat is based on the Dublin Core Metadata Initiative [7], which addresses general electronic resources.

Given the lack of a standard off-the-shelf data catalogue software, public administration and private institutions were utilizing general purpose solutions such as Content Management Systems for publishing their data.

However, the need for an established and stable data catalogue software drove the development of an out-of-the-box open source solution which can be easily customized. The Comprehensive Knowledge Archive Network (CKAN) [8] is the most popular existing solution and is developed by the Open Knowledge Foundation. CKAN fulfills the above mentioned requirements. In terms of interfaces it offers a web access and a well documented REST API based on JSON (JavaScript Object Notation). CKAN's approach to a metadata schema is ambivalent. On one hand, it is clearly inspired by the DCat vocabulary. On the other hand there is no official mapping between DCat and CKAN's core schema. Additionally, CKAN users are invited to amend arbitrary fields within the so-called CKAN "extra" entry. This allows for divergence from Linked Data standards.

IV. Linked Data Store

The management and storage of metadata enables the Open Data platform to only point to data resources, i.e. to integrate them as external links in the overall picture. The multi-data store allows for managing datasets and corresponding resources in the emerging platform itself. This new component should be realized by a powerful software product that allows for storing large amounts of data and scales correspondingly when it comes to accessing and manipulating this data. Furthermore, the component should be realized by a software product that offers standardized or established interfaces. The self-evident solution is to employ a type of database software that is designed and optimized for handling large amounts of data. However, a pure database solution is not optimal given the small, but troublesome, differences in the implementation of standards throughout the database products. Even though this issue can be programmatically relieved by the usage of persistency frameworks, it still remains an obstacle when it comes to achieving a maximum degree of openness trough the multi-data store component.

During the past years, the concept of linked data and the potential it brings for setting up relations among data entities, datasets, and arbitrary objects, has gained on importance within the community. A format of significant importance in that area is the well known and established W3C standard

RDF (Resource Description Framework). It is an XML based format that can be queried based on SPARQL queries, and processed with different tools ranging from "primitive" XML parsers (DOM, SAX) to sophisticated RDF parsing frameworks (e.g. Jena [9]) and tools. That way, we aim at achieving a maximum degree of machine readability and processability of the data within the platform. For the technical realization of the component, a linked database is required. The focus was set on Virtuoso [10] [11] for several reasons. First, Virtuoso makes it possible to store RDF descriptions (graphs) which can be queried using SPARQL. Thereby, Virtuoso provides the corresponding APIs for issuing SPARQL queries on RDF graphs. Furthermore, the graphs as a whole are made accessible over the WebDAV protocol. The data from these graphs can also be consumed in a JSON format. JSON is widely spread within the community of web application developers, since it is quite easy to process by standard libraries and still provides a structured view on the data. Moreover, Virtuoso can also act as a relational or XML database, which allows to switch to such technologies if required in the future. Hence, Virtuoso has the potential to be a sustainable long-standing solution for the emerging OD platform.

V. *OpenCityGuide* Application

OpenCityGuide is a cross-platform mobile application built on top of the OD platforms of different cities which offer touristic data as linked open data. The application accesses the touristic data located at the different open data platforms using SPARQL queries over the Virtuoso interface (see section IV) and aggregates them in a single application. The *OpenCityGuide* consists of several components on the backend and frontend as depicted in figure 2.

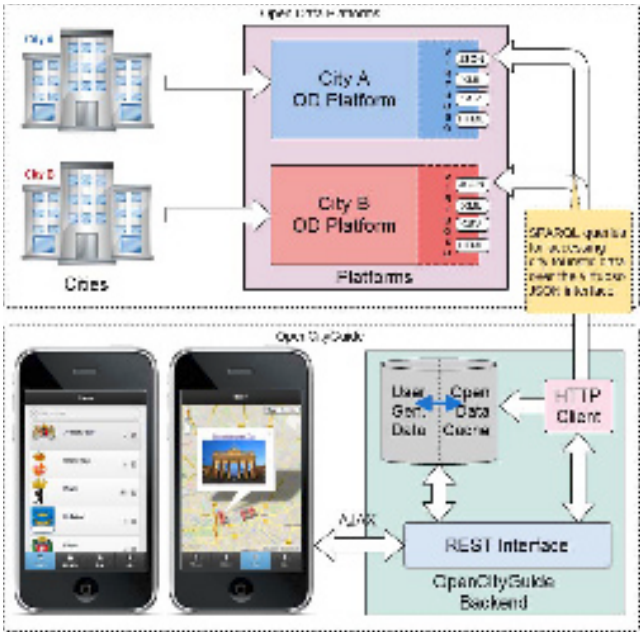

Fig. 2. OpenCityGuide Application

The backend includes a database where all data generated by end-users such as comments, ratings, checkins, etc. are stored. Furthermore, the backend provides a cache for the touristic data of each city. This cache is always synchronized with the connected open data platforms and will be updated either on-demand or periodically using a HTTP client that accesses the touristic data over the JSON REST interface provided by Virtuoso. The main advantage of this cache is that it makes the application faster on one hand and reduces the network traffic between the backend and the open data platforms on the other hand. Regarding the database, we decided to use the document oriented database mongoDB [12] such that there is no need to convert the JSON data returned by Virtuoso into other formats. The frontend of the OpenCityGuide is available as mobile web application optimized for most smartphone browsers as well as a native application for major mobile platforms such as iOS, Android and Windows Mobile. The mobile applications (web and native) interact with the REST interface on the backend side using AJAX technology and JSON as a common data exchange format. The main frameworks selected for the implementation of the mobile application are *PhoneGap* [13], as a cross-platform application development toolkit, and *jQuery Mobile* [14] as a mobile web application framework. The reasoning behind these decisions is that a combination of both technologies offers a broad cross-platform coverage of the application with little adaptation efforts. The most important aspects are: 1) *jQuery Mobile*'s progressive enhancement approach, which ensures support for the largest number of devices, 2) *PhoneGap* integrates well with *jQuery Mobile* and offers access to native functionality where it is not supported out of the box, 3) *jQuery Mobile* uses standard web programming techniques, which means that most concepts are already familiar to web developers and no extensive APIs have to be learned. Back to the design of the application, the first view of the *OpenCityGuide* App lists all involved cities. The user can select a city just by clicking on the corresponding item in the list. Furthermore, the application is able to detect a city automatically by accessing the GPS module in case the user allows the application to access his location. In the next view the application shows all touristic attractions for the selected city either on a map or as a list. The application shows more details if the user selects a point of interest on the map or from the list. In the details view the user is able to do the following actions: 1) add new comment or show comments from other users related to the selected place, 2) add new rating or show the average rating of the selected place, 3) checkin in the selected place. The last feature allows users to document and manage all visited places. For future releases of the application we are planning to add more features such as offline capability, social media integration and personalized recommendation.

VI. Conclusion

Open Data is a key concept for providing transparency with respect to public sector information. In order to facilitate that concept, new tools and platforms are required that cope with

the emerging technical challenges. In that line of thought, the need for a platform dedicated to publishing and sharing Open Data was identified. Correspondingly, the required building blocks were specified and an analysis of freely available software for implementing these components was conducted. This allows for specification and design of an early prototype based on the integration of these components. This process indicates that the proposed platform solution is indeed realizable, and allows to exemplify the utilization of the platform through a mobile application for touristic data.

In the future, we will mainly focus on the full scale implementation of the platform in the course of EU "Open Cities" project. This implementation will include the development of a metadata schema based on a survey in the scope of the project and considering existing standards in this domain. We implement the platform to support multi-language features. That way, it can applied for the participating cities in the "Open Cities" project. This would allow for integrating different datasets, developing innovative applications based on the emerging platform, and executing different case studies. Finally, on the level of concepts, the plan is to investigate the explicit use of existing vocabularies and ontologies over a dedicated repository.

REFERENCES

[1] "Open Data Definition," http://opendefinition.org/, 2012, accessed 15 Feb. 2012.

[2] D. Lathrop and L. Ruma, *Open Government: Collaboration, Transparency, and Participation in Practice*, 1st ed. O'Reilly Media, Inc., 2010.

[3] "Ten Principles for Opening Up Government Information," http://sunlightfoundation.com/policy/documents/ten-open-data-principles/, 2012, accessed 15 Feb. 2012.

[4] J. Höchtl and P. Reichstädter, "Linked open data: a means for public sector information management," in *Proceedings of the Second international conference on Electronic government and the information systems perspective*, ser. EGOVIS'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 330–343. [Online]. Available: http://dl.acm.org/citation.cfm?id=2033665.2033700

[5] F. Maali, R. Cyganiak, and V. Peristeras, "Enabling interoperability of government data catalogues," in *EGOV*, ser. Lecture Notes in Computer Science, M. Wimmer, J.-L. Chappelet, M. Janssen, and H. J. Scholl, Eds., vol. 6228. Springer, 2010, pp. 339–350.

[6] "DCAT," http://dvcs.w3.org/hg/gld/raw-file/default/dcat/index.html, 2012, accessed 15 Feb. 2012.

[7] "Dublin Core," http://dublincore.org, 2012, accessed 15 Feb. 2012.

[8] "CKAN," http://ckan.org/, 2012, accessed 15 Feb. 2012.

[9] B. McBride, "Jena: a semantic web toolkit," *Internet Computing, IEEE*, vol. 6, no. 6, pp. 55 – 59, nov/dec 2002.

[10] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A Nucleus for a Web of Open Data," in *The Semantic Web*, ser. Lecture Notes in Computer Science, K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudr-Mauroux, Eds., vol. 4825. Springer Berlin / Heidelberg, 2007, pp. 722–735.

[11] O. Erling and I. Mikhailov, "RDF Support in the Virtuoso DBMS," in *Networked Knowledge - Networked Media*, ser. Studies in Computational Intelligence, T. Pellegrini, S. Auer, K. Tochtermann, and S. Schaffert, Eds., vol. 221. Springer Berlin / Heidelberg, 2009, pp. 7–24.

[12] "MongoDB," http://www.mongodb.org/, 2012, accessed 15 Feb. 2012.

[13] "PhoneGap," http://phonegap.com/, 2012, accessed 15 Feb. 2012.

[14] "jQuery Mobile," http://jquerymobile.com, 2012, accessed 15 Feb. 2012.