Improving the usability of Open Data portals from a business process perspective

Claudio Di Ciccio, Javier D. Fernández, Jürgen Umbrich
Vienna University of Economics and Business, Vienna, Austria

Open Data portals are considered to be the cornerstones of the Open Data movement, as they offer an infrastructure to publish, share and consume public information. From a business perspective, such portals can be seen as a non-profit data marketplace, in which users try to satisfy their demand and offer requirements in several different processes. In this work, we argue that studying these so far unexplored interaction processes bears the potential to make the portals more effective. We first outline a research roadmap to better understand the behaviour of consumers and publishers by mining the interaction logs of Open Data portals. Then, we discuss potential services on the basis of these outcomes, which can be integrated in current portals to optimize the interaction, improve data quality and user experience.

1 Motivation

It is commonly agreed, that the Open Data (OD) agenda has the potential to enhance transparency in public administrations and can have a significant contribution to economic growth and productivity (Manyika et al., 2013). Core building blocks in the OD landscape are the Open Data (OD) portals, which serve as connection hubs or "non-profit" data marketplaces to share and consume information. These portals are powered by data management systems such as the CKAN open source software[1] or the Socrata system[2]. Such frameworks continuously develop and provide new technical solutions to improve the user experience, e.g., offering plugins to directly visualise information, or serving APIs to query or embed datasets on third party websites.

However, little attention has been drawn so far to analyse and improve the interaction processes between the various users of an OD portal. In this paper, we argue that the interaction workflow in an OD portal can be interpreted from a business process perspective. In the light of this, we gain insights on the interaction behaviours, in order to enhance the overall user experience.

The actors involved are the depicted users in Figure 1, namely:

Data providers who interact with a portal in a data driven **publishing process**, coming from the public as well as the private sector. On the one hand, public administrations

[1] http://ckan.org/
[2] http://www.socrata.com/products/open-data-portal/

Pre-print copy of the manuscript published by Network of Excellence in Internet Science

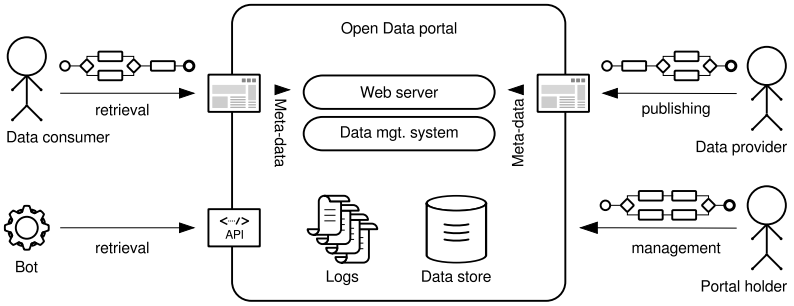

Figure 1: Conceptual architecture of the user interaction with Open Data portals

make their public information available and reusable, following diverse governmental directives to embrace transparency, to foster e-participation, to enhance value generation, etc. On the other hand, data providers from the private sector aim to advertise their data with the hope to create some kind of revenue (e.g., increasing their visibility or to trigger innovations).

Data consumers who use the portals in an information-discovery-driven **retrieval process**; their main goal is to find use-case relevant public information, e.g., to complement their datasets, to transform and analyse raw data or to feed novel applications and services, etc.

Furthermore, we stress the central role of the **portal holders**, i.e., the stakeholders providing all the necessary infrastructure and services for the previous parties and set up **management processes** to coordinate their interactions. Bots and other automated information collecting engines are not of interest to us, as their aim is crawling information to general indexing and storing purposes.

In detail, the main objectives of our proposal are *(i)* the *discovery* and characterisation of user interaction behaviours, from the perspective of both the publishing processes and the retrieval processes, *(ii)* the *understanding* of the interplay among the different interaction behaviours, expressed as communication patterns, and *(iii)* the *recommendation* of optimizations in the interaction design and for improving the data quality, based on insights gathered from the previous steps. According to the categories that Zuiderwijk et al. defined in their analysis about current socio-technical impediments of open data (Zuiderwijk, Janssen, Choenni, Meijer, & Alibaks, 2012), we aim at tackling the issues of *(i)* availability and access of information, *(ii)* findability of information, *(iii)* linking and combining data, and *(iv)* interaction with the data providers. In fact, the information availability would be favoured by a post-hoc analysis of sessions of data consumers, by suggesting a better indexing, or insertion, of knowledge items that have the worst search-frequency/no-answer ratio. The information would thus be made also more findable, by suggesting the highlight of most searched, and most updated, topics. The linking strategy would be enhanced by the analysis of how knowledge items are retrieved, in a sequence of consecutive searches. The interaction with the data providers would be highly improved, as a report-based feedback mechanism would be periodically drawn to their attention, on how data consumers browse and request the content they publish. In the remainder of this paper, we will delve deeper into the detail of how we manage to gather useful information from data available in OD portals, and to which extent we envision to manage such information, to achieve our goals.

2 Methodology

Our approach starts from the discovery task, carried out in a non-intrusive evidence-based way by exploiting machine-readable information, which is recorded by OD portals. The outcome of the discovery phase is exploited as an input for the subsequent understanding of communication patterns with the OD portal. Finally, the recommendation phase aims to improve the overall process encompassing both users and OD portal perspectives, and is discussed in detail in Section 3.

Available Data. In contrast to other data markets, most OD portals do not store datasets locally, but they hold the metadata to discover external datasets from providers. Acknowledging this fact, we identify the following three sources that serve as input data for the highlighted tasks, namely: *(i)* **HTTP access logs**, which contain high-level information about every interaction between the users and the portal at any time, *(ii)* **application-specific logs**, which contain information about portal internal operations such as queries to the datastore, and *(iii)* **data-store contents**, which are the catalogue contents of the portal.

Techniques. We envision to manage the data at hand by utilising and combining techniques from the following research areas: *(i)* process mining (van der Aalst, 2011), *(ii)* query log mining (Silvestri, 2010), *(iii)* text mining (Aggarwal & Zhai, 2012) and *(iv)* data mining (Bramer, 2013). Process mining techniques investigate the temporal and causal relations among performed activities, in order to build a global behavioural description of the process behind the enactment of recorded executions. Such techniques would be adopted for the elaboration of HTTP access logs. Although well established in the context of business process management, works such as (Poggi, Muthusamy, Carrera, & Khalaf, 2013) indeed already outline viable frameworks to extend process mining on HTTP logs analysis. Query log mining aims at discovering interesting patterns out of sequences of registered queries on the web. In our context, related techniques would be adopted in order to process the subsequent search queries within the OD portal, which would be retrieved by HTTP access logs and application-specific logs. Text mining refers to the machine-aided extraction of information from text. In our context, it would provide theoretic bases and implemented algorithms to derive information objects which are searched or modified, out of queries meta-data and low-level formulations, which we would retrieve from application-specific logs and data-store contents . We propose to build upon data mining methods such as correlation analysis and frequent itemset mining to find semantic links and usage relations within data stores.

Discovery. The discovery phase starts by studying the HTTP access logs. It helps us to determine when the user is accessing the site, searching or browsing for information, filtering or refining a search, and so forth. HTTP access logs thus reveal the sequence of elementary interaction tasks ordered by the available timestamps. In turn, analysing the application-specific logs helps us to discover query features and details such as the content of HTTP `POST` requests (which is usually not reported in HTTP access logs). In addition, these logs can contain information about which data was actually added, updated or deleted – which is usually not reported in the HTTP logs. In case these logs do not contain such information or they remain private, we consider to apply data monitoring methods to track modifications and keep snapshots of the datasets in the OD portal. Processing the data-store contents reveals the concrete accessed data and the returned results for every query.

Understanding. The aim of this phase is to understand not only the sequence of actions (unveiled in the discovery phase), but the underlying objectives of portal users. To do so, we

integrate the outcomes of the discovery tasks, which provides a holistic view of the interplay between user-portal interactions and portal data access or modification.

For instance, HTTP logs can suggest that the same user performed two queries in a row. Linking this information with the application-specific logs can reveal that the second query was a refinement of the first one. Inspecting the related data-store contents sheds light on the aim of such refinement, e.g., the user changed the query to narrow down the results.

We remark here that users are required to express their agreement on the usage of their session data, in order to comply with privacy concerns. We are confident that such permission would not be hard to obtain, as they would not be tracked and identified in person, but rather collectively, and after anonimization.

3 Services

Our proposed roadmap finally aims to develop a *service-based* framework that can be integrated within current OD portals (e.g., CKAN) to recommend optimizations. We can distinguish two main set of services that exploit the outcomes from the previous phases, focused on (i) optimizing the interaction in OD portals and (ii) improving the data quality in such portals.

Interaction optimization services are responsible of enriching the user experience with the OD portals. We foresee two main features:

- **Adaptive User Interface.** This service suggests user interface modifications to improve the content discovery experience. The aim is to conform to the consumer needs, either in an automatic fashion (e.g., reorganising query results to match user needs or highlighting related topics), or recommending future improvements to the OD portal holder (e.g., new filtering capabilities based on user interactions, shortcuts, new tag groups, etc.).

- **Change monitoring, usage audit and control.** In addition to commercial services such as Google Analytics, this service uses the aforementioned characterisation of the user behaviour to enable OD portal holders to audit the real usage of the portal over time and judge the effectiveness of different process optimisations. In turn, data providers can keep track of their reachability, make the appropriate amendments and control their effectiveness. This service also provide them with a powerful tool to reveal novel open data markets, once unresolved user questions, topics and issues can be detected.

Data quality improvement services provide publishers with means to improve the quality of their meta-data with the aim of attracting the attention of potential consumers. Two main features are identified:

- **Meta-data cleaning, enrichment and reachability improvement.** In contrast to passive data catalogues, this service helps the data provider to understand why consumers reach (or miss) their data, and how meta-data should be improved. The system identifies duplicated information and misleading meta-data, and suggests new tags or meta-data features to better satisfy the analysed user needs.

- **Data aggregation.** This functionality is meant to recommend to establish new relationships or dependencies between different datasets, if, e.g., they are commonly

accessed in the same user session, or described with similar meta-data (such as the same geolocation).

Finally, notice that a wider comparison between different OD portals can also support for identifying global trends, cross-domain links and other potential sources of improvements. Our initial work on monitoring OD portals represents a first important step in this direction.[3]

4 Conclusions

Most Open Data portals were founded with the idea of maintaining a "passive" virtual marketplace of OD, providing limited navigation, browsing and searching facilities for consumers, and simple publishing methods for providers. In this paper we argued that OD portals should play an active role in the OD scenario at large scale, offering more usable and adaptable interfaces for the consumers, and continuous feedback to data providers. We show that the combination of several mining techniques would provide the necessary infrastructure to tackle these problems. We also discussed a practical *service-based* framework that can be integrated within current OD portals, and how OD stakeholders would benefit from such framework.

Acknowledgements. The authors want to thank Univ.Prof. Dr. Axel Polleres and Univ.Prof. Dr. Jan Mendling for their valuable feedback and discussions. The research work of Javier D. Fernández is funded by Austrian Science Fund (FWF): M1720-G11. The work of Claudio Di Ciccio has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement 318275 (GET Service).

References

Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining text data*. Springer.

Bramer, M. A. (2013). *Principles of data mining, second edition*. Springer.

Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., & Doshi, E. A. (2013). *Open data: Unlocking innovation and performance with liquid information* (Tech. Rep.).

Poggi, N., Muthusamy, V., Carrera, D., & Khalaf, R. (2013). Business process mining from e-commerce web logs. In *Proc. of BPM, 2013.* (pp. 65–80).

Silvestri, F. (2010). Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, *4*(1-2), 1–174.

van der Aalst, W. M. P. (2011). *Process mining: Discovery, conformance and enhancement of business processes*. Springer.

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Alibaks, R. S. (2012). Socio-technical impediments of open data. *Electronic Journal of eGovernment*, *10*(2), 156–172.

[3] http://data.wu.ac.at/portalwatch/

This document is a pre-print copy of the manuscript
(Di Ciccio, Fernández, and Umbrich 2015)
published by Network of Excellence in Internet Science.

References

Di Ciccio, Claudio, Javier D. Fernández, and Jürgen Umbrich (2015). "Improving the Usability of Open Data Portals from a Business Process Perspective". In: *ODQ*. Politecnico di Torino. Network of Excellence in Internet Science, pp. 6–10.

BibTeX

```
@InProceedings{   DiCiccio.etal/ODQ2015:ImprovingUsabilityOpenDataPortals,
  author        = {Di Ciccio, Claudio and Javier D. Fern{\'{a}}ndez and
                   Umbrich, J{\"{u}}rgen},
  title         = {Improving the Usability of Open Data Portals from a
                   Business Process Perspective},
  booktitle     = {ODQ},
  year          = {2015},
  pages         = {6--10},
  crossref      = {ODQ2015}
}
@Proceedings{     ODQ2015,
  title         = {Open Data Quality: from Theory to Practice Workshop, {ODQ}
                   2015, Munich, 30 March 2015},
  year          = {2015},
  organization  = {Politecnico di Torino},
  publisher     = {Network of Excellence in Internet Science}
}
```