Semantic Cities
AAAI Technical Report WS-12-13

QuerioCity: Accessing the Information of a City (DEMO)

Spyros Kotoulas and **Vanessa Lopez** and **Marco Luca Sbodio** and
Martin Stephenson and **Raymond Lloyd** and **Aris Gkoulalas-Divanis** and **Pól Mac Aonghusa**

Smarter Cities Technology Center
IBM Research
Damastown Industrial Estate
Dublin 15, Ireland

Introdution

QuerioCity aims at creating an ecosystem for managing and accessing the information of a city, with a particular focus on transforming, integrating and querying heterogenous semi-structured data in an open environment.

This raises unique challenges in terms of:

- *Fitness-for-use.* The users of the system are not data integration experts and not qualified to use industry data integration tools. Furthermore, they are not able to query data using structured query languages.
- *Domain modeling.* The domain of the information is very broad and open. As such, generating and mapping data to a single model is infeasible or too expensive.
- *Global integration.* Addressing the information needs for solving problems in an urban environment requires integration with an open set of external datasets. Furthermore, it is desirable that city data becomes easily consumable by other parties.
- *Scale.* The data in a city changes often (streams), is potentially very large and it is interlinked with an open set of external data.

Other approaches in this domain have several limitations: *content portals* (e.g. twc open gov data portal, data.gov.uk, IOGDC (Ding et al. 2011)) lack sufficient capabilities to explore, navigate and query across collections of multi-domain data; *enterprise data integration* platforms require significant technical expertise and effort from the user; *technical tools for data cleaning and integration* (e.g. in (Gonzalez et al. 2010), (Huynh and Mazzocchi)) require some technical skills and lack semantic depth to be able to answer complicated queries.

QuerioCity uses and augments technologies from the fields of Linked Data and Semantics-based Integration. We are tapping on research output from the fields of Pay-as-you-go data integration (based on non-expert input) (Bizer et al. 2009), (Madhavan et al. 2007), RDF stores, Information provenance and Anonymization.

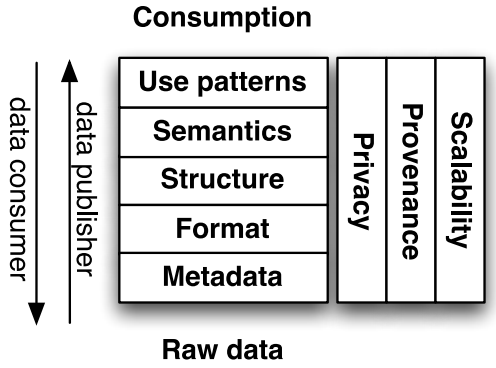

Figure 1: The QuerioCity approach

The QuerioCity Approach

QuerioCity takes a novel approach on managing highly heterogenous data. Instead of depending on integration experts, it lowers the entry threshold for domain experts and users of the system.

Figure 1 illustrates the overall approach taken in QuerioCity. Vertically, we illustrate the progression from raw source data to consumption. Initially, the system archives and catalogs *metadata* (for example, keywords, publishing date, corresponding data files), enabling rich queries over this meta-information, possibly in combination with full-text search. However, this is not enough when we also need to answer queries that span various datasets. By transforming the input into a uniform *format*, we are allowing queries where the data consumer knows the structure and the content of each file (possible by manual examination of the input). By harmonizing the *structure* of the various inputs, the user is able to query based on manually-mapped properties across datasets. Integrating the data into a common *semantic structure* allows transparent querying across datasets, without the need for manual mappings. Finally, given the heterogeneity (and consequently the semantic complexity) of the data, inferring *use patterns* allows users to create and share meaningful "views" over the data.

We also have the two major roles in the system: *data publisher* and *data consumer*. Our system takes a hybrid approach, where the integration effort is split between the two roles. QuerioCity provides tools for the data publishers to

"lift" the data as much as possible, as described in the previous section. On the other side, data consumers pull the data in order to fulfill their needs.

There are some issues that span all aspects of the system. Firstly, given the large volume of the data, the number of datasets and the potential number of users, the system has to provide good *scalability characteristics*. Second, given the open nature of the integration process, it is imperative that the system records *information provenance*. Finally, given the potentially sensitive nature of the information, the system provides functionality for detection of *privacy threats* (Fung et al. 2010) and tools for *anonymization*.

Demo

We will perform a demonstration of a prototype version of the QuerioCity platform, as used in the upcoming update of the dublinked.ie [1] website.

QuerioCity demo is based on static and dynamic data spanning a broad range of domains, including transportation, communication, housing, utilities, etc. Such data is provided by Dublin city.

The platform aims at taking raw city data in various formats from data publishers and making them readily accessible as linked open data. In order to do that the platform assits publishers in lifting their data and semantically annotating it. The assisted annotation allows for the creation of meaningful metadata by reusing existent vocabularies, such as dublin core[2], public sector ontologies [3] and DBpedia (Bizer et al. 2009). The metadata can help us in identifying the type of the content. For instance the content of a dataset annotated with the keyword "parking" can be further refined with the annotations "car park", "car parking permits", "resident parking", "disabled parking", "parking fines", "parking meters", etc. Good metadata allow users, who may not be an expert in a specific domain, to easily explore and mash up data. The platform leverages semantic data types (location, dates, unit of measure) to achieve a uniform type format by automatic conversions. Data can often be contextualized by annotating it with existing vocabularies and by linking it to semantically related entities from other existing datasets (e.g., two datasets about lighting poles in Dublin and Fingal).

On the data consumer side, beyond providing a SPARQL end-point and RESTful APIs to search on the metadata and associated data files, QuerioCity aims at providing a foundation platform for building complex visualisation and analytics explorations of public city data. The platform supports multiple interaction paradigms helping users in navigating information, reducing the search space, exploring data and enabling natural queries that scale. The system helps to discover content and fuse information sources, showing the intersections and relations between previously isolated datasets. Different types of data can be interpreted through different visualisations (Dadzie and Rowe 2011), such as timelines, charts, etc. As city data is situated on a specific temporal and geographical context, further insight is given by comparing datasets through spatial-temporal visualizations or heat maps (optionally points of interest can be extracted from external sources such as Linked open maps [4]).

As such, with QuerioCity the Web of data is brought to its full potential by enriching city data with external data sources, discovering relevant connections across heterogeneous data sources and domains, and understanding the interaction of human behavior with the system and the data (Gonzalez et al. 2010).

[1] http://www.dublinked.ie

[2] http://dublincore.org/

[3] http://doc.esd.org.uk/IPSV/2.00.html

[4] http://linkedgeodata.org/About

References

Bizer, C.; Lehmann, J.; Kobilarov, G.; Auer, S.; Becker, C.; Cyganiak, R.; and Hellmann, S. 2009. Dbpedia - a crystallization point for the web of data. *Web Semant.* 7(3):154–165.

Dadzie, A.-S., and Rowe, M. 2011. Approaches to visualising linked data: A survey. *Semantic Web* 2(2):89–124.

Ding, L.; Lebo, T.; Erickson, J. S.; DiFranzo, D.; Williams, G. T.; Li, X.; Michaelis, J.; Graves, A.; Zheng, J. G.; Shangguan, Z.; Flores, J.; McGuinness, D. L.; and Hendler, J. 2011. Twc logd: A portal for linked open government data ecosystems. *Web Semantics: Science, Services and Agents on the World Wide Web* In Press, Accepted Manuscript.

Fung, B. C. M.; Wang, K.; Chen, R.; and Yu, P. S. 2010. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.* 42(4):14:1–14:53.

Gonzalez, H.; Halevy, A.; Jensen, C. S.; Langen, A.; Madhavan, J.; Shapley, R.; and Shen, W. 2010. Google fusion tables: data management, integration and collaboration in the cloud. In *Proceedings of the 1st ACM symposium on Cloud computing*, SoCC '10, 175–180. New York, NY, USA: ACM.

Huynh, D., and Mazzocchi, S. Google Refine. *http://code.google.com/p/google-refine/*.

Madhavan, J.; Jeffery, S.; Cohen, S.; Dong, X.; Ko, D.; Yu, G.; and Halevy, A. 2007. Web-scale data integration: You can only afford to pay as you go. In *Proceedings of the Third Biennial Conference on Innovative Data Systems Research*.