Data Bridges: Data Integration for Digital Cities

Melanie Herschel, Ioana Manolescu

DataBridges: Data Integration for Digital Cities

Melanie Herschel
Université Paris-Sud & Inria Saclay
melanie.herschel@lri.fr

Ioana Manolescu
Inria Saclay & Université Paris-Sud
ioana.manolescu@inria.fr

ABSTRACT

The European Union has created the European Institute of Technology (EIT), within which ICT Labs focuses on fostering exchange and new result creation in the sphere of Information and Communication Technology, across the areas of research, higher education, and industrial innovation. "Digital Cities of the Future" is an action line (or chapter) of EIT ICT Labs. Within this action line, we coordinated an activity called "DataBridges: Data Integration for Digital cities", whose aim is to produce, link, integrate and exploit open data in the Digital Cities data space, for the benefit of both citizens and administrations. In that context, DataBridges addresses many research challenges such as acquiring (or producing) Open City Data in RDF, data linking and integration, data provenance, and visualization.

This position paper describes our efforts within DataBridges to integrate City Data. We provide a brief introduction to EIT ICT Labs, its goals and structures, and how DataBridges fits them. We then detail the activity and some selected results from 2011 - 2012, and plans for 2013. We conclude with some research challenges we plan to adress in the future.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.12 [**Interoperability**]: Data Mapping

General Terms

Algorithms, Design, Human Factors, Performance

Keywords

Open Data, Digital Cities, Data Integration, Data Mapping, Provenance

1. DIGITAL CITIES IN EIT ICT LABS

EIT ICT Labs (`http://eit.ictlabs.eu`) is one of the first three Knowledge and Innovation Communities (KICs) selected by the European Institute of Innovation and Technology (EIT) to accelerate innovation on Information and Communication Technologies (ICT) in Europe. Its goal is to foster innovation across Europe and to facilitate entrepreneurial research to rapidly bring research results to market. Within the area of Information and Communication Technologies, EIT ICT Labs has identified several *action lines* that are considered key societal issues in a number of selected areas. One of these action lines is *Digital Cities of the Future*. DataBridges is one activity in this action line.

To achieve its goal of more efficient and effective innovation, EIT ICT Labs brings together universities, research institutes, and companies within Europe to work on education, research, and industrial applications. In a research context, so called *activities*, such as the DataBridges activity, are multi-partner projects that span several institutions within EIT ICT Labs and that bring added-value to existing research projects of each activity partner through collaboration, facilitated people and result exchanges, help during standardization efforts, etc.

More specifically, each activity has a set of partners, and each partner has some *carrier projects*, funded by their own organization, by the government, or other sources. On top of these research carrier projects, *catalyst tasks*, funded by EIT ICT Labs, create added-value, as defined by the overall goals of EIT ICT Labs. In our discussion of DataBridges, we will not make the distinction between carriers and catalysts and simply present the goals and results of our activity.

2. OVERVIEW OF DATABRIDGES

Digital cities are tremendous marketplaces for sharing and producing digital information. The overall goal of DataBridges is to empower both individuals and institutions to make the best use of available data leading to exploitable information. Our vision is enabling applications that serve both citizens and administrations, showcased by the following scenario that we will use as testbed for our contributions.

2.1 Goals

Let us first illustrate our goals based on an example application that servers as an overall testbed for the technologies developed in DataBridges. By linking data from social media and Open Data, we will create a data set that allows citizens to plan what may be called "the perfect evening", e.g., find a restaurant with outside dining - if weather permits - having their favorite dish on their menu with reasonable price and quality and a night club not too far away to go after dinner, both reachable by public transportation at the intended times of travel. This requires data on points of interests (POIs), geographical coordinates, menus, weather, comments on the restaurants, public transportation, etc. The same data set, together with searches issued by citizens can be further analyzed to

CDMW'12, October 29, 2012, Maui, Hawaii, USA.

Partner	Research topics
2011 – 2012	
Aalto U	Hackathon events, best practices for data production (G4)
Bell Labs	Augmenting social data with semantics (G1)
DataPublica	Open City Data production (G1, G2)
Inria	RDF data management (G3)
TU Delft	User profile management (G4), RDF data integration (G1)
DFKI	Intelligent maps (G4), facetted browsing on City Data (G4)
KTH	Hackathons applied to City Data (G4)
U Paris-Sud	Data linking (G2)
2013	
FBK	Enrichment with context data (G1)
IBM	From relational data to Open City Data (G1)
Pol. Milano	C-SPARQL to process stream data (G1) Augmented reality mobile applications (G4)
Siemens	Machine-learning based recommendation (G4)
U. Twente	Point of Interest harvesting & linking (G1, G2)
TU Delft	RDF data integration (G1), RDF data provenance (G2)
DFKI	Schematic maps (G4), information extraction (G1)
KTH	Hackathons applied to City Data (G4)
U Paris-Sud	RDF data management (G3) RDF data provenance (G2), Data quality (G2)

Table 1: DataBridges partners and contributions

help administration in optimizing public transportation, identifying where POIs should be favored for city development, etc.

In this context, our activity has the research goals detailed below. Further goals include building a community through a series of workshops, participating to standard developments, and deploying our research in real applications.

G1 - Rich data. Provide rich data using the W3C's Resource Description Format (RDF for short), by leveraging information available in heterogeneous formats (relational, text, Excel etc.) to be mapped to RDF and by semantically enriching City Data with this new data. We also consider context information (time, location, environmental aspects) collected by devices such as SmartPhones or sensors to further enrich the data sets.

G2 - High quality linked data. Produce high quality linked RDF data sets to the Open City Data domain. Indeed, high quality data is essential for analytical applications on which administrative decisions (resource planning, water management, power plant planning, etc.) are based.

G3 - RDF warehousing and analytics. Develop technologies for scalable RDF data warehousing and analytics to actually perform Business Intelligence style analysis over RDF data.

G4 - User-centric applications. Design user-centric applications by using user-friendly visualization techniques, intuitive and effective information retrieval targeted towards City Data, and focused on user needs based on social data analysis and exploitation and recommendations.

The beneficiaries of our results will be citizens, administrations and companies, especially SMEs that will profit from gaining access to high quality Open Data.

2.2 Partners

Several partners were/are involved in the DataBridges activity, summarized in Table 1 together with the main areas each partner is involved in, and the goal their work contributes to (G1 – G4). The following discussions provide more details on this work.

3. ACTIVITY RESULTS (2011- 2012)

The activity focus in 2011-2012 has been oriented on three main axes. First, we focused on producing Open Data and enriching it as a means to exchange information within a Digital Cities setting. Finally, we considered issues raised by efficiently and effectively exploiting (sharing) such data.

3.1 Producing Open City Data

Open Data is a wide-encompassing movement around the idea that data should be freely shared and used to build added-value applications. From the bare data interoperability perspective, RDF is generally agreed upon as having sufficient expressive power and flexibility to serve as the format of choice for encoding Open Data.

A first task studied in this context concerned the *production* of Open Data for digital city contexts. Public administration data was one of the foremost applications of the Open Data movement, pioneered in the respective projects `http://data.gov` for the USA, and `http://data.gov.uk` in the UK. Information on the cultural attractions, public equipments and services, businesses available within cities etc. are naturally part of this effort. Producing Open Data from proprietary databases, however, takes significant effort, mostly because of the variety of native formats in which the data was produced and stored by the various stakeholders.

Best practices. The Aalto University partner, in collaboration with the City Region of Helsinki, had established the Open Data portal of the Greater Helsinki Area, `http://www.hri.fi`. Within DataBridges, Aaalto contributed from a software engineering perspective, a set of lessons learned on the most effective software tools and methodologies for producing Open Data out of proprietary databases.

Open City Data production. The DataPublica partner (`http://datapublica.fr`) is a French start-up which has pioneered the Open Data field in France by being the first to establish a large catalog of the Open Data produced by the national French administration. In France, important areas such as health, education, police, justice etc. are entrusted to the central government, who collects numerous statistics, most of which have been collected by DataPublica. They proceeded by manually establishing a catalog of public administration data portals and then semi-automatically feeding them into DataPublica's warehouse as a collection of *datasets*. A dataset corresponds to a specific category of information aggregated at a specific level over a specific period of time. For instance, *premature children born in each department of France in 2005* is a dataset, as well as *fruit production per kind of fruit in Languedoc-Roussillon in 2000-2008*. Such datasets are the natural unit at which public administration tends to gather and consolidate data. Unfortunately, most of these datasets come under the form of Excel tables, whose exploitation is not obvious. DataPublica is currently exploring a collaboration with a machine learning team (LIRMM, France) aiming at automatically extracting schemas from such Excel tables; conversion of the relations thus obtained into RDF would be the next step. For the time being, DataPublica relies on Google's DSPL (DataSet Publishing Language, `https://developers.google.com/public-data/`) to organize its datasets. One advantage in doing so is the possibility to take advantage of Google's data visualization facilities.

3.2 Integrating and enriching data

Open data sets by themselves can be considered data islands, and one goal of DataBridges is to build bridges between these islands to contribute to the Linked Open Data (LOD) vision. Linked Open Data aims at combining Open Data sources, by encouraging the publication of connections between facts within or across data sources. In digital cities, LOD may be issued by companies and local organizations, but also by the governing bodies of the city or of the country, in the spirit of the US initiative data.gov, for instance. DataBridges considers linking Open Data, but goes even further

by actually integrating and enriching City Data to obtain rich and clean data sets.

Data linking. University of Paris Sud researchers have addressed the *linked* aspect of the LOD concept, by seeking to connect data from structured tables, to concepts from a given ontology. The motivation is quite obvious. Once structured information has been extracted from the Web under the form of tables (such as can be found in the Google Fusion Tables project http://www.google.com/fusiontables, or the kind of tables extracted from DataPublica's Excel datasets as explained above), we may try to analyze and interpret information at the granularity of a single cell, in order to relate the content of the cells to concepts and values from a known domain ontology. Doing so enables one to combine (join) the data from the table, with other data sources, in order to analyze or enrich it etc. The right ontology to use can be manually supplied; this works well when a large number of (similar) datasets are extracted from a well-known, focused source (e.g. "The Ministry of Agriculture Yearly Production Statistics"). Alternatively, one may automatically choose the ontology by comparing attribute names and/or the data values against a large ontology, such as DBPedia etc.

Paris Sud has developed a geographic data type inference module, starting from geographical information from INSEE (http://www.insee.fr), the French national statistics institute, and the DBPedia ontology. A second dataset which has been similarly semantically enriched consisted of museum and restaurant information.

RDF data integration. The RDFGears project (http://wis.ewi.tudelft.nl/imreal/u-sem/rdfGears/) at TU Delft allows the specification of workflows that extend and combine existing RDF datasets. RDFGears provides a language with formal syntax and semantics, an intuitive user interface and an efficient implementation based on algebraic dataflow optimizations. The platform can extract RDF data from external and internal data sources, and make use of a database back-end to store potentially large intermediate results.

3.3 Exploiting Open Data

Being able to navigate, search, and interpret the data is essential to extract valuable information from Open Data.

RDF data management. Inria has focused on the efficient querying of LOD, or more generally, large RDF databases. A first contribution brought in this area concerns a framework for automatically selecting views to materialize in an RDF database in order to speed up the evaluation of a given workload of RDF queries [8]. A second contribution explores the possibility to host a large RDF database in a cloud setting, in particular within the Amazon Web Service (AWS) platform (http://aws.amazon.com) [4].

User profile management. U-Sem [1] is developed within this action out of the European projects Grapple (http://www.grapple-project.org/) and ImREAL (http://www.imreal-project.eu/) and is building an infrastructure that allows to capitalize on social and personal data of citizens to allow for advanced user modeling, e.g., for personalization in digital City Data management applications. We focused our work mainly on Social Web datasets, in particular Twitter datasets, i.e., datasets of citizens tweeting via Twitter. It features a set of strategies that allow for semantic enrichment of Twitter messages (tweets) and relates tweets to external (Semantic) Web resources [2]. It offers functionality such as entity recognition or discovery of relations between entities.

Facetted browsing. A facetted browsing tool implemented by the DFKI partner has been applied on the dataset about museums and restaurants obtained as a result of Open Data linking (http://digitaleveredelung.lolodata.org:8080/DigitalCities/page/). The interest of facetted browsing in this context is first, to allow users to explore the data by dimensions considered in the order that is most relevant to them, and second, to enable on-demand browsing through the heterogeneous, irregular data, ill-suited to more traditional search (based on fixed-schema forms).

Intelligent maps. A distinct DFKI project focused on integrating Open Data about POIs, into user-personalised, interactive maps. These are able to plan detailed trips within the city, finding objectives according to the user's specifications (e.g., children movie in the morning, restaurant at noon, park in the afternoon) and planning the trip while taking into account trip durations etc.

4. PERSPECTIVE FOR 2013

Our work plan for 2013 follows the same three research axes as in 2011-2012, i.e., data production, integration, and exploitation.

4.1 Producing Open City Data

In 2013, we shift our focus to exploit a large variety of sources to gather more Open Data, i.e., text, multimedia, relational, and streaming data. More specifically, the following efforts contribute to this goal.

Multimedia data. LiveMemories (http://www.livememories.org) collects, analyzes, and preserves digital memories (audio, video, text, etc.) at a large scale for a city community, thus facilitating and encouraging the preservation of such community memories to ultimately enrich cultural and social heritage. LiveMemories has focused on achieving these goals for the region of Trento, and in the context of DataBridges, the produced data set will be further enriched with LOD.

Text data. Information extraction techniques allow to extract important information hidden in text, in our sample scenario, we are for instance interested in information hidden in user comments on restaurants, menus in pdf formats, news articles about closed roads, etc. This information will then be used to enriched City Data sets, hence, it will be mapped to RDF and linked to other RDF data we produce or that are readily available. The information extraction methods we will employ are (i) extensions of KnowledgeStore [6], developed as part of the LiveMemories project to cover semi-supervised relation extraction and (ii) unsupervised information extraction engines developed in the Excitement (www.excitement-project.eu) and THESEUS RadSpeech (http://www.dfki.de/RadSpeech/) projects.

POIs are of particular interest in the City Data domain. As part of COMMIT (http://www.commit-nl.nl), van Keulen et al. are investigating methods to automatically harvest POI information from the Web, an independent contribution to the generation of Open City Data.

Data streams. C-SPARQL [3] is a continuous extension of SPARQL, demonstrated to be effective in domains where a large number of heterogenous data streams (e.g., mixing sensor and social sensing) have to be processed to detect complex events. Politecnico di Milano investigated the Stream Reasoning concept, specified the C-SPARQL language, prototyped the C-SPARQL engine and deployed it in BOTTARI [7], winner of the Semantic Web Challenge 2011. BOTTARI is an augmented reality mobile application that permits the personalized and localized recommendation of restaurants in Insadong (a district of Seoul) based on the temporally weighted opinions of the community. We will further study

and apply the capabilities of C-SPARQL in processing streaming data in the DataBridges context.

Relational data. IBM CAS Italy intends to roll out an experimental semantic data integration platform named ONDA [5], which is capable of integrating legacy relational sources by mapping them with a conceptual schema (ontology). Using ONDA in DataBridges has the potential of extending the type of data exploited by Digital Cities applications, as it adds relational data to the possible set of data sources.

4.2 Integrating and enriching data

Data linking and integration. Concerning data linking and integration, our efforts in 2013 will on one hand focus on extending or consolidating the contributions from previous years. More specifically, we will invest further effort in RDFGears to demonstrate its general applicability and to scale RDF data integration to large volumes of data. Linkage techniques as those developed at Paris Sud or [10] will be used and adapted to link the POI data with social media pages in the COMMIT project.

Transformation and data quality. The ultimate goal of DataBridges is to enable smart applications on top of Open City Data. One prerequisite for this is to use and produce high-quality data. Within DataBridges, we will first study processes lowering data quality in the City Data context to later on develop and apply appropriate data cleaning methods at scale. Sources of errors in result data are in general either errors in the data transformations leading to that data or errors in the data itself. For debugging purposes in this context, we will adapt why- and why-not provenance [9] techniques developed in the Nautilus project (`http://nautilus-system.org`) for relational data to RDF data and will implement these in RDFGears.

Data Contextualization. The Contextualized Knowledge Repository [11], another contribution within the LiveMemories project, is an RDF repository that supports the management of context dependent data (i.e., data of which the validity depends on the context given by time, location, topic, agent, etc.). Context dimensions are important to represent time/location dependent data, domain specific data or subjective data. Clearly, contextual data is relevant in the Digital City domain, as context on weather and road conditions given to news articles on accidents or live reports on traffic jams can improve the interpretation and action upon these incidents (reduce speed limit when raining, etc.). In DataBridges, we will enrich our data sets with contextual data. In the DataBridges setting, such context data will for instance be produced by CrowdSearch, contributed by Politecnico di Milano, that represents a novel search paradigm that embodies crowds as first-class sources for the information seeking process. CrowdSearch aims at filling the gap between generalized search systems, which operate upon world-wide information - including facts and recommendations as crawled and indexed by computerized systems – with social systems, capable of interacting with real people, in real time, to capture their opinions, suggestions, and emotions.

4.3 Exploiting Open Data

In 2013, we plan to pursue our work on the faceted search component previously developed, as well as schematic maps and visualization (DFKI). We further continue our work on devising methods and algorithms for efficient and scalable RDF data warehousing (Paris Sud). Moreover, we intend to deploy BOTTARI (see Section 4.1) in Europe.

As a testbed combining the technologies developed by activity partners, we will also set up an actual application resembling the one described in the introductory example of Section 2.1.

5. CONCLUSION AND PERSPECTIVES

We have described the DataBridges activity, connecting partners from academia and industry to generate rich and high quality City Data, define methods for RDF data warehousing and analytics, and build user-centric applications for Smarter Digital Cities. The exploitation of City Data is still in its infancy in DataBridges, which is natural since we first need to have rich, high quality data sets to exploit. In the future, we plan on devising more advanced analytics, applications, visualizations, etc. to get from this raw-data to actual information of interest to the user.

Moreover, we will invest further efforts in making our methods scale to large volumes of data. Possible avenues of research include extending our work to the cloud or defining new algorithms and tools, e.g., for fully automatic linking, quality assessment, and visualization.

Acknowledgements: We thank all partners of the activity (past and prospective), namely the task leaders Jan Hidders (TU Delft), Luciano Serafini (FBK), Guido Vetere (IBM), Yi Hang (Siemens), Bernardo Magnini (FBK), Daniel Sonntag (DFKI), Chantal Reynauld (Paris Sud), Maurice van Keulen (U. Twente), Dominikus Heckmann (DFKI), Yi Hang (Siemens), Inessa Seifert (DFKI), Emanuele Della Valle (Milano), Hannes Ebner (KTH), François Goasdoué (Paris Sud), Geert-Jan Houben (TU Delft), Hakim Hacid, (Alcatel Lucent Bell Labs), Petri Kola (Aalto), as well as all those who have participated to the individual tasks.

This work is partially funded by EIT ICT Labs activity 10803 and the French national grant ANR-11-EITS-003.

6. REFERENCES

[1] F. Abel, I. Celik, C. Hauff, L. Hollink, and G.-J. Houben. U-sem: Semantic enrichment, user modeling and mining of usage data on the social web. In *WWW Workshop on Usage Analysis and the Web of Data (USEWOD)*, 2011.

[2] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Semantic enrichment of twitter posts for user profile construction. In *Extended Semantic Web Conference (ESWC)*, 2011.

[3] D. F. Barbieri, D. Braga, S. Ceri, E. D. Valle, and M. Grossniklaus. C-sparql: a continuous query language for rdf data streams. *International Journal on Semantic Computing*, pages 3–25, 2010.

[4] F. Bugiotti, F. Goasdoué, Z. Kaoudi, and I. Manolescu. RDF Data Management in the Amazon Cloud. In *EDBT/ICDE Workshop on Data analytics in the Cloud (DanaC)*, 2012.

[5] P. Cangialosi, C. Consoli, A. Faraotti, and G. Vetere. Accessing data through ontologies with onda. In *Conference of the Center for Advanced Studies on Collaborative Research (CASCON)*, 2010.

[6] R. Cattoni, F. Corcoglioniti, C. Girardi, B. Magnini, L. Serafini, and R. Zanoli. The knowledgestore: an entity-based storage system. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, 2012.

[7] I. Celino, D. Dell'Aglio, E. D. Valle, Y. H. Marco Balduin and, T. Lee, S.-H. Kim, and V. Tresp. Bottari: Location based social media analysis with semantic web. In *ISWC Semantic Web Challenge*, 2011.

[8] F. Goasdoué, K. Karanasos, J. Leblay, and I. Manolescu. View Selection in Semantic Web Databases. *Proceedings of the VLDB Endowment (PVLDB)*, 5(2), 2011.

[9] M. Herschel and M. A. Hernández. Explaining missing answers to SPJUA queries. *Proceedings of the VLDB Endowment (PVLDB)*, 3(1), 2010.

[10] M. Herschel, F. Naumann, S. Szott, and M. Taubert. Scalable iterative graph duplicate detection (preprint). *Transactions on Knowledge and Data Engineering*, 2011.

[11] M. Homola and L. Serafini. Contextualized knowledge repositories for the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 12(0), 2012.