SPUD: Semantic Processing of Urban Data

Spyros Kotoulas, Vanessa Lopez, Raymond Lloyd, Marco Luca Sbodio, Freddy Lecue, Martin Stephenson, Elizabeth Daly, Veli Bicer, Aris Gkoulalas-Divanis, Giusy Di Lorenzo, Anika Schumann, and Pól Mac Aonghusa

Smarter Cities Technology Centre, IBM Research, Ireland

Abstract. We present SPUD, a semantic environment for cataloguing, exploring, integrating, understanding, processing and transforming urban information. Such information comes from the Web, government authorities, social networks and Linked Data sources. We identify a scenario consisting of the following cases: (a) *Publication* of a dataset, focusing on privacy protection and semantic annotation of the dataset, (b) *Reporting and Consolidation* of multi-faceted information, focusing on searching and visualizing heterogenous data from several sources, including social media, linked data and government data, (c) *Diagnosis*, based on highly expressive reasoning. Through our demonstration, we show that semantic technologies can be used to obtain business results in an environment with hundreds of heterogenous real datasets coming from different data sources and spanning multiple domains.

1 Introduction

Urban data comes in many forms, shapes and sizes. *Government agencies* are increasingly making their data accessible to promote transparency and economic growth. Since the first data.gov initiative launched by the US government, many city agencies and authorities have made their data publicly available through content portals: New York City, London, San Francisco, Boston, and Dublin, to name a few. In the meanwhile, *Linked Data* has emerged as a way to integrate information across sources and domains. Open and Linked data require that publishers put significant resources. A critical question for government agencies is what *return-on-investment* they are getting for resources spent in making their data open. This may come as an increase in economic activity in their constituencies, decrease in administration costs and increased transparency. *User generated content* can provide information outside of the scope of traditional data sources. For example, a traffic jam that emerges due to an unplanned protest may be captured through a twitter stream, but missed when examining weather conditions, event databases, reported roadworks, etc. Additionally, weather sensors in the city tend to miss localised events such as flooding. These views of the city combined however, can provide a richer and more complete view of the state of the city, by merging traditional data sources with messy and unreliable social media streams.

The urban data emerging from such sources may be used to support various operations such as exploration, visualisation, querying and diagnosis. Nevertheless, the cost associated with integrating *all* of this information is prohibitive. Our claim is that semantic technologies can be used to drastically lower the entry cost to accessing the information of a city. We demonstrate a technology platform to address key business challenges for urban information management: (a) *Publication* of a dataset, focusing on privacy protection and semantic annotation, (b) *Reporting and Consolidation* of multi-faceted information, focusing on searching and visualizing heterogenous data from several sources, including social media, linked data and government data and aggregating this information into a single view and (c) *In depth analysis* of this information, to derive conclusions with significant business value. Example of such conclusions include detection and diagnosis of events or anomalies.

The novelty of `SPUD` lies in the ability of the system to ingest highly heterogenous data and process it in an incremental manner. Unlike other approaches, the cost of entry is minimal (i.e. datasets can be imported as they are), and processing (annotation, linking, integration) can be done incrementally, while fully exploiting the power of semantic technologies. In addition, we are showing how a stack based on semantic technologies can go a long way, without the need for global integration, or even linking the entire input. We demonstrate `SPUD` using hundreds of real-world datasets, published by 4 local authorities, datasets from the Semantic Web and data retrieved from other Web sources.

This report is focused on the use-cases and the software platform for `SPUD`. Although some research in this report is still unpublished, we provide references to the relevant literature [1,4,2] when available.

2 Scenario

We are presenting `SPUD` through a series of business cases pertaining to ambulance response times in Dublin. The target audience has various roles within public administration (or contracted entity working with public authorities) and varied competency with regard to semantic technologies. For each use-case, we are outlining how `SPUD` addresses the related challenges.

2.1 Publish Data

A city official wants to publish a dataset about ambulance callouts [1]. The *capability* of the user is limited to Web browsing and using spreadsheets. The *goals* in this case is that publication should be easy while conveying as much semantics as possible and protecting privacy-sensitive information. In addition, given the cost associated with publishing data, the city authorities need to evaluate the return

[1] due to privacy considerations, we can not provide the original dataset, and have thus generated a synthetic dataset based on realistic values. An example of an anonymized dataset from Dublinked can be found in `http://dublinked.ie/datastore/datasets/dataset-027.php`

on investment for the publication of each dataset. In this particular case, this can be measured by how much the information or any information derived from it has been used. *This case is facilitated by an easy-to-use, form-based interface with* recommendations *for metadata terms, taken from the Semantic Web (e.g. IPSV, DBPedia, Dublin Core). In addition, we provide functionality to identify and* protect *sensitive information and automatically extract some semantic information (e.g. geographical coordinates). We provide a ranking of original or derived data sources by their use as a measure of their value to the community .*

2.2 Report and consolidate multi-faceted urban information

A city executive reads an article about ambulances missing targets for response times [2] and tasks a city official with creating a thorough report, including locations of critical infrastructure, problem points, citizen-centric information and some Key Performance Indicators. The *capability* of the user is as in the previous case, but the user has better insight in the organization of the municipal authorities. The *goal* in this case is to retrieve all relevant information. In addition, this information should be composed so as to get a single, thorough, view (for example overlaying delayed ambulances with known traffic jams from traffic systems and citizen reports from social media). In addition, to be able to process data in a quantitative manner, it should be possible to merge data into a single view. *Our exploration panel allows searching on content and metadata, on multiple data sources (including social media) and matches information on multiple levels - lexical, spatial and semantic. The retrieved information can be visualized on tables, maps and charts.*

2.3 Analyze and correlate information

As input to emergency authorities and traffic systems, the city authorities want to move further into understanding *why* a given traffic situation exists. Building on the output of the previous case, a data modeling engineer is tasked with creating a diagnosis component for traffic situations around sensitive infrastructure. The *goals* in this case are to fuse and semantically lift the data which is in turn fed to an automated reasoning component. *Our system analyses and ingests data from social streams, linked data sources and information published by municipal authorities. This information is either used as input or for validation of results obtained by the diagnosis reasoning engine. The result is a substantiated view of possible causes of traffic problems around sensitive health infrastructure.*

Through the aforementioned cases, we demonstrate how we can get value out of open, linked and social data in an urban setting. `SPUD` facilitates the entire data processing lifecycle, from information publication to gaining insight through highly expressive reasoning.

[2] `https://ibm.biz/Bdx2FJ`

3 Technology

SPUD uses semantic technologies in a plethora of ways: for representation of semi-structured data, as an interchange format between component, for data fusion (by mapping to common properties and by recommendations from external ontologies), for taxonomy generation, for automated diagnosis using reasoning and as an import/export format. In this section, we outline the technology palette in SPUD.

Cataloguing We provide a rich publishing interface that allows annotating datasets with relevant metadata from any vocabulary (we currently use IPSV, DBpedia, and Dublin Core). Our interface helps the user select appropriate terms by providing an semantics-augmented autocomplete function and a contextual view of the selected terms. In addition, SPUD semantically lifts the metadata already in Dublinked through the techniques presented in [1]. A panel gives an overview of the available datasets and allows navigation based on publisher, categories, provenance etc.

Data model and provenance tracking In [1], we are describing a flexible data model that allows incremental semantic lifting of data. We capture provenance (using the latest W3C working draft) by making views over datasets immutable and tracking differences between views (effectively allowing sharing of data between views). The meta-information pertaining to provenance and each data view is itself stored in RDF and can be made externally visible .

Data integration SPUD is using data integration techniques for linking data (using standard techniques known from LOD) and for semantically processing semi-structured input from social media. For example, one source of information in our scenario is the social media data obtained from the LiveDrive traffic update service that provides information about the city in the form of messages. This data is lifted into an events ontology and a domain-specific sub-ontology is automatically generated using a hierarchical clustering technique.

Data privacy / anonymization Privacy-preserving data publishing [3] is of great importance when dealing with city data [1]. In SPUD, we support vulnerability identification, data masking and anonymisation tools that enable data publishers to protect their data from re-identification and sensitive information disclosure attacks, prior to data publishing. SPUD operates by first identifying privacy vulnerabilities in the data and then sanitizing the data based on the datatype and the intended purpose of use. We showcase part of this functionality for the case of tabular data and anonymisation of location points in a map.

Diagnosis The diagnoser component [4], illustrated in a road traffic scenario, focuses on explaining anomalies (here ambulance delays) in the real-world context of Dublin City. Static and stream data from the road traffic domain (city events, road works, social feeds) are exploited. Our approach couples pure AI diagnosis approaches with semantic web technologies and ontology stream reasoning for accurate and quasi real-time diagnosing in an open-world context of heterogeneous and large data. Description Logic subsumption has been applied for inferring anomaly-diagnosis causality and evaluating historical diagnosis with respect to real-time conditions. Abductive reasoning is performed to construct the diagnosis report.

Social Media Mining Social media provides insight into the current and dynamically emerging status of a city. In the context of traffic, users send real-time traffic updates to authorities describing congestion, accidents and obstructions. However many of these updates come in the form of free text and are not geocoded. Natural language processing is employed to geocode these user contributed updates and merged with more traditional information sources such as city wide events and planned roadworks in order to i) validate semantically derived traffic diagnosis ii) provide additional detected events such as accidents.

Trajectory Miner The trajectory mining app is using geo-located tweets to mine user trajectories and give insight to the distribution and mobility of citizens. The application visualizes the intensity of users (tweets) activity in each specific region every 15 minutes. In addition, the origin-destination flow is mined and visualized for different times to day, and particular events, illustrating insight that is typically not captured by government sources such as censuses [2].

HTML5-based user Interface We are making extensive use of HTML5 and the Dojo toolkit to produce an attractive user interface. We are exploiting functionality such as drag-and-drop and context menus. Figure 1 outlines the main functionality in the user interface of `SPUD`.

Deployment Our technology stack is based on well-established commercial components from IBM. Figure 2, outlines the main components in our system. Critical components (HTTP/Application Server, RDF Store, Storage) can be clustered as required to ensure scalability and robustness. An alpha version of `SPUD`, intended for demonstration purposes only, is located at `http://www.dublinked.ie/sandbox/SemanticWebChall/`.

4 Conclusions

We have presented `SPUD`, a Semantic platform and set of enterprise apps for cataloguing, exploring, integrating, transforming and understanding urban information, coming from government sources, the Semantic Web and Social Media. The

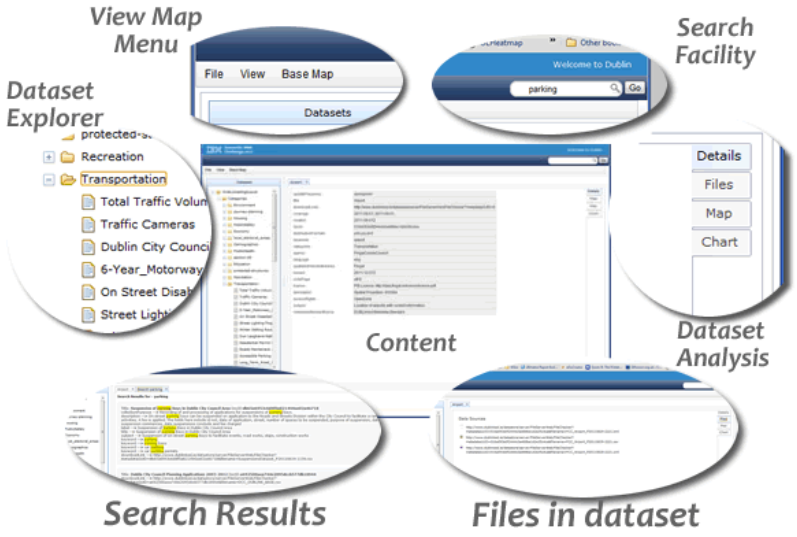

Fig. 1: Screenshot

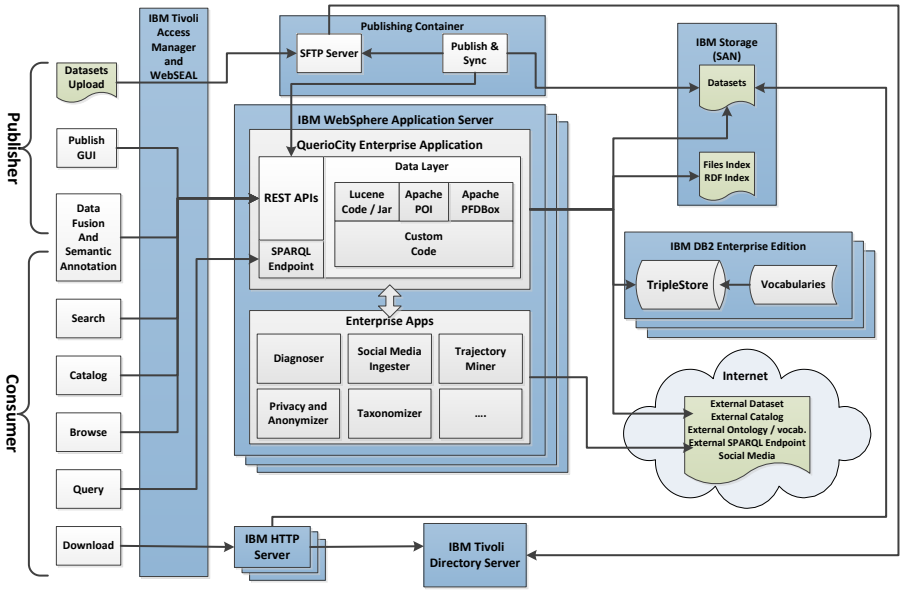

Fig. 2: System Architecture

novelty of SPUD lies in its ability to handle highly heterogenous data from various sources and support incremental integration. In addition, it deploys a palette of semantic techniques, ranging from data ingestion and integration techniques to highly expressive reasoning for anomaly diagnosis. SPUD demonstrates that semantic technologies can indeed be used to process urban information at scale.

We would like to thank the municipal authorities in Dublin and NUI Maynooth for their collaboration in Dublinked and Denis Patterson for facilitating the deployment of SPUD on the IBM systems.

References

1. Lopez, V., Kotoulas, S., Sbodio, M., Stephenson, M., Gkoulalas-Divanis, A., Mac Aonghusa, P.: Queriocity: A linked data platform for urban information management. In: Proceedings of the 11th International Semantic Web Conference. ISWC '12 (2012)
2. Calabrese, F., Pereira, F.C., Lorenzo, G.D., Liang, L., Ratti, C.: The geography of taste: Analyzing cell-phone mobility and social events. In: Pervasive. (2010) 22–37
3. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys **42**(4) (2010) 14:1–14:53
4. Lecue, F., Schumann, A., Sbodio, M.: Applying semantic web technologies for diagnosing road traffic congestions. In: Proceedings of the 11th International Semantic Web Conference. ISWC '12 (2012)

APPENDIX

Criterion	Rating	Explanation
End-User Application	H	Web-based application with users ranging from computer-literate professionals to experts
Diverse ownership or control of sources	H	We use data from social media websites, linked datasets and 4 different municipal authorities in Dublin.
Heterogenous sources	H	The input to our system consists of Social media streams, Map files, CSV files, proprietary formats like Office documents, RDF files etc.
Real-world data	H	At this time, we are using 228 real datasets represented in 1656 files and more than 200M triples.
Use of Semantic Web Technologies	H	Data is mainly represented as RDF. We use ontologies and reasoning.
Data processed to derive useful information	H	We show the business value of our system through 3 use-cases.
Suitability of Semantic information processing	H	Semantic technologies are used for flexible data representation and incremental integration and modeling. Traditional data management tools (e.g. RDBMS) would be very cumbersome to use, since we are lacking a global schema. Semantic technologies allow expressive reasoning with little a-priori processing and cleansing.

Criterion	Rating	Explanation
Attractive and functional Web interface	H	We make extensive use HTML5 and geo-mapping tools to produce an attractive interface.
Scalable application	H	The enterprise software components in our application can be clustered for scalability. `SPUD` can use arbitrary ontologies.
System evaluation and validation of results	M	We validate our approach through a set of business use-cases. We have not yet carried out a performance evaluation of our system.
Novelty in applying semantic technology	H	We demonstrate the usefulness of semantic technologies during the full circle from data publishing to getting a business result.
Functionality goes beyond information retrieval	H	Our system is used to integrate, understand and do deep processing of data.
Commercial potential	H	We show the business value of our system through 3 use-cases.
Contextual information for ratings or rankings	H	Usage of context is prevalent in ranking search results, ranking related dataset recommendations, correlating diagnosis results with social media and in processing information about the same event from multiple sources.
Multimedia documents	No	-
Use of dynamic data	H	Users can upload new data. Some data in our system is streaming (e.g. travel times are updated every minute).
Accurate results (i.e. use ranking)	H	Ranking is used in many places in our application: Searching for datasets, concept mapping, information integration, ranking of traffic diagnosis results etc.
Support for multiple languages and accessibility on a range of devices	No	-