The 3cixty Knowledge Base for Expo Milano 2015

Enabling Visitors to Explore the City

Giuseppe Rizzo
EURECOM, Biot, France
giuseppe.rizzo@eurecom.fr

Oscar Corcho
Universidad Politécnica de Madrid, Madrid, Spain
ocorcho@fi.upm.es

Raphaël Troncy
EURECOM, Biot, France
raphael.troncy@eurecom.fr

Julien Plu
EURECOM, Biot, France
julien.plu@eurecom.fr

Juan Carlos Ballesteros Hermida
Universidad Politécnica de Madrid, Madrid, Spain
jcballesteros@localidata.com

Ahmad Assaf
EURECOM, Biot, France
ahmad.assaf@eurecom.fr

ABSTRACT

In this paper, we present the 3cixty Knowledge Base, which collects and harmonizes descriptions of events, places, transportation facilities and user-generated data such as reviews of the city and Expo site of Milan. This knowledge base is used by a set of web and mobile applications to guide Expo Milano 2015 visitors in the city and in the exhibit, allowing them to find places, satellite events and transportation facilities around Milan. As of July 24th, 2015 the knowledge base contains 18665 unique events, 225821 unique places, 94789 reviews, and 9343 transportation facilities, collected from several static, near- and real time local and global data providers, including Expo Milano 2015 official services and numerous social media platforms. The ontologies used as a backbone for structuring the knowledge base follow a rigorous development method where the design principle has generally been to re-use existing ontologies when they exist. We think that the lessons learned from this development will be useful for similar endeavors in other cities or large events around the world with a similar ecosystem of data provisioning services.

Keywords

Knowledge base; Data Integration; Data Reconciliation; Expo 2015; Smart City; 3cixty

1. INTRODUCTION

In our information society, we often refer to data as the new oil. The Linked Open Data cloud (LOD) has captured this opportunity, collecting a vast amount of heterogeneous knowledge bases that are growing in number and density of links every year, covering different topical domains such as encyclopedic, medical, governmental, and environmental among others. Knowledge bases can be seen as a structured representation of the world that is machine processable, enclosing both data semantics and data instances. Nowadays, they are essential in building intelligent systems able to take actions according to the prior knowledge at their disposal, or crucial to be linked to for augmenting the human user experience of particular tasks. While LOD depicts a flourishing scenario of available knowledge bases, the ones at our disposal were only shallowly covering the topical domain of smart cities[1] as we needed them in the context of our knowledge base development for the 3cixty project.[2]

In the recent past, public administration bodies opened up action lines for releasing open data in order to allow data reuse for commercial and non-commercial purposes, as well as to increase transparency towards citizens. Unfortunately, such data is not always semantically interoperable across public administrations (even across datasets within the same organization) and with data owned (and sometimes released) by private companies, that results into data silos. In the specific case of cities, such open data typically contains statistical information about the city (e.g. GDP, population census, flooding alerts, election results, energy consumption, parking slots, parks, administration personnel), but more rarely points of interest (hotels, bike stations, tram, metro or train stations, museums, tourism attractions, restaurants, shops) and even less events. Given the statistical nature of such data, the live component of city life is often neglected and it is up to private companies to cover this gap. Nevertheless, both public administration bodies and private companies have difficulties in elaborating and aggregating data to provide new services, even if they could have a strong relevance in improving the citizens' quality of life and services, given the data heterogeneity and the limited access to it.

The proposal of building an integrated knowledge base for a city by combining open and private data, including data verification, reconciliation and validation, is not new. For instance, the KM4City (Knowledge Model for City) ontology [1] is the output of an expert study over smart city datasets, focused on creating a comprehensive ontology, which can answer the need of the community, and become a stan-

[1] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state
[2] https://www.3cixty.com

dard. The ontological work carried out in the 3cixty Knowledge Base (or simply 3cixty KB) overlaps with KM4City partially, the main difference being in the choice of reusing existing ontologies, instead of creating a new one. For example, in KM4City, the authors created the class `km4city:Service`[3] to represent points of interest, while in our knowledge base, we simply adopted `dul:Place`. The rationale for this choice stands in a longer maintained and largely used class, which, at the loss of subtleties, perfectly fits our scenario. We have also used the catalogue of datasets and ontologies of the Ready4SmartCities [2], as well as other sources that are explained in Section 2. The reuse of knowledge resources has been largely debated in the community: as an immediate consequence it optimizes cost and quality [6]. Turning cities' data ready to be consumed by web developers is the goal of the CitySDK project,[4] which offers a suite of uniform APIs to access numerous data directories. The heterogeneity is then addressed in a multitude of *ad-hoc* tools collecting data about citizen participation, mobility, and tourism. Similarly, in our work we collect data from different data sources such as the so-called E015 services,[5] a digital ecosystem of services tailored for the Expo Milano 2015, and from social media platforms such as Facebook, Foursquare, Google Places and Yelp.

The collected data is then harmonized according to the proposed data model and it populates the 3cixty KB that acts as a data marketplace for developers. It is accessible at `http://3cixty.eurecom.fr/sparql` and contains 284946 links to external resources on the Web. It covers data about past and upcoming events, happening in Milan, that are illustrated by media (photos and videos), places of interest (hotels, restaurants, bars, theaters, sights) that have reviews, transportation facilities (bus, metro, train, bike stations), as well as information about the estimated time that it takes to reach one place from another using public transportation. Overall, the 3cixty KB contains, as of 24/07/2015, 18789 events, 758 artists, 225821 places, 9343 transportation facilities, 96377 photos, and 94789 reviews contributed by 81944 users.

The remainder of this paper is organized as follows: in Section 2, we present an overview of the ontologies being used for modeling the 3cixty KB. In Section 3, we describe the Expo Milano 2015 use case, listing the data sources, data collection mechanism, and the approach we followed for the reconciliation. We conclude with discussions and an outlook on future activities in Section 4.

2. YET ANOTHER ONTOLOGY? THE 3CIXTY DATA MODEL

Our ontology design principle has focused on optimizing the coverage of the terminology in the context of city exploration. For each entity to model, we looked for existing knowledge resources (keyword search) in LOV,[6] Swoogle,[7] Watson,[8] and the Smart City catalogue[9] while the selection criteria are the popularity of properties based on usage data and favoring schema.org when suitable. We established a rigid search mechanism where two domain experts analyzed the knowledge resources that resulted from the search. Once consensus was reached, ontologies were taken and added to the 3cixty data model, which therefore consists in a constellation of existing ontologies. In `http://3cixty.eurecom.fr/ontology`, we have also published one additional ontology used to model time travel distance. The documentation about the complete data model is available at `http://3cixty.eurecom.fr/documentation`.

In details, we re-used some concepts and properties from the following ontologies: `dul`,[10] `schema`,[11] `dc`,[12] `lode`,[13] `geo`,[14] `transit`,[15] and `topo`.[16] A few additional classes and properties have been created to describe travel distances: we defined origin, distance, travel time, the nearest metro station and bike station.[17] The data model design is presented along with examples in the documentation.

3. 3CIXTY KB POPULATION: THE EXPO MILANO 2015 USE CASE

Expo Milano 2015 is the international exhibit, that has started on May 1st, 2015, and that should welcome up to 20 million visitors. We have prepared the knowledge base to enable the exploration of Milan by its visitors. The creation of the knowledge base goes through the selection of the prominent data sources, and timely and automatic process of data collection and the de-duplication solutions put in place to offer a high data quality. The knowledge base is publicly available via a SPARQL endpoint at `http://3cixty.eurecom.fr/sparql` and via web and mobile user interfaces at `https://www.3cixty.com`. A set of queries available at `http://3cixty.eurecom.fr/queries` provide examples of what type of questions can be answered by the knowledge base while a dashboard, which is available at `http://3cixty.eurecom.fr/dashboard`, has been developed to show some statistics and the evolution of the KB.

3.1 Data Sources

The 3cixty KB contains information about events, artists, places, transportation, and user-generated content such as media and reviews. It is built using three types of data sources: E015 data services, social media platforms, and Expo Milano 2015 manually curated data. The E015 services have been created with the goal to ease the sharing of data and services among different independent entities, with particular reference to data that can be used to support the Expo Milano 2015 visitors. A large portion of the data available through E015 concerns info-mobility. However, a growing share of the data now offered is of different nature, and concerns information about hotels, restaurants and events. More precisely, the most interesting categories of data available through E015 are: i) mobility information concerning the Milan municipality, such as local public transportation

[3] KM4City, v1.4 `http://www.disit.org/km4city/schema`.
[4] `http://www.citysdk.eu`
[5] `http://www.e015.expo2015.org`
[6] `http://lov.okfn.org/dataset/lov`
[7] `http://swoogle.umbc.edu`
[8] `http://watson.kmi.open.ac.uk/WatsonWUI/`
[9] `http://smartcity.linkeddata.es`
[10] `http://www.loa-cnr.it/ontologies/DUL.owl`
[11] `https://schema.org`
[12] `http://purl.org/dc/elements/1.1/`
[13] `http://linkedevents.org/ontology`
[14] `http://www.w3.org/2003/01/geo/wgs84_pos`
[15] `http://vocab.org/transit/terms`
[16] `http://data.ign.fr/def/topo`
[17] `http://3cixty.eurecom.fr/ontology`

routes, stops and timetables (bus, metro, train), current status of the metro and bus lines, current status of the bike sharing stations, etc.; ii) information concerning hotels and restaurants in Milan; iii) information about cultural events such as concerts, theater plays or museum exhibitions. We have then decided to combine such a wealth of data with web content usually available on social media platforms, the rationale being to increase the coverage of points of interests and to complement the description of existing events with user-generated activities. Finally, our last data source is about the Expo itself, the newly built physical site and its official events including what is happening in each national pavilions. An editorial team composed of five people on-site have mapped the geographical area and are continuously scraping the numerous agenda of events per pavilion.

Two teams of investigators analyzed and ranked the data sources at disposal to decide which ones were important to be selected for being included in the knowledge base. A strict procedure has been followed by the two teams, who maximize a 3-objective function: data semantics, instance coverage, and real-time update. The output of such an investigation led to a survey, which has been cross-validated by two domain experts who decided by consensus. An agile process has been established, granting to continuously update the list of sources. The current list, composed of 19 different data sources being used to populate the knowledge base, is listed below (the E015 data provides are marked with a *, the Editorial's data with $^+$):

Events
`http://3cixty.com`$^+$, `http://www.evensi.com`, `http://eventful.com`, `http://www.expoincitta.com`*, `http://www.fieramilano.it`*, `http://fondazionearnaldopomodoro.it`*, `http://www.leonardo-ambrosiana.it`*, `http://www.lastfm.com`, `http://www.teatripermilano.it`*;

Places
`http://dati.comune.milano.it`*, `http://developers.google.com/places`, `http://www.expedia.com`, `http://www.facebook.com`, `http://www.foursquare.com`, `http://www.isnart.it`, `http://www.vaxita.com`*, `http://www.yelp.com`, as well as the event data sources;

Media
`https://www.flickr.com` and the event and place data sources;

Transportation facilities
`https://www.bikemi.com`*, `http://dati.comune.milano.it`*;

User-related data
`https://developers.google.com/places`, `https://www.flickr.com`, `https://www.foursquare.com`, `http://www.lastfm.com`, `http://www.yelp.com`.

3.2 Grid Data Sampling and Data Collection

The data collection process uses, depending on the data source, a real-time and/or a batch procedure. To adapt the collection of data instances over a specific geographical extent, we use a grid sampling using as input the Milano Grid proposed in the 2014 BigData Challenge.[18] The grid, which has its focal point in the Milan center, is composed of 10K squared (approximately) cells, each having an edge of 234m. The cell identifier (via the property `locationOnt:cell`)[19] is attached to all the generated instances. This allows an indexing of the knowledge base to speed up query evaluation times when computing travel distance queries.

The data collection process takes as inputs a source type, a source name, a time coverage, and the collection mechanism. It materializes the collected instances according to the data model used for describing the entity type that the data source offers (each instance is marked with a provenance attribution using the `dc:publisher` property), and it generates a dump serialized in the Turtle syntax. Each dump is attached to a dedicated named graph which eases the data management. For the batch procedure, the data collection is prepared on a weekly basis. To optimize the expensive collection process, it collects only the delta of new instances available in each data source. For handling real-time data (data about availability of transport facility or hotel rooms), the collection is continuous.

3.3 Data Reconciliation

We have addressed the data reconciliation problem by performing: automatic instance reconciliation and manual category reconciliation. The rationale of having both types of reconciliation is to improve data quality by removing duplicates, as well as by combining information from alternative sources, hence easing the exploitation of the knowledge base for any data consumer. In fact, the category reconciliation has the objective to reduce the duplicates of categories when faceted browsing the data, while the instance reconciliation avoids having exact duplicates of the same individuals. In both cases, the reconciliation processes have been applied to the two main topical types of entities in the knowledge base: Events and Places.

3.3.1 Instance reconciliation

The different data sources that have been used to populate the knowledge base overlaps (e.g. the same hotel exists in several data sources). This concerns 0.67% of Events and 26.24% of Places. Given two data sources, namely A and B, an instance reconciliation process looks at identifying data instances that are similar according to their semantics. Generally, the output has the objective to generate *sameAs* links from two data sources. Using SILK [3], we developed our tailored settings grounding on the experimental findings discussed in [4], hence developing a better performing system in such a context as highlighted by the achieved results. The custom settings have been further tuned according to the result analysis in our benchmark scenario. For each of the topical classes, we performed a pairwise reconciliation of all the combinations without repetitions of the data sources, applying Eq. 1 for places, and Eq. 2 for events.

$$f_p = sum(\alpha_p * label_p, \beta_p * geo_p, \gamma_p * address_p) \quad (1)$$

where $\alpha_p = \frac{1}{2}$, $\beta_p = \frac{1}{3}$, $\gamma_p = \frac{1}{6}$. $label_p = jaroWinkler(labels)$ with a distance threshold of 0.2, with normalized labels (they are lowercased, we removed any special chars and stopwords).[20] Since all places are assigned to the corresponding cell of a map, we apply the $geo_p = exact(cell)$ with a distance threshold of 0. $address_p = tokenwiseDistance(address)$, with a distance threshold of 0.2, applying the same normalization as we did for the labels.

$$f_e = sum(\alpha_e * label_e, \beta_e * geo_e, \gamma_e * time_e) \quad (2)$$

[18] `https://dandelion.eu/datamine/open-big-data`

[19] This property is defined in the namespace `http://data.linkedevents.org/def/location#`.

[20] Comprehensive list: Milano, Milan, Italy, Italia, Lombardia, Lombardy, @it, @en.

where $\alpha_e = \frac{1}{2}$, $\beta_e = \frac{1}{4}$, $\gamma_e = \frac{1}{4}$. $label_e = levenshtein(labels)$ with a distance threshold of 0.2, with normalized labels (they are lowercased and we removed any special chars and the stopwords used for places). All events are also attached to a cell and we apply the $geo_e = exact(cell)$ with a distance threshold of 0. $time_e$ measures the timing distance between two events and we set the threshold distance to 6 hours. We then used as confidence scores $s_p = 0.5$ for places, and $s_e = 0.6$ for events, both in the range of $[0, 1]$.

A manual curation of the automatic output results of the reconciliation has led to the empirical values listed above. We evaluate the performances of this process on a gold standard composed of 100 instances for each type, uniformly sampled. The reconciliation procedure performs with an accuracy of 99.6% for Event instances and 90.24% for Place instances.

We then applied a resolution mechanism that attempted to prioritize the choice of the instances to resolve duplicates depending on the contribution of the data source each instance has been collected as defined by the two experts. To improve the efficiency of data consumers, we decided to build two graphs where to store the unique individuals, namely `http://3cixty.com/places` and `http://3cixty.com.events` respectively. Such sets are obtained as the distinct union of the outputs of the resolution mechanism and all the instances which do not hold any sameAs link.

3.3.2 Category reconciliation

Reconciling categories has the objective to reduce sparsity in the use of different labels for the same category groups. We addressed the process by using two category thesauri (implemented in `skos`) as pivots: the Foursquare taxonomy[21] for Places, and the taxonomy used in [4] for Events. The alignment, led by two experts of the domain, has established a set of links from the gathered categories, using `skos:closeMatch` and `skos:broadMatch`. An automatic process is then used to identify links according to the exact match of the found categories with the alignment defined by the experts.

4. DISCUSSION AND OUTLOOK

The Expo Milano 2015 offers a unique testing scenario for our knowledge base. 20 million people visitors are expected for this international exhibit, and a large portion of them may be able to use the 3cixty KB via the ExplorMI 360 application that has been endorsed by Expo.[22] We have established a weekly mechanism of data release and knowledge base updates. A central repository is designed to welcome discussion and record issues. In addition, team members are regularly participating to W3C discussions on the topics of smart city, event modeling, spatio-temporal representation, and open annotation, as well as on more local initiatives. This brings daily discussions on the standard approaches for modeling the data.

The 3cixty KB is the output of a specific need in exploring new opportunities for Expo Milano 2015 visitors to exploit transportation, business, cultural and touristic point of interests offered by Milan in a more personal and environmental manner. The research results, data model, and engineering solutions are ready to be deployed in other contexts (e.g. cities and worldwide events) with a minimal effort. In fact, the approach of having two streams of data (specific to city-related data providers and global social media platforms) allows being agile in building a knowledge base specifically tailored for any other city, improving the precision of the instances collected from social media platforms by using specific data feeds provided by cities themselves. The modular design makes it extensible to any city in the world, and it will be made sustainable since it will be used as the basis for research and commercial activities that will start during 2016. The 3cixty KB is also offered to the users as a curated data marketplace. This, in the exploitation activities currently under investigation, enables API developers to build new visual interfaces powered on top of the 3cixty data model and data resources.

In parallel, we are investigating how to mine the knowledge base to automatically generate new knowledge. Thanks to the rigorous mechanism of data source selection, we offer data quality, further validated by data instance selection and a rigid procedure of rdf-ization. Mining such data becomes of an unexploited value for research and commercial initiatives. A pilot research line that we are currently exploring is focused on generating automatically the geographic fingerprints of the Milan extent [5] that soon will be offered as another type of entity in the knowledge base. Another pilot research line is focusing on using association mining rules to align categories, which is currently performed by hand.

[21] `https://developer.foursquare.com/categorytree`

[22] A web application and mobile companion guides are accessible at `https://www.3cixty.com`.

Acknowledgments

This work was partially supported by the Innovation Activity 3cixty (14523) of EIT Digital (`https://www.eitdigital.eu`). The authors would like to thank all of the activity partners who contributed in daily discussions to shape and further refine the 3cixty KB.

5. REFERENCES

[1] P. Bellini, M. Benigni, R. Billero, P. Nesi, and N. Rauch. Km4city ontology building vs data harvesting and cleaning for smart-city services. *Journal of Visual Languages & Computing*, 25(6):827 – 839, 2014.

[2] R. García-Castro, A. Gómez-Pérez, and O. Corcho. READY4SmartCities: ICT Roadmap and Data Interoperability for Energy Systems in Smart Cities. In *11th Extended Semantic Web Conference (ESWC'14)*, 2014.

[3] R. Isele and C. Bizer. Active learning of expressive linkage rules using genetic programming. *Web Semantics: Science, Services and Agents on the World Wide Web*, 23(0):2 – 15, 2013.

[4] H. Khrouf, V. Milicic, and R. Troncy. Mining events connections on the social web: Real-time instance matching and data analysis in EventMedia. *Web Semantics: Science, Services and Agents on the World Wide Web*, 24(0):3 – 10, 2014.

[5] G. Rizzo, G. Falcone, R. Meo, R. Pensa, R. Troncy, and V. Milicic. Geographic Summaries from Crowdsourced Data. In *11th Extended Semantic Web Conference (ESWC'14)*, 2014.

[6] E. Simperl. Reusing ontologies on the Semantic Web: A feasibility study. *Data & Knowledge Engineering*, 68(10):905 – 925, 2009.