To trust, or not to trust: Highlighting the need for data provenance in mobile apps for smart cities*

Mikel Emaldi
Deusto Institute of Technology - DeustoTech
m.emaldi@deusto.es

Oscar Peña
Deusto Institute of Technology - DeustoTech
oscar.pena@deusto.es

Jon Lázaro
Deusto Institute of Technology - DeustoTech
jlazaro@deusto.es

Diego López-de-Ipiña
Deusto Institute of Technology - DeustoTech
dipina@deusto.es

Sacha Vanhecke
Ghent University - iMinds - Multimedia Lab
sacha.vanhecke@ugent.be

Erik Mannens
Ghent University - iMinds - Multimedia Lab
erik.mannens@ugent.be

ABSTRACT

The popularity of smartphones makes them the most suitable devices to ensure access to the services provided by smart cities; furthermore, as one of the main features of the smart cities is the participation of the citizens in their governance, it is not unusual that these citizens generate and share their own data through their smartphones. But, how we can know if are they reliable? How can we know if can a given user and, consequently, the data generated by him/her can be trusted? On this paper, we present how IES Cities platform integrates PROV Data Model and the related PROV-O ontology and Uncertainty Provenance set of attributes, allowing the exchange of provenance information about user-generated data in the context of smart cities.

1. INTRODUCTION

According to the "Apps for Smart Cities Manifesto"[1], smart city applications could be sensible, connectable, accessible, ubiquitous, sociable, sharable and visible/augmented. It is not a coincidence that all of these features can be found in a standard smartphone: the popularity of these devices makes them the most suitable device to ensure access to the services provided by smart cities. As one of the main features of the smart cities is the participation of the citizens in their governance, it is not unusual that these citizens generate and share their own data through their smartphones. Reviewing the literature, we can find many examples of apps that deal with user generated data like Urbanopoly [6], Urbanmatch [7] or the popular apps related to the 311 service in cities like Calgary, Minneapolis, Baltimore or San Diego, all of them available in Google Play[2]. The IES Cities project goes one step beyond, providing an entire architecture to allow citizens to develop apps based on Linked Open Data [3] provided by government, through user-friendly JSON APIs. All of these works that manage user-generated data have the same worry about these data: are they reliable? How can we know if can a given user and, consequently, the data generated by him/her can be trusted? Recently, the W3C has created the PROV Data Model [13], for provenance interchange on the Web. This PROV Data Model describes the entities, activities and people involved in the creation of a piece of data, allowing the consumer to evaluate the reliability of the data based on the their provenance information. Furthermore, PROV was deliberately kept extensible, allowing various extended concepts and custom attributes to be used. For example, the Uncertainty Provenance (UP) [9] set of attributes can be used to model the uncertainty of data, aggregated from heterogeneously divided trusted and untrusted sources, or with varying confidence. On this paper, we present how IES Cities platform integrates PROV Data Model and the related PROV-O ontology [12], and Uncertainty Provenance set of attributes, allowing the exchange of provenance information about user-generated data in the context of smart cities. The final aim is to enrich the knowledge gathered about a city not only with government-provided or networked sensors' provided data, but also with high quality and trustable data coming from the citizens themselves.

The remaining of the paper is organized as follows: in Section 2 the current state of the art on apps that deal with user data in the context of smart cities is presented. Section 3 outlines the main concepts about IES Cities project. Sections 5 and 4 describes the metrics to calculate the reliability of the data and its semantic representation, through a use cases, respectively. Finally, in Section 6 the conclusions and the future work are presented.

2. RELATED WORK

*This research is founded by project CIP-ICT-PSP-2012-6 "IES Cities: Internet Enabled Services for the Cities accross Europe", funded under "The Information and Communication Technologies Policy Support Programme". More info at http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=325097

[1] http://www.appsforsmartcities.com/?q=manifesto

[2] https://play.google.com

Regarding to smart cities' apps that use user generated content, we highlight the following works. Urbanopoly [6] presents an app for smartphones which combines Human Computation, *gamification* and Linked Open Data to verify, correct and collect data about tourism venues. To achieve this, Urbanopoly offers different games to the users, like quizzes, photo taking, etc. Similar to Urbanopoly, we can find Urbanmatch [7]. Urbanmatch presents a game in which the user takes photos about some tourism venues, to be published as Linked Open Data by the system. Another work that uses Human Computation for movie-related data curation is Linked Movie Quiz[3]. In [5], the authors present *csxPOI*, an application to allow its users to *collaboratively create, share, and modify semantic POIs*. These *semantic POIs* are modelled through a set of ontologies, developed for fulfilling of this specific task; and published as Linked Open Data. csxPOI allows users to create custom ontology classes, modelling new POI categories, and to establish subclass, superclass or equality relationships among them. In addition to create new classes, users can link these categories to a concept extracted from DBPedia [1]. In order to detect duplicate POIs, csxPOI clusters these POIs with the aim of finding similarities among them.

As can be seen, the authors that work with user-generated Linked Open Data have to deal with some replication, mismatching and data enrichment issues; and, as we have described before, the user is the most important agent in smart cities. In the next sections we explain how IES Cities project uses Provenance Data Model to represent provenance information about user-generated data.

3. IES CITIES

'IES Cities'[4], is the last iteration in a chain of inter-related projects promoting user-centric and user-provided mobile services that exploit both Open Data and user-supplied data in order to develop innovative services.

The project encourages the re-use of already deployed sensor networks in European cities, the existing Open Government related datasets, envisaging smartphones as both a sensor-full device in each citizen's pocket and a browser with increasing computation capacities to provide smart data to their owners.

IES Cities' main contribution is to design and implement an open technological platform to encourage the development of Linked Open Data based services, which will be later consumed by mobile applications. This platform will be deployed in 4 different European cities: Zaragoza and Majadahonda (Spain), Bristol (United Kingdom), and Rovereto (Italy), providing citizens the opportunity to get the most out of their city's data.

It is worth mentioning that no project before has considered so deeply the impact citizens may have on improving, extending and enriching the data these services will be based upon, as they will become leading actors of the new open data environment within the city. Nonetheless, the quality of the provided data may significantly vary from one citizen to another, not to mention the possibility of someone's interest in populating the system with fake data.

Thus, the need for evaluating the value and trust of the data requires the inclusion of a validation module [11]. In other words, we should be able to express special meta-information about the data delivered by IES Cities' users. The idea that a single way of representing and collecting provenance could be adopted internally by all systems does not seem to be realistic today, so the actual approaches are based on heterogeneous systems which export their provenance into a core data model, and applications that need to make sense of provenance information can then import it, process it, and reason over it [8].

In addition, when considering user-provided data measures for data consolidation have to be considered. Contributions from one user have to be cross-validated with contributions from other users in order to avoid information duplication and foster validation of other's data. Thus, data contributions from different users presenting spatial, linguistic and semantic similarity should be clustered [4]. Before a user contributes with new data, other user's contributions at nearby locations should be shown to avoid editing already existing data and encourage additions and enhancements to be applied to the existing data. After contributing with new data, the data providing user should be presented with earlier submitted similar contributions both in terms of contents and location in order to confirm whether their new contribution is actually a new contribution or it is amending an earlier existing one. This is, aids before and after editing new entries have to be provided and a two phase commit process for data put in place to ensure that contents of the highest quality are always added. Future work in IES CITIES will tackle these issues by providing REST interfaces to invoke services for clustering data entries and to retrieving related entries associated to a given one.

4. SEMANTIC REPRESENTATION OF PROVENANCE

To illustrate the semantic representation of trust and provenance data through Provenance Ontology and Uncertainty Provenance set of attributes, a use case is presented: 311 Bilbao. 311 Bilbao uses Linked Open Data to get an overview of reports of faults in public infrastructure. From the data owner's point of view, enrichment of their datasets by third parties, such as users of the 311 Bilbao application, revealed two problems: 1) the fact that data does not need to be approved before being published and that there is no mechanism to control the amount of data a citizen can add and 2) there is still the need for a way to differentiate the default trustworthiness of the different authors such as citizens and city councils. At the following code, the representation of the provenance of a user generated report is shown[5]:

```
1  @prefix foaf: <http://xmlns.com/foaf/0.1/> .
2  @prefix prov: <http://www.w3.org/ns/prov#> .
3  @prefix iesc: <http://iescities.eu/ont#> .
4  @prefix up:   <http://users.ugent.be/~tdenies/up/> .
5  @prefix :     <http://bilbao.iescities.org#> .
6
7  entity(:report_23456, [ prov:value="The paper bin is
8  broken" ])
9  wasGeneratedBy(:report_23456, :reportActivity_23456)
10 wasAttributedTo(:report_23456, :jdoe)
11 wasInvalidatedBy(:report_23456, :invActivity_639,
12 2013-07-22T03:05:03)
13
14 activity(:reportActivity_23456, 2013-07-22T01:01:01,
```

[3] http://lamboratory.com/hacks/ldmq/

[4] http://iescities.eu

[5] The provenance data is represented using Provenance Notation (PROV-N). More information at http://www.w3.org/TR/prov-n/

```
15  2013-07-22T01:05:03)
16  wasAssociatedWith(:reportActivity_23456, :jdoe)
17
18  agent(:jdoe, [ prov:type='prov:Person', foaf:name=
19  "John Doe", foaf:mbox='<mailto:jdoe@example.org>' ])
20
21  entity(:report_23457, [ prov:value="It is incorrect,
22  another paper bin has replaced the old one, but 2
23  meters beyond" ])
24  wasAttributedTo(:report_23457, :jane)
25  wasDerivedFrom(:report_23457, :report_23456,
26  :invActivity_639, -, -, [ prov:type='prov:Revision' ])
27
28  activity(:invActivity_639, 2013-07-22T02:58:01,
29  2013-07-22T03:04:47)
30  wasAssociatedWith(:invActivity_639, :jane)
31
32  agent(:jane, [ prov:type='prov:Person', foaf:name=
33  "Jane", foaf:mbox='<mailto:jane@bilbao.iescities.org>'
34   ])
35  actedOnBehalfOf(:jane, :bilbao_city_council)
36
37  agent(:bilbao_city_concil, [ prov:type=
38  'prov:Organization', foaf:name="Bilbao City Council"
39  ])
```

On this piece of semantic information the `:report_23456` resource represents the report made by the user. This report is identified by its own and unique URL and provides information about the user that has made it and which activity that has generated this report (lines 7-12). The `:reportActivity_23456` shows details about the activity that generated the report, like when the user starts to reporting the issue and when has ended. At line 18 the information about "John Doe", the user that reported the fault, can be seen. In the example given, another user, Jane (lines 32-35), has revised the report made by John (lines 21-30). As the `actedOnBehalfOf` asserts, Jane is some kind of municipal worker of Bilbao City Council (line 37). As Jane's report has more authority agains John's report, John's report is invalidated as `wasInvalidatedBy` asserts. Allowing the semantic descriptions of the provenance of the reports made at 311 Bilbao app, the data generated by a concrete user can be reached through SPARQL [14] language queries.

5. PROVENANCE DATA BASED RELIABILITY

There exist some approaches on how to calculate trust in semantic web using provenance information. IWTrust [**?**] uses provenance in the trust component of an answering engine, in which a trust value for answers is measured based on the trust in sources and in users. In [**?**] provenance data is used to evaluate the reliability of users based on trust relationships within a social network. [**?**] present an assesment method for evaluating the quality of data on the Web using provenance graphs, and provides a way to calculate trust values based on timeliness. In [**?**] the authors propose procedures for computing reputation and trust assessments based on provenance information.

In [] the authors identify some factors that affect on how users determine trust in content provided by web information sources. We have selected some of these parameters according to r

$$confidence = confidence_{assertion} * confidence_{content} \quad (1)$$

$$confidence_{content} = \frac{\sum_{p=1}^{n} \alpha_p * score_p(report)}{n} \quad (2)$$

$$score_{authority} = \begin{cases} 0 & if user \neq admin \\ 1 & if user = admin \end{cases} \quad (3)$$

$$score_{distance} = \frac{1}{geodistance(loc_{report}, loc_{reportedplace})} \quad (4)$$

Timeliness is another metric to be taken into account to evaluate data quality, and can be defined as the the up-to-date degree of a data item in relation with the task at hand. We propose and adaption of [10] formula to measure timeliness, based on the work described in [2]:

$$timeliness = (max(1 - \frac{currency}{volatility}), 0)^{sensitivity} \quad (5)$$

where *currency* is the difference between the time data is presented to the user and the time it was reported to the system. *Volatility* refers to the maximum amount of time a given report time should be active (for example, if a broken street lamp is reported, it should be repaired within a month at most), and *sensitivity* may change its value by observing the updates made over the status of the report: it would adopt a high value for data being constantly updated, and a low value for data that does not change often.

6. CONCLUSIONS AND FUTURE WORK

7. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, and Z. Ives. DBpedia: a nucleus for a web of open data. *International Semantic Web Conference*, pages 11—15, 2007.

[2] D. Ballou, R. Wang, H. Pazer, and G. K. Tayi. Modeling information manufacturing systems to determine information product quality. pages 462–484, 1998.

[3] C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[4] M. Braun, A. Scherp, and S. Staab. Collaborative semantic points of interests. In *The Semantic Web: Research and Applications*, page 365–369. Springer, 2010.

[5] M. Braun, A. Scherp, S. Staab, et al. Collaborative creation of semantic points of interest as linked data on the mobile phone. 2007.

[6] I. Celino, D. Cerizza, S. Contessa, M. Corubolo, D. DellAglio, E. D. Valle, and S. Fumeo. Urbanopoly – a social and location-based game with a purpose to crowdsource your urban data. In *Privacy, Security, Risk and Trust*, page 910–913, Amsterdam, 2012.

[7] I. Celino, S. Contessa, M. Corubolo, D. Dell'Aglio, E. D. Valle, S. Fumeo, and T. Krüger. UrbanMatch - linking and improving smart cities data. In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, editors, *Linked Data on the Web*, volume 937 of *CEUR Workshop Proceedings*. CEUR-WS, 2012.

[8] D. Ceolin, P. T. Groth, W. R. van Hage, A. Nottamkandath, and W. Fokkink. Trust evaluation through user reputation and provenance analysis. In

Uncertainty Reasoning for the Semantic Web, volume 900, page 15–26. CEUR-WS, 2012.

[9] T. De Nies, S. Coppens, E. Mannens, and R. Van de Walle. Modeling uncertain provenance and provenance of uncertainty in W3C PROV. In *International World Wide Web Conference*, page 167–168, Rio de Janeiro, Brazil, 2013.

[10] O. Hartig and J. Zhao. Using web data provenance for quality assessment. In *In: Proc. of the Workshop on Semantic Web and Provenance Management at ISWC*, 2009.

[11] O. Hartig and J. Zhao. Publishing and consuming provenance metadata on the web of linked data. In D. L. McGuinness, J. R. Michaelis, and L. Moreau, editors, *Provenance and Annotation of Data and Processes*, number 6378 in Lecture Notes in Computer Science, pages 78–90. Springer Berlin Heidelberg, Jan. 2010.

[12] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. Prov-o: The prov ontology. *W3C Recommendation, http://www. w3. org/TR/prov-o/(accessed 30 Apr 2013)*, 2013.

[13] L. Moreau, P. Missier, K. Belhajjame, R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, et al. Prov-dm: The prov data model. *Candidate Recommendation*, 2012.

[14] E. Prud'hommeaux and A. Seaborne. SPARQL query language for RDF, 2008.