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 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 can we know if these data are reliable? How can identify if a given user and, consequently, the data generated by him/her, can be trusted? On this paper, we present how the IES Cities' platform integrates the PROV Data Model and the related PROV-O ontology, 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 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, some examples of apps that deal with user generated data can be found, like Urbanopoly [4], Urbanmatch [5] or popular mobile apps related to the 311 service in cities like Calgary, Minneapolis, Baltimore or San Diego, all of them available in Google Play. The IES Cities project goes one step beyond, providing an entire architecture to foster the development of urban apps based on Linked Open Data[2] 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 [14], 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) [8] 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 [13], 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 4 and 5 describe the semantic representation of the provenance through a use case and the metrics to calculate the reliability of the data, respectively. Finally, in Section 6 the conclusions and the future work are presented.

2. RELATED WORK

The following works can be highlighted regarding smart cities' mobile applications. Urbanopoly [4] presents an app

*This research is funded by project CIP-ICT-PSP-2012-6 "IES Cities: Internet Enabled Services for the Cities accross Europe", 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] http://linkeddata.org/

for smartphones which combines Human Computation, *gamification* and Linked Open Data to verify, correct and gather data about tourism venues. To achieve this, Urbanopoly offers different games to the users, like quizzes, photo taking contests, etc. Similar to Urbanopoly, Urbanmatch [5] can be found, a game in which the user takes photos about some tourism venues, in order 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 [3], the authors present *csxPOI*, an application that allows its users to *collaboratively create, share, and modify semantically annotated POIs*. These *semantic POIs* are modelled through a set of ontologies developed to fulfill this specific task; and published following the Linked Open Data principles. 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 concepts extracted from DBpedia[4]. In order to detect duplicate POIs, csxPOI clusters the available 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 duplication, missclasification, mismatching and data enrichment issues; and, as previously described, the end-user has arisen as the most important agent in smart cities' environments. In the next sections we explain how the IES Cities project uses the Provenance Data Model to represent provenance information about user-generated data.

3. IES CITIES

'IES Cities'[5], 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 and the existing Open Government related datasets. It envisages smartphones as both a sensors-full device and a browser with increasing computational capabilities which is carried by almost every citizen.

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.

Remarkably, IES Cities wants to analyse the impact that 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 user contributed data requires the inclusion of a validation module [12]. In other words, we should be able to express special meta-information about the data submitted by IES Cities' users. The idea that a single way of representing and collecting provenance could be internally adopted by all systems does not seem to be realistic today, so the actual approaches modelling their provenance information 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 [6].

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 others' data. Thus, data contributions from different users presenting spatial, linguistic and semantic similarity should be clustered [2]. Before a user contributes with new data, other user's contributions at nearby locations should be shown to avoid recreating 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. In essence, aids before and after editing new entries have to be provided and a two phase commit process for user provided data should be 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 the Provenance Ontology, a use case is presented: 311 Bilbao. This app uses Linked Open Data to get an overview of reports addressing faults in public infrastructures. From the data owner's point of view, the enrichment of datasets carried out by third parties (such as users of the 311 Bilbao app), 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 council's staff. The following code represents the provenance of a user-generated report[6]:

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

[3] http://lamboratory.com/hacks/ldmq/
[4] http://dbpedia.org
[5] http://iescities.eu

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

```
15 activity(:reportActivity_23456, 2013-07-22T01:01:01,
16 2013-07-22T01:05:03)
17 wasAssociatedWith(:reportActivity_23456, :jdoe)
18
19 agent(:jdoe, [ prov:type='prov:Person', foaf:name=
20 "John Doe", foaf:mbox='<mailto:jdoe@example.org>' ])
21
22 entity(:report_23457, [ prov:value="It is incorrect,
23 another paper bin has replaced the old one, but 2
24 meters beyond" ])
25 wasAttributedTo(:report_23457, :jane)
26 wasDerivedFrom(:report_23457, :report_23456,
27 :invActivity_639, -, -, [ prov:type='prov:Revision' ])
28
29 activity(:invActivity_639, 2013-07-22T02:58:01,
30 2013-07-22T03:04:47)
31 wasAssociatedWith(:invActivity_639, :jane)
32
33 agent(:jane, [ prov:type='prov:Person', foaf:name=
34 "Jane", foaf:mbox='<mailto:jane@bilbao.iescities.org>'
35  ])
36 actedOnBehalfOf(:jane, :bilbao_city_council)
37
38 agent(:bilbao_city_concil, [ prov:type=
39 'prov:Organization', foaf:name="Bilbao City Council"
40 ])
```

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 8-13). The `:reportActivity_23456` shows details about the activity that generated the report, like when the user started reporting the issue and when it ended. At line 19 the information about "John Doe", the user that reported the fault, can be seen. In the example given, another user, Jane (lines 33-36), has revised the report made by John (lines 22-31). As the `actedOnBehalfOf` asserts, Jane is some kind of municipal worker of Bilbao City Council (line 38). 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 [15] language queries.

5. PROVENANCE BASED RELIABILITY

There exist some approaches on how to calculate trust in semantic web using provenance information. IWTrust [16] 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 [10] provenance data is used to evaluate the reliability of users based on trust relationships within a social network. [11] presents 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 [7] the authors propose generic procedures for computing reputation and trust assessments based on provenance information.

In [9] the authors identify 19 parameters that affect how users determine trust in content provided by web information sources, such as the authority of the creator of the information or the popularity and recency of that information, among others. Based on these factors, we have built a generic model for the measurement of a trust value in the context of IES Cities, in which the trust according to each factor is calculated independently:

$$trust(report) = \frac{\sum_{p=[auth,agree...]}^{n} \alpha_p * trust_p(report)}{n} \quad (1)$$

where p is the measured property and n is the total number of measured properties. α is a value between 0 and 1 to denote the relevance of this property, making the measure based on a certain property more or less relevant. $trust_p$ is a function that returns a value between 0 and 1 determining the trust of a given report according to a certain property.

Both the α values and the $trust_p$ functions can be defined by the developers using IES Cities platform, because both of them are dependant on the context and the need of the application domain.

To clarify, we are using this model in the 311 Bilbao use case. To that end, we have selected the most relevant trust-properties concerning our use case:

Authority: It refers to the fact that if a resource is created by an authority in a given context, this information is more reliable. For our use case a basic function like the following can be used:

$$trust_{authority} = \begin{cases} 0 & if user \neq authority \\ 1 & if user = authority \end{cases} \quad (2)$$

in which being authority can be checked with a SPARQL ASK query:

```
1 PREFIX prov: <http://www.w3.org/ns/prov#>
2 ASK { :jane prov:actedOnBehalfOf :bilbao_city_concil }
```

Popularity: The number of references and uses of a piece of information is a key aspect to determine its trust. In the case of 311 Bilbao we measure the popularity of a report based on the number of visits that the report receives, with the following formula:

$$trust_{popularity} = \frac{visits_{report}}{visits_{open\ reports}} \quad (3)$$

in which the number of visits of the report is normalized with the number of overall visits of opened reports at the moment.

Recommendation: Recommendation refers to importance that the ratings that other users gives to a given resource has in its trust. The function to measure the relevance of user ratings can be as sophisticated as the developer wants, but for our case we have selected a very naive and simple one, in which other users can vote the reports with +1 / -1 buttons and the trust value is calculed with this formula:

$$trust_{recommendation} = \frac{positive\ votes_{report}}{total\ votes_{report}} \quad (4)$$

Provenance / Reputation: In this case, *provenance* refers to the trust that the entities responsible for generating a piece of information may transfer information itself. A key aspect to measure the trust in a publisher is the reputation. There exist many approaches to measure the reputation of a user; some of them measure the reputation based on trust relationships between users [10], while some others like [7] are based the historical evidence of each user. For the our use case, we propose using the three-step procedure presented in [7]. In the 'evidence selection' step every report made by a given user are retrieved, in the 'evidence weighting' step the *recommendation* trust function is executed for every report, and in the last step all these trust values are aggregated through subjective logic to get the trustworthiness of a given user.

Recency / Timeliness: Timeliness 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 [11] formula to measure timeliness, based on the work described in [1]:

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

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.

Other trust factors: Apart from the aspects identified in [9], the model is flexible enough to include other factors affecting the trust. In the case of 311 Bilbao mobile app, the geographical distance could be a key aspect of the truth, as reports talking about events happening near to where the user sends the report would be more reliable.

$$trust_{distance} = \frac{1}{geodistance(loc_{report}, loc_{reportedplace})} \quad (6)$$

The function for the calculus of the geographical distance has as input the geographic coordinates of the report, retrieved from the smartphone GPS sensor, and the geographic coordinates of reported place, obtained with geolocation services like Nominatim[7].

After applying our model we will get a trust value between 0 and 1, that could be inserted in the provenance graph with a triple, assuming the confidence level was '0.6', like *:report_23456 up:contentConfidence '0.6'* [8].

6. CONCLUSIONS AND FUTURE WORK

The proposed approach in this article will allow to evaluate the provenance of user-submitted data in IES Cities' platform. The metrics proposed will measure data trustworthiness level, providing an extra confidence layer in the project's framework. City council staff and platform administrators will be able to query data quality through SPARQL queries, retrieving only those results with a confidence level above a parameterised threshold.

The evaluation and validation of the proposed metrics against other implementations following the PROV-O ontology will be left for a future iteration on IES Cities, aggregating other significant metrics should they improve the provenance of the generated data.

7. REFERENCES

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

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

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

[4] 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.

[5] 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.

[6] 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.

[7] D. Ceolin, P. T. Groth, W. R. van Hage, A. Nottamkandath, and W. Fokkink. Trust evaluation through user reputation and provenance analysis. In F. Bobillo, R. N. Carvalho, P. C. G. da Costa, C. d'Amato, N. Fanizzi, K. B. Laskey, K. J. Laskey, T. Lukasiewicz, T. Martin, M. Nickles, and M. Pool, editors, *URSW*, volume 900 of *CEUR Workshop Proceedings*, pages 15–26. CEUR-WS.org, 2012.

[8] 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.

[9] Y. Gil and D. Artz. Towards content trust of web resources. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(4):227–239, 2007.

[10] J. Golbeck. Combining provenance with trust in social networks for semantic web content filtering. In *Provenance and Annotation of Data*, pages 101–108. Springer, 2006.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] I. Zaihrayeu, P. P. Da Silva, and D. L. McGuinness. Iwtrust: Improving user trust in answers from the web. In *Trust Management*, pages 384–392. Springer, 2005.

[7] `http://wiki.openstreetmap.org/wiki/Nominatim`