UrbanMatch – linking and improving Smart Cities Data

Irene Celino
CEFRIEL, Italy
irene.celino@cefriel.it

Simone Contessa
CEFRIEL, Italy
simone.contessa@cefriel.it

Marta Corubolo
Politecnico di Milano, Italy
marta.corubolo@polimi.it

Daniele Dell'Aglio
CEFRIEL, Italy
daniele.dellaglio@cefriel.it

Emanuele Della Valle
Politecnico di Milano, Italy
emanuele.dellavalle@polimi.it

Stefano Fumeo
CEFRIEL, Italy
stefano.fumeo@cefriel.it

Thorsten Krüger
SIEMENS, Germany
thorsten.krueger@siemens.com

ABSTRACT

Urban-related data and geographic information are becoming mainstream in the Linked Data community due also to the popularity of Location-based Services. In this paper, we introduce the UrbanMatch game, a mobile gaming application that joins data linkage and data quality/trustworthiness assessment in an urban environment. By putting together Linked Data and Human Computation, we create a new interaction paradigm to consume and produce location-specific linked data by involving and engaging the final user. The UrbanMatch game is also offered as an example of value proposition and business model of a new family of linked data applications based on gaming in Smart Cities.

1. INTRODUCTION

Urban environments are experiencing a progressive digitization that is leading to the creation and release of large amounts of data: information about interesting aspects of cities – ranging from street topology and traffic conditions to business activities, from points of interest (POI) to events, from environmental measures to people life-logs – are increasingly present on the Web and even proactively fed by the open community. The growing attention given to Smart Cities themes and problems and the ever increasing popularity of location-based applications make urban ecosystems at the center of the research and innovation agenda of public authorities and big industrial players such as IBM with its Smarter Planet initiative[1], and CISCO with its Smart+Connected Communities initiative[2].

Also the Linked Data world turned to urban data, especially in the area of spatial information. One of the most curated dataset available in this field is GeoNames[3]. Recently, open Web APIs like Open Street Map[4] and geographic datasets from public administrations were (partially) turned into a Linked Data form by efforts like LinkedGeoData[5] [2] and the Spanish GeoLinkedData.es[6].

Still this massive bulk of urban data is largely unexplored and poorly exploited [15]. Two main drawbacks hamper a larger adoption of Linked Data in Smart Cities scenarios: the *doubtful quality* of the available information and the *lack of user-centred tools* to consume such data. Those two (well-known) problems create a vicious cycle: the unreliable quality of data makes people distrust Linked Data content and the lack of usable tools prevent people from contributing to the Linked Data improvement.

Our research hypothesis consists in employing user-friendly mobile gaming applications to engage people in mobility; through those games, Linked Data related to urban environments are consumed, created, improved and corrected. While a similar approach was already partially explored for Linked Data at large [22], we believe that the popularity of location-based services (LBS) can make this approach successful for urban-specific Linked Data: people are more and more used to "check-in" physical places with their mobile devices and to add small bits of information related to their activities and actions in the physical world.

In this paper we present our first experiment to prove our hypothesis: UrbanMatch is a mobile and location-aware Game with a Purpose that engages players to provide information related to the city of Milano. Specifically, UrbanMatch is aimed at linking points of interests in the city with the most representative photos retrieved from Web sources.

The remainder of this paper is organized as follows. Section 2 introduces motivation and related work; Section 3 explains the mechanics and purpose of the UrbanMatch game, while the evaluation methodology is shown in Section 4; finally Section 5 draws some conclusions and future works.

2. MOTIVATION AND BACKGROUND

Our work focuses on link creation and quality assessment for Linked Data in urban scenarios. In this section, we il-

[1] http://www.ibm.com/uk/smarterplanet
[2] http://www.cisco.com/web/strategy/smart_connected_communities.html
[3] http://geonames.org/
[4] http://www.openstreetmap.org/
[5] http://linkedgeodata.org/
[6] http://geo.linkeddata.es/

LDOW2012 April 16, 2012, Lyon, France.

lustrate the motivation problem, explain the specificity of Smart City solutions and introduce and discuss related work.

2.1 Data Quality and Linked Data

Data Quality [19] is the discipline that studies the most appropriate and relevant features to describe the value of data. Examples of dimensions defined by Data Quality are consistency, completeness, accuracy, and relevance.

A key point of Data Quality is that the quality measurements are context-dependent: given a dataset, its quality can be very high with respect to the fulfilment of some tasks but very bad for other ones. In other words, it is not relevant (and not always possible) to define absolute quality values [18]. As pointed out in [16]: "The perception of information quality on the WWW is highly dependent on the *fitness for use* being relative to the specific task that users have at their hands".

For the last years, the Linked Data community has started to follow this topic with growing interest: the birth and growth of the Linked Open Data (LOD) cloud introduced a huge amount of RDF data distributed across several datasets. The Linked Data best practices alone [14] assure more quality than "raw data" in closed databases because: *a*) data becomes *accessible over the Web* rather than being closed up in silos; *b*) the use of *shared vocabularies* makes the data both easier to "read" (i.e. user information needs can be satisfied by a single SPARQL query instead of requiring many dataset-specific queries) and easier to "interpret" (i.e. shared vocabulary semantics can be used to verify data integrity and/or infer implied data); *c*) the *presence of links* makes it also possible to verify consistency across different sources.

Still, problems related to the available data quality soon arose in the LOD cloud. While at the beginning the number of statements and the number of links were used to estimate the relevance of the published data sets, more recently the definition of more detailed and expressive metrics to describe the available data has become more and more important.

Flemming worked on the definition of quality criteria for linked data sources [10]. Fürber and Hepp studied how to integrate the Data Quality Management processes into the Semantic Web. In [12] they presented their approach to use SPARQL and SPIN to model Data Quality rules and execute them, while in [13] they defined an ontology (named DQM ontology) to describe Data Quality dimensions in RDF.

However, the assessment of data quality factors like accuracy, timeliness, completeness, relevance and comprehensiveness of data is intrinsically a hard task that Linked Data technologies do not make any easier.

2.2 Urban Linked Data

In the last years we have been largely experimenting with urban related linked (and non-linked) data in the development of a number of Smart City demonstrators using Linked Data related to urban environments. The Urban LarKC[7] [7] and then the Traffic LarKC[8] [3] integrated DBpedia, an Eventful wrapper and two Milano municipality's datasets with all Milano streets and three years of traffic sensors data; those applications made it possible to answer queries like "which are the modern art exhibitions that I can reach today in less than 25 minutes if I can get into my car this afternoon at 4pm?" [8]. The Seoul Road Sign Management [17] integrated LinkedGeoData POI, OpenStreetMap streets and a private dataset describing Seoul road signs to check the validity of all road sign information. Finally, the mobile app BOTTARI [4] explored social media to provide location-based recommendations for POIs.

These experiences allow us to assert that: *a*) the *quality* of urban Linked Data *unpredictably ranges* from very good to very poor; *b*) it is possible to detect *missing information or inconsistencies* by cross-validating datasets that describe the same urban space from different points of view (e.g., if a road sign tells to turn right to reach a POI, and the POI is nearby, but no street reaches it, a street is missing in the topology dataset); *c*) when inconsistencies or missing data are detected, data quality can be easily increased by a small amount of manual work that does not require specific skills, but often the *physical presence* in the urban environment.

Those three assertions may not be valid for Linked Data in general, but appear to be valid for urban Linked Data. The third point suggests us that a part of the manual work required to fix urban Linked Data could be "crowd-sourced". The recent popularity of Location-based Services (LBS) like Foursquare demonstrates that people are willing to share small bits of information on-the-go by exploiting their mobile devices capabilities. "Unlock your city" – the slogan of Foursquare – reveals that LBS can be considered an effective means to collect useful information for Smart City applications.

2.3 Related work

Assessing data quality is a hard problem for computers. We, as humans, are perfectly capable of it, but we are not necessarily willing to. Human Computation [25] and Social Computing [6], however, demonstrated that a number of different "computations" can be carried out by groups of people. In this area Games with a Purpose [24] (GWAP) emerged as a means to engage people to perform activities that are almost trivial for humans and very complex for computers; these tasks range from labelling images to improve web searching, from transcription of text (where OCR software fails) to any activity requiring common sense or human experience.

The incentives to make people contribute to Human Computation can be of different kinds: they can give the participant an explicit and concrete reward (like in the popular Amazon Mechanical Turk – also named MTurk[9] – in which people are paid to perform small and simple tasks) or they provide a different kind of implicit or more abstract return, for example by means of entertainment like in GWAPs.

In the Semantic Web community, Games with a Purpose were already used to cover the complete Semantic Web lifecycle [21]. A dedicated community portal was recently set up[10] to collect those games. A good showcase is the Linked Data Movie Quiz[11] [1], that builds a cinematographic game based on the available movie-related Linked Data showing that "the answers are out there; and so are the questions". Recently, [20] investigated which Linked Data management tasks can be easily and semi-automatically turned into crowdsourcing assignments for the MTurk platform.

[7] http://larkc.cefriel.it/alpha-Urban-LarKC/
[8] http://larkc.cefriel.it/traffic-larkc/
[9] http://mturk.com/
[10] http://www.semanticgames.org/
[11] http://lamboratory.com/hacks/ldmq/

3. THE URBANMATCH GAME

UrbanMatch[12] is a mobile location-based Game with a Purpose that aims at selecting the most representative photos related to the points of interest (POI) in an urban environment; more specifically, UrbanMatch is oriented to link the monuments and relevant places of the city of Milano with their respective photos as retrieved from social media Web sites and to "rank" those links, so to identify the most characteristic ones and to discard the others, thus improving the quality.

3.1 Data input

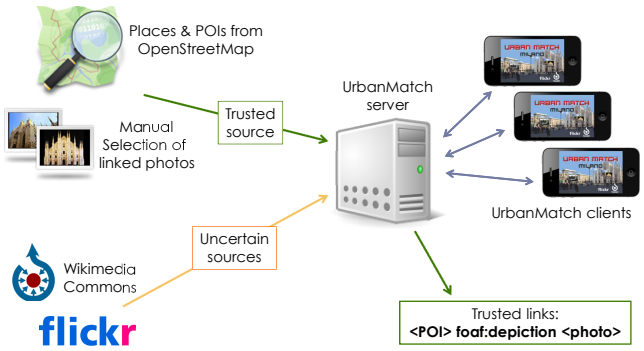

Figure 1: Input data sources and output links in UrbanMatch.

The input data come from available Web sources (cf. Figure 1). Points of interest in Milano were collected and chosen among those available from OpenStreetMap; an RDF description of those POIs is also available in LinkedGeoData, the linked data version of OpenStreetMap. For each of the 34 POIs of this set, we manually selected 5-6 photos depicting them (some chosen on the Web, some taken by ourselves). In this way we built a "*trusted set*" of 196 links that relate the POIs with their respective images; those links are expressed in the form:

```
<POI> foaf:depiction <photo> .
```

A much higher number of photos of Milano POIs was collected from Wikimedia Commons[13] – the media collection of Wikipedia – and from Flickr[14], probably the most popular social media sharing site dedicated to photos. The images were collected either by keyword/concept search (i.e., photos explicitly related to Milano POIs) or via location-based queries (e.g., search by geographical coordinates). Among the collected photos, we considered only those released with an open license, allowing for a free reuse of the image (like CreativeCommons "Attribution" license).

This second set of information is considered – for the game purpose – an "*uncertain source*": it consists of more than 37,000 "candidate" links that relate the POIs with the images that potentially depict them. This link-set is uncertain or untrusted because the retrieved photos can be incorrect (i.e. they are not related to Milano POIs even if returned by the search), their metadata can be wrong or incomplete or they cannot be considered as representative for the game purpose (e.g., a photo is actually taken in the proximity of a POI, but it does not depict it or it is focused on an irrelevant detail).

Those candidate links are expressed as RDF links using the `foaf:depiction` predicate as explained before; however, those links are further annotated with a *confidence value* that expresses the lack of certainty about their trustworthiness (e.g., the initial confidence of links to Wikimedia images is set to 60%, links to Flickr to 40%).

3.2 Gameplay

The UrbanMatch game is a photo coupling game. The game mechanics respects the best practice of casual games and Games with a Purpose [11]: it consists in a simple and intuitive interface that presents the player with 8 photos of POIs in the vicinity of the player and asks for their coupling (cf. Figure 2).

Figure 2: Screenshots of UrbanMatch.

The links between the surrounding POIs and the presented photos is not the same for all the presented 8 photos: some links are *certain*, because they come from the trusted source; some are *uncertain*, because they are taken from the set of candidate links; finally, some are *distractors*, i.e. they are not related to the surrounding POIs and are used to check the reliability of players. Each game is organized in multiple levels of increasing difficulty, i.e. with a varying number of certain/uncertain/distracting links (the higher the level, the greater the uncertainty degree).

The user geographic position taken from the mobile device sensors computes the user proximity to the "playable places", i.e. the game locations; even if the game allows for playing from any place, the user location is used to distinguish between the choices operated "on site" – based on the player's *experience knowledge* – and couples selected on the basis of the player's *domain knowledge*.

The UrbanMatch game is available at `http://bit.ly/um-itunes` on the iTunes store. Being a research prototype, the game will be available only for a few months.

[12] `http://bit.ly/urbanmatch`
[13] `http://commons.wikimedia.org/`
[14] `http://www.flickr.com/`

3.3 Game purpose and Data Analysis

Through the UrbanMatch game we aim at identifying and selecting the POI-photo "correct" links among the candidate links in the uncertain input source. Through a Human Computation approach, we aim at collecting evidences of players decisions to correlate images.

The approach we follow to post-process the collected data is similar to that of other Games with a Purpose [24]. From the collection of all evidences of image coupling as performed by the game players, with majority voting and other statistically-relevant algorithms [5], we alter the confidence value of each POI-photo link.

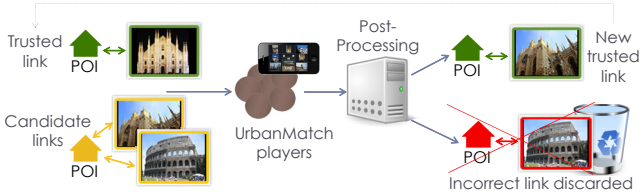

Figure 3: Computing UrbanMatch links.

Intuitively, the game purpose processing is represented in Figure 3: in each game level, trusted POI-photo links (the green one on the left with Milano's Duomo picture) are presented together with a number of candidate links related to the same POI (the yellow ones on the left, with uncertain photos retrieved from Wikimedia Commons and Flickr as being related to Milano's Duomo) and with a number of distractors (for simplicity not drawn in the figure).

If players couple the same two photos, those photos reasonably belong to the same POI: thus, if a player associates a trusted photo with an uncertain photo, the candidate link related to the latter is given a sign of "trust" and its confidence value is increased.

On the contrary, if there is no evidence of the association between an uncertain photo and any other one, the candidate link to that photo is not validated: the lack of coupling actions is considered as a sign of "distrust" and decreases the confidence value of the candidate link.

The same POI-photo candidate link is given as input to multiple users; each player action modifies the link confidence value, by increasing or decreasing it. After a variable number of played games, this confidence value crosses some thresholds, thus leaving its uncertainty status and becoming either a trusted link or an incorrect one.

When the confidence value of a link becomes greater than a given upper threshold (e.g., 70%), the POI-photo link becomes "trustable" and is inserted in the trusted link-set; in the figure, the green link on the right is associated to Milano's Duomo and it can be used to validate other candidate links. Similarly, if the confidence value of a POI-photo link becomes smaller than a lower limit (e.g., 20%), the link is discarded and is no more given as input to other players, like for the red one on the right of Figure 3 (in which the photo evidently depicts Roma's Colosseum).

4. EVALUATION

The evaluation of the UrbanMatch game is currently being performed and it is aimed at two different results: the assessment of the game "purpose" – the ability of our Human Computation approach to actually derive meaningful and quality links between Milano POIs and their related photos – and the appraisal of the game "playability" – the intrinsic fun or entertaining characteristic of the game.

Regarding the assessment of the game purpose, we identified a number of metrics to measure the UrbanMatch capability to improve the "fitness-for-use" quality [19] of the urban-related data involved in the game. More specifically, we measure the completeness and the accuracy of the new trusted links produced by the game.

We define *completeness* as the capability of the game to assess all the input candidate links, deciding if they are either trustable or incorrect. The completeness is calculated by dividing the number of assessed links (i.e. the links that became either trusted or incorrect after the gameplay) by the total number of input uncertain links.

We define *accuracy* as the capability of the game to make correct assessments about the input links, minimizing the "false positive" outcomes (i.e., POI-photo links considered trustable but actually incorrect) and "false negative" outcomes (i.e., POI-photo links considered incorrect but actually trustable). To measure the game accuracy, we need to know the ground truth, thus we manually check the assessed links to identify the false positive/negative items. The accuracy is then calculated by dividing the number of correct assessments (true positive and true negative items) by the total number of input uncertain links.

Our preliminary evaluation is based on an early set of played games: we collected evidences from 54 unique players (not including the development team), who played 290 games for a total of 781 levels, in which they tested 2,006 uncertain links. Setting the thresholds on the confidence value to 70% and 20% for the upper and lower limits respectively, the game assessed the correctness/incorrectness of 1,284 uncertain links, getting to an improvement of the global completeness from 1.54% to 4.98% with a final accuracy of 99.4% (4 false positive and 8 false negative links).

On the other hand, it is clear that an important success factor for Games with a Purpose lies in the gaming feature: the more engaging and entertaining the game, the higher the number of participant players and thus the collected data. For those reasons, we believe that an important part of the UrbanMatch evaluation consists in assessing the "playability" of the game itself.

As suggested by several studies about traditional games [9, 23] and taking into consideration the peculiarities of UrbanMatch, we built an evaluation questionnaire that UrbanMatch players can find at `http://bit.ly/um-survey`. The questions are oriented at assessing the game characteristics as well as "measuring" how much the purpose is hidden and immersed within the gameplay.

At writing time, we collected the feedbacks of 12 players whose opinion was quite positive: UrbanMatch was evaluated to be easy (91%) and clear (61%); most players "spread the word" suggesting their friends to play (64%) and supported our hypothesis that the physical presence in the urban environment makes the gameplay easier (55%).

5. CONCLUSIONS

While Linked Data research is continuously evolving and improving, new ways to interlink information from different independent sources are being formulated and explored. In this paper, we presented our UrbanMatch application, a mobile and location-based Game with a Purpose, oriented to

create high-quality links between existing datasets, namely OpenStreetMap, LinkedGeoData and Flickr. UrbanMatch is our first experiment to prove that mobile gaming applications can be successfully employed to consume, create and improve urban-related Linked Data; our early experience seems to confirm our research hypothesis, even if further evaluation is needed.

It is also worth noting that the data gathered via UrbanMatch have a clear business value: linked and ranked photos of places represent a valuable dataset which can be used to improve a number of services ranging from image search to geo-marketing. More generally, Games with a Purpose aimed at linking, collecting or correcting Linked Data related to urban spaces and Smart Cities constitute a potential business for a number of different stakeholders: local public authorities, local businesses (shops, transportation companies, tourism actors, etc.) and citizens.

Different datasets can be created or improved via Human Computation approaches for local services, trade, tourism, traffic optimization, environmental sustainability or simple information. This includes both urban related Linked Data and non-linked data such as municipality datasets or other private datasets related to Smart Cities. We believe that applications like UrbanMatch show a clear business model to be exploited with special regards to location-aware Linked Data.

A final note on the "gaming" approach: data linkage and content creation or assessment tasks need an active and productive engagement of users, but not all crowdsourcing campaigns are successful and effective. Shaping the Human Computation missions as gaming activities (like in Games with a Purpose) is a potential solution to this issue. A further research question could be oriented to prove that a funny or entertaining flavour can accomplish more than an extrinsic reward: finding a good balance between the game rules and the purpose achievement is often a hard task.

Acknowledgements

This work was partially supported by Siemens Corporate Research and Technologies and by the PlanetData EU project (FP7-257641).

6. REFERENCES

[1] G. Álvaro and J. Álvaro. A Linked Data Movie Quiz: the answers are out there, and so are the questions. `http://bit.ly/linkedmovies`, Aug. 2010.

[2] S. Auer, J. Lehmann, and S. Hellmann. LinkedGeoData - Adding a Spatial Dimension to the Web of Data. In *Proceedings of ISWC 2009*, 2009.

[3] I. Celino, D. Dell'Aglio, E. Della Valle, R. Grothmann, F. Steinke, and V. Tresp. Integrating Machine Learning in a Semantic Web Platform for Traffic Forecasting and Routing. In *Proceedings of IRMLES Workshop 2011*, 2011.

[4] I. Celino, D. Dell'Aglio, E. Della Valle, Y. Huang, T. Lee, S. Park, and V. Tresp. Making Sense of Location-based Micro-posts Using Stream Reasoning. In *Proceedings of #MSM Workshop 2011*, 2011.

[5] K. T. Chan, I. King, and M.-C. Yuen. Mathematical modeling of social games. In *Proceedings of IEEE CSE Conference 2009*, volume 4, pages 1205–1210, 2009.

[6] C. Charron, J. Favier, and C. Li. Social Computing – How Networks Erode Institutional Power, And What to Do About It. Forrester Research, 2006.

[7] E. Della Valle, I. Celino, and D. Dell'Aglio. The Experience of Realizing a Semantic Web Urban Computing Application. *Transactions in GIS*, 14(2), 2010.

[8] E. Della Valle, I. Celino, D. Dell'Aglio, F. Steinke, R. Grothmann, and V. Tresp. Semantic Traffic-Aware Routing for the City of Milano using the LarKC Platform. *IEEE Internet Computing*, 15(6), 2011.

[9] H. Desurvire, M. Caplan, and J. A. Toth. Using heuristics to evaluate the playability of games. In *Proceedings of CHI Conference 2004*, 2004.

[10] A. Flemming and O. Hartig. Quality Criteria for Linked Data sources. Online at `http://bit.ly/ld-quality`.

[11] T. Fullerton. *Game Design Workshop, A Playcentric Approach to Creating Innovative Games*. Gama Network Series. Morgan Kaufmann, 2008.

[12] C. Fürber and M. Hepp. Using SPARQL and SPIN for Data Quality Management on the Semantic Web. In *BIS*, pages 35–46, 2010.

[13] C. Fürber and M. Hepp. Towards a Vocabulary for Data Quality Management in Semantic Web Architectures. In *Proceedings of LWDM Workshop 2011*, March 2011.

[14] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool, 2011.

[15] P. Jain, P. Hitzler, P. Z. Yeh, K. Verma, and A. P. Sheth. Linked data is merely more data. In *AAAI Spring Symposium: Linked Data Meets AI*, 2010.

[16] S. A. Knight and J. Burn. Developing a framework for assessing information quality on the World Wide Web. *Informing Science*, 8, 2005.

[17] T. Lee, S. Park, Z. Huang, and E. Della Valle. Toward Seoul Road Sign Management on the LarKC Platform. In *Posters and Demos of ISWC 2010*, 2010.

[18] F. Naumann and C. Rolker. Assessment Methods for Information Quality Criteria. In B. D. Klein and D. F. Rossin, editors, *IQ*. MIT, 2000.

[19] T. C. Redman. *Data Quality: The Field Guide*. Digital Press, 2001.

[20] E. Simperl, B. Norton, and D. Vrandecic. Crowdsourcing Tasks within Linked Data Management. In *Proceedings of COLD Workshop 2011*, 2011.

[21] K. Siorpaes and M. Hepp. Games with a Purpose for the Semantic Web. *Intelligent Systems*, 23:50–60, 2008.

[22] K. Siorpaes and E. Simperl. Human Intelligence in the Process of Semantic Content Creation. *World Wide Web*, 13(1-2), 2010.

[23] P. Sweetser and P. Wyeth. Gameflow: a model for evaluating player enjoyment in games. *Computers in Entertainment*, 3(3):3, 2005.

[24] L. von Ahn. Games with a Purpose. *IEEE Computer*, 39(6):92–94, 2006.

[25] L. von Ahn. Human computation. In G. Alonso, J. A. Blakeley, and A. L. P. Chen, editors, *ICDE*, pages 1–2. IEEE, 2008.