Geospatial Dataset Curation through a Location-based Game

Description of the Urbanopoly Linked Dataset

Irene Celino

CEFRIEL – ICT Institute – Politecnico di Milano
via Fucini 2, 20133 Milano, Italy
E-mail: irene.celino@cefriel.it

Abstract. The Urbanopoly dataset contains the results of a data curation campaign on available geospatial open datasets like OpenStreetMap. The curation effort is conducted through a location-based Game with a Purpose inspired by the monopoly board game. The paper describes the dataset: we illustrate the genesis and life-cycle of Urbanopoly data; we explain the modelling choices by introducing the provenance-based Human Computation ontology and by giving examples of the dataset content; we describe the dataset publication on the Web as Linked Data and the cross-links to the curated datasets; finally, we indicate the possible uses of the dataset as well as its envisioned re-use.

Keywords: Geospatial Data, Data Curation, Human Computation, Volunteered Geographic Information, Provenance

1. Introduction

Wiki-like and collaborative approaches to collect geospatial information are on the rise, as testified by the popularity of Volunteered Geographic Information (VGI) [8] initiatives. The most celebrated example is OpenStreetMap[1], the free editable map of the world.

Additionally, geospatial datasets are increasingly present in government open data portal and in the Linked Data Cloud; notable examples are LinkedGeoData [18] – the linked data version of OpenStreetMap – and GeoLinkedData.es [19] – the Spanish initiative on geospatial linked data. Also ontologies and vocabularies like NeoGeo [17] and query languages like GeoSPARQL [15] are attracting a growing interest from the Semantic Web community.

In both cases of official government datasets and of collaboratively-collected information, geospatial datasets are not necessarily trustworthy and change over time. Thus, a data curation approach is required, on the one hand, for quality assurance and, on the other hand, to correct and update the dataset, in order to take data dynamics into account.

In this paper, we present the linked dataset resulting from a Human Computation-based data curation approach over pre-existing geospatial datasets. Data management tasks are embedded in a location-based Game with a Purpose that exploits players' physical presence in the environment. The remainder of the paper is organized as follows: Section 2 explains the Urbanopoly game; the resulting dataset is described in details in Section 3, with data sources, ontologies and modelling examples; Section 4 is devoted to illustrate the Urbanopoly dataset exploitation, in terms of dataset life-cycle and uses/re-uses of its content; fi-

---

[1]Cf. http://www.openstreetmap.org/.

nally, Section 5 concludes the paper with some fore-sights.

2. The Urbanopoly application

Human Computation [13] is the paradigm to leverage human capabilities to solve tasks that computers are not yet able to properly undertake. It is often used to address the quality assurance problem and Games with a Purpose (GWAP) [20] are employed to provide an entertaining incentive to the task solution. To be effective, a GWAP should be carefully designed (a) to provide an effective mechanism to address the Human Computation task and (b) to assure a continuous involvement and contribution of users/players.

Our research investigation is oriented to discover if the physical presence in the urban environment, together with location-based technologies, can provide a valuable contribution to Human Computation tasks related to geospatial information. While traditional Human Computation approaches exploit users' background knowledge or domain expertise, we argue that the direct experience and "human sensing" can play an important role in solving tasks related to the physical space. Thus, we built Urbanopoly, a mobile and location-based GWAP whose purpose is *quality assurance on geospatial (linked) data* by exploiting a social sensing approach via Human Computation.

From the gameplay point of view, Urbanopoly [5] is inspired by the monopoly board game[2]. Taking as input open geospatial datasets, Urbanopoly challenges its players to play mini-games in the form of questions, quizzes or quests in order to conquer venues and become a rich "landlord". The different mini-games are the expedient to insert different challenges within the app: some missions are data collection tasks, some other actions require to solve data validation tasks. An aggregation algorithm (cf. Section 4.1) combines players' actions to consolidate up-to-date and reliable information. The gameplay and the competition with friends, on the other hand, provide the long-term incentive for players.

The workflow of the Urbanopoly game is sketched in Figure 1. Geospatial data describing urban points of interest (POIs) are taken from open datasets (cf. Section 3.1) as starting point for the game. Players play Urbanopoly and, in order to be successful in the
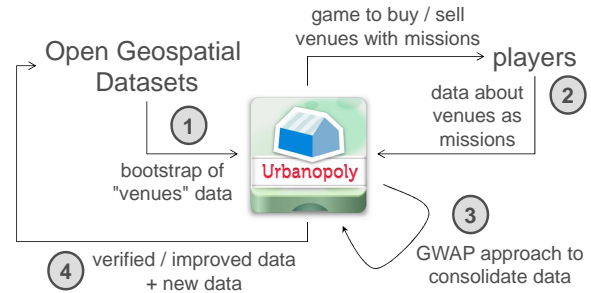

Fig. 1. High level view of the workflow of the Urbanopoly game.

game, they face different "mini-games" (cf. Figure 2): some challenges are aimed at validating the existing data from the original sources, other challenges require them to contribute new data. All players' contributions are collected and a GWAP approach is adopted to consolidate the different evidences by applying an aggregation algorithm. Finally, consolidated information is published together with provenance metadata as Linked Data [9], properly linked to the original datasets; those links allow for the extension and correction of those open geospatial sources with new or improved data (cf. Section 4.1).

The evaluation on the data curation results [6] of Urbanopoly is very good in terms of both precision/accuracy – around 92% – and engagement of players – the Average Life Play metrics (ALP [20], computed as ratio between the total played time and the number of active users) is around 100 minutes, which means that players enjoyed the game very much and returned several times to play it again.

3. The Urbanopoly dataset

This section is specifically devoted to the description of the Urbanopoly linked dataset. We specify the original sources of geospatial data taken into consideration; then, we illustrate the relevant ontologies – both pre-existing and developed on purpose – used in the dataset; finally, we give some examples of the entities described in the dataset.

3.1. Original data sources

As in the monopoly board game, the Urbanopoly player is a "landlord" whose aim is to create a rich portfolio of "venues"; those venues are real places in the surrounding of the player, like shops, restaurants, monuments, etc. The initial information about

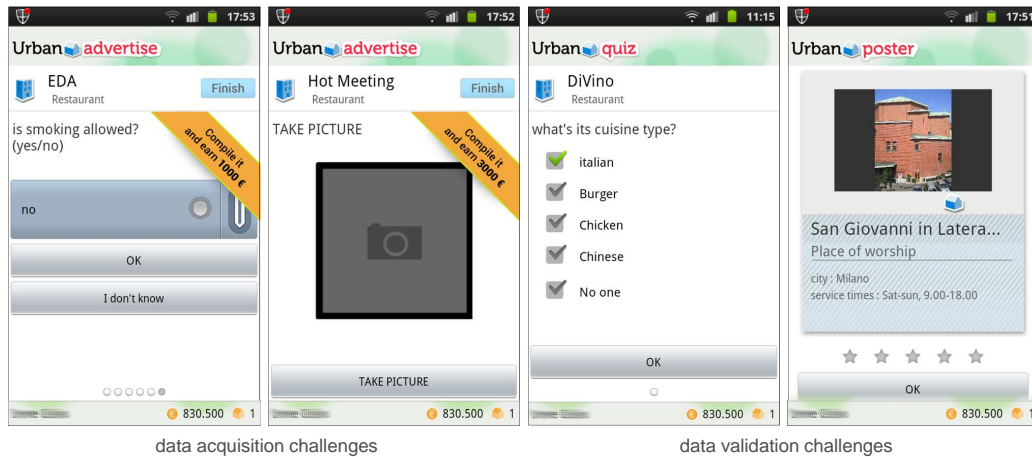data acquisition challenges   data validation challenges

Fig. 2. Screenshot of the Urbanopoly game showing the "mini-games" to acquire or validate data.

the venues was taken from available open geospatial data sources: a well-known VGI collaborative wiki effort – OpenStreetMap – and, for what regards the area around the city of Milano, the Open Data Portal[3] of "Regione Lombardia", the local regional public authority. In the initial dataset, geographic coordinates of venues are considered stable; all other properties of venues are collected or validated through the game.

Data from OpenStreetMap were obtained through LinkedGeoData [18], the linked data version of this VGI dataset. We selected a subset of classes representing POIs like shops, monuments, public transportation stops, etc. For each class we selected a number of properties to describe venues' features that Urbanopoly players can provide in the game: apart from name, category and basic address information – which are common features for all venues – restaurants are described by the cuisine type, bus stops by the line numbers, banks by the availability of an ATM, etc. Similarly, we retrieved from Lombardia Open Data Portal the information about "agriturismo" venues, quite popular in Italy. For each venue of this type, we collected their properties, like the services and products they offer to tourists.

This bunch of information is thus ready to be used in the game: the aim is to gather a high-quality set of triples in the form:

```
<venue> <feature> <value> .
```

in which the feature is the property and the value is its filler w.r.t. the venue.

We initially included 36,897 venues from Lombardia in Italy, then we added 6,749 venues from the Amsterdam area; finally, we included also 7,817 venues from Boston for a total of more than 50,000 venues.

Each venue in Urbanopoly is given a URI identifier with a namespace in our Web domain[4], so to ease the publication of Urbanopoly results as linked data (cf. Section 4.1). For what regards the data from OpenStreetMap/LinkedGeoData, we preserved the connection back to the original sources in the form of RDF links. More specifically, we created `owl:sameAs` relations to LinkedGeoData URIs and `rdfs:seeAlso` links to OpenStreetMap URLs (since OpenStreetMap identifiers relate to Web pages rather than to the POIs described on those pages), as shown in Listing 1.

```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix lgd: <http://linkedgeodata.org/triplify/> .
@prefix osm: <http://www.openstreetmap.org/browse/> .
@prefix u:
    <http://swa.cefriel.it/linkeddata/urbanopoly/> .

u:venue3116
    owl:sameAs    lgd:node959824653 ;
    rdfs:seeAlso  osm:node/959824653 .
```

Listing 1 Cross-links between the Urbanopoly dataset, LinkedGeoData and OpenStreetMap.

3.2. PROV-O and the Human Computation ontology

For the last years, the Semantic Web community has been working on the issue of provenance capture based on knowledge representation. In 2009, the W3C set up a Provenance Incubator Group, whose activity resulted in its final report [7]; given the promising results, that activity was turned in 2011 into an official W3C Working Group, which is standardising the PROV specification for provenance on the Web [1].

The PROV model is based on three main concepts – *entity*, *activity* and *agent* – and their relations (cf. Figure 3). The Provenance Ontology (PROV-O [2]) provides an ontological formalization of PROV in OWL [10].
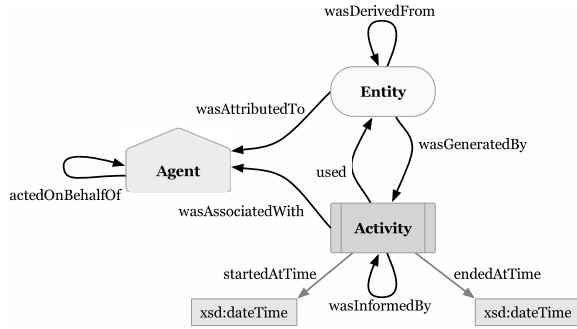

Fig. 3. Overview of the main PROV primitives (source: [2]).

We modelled a specialization of PROV-O that is specifically intended as the ontological formalization of provenance in relation to a Human Computation approach [13]. Figure 4 illustrates the main concepts and predicates of our Human Computation ontology.

The relevant entities are *contributions* – the outputs of human workers – and *consolidated information* – the result of the aggregation algorithm. The respective activities are *Human Computation tasks* – solved by the *contributor* agents – and the *Human Computation algorithm* that consolidates the information contributed by the human participants.

In Description Logics [12], we can formalize the subsumption relationships between our ontology primitives and PROV-O concepts as follows:

$$
\begin{aligned}
\text{hc:Contribution} &\sqsubseteq \text{prov:Entity} \\
\text{hc:ConsolidatedInformation} &\sqsubseteq \text{prov:Entity} \\
\text{hc:HumanComputationTask} &\sqsubseteq \text{prov:Activity} \\
\text{hc:HumanComputationAlgorithm} &\sqsubseteq \text{prov:Activity} \\
\text{hc:Contributor} &\sqsubseteq \text{prov:Agent}
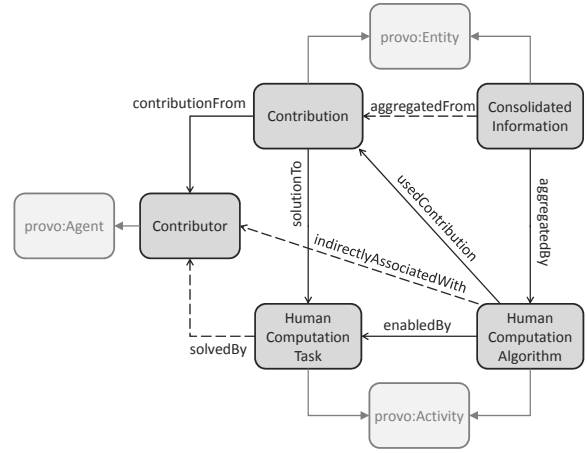\end{aligned}
$$

Fig. 4. Graphical representation of the Human Computation ontology (lighter grey arrows indicate subsumption, dashed arrows indicates derived relations).

in which the *hc* prefix is used to indicate terms from our Human Computation ontology, available on-line at http://swa.cefriel.it/ontologies/hc.

Similarly, the predicates in our ontologies can be derived from PROV-O properties as follows:

$$
\begin{aligned}
\text{hc:contributionFrom} &\sqsubseteq \text{prov:wasAttributedTo} \\
\text{hc:solutionTo} &\sqsubseteq \text{prov:wasGeneratedBy} \\
\text{hc:aggregatedBy} &\sqsubseteq \text{prov:wasGeneratedBy} \\
\text{hc:usedContribution} &\sqsubseteq \text{prov:used} \\
\text{hc:aggregatedFrom} &\sqsubseteq \text{prov:wasDerivedFrom} \\
\text{hc:enabledBy} &\sqsubseteq \text{prov:wasInformedBy} \\
\text{hc:solvedBy} &\sqsubseteq \text{prov:wasAssociatedWith}
\end{aligned}
$$

The dashed lines in Figure 4 indicates relations that can be derived using role composition (i.e., property chain axioms in OWL [10]). A Human Computation task was *solved by* a contributor if the contribution given as solution to the task was contributed by that agent:

$$
\text{hc:solvedBy} \sqsubseteq \text{hc:solutionTo}^{-} \\
\circ \ \text{hc:contributionFrom}
$$

An aggregation algorithm can be *indirectly associated with* a contributor if the algorithm was enabled by a task solved by that contributor:

$$
\text{hc:indirectlyAssociatedWith} \sqsubseteq \text{hc:enabledBy} \\
\circ \ \text{hc:solvedBy}
$$

Finally, consolidated information was *aggregated from* contributions if the aggregated data were generated by a Human Computation algorithm that used those contributions:

hc:aggregatedFrom ⊑ hc:aggregatedBy
∘ hc:usedContribution

3.3. Modelling examples from the Urbanopoly dataset

The full dataset is published on the Web according to the Linked Data principles [9]. A human-readable version of the dataset is browsable from `http://swa.cefriel.it/linkeddata/`. In the following, we provide concise examples of the data modelling using the ontologies presented in Section 3.2.

A *contribution* from a Urbanopoly player is modelled as illustrated in Listing 2 (identifiers are forged to make the example readable). Since the data collected are a player's contribution and the player could be wrong or could cheat, we cannot directly assert the specific statement; thus we make use of RDF reification [14] to express the actual content of the contribution.

```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix hc: <http://swa.cefriel.it/ontologies/hc#> .
@prefix vgi: <http://example.org/vgi#> .

# the individual contribution
vgi:MarioContribution123 a hc:Contribution;
  # it was provided by player 'Mario'
  hc:contributionFrom vgi:Mario;
  # it was created during the gameplay
  hc:solutionTo vgi:MarioUrbanopolyTaskABC;
  # it was collected in a specific moment
  prov:generatedAtTime
      "2013-03-06T14:00:00"^^xsd:dateTime;
  # the actual content of Mario's contribution
  vgi:providedInformation [
    # Mario is describing this POI
    rdf:subject vgi:CentralStation ;
    # Mario is giving information about the POI name
    rdf:predicate foaf:name ;
    # this is the POI name attributed by Mario
    rdf:object "Stazione di Milano Centrale" ;
  ];
.
```

Listing 2 Example of Urbanopoly contribution.

Consolidated information is created by Urbanopoly when different contributions are cross-checked – also with the data from the original source, if present; the Human Computation aggregation algorithm (cf. Section 4.1) combines the contributions from the Ur-

banopoly players into a consolidated set. Depending on the number of evidences and the reliability of players, the algorithm gives also a score to each piece of consolidate information. An example is illustrated in Listing 3.

```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix hc: <http://swa.cefriel.it/ontologies/hc#> .
@prefix vgi: <http://example.org/vgi#> .

# the algorithm is triggered by different
        contributions
vgi:AggregationAlgorithm
  a hc:HumanComputationAlgorithm;
  # it is enables by the gameplay of two players
  hc:enabledBy  vgi:MarioUrbanopolyTaskABC,
           vgi:LuigiUrbanopolyTaskDEF;
  # it uses the contributions from the players
  hc:usedContribution vgi:MarioContribution123,
               vgi:LuigiContribution456;
.


# the resulting aggregated information
vgi:AggregatedInformation
  a hc:ConsolidatedInformation;
  # it is produced by the algorithm above
  hc:aggregatedBy  vgi:AggregationAlgorithm;
  # the aggregation has a confidence score
  hc:confidence      "0.75"^^xsd:float;
  # it was computed in a specific moment
  prov:generatedAtTime
      "2013-03-07T08:20:00"^^xsd:dateTime;
  # the content is the same of the two contributions
  vgi:providedInformation [
    rdf:subject    vgi:CentralStation ;
    rdf:predicate  foaf:name ;
    rdf:object     "Stazione di Milano Centrale" ;
  ];
.
```

Listing 3 Example of Urbanopoly consolidated information.

Whenever the confidence score overcomes a "reliability" threshold, the consolidated information can be considered correct and thus can be explicitly asserted as shown in Listing 4.

A full auto-contained and commented example of the Urbanopoly dataset is available on-line at `http://bit.ly/prov-ex2`.

4. Exploitation of the Urbanopoly dataset

In this section, we first describe the life-cycle of the dataset (i.e., how data are generated, consolidated and

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix vgi: <http://example.org/vgi#> .

# the consolidated information can be asserted
vgi:CentralStation
    foaf:name  "Stazione di Milano Centrale" .
```

Listing 4 Sample Urbanopoly venue with consolidated information.

published as linked data) and then we give examples of further use of the dataset, in terms of queries and inferences and possibly unexpected reuse.

4.1. Dataset life-cycle

Urbanopoly users launch the mobile app at any time and then are allowed to play with the close-by venues, as detected by the positioning service of the mobile device. During the gameplay, each player has to face the data-centred challenges introduced in Section 2; for each solved mini-game, the player's device sends back the outcome to the Urbanopoly server, that stores the contributions, described according to the Human Computation ontology introduced in Section 3.2.

Urbanopoly applies its aggregation algorithm [4] to consolidate the contributions coming from different players; similarly to [3], Urbanopoly's algorithm harmonizes and combines contributions through a scoring function based on different elements: difficulty to provide the piece of data, player's reputation and distance to the venue at contribution time. When the computed score overcomes a threshold, the piece of data is considered "stable" and can be published as consolidated information, again according to the Human Computation ontology (cf. Listings 3 and 4).

The individual contributions and the consolidated information are then published on the Web at `http://swa.cefriel.it/linkeddata/` according to the Linked Data principles [9] as 5-star open data. Besides the machine-accessible data, we also provide a human-friendly data navigation with Pubby[5]. To decouple data processing from data access and to avoid unnecessary or undesired interference with the game, Urbanopoly updates the published linked data periodically. All data are published with an open data license (cf. Section 4.3) to respect the original sources' licenses.

---

[5]Cf. `http://www4.wiwiss.fu-berlin.de/pubby/`.

4.2. Use of the dataset

Since data are expressed in RDF with respect to an OWL ontology, it is possible to compute analysis and statistics by running SPARQL [16] (and possibly GeoSPARQL [15]) queries. For example, it is possible to express queries like: the most active contributors or the contributors whose inputs lead to the greatest amount of consolidated data; the information elements confirmed by the highest number of users or consolidated more recently; the locations from which users provide the highest number of contributions. Listing 5 presents some sample queries.

```
# top 3 most active contributors
PREFIX hc: <http://swa.cefriel.it/ontologies/hc#>
SELECT ?c
WHERE {
    ?c a hc:Contributor .
    ?x a hc:Contribution ;
       hc:contributionFrom ?c .
}
GROUP BY ?c
ORDER BY DESC(COUNT(DISTINCT ?x))
LIMIT 3
```

```
# top 3 contributors with the most consolidated data
PREFIX hc: <http://swa.cefriel.it/ontologies/hc#>
SELECT ?c
WHERE {
    ?c a hc:Contributor .
    ?d a hc:ConsolidatedInformation ;
       hc:aggregatedBy [
          hc:usedContribution [
             hc:contributionFrom ?c
          ]
       ] .
}
GROUP BY ?c
ORDER BY DESC(COUNT(DISTINCT ?d))
LIMIT 3
```

Listing 5: Sample SPARQL queries on the dataset.

The Human Computation ontology introduced in Section 3.2 defines also relations as role compositions. Thus, some of the queries above could be simplified or enabled by those simple inferences (e.g. the second SPARQL query in Listing 5 can be shortened by the use of the "*aggregated from*" or "*indirectly associated with*" relations, cf. Figure 4).

Moreover, given the queries and inferences proposed above, this dataset could be used to change, adjust or improve the game experience. For example, each Human Computation task is described with the

type of mini-game solved by the player; if from the dataset analysis, we discover that a specific player is very good at one mini-game and very bad at solving another type of challenge, the gameplay could be updated either to present that player with the mini-games he likes best (so to give a positive feedback and a stimulus to continue playing) or, on the contrary, to challenge the player to improve himself with the other mini-games (so to make the game more demanding and to keep the player attentive).

4.3. Re-use of the Dataset

The Urbanopoly dataset is released under an open data license, specifically under the Open Data Commons Open Database License (ODbL[6]). This license respects the original sources – both OpenStreetMap/ LinkedGeoData and open data from Lombardy region – and allows for a possible integration of Urbanopoly results back to those datasets. Since Urbanopoly is conceived as a data curation Game with a Purpose, its consolidated information should be used to update and improve the accuracy of the geospatial datasets it refers to. The cross-links to both OpenStreetMap and LinkedGeoData ease this process (cf. Section 3.1).

The publication of the Urbanopoly linked dataset with an open license has also another interesting consequence. We do not only publish the consolidated information, but also the individual players' contributions. This enables the comparison of the Urbanopoly aggregation algorithm with different algorithms and techniques. Thus, the dataset is openly available for any interested researcher.

5. Conclusions and Foresight

In this paper, we described the dataset produced by the Urbanopoly mobile and location-based Game with a Purpose. Since the goal of the Urbanopoly app is to quality check, verify, update and enrich existing geospatial datasets, the presented work can be seen as the result of data curation over OpenStreetMap (or LinkedGeoData); for those reasons, the Urbanopoly dataset is released with an open data license and is linked back to the original geospatial sources.

In our view, Urbanopoly is a successful case of a broader discipline that applies the power of Human Computation [13] to Citizen Science [11]. We name this research field Citizen Computation and we believe that it can bring effective tools for geospatial data curation by exploiting the physical presence of the contributors in the environment.

Acknowledgement

References

[1] Belhajjame, K., Deus, H., Garijo, D., Klyne, G., Missier, P., Soiland-Reyes, S., & Zednik, S. (2012a). PROV Model Primer. W3C Working Draft.

[2] Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., & Zhao, J. (2012b). PROV-O: The PROV Ontology. W3C Working Draft.

[3] Bishr, M. & Kuhn, W. (2007). Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In *AGILE Conf.*, LNG&C, pages 365–387. Springer.

[4] Celino, I. (2013). Human Computation VGI Provenance: Semantic Web-based Representation and Publishing. *IEEE Transactions on Geoscience and Remote Sensing (to appear)*.

[5] Celino, I., Cerizza, D., Contessa, S., Corubolo, M., Dell'Aglio, D., Della Valle, E., & Fumeo, S. (2012a). Urbanopoly – a Social and Location-based Game with a Purpose to Crowdsource your Urban Data. In *Proceedings of the 4th International Conference on Social Computing*, pages 910–913.

[6] Celino, I., Cerizza, D., Contessa, S., Corubolo, M., Dell'Aglio, D., Della Valle, E., Fumeo, S., & Piccinini, F. (2012b). Urbanopoly: Collection and Quality Assessment of Geo-spatial Linked Data via a Human Computation Game. Finalist to the 10th Semantic Web Challenge.

[7] Gil, Y., Cheney, J., Groth, P., Hartig, O., Miles, S., Moreau, L., & Pinheiro da Silva, P. (2010). Provenance XG Final Report. W3C Incubator Group Report.

[8] Goodchild, M. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, **69**, 211–221.

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

[10] Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., & Rudolph, S. (2009). OWL 2 Web Ontology Language. W3C Recommendation.

[11] Irwin, A. (1995). *Citizen science: a study of people, expertise and sustainable development*. Routledge.

[12] Krötzsch, M., Simancik, F., & Horrocks, I. (2012). A Description Logic Primer. *ACM Computing Research Repository (CoRR)*.

[13] Law, E. & von Ahn, L. (2011). *Human Computation*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

---

[6]Cf. `http://opendatacommons.org/licenses/odbl/`.

[14] Manola, F. & Miller, E. (2004). RDF Primer. W3C Recommendation.

[15] Open Geospatial Consortium (2011). OGC GeoSPARQL – A Geographic Query Language for RDF Data. Technical report.

[16] Prud'hommeaux, E. & Seaborne, A. (2008). SPARQL Query Language for RDF. W3C Recommendation.

[17] Salas, J. M. & Harth, A. (2012). NeoGeo Vocabulary Specification – Madrid Edition. Technical report.

[18] Stadler, C., Lehmann, J., Höffner, K., & Auer, S. (2011). LinkedGeoData: A Core for a Web of Spatial Open Data. *Seman-tic Web Journal*.

[19] Vilches-Blázquez, L. M., Villazón-Terrazas, B., Saquicela, V., de León, A., Corcho, O., & Gómez-Pérez, A. (2010). GeoLinked data and INSPIRE through an application case. In *Proceedings of the 18th SIGSPATIAL International Conference*, pages 446–449. ACM.

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