Geospatial Dataset Curation through a Location-based Game

Description of the Urbanopoly Linked Datasets

Editor(s): Jens Lehmann, University of Leipzig, Germany; Oscar Corcho, Universidad Politécnica de Madrid, Spain
Solicited review(s): Marta Sabou, MODUL University Vienna, Austria; Victor de Boer, VU Amsterdam, The Netherlands; Willem Robert van Hage, SynerScope B.V., The Netherlands

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

Abstract. The Urbanopoly datasets contain 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 generated datasets: 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 datasets' content; we describe the datasets' publication on the Web as Linked Data and the cross-links to the curated datasets; finally, we indicate the envisioned uses of the datasets as well as their possible re-uses.

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) [10] 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 [19] – the linked data version of OpenStreetMap – and GeoLinkedData.es [20] – the Spanish initiative on geospatial linked data. Also ontologies and vocabularies like NeoGeo [18] and query languages like GeoSPARQL [17] 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 datasets 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 and its data curation workflow at high level; Sections 3–5 detail the first three steps of the workflow, also discussing modelling choices and presenting explanatory examples; the publication, uses and possible re-uses of the described datasets are ex-

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

1570-0844/13/$27.50

plained in Section 6, as final step of the Urbanopoly data curation workflow; finally, Section 7 concludes the paper with some foresights.

2. Data Curation Workflow of Urbanopoly

Human Computation [15] 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) [21] 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 [6] 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". An aggregation algorithm 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 as starting point for the game. Players play Urbanopoly and, in order to be successful in the game, they face different "mini-games": some challenges are aimed at validating the existing data from the original sources, other challenges require them to contribute new data.

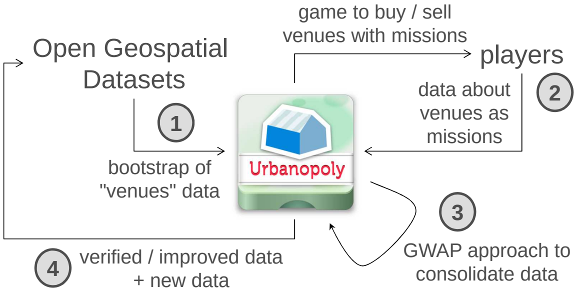
Open Geospatial Datasets
game to buy / sell venues with missions
players
data about venues as missions
2
1
bootstrap of "venues" data
Urbanopoly
3
GWAP approach to consolidate data
4
verified / improved data + new data

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

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 [12], properly linked to the original datasets; those links allow for the extension and correction of those open geospatial sources with new or improved data.

Each step of the workflow illustrated in Figure 1 is explained in the following sections in terms of used or generated datasets.

3. Step 1: Geospatial Input Datasets

As in the monopoly board game, the player is a landlord whose aim is to create a rich portfolio of "venues"; Urbanopoly venues are real POIs in the surrounding of the player, like shops, restaurants, monuments, etc. This section describes the input datasets employed to create the game's venues.

3.1. Original data sources

The initial information about the venues is 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 can be obtained through LinkedGeoData [19], the linked data version of this VGI dataset. We selected a set of classes representing POIs and, for each class, a number of properties to de-

[2] Cf. http://en.wikipedia.org/wiki/Monopoly_(game).

[3] Cf. https://dati.lombardia.it/.

scribe venues' features. Similarly, we retrieved from Lombardia Open Data Portal the information about "agriturismo" venues, quite popular in Italy. That kind of venues is not fully covered in OpenStreetMap, thus this open data source was a very nice addition to the initial dataset.

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.

3.2. Modelling of venues' categories, features and values

Since OpenStreetMap includes a very large and heterogeneous set of geospatial entities, we chose only a subset of the available data. We selected a restricted collection of 82 venues "categories" – like shops, monuments, public transportation stops, etc. –, picking those that we considered meaningful for the game play (i.e., worth of a monopoly-like "conquer"). Those categories are included in the Urbanopoly ontology, at `http://swa.cefriel.it/ontologies/urbanopoly`.

Each Urbanopoly category is linked both to the respective LinkedGeoData class(es) and to the key-value pair(s) used in OpenStreetMap to describe it. An example is given in Listing 1; since in most cases there is no one-to-one mapping between Urbanopoly categories and the original OpenStreetMap or LinkedGeoData sources, we preferred to re-define those categories in our data model[4].

```
@prefix lgdo: <http://linkedgeodata.org/ontology/> .
@prefix uo:
    <http://swa.cefriel.it/ontologies/urbanopoly#> .

uo:Museum a owl:Class ;
      rdfs:label "Museum"^^xsd:string ;
      # link to LGD classes
      rdfs:seeAlso
          lgdo:HistoricMuseum , lgdo:TourismMuseum ;
      # link to OSM key-value pair definition
      uo:comesFromOSMQuery
          "tourism=museum"^^xsd:string .
```

Listing 1 Modelling of venues' categories in Urbanopoly.

For each category we selected a number of 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.

To select the set of features of Urbanopoly, we adopted a mixed approach: after an automatic selection of the most frequent keys used in OpenStreetMap to describe the chosen venues' categories[5], we handpicked those that we considered meaningful for the game (i.e., those that can be provided by a player physically close to the venue while playing the game). In total, we modelled 107 features; on average each category has 11 features, with a minimum of 6 (corresponding to naming and addressing information) and a maximum of 16. Listing 2 presents a sample feature definition.

```
@prefix uo:
    <http://swa.cefriel.it/ontologies/urbanopoly#> .

uo:atm a  owl:DatatypeProperty ;
  # correspondence to OSM key
  uo:comesFromOSMTag "atm"^^xsd:string ;
  # domain and range definition (see also below)
  rdfs:domain
    [ owl:unionOf (uo:Bank uo:PostOffice) ] ;
  rdfs:range uo:ClosedDatatype_YesNo .
```

Listing 2 Modelling of features in Urbanopoly.

Finally, since Urbanopoly is a mobile game and players are expected to play on the move, we had the need to minimize the effort required to enter features' values when playing the game (cf. also Section 4). Therefore, we decided to provide – whenever possible – pre-defined values for the selected features. Some features can assume a closed set of values, while some others have an extensible range of values; to address this difference, we defined closed and semi-closed datatypes to be used as features' ranges. Listing 3 illustrate the case of yes/no values (a closed datatype). The player is always given the possibility to type in a custom value in case of semi-closed datatypes.

This features' values definition was the most challenging part of the modelling, since not only values depend on the feature, but they also change with respect to different places. For example, the "bank operator" feature – which indicates the name of the fi-

[4] In all listings, we do not display the definition of RDF, RDFS, OWL, XSD, etc. namespaces, for sake of readability.

[5] Interesting key/value statistics about OpenStreetMap are available at `http://taginfo.openstreetmap.org/`.

```
@prefix uo:
    <http://swa.cefriel.it/ontologies/urbanopoly#> .

uo:ClosedDatatype_YesNo a rdfs:Datatype ;
  rdfs:subClassOf uo:ClosedDatatype ;
  owl:equivalentClass [
    owl:oneOf ("yes"^^xsd:string "no"^^xsd:string)
  ] .
```

Listing 3 Modelling of a closed datatype for yes/no values.

nancial institution – takes different values on different locations: some banks are multi-national, thus it is possible to find them world-wide (e.g. Barclays), while some others are national or even regional, thus it would be better to suggest their name only to the Urbanopoly players in the specific area. To solve this issue, we created location-specific (at country level) semi-closed datatypes that include pre-defined lists of possible values; those values were derived from the most frequently used values in the location-specific dump of OpenStreetMap.

3.3. Representation and cross-linking of venues data

Each venue in Urbanopoly is given a URI identifier with a namespace in our Web domain (`http://swa.cefriel.it/linkeddata/urbanopoly/`), so to ease the publication of Urbanopoly results as linked data (cf. Section 6). 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 4.

```
@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 4 Cross-links between the Urbanopoly dataset, LinkedGeoData and OpenStreetMap.

4. Step 2: Urbanopoly Game Dataset

Urbanopoly users can 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. The aim of the Urbanopoly mini-games 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.

During the gameplay, each player has to face the data-centred challenges (shown in Figure 2). Data acquisition mini-games are useful when a value is missing; they require the player to insert a value or to pick a pre-defined option from a drop-down list. Data validation mini-games are useful to confirm or confute a previously-inserted value; in this case, players are required to select a value from a closed list (among which the value to be checked is inserted); from the player's point of view there is no great difference between acquisition/validation mini-games; it is the player's response that is treated in different ways by the Human Computation algorithm (cf. Section 5.1).

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 ontology introduced hereafter.

4.1. 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 [9]; 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 [13].

We modelled a specialization of PROV-O, available at `http://swa.cefriel.it/ontologies/hc`, that is specifically intended as the ontological formalization of provenance in relation to a Human Computation approach [15]. To our best knowledge, this is the first attempt to model a Human Computation ontology. We chose to model it by extending PROV-

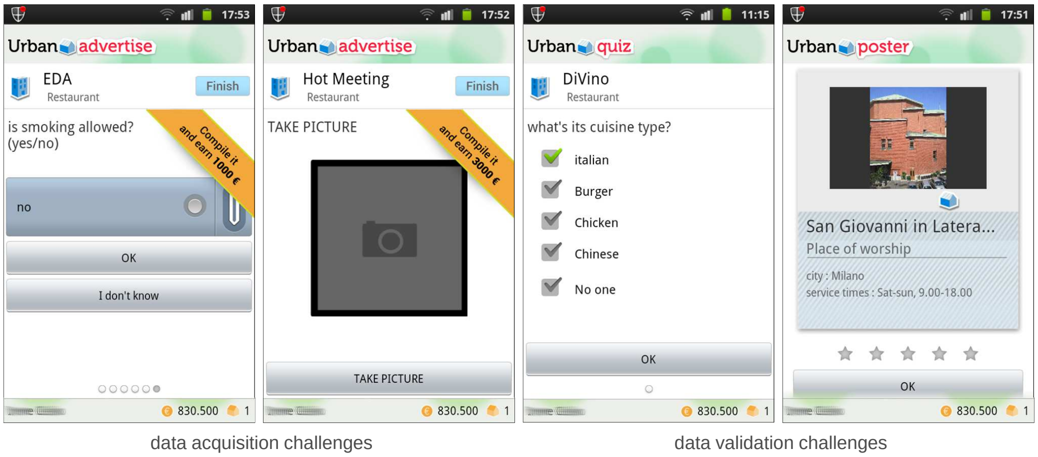

Fig. 2. Screenshot of the Urbanopoly game showing the "mini-games" to acquire or validate data.

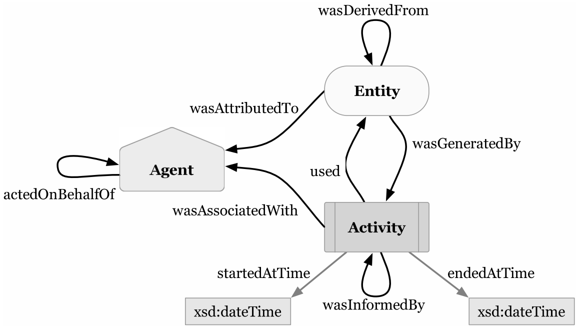

Fig. 3. Overview of the main PROV primitives (source: [2]).

O, because Human Computation approaches need to trace the provenance of the "entities" produced by the respective involved "agents". This ontology can be reused to describe any Human Computation effort, because it is not specific to our Urbanopoly game.

In the domain of data curation, probably the most popular ontology is the Data Quality Management Vocabulary (DQM [8]), that is aimed to represent "data requirements, data quality assessment results, data cleansing rules, and data requirement violations connected to their origin"; its application scope is "data quality monitoring, data quality assessment, and data cleansing"[6]. By addressing mainly automatic approaches to detect and correct data quality issues, DQM was not suitable to model our Human Computation scenario: it would be of course possible to extend DQM to include human-based data quality rules, but the DQM vocabulary do not (currently) include the modelling of the "agent" involved in the data quality task. For those reasons, we found PROV-O more appropriate for our modelling needs.

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.

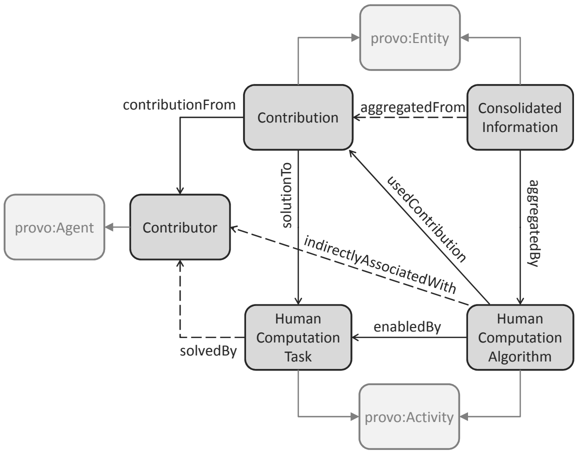

Fig. 4. Graphical representation of the Human Computation ontology (lighter grey arrows indicate subsumption, dashed arrows indicates derived relations).

Classes and properties of our Human Computation ontology are – respectively – sub-classes and sub-properties of the those defined in PROV-O. The dashed lines in Figure 4 indicates relations that can be derived

[6] Cf. http://purl.org/dqm-vocabulary/v1/dqm.

using role composition (i.e., property chain axioms in OWL [13]).

4.2. Examples of Urbanopoly Game data

Every time Urbanopoly players solve a mini-game, their *contribution* is recorded. A contribution is modelled by using Contribution, Contributor and Human Computation Task concepts and their predicates, as illustrated in Listing 5 (identifiers are forged to make the example readable). The example shows a contribution, its link to the player/contributor and the task/mini-game in which it was generated, the timestamp and the actual piece of geospatial information provided by that player.

```
@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 uo:name ;
    # this is the POI name attributed by Mario
    rdf:object "Stazione di Milano Centrale" ;
  ];
.
```

Listing 5 Example of Urbanopoly contribution.

Since the collected data are a player’s contribution and the player could be wrong or could cheat, we cannot directly assert the specific geospatial statement (i.e., the name of the venue); thus – at this stage – we make use of RDF reification [16] to express the actual content of the contribution.

5. Step 3: Human Computation and Geospatial Output Datasets

Recording players’ activity is not the only task for Urbanopoly. On the contrary, on the basis of the initial information (e.g., the original data from OpenStreetMap) and the contributions received during the gameplay, Urbanopoly “assesses” what the correct piece of information is, by applying a Human Computation aggregation algorithm on the collected data.

In this section we illustrate the generation of the main output datasets: the Human Computation information (that complements what illustrated in Section 4) and the consolidated geospatial information.

5.1. Aggregation algorithm

Human Computation approaches adopt an algorithm – usually named *aggregation algorithm* [15] – to decide when and how to combine the contributions from the human workers. Algorithms can be as simple as majority voting or as complex as an intricate method that keeps into account a wide set of parameters.

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 there are at least two evidences (two contributions from two different players or the original source information and a player’s contribution) and when the computed score overcomes a threshold, the piece of data is considered “correct” and can be published as consolidated information, again according to the Human Computation ontology, as explained below.

5.2. Examples of Human Computation output data

When Urbanopoly applies its aggregation algorithm, it evaluates the received contributions (cf. Listing 5) and produces the *consolidated information*, which is again represented by using our Human Computation ontology (cf. right part of Figure 4).

An example of consolidated information is illustrated in Listing 6: the aggregation is triggered by the contributions of two different players (Mario and Luigi) and the resulting information is annotated with a timestamp and a confidence score, that represents the result of the scoring function.

Again, the actual content of the consolidated information is represented via RDF reification, because until the confidence value overcomes a threshold the venue-feature-value statement cannot be asserted.

```
@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  uo:name ;
     rdf:object     "Stazione di Milano Centrale" ;
  ];
.
```

Listing 6 Example of Urbanopoly consolidated information.

A full auto-contained and commented example of the Human Computation output dataset is available online at `http://bit.ly/prov-ex2`.

5.3. Examples of Geospatial output data

As explained in the introduction, the main purpose of Urbanopoly is data curation over pre-existing geospatial data sources. Therefore, the most important output dataset of Urbanopoly is the set of curated geospatial information.

Whenever the confidence score overcomes a "reliability" threshold, the consolidated information is considered "correct" by Urbanopoly. The venue-feature-value statement can thus be explicitly asserted, instead of being reported only through reification, as shown in Listing 7.

```
@prefix vgi: <http://example.org/vgi#> .

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

Listing 7 Sample Urbanopoly venue with consolidated information.

As a consequence, the full information that Urbanopoly "knows" about the venue includes both the initial data (positioning and interlinking information) and the consolidated data; the venue full description included in the output geospatial dataset is thus in Listing 8.

```
@prefix lgd: <http://linkedgeodata.org/triplify/> .
@prefix osm: <http://www.openstreetmap.org/browse/> .
@prefix vgi: <http://example.org/vgi#> .

vgi:CentralStation
  # information from original sources
  geo:lat "45.4844939"^xsd:float ;
  geo:long "9.2029139"^xsd:float ;
  owl:sameAs    lgd:node91567650 ;
  rdfs:seeAlso  osm:node/91567650 ;
  # information "curated" by Urbanopoly
  uo:name  "Stazione di Milano Centrale" .
```

Listing 8 Full description of a sample Urbanopoly venue in the geospatial output dataset.

5.4. Evaluation of the aggregation algorithm

To assess the quality of the resulting geospatial dataset and, thus, of the Urbanopoly data curation approach, we conducted an evaluation campaign [7]. With regards to the precision of the consolidated data, we manually checked the correctness of the output of the aggregation algorithm, one month after the release of the game on Google Play; we restricted this manual check to those venues that either we knew very well or that we could "physically" go to and control. The computed accuracy is around 92%, which appears to be a very good result. To measure the engagement potential of the game, we considered a typical Game with a Purpose metrics, the Average Life Play (ALP [3], computed as ratio between the total played time and the number of active users). Urbanopoly ALP is around 100 minutes, which means that players enjoyed the game very much and returned several times to play the game.

6. Step 4: Publication, use and re-use of Output Datasets

The geospatial and Human Computation output datasets are published on the Web according to the Linked Data principles [12] as 5-star open data. Besides the machine-accessible data, we also provide a human-friendly data navigation with Pubby[7], browsable from `http://swa.cefriel.it/linkeddata/`.

Those datasets are released under an open data license, specifically under the Open Data Commons Open Database License (ODbL[8]). This license respects the original sources – both OpenStreetMap/ LinkedGeoData and open data from Lombardy region – and allows for the integration of Urbanopoly results back to those datasets (cf. Section 6.2).

To decouple data processing from data access and to avoid unnecessary or undesired interference with the game, Urbanopoly updates the published linked data periodically. There is no other specific versioning mechanism in place because, since all Human Computation data are time-stamped, the output data can be generated by (re)applying the aggregation algorithm to (a subset of) the collected contributions.

6.1. Statistics

At the time of writing, Urbanopoly collected 17,139 contributions from 182 active players and about 1,208 distinct venues (thus, more than 14 contributions per venue). The aggregation algorithm produced 2,351 consolidated statements (venue-feature-value triples), related to 634 distinct venues (thus, 52.5% of played venues, with an average of nearly 4 statements per venue). In total, the Human Computation dataset (contributions and consolidation) together with the geospatial output dataset contains about 280,000 triples.

Analysing the geospatial output dataset, other interesting facts can also be observed. The consolidated statements make use of 64 different features as triple predicates, out of the 107 features in our ontology (nearly 60%). The most frequent features are the venue “name” (27.7%), the venue “category” (12.9%), the denomination of the street where the venue is located (12.5%) and the “transport type”, i.e. the typology of public transport means that pass by a stop (e.g., bus, tram, subway; 4.5%). The frequency of those features indicates also that those are the easiest and least controversial data to collect: if a statement is consolidated, it means that (a) multiple players provided that contribution and (b) different players agreed on its meaning. Further evaluation details are included in the report [5] related to the results of the first two months after the Urbanopoly app release.

6.2. Re-use of the Geospatial Dataset

Since Urbanopoly is conceived as a data curation Game with a Purpose, its geospatial output dataset should be used to update and improve the accuracy of the original datasets it refers to. The cross-links to both OpenStreetMap and LinkedGeoData (cf. Listing 4) are included to ease this process.

The additional or modified data as consolidated by Urbanopoly can be sent back to OpenStreetMap, thanks also to the references to the keys and key-value pairs used in OpenStreetMap, as shown in Section 3.2. Thus, if Urbanopoly outputs a statement saying that a venue has the category `uo:Museum`, this information can be added back to OpenStreetMap with the `tourism=museum` key-value pair (cf. Listing 1); similarly, if Urbanopoly generates the information that a bank venue offers an ATM facility, that venue on OpenStreetMap can be enriched with the `atm=yes` key-value pair (cf. Listing 2).

While at this stage we do not have an automatic feedback mechanism to the original sources, it would not be difficult to implement it as described above. Contributing back to OpenStreetMap would be indeed meaningful to improve the accuracy of this dataset; our early evaluation [5] showed that, even if OpenStreetMap data is already of good quality, Urbanopoly is helpful because either it provides previously-missing details or it reflects data dynamics (e.g. a shop changing trade name).

6.3. Use of the Human Computation dataset

Since data are expressed in RDF with respect to an OWL ontology, it is possible to compute analysis and statistics by running SPARQL [11] (and possibly GeoSPARQL [17]) 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 9 presents some sample queries.

[7] Cf. `https://github.com/cygri/pubby`.

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

```
# 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 9 Sample SPARQL queries on the dataset.

The Human Computation ontology introduced in Section 4.1 defines also relations as role compositions. Thus, some of the queries could be simplified or enabled by those simple inferences (e.g. the second SPARQL query in Listing 9 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).

6.4. *Re-use of the Human Computation Dataset*

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.

The Human Computation dataset lends itself also to other kinds of analysis. With regards to the gathered information, the dataset can reveal what venues' features are the most popular or the easiest to collect or, on the contrary, the hardest to acquire or the most frequently mistaken.

Further analysis can be performed to discover the areas or places where players are more frequently located while playing, or to determine the time required to contribute different types of information, etc. This last type of analysis is better suited to understand and improve the mechanics of the contribution effort and, as a consequence, to investigate the role of the Urbanopoly game to address the geospatial data curation effort.

7. 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 geospatial dataset is released with an open data license and is linked back to the original geospatial sources. Our next step is to close the loop by automatically sending new or modified data back to OpenStreetMap.

In our view, Urbanopoly is a successful case of a broader discipline that applies the power of Human Computation [15] to Citizen Science [14]. 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

This work was partially supported by the PlanetData project, co-funded by the European Commission (FP7-257641). The author would like to thank Dario Cerizza, Marta Corubolo, Simone Contessa, Daniele Dell'Aglio, Emanuele Della Valle and Stefano Fumeo for the valuable contribution to the discussion.

References

[1] Belhajjame, K., Deus, H., Garijo, D., Klyne, G., Missier, P., Soiland-Reyes, S., & Zednik, S. (30 April 2013a). *PROV Model Primer*. W3C Working Group Note. Available at http://www.w3.org/TR/prov-primer/.

[2] Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., & Zhao, J. (30 April 2013b). *PROV-O: The PROV Ontology*. W3C Recommendation. Available at http://www.w3.org/TR/prov-o/.

[3] Bishr, M. & Kuhn, W. (2007). Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In S. I. Fabrikant and M. Wachowicz, editors, *The European Information Society, Proceedings of the 10th AGILE Conference*, Lecture Notes in Geoinformation and Cartography, pages 365–387. Springer (Berlin Heidelberg New York).

[4] Celino, I. (2013). Human Computation VGI Provenance: Semantic Web-based Representation and Publishing. *IEEE Transactions on Geoscience and Remote Sensing*, **51**(11), 5137–5144. IEEE.

[5] Celino, I., Cerizza, D., Contessa, S., Corubolo, M., Dell'Aglio, D., Della Valle, E., Fumeo, S., & Piccinini, F. (2012a). Deliverable D9.2 – UrbanGames Prototype and Evaluation. Technical report, PlanetData EU project. Available at http://www.planet-data.eu/.

[6] Celino, I., Cerizza, D., Contessa, S., Corubolo, M., Dell'Aglio, D., Della Valle, E., & Fumeo, S. (2012b). Urbanopoly – a Social and Location-based Game with a Purpose to Crowdsource your Urban Data. In A. Nijholt, A. Vinciarelli, B. Schuller, and M. Smith, editors, *Proceedings of the 4th International Conference on Social Computing*, pages 910–913. IEEE Computer Society (Washington, DC, USA).

[7] Celino, I., Cerizza, D., Contessa, S., Corubolo, M., Dell'Aglio, D., Della Valle, E., Fumeo, S., & Piccinini, F. (2012c). Urbanopoly: Collection and Quality Assessment of Geo-spatial Linked Data via a Human Computation Game. In D. Maynard and A. Harth, editors, *10th Semantic Web Challenge*. Available at http://challenge.semanticweb.org/2012/.

[8] Fürber, C. & Hepp, M. (2011). Towards a Vocabulary for Data Quality Management in Semantic Web Architectures. In R. De Virgilio, D. Bianchini, and V. De Antonellis, editors, *Proceedings of the 1st International Workshop on Linked Web Data Management*, pages 1–8. ACM (New York, NY, USA).

[9] Gil, Y., Cheney, J., Groth, P., Hartig, O., Miles, S., Moreau, L., & Pinheiro da Silva, P. (08 December 2010). *Provenance XG Final Report*. W3C Incubator Group Report. Available at http://www.w3.org/2005/Incubator/prov/XGR-prov/.

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

[11] Harris, S. & Seaborne, A. (21 March 2013). *SPARQL 1.1 Query Language*. W3C Recommendation. Available at http://www.w3.org/TR/sparql11-query/.

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

[13] Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., & Rudolph, S., editors (27 October 2009). *OWL 2 Web Ontology Language: Primer*. W3C Recommendation. Available at http://www.w3.org/TR/owl2-primer/.

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

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

[16] Manola, F. & Miller, E. (10 February 2004). *RDF Primer*. W3C Recommendation. Available at http://www.w3.org/TR/rdf-primer/.

[17] Matthew Perry, J. H. (2011). *OGC GeoSPARQL – A Geographic Query Language for RDF Data*. Open Geospatial Consortium. Available at http://www.opengeospatial.org/standards/geosparql.

[18] Salas, J. M. & Harth, A. (2012). *NeoGeo Vocabulary Specification – Madrid Edition*. GeoVocab.org. Available at http://geovocab.org/doc/neogeo/.

[19] Stadler, C., Lehmann, J., Höffner, K., & Auer, S. (2012). LinkedGeoData: A Core for a Web of Spatial Open Data. *Semantic Web Journal*, **3**(4), 333–354.

[20] 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 A. El Abbadi, D. Agrawal, M. Mokbel, and P. Zhang, editors, *Proceedings of the 18th SIGSPATIAL International Conference*, pages 446–449. ACM (New York, NY, USA).

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