Interaction Design and Architecture(s) Journal - IxD&A, N.31, 2016, pp. 43-58

Data Profiling in a Mobile Touristic Augmented Reality Application for Smart Environments based on Linked Open Data

Silviu Vert, Radu Vasiu

Multimedia Research Centre, Politehnica University of Timisoara, Piata Victoriei 2, 300006 Timisoara, Romania
{silviu.vert, radu.vasiu}@cm.upt.ro

Abstract. Data profiling is an important step in understanding the nature of the datasets that belong to the Web of Data. In this paper, we plan to analyze the appropriateness of exploiting this type of data in an augmented reality application to be used by tourists in smart environments. We build on top of previous work and we analyze the data used in a case study done on integrating several user-generated and governmental open datasets into a mobile augmented reality application called LOD4AR. The results show that the data found on the Web of Data is appropriate for augmented reality applications, despite its shallowness. Governmental open data, especially data that is following several guidelines specified in this paper, can complement the Web of Data and improve the overall integrated data.

Keywords: Mobile Augmented Reality, Linked Open Data, Data Profiling.

1 Introduction

The Web of Data has grown significantly in the past several years and has become a major source of information for various applications that use data extensively.

The exploitation of this kind of data is extremely prevalent nowadays in fields including smart learning and smart environments, due to its ability of adding a layer of "intelligence" to the learning ecosystems and its contribution, with the help of supporting technologies, to citizen and territorial development.

However, due to its open nature and hectic growth, especially of user-generated data, it is not easy to assess the precise structure and content of the Web of Data and its appropriateness in specific fields of interest.

Within the Web of Data, a major source of data comes these days from governments worldwide, in the form of public open data. While this data is published in a more organized way, it is still unclear if and how it complements and adds further value to the user-generated data and if, together, they provide a more appropriate source of information for the targeted domain.

This was also the challenge that we faced in our research on integrating these types of data in a mobile application that is based on augmented reality technologies and

which is intended for tourists exploring data-rich smart environments. Our previous work is described in more detail in section 2.

Section 3 is an overview of the current approaches in the research literature for profiling the Web of Data. We present our approach in section 4, where we show how to process the data to get comparable datasets and where we show detailed statistics on the structure and the content of the data in the analyzed datasets. Section 5 presents a comparison of similar applications, while section 6 contains a discussion of the statistics presented before. Section 7 concludes the paper.

2 Background on our previous work

This paper builds on work done previously by the authors. Our research started by acknowledging that the current mobile augmented reality applications are too limited in their content, which is rather static, comes from one source of information, usually a relational database in the network, and is poorly linked to further sources of information [1]. We identified and proposed Linked Data as a suitable form of overcoming these limitations in such applications but, along the way, we identified some challenges that need to be tackled in order to properly integrate linked data in mobile augmented reality applications, namely geodata integration, data quality, provenance and trust [2].

We proceeded with proposing a model for straightforward integration of linked open data in mobile augmented reality applications for tourists [3]. This model was applied in a case study, which implied the integration of two user-generated open datasets, namely DBpedia and LinkedGeoData, and one governmental open dataset, consisting in the list of museums in Romania, from the National Romanian Open Data Portal, using the LDIF powerful framework [4]. The consolidated dataset is exploited by the prototype of a mobile augmented reality application that we developed, named LOD4AR [5], which is based on the awe.js[1] library and, consequently, works entirely in the browser of the smartphone.

[1] https://github.com/buildar/awe.js/

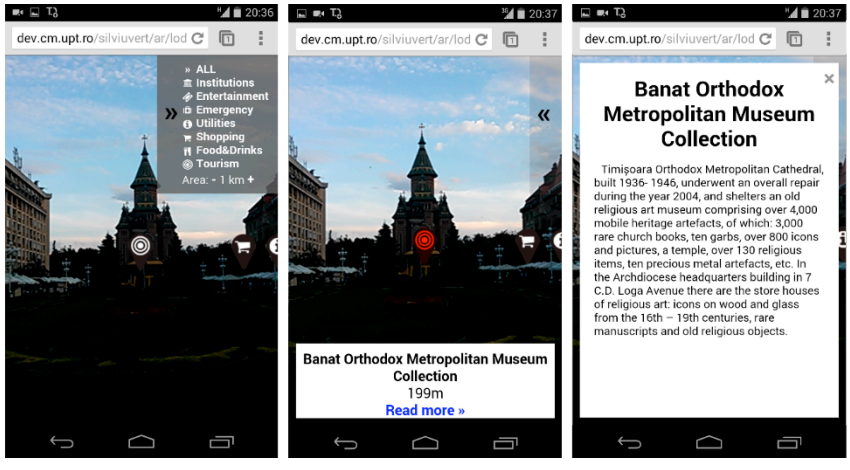

Fig. 1. Screenshots of the LOD4AR application

During the research, we manually analyzed the sources of data that we exploited to be able to choose the optimal parameters for the integration process. In this paper, we want to extend the analysis of the sources of data with some more detailed statistics to understand how well structured and complete these linked open datasets are for exploitation in an augmented reality application and to propose some recommendations for publishing of open data that is appropriate for augmented reality experiences.

3 Related work on profiling the Web of Data

Our desired analysis is part of the challenge of profiling the Web of Data that has been approached at some extent already in the research literature. Profiling data is one of the important prerequisites for data integration, along others such as query optimization, data cleansing, scientific data management and data analytics [6]. Profiling data is a much-needed task, because of the open nature of the web, which allows anyone to say anything online. Consequently, usually, there is no certainty about the content and the structure that one will find in the data.

Researchers are studying some typical challenges that are encountered in profiling the Web of Data [7]. This type of profiling differs in many aspects from profiling the classic relational data, for which well-established tools and methods already exist. Commercial tools for relational data include IBM InfoSphere Information Analyzer[2], Microsoft Integration Services[3] or Informatica Data Explorer[4]. In the Web of Data world, the usual means for describing a dataset are the voiD [8] vocabulary and the

[2] http://www-03.ibm.com/software/products/en/ibminfoinfoanal
[3] https://msdn.microsoft.com/en-us/ms141026.aspx
[4] https://www.informatica.com/content/dam/informatica-com/global/amer/us/collateral/data-sheet/data-explorer_data-sheet_7011.pdf

Semantic Sitemaps [9], but they do not describe the information extensively. What is missing is a more detailed insight into these datasets.

An early project dealing with profiling the Web of Data is RDFStats [10], which is a framework for generating statistics both on RDF documents and on SPARQL endpoints. The tool was published as an open source project and was built to be extensible in the future. Although it was primarily developed for the SemWIQ federator and optimizer [11], it can be integrated in other platforms as well.

Another early project is ExpLOD [12], which allows the user to explore summaries of RDF data in a dataset along with the interlinking of the dataset with others from the Linked Open Data cloud.

A project that tackles the bigger picture of the Web of Data is LODStats [13], a tool written in Python which calculates 32 different statistical criteria and was developed to be specifically interlinked with the Data Hub, so as to provide a big picture of the Web of Data. It covers quality analysis, coverage analysis, privacy analysis and link target identification.

The first web-based tool for profiling – and not only – the Web of Data is ProLOD++ [14], which is a successor of the ProLOD tool [15]. ProLOD++ is able to perform tasks related to profiling, mining and cleansing arbitrary datasets provided by the user of the tool. One of the basic operations is calculating the number of occurrences of distinct predicates, along with their values. A demo is available online[5] but does not seem to work at the moment with the preloaded datasets, and does not allow uploading of new datasets by the user.

A recent project was the winner of the Open Track at the 2014 Semantic Web Challenge[6] and is called RapidMiner Linked Open Data extension [16], as it is a software module built specifically for the RapidMiner platform[7], which is a powerful analytics platform for data analysis and not only. The extension allows importing of RDF data into the platform (via RDF dumps uploading or SPARQL endpoints querying) and using the wide range of RapidMiner operators to analyze the data in various ways, including generating statistics. The extension also allows one to extend the knowledge on local data by searching and linking to data in the Linked Open Data cloud.

In [17], the researchers propose an automated approach for generating structured profiles that describe the topics covered by linked datasets. These profiles are exposed in a format based on the Vocabulary of Interlinked Datasets (VoID[8]) and Vocabulary of Links (VoL[9]).

In the next section, we describe our approach in profiling the datasets exploited by our augmented reality application.

---

5 https://www.hpi.uni-potsdam.de/naumann/sites/prolod++/app.html
6 http://challenge.semanticweb.org/2014/winners.html
7 https://rapidminer.com/
8 http://www.w3.org/TR/void/
9 http://data.linkededucation.org/vol/

4 Profiling the datasets exploited in LOD4AR

This section starts by presenting a general methodology for data profiling in the industry. Next, we propose an approach to apply this general methodology for profiling and assessing appropriateness of linked datasets for augmented reality scenarios. The remaining part of the section is a report on the detailed application of the methodology on the targeted datasets and the results that were obtained.

4.1 Methodology for data profiling and assessment in augmented reality scenarios

The general methodology for data profiling specifies the following big steps [18]:

1. Prepare for the project
2. Prepare for the analysis
3. Extract and format the data
4. Sampling
5. Analysis

Applying this methodology for profiling and assessing the appropriateness of linked datasets for augmented reality scenarios demands some particularities to be taken into account.

Generally, the first two steps involve deciding on the scope of the activity, training the team, software setups etc. It is important to choose the correct type of software and the people with the right skills for extracting and analyzing linked data for exploitation in augmented reality scenarios. However, these two steps are not the focus of this paper, so we will not get into more details concerning them.

The third step, which consists in extracting and formatting the data, is crucial for correctly analyzing the data afterwards. The targeted linked datasets might be accessible in various ways, such as RDF, SPARQL, REST API or data dump. If the data profiling tool works offline, then it is necessary to create a dump of that data. In addition, the serialization format of the data can vary between RDF/XML, Turtle, N-Triples or JSON-LD. Because there are not so many tools for profiling directly the linked data, one might be constrained to extract only the necessary information from the dataset and save it in a common file format such as CSV. This way, one can use regular data profiling tools to analyze the data.

In augmented reality scenarios for tourism, it is advisable to do the data profiling on a certain geographic area, which is of interest for the tourist using the application. However, given the incompleteness issue of the Web of Data, it is not trivial to identify and download the POIs that are in a certain area.

The fourth step, sampling the data, is required for one to get decent processing times during analysis. Of course, sampling is compulsory only when the dataset is very large. This depends, as stated in the previous paragraph, on the size of the targeted geographic area.

The fifth step is the actual analysis of the data. The targeted linked datasets can be assessed from various points of view, one of them being how well they cover all the

necessary information for a touristical augmented reality application. In addition, because the application integrates data from multiple sources, it is important to measure how well the datasets complement each other.

Typically, content for an augmented reality application for tourists should cover as much as possible of the following categories of information:

- name (official name, nickname etc.)
- description (short description – for a snippet, long description – for full page information etc.)
- picture (thumbnail, full picture, album of pictures etc)
- contact (website, email, phone etc.)
- address (geographic coordinates, address, city, street, number etc.)
- category (structural type, functional type etc.)
- provenance (contributor, source of information, last updated etc.)
- other (accessibility, parking, etc.)

There are many types of information – predicates in RDF – in the Web of Data, but only a few show up in the data consistently, so they can be deemed important. One can ignore the predicates that appear for less than, for example, 1% of the total number of POIs.

4.2 The process for extracting the necessary data

The third step of the methodology consists in extracting and formatting the data. In this subsection, we show how the process was applied in the case of LOD4AR.

For building a consolidated dataset, we used information from DBpedia, LinkedGeoData and the Romanian National Open Data Portal. From the former two we chose the data about various POIs (Points of Interest) in Romania. From the latter one we chose a dataset that contains all the museums in Romania, as provided by the National Institute of Heritage.

The three sources of data were chosen based on a set of criteria that we focused on in this paper: the geographical area that they cover, the quality of the information and, most importantly, the reusability. All the data usage licenses are compatible with the Open Definition [19].

A major problem that we encountered was deciding, for DBpedia and LinkedGeoData, which Points of Interest belong to Romania and which not. The initial download of the information was done by getting all the data that was geotagged with GPS coordinates placed on a circular area centered on the geographical midpoint of the country. Given the shape of the country, it is obvious that in this way, many POIs were downloaded, POIs that belong to neighboring countries. It was not possible to download just the POIs that are inside the borders of Romania, on other criteria than the GPS position, as the information is user-generated and as such incompletely or wrongly tagged with a country name or code. In addition, we were not able to identify in the research literature a well-established method that allows downloading POIs from DBpedia or LinkedGeoData and that belong just to a single country.

To solve this issue, we used a three-step approach (just for for the LinkedGeoData dataset, but the process is similar for DBpedia) that is detailed below.

1) Downloading the data from the server. The downloaded dataset from LinkedGeoData is stored on our university's Sesame server, which has a SPARQL interface. Using a query on this interface, we generated a CSV file from the server, which had three columns: URI of the POI, latitude and longitude. Next, using the OpenRefine[10] tool, we cleaned the CSV file, removing the datatype notations, to be able to process the data further in a straightforward manner.

2) Processing the data to identify the correct country for each POI. To test that a pair of GPS coordinates describes a geographical point that is inside the borders of a country, we used an open source algorithm[11] published on GitHub. The code uses the World Borders dataset, which is available online[12] with a Creative Commons Attribution-Share Alike License.

This algorithm (the essential part being displayed in the code snippet below) runs through the previously cleaned CSV file and returns, for each pair of GPS coordinates, the country code that the corresponding geographical point belongs to, which it writes in another CSV file.

```
cc = countries.CountryChecker('TM_WORLD_BORDERS-0.3.shp')
country = cc.getCountry(countries.Point(float(lat),
float(long)))
```

A summary of the result of running the algorithm shows how many POIs were identified in each country: 4047 in Bulgaria, 3319 in Hungary, 3753 in Moldova, 143 in Poland, 23171 in Romania, 2225 in Serbia, 2690 in Slovakia, 4224 in Ukraine and 194 unidentified). The number of POIs that were retrieved but are in another country than Romania is quite high, so running this algorithm is clearly justified.

As it can be seen, the algorithm could not identify the country for 194 POIs. Because the number was small, we used a manual method to identify the country for these POIs. We overlaid these POIs on a map using Google Fusion Tables. It turned out that most of the unidentified POIs are located in Romania, on the seashore, so it seems that the algorithm has problems in identifying POIs that are in this geographical area. The several (few) POIs that were located in the neighboring countries were tagged by us manually with the correct country code. We tagged the rest of them with the Romanian country code using proper tools from OpenRefine.

3) Generating the country triples for Romania. We exported from OpenRefine a CSV file only with the POIs belonging to Romania. Using the previous CSV file and OpenRefine, we generated RDF triples for each POI, through which we stated that the POI has the country code of Romania (using the LinkedGeoData ontology). Below is an example of a line from the generated CSV file.

```
<http://linkedgeodata.org/triplify/node2498136757>
<http://linkedgeodata.org/ontology/country> "RO" .
```

[10] http://openrefine.org/
[11] https://github.com/che0/countries
[12] http://thematicmapping.org/downloads/world_borders.php

We uploaded this file to the Sesame server through the Sesame Workbench. The initial number of Romanian POIs was calculated by counting the POIs that were tagged in various ways as belonging to Romania (we manually analyzed the data to determine these country-specific triples). We found 891 POIs that had the properties *lgdo:is_in:country*, *lgdo:addr:country*, *lgdo:country* or *lgdo:is_in:country_code* with the values "Romania" (for the first one) and "RO" for the latter ones.

The resulting number of Romanian POIs, after importing the RDF triples which were generated based on the location of the POIs, is 23360, which is more than 25 times the initial value (which was determined, as explained in the previous paragraph, as being 891, based on the user-provided country).

The technique described above can be used to generate correct country tags for all the geotagged POIs, either for DBpedia or for LinkedGeoData, and thus improve these widely used sources of information.

Having the POIs' location correctly identified, we proceeded with analyzing the structure and the content of the information, which corresponds to step five in the general methodology, which is data analysis. The fourth step of the methodology has been skipped, as the dataset obtained previously is small enough to not require sampling.

We calculated some statistics for all the three sources of information and only for the POIs that are placed in Romania, as determined by using the steps mentioned above.

4.3 Statistics for DBpedia

The DBpedia Ontology (dbo) is a shallow, multi-domain ontology which was extracted from Wikipedia through hand-made rules [20]. As a result of replicating Wikipedia, DBpedia contains information about POIs that are rather well-known in their area and are, according to internal rules, notable subjects [21]. As such, it mostly contains information about important institutions and touristic venues, and less about shopping places or utilities.

Analyzing the RDF predicates in the DBpedia dataset, to determine how well the POIs are described, we selected several of them that are significantly interesting for an augmented reality application and categorized them as such: name, description, picture, contact, address, category and provenance. The detailed list of predicates and their occurrences is presented in Table 1.

Table 1. Categorization and number of occurences of RDF predicates in DBpedia

Category	Description	Occurrences	Predicate
Name	label	3417	http://www.w3.org/2000/01/rdf-schema#label
	name	706	http://xmlns.com/foaf/0.1/name
	name	553	http://dbpedia.org/property/name
	official name	122	http://dbpedia.org/property/officialName
	other name	60	http://dbpedia.org/property/otherName

Description	comment	856	http://www.w3.org/2000/01/rdf-schema#comment
	abstract	856	http://dbpedia.org/ontology/abstract
Picture	thumbnail	439	http://dbpedia.org/ontology/thumbnail
	depiction	439	http://xmlns.com/foaf/0.1/depiction
	photo collection	832	http://dbpedia.org/property/hasPhotoCollection
Contact	website	157	http://dbpedia.org/property/website
	homepage	224	http://xmlns.com/foaf/0.1/homepage
	external link	377	http://dbpedia.org/ontology/wikiPageExternalLink
	wikipedia	856	http://xmlns.com/foaf/0.1/isPrimaryTopicOf
Address	city	136	http://dbpedia.org/ontology/city
	address	28	http://dbpedia.org/ontology/address
	address	28	http://dbpedia.org/property/address
	latitude	3564	http://www.w3.org/2003/01/geo/wgs84_pos#lat
	longitude	3564	http://www.w3.org/2003/01/geo/wgs84_pos#long
Category	type	1002	rdf:type
	type	474	http://dbpedia.org/ontology/type
	subject	856	http://purl.org/dc/terms/subject
Provenance	wiki page ID	856	http://dbpedia.org/ontology/wikiPageID

Of course, there were a lot more properties that could be included in these categories (we found 1694 predicates in total). However, in general, we ignored the predicates that occurred extremely infrequently (in less than 1% of the number of POIs that were geotagged, which is 3564).

In DBpedia, while almost all of the POIs have a name, only a minority of them have a shorter (abstract) or longer (comment) description. Even fewer have attached a photo of the POI, an important asset for an augmented reality application, as a photo greatly helps the user to identify the POI that she is searching for. The *address* category lacks information on the actual address of the POI (street, number, house), so the only reliable information consists of the GPS coordinates. For the *contact* category, we notice that no information is included on email or phone numbers, as it almost inexistent in DBpedia. For the *provenance* category, we considered the property *http://dbpedia.org/ontology/wikiPageID*, which points to the ID of the Wikipedia page where the information was generated from. Starting from here, one can find out, theoretically, the user(s) that created the information. As it turns out, this is a very indirect and rather unusable provenance information for an augmented reality application. The *category* category shows that only about half of the POIs are categorized somehow, a fact that hinders one of the most important aspects of a good augmented reality feature, the filtering option.

4.4 Statistics for LinkedGeoData

The LinkedGeoData Ontology (lgdo) is a lightweight OWL ontology that was crafted based on the tags existent for each POI in LinkedGeoData, which work as key-value pairs for the pieces of information [22]. As a result of LinkedGeoData replicating the data from OpenStreetMap, it usually contains information about a great variety of POIs, from ATMs to museums, information which is, however, rather shallowly described.

Analyzing the predicates in the LinkedGeoData dataset, to determine how well the POIs are described, we selected several of them that are significantly interesting for an augmented reality application and categorized them as such: name, description, address, accessibility and parking, contact, category and provenance. The detailed list of predicates and their occurrences is presented in Table 2.

Table 2. Categorization and number of occurences of RDF predicates in LinkedGeoData

Category	Description	Occurrences	Predicate
Name	label	15736	rdfs:label
Description	comment	347	rdfs:comment
Address	city	1591	http://linkedgeodata.org/ontology/addr%3Acity
	street	2432	http://linkedgeodata.org/ontology/addr%3Astreet
	house number	2116	http://linkedgeodata.org/ontology/addr%3Ahousenumber
	latitude	23360	http://www.w3.org/2003/01/geo/wgs84_pos#lat
	longitude	23360	http://www.w3.org/2003/01/geo/wgs84_pos#long
Accessibility and parking	parking	408	http://linkedgeodata.org/ontology/parking
	wheelchair	386	http://linkedgeodata.org/ontology/wheelchair
Contact	homepage	474	foaf:homepage
	phone	463	foaf:phone
	opening hours	600	http://linkedgeodata.org/ontology/opening_hours
Category	type	23360	rdf:type
	type (ro)	109	http://linkedgeodata.org/ontology/_Tip_
Provenance	contributor	23360	http://purl.org/dc/terms/contributor
	last modified	23360	http://purl.org/dc/terms/modified
	source of info	2132	http://linkedgeodata.org/ontology/source
	link to source	465	http://linkedgeodata.org/ontology/source%3Alink

Of course, there were a lot more properties that could be included in these categories (we found in total 483 predicates). However, in general, we ignored the predicates that occurred extremely infrequently (in less than 1% of the number of POIs that were geotagged, which is 23360).

LinkedGeoData features many predicates that vary only slightly by name (e.g. *short_name*, *_Nume_*, *_nume_*, *old_name%3Aen* etc.) which is a result of the fact that users can add tags as they wish when describing a POI in OpenStreetMap. There are no properties for images or pictures of the POIs. The *address* (except GPS coordinates) and *contact* related properties show up rather infrequently. More than half of the POIs have some kind of label, although very few have a description. It is interesting to note the presence of some small pieces of information on the accessibility of the POIs. Contrarily to DBpedia, all the POIs are categorized and provenance is well described is terms of contributor and date of last modification (for all POIs) and source of information and link to it (for some POIs).

4.5 Statistics for the museums dataset from the Romanian Open Data portal

To publish this governmental dataset as linked open data, we employed some parts of the FOAF and Basic Geo vocabularies, as well as a part of the DBpedia ontology, along with custom defined properties.

Analyzing the predicates in the museums dataset, we similarly selected several of them that are significantly interesting for an augmented reality application and categorized them as such: name, description, address, contact and category. The detailed list of predicates and their occurrences is presented in Table 3.

Table 3. Categorization and number of occurences of RDF predicates in the museums dataset

Category	Description	Occurrences	Predicate
Name	label	951	rdfs:label
Description	comment	951	rdfs:comment
Address	city	951	http://tom7.cm.upt.ro/onto/cityvalue
	address	673	http://tom7.cm.upt.ro/onto/addressvalue
	latitude	951	http://www.w3.org/2003/01/geo/wgs84_pos#lat
	longitude	951	http://www.w3.org/2003/01/geo/wgs84_pos#long
Contact	website	534	foaf:homepage
	phone	730	foaf:phone
	opening hours	834	http://tom7.cm.upt.ro/onto/hoursvalue
	email	451	foaf:mbox
Category	type	951	rdf:type

The dataset is well described, with names and descriptions for all 951 geotagged POIs (in total there were 967 POIs in the dataset). Half or more than half have addresses, websites, phones, emails and opening hours specified. All the information is generally given in both Romanian and English.

As a result of the fact that the dataset was released specifically as the list of museums in Romania, the category information is clear: all the POIs are museums. In

addition, as it was published on the National Open Data Portal, this is a good indication, although not complete, of its provenance.

5 Comparison with similar applications

Several projects based on augmented reality visualization techniques have tackled the integration of linked open data sources, mainly from general knowledge repositories. A short overview on their scope and exploited data is shown in Table 4.

Table 4. Overview of similar projects from the international landscape

Project	Year	Scope	Exploited data
Cultural heritage mobile guide [23]	2010	Providing cultural heritage resources for an end user	General knowledge platforms (GeoNames, LinkedGeodata, Freebase, DBPedia) and platforms specialized on cultural heritage (e.g. Art and Architecture Thesaurus, Union List of Artist Names) or platforms of individual cultural institutions
Smart Reality [24]	2012	Young people interested in listening to music and attending concerts	Play.fm (more than 18000 DJ mixes and live recordings); other sources crawled, starting from the URI defined by person who is annotating the poster
Mobile mountain guide [25]	2012	Visualizing mountain-specific data	Geonames, LinkedGeoData
ARCAMA-3D [26]	2013	Generic surroundings discovery focusing on topic experiences	Direct linking to DBpedia

These projects relate on their findings in terms of challenges and approaches in integrating linked open data in augmented reality applications. However, none of them gives a detailed overview of the structure, content and appropriateness of the integrated data for an augmented reality-based application.

6 Discussion

DBpedia and LinkedGeoData are two of the backbones of the Web of Data, in general, and of the world of geo linked data, in particular. They complement each other, due to the scope for which their original counterparts were created: Wikipedia features rich information on notable POIs, while OpenStreetMap strives to equally cover smaller and bigger POIs, although not that deep. Geonames is also regarded as a big player in this field; however, we considered it not to be that interesting for an

augmented reality application for tourists, as it mostly contains information about administrative regions in Romania (counties, cities, villages etc.).

Due to the nature of the open content, these big user-generated hubs of information expose data that is rather unpredictable, incomplete and error-prone.

Except the textual information, these platforms lack the really useful elements for an interactive and eye-catching augmented reality application: images, videos or 3D content. While a small part of the POIs feature an image, videos or 3D content is almost non-existent.

GPS data is poor on user-generated portals of information, a fact which is clearly linked to the technical difficulty for the common contributor in obtaining richer GPS information. Augmented reality applications rely on 3D models of the buildings for proper identification and augmentation of the surroundings, while these big platforms lack even GPS coordinates for the boundaries of the POIs (the only information provided is a single pair of GPS coordinates, as if the POI is just a point on the ground).

On the other side, the governmental dataset is well built, with more complete information, at least for the properties that it features. Also, it is inherently more reliable as a source of information due to its governmental origins. However, the range of properties that such information has is rather limited, as only some small data is usually collected by the government. It also lacks the same useful elements for a good augmented reality application.

Based on the previous statistics and on the discussion above, we can conclude that datasets should contain more specific GPS information (at least the boundary and the height), with a 3D model of the POI being the ideal target, they should feature more photos and videos, as interactive elements, and they should include URIs of the same objects as they are described in the Linked Open Data cloud, for proper linking and information crawling. These guidelines for publishing (governmental) open data suitable for exploitation in augmented reality applications should lead to better and more useful applications for the end user.

7 Conclusions

Integrating linked open data in mobile augmented reality applications has certain benefits, most of them related to the removal of the limitations imposed by classical databases that are used nowadays in augmented reality applications.

In this paper, to assess the appropriateness of the linked open data for augmented reality applications, we proceeded to profile the data in order to get an overview about the structure and the content of the data sources.

To put the effort in context, we reviewed the research literature on profiling the Web of Data, which is clearly needed due to the fact that metadata is very shallow or non-existent in linked datasets. Specific challenges of profiling this data include heterogeneity of vocabularies and performance times for large datasets.

The literature review did not reveal techniques for profiling data that report on the appropriateness of it for exploitation in augmented reality applications. Consequently, we proposed a methodology for assessing this appropriateness based on known data

profiling techniques. We described this methodology, noting the foundation that it is based on and the criteria that we look for in the Web of Data.

To apply the methodology to the datasets that we exploited in the augmented reality application described in the previous section, LOD4AR, we first needed to make sure that the datasets covered the exact same geographic area, so we could make a fair comparison between them. We proposed and described a process for obtaining the data for just one country, in our case Romania, from DBpedia and LinkedGeoData.

Afterwards, consistent with the methodology process, we generated statistics for the properties that occur in the data and categorized them on criteria of relevance for augmented reality applications.

We found that the user-generated hubs of information have shallow GPS data, only a small part features photos (videos and 3D content are non-existent), categorization varies in quality and coverage, hindering proper filtering of information, provenance information depends on how the linked source of information was built, and other information, such as contact, is only partially present.

Open data from governmental sources, although well-built and more reliable, still lacks information necessary for a good and useful augmented reality application, but can complement nicely the user-generated information.

After analyzing our own sources of data, we proceeded to compare LOD4AR with other similar projects, from the point of view of the exploited data. We note there is not enough information in the literature to be able to assess the quantity of the data being exploited in similar projects and, in general, there is almost no assessment about the appropriateness of the data for augmented reality applications, except some experiences of dealing with real-world data.

As further work, we intend to also do a profiling of the final integrated, consolidated dataset and to derive in this way optimal parameters for proper integration of the individual datasets.

Acknowledgments. This work was partially supported by the strategic grant POSDRU/159/1.5/S/137070 (2014) of the Ministry of National Education, Romania, co-financed by the European Social Fund – Investing in People, within the Sectoral Operational Programme Human Resources Development 2007-2013.
This work is based on the PhD Dissertation defended at the Politehnica University of Timisoara on 18 September 2015 by Silviu Vert, under the supervision of Radu Vasiu, and published in the University Dissertation series by the University's printing house.

References

1. Vert S., Vasiu R.: Integrating Linked Data in Mobile Augmented Reality Applications in Dregvaite, G. and Damasevicius, R. (eds.) Information and Software Technologies. vol. 465. pp. 324–333. Springer International Publishing (2014)
2. Vert S., Vasiu R.: Relevant Aspects for the Integration of Linked Data in Mobile Augmented Reality Applications for Tourism in Dregvaite, G. and Damasevicius, R.

(eds.) Information and Software Technologies. vol. 465. pp. 334–345. Springer International Publishing (2014)
3. Vert S., Vasiu R.: Integrating Linked Open Data in Mobile Augmented Reality Applications - a Case Study TEM JOURNAL - Technology, Education, Management, Informatics, 4, (2015)
4. Schultz A., Matteini A., Isele R., Mendes P.N., Bizer C., Becker C.: LDIF-A Framework for Large-Scale Linked Data Integration 21st International World Wide Web Conference (WWW 2012), Developers Track, Lyon, France (2012)
5. Vert S., Dragulescu B., Vasiu R.: LOD4AR: Exploring Linked Open Data with a Mobile Augmented Reality Web Application Proceedings of the ISWC 2014 Posters & Demonstrations Track, within the 13th International Semantic Web Conference (ISWC 2014). vol. 1272. pp. 185–188. , Riva del Garda, Italy (2014)
6. Naumann F.: Data profiling revisited ACM SIGMOD Record, 42, pp. 40–49 (2014)
7. Jentzsch A.: Profiling the Web of Data ISWC-DC 2014 Doctoral Consortium at ISWC 2014, pp. 32
8. Alexander K., Hausenblas M.: Describing linked datasets-on the design and usage of void, the'vocabulary of interlinked datasets In Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (WWW 09. Citeseer (2009)
9. Cyganiak R., Stenzhorn H., Delbru R., Decker S., Tummarello G.: Semantic sitemaps: Efficient and flexible access to datasets on the semantic web, Springer, (2008)
10. Langegger A., Woss W.: Rdfstats-an extensible rdf statistics generator and library Database and Expert Systems Application, 2009. DEXA'09. 20th International Workshop on. pp. 79–83. IEEE (2009)
11. Langegger A., Woss W., Blöchl M.: A semantic web middleware for virtual data integration on the web, Springer, (2008)
12. Khatchadourian S., Consens M.P.: Explod: Summary-based exploration of interlinking and rdf usage in the linked open data cloud The Semantic Web: Research and Applications. pp. 272–287. Springer (2010)
13. Auer S., Demter J., Martin M., Lehmann J.: LODStats–an extensible framework for high-performance dataset analytics Knowledge Engineering and Knowledge Management. pp. 353–362. Springer (2012)
14. Abedjan Z., Gruetze T., Jentzsch A., Naumann F.: Profiling and mining RDF data with ProLOD++ 2014 IEEE 30th International Conference on Data Engineering (ICDE). pp. 1198–1201 (2014)
15. Bohm C., Naumann F., Abedjan Z., Fenz D., Grutze T., Hefenbrock D., Pohl M., Sonnabend D.: Profiling linked open data with ProLOD Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference on. pp. 175–178. IEEE (2010)
16. Ristoski P., Bizer C., Paulheim H.: Mining the web of linked data with rapidminer International Semantic Web Conference (ISWC) (2014)
17. Fetahu B., Dietze S., Pereira Nunes B., Antonio Casanova M., Taibi D., Nejdl W.: What's all the data about?: creating structured profiles of linked data on the web Proceedings of the companion publication of the 23rd international conference on World wide web companion. pp. 261–262. International World Wide Web Conferences Steering Committee (2014)
18. Business Data Quality Ltd: Data Profiling Best Practices, http://www.dvbi.ru/portals/0/DOCUMENTS_SHARE/ARTICLES/34_0_Data_Profiling_Best_Practices_BDQ.pdf, (2010)
19. Open Knowledge: The Open Definition, http://opendefinition.org/
20. Bizer C., Lehmann J., Kobilarov G., Auer S., Becker C., Cyganiak R., Hellmann S.: DBpedia-A crystallization point for the Web of Data Web Semantics: Science, Services and Agents on the World Wide Web, 7, pp. 154–165 (2009)

21. Wikipedia: Notability, http://en.wikipedia.org/w/index.php?title=Wikipedia:Notability, (2015)
22. Stadler C., Lehmann J., Höffner K., Auer S.: Linkedgeodata: A core for a web of spatial open data Semantic Web, 3, pp. 333–354 (2012)
23. Van Aart C., Wielinga B., Van Hage W.R.: Mobile cultural heritage guide: location-aware semantic search Knowledge Engineering and Management by the Masses. pp. 257–271. Springer (2010)
24. Nixon L.J.B., Grubert J., Reitmayr G., Scicluna J.: SmartReality: Integrating the Web into Augmented Reality Presented at the I-SEMANTICS (Posters & Demos) (2012)
25. Zander S., Chiu C., Sageder G.: A computational model for the integration of linked data in mobile augmented reality applications Proceedings of the 8th International Conference on Semantic Systems. pp. 133–140. ACM (2012)
26. Aydin B., Gensel J., Genoud P., Calabretto S., Tellez B.: An architecture for surroundings discovery by linking 3D models and LOD cloud Proceedings of the Second ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems. pp. 9–16. ACM (2013)