A Survey on Environmental Open Data in Austria

Peter Wetz
Doctoral College Environmental Informatics
Vienna University of Technology
Vienna, Austria
peter.wetz@tuwien.ac.at

Amin Anjomshoaa, A Min Tjoa
Institute of Software Technology and Interactive Systems
Vienna University of Technology
Vienna, Austria
{anjomshoaa, amin}@ifs.tuwien.ac.at

Abstract— **Environmental Open Data has the potential to become a key strategy for the global endeavor of preserving the world's ecosphere. Therefore, this paper aims at researching the status quo of such data in Austria and exploring related advantages and shortcomings. First, the presented survey is motivated and put in perspective to relevant related work. Then, the approach to gaining the datasets is described and important parameters are specified. Subsequently, the results of the conducted survey are presented. Positive as well as negative highlights are pointed out. Many datasets with good features have been identified, however, there are still some drawbacks concerning data quality, diversity and content. Dealing with these issues and ultimately providing Linked Open Data would greatly improve the datasets and, ultimately, would allow for significantly better utilisation thereof. The goal of this work is to bring the concept of Open Data into the environmental domain by evaluating the environmental Open Data landscape in Austria.**

Keywords- Open Data; Linked Open Data; Environmental Informatics

I. Introduction

In recent years, problems concerning our environment rose to an extent, which cannot be ignored anymore. There are many different areas in which public awareness must rise in order to prevent irreversible damage to our planet. Such areas include, but are not limited to, concerns about territories, water, air, waste, noise, energy and disasters. The annual report of the United Nations Environment Programme (UNEP), called Global Environment Outlook (GEO), points out that the *"World Remains on Unsustainable Track Despite Hundreds of Internationally Agreed Goals and Objectives"* [1] and [2].

See Table I for Earth Systems, which are under risk according to UNEP. Still, in some cases positive progress has been made.

The concept of Open Data has the potential to partially deal with these global problems by providing a means of making certain data available to the public, and therefore enabling bigger public awareness about our environment. Open Data is the definition of data being freely available for everyone, so that people can look at the data and also reuse and republish it, e.g. as an application. The main idea is to provide data-transparency and also the possibility to create value out of publicly available data. Bizer et al. [3] prove that published governmental data is not of high quality, if published in the Linked Data format. They show, that there is a significant problem in the amount of outgoing links of published governmental Linked Data. Only four percent of the stored triples contain links to other data sources, showing that the main principle of Linked Data is not implemented very well. The main principle is the creation of as many meaningful links to other data as possible to enable reusage of such data.

TABLE I. Earth Systems under Risk

Earth System	Problem Description
Atmosphere	Ozone, Lead in Gasoline, Climate Change, Air Pollution
Biodiversity	Access and Benefit Sharing, Protected Areas, Fish Stocks
Water	Water Quality and Quantity, Ground Water Depletion, Integrated Water Management
Marine Pollution	Number of coastal dead zones, eutrophication, marine litter
Extreme Events	Climate change-induced disasters, floods, drought
Land	Access to food, desertification, deforestation
Chemicals and Waste	Pesticides, radioactive waste, persistant organic pollutants, heavy metals, electronic and electrical waste

Linksvayer [4] defines Open Data as follows: *"A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike."*. Additionally, the Open Data Handbook introduces Open Data as a possibility to answer questions of public interest, e.g. *"What is in the air that you breathe along the way?"* or *"What is the shortest, safest and most scenic bicycle route from your home to your work?"* [5]. Open Data mostly is published by governments since the data they hold is of biggest quantity and centrality, and most of it is public data by law. For instance, Obama, the president of the U.S.A., signed an Executive Order to push future release of U.S. Government Data [6].

Besides Environmental Issues and Open Data there is also a third important term which needs to be addressed. Smart City is a strategy which means that a city utilises its human, social and environmental capital to an extent, which helps in making the city more competitive in certain regards, one of them being the environmental aspect. For instance a city can be made *smart*, if sensors, which measure certain environmental indicators, are placed throughout the city in order to inform the population about the current environmental state. The characteristics and

978-1-4799-0652-9/13/$31.00 ©2013 IEEE

specification of a Smart City are not well defined, yet, but there is consensual thinking that the concept shall be adopted in the future to improve, amongst others, environmental aspects of urban living. The use case of the city Santander can be seen as best practise example in terms of planning and realising the Smart City vision [7].

This work aims to bring the concept of Open Data into the environmental domain by surveying and evaluating the Environmental Open Data landscape in Austria. Additionally, related advantages and shortcomings are explored. As a contribution an aggregated overview over a number of available datasets and their features is presented. The general results and some special datasets are discussed in detail. Furthermore conclusions are drawn and future work will be proposed. The future goal is to establish Linked Open Data (LOD) datasets to enable even better exploitation of available information; hence Section II deals with Related Work in the context of LOD; Section III presents the used approach; Section IV illustrates the results and Section V concludes by discussing the results and proposing future work.

II. Related Work

In the field of LOD some research has already been done with the same target as presented and discussed in this work; i.e. the utilisation of LOD to create more transparency for the population.

Use cases in nonenvironmental domains include Martin et al. [8], who transform Open Data of the European Commission to LOD to gain financial transparency of EU project funding. Furthermore they show the possibility of novel queries on the data supported by the concept of LOD. Hage et al. [9] present an approach converting Open Government datasets of piracy reports to RDF, linking them to external resources, exposing them on the Web and enabling easy analysis and visualisations to answer domain specific questions.

Lopez et al. [10] move more towards Urban Information Management and apply Semantic Web Technologies to integrate heterogeneous data sources to finally provide a Linked Data Platform for querying city data. Also [11] join the concept of Smart Cities with LOD by creating a *Game with a Purpose* in order to validate links between points of interest and their photos. Their approach proves high throughput and accuracy strongly affirming the motivation to leverage open datasets related to urban environments. Zapico et al. [12] argue that it is necessary to bring Open Data concepts to environmental impact information, yielding higher transparency, openness and easier creation of sustainability services on top of it. They present footprinted.org, a web service trying to build upon their presented ideas. The group shares their experiences during the development and report on first usages. Making use of sensor data Phuoc and Hauswirth [13] take the approach of providing an infrastructure specialised on real-time data based on LOD principles. This work is promising since the generated real-time data can then be processed in novel ways, which are not possible with usual static datasets.

As one can see, there has already been done some academic work in the field of Environmental Linked Open Data. Approaches range from publishing Environmental Data as LOD, to realising the Smart City vision by means of Semantic Web Technologies and even utilising sensor data to provide innovative services to citizens. This work can be seen as a first step in diving into this field by surveying the status quo of Environmental Open Data in Austria and proposing ideas to generate significant progress in this area.

III. Approach

The conducted survey was done on eleven available Open Data dataset sources from Austria. All of them can be retrieved via the Austrian Open Data portal[1]. Specifically, all available Open Data portals of Austria were looked up and environmental datasets were identified by belonging to the category “Environment”. The datasets then were investigated on different parameters partly sourced from [14]. This reference states that government data shall be considered open if the data complies with the following principles, which, amongst others, consequently were used as parameters for this survey:

- Data must be complete
- Data must be primary
- Data must be timely
- Data must be accessible
- Data must be machine-processable
- Access must be non-discriminatory
- Data formats must be non-proprietary
- Data must be license free

In addition to that, also the compliance with the well-known five star deployment scheme was evaluated [15]. This scheme has been introduced to assess the quality of LOD datasets. The quality levels are defined as follows:

- Data availability on the web with an open license.
- Data availability as machine-readable structured data.
- Data availability in non-proprietary formats such as CSV, XML, etc.
- Using W3C’s open standards (RDF and SPARQL) to identify entities.
- Linking data to other people’s data to provide data context.

Finally, some general evaluation parameters were defined to allow for more precise analysis of the datasets. These include the following characteristics:

- Data type: This property describes the nature of the data. This can be geodata, Image data, measured data, informational data and statistical data.

[1] http://www.data.gv.at

- Content type: This property describes the data from an environmental point of view. It defines the field, in which the data can be categorised: territory, air, water, waste, energy and geology.
- Static vs. real-time: This property determines, if the data has static or real-time character. The difference here basically is the update frequency. If the data is updated less than once a day, it is considered as static.
- Data format: This property holds the file format in which the dataset is made available.
- Date published: The date when the dataset was first published.
- Age of data: This property describes how old the stored data is; this value, therefore, does not necessarily need to correlate with date published.
- Number of records: This field determines how many data points are available in the dataset. This is basically difficult to determine in a comparable way, since the data is available in different formats and describes different entities. Additionally, some data can be queried via an Application Programming Interface (API) and therefore the input-parameters influence the amount of data points which are returned, resulting in a non-constant amount of provided data points. The initial evaluation of this parameter is described here for the record. It is not used in the final evaluation of the dataset.
- Size of dataset: This field represents the file size of the datasets. Similar comparability issues as above apply here. The initial evaluation of this parameter is described here for the record. It is not used in the final evaluation of the dataset.
- Update frequency: This field describes how often the dataset is updated, as long as this is determined in the dataset description

IV. Results

First, we show some general results regarding the data collection: Out of the eleven data sources, *Federal Environmental Office* (26), *Province of Styria* (25) and *Vienna* (33) have most datasets published in the environmental domain. *Province of Carinthia* is the newest addition, since they released the Beta version of their Open Data portal in June 2013. The first published Open Data portal is *Vienna*, which was launched in May 2011. All datasets coherently are licensed via the "Creative Commons – Attribution 3.0 Austria" license.

Table II shows the discovered data sources, which were subsequently investigated.

TABLE II. Environmental Data Sources

Data source	Date of publication	Dataset count	%
Federal Environmental Office[2]	June 2012	26	19
Federal Geological Institute[3]	February 2013	2	1
Province of Upper Austria[4]	February 2013	4	3
Province of Lower Austria[5]	April 2013	21	15
Province of Styria[6]	April 2013	25	18
Province of Vorarlberg[7]	June 2012	1	1
Province of Tyrol[8]	November 2012	13	9
Province of Carinthia[9]	June 2013	2	1
City of Vienna[10]	May 2011	33	24
City of Graz[11]	April 2012	5	4
Municipality of Großengerwitzdorf[12]	November 2012	7	5
Total	-	139	100

The following tables show aggregated results of all data sources of environmental Open Data datasets from Austria. It is important to note that the city of Linz also has an Open Data portal, but it does not provide environment-specific datasets.

Table III shows the absolute counts of datasets according to data types and content types. Data types are categorised as following:

- Geo: geographically encoded point-, vector- or area-data.
- Image: data, represented as an image.
- Measured: data, which is measured (e.g. sensor data) and presented as plain data values.
- Informational: data, which provides organisational information, e.g. opening hours.

[2] http://www.umweltbundesamt.at/opendata/

[3] http://data.gv.at/auftritte/?organisation=geologische-bundesanstalt

[4] http://www.doris.at/fachinfo/natur/genisys/

[5] http://data.noe.gv.at/Land-Zukunft/Open-Government-Data/Umwelt.html

[6] http://data.steiermark.at/cms/beitrag/11822068/95633323

[7] http://data.vorarlberg.gv.at/ogd/umwelt/umwelt.htm

[8] http://www.tirol.gv.at/applikationen/e-government/data/datenkatalog/umwelt/

[9] http://www.kagis.ktn.gv.at/278870_DE-GEODATEN-OGD-Kaernten_%28beta%29

[10] http://data.wien.gv.at/katalog/umwelt/

[11] http://data.graz.gv.at/daten/umwelt

[12] http://www.data.gv.at/auftritte/?search-term=umwelt&organisation=gemeinde-engerwitzdorf

TABLE III. ANALYSIS OF DATASETS ACCORDING TO DATA TYPES AND CONTENT TYPES

Data type / Content type	*geo*	*image*	*measured*	*informational*	*statistical*	*Total*
territory	81	12	-	-	-	93
air	-	-	23	-	-	23
water	-	-	1	-	2	3
waste	12	-	-	3	2	17
energy	-	-	-	-	2	2
geology	1	-	-	1	-	2
Total	94	12	24	4	6	139

- Statistical: data, which represents statistical information, e.g. energy consumption over the past ten years.

Content types declare the distribution of datasets according to different environmental aspects, which include the following categories:

- Territory: data, which holds information about environmental territory, e.g. nature protection areas.
- Air: information about air, e.g. air quality measures
- Water: information about water, e.g. water level measures
- Waste: information about waste, e.g. amount of produced waste.
- Energy: information about energy, e.g. energy consumption rates.
- Geology: information about the solid earth, e.g. tectonic geodata.

It can be observed that most datasets offer geographical data as a data type (94). This is geospatial data in form of vectors, areas or points, mostly defining nature (protection) zones or specific points of interest. Measured data means basically sensor data, which is updated very frequently, i.e. mostly on an hourly basis. Only four out of eleven data sources (Federal Environmental Office, Province of Styria, Province of Lower Austria, Province of Carinthia) provide this type of measured data, with two of them (Federal Environmental Office, Province of Carinthia) providing solely one dataset of that kind. The remainder of the dataset consists of images (12), statistical data (6) and informational data (4).

Regarding content types Table III also shows that most datasets contain territorial data (93), meaning that these datasets describe the landscape. For instance, this can be nature protection zones or park areas in urban regions. There is also much data containing air quality measures (23), but these datasets are nearly only provided by two sources: *Styria* and *Lower Austria*.

Additionally, a very important indicator of Open Data datasets is the compliance to the five star scheme [15]. As Table IV shows, most of the datasets are three star datasets (84%), a minority of them being only two star compliant (16%). Not a single dataset is four star compliant or above. The three star datasets provide data on the Web under an open license, while making them available as structured data in a non-proprietary format. The reasons for the two star compliant datasets is in one case the usage of Microsoft Excel (xls) as a file format, which is proprietary, and in the other case the data is provided as an image (jpg), therefore being not available in a structured format.

TABLE IV. COMPLIANCE TO FIVE STAR SCHEME

Star count	Dataset count	%
★	0	0
★★	22	16
★★★	117	84
★★★★	0	0
★★★★★	0	0

Concerning compliance to Open Data principles of governmental data, Fig. 1 shows, how many datasets implement the seven mentioned principles. It can be clearly seen that Austrian governmental data meets most of these requirements to a high degree, with only a few outliers not adhering. Sometimes timeliness is an issue; this is when the data itself is outdated and already some years old.

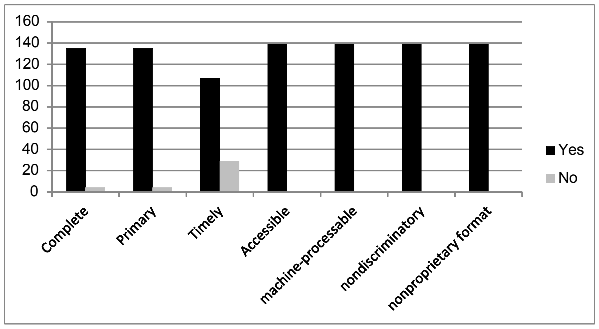

Figure 1. Compliance to Open Data principles

V. DISCUSSION AND FUTURE WORK

As the results show, most datasets contain geographically encoded data. This data is really well processable, but in its nature the data is static and therefore does not need to get updated on a frequent basis. In contrast, measured data, which is offered fairly often (24), obviously is more interesting, because of its real-time nature. Applications which are built on top of such data can be used in a much more flexible way than applications dealing with data which does not change over time. The high amount of published Geodata by each provider suggests that this kind of data is most easily disseminated as Open Data. In contrast to that, we suggest that more datasets containing real-time data and measured data of environmental aspects would significantly increase the value of Open Data portals. Furthermore a higher diversity of content types is recommended. As already stated and motivated in the introduction, there are many more environmental issues than compared to the present coverage at the time of writing this paper. For instance, there is no Open Data covering water or land quality issues.

Concerning adherence to the five star scheme it can be stated that four and five star compliance is recommended. Not a single surveyed dataset reaches this level. Four star compliance implies that each entity in the dataset is denoted by a Uniform Resource Identifier (URI). This would enable linking to or from it on the Web. Furthermore, the data could be bookmarked, reused and integrated with other data, finally leading to five star compliance, resulting in even more enhanced data integration, because of additional linking of the data to external sources. This would yield great advantages, since the data then is discoverable, increasing the value of the published information to great extent. The data should be pushed to a higher star deployment scheme level. As already initially stated, currently the open governmental data is not well linked at all, with only four percent of outgoing links even in LOD domain. This shortcoming proves that more research towards semi-automatic ways of interlinking this data is required.

Finally, the compliance to the Open data principles can be improved by providing better information about timeliness and update frequency to each dataset. Often, information regarding these parameters is missing.

For future work we see the creation of at least four star compliant environmental Open Data datasets as a goal. The more data is available as LOD in the Austrian Environmental sector the more the public and also governments will be able to draw value out of it. Different kinds of applications can be built on top of these datasets, even incorporating very different kinds of datasets by leveraging the principles of LOD. An important aspect of this goal is to take care of high linkage of the environmental datasets to other datasets, to further motivate reuse and creation of more use cases.

Other aspects of future work will be prototypical examination of crowdsourcing efforts to provide Urban Environmental Data. Such approaches are already discussed in academia and seem to create promising results [16, 17].

ACKNOWLEDGMENT

This research was partially funded by the Vienna University of Technology through the Doctoral College Environmental Informatics.

REFERENCES

[1] UNEP, *GEO-5 Assessment.* Available: http://www.unep.org/geo/pdfs/geo5/GEO5-Global_PR_EN.pdf (2013, Jun. 14).

[2] UNEP, *GEO-5 Assessment Full Report.* Available: http://www.unep.org/geo/pdfs/geo5/GEO5_report_full_en.pdf (2013, Jun. 14).

[3] C. Bizer, A. Jentzsch, and R. Cyganiak, "State of the LOD Cloud," *Version 0.3 (September 2011)*, 2011.

[4] Mike Linksvayer. Available: http://opendefinition.org/.

[5] Open Knowledge Foundation. Available: http://opendatahandbook.org/pdf/OpenDataHandbook.pdf (2013, Jun. 14).

[6] White House - Office of the Press Secretary, *Obama Administration Releases Historic Open Data Rules to Enhance Government Efficiency and Fuel Economic Growth.*

[7] J. M. Hernández-Muñoz, J. B. Vercher, L. Muñoz, J. A. Galache, M. Presser, Gómez, Luis A Hernández, and J. Pettersson, "Smart cities at the forefront of the future internet," in *Domingue, Galis et al. (Ed.) 2011 – The Future Internet*, pp. 447–462.

[8] Michael Martin, Claus Stadler, Philipp Frischmuth, Jens Lehmann, "Increasing the Financial Transparency of European Commission Project Funding," *Semantic Web Journal*, no. 1, pp. 1–5, http://www.semantic-web-journal.net/system/files/swj435.pdf, 2012.

[9] Willem Robert van Hage, Marieke van Erp, Véroniqu Malaisé, "Linked Open Piracy: A Story about e-Science, Linked Data, and Statistics," (English), *Journal on Data Semantics*, vol. 1, no. 3, pp. 187–201, http://dx.doi.org/10.1007/s13740-012-0009-6, 2012.

[10] V. Lopez, S. Kotoulas, M. Sbodio, M. Stephenson, A. Gkoulalas-Divanis, and P. Aonghusa, "QuerioCity: A Linked Data Platform for Urban Information Management," in *Lecture Notes in Computer Science, The Semantic Web – ISWC 2012*, P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. Parreira, J. Hendler, G. Schreiber, A. Bernstein, and E. Blomqvist, Eds.: Springer Berlin Heidelberg, 2012, pp. 148–163.

[11] I. Celino, S. Contessa, M. Corubolo, D. Dell'Aglio, E. Valle, S. Fumeo, and T. Krüger, "Linking Smart Cities Datasets with Human Computation – The Case of UrbanMatch," in *Lecture Notes in Computer Science, The Semantic Web – ISWC 2012*, P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. Parreira, J. Hendler, G. Schreiber, A. Bernstein, and E. Blomqvist, Eds.: Springer Berlin Heidelberg, 2012, pp. 34–49.

[12] Zapico Lamela, Jorge Luis, B. Sayan, L. Bonanni, M. Turpeinen, and Y. Steve, "Footprinted.org – Experiences from using linked open data for environmental impact information," in *Proceedings of the 25th EnviroInfo Conference – Innovations in Sharing Environmental Observations and Information*: Schaker-Verlag, 2011, pp. 1–9.

[13] D. L. Phuoc and M. Hauswirth, "Linked Open Data in Sensor Data Mashups," in *Proceedings of the 2nd International Workshop on Semantic Sensor Networks (SSN09) in conjunction with ISWC 2009*: CEUR, 2009.

[14] Florian Bauer, Martin Kaltenböck, "Linked Open Data: The Essentials: A Quick Start Guide for Decision Makers," 2012.

[15] Tim Berners-Lee, *Linked Data - Design Issues.* Available: http://www.w3.org/DesignIssues/LinkedData.html.

[16] S. Kanhere, "Participatory Sensing: Crowdsourcing Data from Mobile Smartphones in Urban Spaces," in *Lecture Notes in Computer Science, Distributed Computing and Internet Technology*, C. Hota and P. Srimani, Eds.: Springer Berlin Heidelberg, 2013, pp. 19–26.

[17] R. Peterová and J. Hybler, "Do-It-Yourself Environmental Sensing," *Procedia Computer Science*, vol. 7, pp. 303–304, 2011.