Metadata Aggregation at GovData.de – An Experience Report

Florian Marienfeld
Fraunhofer FOKUS, Berlin
florian.marienfeld*

Ina Schieferdecker
Fraunhofer FOKUS, Berlin
ina.schieferdecker*

Evanela Lapi
Fraunhofer FOKUS, Berlin
evanela.lapi*

Nikolay Tcholtchev
Fraunhofer FOKUS, Berlin
nikolay.tcholtchev*

ABSTRACT

A key challenge for open data portals is the aggregation of metadata from various data catalogs (on different administrative level or from different application fields) also known as metadata harvesting[1]. This paper describes harvesting at the pilot of the German open government portal GovData.de, which is scheduled to become the data portal for all German public administration levels.

At the launch of the pilot portal in February, eleven federal, state and local data catalogs were integrated, which produced about 2,000 open data sets. In the meantime, the number of data sets increased to over 3,100 mainly due to improved harvesting capabilities of the portal. This paper discusses GovData.de metadata schema and experiences with the different harvesting techniques that are in use at GovData.de: CKAN-Harvest, CKAN-API, CSW-Harvest and JSON-Dump.

1. INTRODUCTION

Open data is an international movement that aims at opening public sector information to maximize reuse. Open in this context usually refers to machine processable online resources that are easy to access and that are put under free licenses. A free license enables the re-use of data by anyone for any purpose at no charge, requiring at most attribution. [Sunglight Foundation, 2010].

A typical implementation to enable ease of access is to collect metadata (i.e., descriptions of online data sets, their corresponding links to the online resources) into central data portals. This offers a "one-stop-shop" experience to data consumers, saving the trouble of collecting data from various portals, authorities or offices with different controls and settings.

*@fokus.fraunhofer.de

[1]Subsequently we call it harvesting only.

WikiSym '13 August 05 - 07 2013, Hong Kong, China

At the 5th German IT Summit, in December 2010, policymakers, public administration, private sector and research community adopted the Dresden Agreement, which states that the next step is to develop a centrally accessible open government platform, that offers open government data with standardized and user-friendly access. Fraunhofer FOKUS together with Lorenz-von-Stein-Institute and Partnerschaften Deutschland analyzed the status of open government data in Germany to lay the foundations for open government data and for planning, setting up and running a prototype of an open government platform operating across all levels of government[2]. This study was published in August 2012 [Bundesministerium des Innern, 2012]. It includes an analysis of the target groups, a compilation of relevant government data, an overview on technical standards, a review of legal frameworks and terms and conditions of use, of possible payment models. In addition, the study presents an operating model for the open government platform and a governance model for federal, state and local cooperation on open government data in Germany.

Subsequently, FOKUS was contracted to prototype the Open Government Data Platform for Germany whose pilot was launched at CeBit 2013 [3].

It turned out to be quite challenging to aggregate metadata in a way that is useful for data consumers. This is mainly caused by the great heterogeneity in terms of standards, schema, practices and semantics [Bundesministerium des Innern, 2012]. To remedy this we created a shallow, minimal schema that is compatible with the predominant data catalog vocabulary and software. Various German open data providers were consulted for this.

GovData.de offers various interfaces to integrate external data catalogs all of which are being used by at least one data provider to make more and more data sets available at the GovData.de prototype.

The remainder of this paper is organized as follows: Section 2 gives an overview of the GovData.de architecture and software stack. In Section 3 we describe the metadata structure that serves as point of convergence. Experiences with the actual import techniques are elaborated in Sections 4. Harmonization efforts and sustainability issues are discussed in Sections 5 and 6. Challenges and outlook are provided in

[2]`http://www.bmi.bund.de/SharedDocs/Pressemitteilungen/DE/2012/01/open_government.html`
[3]`http://www.bmi.bund.de/SharedDocs/Pressemitteilungen/DE/2012/07/opengovernment.html`

Sections 7[4].

2. GOVDATA.DE OVERVIEW

The Government Data Platform for Germany (see https://www.govdata.de/, in short GovData.de) bundles in one web interface metadata about data sets that are maintained in a decentralized for at each data provider's site. It provides a centralized access to all open data sets for anyone, in particular for data journalists, public administration, scientists and business people. GovData.de platform is comprised of two main components: a content management system (CMS) and a metadata catalog. The CMS is used to manage editorial content and enables a consolidated view on the metadata catalog. The catalog stores and manages metadata of all data sets, documents and applications.

The choice of Liferay[5] as CMS and CKAN (Comprehensive Knowledge Archive Network[6]) as metadata catalog are justified in [Bundesministerium des Innern, 2012]. They are setup in a way that the CMS facades the catalog except of the CKAN API. The content of the metadata catalog is displayed using search fields and result lists. This is outlined in Figure 1. Data providers can maintain metadata of new or updated data sets, documents or applications via a CMS's web form. In addition to the web interface, the metadata catalog can be accessed directly via a REST interface.

In alignment with some of the open data criteria only data sets that have an electronic resource, a description and a well-defined license can be published to GovData.de portal. GovData.de classifies sets with machine-readable resources as data, otherwise they are considered as documents. Data , document and app metadata are stored in the same catalog. "App" refers to any work that is derived from open data, suitable for end user presentation and not necessarily fit for reuse itself (e.g., visualizations, web apps of desktop applications).

The prototype went live on February 19th, 2013 and as of May 17th, 2013 hold 3112 open data sets.

3. METADATA STRUCTURE

Metadata is vital for the discovery of data, but what is recorded in addition to the name, description and author in the metadata of open data sets and how? This question arises when capturing the metadata as well as in the automatic import of metadata records, known as harvesting. Only if metadata structure and meaning are sufficiently uniform or self-explanatory, a central portal can be realized, to consolidate various data offers and the contents of existing external metadata catalogs.

Consistent metadata is addressed in various domains through different approaches and priorities, such as environmental or bibliographic data. Within open data innitiatives/communities CKAN is the de-facto standard for metadata catalog software and it is highly aligned to DCAT, the most prominent metadata catalog vocabulary [Maali et al., 2010]. CKAN is broadly used in Europe and recently in US[7].

CKAN exchanges metadata in JSON format. The metadata field name is the only required field, all others are optional. The core fields are title, description, resources (e.g., data files, services), license and contact person. Further details can be stored in a JSON dictionary called extras (i.e., as nested key-value pairs). Focusing on the essentials along with great flexibility are the main reasons why this metadata model has become so widespread. In the early development phase of GovData.de a desire for more structure became apparent, as many data providers and developers were looking for precise instructions on what information must be persisted and in which format. In order to preserve the minimal, flexible character of CKAN and JSON, and to fulfill GovData.de requirements we developed a CKAN-based JSON schema[8] for German public sector information. The structure is maintained on github.com[9]. It is intended not so much as a tool to validate metadata, but rather as a communication tool for those interested, like public decision-makers, data providers, developers and other open data initiatives in the German speaking area. For this reason the schema was published in early beta stage and now developed in public.

The metadata structure supports the description of data sets (including data services), as well as documents and applications. Here is how it is composed: The most important properties are stored at the top level. These include title, identifier, description, responsible and terms of use. Furthermore, the list of resources is essential, which contains pointers to the actual data, documents or applications. The most important property of a resource is its URL. In addition, a description and format can be provided for a resource. This configuration allows capturing related files as one record, possibly for different periods, in different languages or formats. Within the "extras" all other data are stored. These mainly include the temporal and spatial arrangement, and details about the origin of imported items.

4. HARVESTING

With that as a common target structure for unification, four different import techniques were presented: JSON dump, CKAN-CKAN harvesting, CSW-ISO19115-harvesting and CKAN-REST-API.

Using the JSON dump method, the operator of a given remote catalog just names a URL resource, under which a JSON file compliant to the GovData.de schema can be retrieved. The JSON file contains all the data sets and can be retrieved on a daily basis. This procedure has been used in Bremen, Bavaria and Moers. With a few feedback loops, the providers were able to optimize their individual JSON export tools to the extent that a smooth integration of the metadata was possible. The source code for this is published as open source CKAN plug-in `ckanext-govdatade`[10]. This method has proven as quick and easy solution to integrate random catalogs but will have to be refined for performance

[4]Portions of this material were previously posted at `http://open-data.fokus.fraunhofer`

[5]`http://liferay.com`

[6]`http://ckan.org`

[7]`http://data.gov.uk`, `http://data.gov`, cf. `http://ckan.org/instances`

[8]`http://json-schema.org`

[9]`https://github.com/fraunhoferfokus/ogd-metadata`

[10]`https://github.com/fraunhoferfokus/ckanext-govdatade`

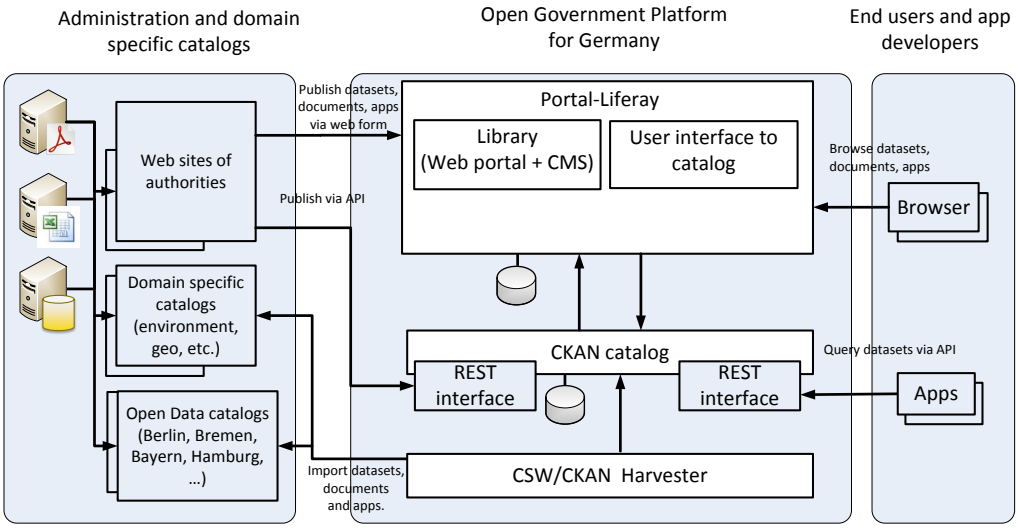

Figure 1: Harvesting Architecture

reasons, as all data sets are overwritten on each poll.

The CKAN-CKAN harvesting is used in the data portals of Hamburg, Berlin, Rostock and Rhineland-Palatinate. Theoretically, it is possible to use the CKAN harvesting extension ckanext-harvest[11] for this task without any further development or configuration. This is feasible because the operator of the remote CKAN instances follow the suggested GovData.de metadata structure. In practice however, it is necessary to take several details into account: 1) the adoption of the categories (CKAN: "groups"), only works with minor tricks; 2) the mapping CKAN.author ⇔ "publishing authority" is not consistently used; 3) the use of CKAN.name and .id have to checked for uniqueness; 4) capital letters and special characters in the tags, or keywords, are not transferred properly. Moreover, additional keywords and titles also have to be supplemented, e.g. the Hamburg metadata catalog does not tag all of the data sets with the word "Hamburg". All these minor issues do not prevent the CKAN-CKAN harvesting from working quite smoothly. In our experience this manifests as efficient regular harvesting.

Importing geospatial metadata which are encoded according to the ISO 19115 standard geographic metadata [International Organization for Standardization, 2003] via the CSW (Catalog Service for the Web) interface is more complicated. This may be due to the fact that geospatial data are distributed and consumed very differently from the usual approach with respect to open data principles. In this context, data are called "products". Frequently these are maps on CDs or paper, which are found on the basis of the metadata. But then usually a bilateral contract is signed and the data is handed over directly from the provider to the contract partner. Thus, the details "online resource" and "license", which are of key importance for open data, only have a very limited level of relevance in terms of both the standard and the use by the data provider. Moreover, the very detailed ISO 19115 meta data model is used with different profiles from federal state to federal state. This means that it is difficult to identify the publishing authority in all the data sets of e.g. German Geoportal.de which covers the whole of Germany. With a public CKAN module for ISO harvesting data sets of the Federal Statistical Office, the Regional Database and the open data sets offered of the Environment Office of Lower Saxony were harvested. In all these cases, the standard was implemented very consistently and the question of the licenses partially clarified. In terms of source code, we branched the CKAN extension `ckanext-spatial`[12]. This fork implements two adaptations. 1) The mapping of ISO metadata to the GovData.de structure, 2) allows for downloading of zipped XML, that some providers offer instead of a CSW endpoint. As for the Environmental Office of Lower Saxony, the open data sets are marked with a special key word, which allows for fetching them with a regular CSW query.

The harvesting architecture is illustrated in Figure 2: The first two importers (JSON and CKAN-CKAN) are only based on the `ckanext-harvest` extension and have therefore been directly installed in the productive CKAN of GovData.de. The ISO 19115 harvester, however, is based on the extension `ckanext-spatial`, which alters CKAN in various ways. In order to keep the main catalog lean it runs on a separate machine. This implies that the metadata is harvested into the extra machine first. In a second step, the data sets are then transferred to the actual metadata catalog.

5. ALIGNMENT WITH EU ACTIVITIES

On European level there are several activities that are strongly interrelated with GovData.de metadata management.

The ISO importer from the previous section nicely illustrates the relationship of open data and the INSPIRE directive. That directive "establishes an infrastructure for spatial information in Europe to support Community environmen-

[11] `https://github.com/okfn/ckanext-harvest`

[12] `https://github.com/fraunhoferfokus/ckanext-spatial/tree/ogpd`

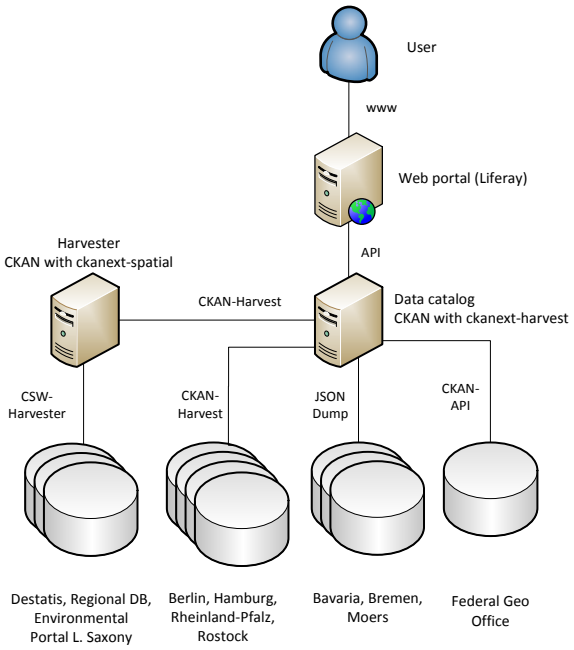

Figure 2: Harvesting Architecture

tal policies" [13]. Effectively, it regulates the registration and provisioning of most geo related data. INSPIRE contributed substantially to the implementation of regional and national metadata registers. These in turn are great inspiration for open data in terms of structure and semantics for re-usable data and metadata. However, from an open data perspective, INSPIRE activities have failed to properly regulate the catalogization of machine processable online resources and the indication of re-use friendly, interoperable license terms. The German INSPIRE community[14] is currently investigating how to tackle these problems with the national ISO profile.

In the German speaking area the metadata structure has been consolidated with the Austrian Open Government Data Initiative[15]. This has – much in contrast to the INSPIRE consolidation – raised little to no controversy, as both the Austrian schema [16] and the German one are eventually CKAN profiles and differ only marginally in syntax and semantics[Höchtl and Habernig, 2013].

The DCAT application profile for data portals in Europe[17] is an activity that aims at harmonizing data catalogs in Europe. The working group is creating a document that declares mandatory, recommended and optional properties on top of the DCAT vocabulary[Maali et al., 2010]. Due to the fact that CKAN and DCAT are mostly JSON and RDF-Versions of each another, German metadata should easy to integrated in a EU portal without significant efforts. A minor point of disagreement in constituted in the fact that the German schema relies on the fact that a data set will always have exactly one license, whereas the European profile allows for individual licenses for each file in a data set.

In February 2013 a prototype of a data portal for the European Commission was launched [18]. It is also CKAN-based and should therefore be well aligned with the activities mentioned above.

6. SUSTAINABILITY

The first experiences of operating this harvesting system already shed some light on how much effort will required to maintain it. So far we have seen that there was some amount of work for both the data providers and the central portal operator each time a new source catalog was connected. After the initial efforts, almost no manual work is required to keep the catalogs in sync. To estimate how much effort will be required in the future, we consider three possible non-exclusive scenarios that are very likely to occur.

Scenario 1) There will be more datasets. The German geospatial data portal for instance holds approximately 100,000 records of metadata. Of course it is hard to say if open data increases to half or double that amount. Nevertheless, given the fact that CKAN relies on scalable database and search engine back ends (by default SOLR[19] and Postgres[20]), the pure amount should never be become a problem of scale. This holds only, though, as long as the central portal only deals with metadata and not with payload data.

Scenario 2) There will also be more data providers. The current mode of bilateral agreements for good harvesting results will obviously not scale to say 1,000 German data providers. A reconsolidated metadata structure and API are key for a more efficient integration of data sources.

Scenario 3) Governmental institutions will change. Over time authorities get merged or split, responsibilities get shifted between departments and officers. Due to this, email addresses, domains and URLs are not quite persistent. As longevity is very important for open data, the system will of course suffer from changes in governmental institutions. This is already manifesting today: data providers demand to be harvested at least daily, because their URLs may change within hours. In the context of these changes there will always be a need for manual correction and adaptation on both sides. However, we believe that the harvesting system presented here can remedy to some extend, since for each data set URLs to online resources and contact email of the responsible party are given. This allows for automatic link checking and notification, which is already taking place at GovData.de and helps maintaining metadata quality.

7. CHALLENGES AND OUTLOOK

Even though these initial results are very promising, we discovered a major challenge that only just starts to manifest. The key problem is the assumption that one can federate metadata strictly from bottom up to the top, i.e. from local, via regional to the national level and further to the European and international level. The first issue is that are

[13] http://inspire.jrc.ec.europa.eu
[14] http://www.geoportal.de/EN/GDI-DE/gdi-de.html?lang=en
[15] http://www.data.gv.at/hintergrund-infos/cooperation-ogd-oesterreich
[16] http://reference.e-government.gv.at/uploads/media/OGD-Metadaten_2_1_2012_10.pdf
[17] https://joinup.ec.europa.eu/asset/dcat_application_profile/description
[18] http://open-data.europa.eu/open-data
[19] http://lucene.apache.org/solr
[20] http://www.postgresql.org

several parallel "federal systems" such as German environmental, geographical, statistical or other metadata catalogs. These perform already some form of harvesting. In addition to passing metadata to the top, it gets often passed "sideways". That means that several routes exist for a piece of metadata to get from the origin via transformation steps to an aggregation portal. For example, a geospatial data set from Bavaria may be transferred to the federal geospatial data portal and the Bavarian open data portal. Consequently, GovData.de is likely to receive at least two possibly quite different copies of metadata for the very same geospatial data set. Dealing with this is more complicated than duplicate detection, because one route may actually be "better" in some relevant sense than another. Moreover, data sets may be split or joined according to the respective policy of bundling resources into data sets.

In addition to this challenge, a set of regular engineering problems surfaced. The fact that data sets get deleted for various reasons is more of a common practice than an exception. Hence, checking if all harvested data sets are still available has to be made a routine task.

Furthermore, GovData.de experiences a large heterogeneity in the metadata quality. Hence, quality assurance is in itself a big issue including e.g. tag harmonization and disambiguation, checks for referenced URL and metadata validation.

Research projects need to be setup to systematically address the challenges discovered in the prototype phase and to lay down a conceptual framework including methods and tools for the efficient and high-quality management of metadata for (open) data sets.

In addition to improving the quality and quantity of metadata on GovData.de, one of the key projects in the near future will be to move from a metadata federation hierarchy to a network of harvesting nodes. This implies that metadata is considered more a message than a static fact. And like emails that carry a list of the servers that they went through, metadata packages would have to carry a list of metadata catalogs that they passed including documentation of the transformations that were performed on them.

8. REFERENCES

[Bundesministerium des Innern, 2012] Bundesministerium des Innern (2012). Open Government Data Deutschland. http://s.fhg.de/od-deutschland-studie.

[Höchtl and Habernig, 2013] Höchtl, J. and Habernig, C. (2013). Gegenüberstellung der OGD-Metadaten Ausarbeitungen von Deutschland und Österreich. http://www.data.gv.at/wp-content/uploads/2012/03/Gegen%C3%BCberstellung_OGD_Metadatenschemata_A_D_20130308.pdf.

[International Organization for Standardization, 2003] International Organization for Standardization (2003). ISO 19115 Geographic Information - Metadata.

[Maali et al., 2010] Maali, F., Cyganiak, R., and Peristeras, V. (2010). Enabling interoperability of government data catalogues. In Wimmer, M., Chappelet, J.-L., Janssen, M., and Scholl, H. J., editors, *EGOV*, volume 6228 of *Lecture Notes in Computer Science*, pages 339–350. Springer.

[Sunlight Foundation, 2010] Sunlight Foundation (2010). Ten Principles for Opening Up Government Information. http://sunlightfoundation.com/policy/documents/ten-open-data-principles.