EGOV-CeDEM-ePart

A Translation Service for Open Data Portals

Sebastian Urbanek*, Sonja Schimmler**

*Fraunhofer Institute for Open Communication Systems, Kaiserin-Augusta-Allee 31, 10589 Berlin, Germany, sebastian.urbanek@fokus.fraunhofer.de
**Fraunhofer Institute for Open Communication Systems, Kaiserin-Augusta-Allee 31, 10589 Berlin, Germany, sonja.schimmler@fokus.fraunhofer.de

Abstract: Open Data portals provide tens of thousands of datasets, which are described by metadata. This metadata is typically available in one or two languages only and, if translations exist, they are usually added manually. To build an inclusive data infrastructure, metadata should be available in as many languages as possible. The paper presents an approach for automatic translation of metadata within Open Data portals that are based on Semantic Web technologies and use the metadata standard DCAT-AP. Based on this approach, new functionalities are possible, such as enabling users to search for datasets in their native language. The approach was implemented for and tested within a practical application in a production environment.

Keywords: translation service, open data, semantic web, linked data, DCAT, DCAT-AP

Acknowledgement: This work has been partially supported by the Federal Ministry of Education and Research of Germany (BMBF) under grant no. 16DII117 ("Deutsches Internet-Institut"). The implementation and provision of the European Data Portal has been funded by the European Commission under contracts DG CONNECT SMART 2017/1123.

CEUR-WS.org/Vol-3049/Paper12.pdf

1. Introduction

A central task of Open Data portals is the provision of data and the possibility to reuse it for other purposes. To make this possible, the data must be findable and accessible by the users. This is ensured by metadata, which describes the properties of the original data. For interoperability, several metadata standards have been established. The Data Catalogue Vocabulary (DCAT) (Albertoni et al., 2020) is a metadata standard for the provision of Open Data. In combination with the Resource Description Framework (RDF) as part of the Semantic Web technology stack, knowledge graph is enabled, to provide context and foster interoperability.

Metadata is not always published in English; many national Open Data portals only list them in the local language. To reduce language barriers and to make data globally accessible to foreign speakers, metadata should be made available in as many languages as possible. Integrating this information into a knowledge graph allows for additional search and filtering options. Making a

Copyright ©2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

SPARQL endpoint available to access such a graph gives access to the translations in addition to the original metadata. This opens up the possibility of developing further applications.

This paper presents an approach for automatic translation of metadata, tailored to Open Data portals that are based on Semantic Web technologies and use the metadata standard DCAT-AP. Translations are performed on top of a knowledge graph representation, which serves as the input and is expanded and updated accordingly. A software component was developed within a data management ecosystem, which uses automatic translation and knowledge graph embedding.

2. Related Work

Today, it is possible to translate texts without extensive knowledge of a target language. **Machine translation engines** do this without the assistance of a single person. Modern translation engines are based on statistical methods (SMT), which analyze the frequency and the similarity of words or phrases in two different languages. If the similarity is high enough, one phrase is considered to be the translation of the other (Koehn, 2010). Neural machine translation (NMT) is also a special type of statistical translation. An artificial neural network analyzes the texts, and the network is trained for the translation with some better results in contrast to SMT (Bahdanau, Cho, Bengio, 2016). This approach is used by the European Commission's translation service, called eTranslation. It emerged in 2017 from a research project called MT@EC and can translate text snippets as well as full documents.

There exist several **standards** for describing data, which build on each other. One widely used metadata standard is the Data Catalog Vocabulary (DCAT), developed by the World Wide Web Consortium (W3C). DCAT uses the Resource Description Framework (RDF) to ensure interoperability between different datasets and portals. An extension for data portals in Europe is the DCAT Application Profile (DCAT-AP), which the European Commission now publishes in version 2.0. It describes the metadata of a dataset and the format in which they are stored and enables Open Data portals to harvest datasets and catalogues from other ones with little effort. Nuffelen (2019) describes a dataset as a collection of data available for access or download in one or more formats. Properties describe each dataset as metadata. Mandatory properties are title and description. The actual data is linked as a distribution to a dataset. Distributions have an access URL as a mandatory property, and a title and a description as optional properties. The datasets and distributions each contain a description and a title that can be translated as literal.

As far as we know, very few Open Data portals have **translations** of their metadata into other languages, and if so, they are not **managed automatically** by a middleware. For instance, the EU Open Data portal[1] has translated metadata, but these translations are done manually. Within research data portals, as provided by universities, metadata is usually provided in a native language and/or in English, and translations are done manually by data curators. In Wikidata (Wikidata, 2020), metadata and data are provided in multiple languages. These translations are contributed by tens of thousands of volunteers.

[1] https://data.europa.eu/euodp/en/about?

3. Design of the Translation Service

3.1 Global Architecture of the Translation Service

Many tools for the machine translation of text snippets or entire documents exist on the market. Very well-known are Google Translate, Bing Translate or DeepL, which are commercial or private services. This is not suitable for every application. In the context of the Connecting Europe Facility (CEF) funding programme[2], the EU supports a machine translation project, **eTranslation**, focusing on translating documents for EU members that do not necessarily have to be done by a qualified translator as a person and are secure on servers in Europe. eTranslation provides a Web interface to upload documents in Microsoft Word or PDF format. Developers can use its API that requires text snippets and some meta-information. This includes the domain, authorisation, source language, and target languages.

Our Translation Service is part of **Piveau** (Kirstein et al., 2020), a data management ecosystem for the public sector. It is designed for Open Data portals based on Semantic Web technologies and uses the metadata standard DCAT-AP. For the translation of text snippets, it integrates eTranslation. Piveau is divided into several microservices that perform different tasks. The most important ones are Piveau Hub, Piveau Consus, Piveau Metrics and Piveau UI. Piveau Hub is the heart of the entire application, managing a data catalogue and providing metadata. It includes an indexing component, enabling search and filter functions. Internal storage is realised with a Virtuoso triplestore[3]. Piveau Consus is importing and processing data and metadata, also from external sources. Piveau Metrics checks the data against various data quality criteria. Piveau UI is a customisable web interface for interacting with the data.

As shown in Figure 1, our **Translation Service** is integrated into Piveau as follows: Being triggered by a scheduler, data is collected by Piveau Harvester. Piveau Hub checks within the internal knowledge graph if the retrieved datasets contain new or changed texts and sends translation requests to the Translation Service. When translations are finished, the Translation Service sends them back to Piveau Hub, which adds them to the internal knowledge graph.

3.2 Components of the Translation Service

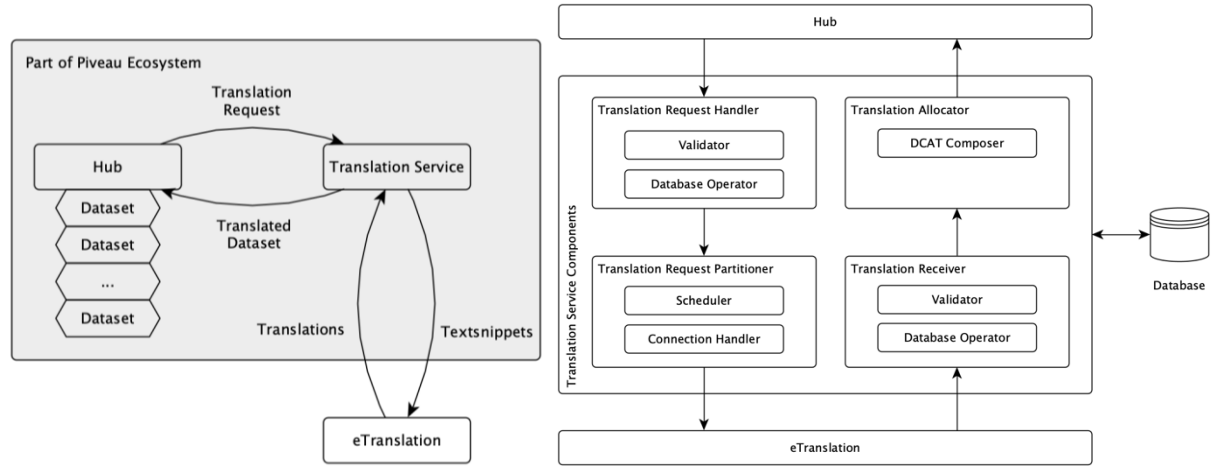
Internally, our Translation Service consists of independently operating components that control all translations simultaneously. It works asynchronously and uses the reactive framework Vert.x, which is well suited for handling large amounts of text snippets. As shown in Figure 2, it consists of four core components, which are triggered one after another: translation request handler, translation request partitioner, translation receiver and translation allocator. It further provides a database for internal storage for fail-safe purposes and integrates eTranslation for the actual translation of text snippets.

[2] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/CEF+Digital+Home
[3] https://virtuoso.openlinksw.com

Translation Request Handler

As described in the previous subsection, new translation requests are sent from Piveau Hub to the Translation Service. The **Translation Request Handler** formally checks the requests. If errors occur, a request is rejected. For instance, text snippets might be empty, or the original language is not supported by eTranslation[4]. Valid translation requests are placed in a queue and temporarily stored in a database for fail-safe purposes.

Figure 1: Overview of Translation Process; Figure 2: Internal Structure of Translation Service



Translation Request Partitioner

The **Translation Request Partitioner** checks periodically if there are new translation requests in the queue. Once translation requests are waiting to be translated, the corresponding datasets are prepared for translation, and text snippets are sent to eTranslation.

It is not possible to send a record in the form of composed RDF triples to a translation engine since such an engine simply translates text into another language. The engine might destroy the syntax or might not be able to translate it at all. Thus, all data records need to be prepared. The individual texts that are meant to be translated are extracted and always remain assignable to the respective datasets by adept ID management. These text snippets are submitted to the translation engine used.

eTranslation provides a REST API that accepts a POST request and JSON as a body. To ensure a successful translation of all texts from the datasets, they are brought into a form that conforms to eTranslation's API. As eTranslation has a character limit, the texts to be translated are divided into several subtexts, translated individually, and reassembled afterwards.

Although the translation requests were already formally checked by the Translation Request Handler, further pre-processing steps are required. Often, the title and description of a dataset are

[4] en, bg, hr, cs, da, nl, et, fi, fr, el, hu, ga, it, lv, lt, mt, pl, pt, ro, sk, sl, es, sv, nb, de

not correctly entered by the data provider. The actual plain text is sometimes an XML or HTML document, sometimes a string of characters resulting from a binary coded file. In the case of XML or HTML documents, a particular form of document translation is used that does not translate tags and alike. In the case of a binary file, no translation is performed. Often, text snippets include URLs and hashtags. These are extracted and saved separately before translation and inserted again afterwards.

Translation Receiver

The Translation Service must also provide an interface for the finished translations, as eTranslation works asynchronously. For each language, the Translation Service receives a separate request from eTranslation. On average, each text is translated into 26 languages. Additionally, each dataset consists of at least four texts - title and description of a dataset and title and description of each distribution. Datasets often have multiple distributions, on average 8 in our usage scenarios, and sometimes hundreds, thus creating a vast number of requests. The **Translation Receiver** manages the finished translations. It checks the properties and language and performs error handling. All translations are stored in a database for fail-safe purposes.

Translation Allocator

The **Translation Allocator** finally prepares the translated texts for integration into the internal knowledge graph. All text snippets are assigned to the corresponding dataset or distribution and provided with semantic features. This includes information about the languages and whether it is a manual translation or a machine translation. A text to be translated according to the DCAT-AP standard is a literal and has an optional language label. The XML data type lang specifies this. "This is an example text."*@en*. The language tag is based on the ISO 639-1 language encoding. If a text is translated, the tag is extended by: *en-t-de-t0-etranslation*. The first language code stands for the actual language, followed by a *t* for the translation from the original language (in this example: *de*). The *t0* signals a machine translation, and the *etranslation* stands for the engine used (Davis et al., 2012). New translations are integrated into the internal knowledge graph together with the generated language tags. The Apache Jena framework[5] is used for interaction between the software and the knowledge graph.

4. Use Case: The European Data Portal

Generally, each EU member has one or more national Open Data portals, separated by topic such as geographic or administrative data. Due to the historical development of Europe, different languages are spoken in these countries. Nevertheless, the EU is trying to foster trans-European exchange and minimise language barriers.

This requirement also exists for the European Data Portal (EDP, 2019). Its central task is making public sector information available as Open Data. It currently provides approximately 1.1 million datasets from 36 European states, harvested from 81 national Open Data portals. These national

[5] https://jena.apache.org/index.html

portals mostly contain datasets in English or the national language. For the goal of being an inclusive portal, all datasets need to be translated, possibly into all other European languages.

In August 2019, we successfully integrated our Translation Service into the EDP. It fully meets the formulated requirements, and apart from some minor adaptions, it is fully functioning since then.

Due to the daily harvesting of national Open Data portals, there are a lot of datasets that need to be translated every day within the EDP. This is because the descriptions or titles change or because new distributions or even entire datasets are added. Usually, between 1000 and 130,000 new data records or translation requests are sent to the Translation Service every day. This requires the Translation Service to produce the translations in a timely manner.

At some point, the metadata has been integrated into a national Open Data Portal by either humans or machines. Thus, the input is very heterogeneous and does not always follow the rules defined by DCAT-AP. This is a major challenge and requires metadata pre-processing.

5. Conclusion & Outlook

The Translation Service has been developed to translate large amounts of metadata automatically. It is designed for Open Data portals that are based on Semantic Web technologies and use the metadata standard DCAT-AP. It allows for the use of different translation engines, which take care of the actual translation. Thanks to integrating the translations into an internal knowledge graph, new functionalities are enabled, e.g., users can search for datasets in their native language. Also, translations are available via a SPARQL endpoint, which can be used for future applications.

Future work on metadata pre-processing can further improve the service. One example is the source language, which is not always stated correctly. Consequently, the text snippet cannot be translated by eTranslation. Automatic recognition of the source language could resolve this issue. Further pre-processing steps of the texts are also required. One example are problems with special characters and character encodings in some languages - including text sections that should not be translated.

References

Albertoni, R., & Browning, D., & Cox, S., & Beltran, A., & Perego, A., & Winstanley, P. (2020). *Data Catalog Vocabulary (DCAT) - Version 2*. W3C Recommendation. Retrieved March 19, 2021, from https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/.

Koehn, P. (2010). *Statistical Machine Translation*. New York: Cambridge University Press, 3rd edition, 81-118.

Bahdanau, D., & Cho, K., & Bengio, Y. (2016). *Neural Machine Translation by Jointly Learning to Align and Translate*.

Nuffelen, B. v. (2019). *DCAT Application Profile for data portals in Europe Version 2.0.0*. EU Commission. Retrieved March 19, 2021, from

https://joinup.ec.europa.eu/sites/default/files/distribution/access_url/2019-12/12f0dc1d-50b6-43e4-90c2-0afe213ac2be/DCAT_AP_2.0.0.pdf.

Phillips, A., & Davis, M. (2009). *Tags for Identifying Languages*. BCP47. Retrieved March 19, 2021 from https://tools.ietf.org/html/bcp47.

Wikidata (2020, August 10). *Help:Multilingual*. Retrieved March 19, 2021 from https://www.wikidata.org/wiki/Help:Multilingual.

Kirstein, F., & Stefanidis, K., & Dittwald, B., Dutkowski, S., & Urbanek, S., & Hauswirth, M., (2020). *Piveau: A Large-Scale Open Data Management Platform Based on Semantic Web Technologies*. In: Harth, A. et al. (eds) The Semantic Web. ESWC 2020. Lecture Notes in Computer Science, vol 12123. Springer, Cham.

EDP (2019). *European Data Portal User Manual. Portal Version 4.3*. Retrieved March 19, 2021 from https://www.europeandataportal.eu/sites/default/files/edp_s1_man_portal-version_4.3-user-manual_v1.0.pdf.

Davis, M., & Phillips, A., & Umaoka, Y., & Falk, C. (2012). *Extension T - Transformed Content*. BCP47. Retrieved March 19, 2021 from https://tools.ietf.org/html/rfc6497.

About the Author

Sebastian Urbanek

Sebastian Urbanek is a researcher at the Fraunhofer FOKUS. Here, he is working on the core topics of Open Data and data management, including the European Data Portal. As a member of the Weizenbaum Institute, he is researching the effects of digitalisation on science. He focuses on databases, data structures and data quality.

Sonja Schimmler

Sonja Schimmler leads the research group "Digitalisation and Science" at the Weizenbaum Institute and at Fraunhofer FOKUS. She is also an associated researcher at the Technical University of Berlin. In her research, she focuses on the digitalisation and opening of science with a special emphasis on research data infrastructures. Her research interests range from semantic web and linked data over data science and artificial intelligence to software engineering and human-centered computing. She is doing fundamental interdisciplinary and application-oriented research.