Walther, Tatiana, & Hauschke, Christian (2018). Integrating data from distributed sources via lookup services. *EuropeanaTech Insight*, *9*. https://pro.europeana.eu/page/issue-9-swib

Integrating data from distributed sources via lookup services

Tatiana Walther, Christian Hauschke (German National Library of Science and Technology (TIB) Hannover

Introduction

Currently global efforts are being undertaken in librarian, cultural and scientific context to make data interoperable and interlinked. The importance of applications and services, which can enable comfortable data integration, has risen.

Descriptive and subject cataloguing can benefit from the use of external controlled vocabularies and thesauri, preferably in Resource Description Framework (RDF). This data is either accessible on the web via various Application Programming Interfaces (APIs) or stored in data dumps.

In this article, we describe an extension of lookup services in the research information system software VIVO, which can be applied for integration of external distributed data - e.g. concepts from subject authorities, available on the web. The integration of non-concepts is planned as well. The lookup service utilizes SPARQL Protocol and RDF Query Language (SPARQL) endpoints, Representational State Transfer (REST) API and the Skosmos tool as a middleware. The extension has been developed at the German National Library of Science and Technology (TIB) Hannover.

Motivation and Goal

VIVO is an open source software, developed at the Cornell University Library. It uses Linked Data technologies and standards, as e. g. RDF, Resource Description Framework Schema (RDFS), SPARQL, Web Ontology Language (OWL), and Simple Knowledge Organization System (SKOS). As a research information system VIVO is generally used to represent scholarly activities of one or more institutions on the Web [1]. A typical installation covers profiles of persons connected with organizations, publications, projects etc.

VIVO delivers a set of default vocabularies, e. g. the Library of Congress Subject Headings (LCSH) to assign concepts as annotations to various information items. Regarding the integration of concepts for subject cataloguing, the aforementioned authorities were not sufficient to fulfill the needs of librarians and end users (mainly scientists) using VIVO. Extending these services is also necessary due to a new and increasingly important reporting standard for research information in Germany – the Research Core Dataset (Kerndatensatz Forschung, KDSF), which determines the use of the „Destatis Fächersystematik“ (Subject Classification of the German Federal Office of Statistics) for assigning subject annotations. To record data in VIVO in conformity with the KDSF we also had to integrate this vocabulary into VIVO.

Either manual or semi-automatic collecting, enriching and converting of a larger amount of concepts or other objects costs a lot of time and resources.

Thus, the objective of our project was to extend the scope of external vocabularies and other bibliographic sources in VIVO, integrated via lookup services.

Implementation

As a pilot we implemented two subject authorities. Besides the „Destatis Fächersystematik“, we opted for the Standard Thesaurus for Economics (Standard Thesaurus Wirtschaft, STW). Integration of non-SKOS data collections is still being planned.

The „Destatis Fächersystematik“ was initially available in a non-machine readable file. We used Skosmos as a means to provide access to the vocabulary for human users and machines alike. Integrating „Destatis Fächersystematik“ in Skosmos and VIVO required conversion to a concept scheme in SKOS, which was done by means of OpenRefine. Subsequently, we have further processed the vocabulary with Skosify [3] for Skosmos to fully interprete the vocabulary, which results in additional features. The STW was already available in SKOS and equipped with a publicly available SPARQL endpoint. Therefore no middleware was required.

Setting up the lookup service involved some work on several Java configuration files.

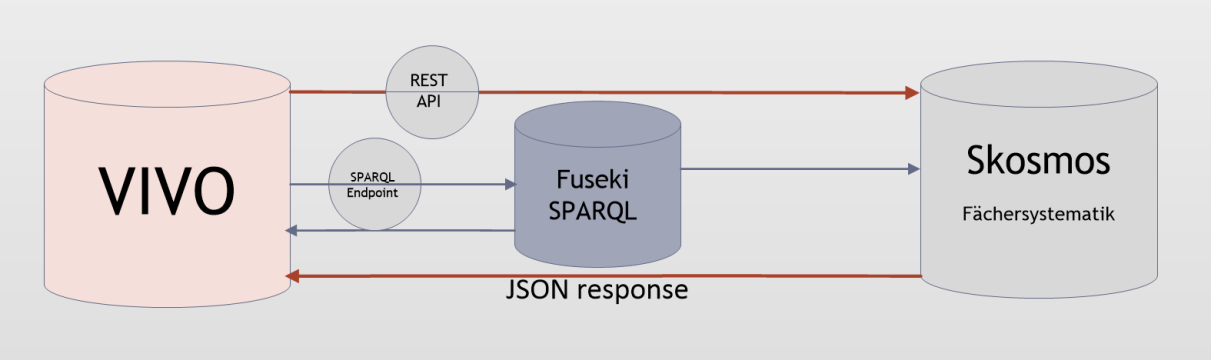

Figure 1: Integration of "Destatis Fächersystematik" from Skosmos into VIVO.

As figure 1 represents, to integrate the Destatis Fächersystematik, we utilized a REST API provided by Skosmos, which offers a SPARQL Endpoint as well. When using the REST API, VIVO sends a request as a URL to the interface of Skosmos. It receives a response in JavaScript Object Notation (JSON) form, which is further processed locally.

In the case of STW, as outlined in figure 2, VIVO communicates with a SPARQL endpoint. VIVO sends a SPARQL query and receives a response in JSON form as well. We chose to try both alternatives to learn about different ways to integrate vocabularies.

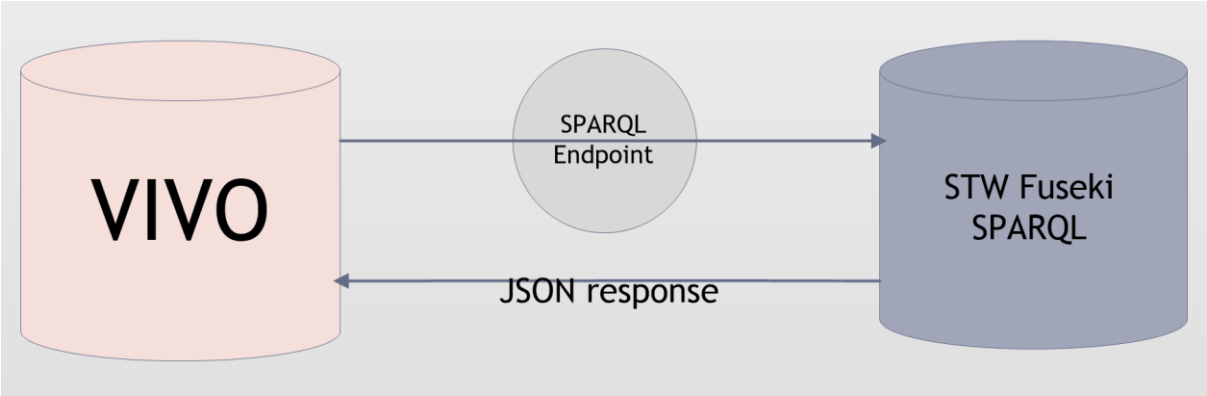

Figure 2: Integration of STW into VIVO

In the case of STW, as outlined in figure 2, VIVO communicates with a SPARQL endpoint. VIVO sends a SPARQL query and receives a response in JSON form as well. We chose to try both alternatives to learn about different ways to integrate vocabularies.

Functioning

Figure 3 shows handling with concepts from external authorities in the user interface. In the list of vocabulary services a user can now select one of the new vocabularies. After a requested term has been typed in the search form it is sent to the selected service. The targeted application checks, if there are any concepts with preferred labels (as for now defined in the configuration file, other settings are possible, too) matching the searched term and sends the response.

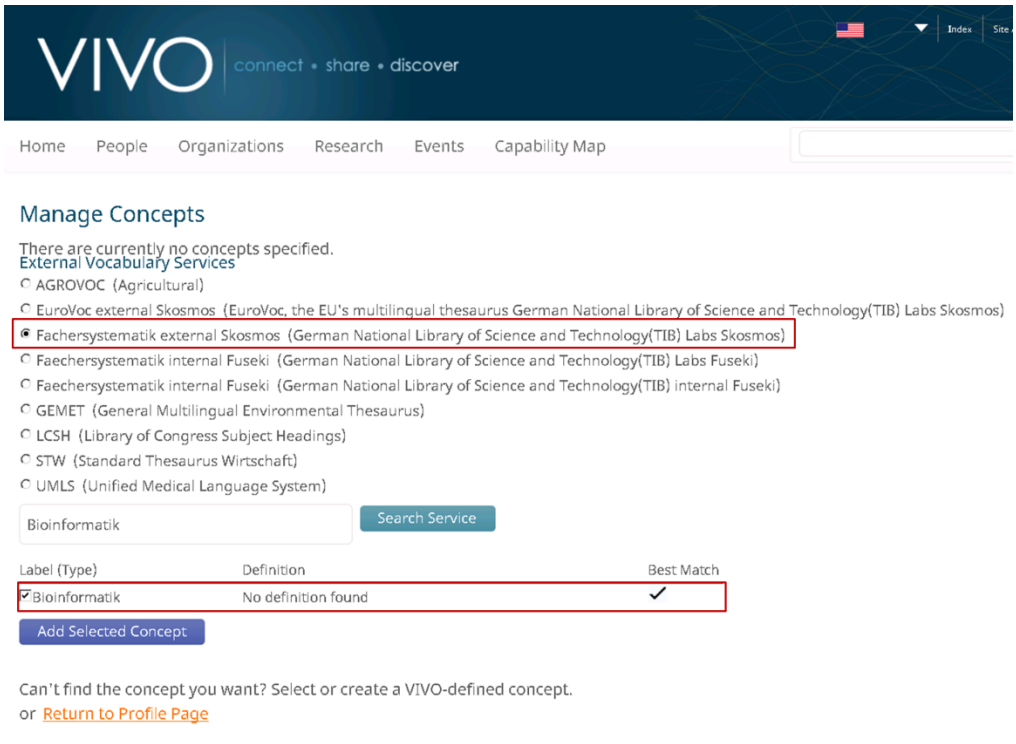

Figure 3: User interface for managing concepts from external sources in VIVO

The result is a list of suggested terms with the best matching concept denoted. The user is now able to add the selected concept to his profile.

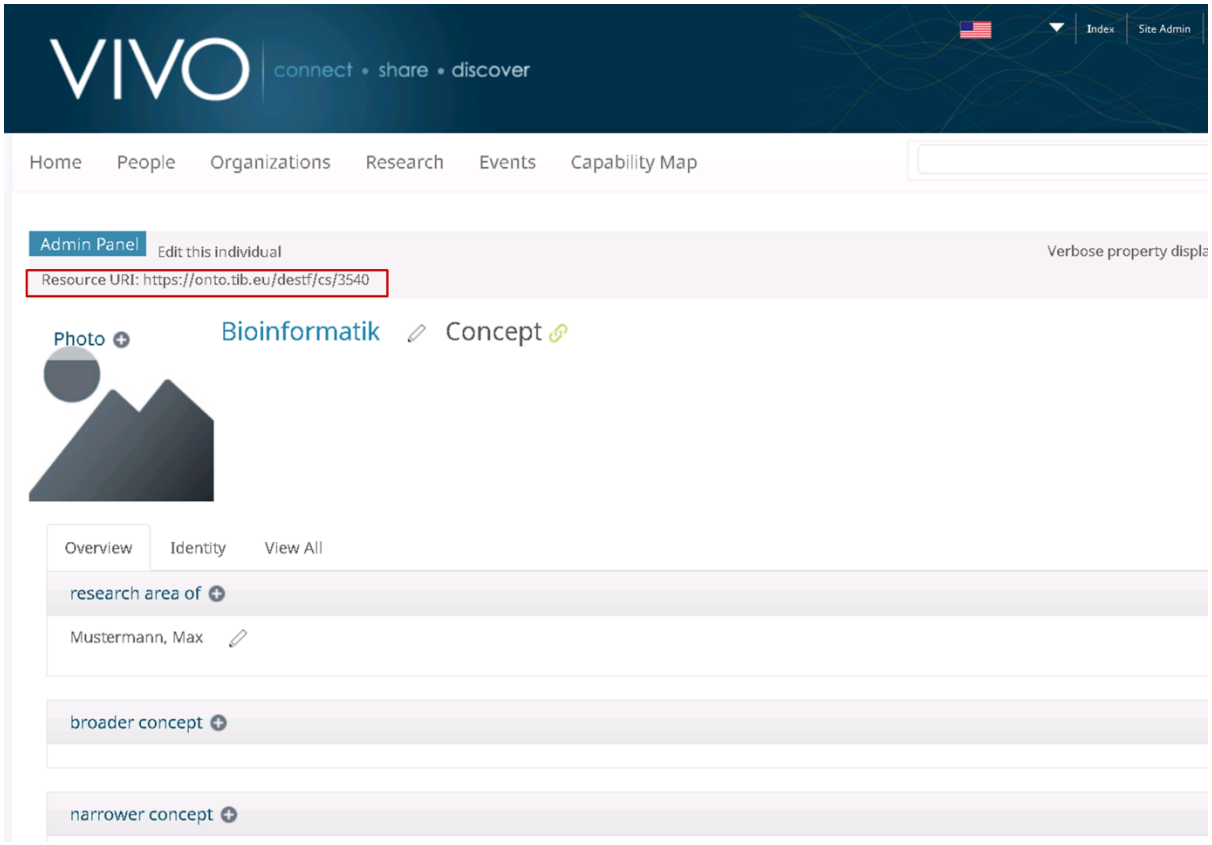

Figure 4: A profile page of the new integrated concept in VIVO

Concurrently a profile page for the concept, shown in figure 4, is being automatically created. The concept preserves its United Resource Identifier (URI) from the external authority, and its metadata, if available, as well.

Conclusion

The lookup service presented here is reliable and meets our expectations of building an infrastructure for vocabularies to be used in VIVO – and other systems, of course. Nevertheless, further conceptual and technical work is necessary to deal with changes in the vocabulary used. In terms of content, the next challenge is to integrate non-SKOS entities from such sources as Wikidata and the Integrated Authority File of the German National Library (GND) [4], which requires mapping models, normalization of data and disambiguation processes in the background. A significant work in this area was done in the scope of Linked Data For Libraries (LD4L) project (see also [5]). A generic user friendly interface for lookup services is another goal to be achieved. Thus, further developments and improvements of the existing lookup services in VIVO to make them more generic and applicable are required.

References

[1] Duraspace, "VIVO," [Online]. Available: http://vivoweb.org/info/about-vivo. [Accessed 23 Februar 2018].

[2] O. Suominen, H. Ylikotila, S. Pessala, M. Lappalainen, M. Frosterus, J. Tuominen, T. Baker, C. Caracciolo and A. Retterath, "Publishing SKOS vocabularies with Skosmos. Manuscript submitted for review," 2015.

[3] O. Suominen and E. Hyvönen, "Improving the Quality of SKOS Vocabularies with Skosify," in Knowledge Engineering and Knowledge Management : 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings, 2012.

[4] T. Walther and M. Barber, "Integrating Data From Distributed Sources Via Lookup Services," 18 December 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117861.

[5] H. Khan and E. L. Rayle, "Linking the Data: Building Effective Authority and Identity Lookup," 2017. [Online]. Available: https://wiki.duraspace.org/display/ld4lLABS/Samvera+%28aka+Hydra%29+Community+Linked+Data+Support?preview=/87458291/90976949/khan_linking-the-data.pdf.