1st Conference on Research Data Infrastructure

Humanities and Social Sciences

https://doi.org/10.52825/CoRDI.v1i.381

Published: 07 Sept. 2023

MetaBelgica Project

A Linked Data Infrastructure Between Federal Scientific Institutes in Belgium

Sven Lieber[1] [https://orcid.org/0000-0002-7304-3787], Ann Van Camp[1] [https://orcid.org/0000-0002-1915-5956], Dieter De Witte[2][5] [https://orcid.org/0000-0001-8480-5719], Eva Coudyzer[3] [https://orcid.org/0000-0002-5985-7231], Erik Buelinckx[3] [https://orcid.org/0000-0003-1831-158X], Els Angenon[4] [https://orcid.org/0000-0002-8888-9662], Hannes Lowagie[1] [https://orcid.org/0000-0002-0671-3568], Julie Birkholz[1][5] [https://orcid.org/0000-0003-1193-0847], Karine Lasaracina[2] [https://orcid.org/0009-0006-1732-6607]

[1] Royal Library of Belgium (KBR), Brussels, Belgium

[2] Royal Museums of Fine Arts (RMFAB), Brussels, Belgium

[3] Royal Institute for Cultural Heritage (KIK-IRPA), Brussels, Belgium

[4] Royal Museums of Art and History (RMAH), Brussels, Belgium

[5] Ghent University, Ghent, Belgium

Abstract:

Keywords: FAIR Data, Linked Data, RDF, Wikibase, GLAM, Belgium

1 Introduction

Trustworthy metadata about entities related to cultural heritage is important to correctly identify bibliographic metadata, attribute works, identify works of public domain (based on contributors' date of death), or to support Named Entity Linking (NEL) for digitised documents. However, in Belgium such data is currently dispersed between numerous institutions, modelled with different ontologies, represented in different formats and curated in different languages. GLAM institutions would profit from data about Belgian entities to correctly annotate their collections. Furthermore, this would facilitate research in Belgium and abroad. The current situation does not only complicate data exchange in a national and international setting, but also leads to duplicate data curation efforts and data quality issues, negatively impacting the users' experience.

This paper introduces the *MetaBelgica* project, coordinated by KBR - The Royal Library of Belgium with the Royal Museums of Fine Arts, the Royal Institute for Cultural Heritage, and the Royal Museums of Art and History, to improve the status quo. The aforementioned *Federal Scientific Institutes (FSIs)*, belonging to the *Belgian Science Policy Office (BELSPO)*, joined forces to develop a shared Linked Data platform for managing shared entities. This platform, based on Wikibase (the technology behind Wikidata [1]), aims to ensure FAIR data about *persons*, *organisations*, *time/events* and *locations* related to Belgian cultural heritage. We will integrate data about millions of

Belgian entities from the four participating FSIs into a Wikibase instance and make it accessible by using Persistent Identifiers. This will not only professionalize our own data management and improve data quality, but according to received project support letters, also impact other regional, national and European institutions.

The aim of this paper is to introduce the project and its methodology as to to obtain early feedback and exchange information with the Research Data Infrastructure community. The *MetaBelgica* project will kick-off in 2023 and will run until the end of 2026. Section 2 presents the methodology - anchored in the state of the art - and section 3 concludes this paper.

2 Methodology

The proposed methodology aims to create a sustainable entity management system for Belgian entities to serve different stakeholders, and therefore covers (i) the integration of existing FSI data, (ii) a platform to manage the data in the long term, and (iii) providing data access to different stakeholders. The proposed methodology consists of five pillars that are translated into iteratively executed work packages and tasks, see fig. 1.

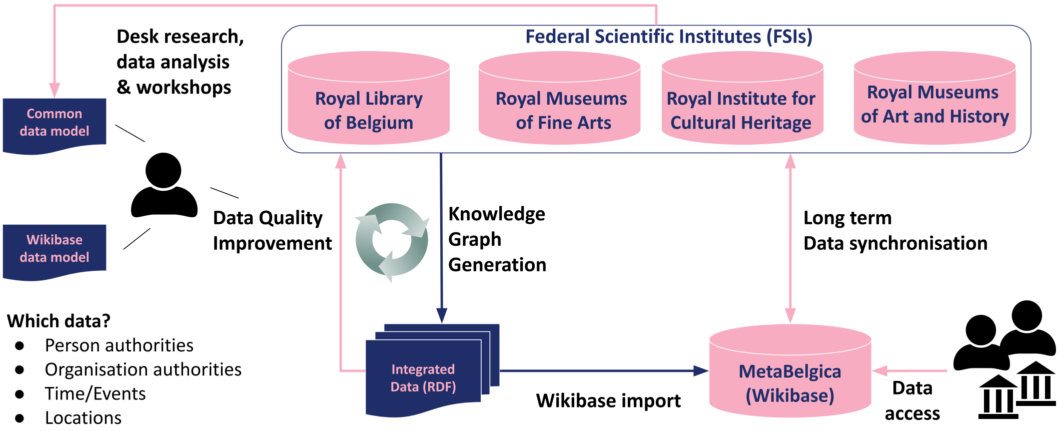

Figure 1. An illustration of the MetaBelgica setup: Data from the four FSIs is integrated according to a common RDF data model. Based on an additional mapping to a Wikibase data model, the data will be ingested into a Wikibase instance from which it can be collaboratively curated by the FSIs and from where it will be accessible.

Review current practices and data, and develop a common data model By using desk research and workshops, we will review the state of the art to represent the selected entities in an interoperable fashion. For example by using regional, national and international standards such as OSLO [2], FedVoc [3], or the Europeana Data Model [4]. Furthermore, we will analyse the existing data and divide it into different quality categories to ease further integration steps and link suitable concepts of international taxonomies.

A divide and conquer approach to integrate data in an iterative fashion By following state of the art knowledge engineering activities, we will iteratively integrate FSI data, e.g. with declarative RML mapping rules [5] for Knowledge Graph Construction of heterogeneous data. We will determine different integration strategies based on the identified quality categories, i.e. semi-automatic integration for high quality RDF data based on identifiers (with a human-in-the-loop [6]) or a manual curation for lower quality data.

Collaborative entity management software To enable coordinated data curation between FSIs, we will set up a collaborative Wikibase system that can be used for data curation in the long term and integration of low-quality data via reconciliation during the project. We develop a Wikibase data model and investigate suitable data ingestion methods such as the recently announced Wikibase REST API or other automatic means [7]. The open source software Wikibase is already tested in other projects [8] and comparable international initiatives such as (i) the *French National Entities File* [9] by the French national library BnF and the Bibliographic Agency for Higher Educcation (ABES), (ii) the *Integrated Authority File (GND)* in Austria, Germany and Switzerland [10], and (iii) The *Shared Authority File* in Luxembourg [11].

Long-term and sustainable data curation platform Longevity of the platform is a core feature of this project. Besides a stronger collaboration between the FSIs, we want to ensure the sustainability of *MetaBelgica* by integrating its operation into the functioning of each institution. We will set up organisational structures and provide internal trainings to ensure coordinated data curation with the new platform. Furthermore, we will implement technical components to ensure data synchronisation in the long term. Therefore we also have chosen for Wikibase: technical personnel is rare in the GLAM field, but data in a Wikibase can be maintained by non-technical personnel such as librarians or curators.

Data access via open license to increase impact We will ensure that the dispersed and heterogeneous data are presented to different users in a uniform and persistent way as open data. This includes surveying relevant stakeholders to provide appropriate FAIR data services. For these activities, we will follow the BELSPO Open Data directive ”as open as possible and as close as necessary”.

3 Conclusion

MetaBelgica has the aim to create FAIR entities of Belgian cultural heritage for worldwide use. The primary stakeholder group is the scientific community, but the platform can also be of use for society in general. Targeted users are scientific (researchers), cultural (GLAM-professionals and a broad public), educational (teachers and academics), technical (data aggregators) and economic (publishing professionals and creators) stakeholders.

With respect to the creation and operation of the presented platform we anticipate different challenges related to legal, organisational and technical interoperability. Challenges, that other Research Data Infrastructure projects (of other domains) likely encounter as well and for which best practices should be established.

Competing interests

The authors declare that they have no competing interests.

Funding

The presented project is funded by the Belgian Science Policy Office (BELSPO) in an Impulse action call to fostering the development of emerging research infrastructures within federal research institutions (INFRA-FED).

Acknowledgements

The authors like to thank the various stakeholders that supported the project proposal with a letter of intent.

References

[1] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, Sep. 2014, ISSN: 0001-0782. DOI: 10.1145/2629489.

[2] R. Buyle, L. De Vocht, M. Van Compernolle, *et al.*, "OSLO: Open standards for linked organizations," en, in *Proceedings of the International Conference on Electronic Governance and Open Society: Challenges in Eurasia*, St. Petersburg Russia: ACM, Nov. 2016, pp. 126–134, ISBN: 9781450348591. DOI: 10.1145/3014087.3014096.

[3] FPS BOSA, *Federal Vocabularies*, English, 2019. [Online]. Available: https://www.belgif.be/page/specification/fedservicesplatform.en.html.

[4] M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini, and H. Van de Sompel, "The Europeana Data Model (EDM)," in *World Library and Information Congress: 76th IFLA general conference and assembly*, vol. 10, 2010, p. 15.

[5] A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, and R. Van de Walle, "RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data," in *Proceedings of the 7th Workshop on Linked Data on the Web*, C. Bizer, T. Heath, S. Auer, and T. Berners-Lee, Eds., ser. CEUR Workshop Proceedings, vol. 1184, CEUR-WS.org, 2014. [Online]. Available: http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf.

[6] Lieber, Sven, Van Camp, Ann, and Lowagie, Hannes, *A LITL more quality: Improving the correctness and completeness of library catalogs with a Librarian-In-The-Loop Linked Data Workflow*, en, Nov. 2022. DOI: 10.5281/ZENODO.7372985.

[7] R. Shigapov, J. Mechnich, and I. Schumm, "RaiseWikibase: Fast Inserts into the BERD Instance," en, in *The Semantic Web: ESWC 2021 Satellite Events*, R. Verborgh, A. Dimou, A. Hogan, *et al.*, Eds., vol. 12739, Cham: Springer International Publishing, 2021, pp. 60–64, ISBN: 9783030804176. DOI: 10.1007/978-3-030-80418-3_11.

[8] J. Godby, K. Smith-Yoshimura, B. Washburn, *et al.*, "Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage," DOI: 10.25333/faq3-ax08.

[9] B. Bober and A. Angjeli, *Assessing Wikibase as the core for the French National Entities File (FNE)*, English, Berlin, Oct. 2019. [Online]. Available: https://upload.wikimedia.org/wikipedia/commons/1/16/Wikibase_for_FNE.pdf.

[10] B. Fischer and S. Hartmann, *GND meets wikibase*, English, Jul. 2020. [Online]. Available: https://www.youtube.com/watch?v=3vJrsUQIbV4.

[11] J. E. Labra Gayo, M. Pfeiffer, A. Waagmeester, *et al.*, *Representing the Luxembourg Shared Authority File based on CIDOC-CRM in Wikibase*, Semantic Web in Libraries Conference 2021, Dec. 3, 2021. [Online]. Available: https://www.youtube.com/watch?v=MDjyiYrOWJQ (visited on 04/21/2023).