VIVO-based Research Information Platform for the Berlin University Alliance

Fadwa Alshawaf, Claudia Adam, Rolf Guescini,
Florian Kotschka, Malte Dreyer
Humboldt University of Berlin, Germany
fadwa.alshawaf@hu-berlin.de

Abstract

This paper describes our efforts on developing a platform for presenting structured, transparent, categorized, and linked information about researchers and their research activities within the framework of the Berlin University Alliance (BUA). This will improve the discoverability of expertise, connect researchers to their work across disciplines and boundaries, and facilitate new research collaborations. The platform is established using the open-source, web-based VIVO software, which uses semantic web techniques to connect research outputs, organisations, people, things, and research activities. This is achieved by developing ontologies that relate and connect these entities within the alliance.

1 Introduction

Research information is an institutional asset that should be utilized to improve its visibility, research quality and range, and attract research cooperation. Research information refers to the metadata about projects, publications, published data sets, infrastructures and persons or teams. This information can be stored in local, internally accessible systems or in structured, transparent, and trusted platforms that connect and present research outputs, organisations, people, things, and research activities. Research data, however, are the products created in the research process such as publications, books, datasets, results, algorithms, and questionnaires are deposited as values, tables or documents in specialized repositories. Research information systems gather distributed research information from administration and science and enable a structured view of the infrastructure, services, and research outputs of an institution and its organisational units [1]. They can be implemented as specialized databases or as web platforms that combine databases with expert profiles such as the VIVO-based system [2].

According to the report published by the European commission in 2021, the research process is undergoing digital transformation, hence a reformed research assessment system is necessary [3]. This includes valuing cooperation, the acquisition of project money, and open access. One of our endeavors to facilitate the discovery and cooperation between individual researchers and across disciplines, is implementing a online portal for research information for the Berlin University Alliance (BUA). The BUA was founded 2018 by four Berlin universities; the Humboldt University of Berlin, the free university of Berlin, The technical university of Berlin, and the Charité (university hospital). The online portal provides a single point of access for information on scholarly activities across disciplines and beyond the boundaries of each organisation. The VIVO software is a member-supported, open source, semantic, web-based tool. It describes and links researchers (profile pages) and research (scholarly records). **Figure 1** show a diagram of a simplified system architecture that starts with the data harvesting, the workshop that contain all the software and the tools to get the system up and running, the backend in which the data are processed and prepared, the front-end where the data are displayed and browsed and the export function.

All information within VIVO is represented in the RDF data format and standard vocabulary via web ontologies with tools to edit existing ontologies and write new ones. The system links lists of various types of entities: People, activities, courses, events, organisations, grants, research and locations. It also has tools to record, edit, report, search, browse, visualize, analyze, adjust, and connect. The data can be ingested to the system manually and automatically using locally-stored files, organisation databases (HR), and online information systems (publication aggregation). VIVO's search returns faceted results for rapid retrieval of the desired, more specific information across disciplines. The research information should be documented in standard format such as the Common European Research Information Format (CERIF) or research core dataset format [4].

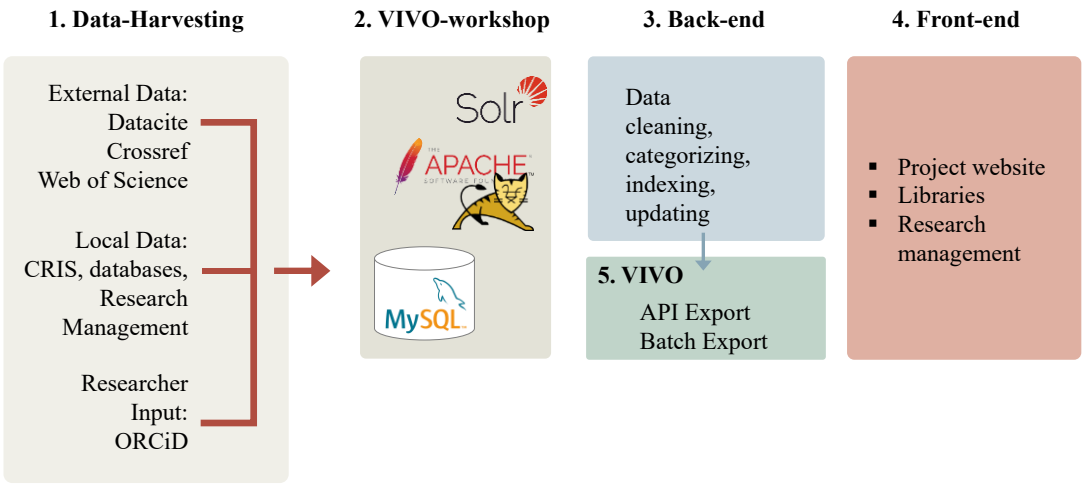

Figure 1: An overview of the VIVO-based platform

In Section 2, we describe a few of the opportunities that our platform provides for BUA researchers as well as some challenges. Then we present a use case in Section 3 followed by some specifications about the ontologies in Section 4. Section 5 provides a brief summary of the paper.

2 Chances and challenges

Information about researchers, research projects and publications (so-called “research information”) at the Berlin University Alliance (BUA) can be found online in various places. On the websites of individual institutes, for example, there are lists of contact information of employees, information about research projects and publications. However, the online distributed information is rarely linked together. For example, research information on websites of projects, funding organisations, ORCID profiles, or social networks may not be linked to the contributing researchers. The open-source software VIVO uses Semantic Web technologies to link information, which makes the information easier to find as well as the colocation of related information.

Using the VIVO software, data from existing systems can be automatically harvested. In this way, research information from different sources can be aggregated in a meta-platform. In addition, the software enables a search function so that the platform acts as a search engine to search for information, aggregated already from other systems ("One Stop Shop"). Since the project aims to develop a research information platform within the Berlin University Alliance (BUA), the described functions are even more effective; not only can information from different sources be aggregated and searched for one organisation, but across the alliance.

The open-source software VIVO provides diverse analyses and visualization tools that go beyond the capabilities of existing research information systems. Functions such as the visualization of the co-author network or the Map of Science can facilitate and enrich reporting as well as research management.

The browse and faceted search functions of the platform facilitates to search for and find relevant information and people - including researchers - and thus promote networking and exchange within one discipline and across different disciplines.

While the platform has substantial benefits for the alliance, it also has to overcome some challenges. The quality of the mapped data is an important success factor for the platform; therefore, the first step for the project team is to obtain an overview of the data available at the individual organisations and which data can be used for the platform. It is important to decide how the data should be accessed, cleaned, harmonized, mapped, and kept up-to-date.

Preparing a data protection concept for the platform is one fundamental challenge, because so far, there are diverse concepts within the BUA for the use of researchers data. In order to ensure and document compliance with the European General Data Protection Regulation (GDPR), an information security and data protection concept for the platform must therefore be developed in the first phase of the project. This is a hindrance we face in developing this platform because even if the information are openly available on organisations websites or online aggregation systems, we need to ask for a consent to import and display the information on the platform. A form to ask each researcher for consent to the use and publish of their information on the platform.

Another issue that should be mentioned within this context is the creation of representative taxonomies and ontologies that match the German academic system, especially within the context of the BUA. At the same time, ontologies for disciplines and interdisciplinary fields should be built, as discussed in section 4.

3 Use Case

The BUA relies on the synergy of partners’ strengths to open up opportunities for new innovative, cooperative research projects. Therefore, the focus within the alliance is put on promoting the

networking of researchers and their research activities and connecting many research groups within the clusters of excellence as well as beyond. The vivo-based research information platform aims at facilitating the digital connectivity and discovery of research and researchers of the BUA. An added value of our project lies in the search and navigation functions, which through the modeling of data and building of ontologies, can suggest and present relevant information that is otherwise not easily found. If we connect and relate BUA institutes, researchers and research activities, particularly the projects, using ontologies, this allows easier and faster discovery of organisations, people and research topics. The required information includes, for example, lists of research projects that are described using their metadata precisely. The metadata contain the project title, identifier, description, duration, keywords, URL, members, funding organisation, award (number/identifier), language, geolocation, cooperation partners (persons/institutions), which is turn have names, affiliations, identifiers, sectors (industry/research), geolocations, roles, etc. The grant projects are linked to organisations, persons, publications, events, which are represented using metadata too. Researcher profiles show contact information, title, affiliation, function, research fields and interests, grants, publications, social-media activities, and more, according to the desire of each researcher (c.f. Ch. 4). For the sake of demonstrating the potential and the benefits of the platform for researchers and administrators, we start with a small group of several hundreds of researchers, which will be extended into several thousands in the future.

The data can be acquired from different resources such as the researchers themselves. However, this method should currently be avoided for time limitations. Another method is to receive files of research information from the research management departments of each BUA partners, which requires manual labor and makes it difficult to keep the system up to date. This method cannot be completely avoided since not all research information are stored in current research information systems or databanks yet. They are delivered in excel spreadsheet and other file formats. The third method is to use data harvesting though APIs, which simplifies the data update, cleaning, and harmonization. The research information can be acquired from online records aggregated by Datacite and ORCiD or local research information systems. To minimize the manual import of data, scripts are developed that aggregate data from different APIs. It is important however to consider that different systems have different structures and semantics; therefore, a mapping of the attributes between the systems is indispensable. To reduce complexity, we choose to limit the attributes that describe each entity (e.g., projects, publications, and persons).

To facilitate the browsing of relevant information, we rely on categorising the research entities under disciplinary and interdisciplinary research. The first is done using the disciplines list created by the B2FIND project [5] and for the latter we use the list of interdisciplinary research fields created for the research core dataset [6]. These are represented using ontologies as described in Section 4. Besides, we use keywords suggested by the researchers themselves or recognized through text mining in the abstracts and brief descriptions of the research projects.

4 Ontology and data ingest

VIVO uses a collection of ontologies to represent and connect researchers to their research activities. In computer science, an ontology refers to how to name and relate types, properties and relationships that exist in a particular domain such that they are interpretable in this specific context not only by human beings but also by machines. The VIVO ontology represents researchers by connecting them to their activities and accomplishments, resources, and institutions. This is done using a set of types and relationships (properties) that represent researchers and the entire context in which they work. These

ontologies model organisations, concepts (research areas), persons, publications, grants, and others. The ontology requires precise identification of people, organisations, and works, and the correct relationships between these three.

The VIVO ontologies were originally created based on the academic system of USA and in the medical environment. Therefore, in addition to the existing VIVO ontologies that we edit to meet our requirements, we first created an ontology that describes the organigram of the alliance, i.e., the Humboldt-Universität zu Berlin, Freie Universität Berlin, Technische Universität Berlin, and the Charité – Universitätsmedizin Berlin. The ontology requires accurate representation of the institutions and elements that do not fully match for all partners. A part of the organigram ontology is shown in **Figure 2**.

In addition, we created ontologies for the academic disciplines and interdisciplinary fields to categorise the scientific work and make it easily findable. The academic disciplines are represented using the list provided by the EUDAT metadata indexing service, B2FIND [5]. The interdisciplinary research fields were represented using the list in [6] that was presented within the frame work of the project: "Developing a classification of interdisciplinary research fields for the research core dataset" [4]. There is a great challenge faced in this regard, namely how to categorise big clusters of excellence that cover a wide scope of research under interdisciplinary research fields.

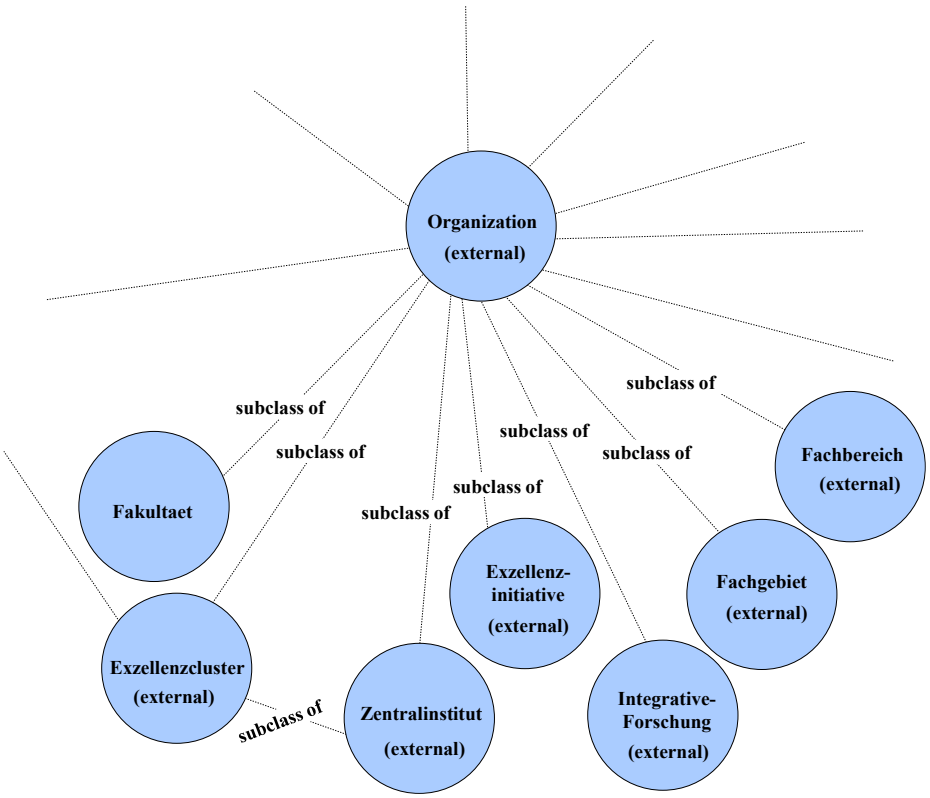

Figure 2: A part of the ontology that represents the BUA organigram.

In order for VIVO to represent the BUA properly, data integrity, i.e., the maintenance and the assurance of data accuracy and consistency, is essential. The importing of data from four organisations with their respective research information systems is challenging. To achieve this task, we have

developed different automated and semi-automated methods to ingest data from different sources into our system. Data pipelines will connect to system APIs of the three universities and the Charité whenever applicable. When the required information is not available in those systems, manual update schedules have to be implemented. Smaller research projects and clusters that are still not connected to bigger research information systems will have to provide their data in excel, json, yaml or other formats or provide a connection to their internal databases. This data will be cleaned, structured and uploaded in VIVO readable formats. Corresponding with EU data protection policies, this project will also have to deal with data that shall not be open.

5 Conclusions and Outlook

In this paper, we gave a brief description of the VIVO-based research information platform that is created to showcase expertise and research activities within the Berlin University Alliance. The platform uses semantic web technologies to best connect researchers and their work to make them easily discoverable within the system and in search engines. This should encourage and facilitate cooperation, main transparency, and build bridges across disciplines.

6 Acknowledgement

Thanks to the Federal Ministry of Education and Research (BMBF) and the state of Berlin under the Excellence Strategy of the Federal Government and the Länder for funding this project.

7 References / Citations

[1] Ebert, B., Tobias, R., Beucke, D., Bliemeister, A., Friedrichsen, E., Heller, L., Herwig, S., Jahn, N., Kreysing, M., Müller, D., & Riechert, M. (2016). *Forschungsinformationssysteme in Hochschulen und Forschungseinrichtungen. Positionspapier.* Version 1.1. Zenodo. https://doi.org/10.5281/zenodo.45564

[2] Mitchell, S., Chen, S., Ahmed, M., Lowe, B., Markes, P., Rejack, N., Corson-Rikert, J., He, B. and Ding, Y. (2011). *The VIVO ontology: enabling networking of scientists.*

[3] European Commission, Directorate-General for Research and Innovation, *Towards a reform of the research assessment system: scoping report*. (2021) Publications Office. https://data.europa.eu/doi/10.2777/707440

[4] Biesenbender, S. and Hornbostel, S., 2016. *The research core dataset for the German science system: Developing standards for an integrated management of research information*. Scientometrics, *108*(1), pp.401-412.

[5] https://github.com/EUDAT-B2FIND/md-ingestion/blob/master/etc/b2find_disciplines.json

[6] https://www.kerndatensatz-forschung.de/index.php?id=forschungsfelder

Fadwa Alshawaf received her Ph.D. degree in Engineering (remote sensing and geodesy) from Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, in 2013. Since 2021, she has been leading the project Research Information Platform using VIVO at Humboldt-Universität zu Berlin. From 2010–2020, she has been working as a researcher at KIT and the German Research Centre for Geosciences. Her research interests include water vapor mapping using interferometric synthetic aperture radar and GNSS, improving the quality of these maps by statistical data fusion with numerical weather models, analysis of water vapor trends across time. Fadwa has a certification in systemic coaching and leadership. Her most important value is to inspire individuals to lead a life of success and contribution.

Claudia Adam contributes to the project "Research Information Platform using VIVO" as a Community Manager. In this role, she handles communications and aims at building a community around the platform for Berlin University Alliance, for example with the online event series VIVO Talks. Claudia studied Media and Communication Studies in Mannheim, Berlin and Toulouse. In her master thesis at Humboldt-Universität zu Berlin, she studied community building through a virtual platform as a mechanism of communication. Claudia brings in various experiences from Journalism, Event Management and Content Marketing, where she has developed strategies, concepts and content as well as managed content creation processes. For more information, please visit Claudia's Linkedin profile.

Rolf Guescini is an Information architect and developer with a background in Humanistic informatics and lingustics. The interest in designing and organizing information brought him to develop a deep interest in the works, ideals and goals of the early hyper text pioneers. The research into the need for free and open distributed information on one hand and the information overload block on the other hand, consequently led into research in semantic web technologies such as Topic Maps and RDF. He sees structured meta data, formal ontologies and folksonomies as possible solutions to our modern informational challenges, especially within the domains of Open Research and Free Digital Learning.

Florian Kotschka is the System Administrator of the 'Research Information System with Vivo' Project at the Berlin University Alliance (BUA) and the IT center (CMS – Computer and Media Service) of Humboldt-Universität zu Berlin, Germany. Furthermore, he is developing research and learning platforms for the department of history of science at HU Berlin. Prior to joining BUA Mr. Kotschka was involved in creating and maintaining the repository for the Cluster of Excellence Topoi in Berlin. Having an academic background in ancient History and Philosophy, he also worked as a freelance Data Scientist on topics connected to open research and open data in the field of digital humanities.

M. Dreyer elected to the EUNIS Board of Directors till June 2020.
He is the Director of the Computer and Media Service of Humboldt University Berlin, Germany. Before he was the Director for the Department of Research and Development at Max Planck Society, Max Planck Digital Library. He designed and developed research and publication data infrastructure for the Max Planck Society's institutes, as well as many research tools. Within several major German, European and international projects he was active in the areas of digital research infrastructure, repositories, virtual research environments and software architecture across many scientific disciplines.
Providing advise on software architecture, he is a member of several technical boards. He was a member of the German national alliance initiative working groups for research data and virtual research environments as well as a member of commission "Zukunft der Informationsinfrastruktur in Deutschland" and DINI e.v. member of the board.
Malte Dreyer's interests now are in the field of scalable information management architectures and infrastructures in the intersection of organisational perspectives on ICT from data centres and information management organisations. Current projects are in the fields of research data management and repositories, linguistics, as well as cloud architectures and integrated cloud services.