Developing Crosslingual Ontologies in WissKI: Transcontinental Research Collaboration in the Africa Multiple Cluster of Excellence

MLO Modern Languages Open

SPECIAL COLLECTION: NEW DIRECTIONS IN DIGITAL MODERN LANGUAGES RESEARCH

ARTICLES – DIGITAL MODERN LANGUAGES

LIVERPOOL UNIVERSITY PRESS

PHILIPP EISENHUTH
MYRIEL FICHTNER
BRITTA FREDE
RÜDIGER SEESEMANN

*Author affiliations can be found in the back matter of this article

ABSTRACT

Since July 2019 the Africa Multiple Cluster of Excellence at the University of Bayreuth has been funding a project dedicated to the creation of an Islamic Cultural Archive (ICA), a database designed for collaborative research in English, French and Arabic connecting researchers based in Germany and four African countries. The ICA pursues individual but interconnected studies revolving around Islamic learning in Africa. With the support of IT specialists, the project team developed a platform that allows us to handle our data through an ontology-based digital research environment building on WissKI, a set of modules that combines the Drupal content management system with semantic web technology. This article discusses the technical implications of our endeavour to connect diverse languages and data from various digital and digitized media. The research team members link their datasets to create synergies between various research foci and interests, ranging from the nexus of Islamic knowledge production, dissemination and acquisition to the socio-religious, political-economic, and cultural dimensions of Islamic learning. The article shows how our platform allows us to collect and archive different types of data, generate metadata through a shared ontology, and connect data beyond language barriers. Most notably, our data description method links the data through multilayered and multilingual tags, as well as through comprehensive cross-references, thus constituting an innovative way of data handling that can benefit researchers in Islamic Studies as well as cultural and literary studies more broadly.

CORRESPONDING AUTHOR:
Rüdiger Seesemann
University of Bayreuth, DE
ruediger.seesemann@uni-bayreuth.de

TO CITE THIS ARTICLE:
Eisenhuth, Philipp, Myriel Fichtner, Britta Frede and Rüdiger Seesemann 2023 Developing Crosslingual Ontologies in WissKI: Transcontinental Research Collaboration in the Africa Multiple Cluster of Excellence. *Modern Languages Open*, 2023(1): 39 pp. 1–16. DOI: https://doi.org/10.3828/mlo.v0i0.445

1 CONTEXT OF THE ISLAMIC CULTURAL ARCHIVE PROJECT

This article presents the collaborative digital research practices and their technical underpinnings as developed in a transcontinental and multilingual research project titled "Toward an Islamic Cultural Archive" (ICA) and conducted under the umbrella of the Africa Multiple Cluster of Excellence at the University of Bayreuth, Germany.[1] The overarching aim of the ICA is the establishment of a dynamic archive of Islamic culture in Africa based on innovative digital working formats. The project team members come from Islamic Studies, Religious Studies, Anthropology and Arabic Literary Studies. Apart from Germany, they are based in Kenya, Senegal, Mauritania and Tunisia. In addition to these countries, research is also conducted in Tanzania.

Although laid out as a long-term project covering the entire range of Islamic cultural expressions in Africa, the thematic focus of the current project phase is on Islamic learning. Each team member contributes to the project by collecting diverse sets of data – notes, interviews, reports, statistics, (text-)books, pamphlets, newspapers, magazines, photographs, videos and many other kinds of written and audiovisual documentation – pertaining to the overarching theme of learning and knowledge transmission in Islamic contexts.

In the current ICA, research is conducted in the following settings and concerned with the following themes and questions:

- The case study by Mohamed Mraja in Tanzania explores how Islamic educational institutions (such as schools, colleges, universities, mosques, Muslim organizations, Sufi orders, media houses, libraries, bookshops and publishing houses) and their networks contribute to the creation, production, dissemination and acquisition of Islamic knowledge.
- The first Kenyan case study by Rüdiger Seesemann concerns the contents and practices of learning in Islamic educational institutions in Kenya. Relevant data include information on the location, history and size of the institutions, as well as the physical and material setting, the subjects studied, the books read and the pedagogies used.
- The second Kenyan case study by Hassan Ndzovu focuses on Islamic institutions in Kenya led by women, with special emphasis on their teaching activities and sermons. Particular attention is paid to the mediatization of the sermons, that is, the way media technologies are used to disseminate the sermons and the concomitant effects on the preachers, the audience and the sermon itself against the backdrop of a male-dominated religious landscape.
- The Senegalese case study by Abdourahmane Seck aims to produce an overview of the initiatives, corpora and infrastructural frameworks, but also of the symbolic, socio-religious, political-economic and cultural issues that characterize the recent and ongoing history of Islamic learning in Senegal. Ethnographic methods are especially prominent here, with the objective of including "lived Islam" in the analysis.
- The Mauritanian part of the project is conducted by Fatimetou Abdel Wahabe and relates to the so-called *madīḥ* circles, that is, the congregational performance of panegyrical poetry in praise of the Prophet Muhammad, which are used as lens to gain insights into learning processes. Of special interest are the sites and occasions of *madīḥ* circles, the genres and texts (including the sources drawn upon), performative aspects, and the dimension of knowledge transmission connected to such congregations.
- Islamic knowledge acquisition among adults is the topic of Britta Frede's comparative study of Mauritanian and Kenyan urban settings, with a particular focus on women (both as students and teachers). In addition to examining the role of learning circles, mosques and libraries in the transmission and acquisition of Islamic knowledge, the study also seeks to understand how adults use and appropriate media for learning purposes.

1 This article comes out of a project pursued in the framework of the Africa Multiple Cluster of Excellence at the University of Bayreuth, which is funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) under Germany's Excellence Strategy – EXC 2052/1 – 390713894. It is the result of a joint writing exercise in which we divided the labour as follows: Philipp Eisenhuth was in charge of section 4, with the exception of sub-section 4.3; Myriel Fichtner authored section 3 and large parts of section 5; Britta Frede wrote sub-section 4.3 and contributed to section 5; Rüdiger Seesemann provided the initial drafts for sections 1 and 2, which were further expanded by Britta Frede. He also put the final touches to the conclusion, based on notes from the other co-authors.

- The educational and biographical itineraries of sub-Saharan scholars and students of Islam in Tunisia is the focus of the first Tunisian case study by Ramzi Ben Amara. Matters of interest also include the development of the religious curriculum in government-run schools at different levels of learning (primary, secondary and university).
- The second Tunisian case study by Franz Kogelmann is devoted to the reforms of Islamic education and its sponsors, with a focus on state-controlled educational institutions. Here, it is crucial to explore the strong influence of the Islamist opposition, especially in the 1980s and 1990s, as well as the new factors that emerged in the field of Islamic education after the revolution of 2011.

By bringing together such diverse topics and data, the research team seeks to open new perspectives on processes, conceptualizations and transformations of Islamic learning in selected African contexts. We use a wider lens on the subject by considering formal as well as informal ways of learning and by identifying the political, social, economic and cultural dynamics at play in this research field. Some of our team members are interested in the nexus of Islamic education and religious politics (Ben Amara, Kogelmann), while others focus on institutional and conceptual aspects of formal Islamic education (Mraja, Seesemann). Two projects foreground a gendered perspective on the subject (Hassan, Frede), and the West African case studies both approach Islamic learning as a lived practice against the backdrop of rapid socio-economic transformation (Seck, Abdel Wahabe).

While all team members are experienced researchers in their respective fields, embarking on this joint project meant breaking new ground in research data management. We soon realized that the challenges we faced required tailor-made solutions. The most important needs were the adoption of a coherent metadata structure and tagging system, the development of IT ontologies, and the building of a crosslingual working environment.

In this article, we will outline how we were able to develop the tools needed to make the ICA database feasible and functional, especially with regard to crosslingual collaboration and IT ontologies. The next section lays out the demands we identified for our joint digital research platform.

2 REQUIREMENTS FOR THE DIGITAL PLATFORM

At the outset of the project, the participating researchers identified six major demands to ensure the functionality of our digital platform. These demands led us to draw on the expertise of two PhD students in computer science, whose support was invaluable in creating a database capable of tackling the specific needs of our transcontinental, multidisciplinary and multilingual research team.

The first and most immediate demand relates to the necessity of establishing links between the sub-projects with their respective focus and approach to Islamic learning. The task thus consists of integrating heterogeneous analogue and digital data, both qualitative and quantitative, into a common digital research platform accessible to all researchers involved. Key to this endeavour is the creation of a joint database, which connects our multiple research perspectives and thus allows us to identify particularities and commonalities of learning processes in the diverse settings under study. Establishing such connections between data can shed new light on the interrelations between informal and popular Islamic knowledge practices on the one hand and the more formalized and scholarly learning environments on the other. Such findings also have the potential to contribute to theory-building.

A research design where the members of eight sub-projects simultaneously collect data in five countries requires a coherent tagging system that can be adapted to diverse disciplinary practices and to the various approaches to the field of Islamic learning. We therefore started with a joint brainstorming exercise designed to help us in “mapping the field” by identifying common concepts that could serve as nodes to interconnect the data collected in the sub-projects. These concepts needed to be kept flexible and open to avoid a common problem of such taxonomical approaches to knowledge, whereby working with predefined concepts tends to influence research trajectories and fix the categories at an early stage, thus narrowing the paths that might lead to innovative findings.

To avoid these pitfalls and keep our concepts open, we organized our approach to the field of Islamic learning along three perspectives: a physical perspective, a meta-perspective and a contextual perspective, which each contributed to describing the content of the data with terms that serve as nodes connecting our heterogeneous datasets. We aligned the concepts along these three perspectives and linked them in a way that allows all team members to import their data and connect them further by adding more conceptual nodes within the broader logical structure.

The second demand to make our database functional relates to the multilingual character of our endeavour (see Figure 1). The members of the research team work with three major languages: English, French and Arabic. In addition to the Latin alphabet, we also use the Arabic script and transliteration letters. We therefore needed a digital tool that allows switching between several working environments and, at the same time, is capable of handling various alphabets. This necessity was the more urgent as we had to adapt our research design to the travel restrictions caused by the COVID-19 pandemic, which led us to create larger local teams whose members collected data in the five research regions. Our multilingual research tool facilitated the integration of researchers and research assistants with limited command of English.

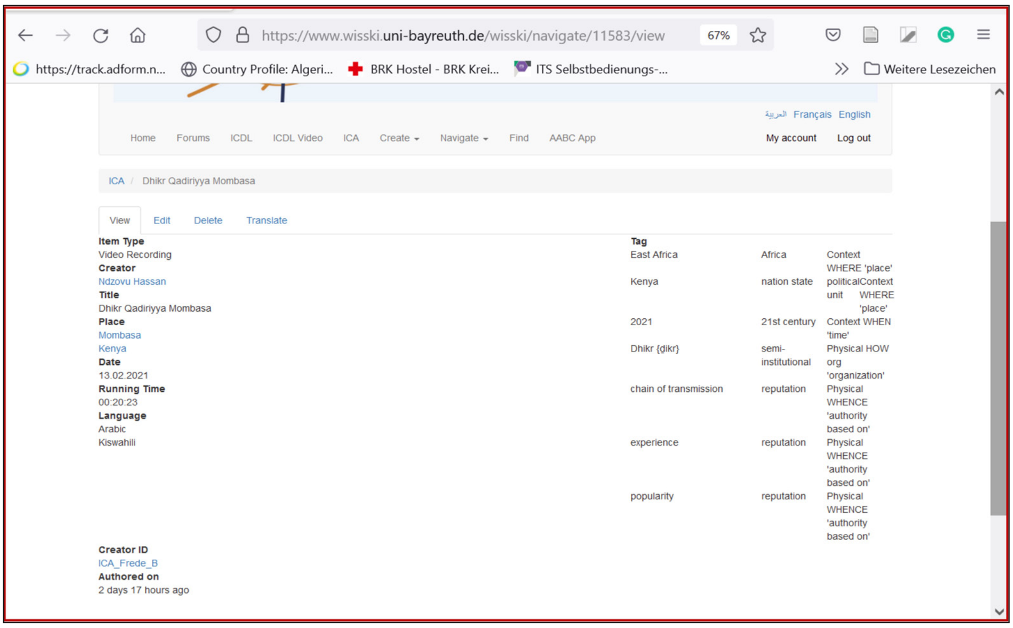

Figure 1 Screenshot of data entry, showing the "translate" function. On the upper right, users may choose the language for the data input mask. After data input, users may click on the "translate" button for an automatic translation of the tags into all three working languages, thus allowing for crosslingual searching of the database.

However, the members of our research team not only use different working languages, they also handle data sources in African languages, reflecting the multilingual character of the African continent. While former colonial languages such as French and English continue to be used among Muslims in Africa, languages such as Wolof and Kiswahili (to mention only those that are relevant in our sub-projects) play a preponderant role, as do local varieties of Arabic in addition to classical and modern standard Arabic. As not all team members are able to work with all languages featuring in the collected data, we require a tagging system that is crosslingual, that is, a system that allows us to search all data in all three working languages. The tagging system also needs to be capable of connecting the data beyond language barriers by referring to the common concepts that cover the broad array of Islamic learning and are shared by the various sub-projects.

The latter requirement leads us to yet another demand our digital working environment needs to fulfil, that is, the creation of a sustainable data management system that supports a structured approach to the handling of data in our research field. This demand not only responds to new standards in good research practice set by funding agencies, but also considers the specific context of research and data collection on the African continent. The state of existing archives in Africa continues to pose significant challenges for researchers. In African Islamic contexts, archival practices rarely go beyond private initiatives, often in connection with manuscripts or smaller libraries that cover a broad range of texts and documents ranging from private letters, accounting records and contracts to religious texts and poetry. Systematic collection and archiving efforts of governmental or academic institutions have rarely fully developed.

In general, archives remain scattered due to financial constraints, lack of research support and inconsistent approaches to cultural policies. Therefore, some archives in Europe have greater holdings of materials from Africa than collections and libraries on the African continent (Madore 7–8). This underscores the need for a more systematic approach to collecting and archiving through the creation of digital resources that can accommodate materials and data of different provenance and, at the same time, offer easy access to researchers from all over the world, especially those from Africa who do not have the financial means to travel to consult collections in the US or Europe.

However, not all data collected are suitable for sharing. There are instances of confidential and sensitive data, such as statements made in interviews or video or audio recordings featuring controversial scenes. Researchers need to protect the dignity and safety of their respondents and interlocutors, and they need to follow the established regulations for storing research data. The endless possibilities for linking and sharing offered by digitization require us to pay even more attention to research ethics, with clearly formulated agreements between project team members, research assistants, interlocutors and respondents.

Apart from dignity and safety issues for the local interlocutors, the structural architecture of academic research environments needs to be taken into account as well. Doctoral researchers and postdoctoral researchers without stable positions especially need to be assured that they can benefit from their collected data first before they are shared with a broader public. Otherwise, exploitative working environments will be (re-)established, thus reinforcing the existing power imbalance in academia that puts into question the often-invoked academic freedom and transforms it into freedom for some researchers at the expense of others.

These constraints do not affect the entire body of the collected data, but in all instances where they do matter, the system needs to enable the researchers to define the extent to which the data may be shared within the team and beyond. This implies establishing clear rules for the fair and transparent handling of access and usage rights, by determining what is publicly accessible and what is not, what can be accessed by the team members and what is limited to the use of the researcher who collected a specific set of data.

These considerations notwithstanding, the database required for the ICA needs to be set up not only as a node between the researchers and sub-projects involved, but also as a link connecting the various institutions that form part of the Africa Multiple Cluster of Excellence. The principal idea of the ICA is to establish a database as a long-term project that will go beyond the thematic focus on Islamic learning. As such, we intend to open it up to researchers at our partner universities and eventually to the broader field of African Studies.

A final demand that guided our decisions in setting up our database pertains to questions of IT ontologies. Usually in IT systems, storage is organized into relational and hierarchical databases and often follows pre-set classification protocols, based on fixed “ontologies”, here referring to “the modes by which knowledge is articulated, expressed, interpreted, and formalized” (Srinivasan 204). While this procedure is feasible in contexts where similar types of data are generated by uniform methods, the ICA’s research data are more varied. Moreover, the structures of conventional databases emerge out of academic abstractions with roots in Western knowledge traditions. Rather than mirroring the fluid and dynamic aspects of actual lifeworlds, they tend to be static and ignore other forms of knowledge production. Our challenge thus lies not only in conjoining the epistemes of different academic disciplines, but also in bridging these and other kinds of knowledge production based on research data and their ontologies, thus going beyond mere “buttonology” (Russell and Hensley). For this purpose, we need to correlate metadata derived from diverse ontologies, thereby allowing for ontologies that overcome rigid classifications. The question of ontologies will be addressed in more detail below.

The next section introduces the virtual research environment in WissKi and elucidates its suitability for the demands just outlined.

3 THE VIRTUAL RESEARCH ENVIRONMENT BASED ON WISSKI

Virtual research environments (VRE) can be described as digital working infrastructures designed to support collaboration between researchers. According to the UK’s Joint Information Systems Committee (JISC), the purpose of a VRE is to help researchers from all disciplines to work collaboratively by managing the increasingly complex range of tasks involved in carrying out research on both small and large scales. Thereby, VREs cover multiple aspects pertaining to cooperation by providing opportunities to share research findings and expertise across organizational and operational boundaries (Candela et al.). This includes functionalities to store, link, share, discuss and browse through research data while considering all aspects of security.

VREs should support work across disciplines over the entire research lifecycle, enabling researchers from different geographical locations to interact with each other (Carusi and Reimer). From a technical perspective, this requirement can be implemented differently, whereby web-based applications currently constitute the most widely used option (Keraminiyage et al.). Even though VREs are used in many disciplines, arts and humanities have come to take a special interest in such digital infrastructures (Carusi and Reimer).

In the arts and humanities, several VREs and digital infrastructures have emerged, including HuNI, Alveo (Cassidy), Parthenos (Frosini et al.), heiMAP (Pfeiffer et al.), and DARIAH (Dariah-EU). Whereas these systems were developed to respond to different requirements and thus have a focus that may vary from one system to another, they all must provide stored or linked data in an appropriate way, thus fulfilling the FAIR principles, that is, findability, accessibility, interoperability and reusability (Wilkinson et al.). Therefore, metadata schemata were developed for the VREs to describe data regarding their structure and content. This includes descriptions of the semantics or meaning of the data. The relations between semantic concepts are structured in so-called ontologies.

In our ICA, building an ontology-based tagging system was one of the major tasks to be accomplished. One of the VREs that fulfils our requirements is WissKI. This is an acronym for the German term *Wissenschaftliche Kommunikationsinfrastruktur*, which can be translated as "scientific communication infrastructure" (Hohmann and Schiemann; Scholz and Görz). The software development for WissKI was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) starting in 2008, and is currently supported by the NFDI4Culture Consortium within the National Research Data Infrastructure. It is a joint product of the Digital Humanities Research Group of the Department of Computer Science at Friedrich-Alexander University in Erlangen, the Department of Museum Informatics at the Germanic National Museum in Nuremberg, and the Biodiversity Informatics Group at Zoological Research Museum Koenig in Bonn. WissKI also receives support from the Association for Semantic Data Processing (IGSD e.V.) and has a continuously growing community of users.

Although WissKI was originally developed to provide solutions for museums, it is now used in many research projects that manage large amounts of data and collections involving a wide range of disciplines, including but not limited to arts, music, history, literature and geography. Colleagues in the research project "Objekte im Netz" based at Friedrich-Alexander University in Erlangen and conducted in collaboration with the Germanisches Nationalmuseum especially encouraged us to work with WissKI. In the "Objekte im Netz" project, WissKI was used to manage research data collections from 25 different fields through the use of ontologies. A major challenge for this project was the digital acquisition, networking and visualization of heterogeneous data across collections and disciplines, while considering their (subject-)specific characteristics.

Technically, WissKI is implemented as a set of modules of the Drupal content management system (CMS). Drupal is a free and open-source web application framework used to manage websites worldwide. A Drupal module can be described as a new feature or functionality that extends the code of Drupal, while each user can create or use modules created by others. Most modules can be combined with each other, thus constituting a system that can be tailored to any needs. Furthermore, such modules benefit from all the characteristics and advantages provided by Drupal, such as detailed permission and user management, a configurable graphic interface, and a highly active community of users.

In technical terms, WissKI provides a knowledge base by extending Drupal with a Triplestore (Hohmann and Schiemann). This is a type of graph database that saves semantic facts by storing data in a three-part subject–predicate–object structure and also supports ontologies (Ontotext). In this manner, WissKI allows for the integration of ontologies by connecting a CMS to a graph database. The CMS (in our case Drupal) manages data by using 1) relational databases to manage administrative data and 2) graph databases (in our case a Triplestore) for the semantic back-end, as illustrated in Figure 2.

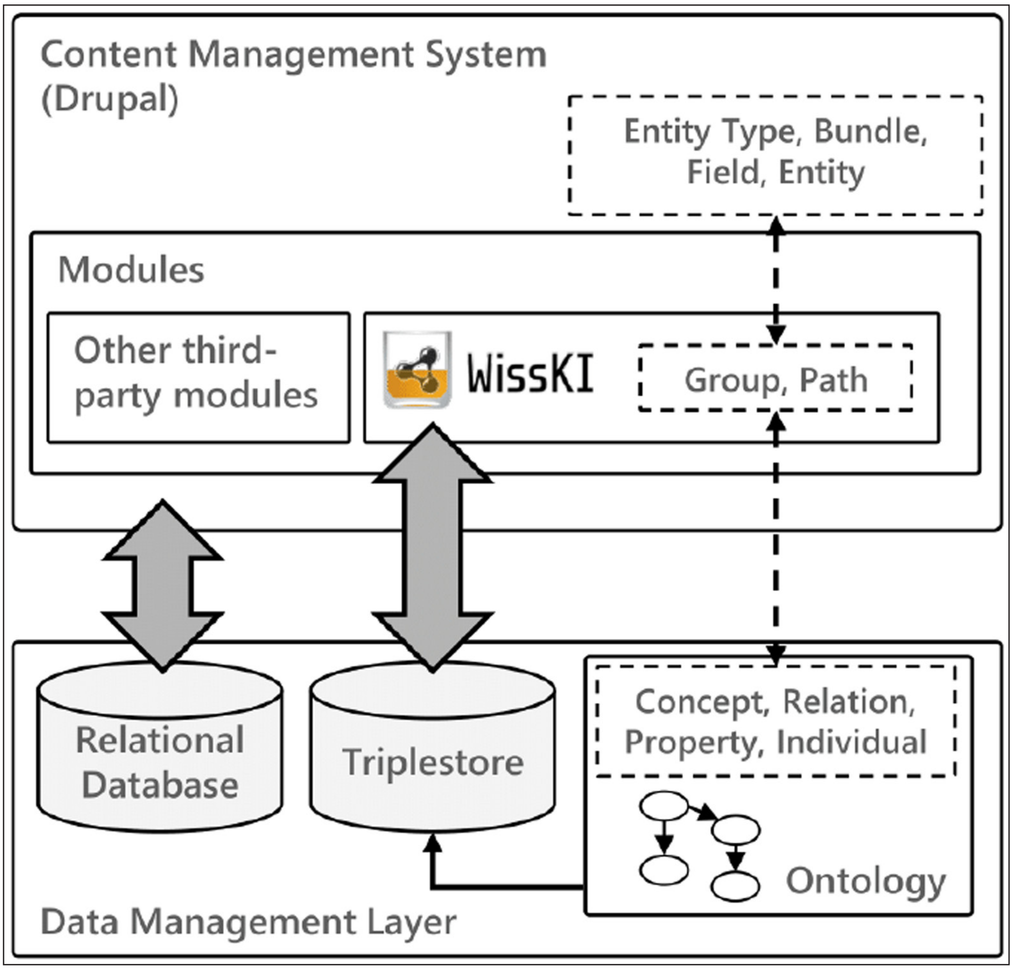

Figure 2 The system architecture of WissKI [Huisl 2022].

In Drupal, each content is named as an entity that constitutes a single instance of a bundle. Bundles are, in turn, defined as a sub-type of entity types, meaning that two bundles that belong to the same entity type behave in the same way. However, bundles differ in the fields they contain, as the fields store different types of information, for example texts or images. To enable the integration of ontologies into Drupal, bundles and fields need to be mapped to the structure of ontologies through their concepts and properties.

This key challenge is met by a core component of WissKI, the so-called pathbuilder. It manages the mapping between the graph-based ontology and Drupal's content structures by providing functionalities to create groups and paths. In other words, the pathbuilder links the entities, that is, research data, to the ontology and therefore enables their semantic connection. A detailed description of this functionality is offered in Fichtner and Ribaud.

4 ONTOLOGIES AND META-TAGGING

4.1 ONTOLOGIES AND CONTROLLED VOCABULARIES

This section gives the basic definitions – mostly from the technical perspective of computer science – to provide a better understanding of the implementation of our system for research data management.

We distinguish different kinds of controlled vocabularies. They vary in their relational complexities and may comprise the entire range from glossaries to thesauri, taxonomies and ontologies (Wohlgenannt 37). Despite some similarities, they are usually used for different purposes. Their capacity to describe complex relationships between the covered terms differs greatly, depending on the availability of semantic constructs. Expanding the basic vocabularies or glossaries, which are simply collections of various term without any relationship, taxonomies comprise a restricted set of predefined constructs. They introduce relations between objects and concepts as "is a", thus categorizing them in hierarchical structures that often consist of classes and their potential sub- or super-classes. Thesauri extend these structures with additional elements by listing synonyms and associations for the terms, included adding relations such as "is equal to" or "is associated with", thereby offering richer semantics. Ontologies allow for arbitrary (i.e., self-defined) relations between the included elements (i.e., the terms, their associated classes, and potentially even the relationships themselves) and are therefore the most expressive type of these three.

Usually, all these vocabularies are used for defining and describing available subjects, which in turn can be used to further describe and classify specific objects and items drawn from the research data (Garshol). Ontologies have an additional capacity of expressing relationships, since our VRE is built upon a semantic back-end in the form of a Triplestore, where the ontologies also serve as the describing schema for the data, as outlined in greater detail below.

Originally coming from the field of philosophy, ontologies were defined as the "science of being" (Smith). They are concerned with the description of abstract concepts, individual entities, and their relationships within any real-world or imaginary setting. This idea can be transferred to the field of computer science, where ontologies are used to model such settings in a (somewhat) formal way. In the remainder of this article, we use the term ontology only in the latter sense. According to a widely quoted definition, ontologies constitute "a formal, explicit specification of a shared conceptualization" (Guarino et al.). This definition foregrounds several important properties: ontologies are meant to be interpreted by computers easily ("formal"), and the elements should be defined clearly and unambiguously ("explicit"). Another major aspect is their usability by many people to generate a common understanding of a domain ("shared").

Concrete ontologies are crafted with the help of designated modelling languages – so-called ontology languages – and are usually built on formal logic, thus rooting them in a strong mathematical foundation. This has multiple benefits, including the possibility of inferring implicit information from explicitly stated facts or checking the ontology for inconsistency with logic-based criteria.

Various technologies are involved in the technical (and formal) implementation of ontologies. The foundation for their representation is the Resource Description Framework (RDF). The RDF provides a data model for organizing information within so-called triples, which can be interpreted by machines (i.e., computers) easily. A triple has a simple three-part "subject–predicate–object" structure. Each element (i.e., the concepts, individuals and relationships) is represented by a Uniform Resource Identifier (URI), which also unambiguously identifies the element in its context. Subjects and predicates are always denoted as a URI, whereas the object can either again be a URI or an actual data value holding concrete information. For example, this data value could be the value of a property from a real-world entity, like the name or age attributed to a person. Figure 3 depicts the structure of the RDF data model.

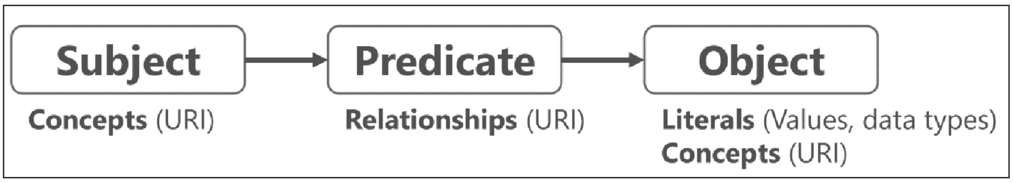

Figure 3 Structure of the RDF.

The RDF has an extension for schema definition, namely the RDF Schema, which already provides many useful predefined elements for constructing ontologies. For example, the "rdf:type" predicate is used to assign particular objects to a class, having the semantics of the "is a" relationship. The "rdfs:subclass" and "rdfs:superclass" predicates are used for defining hierarchical relationships between classes, therefore covering the taxonomy aspects of an ontology. To restrict possible source and target classes of self-defined predicates, the "rdfs:range" and "rdfs:domain" predicates can be used respectively.

The Web Ontology Language (abbreviated as OWL) uses the former two modelling languages as a foundation and offers all elements necessary for defining substantial ontologies. Conceptual elements in OWL are assigned with the type "owl:Class" and can have two different kinds of properties (i.e., specific predicates), which can be used to further describe these elements. The "owl:ObjectProperty" is used to define relationships between two classes, whereas the "owl:DatatypeProperty" is used to define the relationship between a class and some kind of actual value. Both types of relationships are directed, which means that they have a distinct source and a distinct target element connecting them with the above-mentioned domain and range elements of the RDFS. For the former type of property, there is the possibility to define an inverse relationship through a corresponding predicate (i.e., "owl:InverseOf"), which among other things can be used to infer implicit knowledge from existing facts. For example, if individual A is the parent of individual B, which is defined through a corresponding "isParentOf" relationship, and an "isChildOf" relationship is defined as an inverse of the former, it can be automatically inferred that individual B is the child of individual A without the need to explicitly state this fact.

For querying RDF-based data, which is usually stored and managed within a Triplestore, we use the descriptive querying language SPARQL (SPARQL1.1). WissKI also uses this language in its querying components to get the actual data from its underlying Triplestore. Figure 4 shows the relationship between the technologies described here.

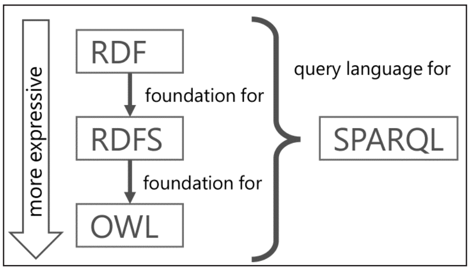

Figure 4 Technical foundations for ontologies.

4.2 USAGE IN WISSKI

In WissKI, ontologies define the basic building blocks for describing the schema for the data of the underlying Triplestore. The comprehensive description of the data with elements from an ontology also creates the foundation for the data to adhere to the FAIR principles.

For the ICA, we created an ontology that can be divided into two parts. The first relates to the description of the data with adequate metadata, whereas the second is designed to describe the subjects of the data items in a more comprehensive way. For the first part, we created a list of attributes that are potentially relevant for the description of our research items. We then connected them to corresponding concepts, their attributes and their possible relationships within the ontology. Through the pathbuilder introduced earlier, WissKI provides a mechanism to define input forms that map the metadata (e.g., a book title or the name of its author) of the data item and link them to corresponding elements of the ontology.

When creating a custom ontology, it is good practice to reuse (parts of) existing ontologies for common concepts. This ensures that objects are not defined multiple times and possibly in different ways despite sharing the same meaning, and it also contributes to reusability.

So far, we have only considered the function of ontologies as a foundation for the schema of our research data. In the following sub-section, we illustrate the second major objective we pursue through ontologies in this project, which is the comprehensive description of the content of the research data through a complex structure of keywords or tags.

4.3 ONTOLOGY OF ISLAMIC LEARNING

Our project is guided by a long-term vision to open our database to the global research community. The database has been developed by researchers for researchers. The tags we use correspond to research questions and include specialized terminologies, broader Islamic concepts, names, places and dates. They also comprise their relation to other technologies and concepts of a wider range of cultural, philological and social studies. As we intend to use tags to create links between data and to make data findable for other researchers, we agreed that the overarching framework of our tagging system needed to be precise and at the same time open enough to enable the extension of the tags. We therefore opted for a controlled vocabulary for tagging purposes, which is managed by an administrator who is a member of our research team. The administrator ensures that spelling remains consistent, alternative spellings are included, and the overall structure remains coherent but adapts to new research perspectives.

One of the main challenges we faced was the question of how to map the field of learning in a way that accommodates a wide variety of perspectives. We wanted to avoid falling back into a static taxonomy based on predefined classes and sub-classes and characterized by a

Modern Languages Open
DOI: 10.3828/mlo.v0i0.445

hierarchical order that integrates only one kind of relation. The diverse and multidisciplinary composition of our research team enabled us to include multiple perspectives on the issue of Islamic learning and helped to avoid strict hierarchical categorizations of the data and their content. We identified three angles for developing our controlled vocabulary: first, a physical angle that describes the content of the data with keywords along its natural attributes; second, a meta-perspective that relates the content to broader concepts pertaining to Islamic currents and schools of thought, cultural practices, learning technologies and media, among others; and third, information related to the context, such as place, time, global events, wars, ecological crisis, international policy programmes or natural disasters, etc.

In order to create common ground, we adopted a trajectory with straightforward questions, including "what", "when", "where", "who", etc. (see Figure 5). This enabled all team members to extend the map of tags, which are not related to one another in a strictly hierarchical structure, but rather in multidirectional ways that can capture the full range of possible relations through the semantic description of the data. A data item is therefore part of several strings of categories, and the categories relate in various ways to the data content.

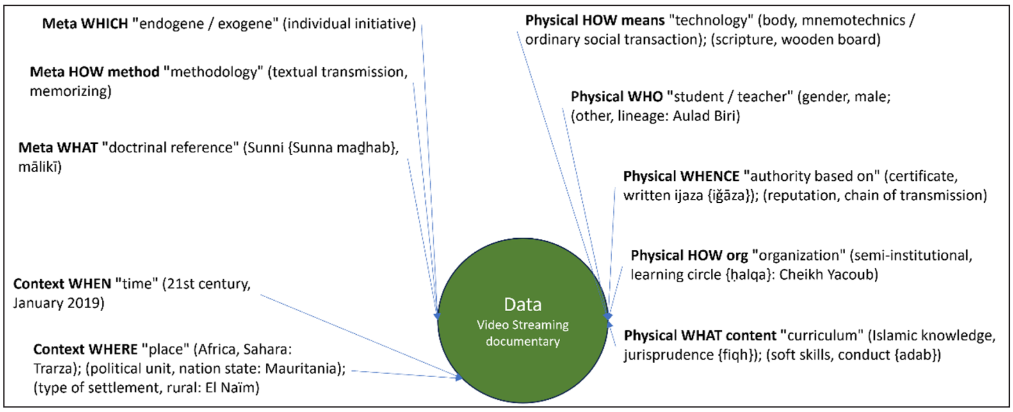

Figure 5 The ontology of Islamic learning: tagging with controlled vocabulary.

The following example will elucidate this further. The data item represented in Figure 5 is a video published on a web-based streaming platform and featuring a short documentary on traditional Islamic schooling in a Mauritanian village. In this educational institution, a specific type of pedagogy is employed. The teachers and their students come from various backgrounds – socially, doctrinally and geographically, as covered in the tagging system. The data can therefore be described with a set of tags that categorizes the content of the data in a systematic manner and integrates a broad range of relational information. Starting with the physical description, we create connections with the following tags: Learning **is organized as** *semi-institutional*, more specifically **as** a *learning circle* (**synonym to** *ḥalqa*) **led by** Cheikh Yacoub. The teacher and students **are identified as** *male*. The **technology** employed in the learning process **is**, on the one hand, a **technique** of the *body*, more specifically *mnemotechnics* and *social transaction*, and, on the other hand, *scripture*, namely **on** *wooden boards*. The curriculum thought **consists of** Islamic knowledge, especially jurisprudence {**associated with** *fiqh*} and further soft skills, such as conduct {**associated with** *adab*}.

If we were to proceed in the same manner with the "Meta" and the "Context" perspective, we would provide a comprehensive and in-depth systematic description of the data's content. Our example illustrates that the controlled vocabulary used here is capable of highlighting various types of relationships: categories, concepts, identities, technologies, associated terms and synonyms. The inclusion of synonyms is important and extends to the three working languages, covering varieties of spellings and transliterations. We also integrated Arabic terms in transliteration by adding associative terms to the tag in parentheses. At the same time, we translated all tags into our working languages (English, French and Arabic) to enable the searching of data in all three languages.

The selection of terms from this comprehensive tagging structure and their connection with research data items is facilitated by a tool that replicates the tree-like structure and allows users to select the desired tags in a convenient manner. The tool also displays relationships to other tags within the structure and therefore supports continuous content-related classification. New elements can be added by creating a new tag (potentially with multiple translations) and then selecting the relevant related tags together with their corresponding relationship, thus arranging them in the overall structure.

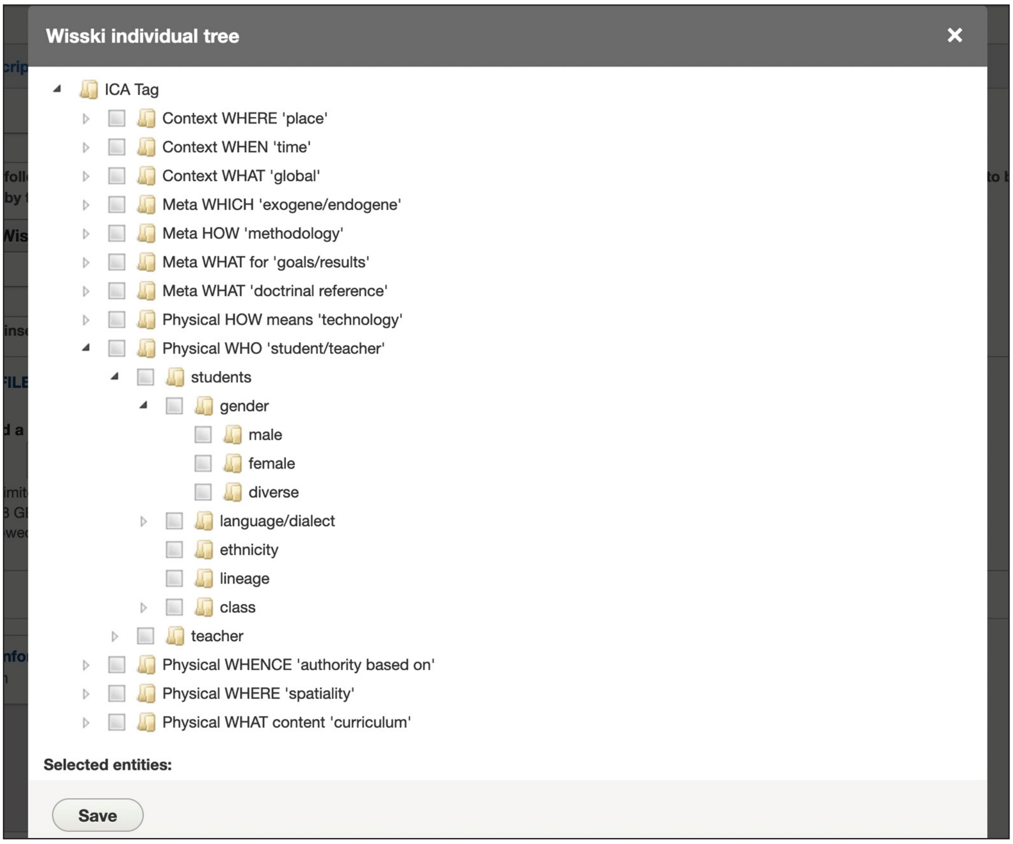

Figure 6 Screenshot of the "tagging tree" allowing for in-depth semantic description of the data.

The fluidity of the ontology is perhaps the most noteworthy feature of the ICA. As our database is designed for long-term use, the controlled vocabulary is set up in a way that can be extended by the users, who may update, add and delete tags in our "tagging tree" (see Figure 6). Thus, the ontologies can grow as research perspectives evolve and research results flow back into the database. In this manner, the mapping of our research field remains flexible and capable of integrating new insights or changes in the state of the art.

5 CROSSLINGUALISM

This section addresses the support of multiple languages, which is one of the main demands that needed to be fulfilled to build the ICA. It is crucial for our virtual research environment to enable collaborative work among team members with different language skills. From a technical perspective, several terms are used in this context, such as multilingualism, translingualism and crosslingualism. According to Mariani, language technologies are "monolingual when they handle a single language, multilingual when the same technology processes several (individual) languages, or crosslingual when they allow for switching and transferring from one language to another" (Mariani 146).

Systems that support multilingualism present content in different languages, but clearly distinguish between the individual languages. For example, content that exists in a system only in Arabic is usually not displayed to users who are using the system in English, as they do not correspond to the requested interface language. This leads to a difference in the amount of data presented to the user. In contrast, trans- and crosslingual systems support constant movement between different languages. Languages are not separated by distinct boundaries, and individuals can cross freely between them instead of switching between discourses ("Translanguaging TCU").

Technically, this means that the content of the system is presented in the language in which it was created (i.e., the original language) if no translation in the selected interface language exists. This behaviour is essential for a collaborative working platform to allow for the connection of research beyond language barriers. In the context of our project, we contributed to the development of WissKI by implementing the support of translations and crosslingual system behaviour. In the following, we give an outline of the conceptual decisions connected to that extension and describe the translation engine of our system.

DOI: 10.3828/mlo.v0i0.445

In general, when it comes to translations in digital environments, a distinction must be made between the translation of the system's user interface and the translation of the content stored in the system, that is, the research data. In our case, the translation of the user interface is semi-automatically provided by Drupal, while WissKI supports the translation of entities. In the data back-end of WissKI, this is managed by functionalities that consider the language tags of literals in RDF graphs. Each literal can be attached with a specific tag to specify the language used for the concrete data value. Figure 7 shows an example of the use of such language tags. The concept which describes a taxonomy term for Africa has two related (and similar) properties. One is annotated with an "@en" tag, expressing that the value has to be interpreted in English, and the other with an "@fr" tag, assigning the French interpretation.

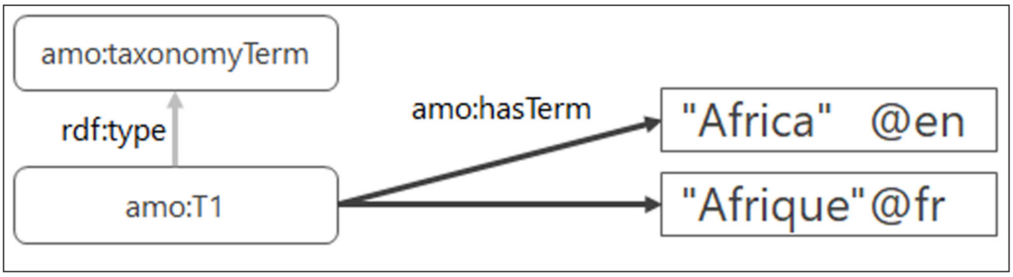

Figure 7 Literals with language tags.

From a user perspective, our implementation mostly affected the functionalities of creating, viewing and searching data. We focus on describing the particularities of these functionalities in terms of the translation-related extensions.

Once a new entry is created in the system, it can be translated. Hence, system administrators need to define for each bundle and corresponding fields whether they should be translatable. If the whole bundle is not translatable, users have no possibility of translating the content related to this bundle, and the corresponding entry does not receive a language tag in the Triplestore. As result, the content will always be displayed in that version, independently of the language that is selected in the interface. If the bundle is translatable, users are able to translate the content of translatable fields within that bundle on their own. This can be done by adding a translation in a language that has been installed in the system. In our case, users can add translations in English, French and Arabic. A precondition is that users have permission to translate content, which is configurable in the extensive role management provided by Drupal. If a role has permission to translate a bundle, users assigned to that role are displayed an additional menu item with that function.

In our system, a view is a listing of entries that belong to the same bundle, for example a list of books. Such views can be fully configured. For example, the fields to be displayed per entry can be set. Furthermore, it is possible to define criteria that can be used to filter the content of a view. One filter criterion considers the display language. Administrators can determine the behaviour of the view by defining a permanent display language as a filter criterion or by making it dependent on the selected interface language. In the first case, the view is always loaded in the permanently configured language, even if another interface language is selected. In the second case, the content of the view is loaded dynamically depending on the selected interface language. Changing the interface language initiates the view to be reloaded in the selected language.

In both cases, only a subset of entries is listed, that is, only entries that have a translation according to the requested language (which is either configured in the view or selected as interface language). To enable crosslingual views, we modified the second case. For this purpose, all entries of the requested bundle which have a translation according to the selected interface language are loaded in this language. All other entries of this bundle are loaded in their original language. In detail, the loading process parses every field of an entry and decides to render the content of the field in its original language (if the field is configured as not translatable) or in the requested language (if the field is configured as translatable and a translation of the field exists).

The number of entries always remains the same, independent of the selected interface language. It should be noted that these changes may cause the views to display different entries in different languages at the same time. This occurs when entries have different original languages but no translation in the requested interface language. According to our project

Modern Languages Open
DOI: 10.3828/mlo.v0i0.445

requirements, this is the desired behaviour, as in some cases the term should be intentionally displayed in its original language. For example, the Arabic term "{سُني {مذهب السُنة" (Sunni {*maḏhab as-Sunna*}) as it exists in its original language is used as a technical term in other languages as well. In such a case, it is only the spelling that is adapted, as the term is integrated in transliteration, but the term itself, including its meaning, is adopted as it cannot be translated appropriately. If, as in the first case, a permanent language is configured for the view, we decided to keep the previous behaviour, showing only entries that have a translation according to the configured language.

The search function is one of the most important functionalities in the ICA. It enables researchers to find relevant content and exploits the potential of the taxonomy. We realized the search by using SOLR, which is a leading open-source search platform provided by the Apache Software Foundation (SOLR). It can easily be integrated into Drupal-based systems by using the module Search API and Search API SOLR.

With the standard configuration of the module, the search returns only content that corresponds to the selected interface language. The following example illustrates this behaviour. Consider a data object that has only one field corresponding to its label. We assume that this data object has the label "Africa" and is stored in the system. Furthermore, a French translation of this data object is created when the label is translated into "Afrique". When the user, having selected English as interface language, searches for "Afrique", the search does not return any results. The reason for this is that only entries in the system with English as the assigned language are searched, that is, those having the language annotation "@en" in the Triplestore. In contrast, the example search term is stored as "Afrique@fr" in the Triplestore and therefore is not considered. This configuration does not meet our project requirements since researchers with different language backgrounds use the search function. A deviation of the search term language from the interface language is a frequent case that is not covered by this configuration.

The Search API configuration includes an option to enable multilingual search. However, this option does not produce the desired outcome either. Enabling this option will include content without an assigned language in the results of custom search queries. Content without an assigned language corresponds to entries in the Triplestore without a language tag. Translations, in contrast, do have language tags, which is why this configuration is not sufficient. Since crosslingualism was not yet supported by default, we had to adapt the functionality of the module. Hence, we modified the implementation of the function that collects the search results based on changed parameters. Instead of choosing only one language, that is, the selected interface language, we modified the function to consider all languages that are installed in the system. Now, the final set of results builds upon all subsets per language. Considering the previous example once again, the user with an English interface who searched for the French term "Afrique" now receives a matching entry. We furthermore configured a view for the search as described in the previous section. Therefore, the matching entry is displayed with its English label "Africa", since a version of the data object now exists in the selected interface language. Switching to a French interface reloads the view, showing the data object with the label "Afrique".

6 CONCLUSION

This article has outlined our experience in developing a crosslingual and ontology-based virtual research environment database utilizing the WissKI system and tailored to the needs of the Islamic Cultural Archive project in the African Multiple Cluster of Excellence. Through the detailed description of the steps we took in selecting our platform, establishing our tagging system, creating our ontology of Islamic learning and building the tool for crosslingual collaboration, we hope that this article has succeeded in laying out a digital solution that is of relevance for other users in the arts and humanities. By early 2024 the team members will have completed the process of uploading and tagging the research data collected in our eight sub-projects of the Islamic Cultural Archive. Adapted to our specific needs, WissKI allows us to integrate multiple research perspectives and to jointly develop a semantic map that describes the research field of Islamic learning. We anticipate that the platform will be expanded further, possibly through the inclusion of data from other collaborative research projects within the Africa Multiple Cluster of Excellence, thus extending our anglophone, francophone and arabophone research environment to other languages.

The WissKI platform turned out to provide a solid technical foundation for our endeavour. The extension of the Drupal content management system with Triplestore data storage enabled us to describe our data using ontologies in the OWL format, which includes the option of language annotations for data values, thus creating the basis for crosslingualism in our system. The system is therefore especially well-suited for collaborative formats, where researchers and institutions work together across continents and languages. Such a collaborative working tool offers great potential for creating new insights into any research field in the arts and humanities, and it may even enrich theoretical debates once large datasets have been made productive through fluid ontologies.

SHORT BIOS

Philipp Eisenhuth studied Computer Science at the University of Bayreuth. He is currently working as a research fellow in the field of heterogeneous data management. Since early 2019, he has been responsible for building and managing the infrastructure of a digital research environment for the Africa Multiple Cluster of Excellence at the University of Bayreuth.

Myriel Fichtner studied Computer Science (M.Sc.) at the University of Bayreuth and joined the Africa Multiple Cluster of Excellence in summer 2019, where she worked as a research fellow until 2023. In the Cluster, she is in charge as Digital Solution Developer and is responsible for the development of the ontology-based digital research environment. She completed her degree in 2023.

Britta Frede holds a PhD in Islamic Studies. Between 2019 and 2022 she was a postdoctoral researcher in the Africa Multiple Cluster of Excellence at the University of Bayreuth, working on Islamic education in contemporary African cities, with particular focus on the multi-layered transformative processes of Islamic learning and their embeddedness in gender and ethnicity-specific environments. Since April 2022 she has substituted the chair of Islamic Studies at the University of Bayreuth.

Rüdiger Seesemann is professor of Islamic Studies and Dean of the Africa Multiple Cluster of Excellence at the University of Bayreuth. He specializes in the study of Islam in Africa with a thematic focus on Sufism, Islam and politics, and theories and practices of Islamic knowledge.

AUTHOR AFFILIATIONS

Philipp Eisenhuth
University of Bayreuth, DE

Myriel Fichtner
University of Bayreuth, DE

Britta Frede
University of Bayreuth, DE

Rüdiger Seesemann orcid.org/0000-0003-0647-3899
University of Bayreuth, DE

REFERENCES

Candela, Leonardo, Donatella Castelli and Pasquale Pagano. "Virtual Research Environments: An Overview and a Research Agenda." *Data Science Journal*, vol. 12, 2013, pp. GRDI75-GRDI81. DOI: https://doi.org/10.2481/dsj.GRDI-013

Carusi A., and T. Reimer. "Virtual Research Environment Collaborative Landscape Study: A JISC Funded Project." *JISC Report*, 2010. DOI: https://doi.org/10.25561/18568

Cassidy, Steve, Dominique Estival, Timothy Jones, Denis Burnham and Jared Burghold. "The Alveo Virtual Laboratory: A Web Based Repository API." *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association, 2014, pp. 1–7. http://www.lrec-conf.org/proceedings/lrec2014/pdf/628_Paper.pdf.

"Dariah-EU." dariah-eu, 2022. https://www.dariah.eu/.

"Drupal." drupal.org, 2022. https://www.drupal.org/.

"FAIR." Go-fair.org, 2022. https://www.go-fair.org/fair-principles.

Fichtner, M., and V. Ribaud. "Paths and Shortcuts in an Event-Oriented Ontology." *Communications in Computer and Information Science*, vol. 343, special issue: Metadata and Semantics Research, 2012. DOI: https://doi.org/10.1007/978-3-642-35233-1_22

Frosini, L., et al. "An Aggregation Framework for Digital Humanities Infrastructures: The Parthenos Experience." *SCIRES-IT-SCIentific RESearch and Information Technology*, vol. 8, no. 1, 2018, pp. 33–50. DOI: https://doi.org/10.2423/i22394303v8n1p33

Garshol, Lars Marius. "Metadata? Thesauri? Taxonomies? Topic Maps! Making Sense of it All." *Journal of Information Science*, vol. 30, no. 4, 2004, pp. 378–91. DOI: https://doi.org/10.1177/0165551504045856

Guarino, N., D. Oberle and S. Staab. "What is an Ontology?" *Handbook on Ontologies. International Handbooks on Information Systems*, edited by S. Staab and R. Studer. Springer, 2009, pp. 127–33. DOI: https://doi.org/10.1007/978-3-540-92673-3_0

Hohmann, G., and B. Schiemann. "An Ontology-Based Communication System for Cultural Heritage: Approach and Progress of the WissKI Project." *Scientific Computing and Cultural Heritage*, edited by H. Bock, W. Jäger and M. Winckler. Springer, 2013, pp. 127–35. DOI: https://doi.org/10.1007/978-3-642-28021-4_14

Huisl, Jonas, M. Fichtner, Ph. Eisenhuth et al. "Towards an Integrated and Contextualized Research Data Management at the Cluster of Excellence Africa Multiple." *Frontiers in African Digital Research: Conference Proceedings*, edited by Anja R. Dreiser and Cyrus Samimi. Institute of African Studies, University of Bayreuth, 2022, pp. 155–72. DOI: https://doi.org/10.15495/EPub_UBT_00005720

"HuNI." huni.net.au, 2022. https://huni.net.au/#/search.

"IGSD: Interessengemeinschaft für semantische Datenverarbeitung." igsd-ev.de, 2022. https://igsd-ev.de/.

Keraminiyage, K., D. Amaratunga and R. Haigh. "Achieving Success in Collaborative Research: The Role of Virtual Research Environments." *Journal of Information Technology in Construction (ITCon)*, vol. 14, special issue: Technology Strategies for Collaborative Working, 2009, pp. 59–69. https://www.itcon.org/2009/7.

Madore, Frédérick. "La Collection Islam Burkina Faso. Promesses et défis des humanités numériques." *Revue d'Histoire Contemporaine de l'Afrique*, 2021, pp. 1–15. DOI: https://doi.org/10.51185/journals/rhca.2021.e610

Mariani, Joseph. "How Language Technologies Support Multilingualism." *NET.LANG. Toward the Multilingual Cyberspace*, edited by Laurent Vannini and Hervé le Crosnier. CF Editions, 2012, pp. 141–59. https://unesdoc.unesco.org/ark:/48223/pf0000216692.

"OWL2 Web Ontology Language Primer (Second Edition)." w3.org, 2012. https://www.w3.org/TR/owl2-primer/.

Pfeiffer, Michelle, Dirk Eller and Cyril Carré. "Toward a Virtual Research Environment for the Spatial Humanities: heiMAP." *Digital Scholarship in the Humanities*, vol. 35, no. 2, 2020, pp. 429–40. DOI: https://doi.org/10.1093/llc/fqz032

"RDF 1.1 Primer: W3C Working Group Note 24 June 2014." w3.org, 2014. https://www.w3.org/TR/rdf11-primer/.

"RDF Schema 1.1: W3C Recommendation 25 February 2014." w3.org, 2014. https://www.w3.org/TR/rdf11-schema/.

Russell, J. E., and M. K. Hensley. "Beyond Buttonology: Digital Humanities, Digital Pedagogy, and the ACRL Framework." *College and Research Libraries News*, vol. 78, no. 11, 2017, pp. 588–600. DOI: https://doi.org/10.5860/crln.78.11.588

Scholz, M., and G. Görz. "WissKI: A Virtual Research Environment for Cultural Heritage." *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI 2012)*, edited by Luc De Raedt et al. IOS Press, 2012, pp. 1017–18. doi: 10.3233/978-1-61499-098-7-1017.

Smith, B. "Ontology." *The Furniture of the World: Essays in Ontology and Metaphysics*, edited by Guilermo Hurtado and Oscar Nudler. Brill, 2012, pp. 47–68. DOI: https://doi.org/10.1163/9789401207799_005

"Solr." sol.apache.org, 2022. https://solr.apache.org/.

"SPARQL1.1 Overview: W3C Recommendation 21 March 2013." w3.org, 2013. https://www.w3.org/TR/sparql11-overview/.

Srinivasan, Ramesh. "Re-Thinking the Cultural Codes of New Media: The Question Concerning Ontology." *New Media and Society*, vol. 15, no. 2, 2013, pp. 203–23. DOI: https://doi.org/10.1177/1461444812450686

"Translanguaging TCU. Exploring a Translingual Approach to Language Difference in Composition." whitneylewjames.com, 2022. http://www.whitneylewjames.com/translanguaging-tcu/translingual-scholarship/.

"What is an RDF Triplestore?" ontotext.com, 2022. https://www.ontotext.com/knowledgehub/fundamentals/what-is-rdf-triplestore/.

Wilkinson, M., M. Dumontier, I. Aalbersberg et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Sci Data*, vol. 3, no. 1, 2016, p. 160018. DOI: https://doi.org/10.1038/sdata.2016.18

"WissKI." wiss-ki.eu, 2022. https://wiss-ki.eu/.

Wohlgenannt, G. *Learning Ontology Relations by Combining Corpus-Based Techniques and Reasoning on Data from Semantic Web Sources*. Peter Lang, 2018. DOI: https://doi.org/10.3726/b13903

TO CITE THIS ARTICLE:
Eisenhuth, Philipp, Myriel Fichtner, Britta Frede and Rüdiger Seesemann 2023 Developing Crosslingual Ontologies in WissKI: Transcontinental Research Collaboration in the Africa Multiple Cluster of Excellence. *Modern Languages Open*, 2023(1): 39 pp. 1–16. DOI: https://doi.org/10.3828/mlo.v0i0.445

Published: 18 December 2023

Modern Languages Open is a peer-reviewed open access journal published by Liverpool University Press.