Wikibase solutions for african literary metadata

Aangenent, G.; Harris, A.; Wu, D.; Maen, A.; Oberst, U.

Journal of Open Humanities Data

Wikibase Solutions for African Literary Metadata

DISCUSSION PAPER

ubiquity press

GIJS AANGENENDT
ASHLEIGH HARRIS
DANN WU
ADAM MAEN
URSULA OBERST

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

ABSTRACT

The African Literary Metadata (ALMEDA) project is building a Wikibase instance to create a multilingual, linked data repository of African literary materials that often go uncatalogued in libraries and archives. While Wikidata enables a flexible solution for linking data on diverse materials, ALMEDA opted for creating a separate Wikibase instance because the project requires a robust ontology to repeal the effects of colonial library standards on literary description and to enable multilingual functionality. This paper elaborates the reasons for this choice, the challenges the project has faced in implementing a Wikibase instance, and the solutions it has employed to address those problems.

CORRESPONDING AUTHOR:
Ashleigh Harris
Department of English,
Uppsala University, Uppsala,
Sweden
ashleigh.harris@engelska.uu.se

KEYWORDS:
decolonizing metadata; African cultural heritage; literary metadata; African literature; Wikidata; Wikibase

TO CITE THIS ARTICLE:
Aangenendt, G., Harris, A., Wu, D., Maen, A., & Oberst, U. (2025). Wikibase Solutions for African Literary Metadata. *Journal of Open Humanities Data*, 11: 79, pp. 1–10. DOI: https://doi.org/10.5334/johd.433

(1) INTRODUCING THE AFRICAN LITERARY METADATA (ALMEDA) PROJECT

The ALMEDA project[1] transcribes and links metadata on a range of African expressive materials that do not normally get catalogued in libraries or archives. In order to increase the visibility of the rich literary and oratory cultures from the African continent, ALMEDA is developing a linked open metadata repository using the open-source software, Wikibase.[2] This repository links metadata based on a fluid multilingual ontology developed within the project to provide accurate and decolonized descriptions of informal African literatures. As such, ALMEDA makes a decolonising intervention in two related ways: first, the materials that we catalogue are previously omitted from other data repositories and ALMEDA is the first repository committed to describing and making these materials visible and searchable. Secondly, ALMEDA is developing a unique decolonial model for linking literary and expressive forms in ways that enable multilingual searchability and that do not prioritise colonial-language understandings of literary form.

Given that the project is currently at the start of its third year of operation, we are still in the process of resolving problems and addressing challenges. We have, however, started to consolidate major aspects of the project, such as the implementation of the Wikibase instance, the drafting of the first version of our ontology, and the creation of a workflow for uploading data. This paper discusses the reasons for using Wikibase, the challenges the project has faced during the implementation of the software, and some of the methods for resolving them.

(1.1) INVISIBILITY OF AFRICAN LITERARY AND ORAL MATERIALS

The reasons for the exclusion of certain kinds of African literature from library and archive catalogues are complex and multiple. Primarily, this is a consequence of the ways cataloguing systems, which owe their designs largely to 19th century American and European library science, were biased towards the book object as the unit from which metadata was generated (Holden, 2020, p. 84). Given the fact that until the early 19th century most sub-Saharan African cultural expression was orally transmitted, the consequence of this bias has been extensive. Most obviously, the oral expressive forms that dominated the continent before the onset of book culture have been relegated to the fields of sociology, anthropology or ethnography, rather than to literary studies. Secondly, the fact that cataloguing standards were developed to describe collections of books has created a problem for African expressive culture, much of which has been circulated orally or in a variety of print forms that do not easily conform to the book-driven conception of the 'work' – one which is determined by 'Author-Title' metadata employed by most libraries and cataloguing standards.

The consequence of this has been detrimental to the archiving and visibility of African literatures, both written and oral. How, for example, does one catalogue the metadata of oral literatures that are not authored by any one person or by an identifiable collective? The colonial archive dealt with this by ascribing authorship to the transcribers of oral works, or to the ethnographers that later recorded or filmed their performance – a practice of attribution that has contributed to the erasure of African authorship in the formal archive and to a clear practice of European folklorists appropriating African materials. Moreover, lack of linguistic expertise has resulted in African cultural forms being erased as they are catalogued under English terms and ontological structures. For example, the Zimbabwean language, *Shona*, has a strong tradition of oral praise poetry and this tradition has a clear structure of praises, including clan praises, personal praises and boasts. Under personal praises is a form known as *nhetemo dzokunyaradza mwana*, which translates directly as 'phrases of comfort for a baby', but which is sung, and thus can be translated into English as 'lullaby'. A lullaby might be categorised as a song or a children's song, but this would miss the fact that these soothing songs to babies are part of the praise tradition in *Shona* (Hodza and Fortune 1979).

[1] For more information about ALMEDA, visit the project website: https://almedaresearch.org.

[2] Wikibase provides the technical infrastructure to create and manage a linked open knowledge base. Its most well-known implementation is Wikidata, a repository that collects the structured data of Wikimedia projects such as Wikipedia (Vrandečić & Krötzsch, 2014). Another example is the digital humanities project Enslaved: Peoples of the Historical Slave Trade (Shimizu et al., 2023). For an introduction on how to get started with Wikibase, see Varvantakis (2025).

A further factor contributing to the invisibility of metadata about African print literature in the formal archive and catalogue has to do with the relationship between mode of publication or distribution of a work and the metadata it accrues. From the late 19th century through to the late 20th century, the dominant form through which African literature was published and circulated was in newspapers, magazines and pamphlets. Libraries seldom catalogue the content of such ephemeral print materials, thereby obscuring the literary content that is embedded in these forms. An illustrative example can be found in the publication of 300 serialised Swahili novels and novellas in two Kenyan-Tanzanian newspapers between 1960 and 1980 (Taifa Weekly and Baraza Weekly).[3] Of these (exactly) 300 serialisations only 57 appear in the World Catalogue (WorldCat), which is the largest platform for library metadata from thousands of libraries across the globe.

(1.2) A WIKIBASE FOR AFRICAN LITERARY METADATA

Because of the diversity of the materials that ALMEDA is cataloguing – from print ephemera (such as market pamphlets, newspapers, and magazines) to performance-based works (theatrical works, street performance), to online materials (social media written or video works), and potentially in any number of African languages – the project needed to work with a flexible linked open data solution. The data model had to allow for the creation of metadata descriptions of various literary forms generally not captured in bibliographic databases and must have the capacity for multilingual functionality. Since Wikibase met these requirements, thereby offering the possibility to build a flexible and durable knowledge base whilst also allowing for easy data management for researchers with no coding skills, the decision was made to use this software for the ALMEDA repository.

The project made the further choice to use a private Wikibase instance for its repository rather than to create, more simply, a project within Wikidata itself (Rossenova et al., 2022). While Wikidata's extensive reach and pre-existing modelling would have allowed the project to begin entering collected metadata on literary materials easily and quickly, there were several reasons as to why the decision was made to operate ALMEDA as a separate Wikibase project.

First and foremost, ALMEDA is an academic project that must be held accountable for the accuracy and veracity of its data: by allowing data to be editable by any Wikidata community member, such veracity could not be guaranteed. Wikidata addresses this problem with their notability criteria by requiring entities to be verifiable with legitimate and publicly available references, such as university-level textbooks, reference books, academic journals, or newspapers.[4] However, the informal nature of the materials that ALMEDA is collecting and the fact that these materials do not as yet exist in verifiable databases make it difficult to provide such proof. This highlights a problem that emerges when creating structured and verifiable data on highly informal and as yet uncatalogued material: when trying to correct absences or data inequality in global databases, the burden of proof remains on existing formalised data, in the form of external identifiers or references, to prove the veracity of an entity.[5]

A further problem with Wikidata is that the knowledge schema behind many of its classes for literature and culture has inherited the structural and classificatory problems of the very standards that the ALMEDA project wishes to interrogate. One example of this in Wikidata is the genre 'oral literature': though a sub-class of 'literature', Wikidata nevertheless connects this to the Library of Congress external identifier 'Folk Literature', which is a classification that the ALMEDA project rejects.[6] By using a private Wikibase instance, instead, ALMEDA has been able

[3] If it had not been for the indexing work of Richard Marshall Lepine, who published an index of these literary works as part of his doctoral thesis (Lepine, 1988), the ALMEDA project might not have been alerted to the existence of these works.

[4] https://www.wikidata.org/wiki/Wikidata:Notability (last accessed: 2025/10/22).

[5] To address this matter, ALMEDA is working with researchers who formally publish datasets of materials they have collected and verified (see https://zenodo.org/communities/almeda) which are then used as references for the data, thereby enabling a clear line of provenance for data. These datasets are published under Creative Commons Attribution 4.0 International licences, for ease of downloading and reuse even before the repository's search interface is ready for use.

[6] For an account of the colonial implications of the Library of Congress terms for Folk Literature, see Harris (2025).

to determine the configuration of the knowledge base itself and to build a data model based on the inclusive ontology for African expressive forms developed in the project.

(1.2) WIKIBASE FOR FAIR AND CARE DATA

Despite having made the choice to work with what might be perceived as a siloed, private Wikibase, as opposed to the community-driven Wikidata, the ALMEDA project is committed to investing in its data being as interoperable as possible. By aligning the structure of the ALMEDA Wikibase with Wikidata when possible, we aim to enable federated searches in the future across the repositories and to share our data, where possible, with Wikidata.[7] This is part of the project's commitment to implement FAIR (Findable, Accessible, Interoperable and Reusable) principles of data management.

The CARE principles[8] (Collective Benefit, Authority to Control, Responsibility, Ethics) of data management are also core to the ALMEDA project, given that a significant amount of data relevant to this project relates to cultural heritage materials. To ensure that data is for the 'Collective Benefit' of the communities it represents or emerges from, ALMEDA works with researchers from within relevant communities, guaranteeing that the kinds of data captured on the materials are curated with attention to the inherent cultural knowledge models and linguistic logics of the material.[9]

Furthermore, Wikibase enables the perpetual editability of all data collected in ALMEDA: this allows for continued conversations between researchers, communities and other stakeholders, which in turns means that by seeking feedback and collaboration with communities in the structuring and presentation of data, ALMEDA will enable communities to have the 'Authority to Control' the data collected. Through the structure of an appointed and named advisory editorial board, ALMEDA will review all data with the expertise necessary to take 'Responsibility' in transparent and 'Ethical' ways for how that data is presented and for what purposes. The largest contribution that ALMEDA is making towards the CARE principles is to ensure that African language forms are included in the ontology in the languages that these forms are created in.

(2) METHOD

This section discusses the three main technical parts of the ALMEDA project, those being data modelling, implementation of the Wikibase, and the workflow for uploading data.

(2.1) MODELLING

ALMEDA is developing an ontology that models African literary expressions in ways that both resist the impact of colonial library science on the knowledge schema for these works and that allows African language terms their own space in the knowledge model without subsuming them under an English (or other colonial language) term. At the same time, this ontology must operate multilingually, to enable accessibility. To achieve this, the project team has worked extensively with researchers on creating a structure for a wide range of 'Forms of the Creative Work' relevant to the project. We are also working with African language experts who are compiling translated lexicons of these forms of literary expression and adding new forms from their language-context. It is important to note that this is a fluid ontology, which Srinivasan and Huang define as "flexible knowledge structures that evolve and adapt to communities' interest based on contextual information articulated by human contributors, curators, and viewers…" (2005, p. 193). By virtue of our approach, the inclusion of as-yet-uncatalogued forms in the ALMEDA database requires us to keep our ontology open to review as new forms may challenge the structure we are working with.

[7] Bidirectional federation requires whitelisting ALMEDA's SPARQL endpoint and the creation of a property for the ALMEDA identifier in Wikidata. These prerequisites still have to be realized for our instance. For more information on bidirectional queries, see the documentation of the MiMoText project: https://github.com/MiMoText/ontology/blob/main/module13_federation/federation_mimotextbase_wikidata.md (last accessed: 2025/11/17).

[8] https://www.gida-global.org/care (last accessed: 2025/11/18).

[9] We also plan for an external review of our data model by an ethical review panel focussing on African oral and literary culture in the fourth year of the project.

The basis for ALMEDA's ontology is the Library Reference Model, object-oriented version (LRMoo), "intended to capture and represent the underlying semantics of bibliographic information and to facilitate the integration, mediation, and interchange of bibliographic and museum information."[10] ALMEDA's domain of African literary and cultural materials in print, audio, video or digital formats is of relevance to both the library and museum worlds, and the decision to work in alignment with the LRMoo formal ontology ensures future interoperability with these sectors. However, where LRMoo allows for the theoretical modelling of literary works on four levels (Work, Expression, Manifestation, and Item), many of the materials ALMEDA is modelling are only available as single expressions, manifestations, and even items. As such, and in the interests of best use of resources, ALMEDA has adopted Wikidata's two-level model for 'works': that being, 'work' and 'version, edition, translation'.[11] This simplifies the modelling and data upload processes without losing the quality of the data.

Besides drawing on formal ontologies such as LRMoo and pre-existing Wikidata models such as WikiProject Books, ALMEDA is also introducing project-specific elements to the ontology. The most significant of these being the aforementioned types we include under 'Form of the Creative Work', which, while fluid and open for further inclusion of new forms, is a domain-specific list that is unique to the ALMEDA project. These types and their language-independent object properties have an instrumental role in the ontology, allowing discoverability of literary expressions across language and cultural contexts.[12]

While ALMEDA's model for 'Forms of the Creative Work' constitutes a new way of structuring data in the domain of African literature, we nevertheless have designed it to maximize interoperability with existing databases and sectors. This is important for the sustainability and usability of ALMEDA's data, and also because a significant amount of our data, while on informal and ephemeral materials, is likely to intersect with more formal literary data. Some published authors, for example, published works in small magazines, while others have written plays, performed in street environments, based on their published novels. The ALMEDA project highlights the fact that the intersections between formal and informal data are themselves porous and messy. We seek to link our data to existing formal data – such as that found in library catalogues – rather than to create a database disconnected from the formal literary field.

(2.2) WIKIBASE IMPLEMENTATION

ALMEDA uses the open-source software Wikibase Suite Deploy to self-host its linked open knowledge base.[13] The software is divided into smaller units in the form of Docker containers, self-contained packages that include the code, settings, and libraries needed to run a piece of software. Each container fulfils a specific function within the complex structure of Wikibase, such as functions related to the interface, data repository, data entry, search functionality, query services, and web traffic. Compared to Wikibase Cloud,[14] a ready-to-use installation of Wikibase hosted by Wikimedia Germany, Wikibase Suite Deploy requires more technical expertise and computational resources to implement but offers greater control over the instance and customization of its functionalities.

One example of this customization is the possibility to install MediaWiki extensions.[15] These extensions, developed by members of the Wikibase community, provide additional functionalities and even new data types. For example, the WikibaseQualityConstraint extension can be used to define constraints on the properties and check whether they are met in statements, informing the editors of items where properties are incorrectly used.[16] Thanks to Docker, the main components of the Wikibase instance can be deployed with the predefined system configuration and environment, but for the installation of extensions a lot of testing work is required.

[10] https://cidoc-crm.org/lrmoo (last accessed: 2025/11/21).

[11] We have based our basic modelling of literary works on WikiProject Books' data model https://www.wikidata.org/wiki/Wikidata:WikiProjectBooks (last accessed: 2025/11/21).

[12] The ALMEDA ontology of Forms of Creative Work, and how this will operate multilingually, is the topic of a much larger work, which is planned for publication in the final year of the project.

[13] https://github.com/wmde/wikibase-release-pipeline/tree/main/deploy (last accessed: 2025/10/23).

[14] https://www.wikibase.cloud/ (last accessed: 2025/10/24).

[15] https://www.mediawiki.org/wiki/Category:Extensions (last accessed: 2025/10/20).

[16] https://www.mediawiki.org/wiki/Extension:WikibaseQualityConstraints (last accessed: 2025/10/20).

A newly created Wikibase instance is a blank canvas that requires extensive set-up and configuration to fit a project's needs. After the first launch of the instance, the environment variables need to be specified in order for the software to work correctly. These settings specify among other things the communication between the interface and the database, the web address where the Wikibase can be accessed, and the location of services such as the SPARQL query service for the retrieval of data. Other configuration steps include setting up user groups and rights as well as adding support for languages. The majority of the languages relevant to ALMEDA were already provided for in Wikibase. Additional languages can be added as long as the language has an IETF language tag.[17] For ALMEDA, this meant that the Bantu language *Lulogooli* could be added, but not informal urban vernacular languages that do not have such a tag, one example being *Sheng*, a mix of English, Swahili and other Kenyan languages spoken in Nairobi.

The next step involves translating the ontology into the Wikibase data model[18] and populating the instance with items and properties. In Wikibase, items are uniquely identified entities that, in the context of ALMEDA, could represent a creative form, an author, a country, or an event. Each item has a unique identifier, starting with the letter 'Q' followed by a number, as well as human readable labels, descriptions, and aliases in different languages. Properties define specific characteristics or relationships of an item. They also have a unique identifier (starting with the letter 'P' followed by a number), as well as multilingual labels, descriptions, and aliases. Knowledge about an item is expressed through statements, which follow a subject-predicate-object structure: the subject is the item, the predicate a property, and the object is a value, which could for instance be another item, a date, or string. To illustrate this, we can use an example from the ALMEDA Wikibase: the Namibian writer Ndawedwa Denga Hanghuwo who received an award called the Bank Windhoek Doek Literary Award For Fiction (Figure 1). To the item 'Ndawedwa Denga Hanghuwo', a statement is added using the property 'award received' and linking it to the item for the award.

Furthermore, the Wikibase model allows adding contextual information to statements through qualifiers and references. Qualifiers can be used to further specify a statement (e.g. using the qualifiers 'for work' and 'point in time' to express for which work the author received the award and the time of the rewarding as seen in the example presented in Figure 1). References refer to the source of the information. They can take different shapes and refer to academic articles, datasets, encyclopedias, as well as web pages. In the example given here, the webpage of the literary magazine is provided as a reference. Here the work by Ndawedwa Denga Hanghuwo is published alongside the author's biographical information.[19]

A major challenge when setting up a Wikibase instance, is the time-investment required in implementing the data model based on the project's ontology and in populating the instance with properties and data items, such as countries and languages. To facilitate this process, the project reused properties and basic metadata items from the public Wikidata, though this was not always straightforward. In Wikidata, for example, 'human settlement' items often include varied and sometimes haphazard definitions of 'megacities', 'cities', 'big cities' and 'towns'. Johannesburg,[20] for example, is an instance of both a 'city' and a 'big city', while Nairobi[21] is (in addition to being an 'administrative territorial entity of Kenya') a 'big city', but not a 'city' – both cities have similar central populations of around 6 million inhabitants.[22] The messiness of data in the hands of an ever-changing community of editors is a problem for research-based data repositories. As such, ALMEDA has decided to only use one term, 'human settlement', to determine all urban and rural locations and we are populating these items with data from GeoNames rather than Wikidata because of GeoNames' standardization and stability of data.[23]

[17] IETF language codes are based on the ISO-codes for languages, regions and scripts. For more information on adding languages, see https://www.wikidata.org/wiki/Help:Monolingual_text_languages (last accessed: 2025/11/20).

[18] For an introduction to the Wikibase data model, see https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer and https://www.mediawiki.org/wiki/Wikibase/DataModel (last accessed: 2025/11/24).

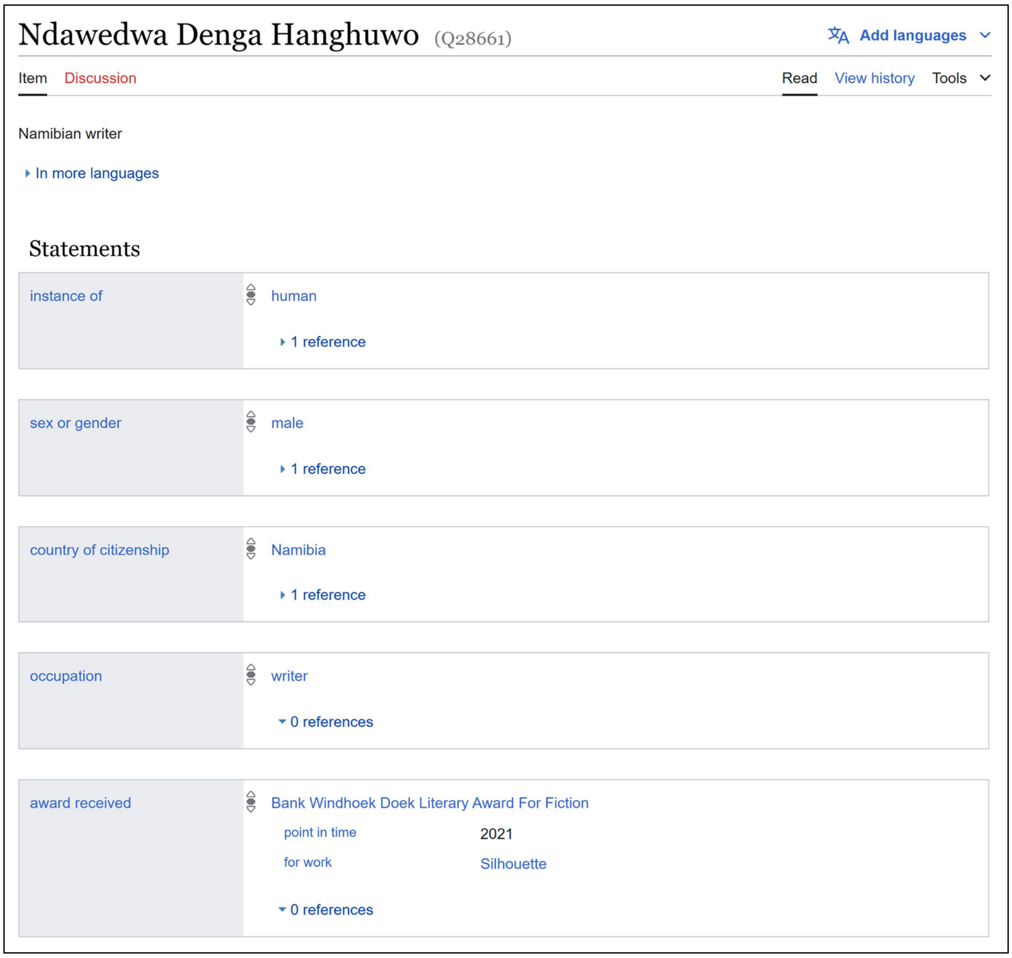
[19] https://doeklitmag.com/silhouette/ (last accessed: 2025/10/24).

[20] https://www.wikidata.org/wiki/Q34647 (last accessed: 2025/10/24).

[21] https://www.wikidata.org/wiki/Q3870 (last accessed: 2025/10/24).

[22] https://worldpopulationreview.com/cities/continent/africa (last accessed: 2025/10/24).

[23] https://www.geonames.org/ (last accessed 2025/10/24).



Ndawedwa Denga Hanghuwo (Q28661)

Item Discussion Read View history Tools

Namibian writer

In more languages

Statements

instance of	human 1 reference
sex or gender	male 1 reference
country of citizenship	Namibia 1 reference
occupation	writer 0 references
award received	Bank Windhoek Doek Literary Award For Fiction point in time 2021 for work Silhouette 0 references

Figure 1 Item of Ndawedwa Denga Hanghuwo in the ALMEDA Wikibase.

Other data was used to 'kick-start' the ALMEDA project, such as literary or related occupations, existing African author data, and data on existing festivals, events, awards and venues. Once the project members agreed on which data could be reused, two programming packages – WikibaseIntegrator[24] and Pywikibot[25] – were used to retrieve these properties and items and insert them into the ALMEDA instance. The imported data was then examined for their qualities, and subsequently simplified, cleaned and standardised for ALMEDA's purposes. Wikidata had much non-essential information for ALMEDA, which was filtered out during the reuse process. The basic information imported from Wikidata items are Labels, Descriptions and Aliases in seven languages. We also kept the links to Wikidata for provenance and reuse purposes. For countries reused, two properties 'coordinate location' and 'inception' are kept. For languages reused, we keep the property 'country' and three external identifiers, 'Ethnologue.com language code', 'ISO 639-3 code' and 'Glottolog code'.[26]

(2.3) WORKFLOW

With the implementation of the ALMEDA Wikibase nearing completion, the project is moving forward with consolidating the workflow for uploading collected metadata to the instance. Metadata is collected in several ways, through case studies by affiliated researchers, data harvesting from available physical and digitized materials at archives and libraries, and dataset donations from collaborating researchers and organizations. New metadata will be collected and added continuously throughout the project. After the project has formally ended, an advisory editorial board will be established to review metadata submissions before they are uploaded.

[24] https://github.com/LeMyst/WikibaseIntegrator (last accessed: 2025/10/27).

[25] https://github.com/wikimedia/pywikibot (last accessed: 2025/10/27).

[26] In the ALMEDA Wikibase, the Ethnologue language code and Glottolog code function as references to information about (African) languages, including where the language is spoken, the number of speakers, and the language family and group it belongs to. The ISO 639-3 code primarily serves a technical purpose for linking language records across datasets and systems.

A question that impinges on the project as a digital humanities initiative is the problem that domain experts from literary studies and related fields have little to no experience of working with datasets, since this is seldom required in the field. This meant that basic training was required to bring researcher expertise in alignment with how data needed to be structured. ALMEDA has addressed this problem by tasking our information specialist with preparing templates for data capture for researchers. These templates facilitate the data-uploading process and ensure that the necessary data fields are present for the data uploader to proceed.

When a new dataset comes in, the data is first checked against the appropriate template. This includes checking whether the necessary data fields are present and the values in the columns align with the formatting standards decided upon. Any inconsistencies in the dataset are corrected, such as the incorrect formatting of dates, page numbers, and author names. If any major problems are encountered that cannot be resolved by the data uploader, the dataset is sent back to the Principal Investigator for referral to the researcher or data collector. Once all the requirements are fulfilled, the cleaned dataset is published on Zenodo, generating the digital object identifier (DOI) used for reference in the items in the Wikibase instance. The data cleaning steps are performed using basic functionalities of spreadsheet editors and the data wrangling tool OpenRefine. This tool offers a range of useful methods for cleaning and transforming data, for example harmonizing spelling variations of author names through clustering.[27]

For the data upload, there are several options to add data to the Wikibase instance. Data can be added through manual entry of individual items, but also in bulk using the previously mentioned python packages. In the case of ALMEDA, OpenRefine has so far been the preferred tool, due to its capacity to clean, reconcile, and upload the data in one place. From the single dataset different items need to be created in the instance, related to the source of publication or venue, the author/creator, the creative work itself, and finally the 'version, edition, or translation' of the work. For each of these a list of required and optional properties is defined.

Reconciliation in OpenRefine is used to match the values in the dataset with already-existing records on the ALMEDA instance.[28] This is done to check whether the work or author already has a presence in our instance and link the values related to 'country of citizenship' or 'form of creative work' with the appropriate item on the Wikibase. Additionally, reconciliation of data about people takes place against Wikidata to see if relevant information regarding e.g., sex assigned at birth, date of birth, awards received is available. This information, if available, is then added to the record alongside the metadata collected by ALMEDA. The link to the Wikidata item is added to the record using the ALMEDA property 'Wikidata link'. Reconciliation is also a required step before data can be uploaded to the Wikibase instance. If the value in the cell cannot be matched to an existing item on the Wikibase instance, a new item is created alongside the other statements.

In order to reconcile and upload data, a Wikibase manifest is needed, a configuration file that specifies how OpenRefine should communicate with the instance. Additionally, the data uploader should be added to the user group of bots to be able to create and edit items in bulk. Furthermore, a Wikibase schema needs to be defined. This schema defines how the columns and values in the dataset are to be transformed into statements.[29] When uploading a dataset, OpenRefine provides the possibility to inspect future edits on the instance in preview. The tool also reports issues with the schema and values, reducing the chance of incorrect uploads.

An overall challenge for the project concerns aligning its different elements, from the data modelling and the implementation of the ontology in Wikibase, to the data upload workflow. The ontology is developed from an ideal theoretical scenario, where all the metadata related to entities is relatively clear and available. The actual datasets collected by researchers or donated by collaborating partners, however, do not mirror the ideal data model as defined in the ontological structure. The datasets include valuable but fragmented and often patchy data that need to be restructured to fit the data model. Because of the informal circumstances in which much of our data is collected, these datasets often miss basic metadata that is normally expected in library catalogues, such as date of birth or affiliation. As a result, the ontology and level of detail of the metadata need to be weighed against the workload for the data uploader.

[27] https://openrefine.org/docs/manual/cellediting#cluster-and-edit (last accessed: 2025/11/20).

[28] https://openrefine.org/docs/manual/reconciling (last accessed:2025/10/20).

[29] https://openrefine.org/docs/manual/wikibase/overview (last accessed: 2025/10/20).

At the same time, ALMEDA has made some choices that, whilst adding extra work for a data uploader, have been retained for the purposes of theoretical clarity and future interoperability of the data with other repositories. One example of this is our two-level modelling of a creative work at both the 'work' and 'versions, edition, and translation' level. Whilst this two-level structure is not strictly relevant for single issue works (which we anticipate will be the vast majority of ALMEDA's content), and while it would be easier to remove this distinction and upload one item for each work rather than two, this could affect the searchability and interoperability of the whole repository.

(3) DISCUSSION

The ALMEDA project still has a long way to go before it will be ready for a full review of the process of implementing our Wikibase instance: we have yet to upload most of our collected metadata and to develop the search interface with the repository. Nevertheless, as a large-scale, research-based, digital humanities project, we feel that our reflections on our experience with Wikibase thus far might be valuable for new projects intending to use Wikibase, so as to accelerate their progress from software choice to data upload.

Wikibase is flexible software and fits ALMEDA's needs when it comes to implementing our own data model. It also formed a useful 'sandbox environment' right from the outset, for testing and discussing different data modelling approaches. Working in an interdisciplinary team, our domain experts/researchers (literary and cultural researchers or linguists), information specialists, engineers, and data curators all approach the object of our shared work – data – with very different understandings. This can lead to miscommunication or misunderstanding, but we have found that the interface of the Wikibase instance and the presentation of items and statements helped to find common ground when discussing important ontological questions.

That said, the project faced some technical difficulties in early implementations and maintenance of the software. The lack of available central documentation on Wikibase installation created numerous technical challenges. When encountering certain bugs, without a clear description of the immediate solution, extra time is spent trying to resolve the issue in isolation. Furthermore, Wikibase and related tools, extensions, and services are continuously updated by the developers and the open-source community. As such, there is a chance that certain components are no longer compatible or require a different version. In the absence of documentation, we did find the community forum[30] very helpful to ask questions and get advice in resolving issues with the SPARQL service of our Wikibase.

Whilst one of the benefits of using Wikibase was the ease of reconciliation against Wikidata and the reuse of items and properties, the messy nature of Wikidata has been a recurring problem in the project. This also has implications for automated import of basic data such as languages, countries, human settlements and occupations. When automating imports from Wikidata into our Wikibase, the irregular nature of Wikidata led to the inclusion of a significant amount of irrelevant data that needed to be subsequently deleted. Considering this, it might be better to switch to more curated data for basic data, relying on gazetteers and other sources of data that are more reliable and stable, when implementing a Wikibase.

These challenges notwithstanding, we are of the opinion that more robust documentation and sharing of experiences in implementing a Wikibase would significantly ease the teething-problems we have experienced during the starting phase of ALMEDA. Ultimately, we are confident that Wikibase was the correct choice of software for this project, given the varied skillset of the team, the need for interoperability, and the need for flexible solutions to complex and messy data. When we have developed our search interface and domain experts in the African literary and cultural sphere start to query our data, we will have a better perspective on the efficacy of this choice.

FUNDING STATEMENT

ALMEDA (African Literary Metadata) has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme. Grant agreement No: 101097763.

[30] https://phabricator.wikimedia.org/ (last accessed: 2025/10/10).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Gijs Aangenendt: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Writing – Original Draft, Writing – review and editing

Ashleigh Harris: Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Supervision, Writing – Original Draft, Writing – review and editing

Dann Wu: Conceptualization, Methodology, Software, Writing – Original Draft

Adam Maen: Conceptualization, Methodology, Software

Ursula Oberst: Conceptualization, Data Curation, Methodology

AUTHOR AFFILIATIONS

Gijs Aangenendt orcid.org/0000-0002-1411-6595
Department of English, Uppsala University, Uppsala, Sweden

Ashleigh Harris orcid.org/0000-0002-6207-3067
Department of English, Uppsala University, Uppsala, Sweden

Dann Wu orcid.org/0009-0007-7720-4502
Centre for Digital Humanities and Social Sciences, Department of ALM, Uppsala University, Uppsala, Sweden

Adam Maen orcid.org/0009-0008-4173-7224
Centre for Digital Humanities and Social Sciences, Department of ALM, Uppsala University, Uppsala, Sweden

Ursula Oberst orcid.org/0000-0003-4168-6742
African Studies Centre Leiden, Leiden University, Leiden, The Netherlands

REFERENCES

Harris, A. (2025). African Literary Metadata and Makerere University's Library. *Research in African Literatures*, *55*(1), 1–27. https://doi.org/10.2979/ral.00042

Hodza, A. C., & **Fortune, G.** (1979). *Shona Praise Poetry*. Clarendon Press.

Holden, C. (2020). The Bibliographic Work: History, Theory, and Practice. *Cataloging & Classification Quarterly*, *59*(2–3), 77–96. https://doi.org/10.1080/01639374.2020.1850589

Lepine, R. M. (1988). *Swahili Newspaper Fiction in Kenya: The Stories of James I. Mwagojo* [Doctoral dissertation, University of Wisconsin-Madison].

Rossenova, L., Duchesne, P., & **Blütmel, I.** (2022). Wikidata and Wikibase as complementary research data management services for cultural heritage. In L. Kaffee, S. Razniewski, G. Amaral, & K. S. Alghamdi (Eds.), *Wikidata 2022: Wikidata Workshop 2022, Proceedings of the 3rd Wikidata Workshop 2022 co-located with the 21st International Semantic Web Conference (ISWC2022)*. https://ceur-ws.org/Vol-3262/paper15.pdf (last accessed: 2025/11/20).

Shimizu, C., Hitzler, P., Gonzalez-Estrecha, S., Goeke-Smith, J., Rehberger, D., Foley, C., & **Sheill, A.** (2023). The Wikibase Approach to the Enslaved.Org Hub Knowledge Graph. In T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, & J. Li (Eds.), *The Semantic Web – ISWC 2023: 22nd International Semantic Web Conference, Athens, Greece, November 6–10, 2023, Proceedings, Part II* (pp. 419–434). Berlin: Springer. https://doi.org/10.1007/978-3-031-47243-5_23

Srinivasan, R., & **Huang, J.** (2005). Fluid ontologies for digital museums. *International Journal of Digital Libraries*, *5*, 193–204. https://doi.org/10.1007/s00799-004-0105-9

Varvantakis, C. (2025). *How to… Wikibase?* https://doi.org/10.5281/zenodo.15828659 (last accessed: 2025/11/11).

Vrandečić, D., & **Krötzsch, M.** (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85. https://doi.org/10.1145/2629489

TO CITE THIS ARTICLE:
Aangenendt, G., Harris, A., Wu, D., Maen, A., & Oberst, U. (2025). Wikibase Solutions for African Literary Metadata. *Journal of Open Humanities Data*, 11: 79, pp. 1–10. DOI: https://doi.org/10.5334/johd.433

Submitted: 26 October 2025
Accepted: 25 November 2025
Published: 22 December 2025

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.