Reimagining Wikimedia Incubation: Integrating Encyclopedic, Structured, and Linguistic Data into the Wikimedia Ecosystem

Daniel Antal
Reprex,
University of Amsterdam

Kata Gabor
INALCO

Ieva Pigozne
University of Latvia,
Institute of Latvian History

Abstract

This project proposes a research-driven overhaul of the Wikimedia Incubator model, based on lessons learned from long-stagnant initiatives like the Livonian and Mari Wikipedias. These examples reveal structural limitations: the current model depends on a critical mass of volunteer editors and formal institutional support—resources that many small or endangered language communities do not have.

Drawing on methods developed in the Horizon Europe–funded Open Music Observatory, we propose a data-centric incubation model using the Wikibase Suite and a Wikibase-centered data sharing space, grounded in the European Interoperability Framework and the Data Governance Act. Our team, an interdisciplinary collaboration working without formal funding, has already demonstrated this alternative approach through a poster and paper presented at the DHNB 2025 conference. Our exploratory, alternative incubation platform for Finno-Ugric minorities is available on https://reprexbase.eu/fu/index.php?title=Main_Page, with interesting traditional collections and contemporary collection of minority language films, Mari ethno-punk, Udmurt folktronica, Seto death metal, and some newly disseminated Commons items.

Rather than replicating encyclopedic content from larger Wikipedias, our model focuses on curating and narrating cultural universes—heritage artifacts, oral histories, and linguistic resources—collected in collaboration with community members, NGOs, and humanities researchers. These are published as structured exhibitions and multilingual datasets that can be enriched and described by the communities themselves (e.g., Võro, Seto, Latgalian, Livonian).

The research will result in:

- Peer-reviewed publications on semantic modeling, applied cultural studies, community data stewardship, and incubation governance;
- Open, multilingual datasets federated with Wikidata and Wikibase;

- A replicable framework for community-led incubation, grounded in legal and semantic interoperability.

This work advances Wikimedia 2030 goals by creating inclusive, interoperable pathways for underrepresented languages and cultures to thrive within the Wikimedia ecosystem.

Introduction

The current Wikimedia Incubator model, while conceptually inclusive, has proven ineffective for many small or endangered language communities. Projects such as the Livonian and Mari Wikipedias have remained stagnant for over a decade, with limited engagement from both speakers and researchers. These outcomes highlight a deeper structural issue: the model relies heavily on sustained volunteer activity, formal institutions, and editorial capacity—elements that are often absent in small-language contexts.

Communities such as the Võro, Seto, Latgalian, and Livonian peoples typically do not have national language institutions or major editorial infrastructures. Their heritage materials are often held by small NGOs, local researchers, or community archives. These groups face significant barriers in using Wikimedia tools, which do not easily accommodate the integration of structured data from private or fragmented sources. Current workflows also lack support for multilingual metadata and culturally sensitive reuse.

Meanwhile, content creation in dominant languages—especially English—is being rapidly amplified by large language models and automated recommendation systems. Without new incubation approaches, these small-language communities risk digital invisibility, even in their own cultural domains.

This project addresses these limitations by testing an alternative, data-centric model of incubation. Rather than requiring a critical mass of editors to launch a new language edition, we scaffold Wikipedia content through structured data, Lexemes, and community-curated metadata. We build a lower-threshold, higher-trust pathway into the Wikimedia ecosystem by enabling small communities to start with their own cultural materials—photographs, garments, oral histories, and linguistic fragments—and build from there, using Wikibase, Sampo-UI, and Wikidata-compatible workflows.

By combining structured data technologies with community-led stewardship, this project tests a replicable, Wikimedia-aligned alternative to conventional incubation. It advances knowledge equity not by scaling content, but by scaling participation—on the communities' own terms.

In doing so, we aim to answer the following research questions:

- How can a federated, Wikibase-based data sharing space support the co-creation of encyclopedic and linguistic content in small-language communities with minimal institutional infrastructure?

- What governance and interoperability models (legal, organizational, semantic) are required to ethically integrate data from GLAM institutions (galleries,

libraries, archives, and museums), private collections, and community contributors in line with Wikimedia and European standards?

- Can structured data workflows (including Lexeme-based content and metadata enrichment) enable an alternative, sustainable incubation pathway for Wikipedia, Wiktionary, and other Wikimedia projects in small languages?

- How can digital humanities researchers and computational linguists collaborate with community members to produce semantically structured, multilingual, culturally rooted knowledge?

- How can linked data and semantic technologies enhance the study of small communities' cultural heritage by improving the integration and representation of traditional dress and folklore across community-held and GLAM institution collections?

- What safeguards are needed to ensure that AI tools used for enrichment, translation, or semantic linking support—not replace—community voice and agency?

These questions guide the development of a Wikimedia-aligned incubation approach that emphasizes sustainability, multilingual inclusivity, and community ownership.

Date: July 1, 2025 - June 30, 2026.

Related Work

This project grows from a sustained, interdisciplinary collaboration between researchers in digital humanities, computational linguistics, open data governance, and Wikimedia outreach.

Our team began working together informally, without funding, to address challenges in Wikimedia incubation for small-language communities. This collaboration has already produced a paper and poster presented at *DHNB 2025* (*Digital Humanities in the Nordic & Baltic Countries* conference), and a functioning prototype: the Finno-Ugric Dataspace. These early results validate our workflows for structured data ingestion, multilingual modeling, and legal interoperability in Wikibase and Wikimedia environments.

Our methodological foundations build on the Horizon Europe–funded *Open Music Europe* project, where we developed a federated data infrastructure for cultural and linguistic metadata. This platform follows the European Interoperability Framework (EIF) and the Data Governance Act (DGA), and integrates tools like Sampo-UI, SPARQL, and Wikibase. Our adaptation of this model for Wikimedia demonstrates its generalizability across sectors and languages.

In the broader research landscape, our work builds on efforts to enrich and link multilingual cultural heritage data using NLP and semantic technologies. Projects in digital humanities have demonstrated the value of semantic annotation, named entity recognition, and multilingual knowledge graph construction for analyzing under-documented languages and

collections (Fiorucci et al., 2020; Münster & Terras, 2020). However, small communities often face barriers in applying these methods due to lack of tools, access, or expertise (Jehangir et al., 2023; Labusch & Neudecker, 2020).

Our approach addresses these gaps by grounding NLP and semantic tooling in community participation and GLAM collaboration. We emphasize use of Wikibase and Lexeme infrastructure not only for research, but also for culturally grounded Wikimedia content creation. Community-led modeling and structured annotation are central to our method.

We also align with calls from the Wikimedia Movement for novel socio-technical solutions supporting underserved communities. Our work reflects the Wikimedia 2030 strategic priority of Knowledge Equity and responds to past limitations of the Incubator model by designing an approach grounded in community capacity, structured data, and semantic clarity. Together, this body of work positions our project at the intersection of applied research, infrastructure innovation, and multilingual community empowerment.

In the applied cultural studies of communities' cultural heritage—particularly their traditional dress and folklore—we base our work on a combination of Dr. Ieva Pigozne's textile classification expertise and the use of structured metadata and image integration. This approach will allow us to document and analyze vernacular heritage objects in a way that supports comparative studies of traditional dress and folklore across communities that are geographically close but historically separated by language, religious affiliation, and administrative borders. Our methodology builds on existing practices in digital heritage documentation while addressing the underrepresentation of these communities in structured and multilingual digital environments (Simeone et.al. 2021; Goodrum et.al. 2017; Skrydstrup 2014).

Methods

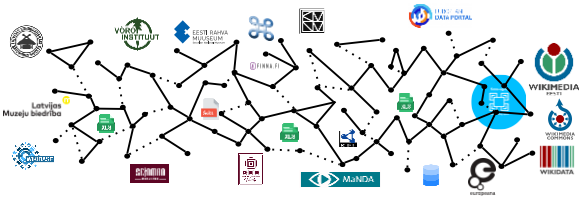
This project builds on tested workflows developed through the Horizon Europe–funded *Open Music Europe* initiative and implemented in our live, operational prototype: the Finno-Ugric Dataspace. This system, based on the Wikibase Suite and aligned with the European Interoperability Framework (EIF), serves as the backbone of our proposed methodology. It demonstrates multilingual, legally grounded, and semantically rich data integration from cultural institutions, private archives, and fieldwork contributions. These foundations now allow us to test a replicable incubation model within the Wikimedia ecosystem.

<u>Data Collection and Curation:</u>
We will collect, digitize, and enrich cultural heritage data relevant to the Võro, Seto, Latgalian, and Livonian communities. Our sources include:

- Public GLAM collections (e.g., Estonian National Museum, Latvian National Library, Finnish and Hungarian archives housing Finno-Ugric material);
- Private and family archives accessible through community partners;
- Open research outputs from previous projects such as TextileBase and Open Music Europe;

- Oral recordings and cultural artifacts gathered via participatory workshops.

We will collect, structure, and publish data on traditional dress and folklore from small communities and GLAM institutions using interoperable, linked, and open data formats to foster knowledge exchange and reuse. The project will apply semantic technologies to connect community-held and institutional sources, enabling integrated, multilingual, and richly contextualized representations of cultural heritage. This methodology supports comparative analysis and long-term accessibility across digital platforms.



On this sematic diagram, horizontal data exhange ↕ results in data being enriched by connected Estonian, Hungarian, Latvian museums, archives and libraries. These state institutions invested into the preservation of Finno-Ugric material and immaterial culture, but their systems are not even searchable in the communities' languages. Together with the communities we will describe their own heritage in their own languages and re-disseminate them (➔ towards the right on the diagram) on the alternative experimental Wikimedia Incubator, and eventually on Wikimedia Eesti's domains, on Commons, Wikidata, and Wiktionary.

All collected materials will be curated for semantic integrity, language attribution, and community relevance. Metadata will follow established ontologies such as CIDOC-CRM, DCTERMS, ESCO, and DDI, and will be aligned with Wikidata's modeling practices.

Structured Data Infrastructure:

The technical foundation of our project is a **Wikibase-centered data sharing space**, already in production and continually updated. This infrastructure supports:

- Multilingual labeling and translation workflows;
- Lexeme extensions for modeling under-documented vocabulary;
- RDF and SPARQL interfaces;
- Open APIs and GUI access via a Sampo-UI frontend;
- Data synchronization with Wikidata and Wiktionary through reconciliation tools.

Custom R-based libraries developed in our previous work (wbdataset, dataset) support seamless transformation from spreadsheets, CSVs, and archival systems into structured, linked data. These pipelines are critical for reproducibility and reusability.

Legal, Semantic, and Organizational Interoperability:

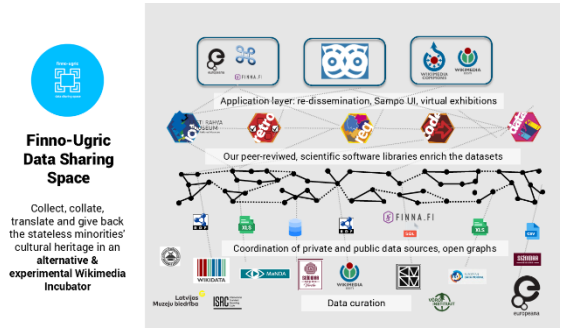
We explicitly implement the four-layer model of the European Interoperability Framework:

- **Legal interoperability**: All content will follow the EU's Data Governance Act and Wikimedia's reuse policies. Reuse agreements and proper attribution models will be established for private collections and community-generated content.
- **Organizational interoperability**: We design workflows that align GLAM metadata processes, research data

practices, and Wikimedia-compatible ingestion.

- **Semantic interoperability**: We harmonize taxonomies across platforms (e.g., NACE, ISCO, VIAF, ISRC) and link cultural data to multilingual identifiers.
- **Technical interoperability**: Our infrastructure adheres to linked open data principles (LOD) and supports federated querying and integration.

This approach ensures legal compliance, data quality, and long-term reuse within Wikimedia and the broader European data ecosystem.



Our alternative and experimental Wikimedia Incubator will provide a different, community and researcher controlled peer review, and a combined incubation to reach a critical mass of not only encyclopedia text, but relevant Commons, Wikidata, Wiktionary, Lexeme assets that are useful for these communities (and their researchers) to invest into a Võro, Seto, Latgalian, or Livonian open knowledge platform that remains long-term curated and sustainable.

Community Engagement and Participation:
At the center of our approach is community stewardship. We co-develop content with the Võro, Seto, Latgalian, and Livonian communities through:

- Local workshops and on-site field visits in Võrumaa and Latgale;
- Training sessions in Wikidata, Wikibase, Lexeme editing, Commons, and Wiktionary;
- Oral documentation initiatives using LinguaLibre and structured Wikibase statements;
- Virtual exhibitions curated collaboratively with community members.

Our goal is not to convert participants into editors in the traditional sense, but to support them as narrators and interpreters of their own heritage. Workshops are led in collaboration with regional actors and trained Wikimedia facilitators.

Lexicographic and Linguistic Modeling:
We import and generate linguistic data relevant to the communities' languages. This includes:

- Stem dictionaries and morphological models from prior research;
- Manually created Lexemes with community input;
- Semi-automated inflection, glossing, and linking via SPARQL and Lexeme APIs.

This component enables Wikimedia-compatible content in Wiktionary and provides language resources for both native speakers and computational linguists working in low-resource contexts.

Experimental Incubation:
The project will test a new model of incubation that bypasses the need for a large editorial community. Instead, we bootstrap Wikimedia

content using:

- SPARQL-based infoboxes and data visualizations;
- Commons galleries tied to cultural metadata;
- Article scaffolding and templates created from structured Lexeme and entity data;
- Multilingual labels and sitelinks generated through curated translation workflows.

We will begin with a focus on documenting community-specific cultural universes—traditional garments, traditional music, oral histories—rather than general encyclopedic entries.

Expected Output

This project will produce tangible, open outputs across four domains: academic research, community tools, Wikimedia content, and policy alignment.

1. Scientific Publications and Presentations:
We plan to publish at least two peer-reviewed journal articles:

One in digital humanities or knowledge representation, building on our DHNB 2025 conference presentation, focusing on multilingual, semantic workflows for Wikimedia environments and another in applied cultural or folklore studies, centered on the modeling of Latgalian and Seto traditional dress using Wikibase and structured metadata.

In addition, we will present results at: **DHNB 2026** (follow-up to 2025 session); **Wikimania 2026** (e.g., sessions on structured data, small-language projects, or Lexemes).
Where possible, we will publish open access via institutional affiliations or community support, minimizing APC costs.

2. Public APIs and Open Datasets:
We will expand the existing Finno-Ugric Dataspace (powered by Wikibase and Sampo-UI) by adding public SPARQL endpoint, federation with Wikidata and TextileBase, and Open-source ingestion tools and templates.
All datasets (Lexemes, media metadata, cultural entities) will be released under **Wikimedia-compatible open licenses** and structured for reuse.

3. Wikimedia Contributions:
We will generate new or enriched content across **Wikipedia** (pilot encyclopedic articles and exhibitions in Võro and Latgalian reflecting our collected data), **Wikidata** (new items, Lexemes, and multilingual labels), **Wikimedia Commons** (culturally significant media linked to structured metadata), and **Wiktionary** (via Lexeme-linked glossaries and oral recordings from community contributors).
Workflows and tools will be documented for reuse by other small-language projects and GLAM-Wiki initiatives.

4. Strategic Insights:
We will compile an "Incubator 2.0" summary report outlining:

- Legal, semantic, and organizational best practices for multilingual content;
- Case study insights from our community pilots;
- Policy-relevant takeaways for Wikimedia, GLAMs, and EU data spaces.

These outputs support a broader shift toward structured, inclusive incubation aligned with both Wikimedia 2030 and European data governance frameworks.

Risks

We anticipate three primary categories of risk in this project: limited community participation, legal/ethical complexities in data reuse, and sustainability beyond the grant cycle. Each is addressed through specific mitigation strategies built into our workflows.

1. Limited Community Participation:
A key risk is insufficient engagement from Võro, Seto, Latgalian, and Livonian language communities. Many of these groups are small, lack institutional infrastructure, and may be unfamiliar with Wikimedia tools.
Mitigation: We work through trusted regional actors like the Võro Institute, Wikimedia Estonia, and the Traditional Culture Centre (Latvia), who have long-standing relationships with these communities. Our workshops emphasize culturally meaningful content and flexible contribution methods (e.g., oral narration, tagging photos) to reduce participation barriers.

2. Legal and Ethical Uncertainty in Data Reuse:
Community-held heritage, private archives, or unstructured GLAM metadata may present challenges related to attribution, licensing, or cultural sensitivity.
Mitigation: Our project applies the European Interoperability Framework (EIF) and follows the Data Governance Act (DGA). We will only reuse materials with clearly defined rights and consent, establish reuse agreements with content providers, and ensure transparency through metadata and documentation.

3. Sustainability and Long-Term Maintenance:
There is a risk that the structured content, tools, and workflows may not continue to be used or updated beyond the grant period.
Mitigation: All tools and data will be developed within Wikimedia-compatible environments (Wikibase, Lexeme, SPARQL). Documentation, multilingual UI (via Sampo-UI), and modular design will support handoff to GLAM-Wiki groups and other small-language communities. Where feasible, institutional hosting will be secured through partners like TextileBase, Reprex, or Wikimedia Estonia.

Community Impact Plan

This project is designed with community stewardship at its core. We aim to support underrepresented language communities—Võro, Seto, Latgalian, and Livonian—not just as data contributors but as co-curators of their cultural and linguistic heritage. Among these, the **Võro community is a primary focus**, due to its more developed linguistic infrastructure, organized institutions, and history of participation in Wikimedia initiatives.

1. Participation Through Curation, Not Just Editing:
Traditional Wikimedia incubation relies on consistent editorial activity, which can be difficult to sustain in small-language contexts. Our approach reduces this burden by:

- Starting from digitized cultural resources (photos, oral traditions, garments, lexemes);
- Offering guided, multilingual interfaces through the Finno-Ugric Dataspace;
- Supporting contributions such as narration, translation, and annotation instead of article writing.

This makes participation accessible and culturally relevant, particularly in communities with active local heritage efforts, such as Võro.

2. Regional Workshops and Local Anchors:
We will host two community workshops:

- One in **Võrumaa (Estonia)**, focusing on the Võro community, which benefits from active organizations like the Võro Institute and Wikimedia Estonia. This site is ideal for piloting lexeme enrichment, cultural metadata annotation, and small-scale Wikipedia seeding.
- One in **Rēzekne (Latvia)**, where we will engage Latgalian speakers and museum partners in content creation and lexeme work.

These workshops will include training in Wikimedia tools (Commons, Lexeme, Wikidata, Wiktionary), with support from experienced facilitators and local Wikimedians. We will also collect oral and visual documentation for structured reuse.

3. Collaboration Across the Wikimedia Ecosystem:
We work closely with:

- Wikimedia affiliates in Estonia, Latvia, and Finland;
- Volunteer developers and editors with experience in structured content and small-language communities;
- GLAM-Wiki partners committed to multilingual metadata;
- Events like the Celtic Knot Conference, where we will share tools and governance models.

Our open-source datasets and workflows are designed for reuse by other small-language initiatives in the Wikimedia ecosystem and beyond.

4. Building Equity and Resilience:
The Finno-Ugric Dataspace is a platform for multilingual engagement—not just preservation. By prioritizing the Võro community, which has both linguistic assets and digital engagement history, we maximize the potential for success and reuse. This model is designed to scale outward to less resourced communities (e.g., Livonian) over time.
In doing so, we advance the Wikimedia 2030 goal of **Knowledge Equity**, supporting content generation in languages and formats meaningful to local communities.

Dissemination

Our dissemination strategy is designed to reach both scholarly and Wikimedia audiences, while

also prioritizing accessibility for the communities involved in the project. We will share outcomes through conferences, open publications, digital platforms, and direct community engagement.

Academic Dissemination:
We will submit at least two peer-reviewed journal articles based on the research: one focused on semantic interoperability and cultural data governance, and another on lexeme workflows and multilingual content modeling. We also plan to return to the *Digital Humanities in the Nordic and Baltic Countries (DHNB)* conference in 2026 with new results. Where possible, publishing fees will be waived through institutional affiliations and funding coverage.
We will organize a scientific symposium at the University of Latvia, bringing together Wikimedia researchers, folklorists, and digital humanists to present results and plan future collaborations.

Wikimedia and Community Events:
We will present findings at **Wikimania 2026 in Paris**, with Asmah Federico and Kata Gábor as primary presenters, and will share additional updates through structured data and small-language sessions (e.g., at Celtic Knot, CEE Meeting). Two public-facing workshops will take place: one in **Võrumaa (Estonia)** and one in **Rēzekne (Latvia)**, co-organized with local partners to showcase community-curated knowledge.

Open Tools and Documentation:
All code, workflows, and documentation will be published under open licenses through GitHub and Zenodo. Multilingual datasets, SPARQL endpoints, and onboarding guides for Lexeme and Commons work will remain accessible through our Finno-Ugric Dataspace. This ensures both technical reusability and long-term benefit to Wikimedia contributors, educators, and regional institutions.

Team and Collaborators:
Our team is an interdisciplinary group of researchers and practitioners who began collaborating independently, without formal funding, to address the challenges of incubating Wikimedia content in small and endangered languages. Our collaboration has already resulted in a paper and poster presented at the *Digital Humanities in the Nordic and Baltic Countries (DHNB) 2025* conference, and is backed by ongoing technical development and community engagement through our **Finno-Ugric Dataspace prototype** (see https://reprexbase.eu/fu/index.php?title=Main_Page)

Team Members:
Daniel Antal – Data scientist and co-founder of Reprex. Daniel brings extensive expertise in legal, organizational and semantic interoperability, open data policy, and data governance. He is the Work Package leader for data infrastructure in the Horizon Europe–funded *Open Music Europe* project, where he develops FAIR-aligned, EU-regulatory-compliant data infrastructures using Wikibase and R-based tooling. He represents Reprex in the Big Data Value Association (BDVA) and works closely with the Data Spaces Support Centre (DSSC). He is also a research fellow at the Information Law Research Institute of the University of Amsterdam, Wikipedian since 2007 and member of WP:WMM.
Dr. Ieva Pigozne – Cultural historian and expert in Baltic and Finno-Ugric traditional culture.

Ieva leads fieldwork, metadata modeling, and narrative structuring for heritage objects. She initiated the TextileBase project and contributes deep regional expertise in vernacular traditions, textile typologies, and community-curated collections.

Dr. Kata Gábor – Computational linguist based at INALCO, Paris. Kata specializes in multilingual metadata harmonization and semantic data integration. She oversees our linguistic interoperability strategies and Lexeme modeling, working across Latvian, Estonian, Hungarian, Finnish, and Russian sources.

Asmah Federico – Community organizer and Wikimedia outreach expert. Based in Tartu, the nearest university town to Setomaa and Võrumaa, Asmah studies at the University of Tartu and volunteers with Wikimedia Estonia's Tartu office. Her work supports community participation in both technical workflows and cultural knowledge sharing.

Edīte Punka – Research assistant. Edīte is a master's student in Library and Information Science at the University of Latvia, with practical experience in museum archiving, metadata curation, and exhibition management. She contributes to data collection, dataset curation, content annotation, and structured ingestion from community and GLAM sources.

The project will be hosted and administered by the **Traditional Culture Center** (**TCC**, *Tradicionālās kultūras centrs*), a Latvian NGO led by Valdis Putniņš and widely known for its long-running initiative *Rīgas Danču klubs*. The organization has over 20 years of experience promoting and researching traditional culture and folklore across Latvia, Estonia, and Lithuania. TCC has an extensive track record of collaboration with local communities and folklore groups, organizing both grassroots events and large-scale international festivals and summer camps. It maintains a well-established network across the Baltic countries and is also actively engaged in documenting and researching contemporary folklore activism in Latvia.

Collaborators:

We are supported by a growing network of institutional and community partners:

- **Wikimedia Estonia**, with whom we coordinate outreach, training, and technical alignment;
- The **Võro Institute**, a regional leader in language preservation and prior host of Wikipedians-in-residence;
- GLAM partners in Estonia, Latvia, and Finland, including custodians of traditional dress, photography, and music heritage;
- **DSSC and BDVA**, through which we contribute to European policy and technical frameworks for data sharing and reuse.

We also maintain close ties with NGOs and grassroots archives in Setomaa, Latgale, and Võro-speaking areas. These relationships are central to the long-term trust, reuse, and cultural relevance of the project.

Evaluation

We will evaluate this project across three core dimensions: research outcomes, community impact, and sustainable integration with Wikimedia projects. These criteria align with

the Wikimedia Research Fund's expectations and our interdisciplinary goals.

1. Research and Scholarly Outputs:
Success indicators: At least two peer-reviewed journal articles (including one following our DHNB 2025 presentation), conference participation (DHNB 2026, Wikimania 2026), and structured datasets published under open licenses.
We will track citations, dataset downloads, SPARQL endpoint use, and reuse by Wikimedia contributors, GLAM institutions, and scholars.

2. Community Engagement and Cultural Contribution:
Success indicators: Two in-person workshops (in Setomaa and Latgale), one academic symposium (at the University of Latvia), and demonstrable community contributions (e.g., oral recordings, lexemes, and media annotations).
We will document participant feedback, measure retention and follow-up contributions to Wikimedia platforms, and log localized content growth (e.g., Võro or Latgalian lexemes or Commons files).

3. Structured Content Integration and Sustainability:
Success indicators: Federation of data with Wikidata and Commons, live SPARQL endpoint, and multilingual UI (via Sampo-UI).
We will document technical workflows and governance models to support future replication and interoperability, contributing to ongoing discussions around small-language support and Wikibase-centered data stewardship.

The project's success will be determined by its ability to demonstrate a viable, community-led model of structured content incubation that is ethically grounded, semantically rich, and sustainable beyond the grant period.

Budget

This project is designed as a 12-month initiative, running from **July 1, 2025 to June 30, 2026**, with funding requested to support personnel, travel, computing resources, dissemination activities, and institutional overhead. Our total request is **$50,000 USD**, the maximum eligible amount under the Wikimedia Research Fund.
Link to the budget spreadsheet: https://docs.google.com/spreadsheets/d/1pnHZfAswxPtsRQmXwzVy4TzXQyMJ8kUXpfjRWZotAME/edit?gid=0#gid=0

Personnel:
Two early-career researchers, **Asmah Federico** and **Edīte Punka**, will be engaged part-time and paid for their contributions. Asmah will lead community workshops, facilitate participatory contributions, and support outreach in Estonia and Latvia. Edīte will contribute to data collection, metadata curation, digitization, and ingestion of cultural collections into our Wikibase platform.
Dr. Ieva Pigozne will be partly compensated for her work leading cultural modeling, field research, and community coordination in Latvia and Estonia.
Dr. Kata Gábor and **Daniel Antal** hold full-time academic and professional positions. They will not receive salaries but will be reimbursed for participation in conferences and community-facing events, including Wikimania and DHNB 2026.

Dissemination and Travel:
Dissemination funds will cover:

- Community events in **Võrumaa** (Estonia) and **Rēzekne** (Latvia), where we will host multilingual workshops and share community-curated data;
- A scientific symposium at the University of Latvia;
- Participation in **Wikimania 2026 (Paris)** with presentations;
- Presentation of new findings at **DHNB 2026**, building on our 2025 contributions;
- Open access publishing for two planned journal articles (where fee waivers are unavailable);
- Translation, editing, and documentation support to ensure multilingual accessibility and cross-platform adoption.

These investments ensure our work is visible within both Wikimedia and scholarly communities, while also returning direct benefits to the communities involved.

Compute Resources and Infrastructure:
Compute resources will fund software tools, digital storage, and maintenance of our Wikibase platform and SPARQL endpoint. This will ensure long-term access to the Finno-Ugric Dataspace, which will continue to support community access and scholarly reuse beyond the project.

Institutional Overhead:
Institutional overhead is budgeted at **15%** of the total, in accordance with Wikimedia Foundation policy.

Summary Budget Breakdown:

- **Personnel Support**: $20,000
- **Dissemination Activities**: $16,000
- **Computer Resources**: $3,500
- **Travel and Conference Participation**: $3,000
- **Institutional Overhead**: $7,500
- **Total Requested**: **$50,000 USD**

All project outcomes—datasets, tools, workflows, and training materials—will be released under open licenses and hosted in multilingual, community-accessible platforms.

References

Antal, D. & Grochal, M. (2024). Building a Music Data Sharing Space with Wikibase. Presented on Wikimedia CEE Meeting 2024. https://doi.org/10.5281/zenodo.8046977

Antal, D. (2017). The growth of the Hungarian popular music repertoire: Who creates it and how does it find an audience. In: Made in Hungary (1st ed., Studies in Popular Music series). New York, NY: Routledge.

Antal, D. (2024). Making Datasets Truly Interoperable and Reusable In: R. The specific case of working with the Wikibase Data Mode 10.32614/CRAN.package.dataset

Antal, D. (2023). Pilot Program for Novel Music Industry Statistical Indicators in the Slovak Republic. https://doi.org/10.5281/zenodo.10372026

Antal, D. (2023). The Planned Slovak Comprehensive Music Database. https://doi.org/10.5281/zenodo.10392759

Antal, D. (2024). A szlovák adatkicserélési tér magyarországi föderációjának lehetőségei. https://doi.org/10.31915/NWS.2024.25

Benkhedda, Y., Skapars, A., Schlegel, V., Nenadic, G., & Batista-Navarro, R. (2024). Enriching the Metadata of Community-Generated Digital Content through Entity Linking: An Evaluative Comparison of State-of-the-Art Models. Proceedings of the 8th Joint SIGHUM Workshop (LaTeCH-CLfL 2024), 213–220.

Curry, E. (2020): Fundamentals of Real-time Linked Dataspaces. In: Real-time Linked Dataspaces: Enabling Data Ecosystems for Intelligent Systems. Springer International Publishing. https://doi.org/10.1007/978-3-030-29665-0_4, ISBN: 978-3-030-29665-0

Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Del Bue, A., & James, S. (2020). Machine learning for cultural heritage: A survey. Pattern Recognition Letters, 133, 102–108.

Gábor, K., Buscaldi, D., Schumann, A., QasemiZadeh, B., Zargayouna H., & Charnois T. (2018). Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In: Proceedings of The 12th International
Workshop on Semantic Evaluation, 679–688.

Gábor, K., Zargayouna H., Buscaldi, D., Tellier, I., & Charnois, T. (2016). Semantic annotation of the ACL anthology corpus for the automatic analysis of scientific literature. In: Proceedings of the Language Resources and Evaluation Conference (LREC).

Gábor, K., & Sagot, B. (2014). Automated error detection in digitized cultural heritage documents. In: Proceedings of the EACL 2014 Workshop on Language Technology for Cultural Heritage, Göteborg, Sweden.

Gábor, K., Zargayouna H., Tellier, I., Buscaldi, D., & Charnois, T. (2016). Unsupervised relation extraction in specialized corpora using sequence mining. In: International Symposium on Intelligent Data Analysis. Springer, 237–248.

García-Mendoza, Juan-Luis, Buscaldi, D., Bustio-Martínez, L., Gábor K., Zargayouna, H., Charnois T. ,& Herrera-Semenets V. (2024). Towards a novel approach for knowledge base population using distant supervision. In: Mexican Conference on Pattern Recognition. Springer Nature Switzerland, 96-106.

Goodrum, A., & Dalrymple, L. (2017). Digitizing Dress: Creating an Image-Based Classification System. Fashion Theory, 21(1), 63–92. https://doi.org/10.1080/1362704X.2016.1147450

Hanna, E., Hughes, L. M., Noakes, L., Pennell, C., & Wallis, J. (2021). Reflections on the Centenary of the First World War: Learning and Legacies for the Future. Arts and Humanities Research Council.

Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023). A survey on Named Entity Recognition—datasets, tools, and methodologies. Natural Language Processing Journal, 3, https://doi.org/10.1016/j.nlp.2023.100017

Kumar, A., Pandey, A., Gadia, R., & Mishra, M. (2020). Building knowledge graph using pre-trained language model for learning entity-aware relationships. IEEE GUCON, 310–315.

Kung, A., Walshe, R., & Wenning, R. (2023). Data Sharing Spaces and Interoperability. BVDA Position Paper. https://bdva.eu/download/92/publications/3841/data-sharing-spaces-and-interoperability-bdva-discussion-paper-december-2023.pdf

Labusch, K., & Neudecker, C. (2020). Named Entity Disambiguation and Linking Historic Newspaper OCR with BERT. In: CLEF Working Notes.

Münster, S., & Terras, M. (2020). The visual side of digital humanities: a survey on topics, researchers, and epistemic cultures. Digital Scholarship in the Humanities, 35(2), 366–389.
Nagel, L., & Lycklama, D. (2020): Design Principles for Data Spaces. Position paper. https://doi.org/10.5281/zenodo.5244997

Open Music Europe. (2023). Open Music Europe (OpenMusE): An Open, Scalable, Data-to-Policy Pipeline for European Music Ecosystems. https://doi.org/10.3030/101095295

Pīgozne, I. (2018). Reuse of Garments and Accessories of Latgalian Clothing in the 19th Century. In: Proceedings of the 60th International Scientific Conference of Daugavpils University: Humanities, 40–48.

Pīgozne, I., & Antal, D. (2024). Linked Open Datasets on Garments from the Latgale Region. https://doi.org/10.5281/zenodo.13971707

Pīgozne, I. (2024). Sacred Footwear: Latvian Perceptions in the 19th Century and Today. Yearbook of Balkan and Baltic Studies, 7(1), 197–212. https://doi.org/10.7592/YBBS7.08

Pīgozne, I. (2020). Tradicionālā apģērba novadu izveides varianti. Latgales piemērs. Latvijas vēstures institūta žurnāls, (2), 5–31. https://doi.org/10.22364/lviz.112.0

Reinsalu, R. (2023). Using Wikipedia for educational purposes in Estonia. Wiki Workshop (10th ed.), Paper 35.

Roach-Higgins, M. E., & Eicher, J. B. (2022). Dress and Identity. In Welters, L., & Lillethun, A. (Eds.). The Fashion Reader (3rd ed.). London: Bloomsbury Visual Arts.

Simeone, M., & Dombrowski, Q. (2021). Metadata as Ethics: Collaborative Cataloging for Cultural Heritage. Digital Humanities Quarterly, 15(3). http://digitalhumanities.org/dhq/vol/15/3/000545/000545.html

Skrydstrup, M. (2014). Digital Repatriation and the Circulation of Indigenous Knowledge. Anthropology Today, 30(6), 12–16.

Toupin, S., & Gendron, M. (2020). Wikipedia and the Representation of Traditional Knowledge. First Monday, 25(8). https://firstmonday.org/ojs/index.php/fm/article/view/10801

Wu, L., Petroni, F., Josifoski, M., Riedel, S., & Zettlemoyer, L. (2020). Scalable Zero-shot Entity Linking with Dense Entity Retrieval. EMNLP 2020, 6397–6407.

Zhang, Z., Liu, X., Zhang, Y., Su, Q., Sun, X., & He, B. (2020). Pretrain-KGE: Learning knowledge representation from pretrained language models. Findings of EMNLP 2020, 259–266.