# Enriching and Futureproofing the Databases of the Hungarian Heritage House

**A short feasibility study and early manuscript for publication (version 0.2.0)**

Antal, Dániel Zagyva, Natália Mester, Anna Márta

## Introduction

> **!** Important
>
> This document serves as a living document for documenting the Reprex-HH documentation to serve as a basis of the forthcoming joint publication.
> This is the version 0.2.0 and it has not been reviewed by *Hagyomanyok Háza* institutionally, it may contain mispresentation of some HH databases.
> DOI: 10.5281/zenodo.17759777; please check for potential later versions before re-using or citing.

### Folklore Databases in Hungary

Folklore databases in Hungary play a crucial role in preserving, systematising, and rendering accessible the country's intangible cultural heritage. The most significant collections and datasets are maintained by leading scientific and cultural institutions that have long served as custodians of ethnographic and musicological knowledge.

The *Ethnological Archives of the Museum of Ethnography* function as the central archival repository of Hungarian ethnography. They hold an extensive corpus of manuscripts, photographs, films, and sound recordings, much of it originating from a nationwide network of community collectors who contributed materials through organised collecting schemes. Through sustained digitisation efforts, the Museum has made a substantial portion of these holdings available as online collections and thematic digital resources. These resources constitute an irreplaceable foundation for contemporary folkloristic research.

The archives of the *Institute of Ethnology* and the *Institute for Musicology* primarily preserve materials accumulated through fieldwork conducted by researchers, along with records arising from archival investigations. They also safeguard the personal papers, manuscripts, and legacies of scholars, folklorists, and composers. Over the past decades, these institutes have developed a number of highly valuable thematic folklore databases, covering the domains of folk music, folk dance, and narrative folklore.

A fourth institution of major importance is the **Heritage House** (Hungarian Heritage House) in Budapest, founded in 2001. Its folklore collections and ongoing development projects are discussed in detail in the following chapter.

For piloting, we chose to work with two data sources.

**500 Folk Music Examples from Székelyföld**: One of the processed collections derives from ethnomusicologist István Pávai's representative selection of folk music from Székelyföld (Szeklerland in Romania.) The examples are organised according to a regional classification system developed by Pávai, based on his field research and tailored to the stylistic and geographical distribution of traditional dance music.

The metadata accompanying the sound recordings include: place of use (macro-region, region, micro-region, settlement); genre; informant (name, age or year of birth); performance mode (instrumentation); date of collection; collector.

The recordings originate from 101 settlements across Székelyföld.

**Folk Tale Inventory**: The second collection is currently unpublished. In 2003–2004, Sára Dala and her colleagues compiled a dataset containing more than 11,000 Hungarian folk tales, intended to serve as the foundation for an online catalogue arranged by geographical area and chronology, supporting the work of researchers, educators, cultural mediators, and storytellers.

Under the working title *Folk Tale Inventory*, the compilers processed the contents of 467 printed volumes. In addition to bibliographic data, the dataset includes the place of use (settlement, county), the informant, the collector, the date of collection, and the tale type. The editors intend to expand the dataset with materials published after 2000, as well as tales available in audio and video formats.

Within the present project, we restrict our work to the material published in 1957 as Háromszék Hungarian Folk Poetry (Háromszéki magyar népköltészet), documenting the collections of Samu Konsza and his students in the Háromszék region. This subset comprises 121 folk tales from 39 Székelyföld settlements.

We chose this collection to show that gradually we can extend our immaterial work model optimised for music to other related immaterial heritage. The transition from musical works to folk tales is natural, because the tales are often preserved as sound recordings, and they have similar textual manifestations as the lyrics of vocal music.

Our aim is to follow the *European Interoperability Framework*[1], and to utilise our work developed in the OpenMusE project, particularly with the creation of the architecture of the *Open Music Observatory*, and the experience gained with its Finno-Ugric federated module. we aim to create not only technically and semantically correct new representations of the databases of Hagyományok Háza (in short: HH) but also workflows (organisational aspect) and data governance (legal aspect) of our work.

[1]The European Interoperability Framework (EIF) not only as a semantic guide but also as a legal and organizational model (Commission and Digital Services 2017; European Commission 2017) to connect various library, museum, archival, geographic information and other services.

In our work, we aim to transfer some of the databases, or their data tables as datasets into a graph format. Particularly suitable data for this task is the person, corporate and geographical name spaces, which need to be permanently updated, and often enriched for further historical or other language variants. Another trivial choice is the representation of a thesaurus in this format, as thesauri per se follow a graph format. Last, but not least, the graph format can help to connect relational databases that are now standing alone, or to re-document and improve the metadata schemas of existing relational databases or their tables.

Linking datasets across domains, institutions, and languages proved more difficult and less successful than the 5-star FAIR model suggests. In practice, findability, interoperability, and reuse are not easily achieved across national metadata practices, multilingual vocabularies, or heritage-specific contexts. For this reason, we adopted the extended **8-star model** of linked cultural data proposed by Hyvönen and Tuominen (Hyvönen and Tuominen 2024), which adds essential dimensions such as provenance, multilingualism, contextualization, participation, and sustainability. This higher standard better reflects the realities of cross-border, community-anchored heritage work and connected unconnected, often less documented databases, Excel sheets, and different information sources.

## Architectural choices

The architectural choices for connecting Hungarian music and folklore datasets to a future-proof, interoperable infrastructure must address three simultaneous challenges:
(1) historically layered and multilingual heritage;
(2) highly heterogeneous and partly undocumented legacy databases; and
(3) limited institutional capacity for large-scale ontology engineering.

For these reasons, we have focused on Wikibase as the core bridging technology. Its suitability is demonstrated most clearly by the National Library of Wales' SNARC (*Shared National Archival Record Catalogue*) architecture, which has already been adopted as a model by cultural-heritage institutions across Europe. SNARC is a directly relevant precedent for HH, because it supports exactly the combination of requirements present in Hungary: distributed collections, multilingual metadata, legacy databases, and active community contributors.

Wikibase has proven effective for multilingual authority control (Bianchini, Bargioni, and Pellizzari di San Girolamo 2021; Fagerving 2023). The Finnish Wikibase pilot of the National Library of Finland demonstrated how it can connect museum, linguistic, and archival records while also supporting community participation (National Library of Finland 2021). Similar experiences in minority-language contexts—such as the Võro User Group (Wikimedians of Võro language User Group 2025) and Wikimedia Norge's Sámi knowledge project (Wikimedia Norge 2020)—show that Wikibase can combine semantic interoperability with respect for knowledge sovereignty and multilingual representation. These experiences are particularly suitable for connecting to less formal civic or community, municipal diaspora

Hungarian collections. These precedents confirm that Wikibase can act as a shared, multilingual curatorial platform for minority and diasporic heritage without imposing a single institutional metadata regime.

The relevance of the Welsh precedent is explicitly confirmed in the documents the National Library has made public and is also reflected in the wider scholarly literature, and given the institutional strength of the Welsh National Library, it is probably the highest scaled subnational example. As summarised in our internal materials:

> "A notable example is the SNARC Wikibase… developed by the National Library of Wales. SNARC allows Welsh libraries, archives, and museums to describe cultural heritage using persistent identifiers and structured metadata in Welsh and English. It supports shared authority work and multilingual reconciliation while respecting institutional diversity and community input."

This architecture provides not only a technical blueprint but also a governance model that suits a heritage ecosystem with many semi-independent contributors, including fieldworkers, regional archives, civic initiatives, and diasporic communities.

Wikibase offers an effective bridge between human-centred curatorial workflows and the formal structures of the semantic web. As an open, collaborative platform, it enables curators, researchers, and community partners to describe heritage materials in a readable, intuitive format, while simultaneously maintaining machine-actionable semantics through RDF export. This duality makes Wikibase particularly suitable for projects that must combine community participation, multilingual metadata, and interoperability with formal ontologies such as HDTO (ECCCH), CIDOC-CRM (museums), Records in Contexts (archives), and DCTERMS)[2].

In contexts characterised by limited archival depth, legacy databases of uneven quality, or undocumented local schema variations, strict ontology-based systems can easily become too rigid for practical curation. Wikibase, by contrast, supports *incremental enrichment*. Incomplete, contradictory, or uncertain statements can coexist with qualified assertions, references, and provenance notes. Curators can therefore record uncertainty—an intrinsic part of folkloristic and ethnographic data—without compromising the coherence of the underlying semantic structures.

Its flexible data model also allows domain ontologies to be introduced progressively: digital-twin classes, heritage-entity relations, component hierarchies, and temporal layers can be added over time without disrupting existing documentation practices. This *progressive modelling* principle is essential for a project like HH, where archival quality, provenance depth, and historical metadata formats vary significantly across collections.

[2]ECCCH: (Commission et al. 2022; ECHOES Ontology Task Force 2025); CIDOC: (Bekiari et al. 2024), RiC (Archives Expert Group on Archival Description 2023), Dublin Core (DCTERMS): (Core 2020).

### The Welsh Model: Why SNARC Is the Most Relevant Precedent

SNARC demonstrates how Wikibase can operate as a national-level authority and reconciliation layer for distributed archives. Its architecture exhibits several properties that map directly onto the needs of Hungarian folklore collections:

1. **Multilingual authority control with community participation**: SNARC maintains bilingual (Welsh/English) labels, descriptions, and authority records, enabling minority and regional languages to coexist with national cataloguing standards. This aligns with HH's need to support Hungarian, regional dialects, and minority languages across the Carpathian Basin.
2. **A federated approach rather than centralisation**: Institutions retain autonomy over their internal databases, while contributing shared authority data into the Wikibase layer. This mirrors the governance model recommended we followed with the prototyping in the Finno-Ugric Data Sharing Space, and fits in with the situation in Hungary: HH must integrate data from diverse field researchers, local cultural centres, and legacy scholarly collections without imposing a single catalogue.
3. **Support for weakly standardised, legacy, and community-derived data**: SNARC deals with inconsistent archival metadata, variant place names, historical spellings, and ambiguous provenance—challenges identical to those found in Hungarian folklore collections.
4. **A sustainable governance model**: Wikibase enables contributors, stewards, validators, and reviewers to participate according to their role. This matches HH's needs: ethnomusicologists, dance researchers, ethnographers, and community collectors can contribute in different capacities.
5. **Compatibility with semantic-web export**: SNARC relies on Wikibase's RDF layer to integrate with external linked-data infrastructures—analogous to our plan to export to HDTO/ECCCH and other European data spaces.

### Wikibase as a Bridging Layer in Heritage Architectures

Architecturally, Wikibase functions as a **bridging layer** between descriptive archives (AtoM, legacy field-record tables, digitised notebooks, manuscript metadata) and semantic infrastructures capable of supporting cross-institutional reasoning. It links curatorial work with graph-based analytics through:

- Federated SPARQL queries
- Multilingual labels and aliases
- Reconciliation services
- Versioned, provenance-rich authority records
- Incremental modelling of heritage ontologies

This architecture has been tested extensively in the **Finno-Ugric Data Sharing Space**, where it supports the representation of complex, multilingual, and historically layered heritage in a transparent, sustainable form. Human curators maintain an accessible knowledge base; automated pipelines generate HDTO-compliant linked-data exports for long-term interoperability.

For the HH pilot, we will therefore:

1. Use the existing, well-curated Finno-Ugric Wikibase instance to test small sample datasets.
2. **E**valuate feasibility and governance requirements for establishing one or two dedicated Wikibase instances:
   – a public-facing authority and catalogue layer,
   – and, if necessary, a more detailed internal-facing instance for sensitive or partially processed data.
3. This staged approach mirrors the Welsh model, where the public Wikibase instance aggregates reconciled data, while institutions retain internal systems tailored to their operational needs.

### Subsidiarity in Data Governance: Proven Suitability for Minority, Regional, and Diasporic Heritage

The suitability of Wikibase for complex cultural ecosystems is supported by comparative examples:
– the National Library of Finland's Wikibase pilot,
– the Võro User Group's regional-language knowledge base,
– Wikimedia Norge's Sámi metadata projects,
– and other community-based knowledge initiatives.

These projects demonstrate that Wikibase can combine **semantic interoperability** with **knowledge sovereignty**, a crucial requirement for representing the diverse, often community-generated folklore and folk-music heritage that HH oversees.

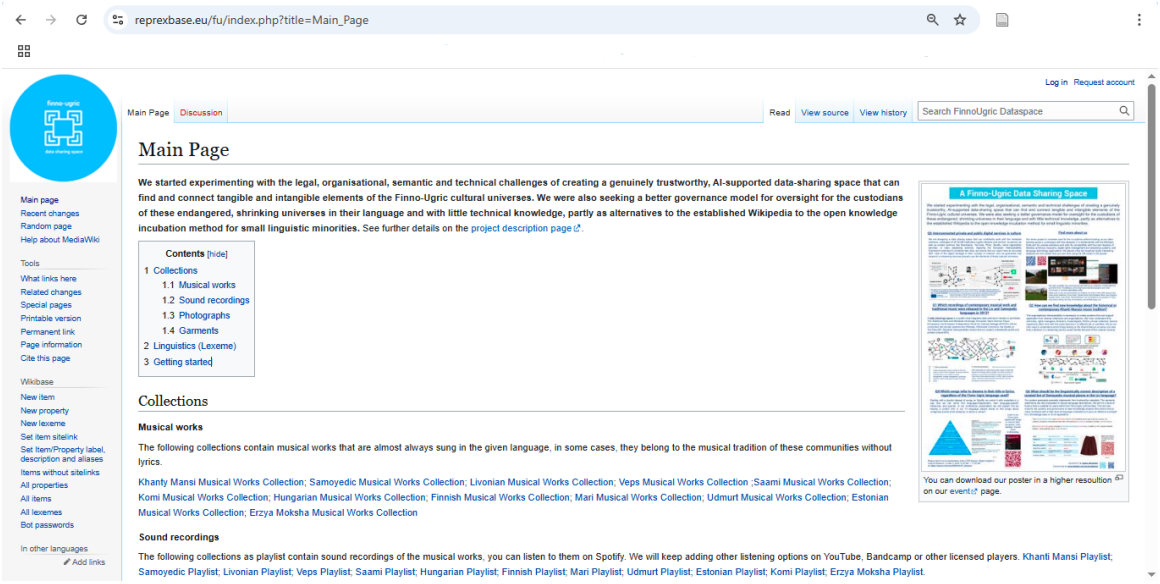



Figure 1: Figure 1: The opening page at https://reprexbase.eu/fu/index.php for the Finno-Ugric Data Sharing Space

Within the *Finno-Ugric Data Sharing Space*, this approach allows complex, multilingual, and historically layered heritage to be represented transparently and sustainably: human curators maintain an accessible knowledge base, while automated exports produce HDTO-compliant datasets for long-term interoperability. We will test this well-developed and curated Wikibase instance for the first phase of testing small data samples from HH, then we review the feasibility of building one or two dedciated (for example, public facing and separate internal facing) instances for HH.

### Alignment with European Research Infrastructures

Reprex aligns the Wikibase data model with the *Heritage Digital Twin Ontology*(HDTO), the semantic backbone of the *European Collaborative Cloud for Cultural Heritage* (ECCCH). This alignment ensures that HH's data will be future-proof, interoperable with European heritage clouds, and ready for integration into AI-assisted digital humanities tools.

In this respect, Wikibase functions not merely as a curatorial interface, but as an enabling architecture for open, interoperable cultural-heritage data within the emerging European research and data-space ecosystem.

## Viability and competences

Because Reprex works with very similar data from the Latvian Folk Archive, and various smaller Nordic and Baltic collections, and Hungarian and Slovak repertoire data, we are importing into the current datasets a small subset of Folk Tales Inventory [Népmeseleltár], in

short: *FTI Example Dataset*, and the *A Székelyföld népi tánczene szempontú táji tagolódása – 500 hangzó példa*, in short, *Szekler dance music dataset* created by Pávai, István to demonstrate on a small scale the value, usability and problems of the broader collections of HH.

These datasets, as much as it is possible, will be converted and added to the Finno-Ugric Data Sharing Space (FUDSS), a non-profit, research data sharing space that connects various Finno-Ugric immaterial and material culture in formats that are interoperable with libraries, archives, museums, Europeana, Daria-H, and the European Culture Heritage Cloud.

FUDSS only stores metadata and does not take away any published HH material; it will link back to the *Szekler dance music dataset*. It will expose a small subset, probably ~50 tales with connected bibliographical entries from *FTI Example Dataset*, which is less than 0.5% of the contents of the original data source.

The aim of this direction is:

- to evaluate to what extent the current data models developed by Reprex with the *Latvian Folk Archive* and other similar organisations can be directly applicable to HH;
- to gain hands-on graph building, reviewing, editing experience to the HH team;
- to expose various professional questions in the entire data curation workflow (technical, semantic, scalability or legal)

At the same time, HH colleagues will write competency questions to the expected graph, which we hope to be able to answer by querying the FUDSS in SPARQL. The competency questions should be organised like this.

### A Székelyföld népi tánczene szempontú táji tagolódása – 500 hangzó példa

Competences expected from importing to the FUDSS graph:

- Which example items were in use in `Gyergyóremete`?
- Which example items were collected by `Pávai, István`?
- What musical instruments appear in Pávai's examples?
- Which instrumental ensembles appear among Pávai's examples?
- In which example does a clarinet appear?
- In which folk region is the dance called korcsos found?
- What cultural micro-regions belong to the Marosszék cultural (meso)region?
- From which Bartók collections (location and date of collection) did István Pávai select his examples?

**After enriching the graph with Linked Open Data**:

- Which collector worked in `Remetea` [with the addition of multi-language gazetteer], for methodology see *Remapping the Livonian Coast: A Multilingual Gazetteer of the Settlements of Northern Kurzeme* (Antal, Pigozne, and Mester 2025).
- Which example items were collected within 25 kms from `46°15'8.96"N, 25°26'5.86"E`. We enriched Abásfalva-Aldea with geographical coordianates and Geonames PID (686517); a simple distance function can be built in a query that answers this question and lists other collection locations that fall into the radius.
- Which publications of the collector Agócs, Gergely are available in OSZK and other libraries?
- Which settlements in the collection have a proportion of Hungarian population below 50% according to the 2021 Romanian census?

**After feasibility accomplished**:

- Which publications of the collector `Bartók, Béla` are available in library of HH
- Which original audio tapes on inventory are connected to the collector `Pávai, István`
- Which collection can be listened to in its entirety in the Folklore Database (*Folklóra-datbázis*)?
- Which original media can be listened to in its entirety in the ZTI Sound Archive?

### Folk Tales Inventory [Népmeseleltár]

We placed about 120 folk tales from Transylvania into our database. These entries are not the abstract tales, but their textual manifestations as they were described in Konsza, Samu: Háromszéki magyar népköltészet [Folk Poetry of Háromszék] and made available originally by the *State Publishing House for Literature and Arts / Állami Irodalmi és Művészeti Kiadó*, and currently by Adatbank (fudss:Q5823)].

We used the following modelling solutions:

1. We placed **Competences expected from importing to the FUDSS graph**:

- Which tales were published in Magyar Nyelvőr in the year 1888?
- Which tales were collected in Mohács?
- Which tales were collected by Ortutay, Gyula?

**After enriching the graph with Linked Open Data**:

- Which collector worked in [requires addition of multi-language gazetteer]
- Which example items were collected within 25 kms from 48°16'36"N, 22°48'28"E
- Which publications of the collector [Ortutay, Gyula] are available in OSZK?
- What did Samu Konsza look like?

- Where was Samu Konsza born and where did he die?

**After feasibility accomplished**:

- Which publications of the collector [Ortutay, Gyula] are available in library the of HH?
- Which original fieldnote documents are connected to the collector Ortutay, Gyula?

### AtoM integration and round-trip validation

*Hagyományok Háza* (HH) aims to migrate its archival descriptions in the Access to Memory system (AtoM). To ensure that the proposed semantic-web workflow is compatible with the planned archival practices, the pilot includes a round-trip interoperability test between AtoM and the *Finno-Ugric Data Sharing Space* (FUDSS).

In this phase, a small number of AtoM records will be exported in standard archival formats (EAD and EAC-CPF) and imported into the FUDSS Wikibase environment. These items will be modelled using the same pragmatic linked-data patterns applied to the *FTI Example Dataset* and the *Szekler dance music dataset*. After modelling and basic semantic enrichment, the records will be exported back to AtoM-compatible representations.

This round-trip test allows us to evaluate whether entity identifiers, personal and geographical names, roles, provenance fields, and relationships survive the full cycle of import, modelling, enrichment, and re-export without structural loss or semantic distortion. Success will be demonstrated if a small number of enriched items can be re-ingested into AtoM and remain usable within its archival description workflows.

The National Library of Wales follows a similar model in its *Shared National Archival Record Catalogue* (SNARC), where Wikibase acts as a reconciling and multilingual authority layer while traditional archival systems continue to provide the core descriptive functions. Our test will assess the feasibility of adopting this architectural pattern at *Hagyományok Háza*.

## Modelling pragmatism

Our solution is the **Finno-Ugric Data Sharing Space**, a lightweight, federated infrastructure that supports participatory metadata repair, semantic enrichment, and multilingual modelling. Technically, our infrastructure connects CIDOC CRM-based museum records, DCTERMS-based library metadata, and Wikibase lexemes through shared patterns, allowing heterogeneous sources to interoperate without enforcing a single monolithic schema.

### Authority control

Robust authority control is essential for maintaining data quality and semantic precision in a multilingual knowledge graph. It ensures that persons, organisations, and places are uniquely identified across systems and over time.

**Personal and Corporate Names**
All creators, collectors, and institutions must be associated with permanent, globally recognised identifiers. We rely on:

- **VIAF** – international library authority file (including OSZK).
- **ORCID** – the preferred identifier for researchers.
- **ISNI** – ISO-standard name registry.
- **Wikidata QIDs** – decentralised identifiers linking VIAF, ORCID, ISNI.
- **Magyar Nemzeti Névtér** – currently inactive but conceptually aligned.

VIAF and ORCID offer strong interoperability with ISNI and Wikidata. For historical bibliographic entities, VIAF is generally most reliable; for living contributors in research workflows, ORCID is preferred. ISNI functions as a high-quality fallback, while Wikidata provides a fully open alternative with community-based validation. Authority files also preserve spelling variants and legacy forms (e.g., *Ferencz*), supporting historical continuity.

Our work draws on established approaches to multilingual authority control (Bianchini, Bargioni, and Pellizzari di San Girolamo 2021; Fagerving 2023). Comparative models—from the *National Library of Wales* and Wikimedia UK (Evans 2025), through the semi-institutionalised Võro User Group (Wikimedians of Võro language User Group 2025), to the fully community-driven Sámi project in Norway (Wikimedia Norge 2020)—demonstrate how multilingual reconciliation can be embedded in institutional, semi-formal, or community-led infrastructures.

In our project, collectors, authors, and—where legally permissible—data subjects were reconciled with relevant authority files. This yielded two primary benefits:

- stronger interoperability with library and research systems;
- identification and correction of inconsistent or duplicate metadata.

> **i** Disambiguation and repair
>
> **Dozdovszky, Zoltán** (fudss:Q6370) illustrates this issue. His name appeared in multiple inconsistent spellings in the source dataset. Authority reconciliation consolidated these variants into a single identity.

Graph-based linking thus enables metadata repair that would be cumbersome in relational databases, where fuzzy matching must be used without authoritative grounding.

### Geographical Gazetteer

Place names show even greater variability than personal names, especially in multi-ethnic regions. To manage this, we constructed a multilingual **geographic reconciliation layer** (gazetteer) to align historical and modern toponyms.

Our initial focus was the *FTI Example Dataset*, covering part of Székelyföld. For example, **Gyergyóremete** (fudss:Q5730) has multiple culturally relevant appellations:

**Gyergyóremete** (fudss::Q5730)
`has label` (en) → Remetea, alias Gyergyoremete
`has label` (ro) → Remetea, alias Gyergyoremete
`has label` (hu) → Gyergyóremete, alias Remetea
`coordinate location` → 46°47'36.71"N, 25°27'1.01"E
`Geonames ID` https://www.geonames.org/668986

We use the current official name (*Remetea*) for English-language systems, while retaining historical and culturally meaningful variants for multilingual discovery. ASCII-normalised aliases ensure compatibility with search interfaces lacking diacritics. Wikibase's UI handles all these variants gracefully through its label–alias lookup.

We treat *Gyergyóremete* as a **place appellation**, reflecting a cultural concept rather than an administrative boundary. Heritage materials relate to this conceptual locality even when municipal borders shift.

For broader regions such as *Háromszék*, midpoint coordinates are insufficient. Boundary polygons (e.g., from OpenStreetMap) or explicit hierarchical relations are required to support reasoning about inclusion.

Our workflow follows a **bottom-up** strategy: starting from raw tabular values and incrementally introducing higher-level spatial hierarchies. This mirrors our **Livonian Coast Gazetteer**, derived from *The Livonian Place Name Catalogue* (Ernštreits, Šuvcāne, and Damberg 2024; Antal, Pigozne, and Mester 2025).

> **OpenRefine**
>
> All reconciliation steps are documented using *OpenRefine* and integrated into our tutorial book.

### Thesaurus and Ontology Patterns

Parallel to geographic reconciliation, we translated and graph-integrated a subset of *HH's internal thesaurus*, aligning it pragmatically with the *Getty Art & Architecture Thesaurus® (AAT)* (Harpring et al. 2020; Silva 2022). We used a label-focused approach, retaining AAT's upper-level conceptual structure while introducing local subclasses where appropriate.

This balances broad discoverability with ontological precision. General terms (e.g., *lakodalom*) serve as human-friendly entry points, while more specific concepts emerge through iterative refinement. This follows the principle of *ontology seeding* (Maria Teresa, Stellato, and Vindigni 2012; Maria Teresa and Stellato 2012), combining automated extraction with expert correction.

Our approach aligns with the Semantic Web "layer cake" model (Mayank, Craig A., and Pedro 2021), where vocabularies like AAT and CIDOC CRM occupy the semantic middle layer. CIDOC CRM itself functions more as a modelling dialect than as a rigid ontology (Bekiari et al. 2024), and archival metadata adds additional complexity as institutions transition from legacy standards, such as ISAD(G), to *Records in Contexts* (RiC) (Archives Expert Group on Archival Description 2023).

To address this heterogeneity, we emphasise modular ontological patterns rather than comprehensive global alignment. Drawing on eXtreme Design and recent work in the Polifonia Ontology Family (Blomqvist, Hammar, and Presutti 2016; Berardinis et al. 2023), we extract reusable fragments that support low-friction cross-domain linking. Rather than "ontological hijacking," we formalise recurring vernacular patterns found in AAT, Wikidata, and community vocabularies alike.

This positions our work at the interface between general-purpose ontologies and community-specific modelling. Our aim is not to replace established standards but to complement them—enabling smaller, underrepresented knowledge systems to participate in linked-data infrastructures without flattening epistemic diversity. Governance, multilingual support, and contextual sensitivity are built into this architecture from the outset.

### Folk tales

We mapped the *Meseleltár* to a text-manifestation class. The `chapter` class is well established in museum, library, and web ontologies (`crm:E31_Document`, `frbroo:F24_Publication_Expression`, schema:Chapter):

`chapter`
  book section (unnumbered) (fudss::Q5818)
  folk tale textual manifestation (fudss:Q5819)

Example:
**The Wolf, the Bear and the Hussar (folk tale manifestation)** → https://reprexbase.eu/fu/Item:Q5820

The corresponding abstract work is modelled as a subclass of literary work without individual authorship. Later iterations will connect abstract tales to international taxonomies (e.g., ATU types), enabling linkage of textual manifestations, audio recordings, and fieldwork variants.

`The Wolf, the Bear and the Hussar (folk tale)` ([fudss::Q5820])
  `has taxonomical entry` → ATU type
  `has manifestation` → folk tale text manifestation ([fudss::Q5826])

`is curated member of` → Meseleltár
`published in` → Háromszéki magyar népköltészet ([fudss:Q5815])
`is heritage of` → Székelys

The property `is heritage of` derives from the *Heritage Digital Twin Ontology* (HDTO), supporting connections to legal and policy frameworks such as Hungaricum registries or national cultural-content classifications.

`Meseleltár` ([fudss:Q5825])
`instance of` → collection

The `is curated member of` relationship provides transparent curatorial provenance. New tales can be added; erroneous entries can be corrected or removed. The publication node (`Háromszéki magyar népköltészet`) is treated as a standard book object and can be linked to external library systems.

For usability, we attach `has access point` directly both to the publication node and to individual tale manifestations so that Sampo-UI displays a direct download link for each item, even though this introduces a small amount of formal redundancy.

## Ethical and legal considerations

Alongside FAIR and the 8-star model, we also considered the **CARE** Principles for Indigenous Data Governance (Collective Benefit, Authority to Control, Responsibility, Ethics) (Carroll et al. 2020). While these principles were formulated in Indigenous contexts such as North America and Oceania, they offer relevant guidance for Finno-Ugric heritage. The situation of these communities is heterogeneous: Hungarians, Finns, and Estonians are nation-states with strong institutional infrastructures; Seto, Võro, Csángó or Szekler communities are integrated minorities with various level of community rights, formal and informal organisations, and ability to use their languages. Our DSS adapts CARE to this environment by prioritizing multilingual metadata accessibility, cultural reassembly, and safe avenues for community review, while acknowledging that not all CARE provisions can be applied in the same way.

Rights management within the FinFAIR framework is designed to be **transparent, interoperable, and machine-readable**.
It combines the Heritage Digital Twin Ontology (HDTO) with practical extensions in Wikibase that allow curators to express not only licences but also contextual usage rules derived from the *Open Digital Rights Language (ODRL)*.

While large digital-heritage platforms such as Europeana or DARIAH rely mainly on single-URI licence models (`edm:rights`, `dcterms:license`), FinFAIR adds an optional layer of structured **use policies** to handle complex, multimodal works such as music or film, where overlapping rights coexist.

### Personal Data and GDPR

Every `hdt:HC1_HeritageEntity` or `hdt:HC2_HeritageDigitalTwin` may link to living agents—performers, collectors, researchers, donors, or other data subjects—through `dcterms:creator`, `prov:wasAttributedTo`, or `crm:P14_carried_out_by`.

Under the **General Data Protection Regulation (GDPR)**, these names constitute *personal data*.

Before publication or linkage, consent and attribution are evaluated:

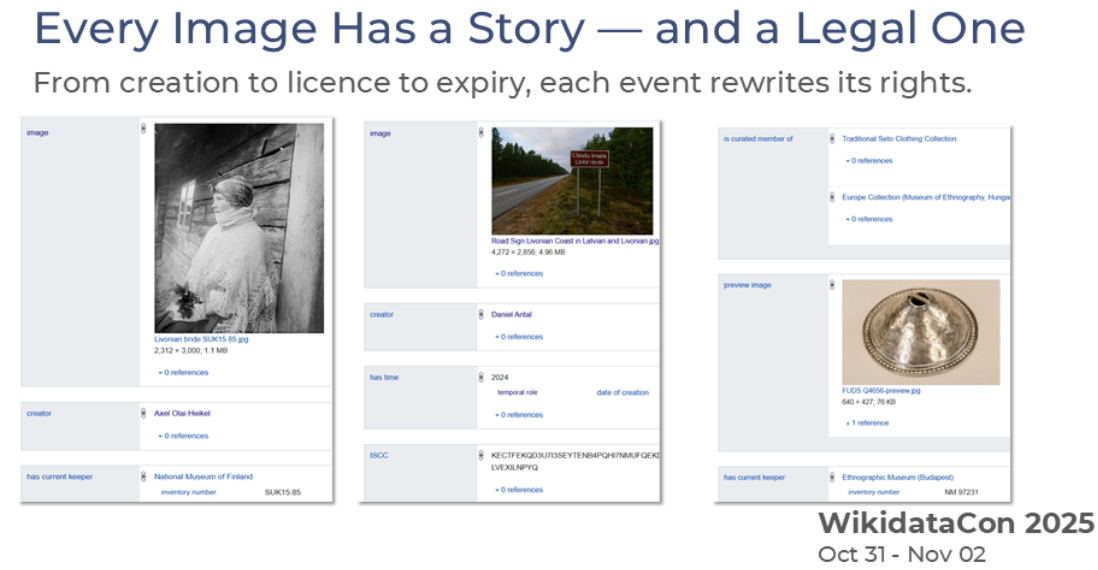
- **Attribution** — if the person consents to credit, we use `cc:attributionName`.
- **Pseudonymisation** — if consent is uncertain, we use neutral labels, e.g. "Field recording performer, 1962".
- **Anonymity** — if the person requests privacy, all identifying information is withheld.
- The selected attribution mode determines the applicable licence.

For example, `CC-BY` assumes credit, while pseudonymised or anonymous materials default to `CC0`, `BY-NC`, or a corresponding `rightsstatements.org` URI.

For older archival materials, FinFAIR follows Europeana's approach: if the data subjects are certainly deceased (typically pre-1925), personal-data restrictions no longer apply.

### Simple Case: One Licence per File

For non-composite objects—photographs, scans, or plain-text documents—the **Europeana Data Model (EDM)** `edm:rights` field provides sufficient legal clarity.
FinFAIR implements this directly via the Wikibase property **`use rights`**, which corresponds to `dcterms:license` and `edm:rights`.




Even these "simple" cases are dynamic. A copyright protection term may expire, making a less accessible item fully reusable under the public domain. Rightsholders may change their minds, and give a more permissive license, for example, change from `CC BY-NC-ND` to `CC-BY-SA`, allowing commercial use and derivate works.

| Type of object | HDTO class | Typical examples |
|---|---|---|
| Digital text | `hdt:HC6_DigitalDocument` | metadata exports, lyrics, PDFs |
| Digital image | `hdt:HC7_DigitalVisualObject` | photographs, scans, blueprints |
| Simple audio file | `hdt:HC5_DigitalRepresentation` | single, unambiguous recordings |

Each digital component carries:

- `use rights` → a URI pointing to a standard licence (e.g. CC-BY, CC0);
- optional provenance (`dcterms:rightsHolder`, `dcterms:accessRights`, `dcterms:provenance`).

The composite digital twin (`HC2`) aggregates multiple components but does not override their licences.
When reused, the **most restrictive** licence among components applies.

### Wikibase Rights Extensions

To align HDTO, EDM, and Wikibase, FinFAIR defines two complementary properties and one supporting class.

> **!** Important
>
> All examples below will be changed this week to HH examples for discussion.

**Property: `use rights` (`wd:P486`)**

| Field | Description |
|---|---|
| **Label** | use rights |
| **Description** | Links a digital object to a URI identifying the applicable licence or rights statement (e.g. rightsstatements.org, Creative Commons). |
| **Aliases** | licence; rights; edm:rights; dcmi:license; datacite:rights |
| **Datatype** | External identifier (URI) |
| **Equivalent properties** | `dcterms:license`, `edm:rights`, `datacite:rights` |
| **Full definition** | https://reprexbase.eu/fu/Special:EntityData/P486.ttl |

**Example:**

- Dēliņi homestead detail: combined barn-shed (detail photo)
  `use rights`: https://creativecommons.org/licenses/by-sa/4.0/deed.en
- The Livonian Community House in Mazirbe (photograph)
  `use rights`: https://creativecommons.org/licenses/by/4.0/deed.en

This property provides one-to-one interoperability with both Europeana and research repositories using DataCite

**Property: `use policy` (`wd:P487`)**

| Field | Description |
|---|---|
| **Label** | use policy |
| **Description** | Links a digital object to a *local* rights or usage policy item within the same Wikibase instance (e.g. "Listen on Spotify"). Used when the applicable terms are described internally rather than via an external URI. |
| **Aliases** | has policy; user rights policy |
| **Datatype** | Wikibase item |
| **(Near) Equivalent class** | `odrl:Policy` |
| **Full definition** | https://reprexbase.eu/fu/Special:EntityData/P487.ttl |

### Class: `use policy` (`wd:Q5606`)

| Field | Description |
|---|---|
| **Label** | use policy |
| **Description** | A simplified representation of `odrl:Policy` within Wikibase. Describes how a digital object may be accessed, used, or shared under specific conditions. Instances act as human-readable policy nodes. |
| **Full definition** | https://reprexbase.eu/fu/Special:EntityData/Q5606.ttl |

**Example instance of this class:**

`Q5608 - Listen on Spotify`
*You may listen to this recording via the Spotify web or mobile player, either for free or with a subscription, under Spotify's End-User Licence.*

This simplified approach to ODRL was first introduced in our *WikidataCon 2025* presentation, where we demonstrated how Wikibase can represent policy semantics ("play", "reproduce", "not download") without external RDF frameworks.

#### Complex Case: Composite Rights

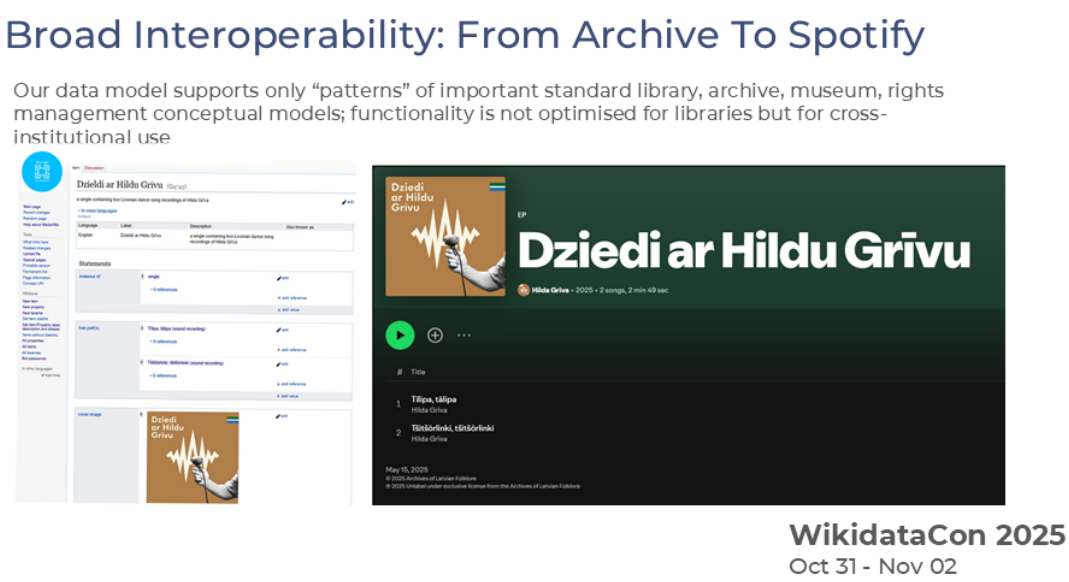
Multimodal cultural objects, such as sound recordings, often involve overlapping rights: composer, lyricist, performer, producer, and distributor.
In such cases, each digital representation (`HC5`) carries its own `use rights` and `use policy`.


This example is taken from *Wikibase as a Data Sharing Space: Connecting Rights, Communities, and GLAM through Federated Infrastructures* (Antal 2025).

Example: *Tšitšõrlinki, tšitšõrlinki*

- **Archival recording** — accessible under `CC-BY-NC`, for non-commercial research. `use policy`: "Non-commercial analysis and reproduction permitted."
- **Spotify release** — available for streaming only under Spotify's end-user licence. `use policy`: "Listen on Spotify."

These are linked within a single digital twin (`HC2`), which aggregates but does not merge or override rights.




Aggregators such as Europeana can display the most restrictive licence, while research or educational users can act according to the specified `use policy`.

### Summary

1. **GDPR compliance** — screen linked agents, pseudonymise or anonymise where required.
2. **Simple case** — one `use rights` URI per file, equivalent to `edm:rights`.
3. **Composite case** — multiple components, each with its own `use rights` and `use policy`.
4. **Interoperability** — full equivalence with EDM, DCTERMS, and ODRL.
5. **Scalability** — lightweight for images and texts, yet expressive for music and audiovisual works.

This model provides a sustainable bridge between **cultural heritage and research data ecosystems**, ensuring that digital twins can circulate legally and ethically within the European Collaborative Cloud for Cultural Heritage.

### References

Antal, Daniel. 2025. "Wikibase as a Data Sharing Space: Connecting Rights, Communities, and GLAM Through Federated Infrastructures. Presentation on Wikidata Conf 2025." Open Music Observatory. https://doi.org/10.5281/zenodo.17496740.

Antal, Daniel, Ieva Pigozne, and Anna Márta Mester. 2025. "Remapping the Livonian Coast: A Multilingual Gazetteer of the Settlements of Northern Kurzeme." Finno-Ugric Data Sharing Space. https://doi.org/10.5281/zenodo.15668712.

Archives Expert Group on Archival Description, International Council on. 2023. "Records in Contexts–Conceptual Model. Version 1.0." International Council on Archives. https://www.ica.org/app/uploads/2023/12/RiC-CM-1.0.pdf.

Bekiari, Chryssoula, George Bruseke, Erin Canning, Martin Doerr, Philippe Michon, Christian-Emil Ore, Stephen Stead, and Velios Athanasios, eds. 2024. "Definition of the CIDOC Conceptual Reference Model." CIDOC CRM Special Interest Group. https://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_7.2.4.pdf.

Berardinis, Jacopo de, Valentina Anita Carriero, Nitisha Jain, Nicolas Lazzari, Albert Meroño-Peñuela, Andrea Poltronieri, and Valentina Presutti. 2023. "The Polifonia Ontology Network: Building a Semantic Backbone for Musical Heritage." In *The Semantic Web – ISWC 2023*, edited by Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li, 302–22. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47243-5_17.

Bianchini, Carlo, Stefano Bargioni, and Camillo Carlo Pellizzari di San Girolamo. 2021. "Beyond VIAF Wikidata as a Complementary Tool for Authority Control in Libraries." *Information Technology and Libraries* 40 (2). https://doi.org/10.6017/ital.v40i2.12959.

Blomqvist, Eva, Karl Hammar, and Valentina Presutti. 2016. "Engineering Ontologies with Patterns – the eXtreme Design Methodology." In *Ontology Engineering with Ontology Design Patterns*, 23–50. IOS Press. https://doi.org/10.3233/978-1-61499-676-7-23.

Carroll, Stephanie R., Ibrahim Garba, Olga L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Sally Materechera, Mark Parsons, et al. 2020. "The CARE Principles for Indigenous Data Governance." *Data Science Journal* 19 (43): 1–12. https://doi.org/10.5334/dsj-2020-043.

Commission, European, and Directorate-General for Digital Services. 2017. *New European Interoperability Framework. Promoting Seamless Services and Data Flow for European Public Administrations*. Luxembourg: Publications Office of the European Union. https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf.

Commission, European, Directorate-General for Research, Innovation, P. Brunet, L. De Luca, E. Hyvönen, A. Joffres, et al. 2022. *Report on a European Collaborative Cloud for Cultural Heritage – Ex – Ante Impact Assessment*. Publications Office of the European Union. https://doi.org/doi/10.2777/64014.

Core, Dublin. 2020. "DCMI Metadata Terms." http://dublincore.org/specifications/dublin-core/dcmi-terms/2020-01-20/.

ECHOES Ontology Task Force. 2025. "Heritage Digital Twin Ontology (HDTO) – First Draft." Technical Report. ECHOES Project / European Collaborative Cloud for Cultural Heritage (ECCCH). https://github.com/ECHOES-ECCCH/HDTO-Heritage-Digital-Twin-Ontology.

Ernštreits, Valts, Baiba Šuvcāne, and Pētõr Damberg. 2024. *Lībiešu Vietvārdu Katalogs = the Livonian Place Name Catalogue*. 1st ed. Rīga: Latvijas Universitātes Lībiešu institūts.

European Commission. 2017. "European Interoperability Framework – Implementation Strategy. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions." European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52017DC0134.

Evans, Jason. 2025. "A Decade of Collaboration Between Wikimedia UK and the National Library of Wales: Building a Sustainable Future for Welsh Culture Online." Wikimedia Diff blog. https://diff.wikimedia.org/2025/01/27/a-decade-of-collaboration-between-wikimedia-uk-and-the-national-library-of-wales-building-a-sustainable-future-for-welsh-culture-online/.

Fagerving, Alicia. 2023. "Wikidata for Authority Control: Sharing Museum Knowledge with the World." *Digital Humanities in the Nordic and Baltic Countries Publications* 5 (1): 222–39. https://doi.org/10.5617/dhnbpub.10665.

Group, DataCite Metadata Working. 2024. *DataCite Metadata Schema for the Publication and Citation of Research Data and Other Research Outputs: Version 4.5*. DataCite e.V. https://doi.org/10.14454/g8e5-6293.

Harpring, Patricia, Robin Johnson, Jon Ward, and Antonio Beecroft. 2020. "Getty Vocabulary Program: Update, ITWG Meeting, Getty Vocabulary Program." Los Angeles: United States of America. https://www.getty.edu/research/tools/vocabularies/000_vocab_updates_itwg2020.pdf.

Hyvönen, Eero, and Jouni Tuominen. 2024. "8-Star Linked Open Data Model: Extending the 5-Star Model for Better Reuse, Quality, and Trust of Data." In *Posters, Demos, Workshops, and Tutorials of the 20th International Conference on Semantic Systems (SEMANTiCS 2024)*, edited by Daniel Garijo, Alessandra Lo Gentile, Ana Kurteva, Andrea Mannocci, Francesco Osborne, and Sahar Vahdati. Vol. 3759. CEUR Workshop Proceedings. Aachen: CEUR-WS.org. http://hdl.handle.net/10138/586661.

Maria Teresa, Pazienza, and Armando Stellato. 2012. *Semi-Automatic Ontology Development : Processes and Resources.* Information Science Reference. https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=440669&site=ehost-live&scope=site.

Maria Teresa, Pazienza, Armando Stellato, and Marco Vindigni. 2012. "A Modular Framework to Learn Ontologies from Domain Specific Texts." In *Semi-Automatic Ontology Development: Processes and Resources*, edited by Pazienza Maria Teresa and Armando Stellato, 1–29. Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-4666-1809-1.ch001.

Mayank, Kejriwal, Knoblock Craig A., and Szekely Pedro. 2021. *Knowledge Graphs : Fundamentals, Techniques, and Applications.* Adaptive Computation and Machine Learning Series. The MIT Press. https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=2517962&site=ehost-live&scope=site.

National Library of Finland. 2021. "Finnish Wikibase Pilot: Connecting Museum, Linguistic, and Archival Records Through Wikibase." Project Report. National Library of Finland.

Silva, Camila da. 2022. "The Ongoing Translation of the Getty Art & Architecture Thesaurus® into Portuguese: An Art Information Access and Retrieval Tool for Cultural Institutions in Portuguese-Speaking Countries." *Getty Research Journal* 16 (16): 209–25. https://doi.org/10.1086/721991.

Wikimedia Norge. 2020. "Samisk Kunnskap På Nett: Wikimedia Norge's Sámi Project 2017–2020." https://no.wikimedia.org/wiki/Prosjekt:Samisk_kunnskap_p%C3%A5_nett/2017%E2%80%932020/en.

Wikimedians of Võro language User Group. 2025. "Wikimedians of võro Language User Group." Meta-Wiki affiliate page. https://meta.wikimedia.org/wiki/Wikimedians_of_V%C3%B5ro_language_User_Group.