



BIBFRAME Workshop in Europe 2026

Linked Library Data transformation pipeline using local LLM to support feature vector methods in work clustering and authority record reconciliation

01 Introduction

Who am I, who are we, why are we doing this?



Mikko Vihonen

Senior Solution Architect

- Design and implementation of digital business solutions
- Solution and enterprise architecture
- Background in digital solutions consulting mobile phone and networking, banking and retail business customers

- Enterprise and Solution architecture design for Helsinki Metropolitan area (HELMET) public libraries
 - OPAC renewal
 - Technical consulting (requirements definition and vendor negotiation) for Library system tendering competition
 - Implementation of low cost Shelfmark printing solution to City of Helsinki and Annif-enabled classification system change
 - Data migration and library system implementation technical ownership
- eKirjasto (Library Simplified) evaluation and enterprise architecture as Nationwide e-library
- Pro bono implementation of BIBFRAME to BFFI (Finnish BIBFRAME adaptation) conversion for National Library of Finland



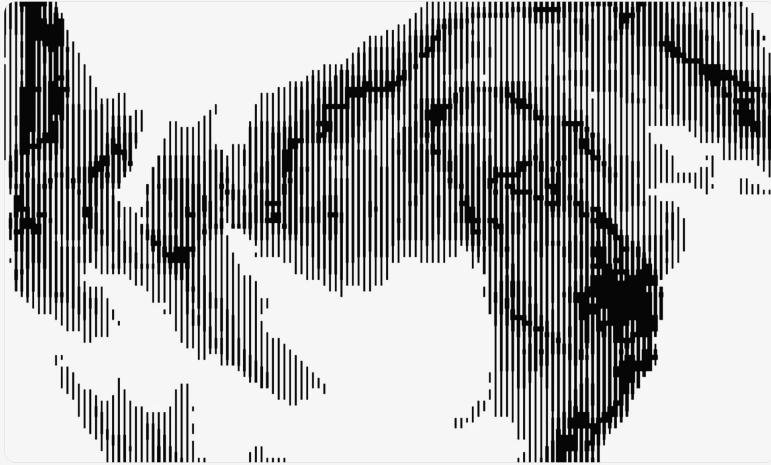


Hello, we are Nitor

Turn your ambition
into lasting, high-impact
solutions for the AI era.

Services

Engineering digital success
for the AI era



Think

Manifest your ambitions into winning digital strategies, CX and AI foresight

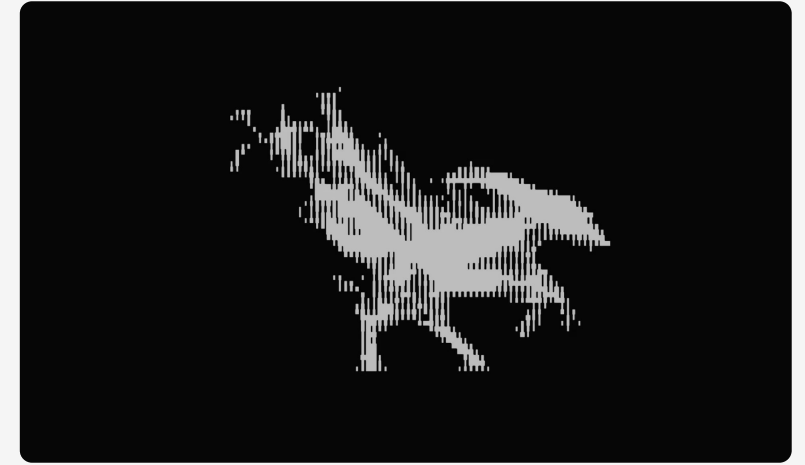
- DIGITAL STRATEGIES AND WINNING CX
- AI FORESIGHT AND TRANSFORMATION
- TECHNOLOGY AND DATA ADVISORY



Build

Hands-on execution to value: data, commerce & AI solutions

- AI-EMBEDDED DIGITAL PRODUCTS
- DIGITAL AND AGENTIC COMMERCE
- DATA AND AI SOLUTIONS



Run

Optimise to scale, operate and run your game 24/7

- PLATFORM AND CLOUD ENGINEERING
- AI OBSERVABILITY AND OPERATIONS
- 24/7 RELIABILITY AND SECURITY



Premium engineering in every touchpoint

300+

Strategists, designers, AI
engineers, data scientists

40M+

Net sales in 2025

Think
Build
Run

AI-era digital success

1.

Best customer satisfaction
2012-2026 (Onway)



02 Motivation

Implementing format
conversion to be able to assist
others implementing it



Motivation

Why did I implement this as a pro bono project?

- Helsinki Metropolitan (HELMET) public libraries have put considerable emphasis to BIBFRAME (BFFI) support in their tendering competition requirements. It is a must win battle.
- Vendors require reference implementation and real-world examples of the bibliographic records so that they can verify round-trip support of their solutions



03 First try

Implement full conversion,
clustering and reconciliation
pipeline to gain insights



Data at hand

How to map the sparse and sometimes typo-riddled metadata of HELMET libraries into the national subject and value vocabularies and agent authority records?

How to determine if the bibliographic records are instances of the same works (or, in case of BFFI, expressions and manifestations of the same works)?

Dataset	Source	Challenges
Bibliographic data	HELMET records	Sedimentation across multiple MARC-based system migrations. Sparse data.
Subject and value vocabularies	YSO, KAUNO, MUSO	No URI references in bibliographic records, use of outdated sets
Nationwide authority records	ASTERI, KANTO	No URI references in bibliographic records
International authority records	VIAF	Very few references in the bibliographic records
Reconciliation support dataset	Wikidata	Agent information doesn't necessarily overlap with those of HELMET records



Methods considered

Deterministic rules were easy to set up and seemed to solve most problems with ontologies and value vocabularies

If I had had more time and resources to spend on training, I would have created training dataset also leveraging capabilities of NLF automatic indexing tool Annif, but unfortunately I didn't have much time to spend for this pro-bono project.

Embedding models and using LLM judge had potential I wished to explore while implementing the reference data transformation pipeline.

	Method	Role / Strength	Practical Limitation in My Context
Known	Deterministic Rules (Levenshtein, Blocking)	Fast, exact, 100% explainable	Brittle against typos, variations, and homonyms
Annif territory	Unsupervised ML (K-Means, TF-IDF)	Groups data without manual labels	Lacks domain reasoning & format-to-role context
	Supervised Classifiers (Random Forests, GCNs)	High precision on learned features	Requires large labeled training datasets
Unknown potential	Neural Embeddings (BGE-M3 + FAISS HNSW)	Microsecond multilingual recall across millions	Captures semantic similarity, but cannot perform multi-step deduction
	LLM Judge (Zero-Shot Backstop)	Deep world knowledge & domain reasoning	High per-query compute



Clustering and reconciliation methods for agent authority record linking

Clustering Aggregates Context

- Grouping records creates a rich composite profile before linking.

Reconciliation Anchors Clusters

- Resolving the composite profile links the entire cluster to a canonical URI at once

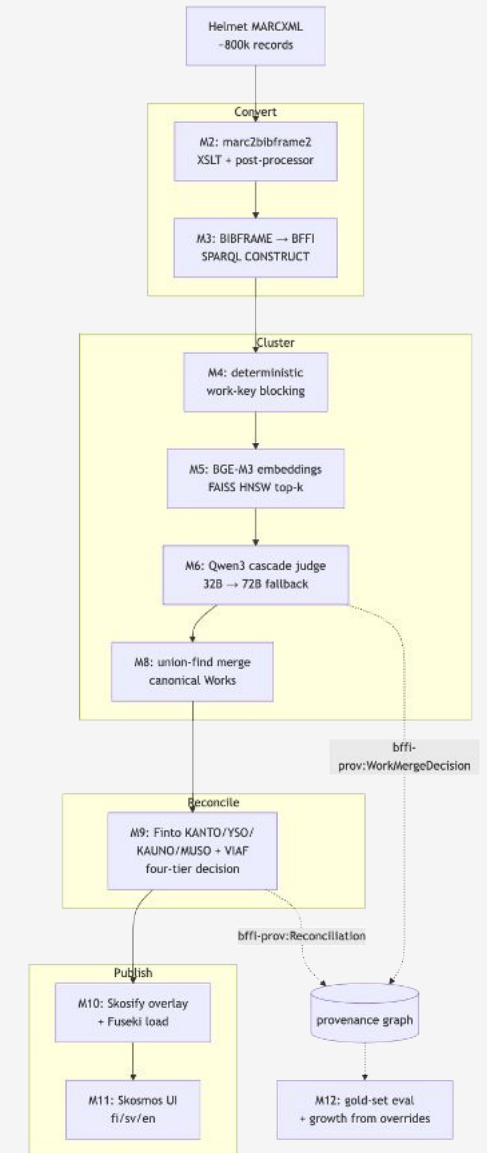
Dimension	Phase 1: Clustering Entity Grouping	Phase 2: Reconciliation Entity Linking
Core Question	"Which unlinked mentions refer to the same real-world entity?"	"Which canonical authority record (KANTO, VIAF, ISNI, Wikidata) does this match?"
Input - Output	Raw catalog records to Disjoint entity clusters	Clustered records / mention to Persistent URI
Ground Truth	Internal consistency is key, no external target needed	External standard, requires target authority file
Primary Method	Vector similarity (BGE-M3)	Target candidate retrieval (FAISS HNSW) + LLM Judge



First attempt: clustering and reconciliation as a part of conversion pipeline

- Convert everything to target format first
- Cluster everything
- Reconcile everything
- Add Skosmos overlay and publish to data store
- Use Skosmos and TSV report as reviewer surface to data

- Experiment was run on a 16" Macbook Pro M2 Max 64GB RAM
- Emphasis was put on clustering works and using the works as the context for subject and value vocabulary, and agent authority record link resolution
- This approach worked with smaller runs, but had trouble scaling to 800k records due to too many requests hitting the LLM cascade judge and the system running out of memory
- It did however show the extent a completely local can help in clustering and reconciliation



Lessons learned

- Conversion to BIBFRAME with marc2bibframe lost precious local context
- Especially the lost local classification fields made clustering and reconciliation more difficult
- Best to apply corrections when MARC 21 context is present
- LLMs are expensive, but bridge the gap when agent authority record features and those on the bibliographic records have not been mapped to each other

Case Study 1: Role-to-Format Plausibility (Zero-Shot Reasoning)

Input: Ambiguous name matching two distinct authority IDs:
Physician vs. **Opera Singer**.

Context: Format = *Audio Recording / Performance*.

LLM Decision: Reasoned zero-shot that an opera singer is substantially more plausible to be credited on a musical recording than a physician—bypassing the need to train a schema classifier.

Case Study 2: Parametric Knowledge & Catalog Anomaly Detection

Input: Performer: Anssi Karttunen | Instrument: Harpsichord.

Vector Failure: Near-1.0 match to *Anssi Karttunen* (well-known Finnish musician).

LLM Anomaly Catch: Flagged the record typo: *Anssi Karttunen* is a **cellist**, whereas *Assi Karttunen* is the **harpsichordist**. Caught a legacy catalog typo that standard pipelines would have codified.



03 Where are we now?

Format conversion
implemented. Lessons learned
to be applied to the pipeline
next.



What's ready?

MARC21 BFFI conversion via BIBFRAME is being tested in the National Library of Finland

- We needed reference material to help Systematic, the HELMET library system vendor
- Conversion pipeline implemented as open source under MIT license
- <https://mikkovihonen.github.io/bffi-conversion-pipeline/>

The National Library of Finland is now testing the schema conversion pipeline which was implemented by AI coding agents on a feedback loop

Agents (Claude Code, Pi with local LLM) were tasked to independently find conversion routes from BIBFRAME into BFFI with a ruleset and find also conversion route back to MARC21 from BFFI

- **Claude Code:** Succeeded technically in the loop, but consumed \$1,000 of my complimentary tokens overnight
- **Pi Coding Agent + local Qwen 3.6 MoE:** Delivered comparable grounded iteration success locally at zero marginal API cost.



What's next?

Fix references to subject and value vocabularies into linked form in MARC21.

Attempt to link as many agent authority records as possible given the full context of MARC21

Corpus internal
Work/Expression level clustering and reconciliation after schema conversion

- This approach will directly benefit HELMET libraries' data migration from Sierra ILS into their new Cicero ILS.
- I believe it is also the approach that will bring best results to clustering and reconciliation of the data in the conversion pipeline reference implementation.



Conclusions

Key takeaways

Toolset used in development

- <https://mikkovihonen.github.io/pi-container/>

BFFI conversion pipeline

- <https://mikkovihonen.github.io/bffi-conversion-pipeline/>

- Even locally deployed large language models can help in classification tasks that would otherwise require considerable training
 - If there is time and resources available to train, the traditional models are more efficient than LLMs
 - Heuristics are mostly enough to reconcile ontologies and value vocabularies
- Agentic development using LLMs allows development of format conversions even without access to frontier models
 - A prerequisite for this is a feedback loop for the agent
- It would be of great help if marc2bibframe offered a means to convert also local use MARC21 fields into BIBFRAME
 - Maybe as bflc:marcKey literals?



Engineering digital success

PUBLIC

