Best Practices and Evaluation Criteria for Constructing Multilingual Structured Vocabularies for LGBTQ+ Libraries and Heritages

SHUAI WANG

ORCID: 0000-0002-1261-9930

Vrije Universiteit Amsterdam

Semantic Web in Libraries Conference 2025

OUTLINE

- Introduction
- Current Situation and Motivation
- **Best practices**
- **Evaluation Criteria: Indicators**
- **Evaluation Criteria: Self-assessment report template**

From the community, for the community

MULTILINGUAL LGBTQ+ CONCEPTS IN THE SEMANTIC WEB

YAGO

DBpedia

Homosaurus
English/Dutch
labels in around 55 languages

LCSH

Wikidata
507

GSSO
77

QLIT
Swedish

	Homosaurus	QLIT	GSSO	Wikidata
Purpose	LGBTQ heritage & libraries	Online literature database	Biomedical research + GLAM	Generic
Used with	LCSH	Svenska ämnesord and/or Barnämnesord	MeSH, SNOMED-CT, etc.	Too many :)
Languages*	English, Dutch (recently Spanish, etc.) → Around 55 languages	Swedish	Multilingual (77)	Multilingual (507+)
Type	Thesaurus	Thesaurus	Ontology	Knowledge base / Knowledge graph
Published as	linked open data	linked open data	linked open data	linked open data
Access	Website + API (SPARQL)	Website + APIs	GitHub + APIs	SPARQL, etc.
Maintenance	A new version every half a year. Now version 4!	Only one version but actively watched	Maintained 3 years ago	Actively updated by community members
Licensing	CC BY-NC-ND	CC0 1.0	CC BY-NC-ND	CC0 1.0
Documentation/resources	Yes but outdated (2022 and 2023)	https://queerlit.dh.gu.se/om/en/thesaurus/	Outdated readme, publication, wiki, etc.	Not pecifically about LGBTQ+-related concepts.
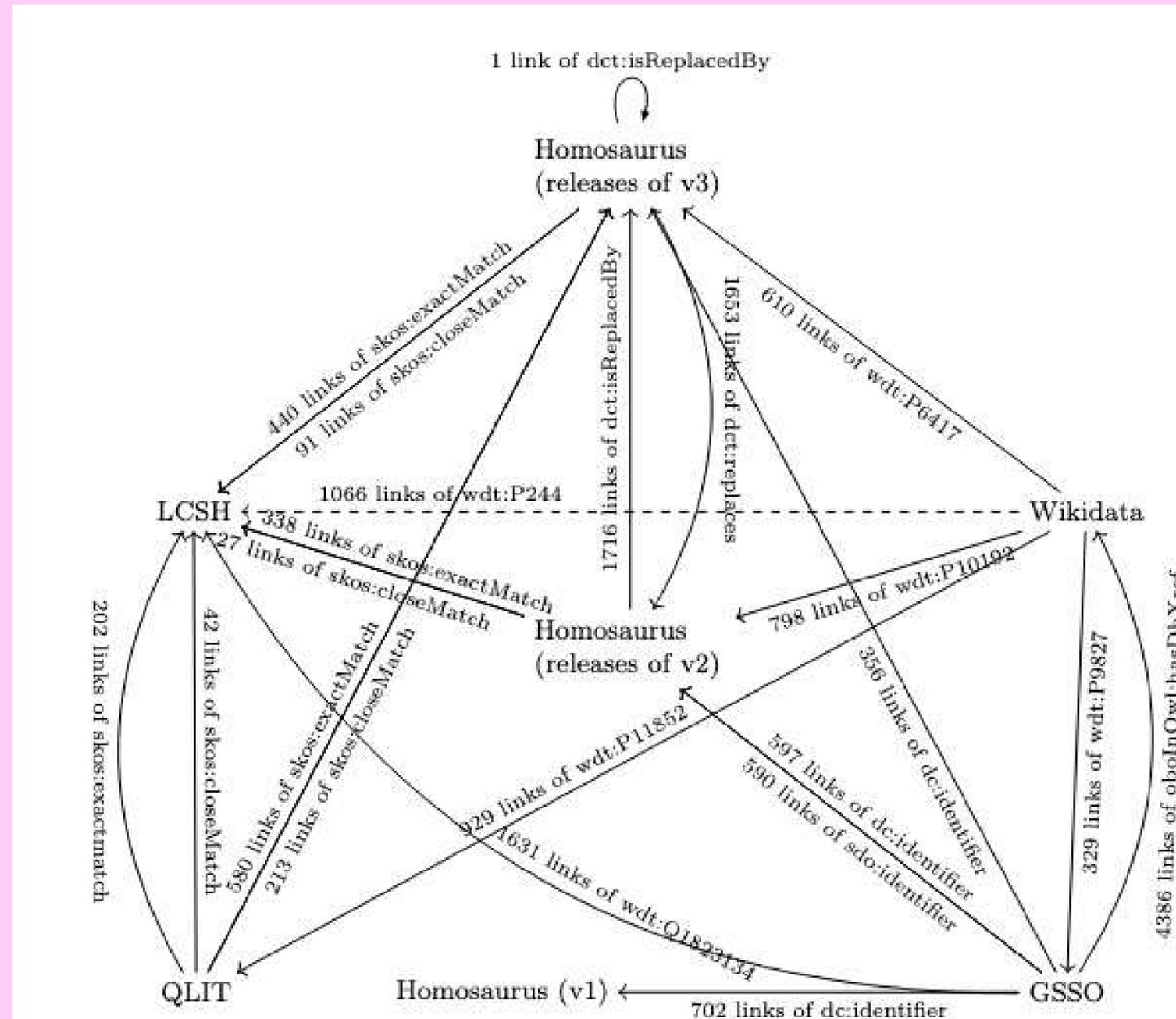
**Fig. 1.** Conceptual models and their extracted links. The dashed edge indicates that only edges about LCSH entities that appear in the rest of the selected concept models were chosen in this study for further integration and analysis.

Shuai Wang, Maria Adamidou, *Examining LGBTQ+-Related Concepts in the Semantic Web: Link Discovery, Concept Drift, Ambiguity, and Multilingual Information Reuse.* EKAW 2024

	Homosaurus (4.1.0): @ar, @bla, @bn, @bug, @ceb, @chy, @egy, @en, @en-gb, @en-us,@es, @fij, @fil, @haw, @he, @hil, @id, @ja, @km, @kn, @lat, @lkt, @mdy, @mg, @mi, @mov, @ms, @mzk, @nap, @ne,niu, @nv, @oj, @pa, @pmy, @qu, @rar, @ru, @smo, @sq, @srn, @sw, @ta, @tah, @tha, @tl, @ton, @tr, @tsg, @ur, @wba, @zap, @zh, @zmw, @zun							QLIT
	English (en)	Dutch(nl)	French(fr)	Spanish(es)	Bengali (bn)	Hindi(hil)	Others	Swedish(sv)
Translated by	Homosaurus board	Jack van der Wel (IHLIA)	Sandrine Vachon (and team?)	Ernesto Cuba, volunteers, and team	Prof. Partha Sarathi Mukhepadhyay, and volunteers	Prof. Partha Sarathi Mukhepadhyay and volunteers	Unknown	Olov Kriström (and colleagues)
Translation methods (labels)	-	Manual translation	Manual translation	Manual translation of 5499 labels	Manual translation	Manual translation	Unknown	Manual translation
Translation methods (context: comment/scopeNote)	-	Manual translation	Manual translation with volunteers	Manual translation with volunteers: 3011 comments, no scopenotes detected.	Manual translation with volunteers	Manual translation with volunteers	unknown	Manually translated scopenotes, no comments
Glossary involved?	unknown	unknown	unknown	unknown	unknown	unknown	unknown	Unknown
Evaluation	Manual	unknown	unknown	unknown	unknown	unknown	unknown	None
Indicators	None	None	None	None	None	None	unknown	None
Included in v4.1.0 release?	Yes	No	No	Yes	Yes, 1 label	Yes, 8 labels.	labels in 55 languages	Not applicable

MOTIVATION

- Not transparent: not always clear how, who, why, funder, outdated, what bias, what resources (dictionary, references) used, current status?
- Some translations were integrated and published, but the quality might be questionable: without standardized glossaries, evaluation criteria, documentation, etc.
- Ethical and political issues not highlighted: LCSH: sexual deviation, illegal aliens, and undocumented immigrants, etc.
- Relations between entities are being updated at each release → documented by Homosaurus!
- Concept changes (convergence, divergence, drift) have not been properly analysed.
- Missing Indicators, measures, "numbers", etc.
- The use of A.I. in translation?
- Standardisation and specificity.
- Availability (many are not FAIR).
- Sustainability.
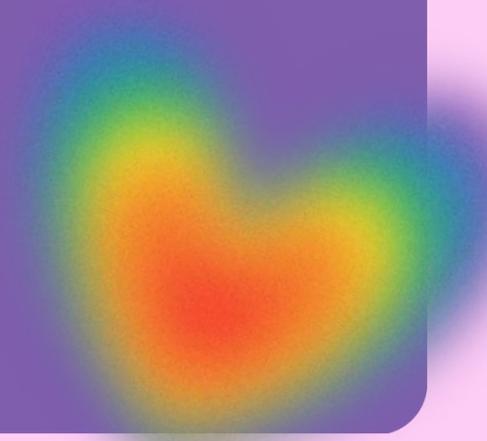
BEST PRACTICES EVALUATION CRITERIA

1. Indicators
2. Self-assessment Report Template

BEST PRACTICES

- Clarity and Accuracy
- Consistency
- Cultural and Contextual Sensitivity
- Inclusivity and Ethical Considerations
- Transparency and Community Contribution
- Documentation, Publishing, and Maintenance

Clarity and Accuracy

- Specified Writing System
- Scope: QLIT (LGBTQI) v.s. Homosaurus (LGBTQ+)
- Translation of Acronym: QLIT (hbtqi) v.s. Homosaurus (LGBTQ+)

Consistency

- Missing glossary. Ex. "legal name" → Naamrecht but "legal name change" → Wettelijke naamswijziging
- Syntactic consistency: African American queer people
- Translation of Pronouns: Co/Co/Coself and Cy/Cyb/Cyberself. Many could not be translated.

Cultural and Contextual Sensitivity

- Preserve the meaning of slang, e.g., bears.
- Localization: no direct translation of "gay parents" in Chinese.

Transparency, Documentation, FAIR, etc.

- Transparent about the editing process.
- Knowledge representation, metadata, etc.

EVALUATION CRITERIA: INDICATORS

- The number of concepts translated (and their sources)
- number of terms about slangs/slurs, pronouns, new terms, biased, culturally sensitive, etc.
- The ratio of preferred/alternative labels, comments, scopenotes, links, etc.
- Simplicity/readability indicators:
    - Bilingual Evaluation Understudy (BLEU)
    - Word Error Rate (WER)
- Levenshtein distance (between the initial version and the refined version)

EVALUATION CRITERIA: SELF-ASSESSMENT REPORT TEMPLATE

CHECKLIST 1: GLOSSARIES AND GUIDELINES OF TRANSLATION

1. The writing system of the target language needs to be specified.
2. There is a glossary to provide a <u>standardized</u> translation of selected tokens and terms.
3. There are <u>criteria for the inclusion and exclusion</u> of terms for the selection of terms to be translated.
4. There is a set of rules for the translation of syntactically similar or analogical expressions to maintain the <u>syntactic structure</u> where possible.
5. There is a guideline on translating <u>historical</u> terms.
6. There is a list of selected <u>pronouns</u> with their translation specified.
7. If any translation is exceptional or counter-intuitive, the reason or explanation should be included in the comments or scope notes.

CHECKLIST 2: DOCUMENTATION AND PUBLICATION

1. **WHO**: The documentation should include information about all the members and their affiliations, if possible.
2. **HOW**: The frequency of maintenance should be specified in the documentation. The same for the frequency of release of the sources used.
3. **WHAT**: Each term is associated with information about publication and modification.
4. **FAIR**: The published dataset with a persistent identifier (PID). If possible, associate each version with a PID.
5. **FAIR**: The metadata should be updated and enriched, made publicly accessible, and uploaded to search engines (e.g., Google Dataset Search) and portals (e.g., the ODISSEI portal).
6. **Usability**: The documentation should include instructions for use and application as well as a list of related resources and use cases in the documentation (if available).

THANK YOU!

https://zenodo.org/records/15082539

Speical thanks to Wilfred van Buuren (IHLIA), Gerrit Wessel (IHLIA), Andrei Nesterov (CWI), and Martin Borissov Mashalov (UvA) for reviewing.

Towards Semi-automatic Construction of Multilingual LGBTQ+ Conceptual Models

Int. Conf. Multilingual Digital Terminology Today (MDTT), Thessaloniki, Greece, June 2025

Examining LGBTQ+-related Concepts in the Semantic Web: Link Discovery, Concept Drift, Ambiguity, and Multilingual Information Reuse

International Conference on Knowledge Engineering and Knowledge Management (EKAW), Amsterdam, the Netherlands, Nov 2024

with kind contribution from Wilfred van Buuren (IHLIA), Gerrit Wessel (IHLIA), Andrei Nesterov (CWI/Dataout Foundation), and Martin Borissov Mashalov (UvA)

Email: shuai.wang.vu@gmail.com

AN OVERVIEW OF THEIR STATUS

- Initially developed by IHLIA as a thesaurus in the Netherlands.
- Published as a linked open vocabulary, with a SPARQL endpoint.
- Available in English and Dutch. This year, it was enriched with labels of Spanish, French, etc.
- Used together with LCSH in IHLIA, and many LGBTQ+ and generic libraries.
- Official website: https://homosaurus.org/v4, with a SPARQL endpoint.
- CC-BY-NC-ND license

QLIT

Queerlit

- QLIT = Queer Literature Indexing Thesaurus
- Oriented for indexing and searching for the online literature database QueerLit.
- A significant number of terms were translated from that of Homosaurus.
- QLIT has been integrated into Libris XL.
- Queerlit website: The https://queerlit.dh.gu.se/qlit/v1
- LIBRIS website: https://libris.kb.se/bib/k1m44fr0hr20v690

GSSO

- GSSO = Gender, Sex, and Sexual Orientation
- An ontology initially created for the (bio)medical domain, seeing application in GLAM lately.
- https://github.com/Superraptor/GSSO; last maintained 3 years ago.
- Available on its webpage@CCHMC, Ontobee, NCBO BioPortal, EMBL-EBI OLS, etc.
- CC-BY-NC-ND license

Wikidata

- Generic purposes; contributed and maintained by the community
- Contains multilingual LGBTQ+ terms.
- SPARQL endpoint: https://query.wikidata.org/sparql