

24th International Conference on
Knowledge Engineering and
Knowledge Management (EKAW)

Examining LGBTQ+-related Concepts in the Semantic Web:

Link Discovery, Concept Drift,
Ambiguity, and Multilingual
Information Reuse

Shuai Wang, Maria Adamidou

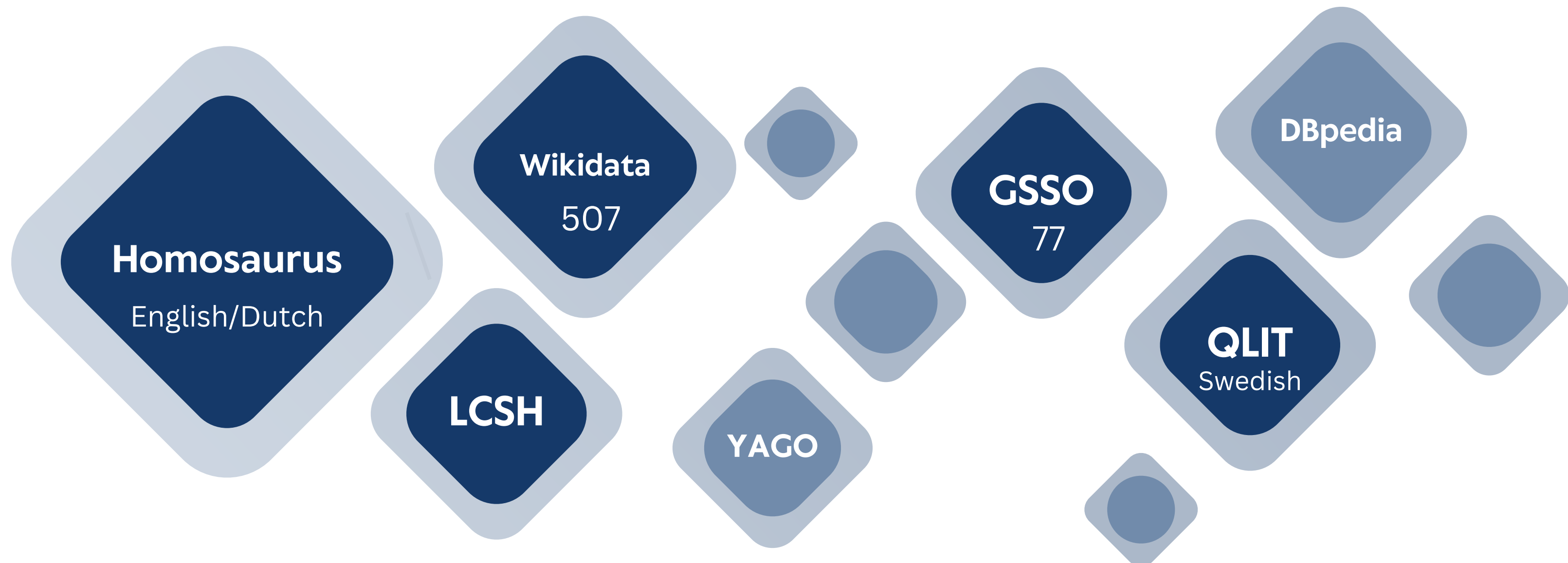


Introduction

LGBTQ+-related concepts widely exist in the semantic web
Many of them are published as linked data and are interlinked

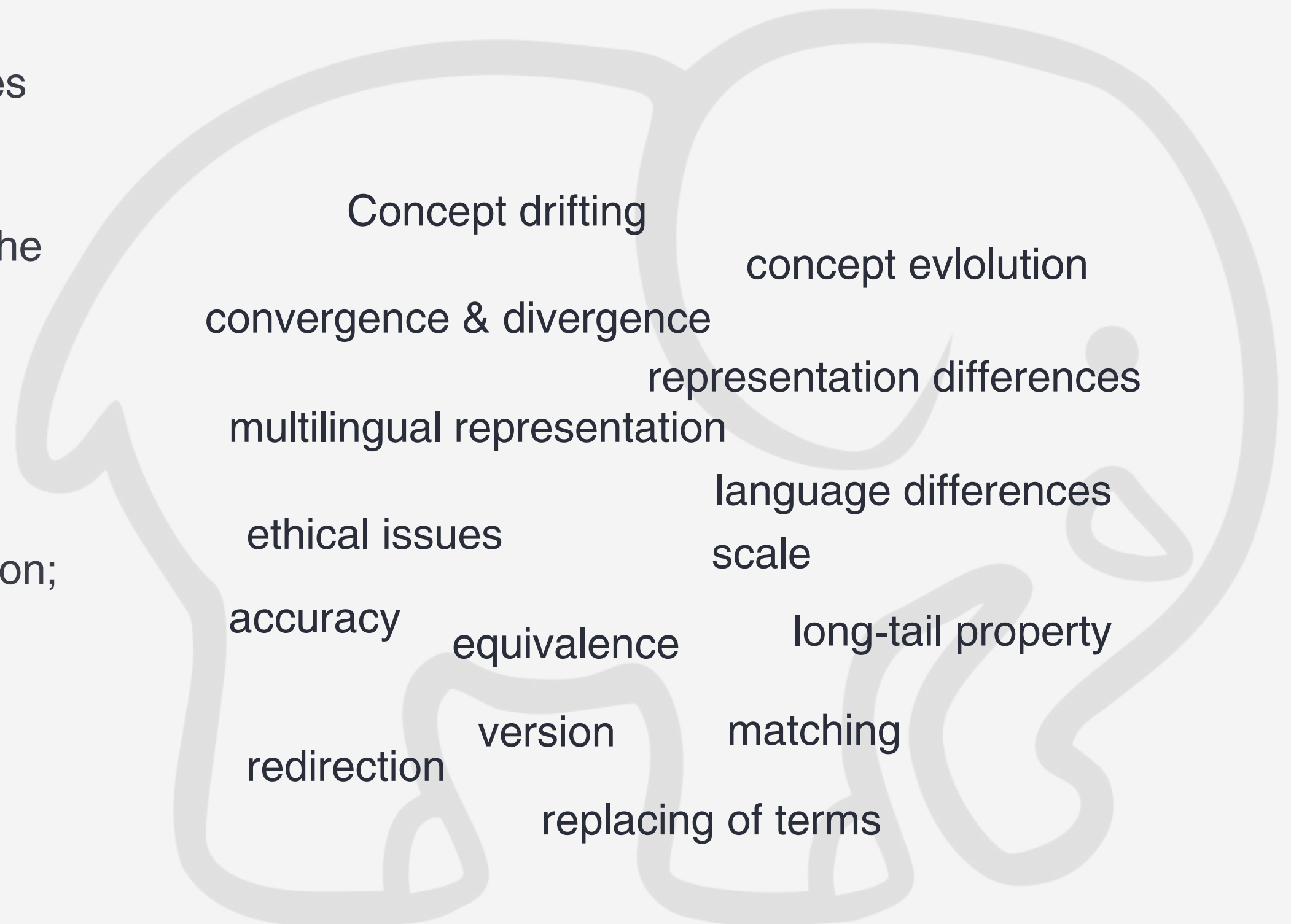
Conceptual Models

Our study excludes glossaries, vocabularies, and thesauri that are not linked data



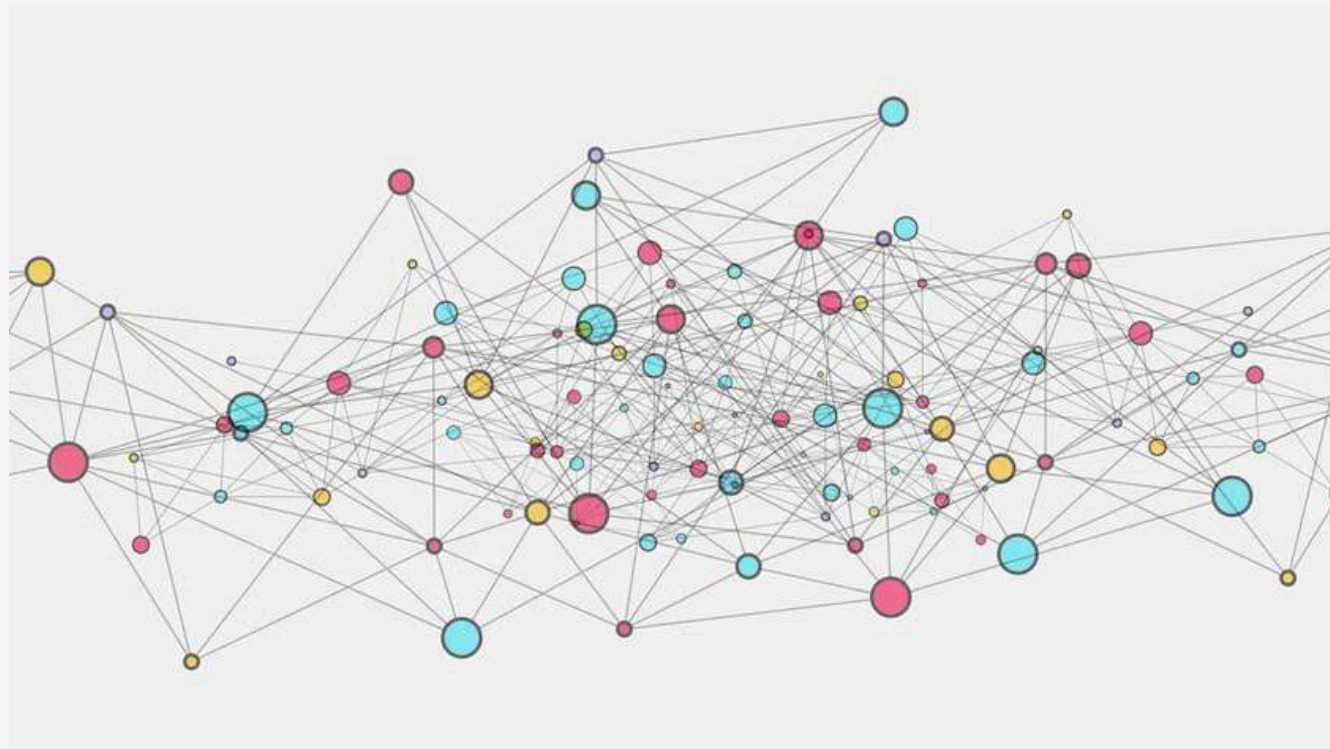
Related Work

- ◆ Braquet briefly examined LGBTQ+ entities within library settings [4].
- ◆ Dobreski et al. compared the overlap of the Homosaurus, LCSH, and LCDGT [5].
- ◆ Others reported mappings from Homosaurus terms to outdated terms in LCSH can lead to problems [5, 16].
- ◆ Redirection and concept drifting is common; has not been systematically studied yet.
- ◆ Identity crisis in the semantic web



Our Approach

has its roots in the needs of community



**Construct a KG with extract links:
identity-related links and
those about identity changes**



Link Discovery



**Concept Drift
and Ambiguity**



**Multilingual
info reuse**

Data Engineering

1. Relation extraction
2. Refinement (wd:Q1823134->wdt:P244)
3. Redirection

- skos:exactMatch/closeMatch
- dc:identifier
- dct:isReplacedBy/replaces
- oboInOwl:hasDbXref
- wdt:P6417/P10192/etc.
- redirection (61+2): meta:redirectedTo
- ~~owl:sameAs~~

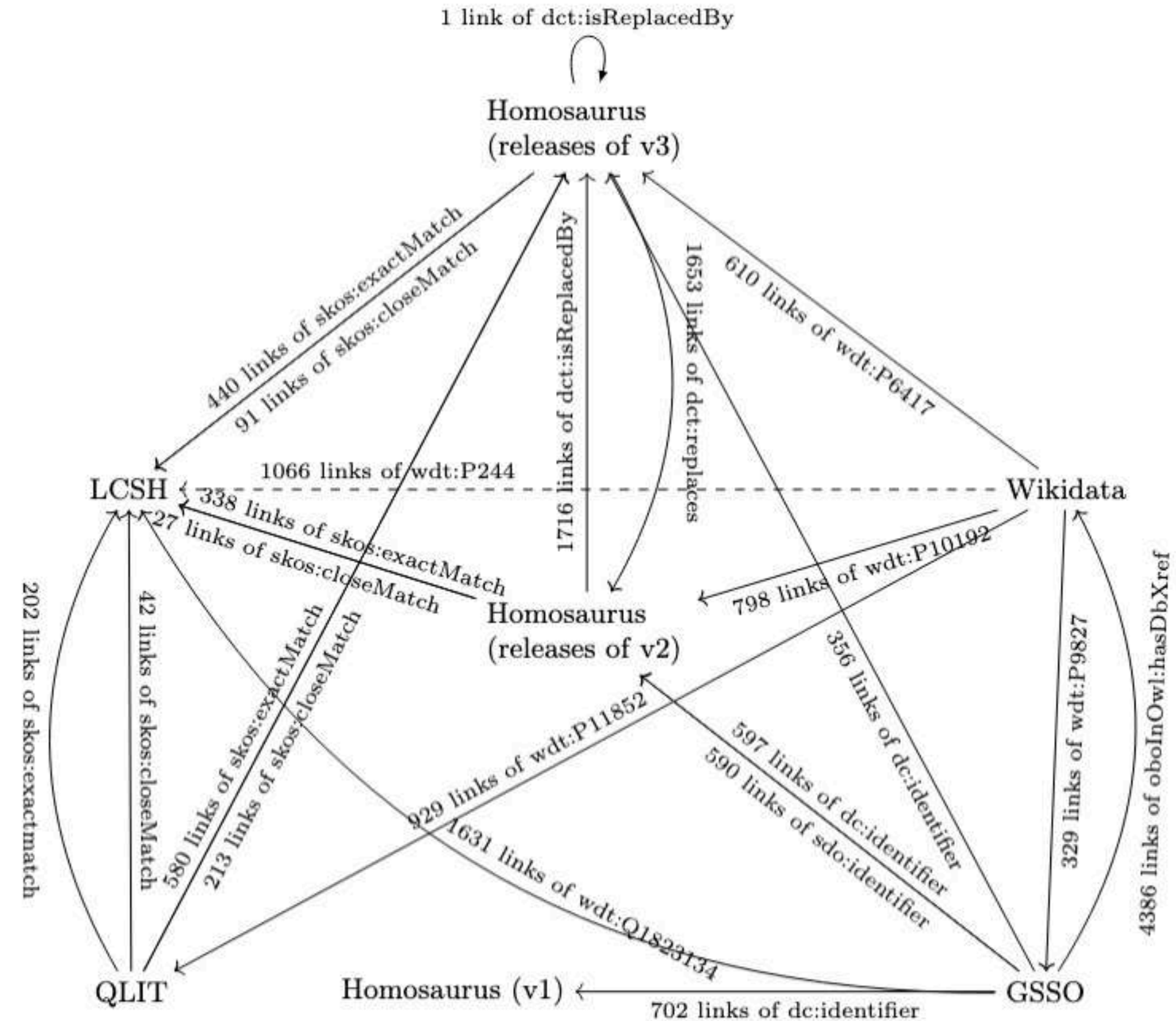


Fig. 1. Conceptual models and their extracted links. The dashed edge indicates that only edges about LCSH entities that appear in the rest of the selected concept models chosen in this study for further integration and analysis.

Weakly Connected Components

A Weakly Connected Component (WCC) for a directed graph is the maximal subgraph s.t. the vertices can reach each other by ignoring the direction of edges

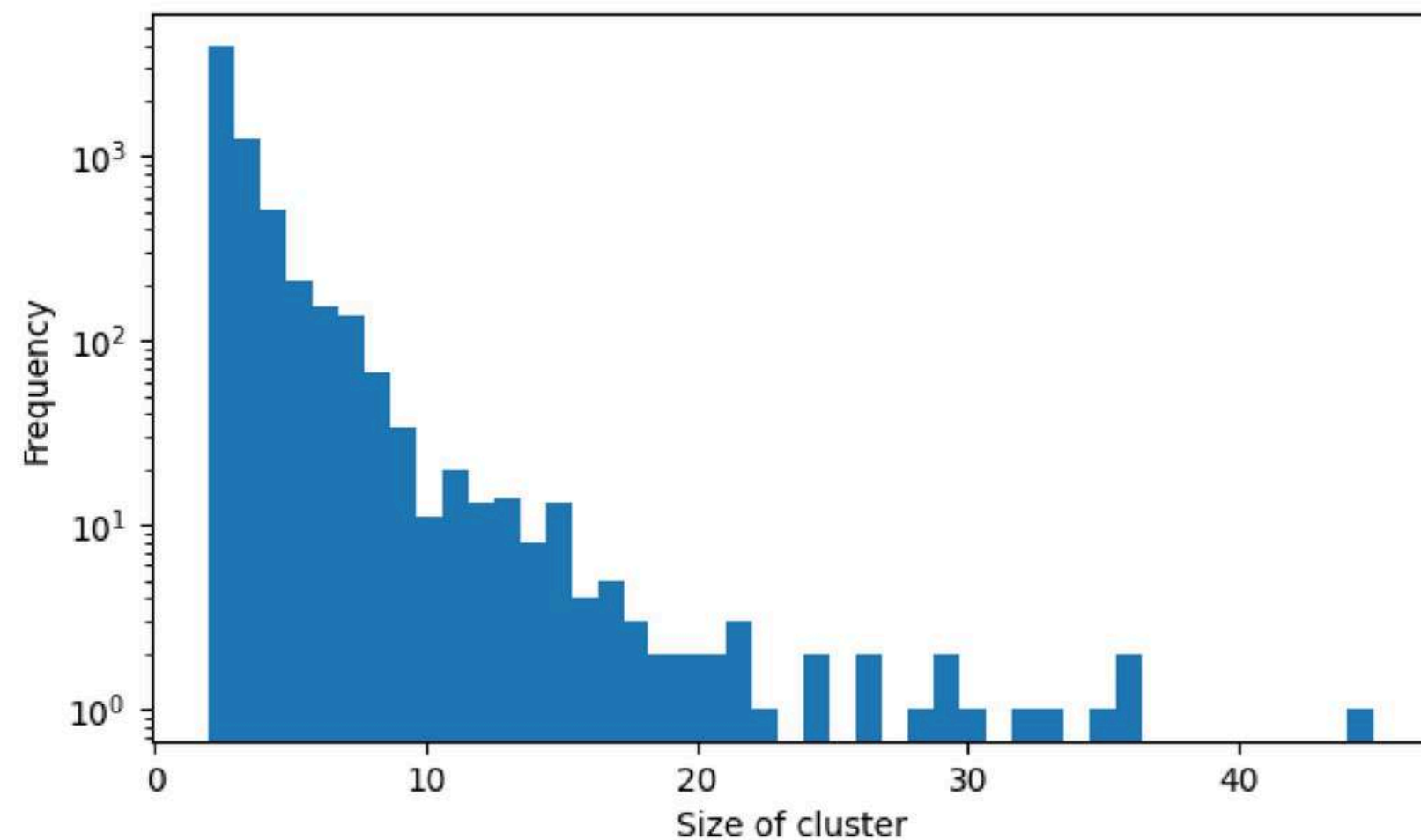


Fig. 2. Frequency histogram of the size of clusters

Entities

19.2K

Links

17.8K

Weakly Connected Components

6.4K

The size of the biggest WCCs

45, 36, 36, 35

Link Discovery

- Often take each other as references
- But not good at keeping track of the sources
- The intuition of our approach:

In a WCC, if there is exactly one entity from two conceptual models each, there could be a link added.

Kind of
"Ambiguitation-free"

Homosaurus -> LCSH

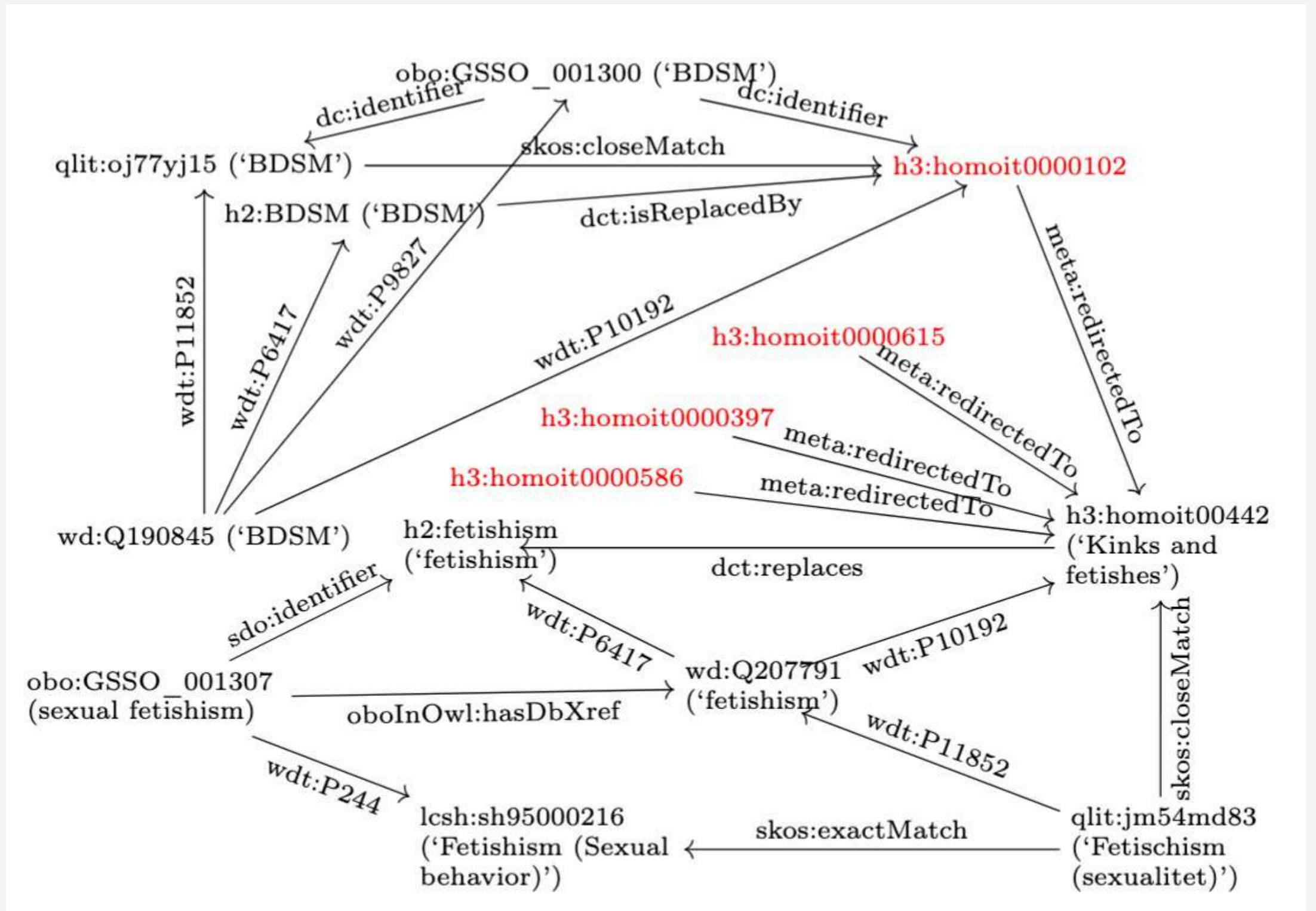
25 newly discovered links despite 531 existing links between Homosaurus and LCSH.

QLIT -> LCSH

- 105 potential links newly found (QLIT has 244)
- 38 (36.19%) can be included using skos:exactMatch
- 38 (36.19%) using skos:closeMatch.
- Thus, 72.38% could be included
- The remaining are incorrect

Concept Drift and Change

- ◆ Entities in red are no longer in the latest version of Homosaurus. Redirection links were not captured.
- ◆ "BDSM" has a broader scope now, resulting in concept convergence.
- ◆ Visualizing experts' perception and the drift of (sexual) fetishism, kinks, and BDSM.
- ◆ Ambiguity, concept drift & change, and the change of scope is often blended into complex cases which requires manual revision



Queer, Queen, Wolf, Daddy, etc.

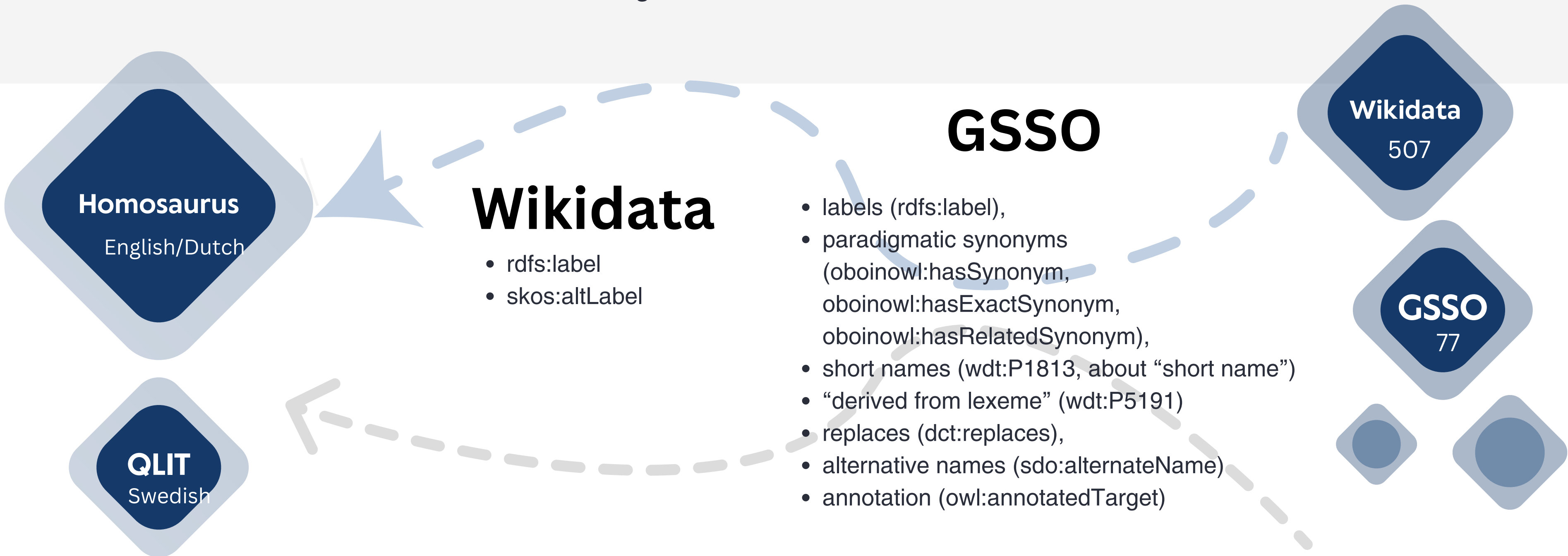
SUBMIT

Multilingual info reuse

Community's needs for multilingual labels.

Manually searching for alternative labels.

The reuse of multilingual labels for entities in the same WCCs



Reuse! But how much?

Homosaurus



GSSO

Only 48 entities can be enriched
top 3: English, Danish, and French

Homosaurus



Wikidata

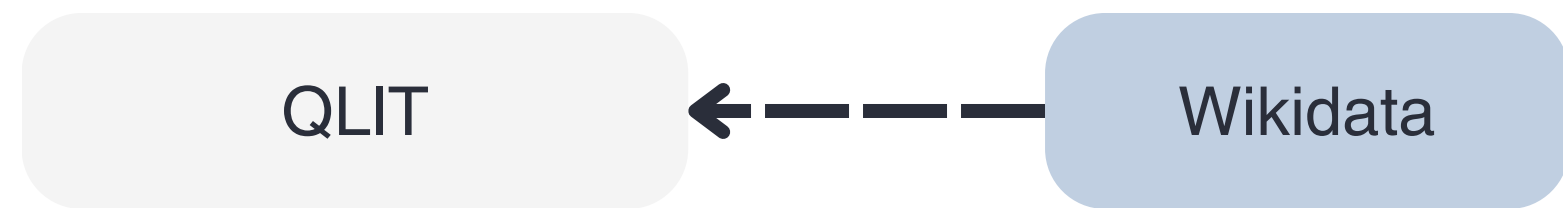
429 entities can be enriched
top 4 languages:

- English (1,692 labels for 429 entities)
- Spanish (951 labels for 333 entities)
- Chinese (893 labels for 287 entities)
- Portuguese (881 labels for 299 entities)

SUBMIT

QLIT:

- 914 entities with one prefLabel each
- only 480 altLabels



775 Swedish labels in Wikidata (524 prefLabels and 251 altLabels) for 524 entities.

SUBMIT

imagine if they had these labels from the beginning!

Discussion

- 1 WCC-based approach has its limit when used for link discovery: there has to be a path.
- 2 The quality of the newly found links of Homosaurus remains unknown. It can take quite some time for experts to manually examine them. Lack of awareness.
- 3 Wikidata provides more labels in more languages than GSSO. But the quality remains unknown.
- 4 The suggested labels could violate the formatting of terms in the selected conceptual models. The overlapping and quality of the suggestions remains unknown.
- 5 owl:sameAs, owl:differentFrom, the Unique Name Assumption, etc.
- 6 Licencing issues.

Conclusion & Future Work

- Constructed a knowledge graph with selected relations about identity and their change.
- WCC-based approach for the analysis of concepts and their change.
- Three different scenarios of LGBTQ+-related concepts for the needs of the community.

- A new approach with some potential to be adapted to others domains.
- Some resources (new links, WCCs, multi-ling' labels)
- Visual representations
- Combinig with NLP (LLM)?
- Reuse of multilingual labels for fast developement of multilingual Homosaurus and others.
- Other conceptual models such as DBpedia, YAGO.



Zenodo

DOI [10.5281/zenodo.12684870](https://doi.org/10.5281/zenodo.12684870)



Github

https://github.com/Multilingual-LGBTQIA-Vocabularies/Examining_LGBTQ-related_Concepts

Acknowledgement

Andrei Nesterov

CWI

Jack van der Wel

IHLIA/Homosaurus

Clair Kronk

GSSO

Siska Humlesjö

Olov Kriström

QLIT



Email

shuai.wang@vu.nl



Twitter/X

shuai_wang_ai



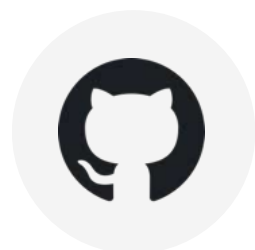
LinkedIn

@shuai-ai



Zenodo

DOI **10.5281/zenodo.12684870**



Github

https://github.com/Multilingual-LGBTQIA-Vocabularies/Examining_LGBTQ-re



See you at the ODISSEI Conference

References

1. Trans & Gender Diverse LCSH (2024), <https://translcsch.com/>, The list was last accessed on 25th May, 2024.
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: international semantic web conference. pp. 722–735. Springer (2007)
3. Bergenmar, J., Golub, K., Humelsj., S.: Queerlit database: Making swedish lgbtqi literature easily accessible. In: DHNB 2022: The 6th Digital Humanities in the Nordic and Baltic Countries Conference 2022. pp. 433–437. CEUR-WS. org (2022)
4. Braquet, D.: Chapter 2 LGBTQ+ Terminology, Scenarios and Strategies, and Relevant Web-based Resources in the 21st Century: A Glimpse, pp. 49–61 (05 2019). <https://doi.org/10.1108/S0065-283020190000045009>
5. Dobreski, B., Snow, K., Moulaison-Sandy, H.: On overlap and otherness: A comparison of three vocabularies' approaches to lgbtq+ identity. *Cataloging & Classification Quarterly* 60(6-7), 490–513 (2022). <https://doi.org/10.1080/01639374.2022.2090040>
6. Ihrmark, D.O., Golub, K., Tan, X.: Subject indexing of lgbtq+ fiction in sweden and china. In: Knowledge Organization for Resilience in Times of Crisis: Challenges and Opportunities. pp. 379–384. Ergon-Verlag (2024)
7. Jagose, A.: *Queer theory: An introduction*. NYU Press (1996)
8. Kazarian, A.M., Wang, S.: Evaluating Automated Machine Translation of LGBTQ+ Terms: Towards Multilingual Homosaurus (Mar 2024). <https://doi.org/10.5281/zenodo.10523283>
9. Kronk, C.A., Dexheimer, J.W.: Development of the gender, sex, and sexual orientation ontology: Evaluation and workflow. *Journal of the American Medical Informatics Association* 27(7), 1110–1115 (2020)
10. Lynch, K.E., Alba, P.R., Patterson, O.V., Viernes, B., Coronado, G., Du-Vall, S.L.: The utility of clinical notes for sexual minority health research. *American Journal of Preventive Medicine* 59(5), 755–763 (2020). <https://doi.org/https://doi.org/10.1016/j.amepre.2020.05.026>, <https://www.sciencedirect.com/science/article/pii/S0749379720302774>
11. Matsson, A., Kristr.m, O.: Building and serving the queerlit thesaurus
12. Nasim, I., Wang, S., Raad, J., Bloem, P., van Harmelen, F.: What does it mean when your URIs are redirected? Examining identity and redirection in the LOD cloud. In: Proceedings of the 8th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) (2022)
13. Office of Communications and Marketing: An LGTBQ language thesaurus is translated to spanish (2024), <https://www.gc.cuny.edu/news/lgbtq-language-thesaurus-translated-spanish>, accessed on May 19, 2024
14. Peterson, R.: Library of congress subject headings for lgbt studies (8 2023), <https://guides.libraries.emory.edu/main/queerlcsch>
15. Tai, J.: Cultural humility as a framework for anti-oppressive archival description. *Reinventing the Museum: Relevance, Inclusion, and Global Responsibilities* p. 349 (2023)
16. The Homosaurus editorial Board: Homosaurus vocabulary site (2024), <https://homosaurus.org/about>, Its documentation was last accessed on 24th May, 2024.
17. Vrandečić, D., Kr.tzsch, M.: Wikidata: a free collaborative knowledgebase. *Communications of the ACM* 57(10), 78–85 (2014)
18. Wang, S., Schlobach, S., Klein, M.: Concept drift and how to identify it. *Journal of Web Semantics* 9(3), 247–265 (2011). <https://doi.org/https://doi.org/10.1016/j.websem.2011.05.003>, semantic Web Dynamics Semantic Web Challenge, 2010
19. Wang, S., Maineri, A., Singh, N., Kuhn, T.: FAIR implementation profiles for social science. In: Garoufallou, E., Sartori, F. (eds.) *Metadata and Semantic Research*. pp. 284–290. Communications in Computer and Information Science, Springer Science and Business Media Deutschland GmbH, Germany (2024). https://doi.org/10.1007/978-3-031-65990-4_26
20. Wang, S., Raad, J., Bloem, P., van Harmelen, F.: Refining large integrated identity graphs using the unique name assumption. In: European SemanticWeb Conference. pp. 55–71. Springer (2023)
21. Watson, B.M.: “there was sex but no sexuality*” critical cataloging and the classification of asexuality in lcsch. *Cataloging & Classification Quarterly* 58(6), 547–565 (2020)

Homosaurus:

- Version issues
- Entities that are missing in the latest versions
- Lack of links between the old and new versions
- Newly added terms
- Merged/removed terms

Table 2: A summary of the version updates of Homosaurus

Version	Release Date	#Terms newly added	#Terms re-removed	#Terms with labels changed	Available	Comment
v2.1	Jun 2020	99	0	0	No	no information found for earlier versions
v2.2	Dec 2020	23	0	0	No	
v2.3	Jul 2021	69	2	3	Yes	newly added terms include 45 pronouns-related terms
v3.0	Sep 2021		-	-	No	a new release
v3.1	Dec 2021	77	3	276	No	'LGBTQ' changed to 'LGBTQ+' in all terms. All terms formatted as [Term] (LGBTQ) changed to LGBTQ+ [term]
v3.2	Jun 2022	263	0	148	No	8 terms were redirected
v3.3	Dec 2022	308	0	32	Yes	25 terms replaced some older terms
v3.4	Jun 2023	530	0	1	Yes	
v3.5	Jan 2024	255	0	24	Yes	the latest release

Table 1. Extracted relations from sources and the number of triples

Source	Relation	#Triples	Comments
Homosaurus	dct:isReplacedBy and dct:replaces	3,370	Mostly links about replacing between version 2 and version 3.
	skos:exactMatch and skos:closeMatch	896	Links to entities in LCSH extracted from Homosaurus v2 and v3.
	meta:redirecsTo	63	Links representing redirection between entities in Homosaurus v3. Redirects for v2 were not included.
GSSO	wd:Q1823134	1,827	Links from entities in GSSO to subject headers in LCSH. It is mistaken to use wd:Q1823134. It was replaced by wdt:P244 in the integrated graph.
	oboInOwl:hasDbXref	4,643	Links from entities in GSSO to entities in Wikidata
	dc:identifier and sdo:identifier	2,245	Links from entities in GSSO to entities in Homosaurus (all three versions)
QLIT	skos:exactMatch and skos:closeMatch	793	There are only links to Homosaurus v3.
	skos:exactMatch and skos:closeMatch	244	Links from QLIT to LCSH
Wikidata	wdt:P244	1,066	Selected links from Wikidata to LCSH
	wdt:P6417 and wdt:P10192	1,408	Links from Wikidata to Homosaurus 2 and 3
	wdt:P11852	929	links from Wikidata to QLIT
	wdt:P9827	328	links from Wikidata to GSSO
Overall		17,812	The integrated graph involves 19,200 entities.

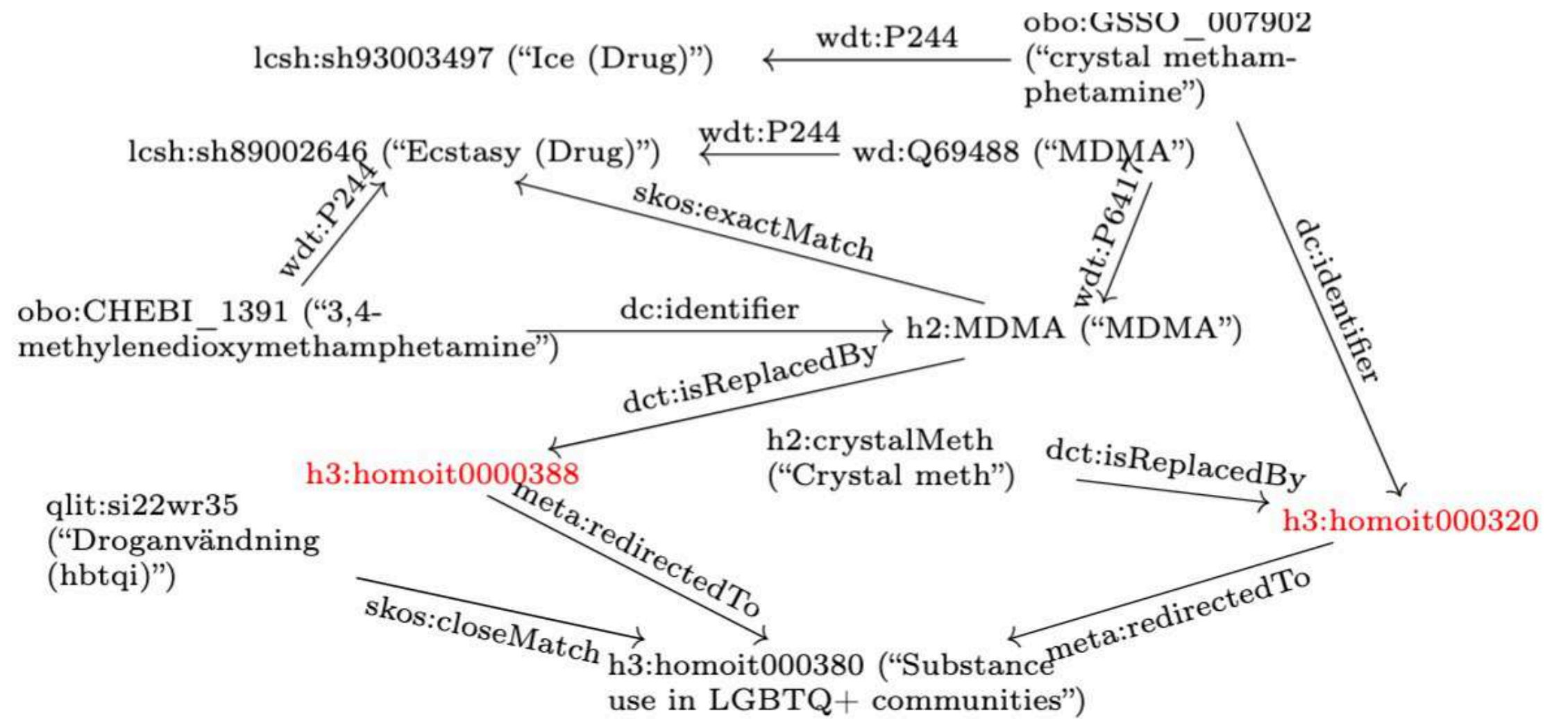


Fig. 4. An example of concept drift and change involving 3,4-methylenedioxymethamphetamine, MDMA, Crystal Meth, Ecstasy, Ice, Substance use in LGBTQ+ communities, etc. Some entities and links are not included for clear visualization. Highlighted in red are two entities in Homosaurus but not in v3.

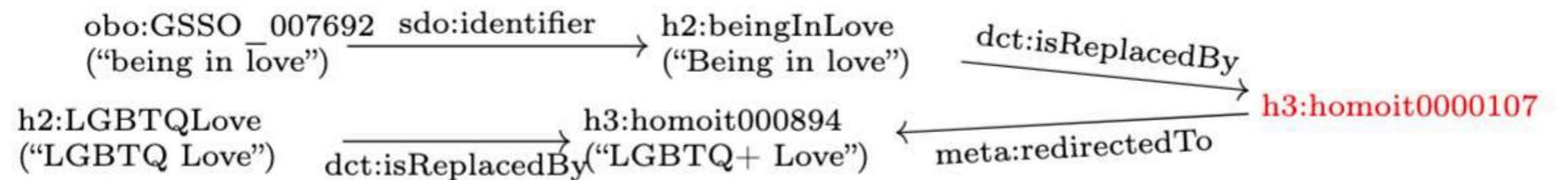


Fig. 5. Convergence of "Being in love" and "LGBTQ love" to "LGBTQ+ love". Following the links could lead to a change in scope. Highlighted in red is an entity no longer maintained in Homosaurus v3. Not all entities and links in the WCC were illustrated.

Acknowledgement

Thank you so much for your help!

Andrei Nesterov

CWI

Jack van der Wel

IHLIA/Homosaurus

Clair Kronk

GSSO (Gender, Sex, and Sex Orientation ontology)

Siska Humlesjö and Olov Kriström

QLIT