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

unknown

No Institute Given

Abstract. Recent years have seen a notable increase in the use of LGBTQ+ ontologies and structured vocabularies in library systems, digital archives, online databases, heritages, etc. Many were published as linked data, including Homosaurus, QLIT, GSSO, etc. However, little has been reported about the links between the concepts captured by them. We retrieve all their published mappings as well as relevant information to form an integrated knowledge graph. Taking advantage of its weakly connected components, we study the discovery of missing links between entities. We analyze concept drift and change by providing examples of concept convergence, ambiguity, and scope change. Moreover, we study how multilingual information from other resources can enrich entities using Homosaurus as an example. Finally, we discuss potential challenges and the practical implications of our findings.

Keywords: Homosaurus · concept drift · identity management.

1 Introduction

Although the LGBTQ+ demographic group has gained more visibility and acceptance, there is still a considerable disparity between the intricate nature of their lived experiences and the current structured vocabulary and other knowledge representation initiatives. Even with the advancements made by several projects with different goals in recent years, many still fall short in capturing the diverse and nuanced realities of the evolving vocabulary used in the community, leaving gaps when aligning concepts between these attempts. This may lead to interoperability issues when projects become outdated, requiring revisions of their compatibility with other projects.

For the sake of clarity, in this paper, we use the term *conceptual models*[1] as the umbrella term for thesauri, structured vocabularies, ontologies, subject headings, and knowledge bases. More specifically, we focus on conceptual models published as linked data in the semantic web. Concepts are represented as entities with (multilingual) *labels*. *Links* between entities in conceptual models are in

[1] An alternative umbrella term used in the community is ‘knowledge organization systems’ (KOS). However, here we study concepts’ change in semantics and relations rather than the system. Thus, we opt for the umbrella term of ‘conceptual model’.

the form of a triple with subject, predicate, and object. A *mapping* is a set of links between entities from two different conceptual models. By *community*, we specifically refer to the individuals who are engaged in the creation and development of the conceptual models to be studied below. In this paper, we focus on mappings between selected conceptual models and some related links as well as their multilingual labels. Other aspects such as comments, scope notes, and definitions are excluded from this study.

Homosaurus[2] [16] was initiated by librarians in the IHLIA LGBTI Heritage[3] and has become a popular conceptual models in LGBTQ+ libraries and archives. There are mappings between Homosaurus, the QLIT thesaurus [3], the Gender, Sex, and Sexual Orientation (GSSO) ontology [9], the Library of Congress Subject Headings (LCSH) [14], and the Wikidata knowledge base [17]. Many of these terms have evolved or are mistakenly linked to reclaimed terms. For example, 'wolves' is reclaimed as a slang term[4] for 'masculine gay men who are often characterized as having hairy bodies and facial hair'. However, it has a link of `skos:exactmatch` to a term[5] in LCSH, which is about the animal wolf. Moreover, since these entities and links are published in the semantic web, mistakes and accumulated subtle changes can result in errors and complex concept drift which demands careful examination from multiple parties, especially in a multilingual setting. Recently, the need for LGBTQ+ conceptual models that support multiple languages has grown, such as Spanish [13] and Chinese [6]. Considering the suboptimal efficacy of machine translation [8], experts must put considerable effort into manually translating terms and their scope notes. It remains an open question whether the reuse of pre-existing terms in specific languages from relevant resources could facilitate the development of bilingual or multilingual conceptual models. Therefore, in this paper, we construct a knowledge graph based on links between conceptual models and relevant information on identity management. Using it, we study the evolution and drifting of LGBTQ+-related concepts in the semantic web. Finally, we explore how multilingual information from one conceptual model can be used to enrich entities of another.

In this paper, we retrieve the mappings and related links that contain identifying information, as well as links about concept replacement and redirection. We aim for a preliminary quantitative and qualitative analysis and propose possible solutions. More specifically, our contributions are as follows. a) We provide our code and all the data[6] using Wikidata and QLIT to reproduce our experimen-

[2] `https://homosaurus.org/`.

[3] `https://ihlia.nl/en/collection/homosaurus/`

[4] `https://homosaurus.org/v3/homoit0001508`

[5] `http://id.loc.gov/authorities/subjects/sh85147257`

[6] The Python scripts, SPARQL queries, detailed explanation of experiments, and the analytical results are in the supplementary material at: `https://figshare.org/0123456789`. Wikidata and QLIT are with the license CC0. Thus, only datasets extracted from them are provided. Due to the strict CC-BY-NC-ND license of GSSO and Homosaurus, the remaining datasets are only accessible upon request from IHLIA as well as the developers of GSSO and QLIT. See the github repository for the latest version of the Python scripts: `link.anonymised`.

tal results as well as for future use. b) We demonstrate a detailed examination of LGBTQ+-related concepts and their links in the semantic web and their multilingual information from three aspects: link discovery, concept drift and (multilingual) ambiguity, as well as the reuse of multilingual information. c) By providing detailed analysis and examination of the labels of entities and related multilingual information, we address issues that could be fixed and aspects that the community can improve using our data. The orientation of this paper is knowledge representation and knowledge engineering. Bias, sensitivity, ethics, and political issues are beyond the authors' expertise and are not covered.

The paper is organized as follows. Section 2 introduces the conceptual models and presents the related work. In Section 3, we present details of data engineering. We evaluate the integrated graph in use in three scenarios in Section 4. Finally, we discuss the results in Section 5 with the conclusion in Section 6.

2 LGBTQ+ Conceptual Models and Related Work

As far as the authors are aware, there is no prior work directly related to the analysis of concept drift, and multilingual information of LGBTQ+-related concepts in the semantic web. Therefore, in this section, we summarize the updates of terms in the release notes of conceptual models and present some related research, but not all are exclusively about LGBTQ+ concepts.

Homosaurus is a linked data vocabulary focusing on LGBTQ+ terminology, aimed at enriching general subject term vocabularies, and it undergoes updates every six months. It was intended as a companion to LCSH [16]. In recent years, three versions of Homosaurus have been released with updates every half a year. It contains English terms along with their corresponding translations in Dutch, offering a valuable bilingual dimension to its utility. Serving as a robust and state-of-the-art conceptual model widely used in libraries and heritages, Homosaurus significantly enhances the findability of LGBTQ+ resources and information. Furthermore, Homosaurus offers a SPARQL endpoint[7] for accessing its data. At each release, information on updates of "labels" and newly added terms are provided on the website.[8] Despite some statistical analysis on terms in Homosaursus and how they overlap with others such as LCSH [5], to the authors' knowledge, there is no systematic examination or literature on the evolution of terms in Homosaurus and how its links to other conceptual models change.

The QLIT (Queer Literature Indexing Thesaurus) [11] is a recent Swedish thesaurus dedicated to indexing literature with LGBTQI themes. It was mainly used in Queerlit[9] [3], a bibliographic database on Swedish fiction. More than half of the terms in QLIT were translated from the English terms of Homosaurus

[7] https://data.ihlia.nl/PoolParty/sparql/homosaurus. Note that this endpoint may be delayed compared to the latest release on the Homosaurus website.

[8] See for example https://homosaurus.org/releases/show/3. This latest release in January 2024 added 255 new terms and changed the 24 terms. One term has been replaced.

[9] https://queerlit.dh.gu.se/

(v3.3). QLIT has mappings to the two main Swedish library thesauri: Svenska ämnesord (SAO) and Barnämnesord (Barn) [11].

The Gender, Sex, and Sexual Orientation (GSSO)[10] ontology was designed to facilitate communication in gender, sex, and sexual orientation research and assist knowledge discovery in literature [9]. Its second version includes 10,060 entries, an increase from 6,250 in its first version. Its application ranges from clinical studies [10] to archives [15].

The three conceptual models mentioned above have links to LCSH (Library of Congress Subject Headings) [14]. Despite the fact that it includes some LGBTQ+-related terms, it was reported to have flaws and can be influenced by politics [21], which is beyond the authors' expertise. We study it from a semantic web point of view. Wikidata [17] contains identifiers of GSSO, QLIT, and Homosaurus but, to the best of the authors' knowledge, have not been analyzed from this perspective.

These conceptual models serve distinct purposes, were revised at various times, are managed by teams with different expertise, and have not always been developed with full awareness of the changes of each other. Therefore, a perfectly unified representation of concepts and their relations is not possible. Braquet [4] briefly examines the provision of support for LGBTQ+ patrons within library settings, offering insights through various library-based scenarios. Dobreski et al. compared the overlap of the Homosaurus, LCSH, and Library of Congress Demographic Group Terms (LCDGT) [5]. They examined an old version of Homosaurus with 1,754 terms and found 618 terms related to identity. They reported 153 matches in the LCSH (exact matches and closest matches). Similarly, they found 176 matches in LCDGT, including faceted matches. Furthermore, it has been reported that there are outdated terms in LCSH, which leads to problems with the mapping of terms between Homosaurus and LCSH [5,16]. However, to the authors' knowledge, there is no systematic report on the quality and reliability of these links and what kind of consequences would there be following erroneous links or involving ambiguous entities. A comprehensive comparison of all the entities released and their relations in these conceptual models is missing.

We observed redirection when resolving URIs of Homosaurus. Prior examinations of entities indicate frequent redirections among entities in identity graphs, with an estimate of 45% to 83% maintaining the semantics of identity [12]. However, no research has been done to study how many URIs have been redirected among those corresponding to LGBTQ+-related concepts.

Concept drift refers to the phenomenon where the meanings or nuances of terms, concepts, or language evolve over time [18]. In the context of LGBTQ+ vocabularies, concept drift occurs as societal attitudes, understandings, and discussions about gender identity, sexual orientation, and related topics change and progress. An example is a term like "queer" which has changed in meaning over time. "Queer" was used as a slang for homosexuals as well as a term for homophobic abuse, but it has been reclaimed as an umbrella term for a coalition of culturally marginal sexual self-identifications in recent years [7]. The term

[10] `https://github.com/Superraptor/GSSO`

“homosexual” is now considered somewhat “clinical” [4]. When it comes to the analysis of concept drift at scale, a method for large knowledge bases with instances of classes was proposed by Wang et al., but it does not apply to our data due to the missing of instances [18]. As far as the authors are aware, there is no systematic report on the concept drift and change in the field.

3 Data Engineering

In this section, we present details of selected conceptual models and links extracted for our analysis in Section 3.1. Section 3.2 includes details of multilingual labels extracted for the study of information reuse. Moreover, it was observed that some outdated URIs were redirected to new URIs, but not explicitly captured in the conceptual models. Thus, we present how these redirection relations were obtained in Section 3.3. Finally, we integrate all the links in Section 3.4.

3.1 Dataset selection and data preprocessing

Since its version 2.1 in June 2020[11], Homosaurus experienced 8 updates (on average twice per year). In this study, we focus on the last release of version 2 (v2.3) and the latest release (Janurary 2024, v3.5), which captures 3,149 terms. It was noticed that some URIs[12] were no longer maintained in version 2 and were therefore replaced. Although replacement does not necessarily imply equivalence, we include this type of relation in our study to capture the evolution of conceptual models. Replacement was captured in Homosaurus using the relation `dct:isReplacedBy` and its inverse `dct:replaces`. Each entity is accompanied by an identifier (e.g. `homoit0002950`), a preferred label using `skos:prefLabel` and some using `skos:altLabel` as well as a scope note using `rdfs:comment`. Additionally, we noticed that some outdated URIs were redirected to newer ones. However, this information was not explicitly stated in the latest version. In Section 3.3, we extract these relations by resolving the URIs and capture this redirection relation for further analysis. Homosaurus has links to LCSH[13]. They are being used together in libraries and heritages. Our examination shows that none of the entities in LCSH is outdated.

Recall that QLIT[14] was developed mostly based on translating terms in Homosaurus. Among its 914 terms, 774 exhibit either an exact match (`skos:exactMatch`) or close match `skos:closeMatch` to terms in Homosaurus, with an additional 140 terms not mapped to terms in Homosaurus, some of which are exclusive to

[11] Version 1 is no longer available on the official website. v2.1, v2.2, v3.0, v3.1, and v3.2 are no longer available on their website.

[12] In this paper, we use the prefix `h2` for the namespace `http://homosaurus.org/v2/` and `h3` for the namespace `https://homosaurus.org/v3/`.

[13] The N-Triple file of LCSH was obtained on 9th May, 2024 from `https://id.loc.gov/authorities/subjects.html`. For fast analysis of its entities, it was converted to its HDT format. Both were included in the supplementary material. We use the prefix `lcsh` for the namespace `http://id.loc.gov/authorities/subjects/`.

[14] We use `qlit` for the namespace `https://queerlit.dh.gu.se/qlit/v1/`.

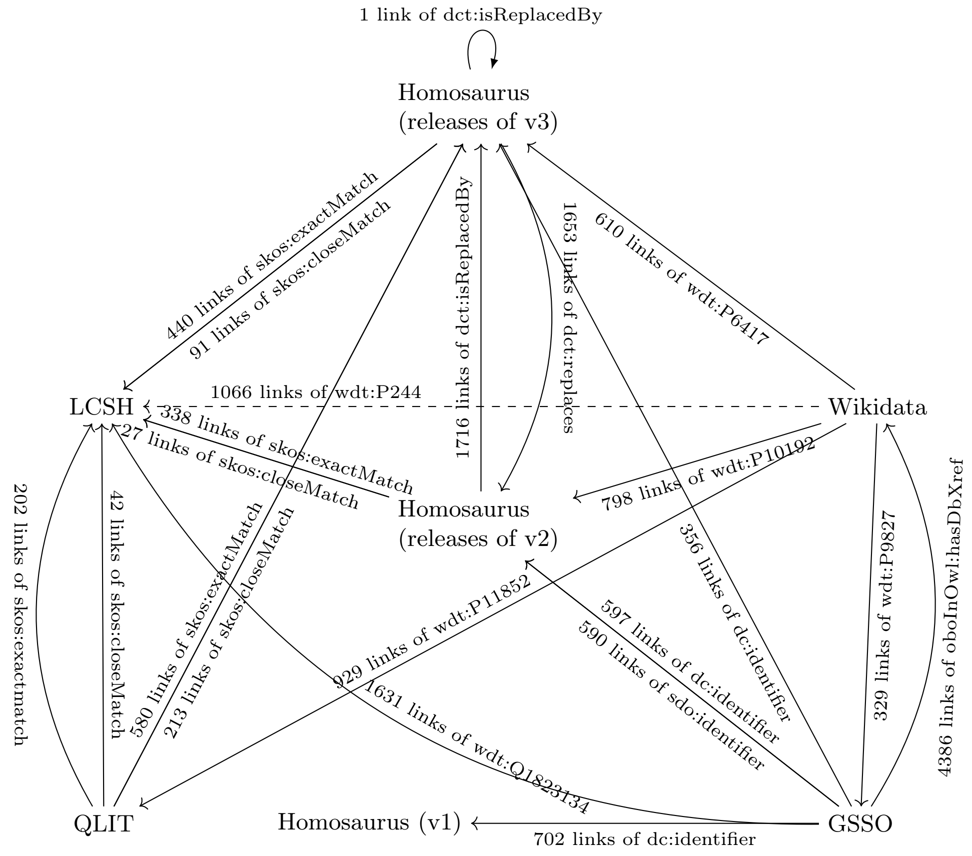
1 link of dct:isReplacedBy
Homosaurus (releases of v3)
440 links of skos:exactMatch
91 links of skos:closeMatch
1716 links of dct:isReplacedBy
1653 links of dct:replaces
610 links of wdt:P6417
LCSH
1066 links of wdt:P244
Wikidata
338 links of skos:exactMatch
27 links of skos:closeMatch
Homosaurus (releases of v2)
798 links of wdt:P10192
202 links of skos:exactmatch
42 links of skos:closeMatch
580 links of skos:exactMatch
213 links of skos:closeMatch
929 links of wdt:P11852
1631 links of wdt:Q1823134
356 links of dc:identifier
597 links of dc:identifier
590 links of sdo:identifier
329 links of wdt:P9827
4386 links of oboInOwl:hasDbXref
QLIT
Homosaurus (v1)
702 links of dc:identifier
GSSO

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.

QLIT. Only 244 links were found when examined against LCSH. Missing links will be discussed in Section 4.1. Moreover, its scope is limited to LGBTQI instead of LGBTQ+, thus the translation misses '+' in its Swedish labels.

Next, we extract links between GSSO and Homosaurus. It was noticed that it uses both `sdo:identifier` and `dc:identifier`. Since its publication in September 2022, the links from GSSO to Homosaurus version 2.2 and version 3.1 have not been updated. GSSO has 597 links of `dc:identifier` and 590 `sdo:identifier` to version 2 of Homosaurus. In comparison, there are only 356 `dc:identifier` links to version 3. Moreover, these links are using version 2.2 and version 3.1, which are outdated. This is because GSSO has not be updated since September 2022. Moreover, GSSO has 1,830 links to LCSH. The links of Wikidata and the two versions of Homosaurus are mostly asserted and maintained by experts and members of the Wikidata community. This demands significant human labor.

Links from Wikidata[15] to Homosaurus were provided using specified relations: `wdt:P10192` for entities in version 3 and `wdt:P6417` for entities in version 2. There are 610 and 798 links between Wikidata and versions 2 and 3 of Homosaurus, respectively. Similarly, for links from Wikidata to GSSO, a specific relation `wdt:P9827` was used. Only 329 links were found. As far as the authors know, links from Wikidata to GSSO and Homosaurus are maintained by hand by members of the Wikidata community without any use of automation. In total, 929 links were found for entities in QLIT, corresponding to the Wikidata property `wdt:P11852`. A total of 55,980 links were found between Wikidata and LCSH. Given that we study only LGBTQ+-related concepts. We restrict the entities to only those that appear in the links between conceptual models. Thus, only 1,066 links from Wikidata to LCSH were to be integrated and studied in the next steps. We obtain the complete URIs for the entities GSSO, Homosaurus v2 and v3, LCSH, and QLIT.[16]

3.2 Multilingual Information Extraction

GSSO consists of labels of 77 languages, while entities of Wikidata in the integrated graph (see Section 3.4) are associated with 507 languages. For the study of reuse of multilingual information to be presented in Section 4.3, we extract also multilingual information in GSSO[17] regarding labels (`rdfs:label`), paradigmatic synonyms (using `oboinowl:hasSynonym`, `oboinowl:hasExactSynonym`, and `oboinowl:hasRelatedSynonym`), short names (`wdt:P1813`, about “short name”), (`wdt:P5191`, about “derived from lexeme”), replaces (`dct:replaces`), `sdo:alternateName`, and `owl:annotatedTarget` in all its languages. Similarly, 595,167 multilingual labels using `rdfs:label` and `skos:altLabel` about the entities in the integrated graph were extracted from Wikidata.

3.3 Redirection

Our analysis showed that Homosaurus switched its protocol from HTTP to HTTPS. Thus, 1,738 URIs were redirected to their HTTPS equivalent in version 2. No redirection was found from version 2 to 3. Another 63[18] redirec-

[15] We use the prefix `wdt` for the namespace `http://www.wikidata.org/prop/direct/` and `wd` for the namespace `http://www.wikidata.org/entity/`.

[16] Using the Wikidata SPARQL endpoint (`https://query.wikidata.org/sparql`), we can obtain the corresponding identifiers of Homosaurus, QLIT, and GSSO. To obtain the full URI, we process these identifiers using the “frommatter” as specified on their pages. Take GSSO for example, 002171 is the identifier. Using the formmater `http://purl.obolibrary.org/obo/GSSO_$1`, we can replace the place-holder and get the full URI: `http://purl.obolibrary.org/obo/GSSO_002171`. In this paper, we use the prefix `obo` for the namespace `http://purl.obolibrary.org/obo/`. The preparation of Wikidata and links was done between 5th May and 8th May, 2024.

[17] We use the prefix `oboinowl` for the namespace `http://www.geneontology.org/formats/oboInOwl#` and `sdo` for the namespace `https://schema.org/`

[18] We include the entities that no longer exist in the latest version of Homosaurus but were still referenced in GSSO.

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.

tions were found in version 3.[19] None was covered by existing replace relations (`dct:replaces` or `dct:isReplacedBy`). Among them, 61 Homosaurus entities could be from outdated release(s) of Homosaurus and were redirected to entities in the latest release (v3). Only 2 redirections were found between entities in Homosaurus (v3). Moreover, we noticed that some URIs in version 2 cannot be resolved anymore, such as `h2:aromantic`. We use the redirection relation `https://krr.triply.cc/krr/metalink/def/redirectedTo` (`meta:redirectedTo` in short) [20].

3.4 Integrating Extracted Links

Table 1 presents the components of the integrated graph. We noticed a mistake that, for links from GSSO to LCSH, `wd:Q1823134` (representing the LCSH controlled vocabulary, rather than a property about links to their identifiers) was mistakenly used as a property. For consistency with the representation elsewhere, we change it to `wdt:P244`. When examining its links to Wikidata, it was also observed that GSSO has the URLs of webpages rather than the entities in Wikidata.[20] For example, `http://www.wikidata.org/entity/Q190845` should not have been used as `https://www.wikidata.org/wiki/Q190845` in its published data. We have corrected this in the integrated dataset. We further

[19] All redirect relations were obtained using the *webdriver* of the *selenium* Python package (`https://selenium-python.readthedocs.io/`) on 7th May, 2024.

[20] See `https://www.wikidata.org/wiki/Wikidata:Identifiers` for details.

double-checked that all the 2,085 LCSH entities in the integrated file are in the latest version of LCSH except one due to a suffix of '.html' from GSSO, which was corrected in the integrated file. The integrated graph involves 19,200 entities with 17,812 links. Its N-Triple file is 2.4MB. Considering only entities with the above-mentioned links are in the integrated graph, no singleton is present.

3.5 Clustering

We compute its *Weakly Connected Component* (WCC) for our integrated directed graph. A WCC of a graph is a subgraph with maximal entities where there is a path between any of its entities regardless of the direction of edges. In this paper, WCCs are clusters of entities that mostly share a similar or related meaning. For our integrated graph, there are 6,406 WCCs. Figure 2 shows that the largest four WCCs consist of 45, 36, 36, and 35 entities respectively. The largest one consists of 12 entities from Wikidata, 6 from GSSO, 5 from QLIT, 7 from Homosaurus v2, 6 from Homosaurus v3, 4 from LCSH, etc. More specifically, it involves related concepts about "human sexuality" (e.g. `wd:Q154136`), "Sexual intercourse" (e.g. `h3:homoit0000662` and `lcsh:sh85120739`), "Sex (Act)" (e.g. `h3:homoit0001267`), "fucking" (e.g. `h2:fucking`), "gender"(`h2:gender`), and "sexuality" (e.g `wd:Q3043188`). This example shows how the ambiguity of these entities accumulates into bigger clusters. This is in line with the intuition that larger weakly connected components may have more potential errors. The size of these clusters is within the capability of manual revision, but the number of these WCCs remains significant.

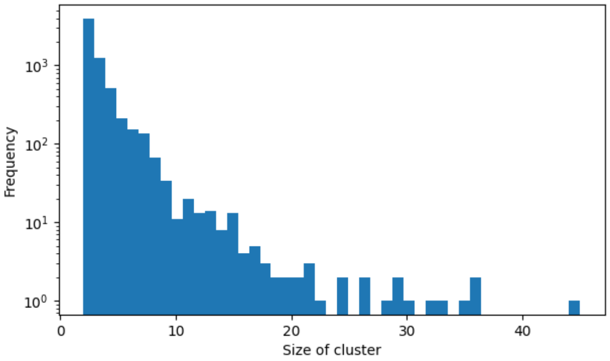

Fig. 2. Frequency histogram of the size of clusters

4 Evaluation

In this part, we showcase the integrated data in application scenarios guided by community requirements for practical evaluation. We assess the data's usefulness by showcasing three scenarios, illustrating how they can facilitate the tasks developers in the community encounter to ease the maintenance of their conceptual models and their links. More specifically, we demonstrate how our data can ease the maintenance of links by automatically detecting missing and outdated links in Section 4.1. Given that these conceptual models are developed for different purposes, we do not study which terms should be included or excluded.

4.1 Scenario 1: Link Discovery

Relying on the WCCs, we propose to discover missing links between two conceptual models. The intuition is that, if entities from different concept models are in the same WCC and they are unique of its conceptual model in the WCC, they are likely to refer to the same or related things. A link could be added after manual examination. In practice, for the case of Homosaurus, we take into account exceptions such as redirection, replacement, and entities no longer maintained. Next, we report our findings for Homosaurus v3 and QLIT.

Only 531 links were observed between Homosaurus v3 and LCSH. However, 2,085 LCSH entities were found in the integrated graph.[21] Using the method mentioned above, 25 links were found. These discovered links have been submitted to the experts in Homosaurus for consideration before the next release.

Similarly, QLIT has only 244 links to LCSH. Using the same method, we found 105 potential missing links between QLIT and LCSH, which require further manual revision by Swedish-speaking experts. Given that some entities were redirected in Homosaurus v3, we also found one outdated link and its corresponding entity in the latest Homosaurus v3 (see `qlit:oj77yj15` in Figure 3).

4.2 Scenario 2: Concept Drift and Change

It has been addressed in related work that existing measures do not apply to this multilingual case. Furthermore, the evolution of concepts was captured by different conceptual models at various levels, resulting in a complex co-evolution. The weakly connected components can be used for manual examination of drift of concepts and understanding of ambiguity. Next, we present three scenarios for concept convergence, ambiguity, and scope change, respectively.

First, we use an example in Figure 3 to demonstrate how missing identity links, redirection relations, evolving concepts, as well as no longer maintained URIs can result in ambiguity and difficulty in identifying erroneous links. It was observed that `h3:homoit00442` replaced `h2:fetishism` and is the target of a few redirected URIs, including `h3:homoit0000102`, which was linked by many. It could be that `h3:homoit00442` is a merge of the concept of 'BDSM' and 'fetishism'. As a result, two clusters of entities about BDSM and fetish/kinks are in the same connected component. Further examination shows that, 'BDSM' is now an alternative label (`skos:altLabel`) for `h3:homoit00442` in the latest

[21] This little overlap of concepts has been considered evidence by many that Homosaurus can be used as a complementary conceptual model of LCSH.

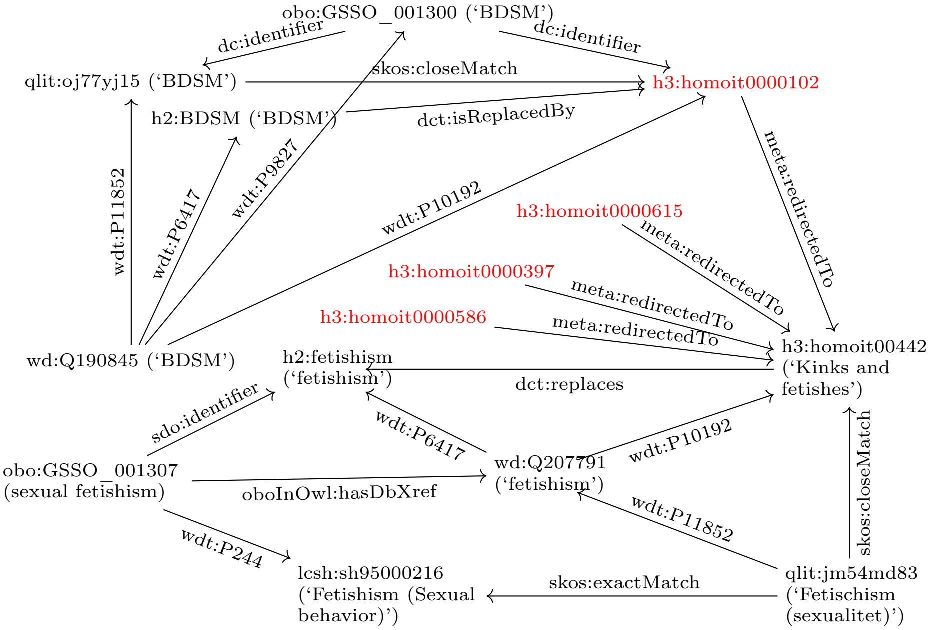

Fig. 3. A subgraph of the seventh largest weakly connected component with 30 entities and 44 edges including concepts related to BDSM and sexual fetish. Labels that can be found are included. Some links and entities were omitted for clear visualisation. Entities in red are no longer in the latest version of Homosaurus.

version. It was also noticed that a few URIs (highlighted in red) no longer appear in the latest version of Homosaurus, which leads to missing label information and relations. This example shows how multiple parties can have different views on concepts as conceptual models develop and the consequences of such changes.

In practice, it was observed that concept drift is often coupled with version changes, errors, and ambiguity, leading to a complicated network that requires careful manual evaluation. Next, we showcase the complexity by focusing on entities in GSSO and Homosaurus. Figure 4 is an example of concept drift involves the GSSO term "3,4-methylenedioxymethamphetamine" with the URI `obo:CHEBI_1391`, linking to the Homosaurus version 2 term "MDMA" with the URI `h2:MDMA`, which is replaced by the URI `h3:homoit0000388` in Homosaurus version 3. However, the URI no longer exists in the latest version of Homosaurus, and was automatically redirected to the identifier `h3:homoit0000380`, which corresponds to the term "Substance use in LGBTQ+ communities" with a few other labels (`skos:altLabel`) including "Drug use (LGBTQ)", "Alcohol use in LGBTQ+ communities", etc. Therefore, the term "3,4-methylenedioxymethamphetamine" from GSSO is the same cluster as the term "Substance use in LGBTQ+ communities" in Homosaurus version 3, which can be inaccurate and misleading. MDMA is an abbreviated form of methylenedioxymethamphetamine

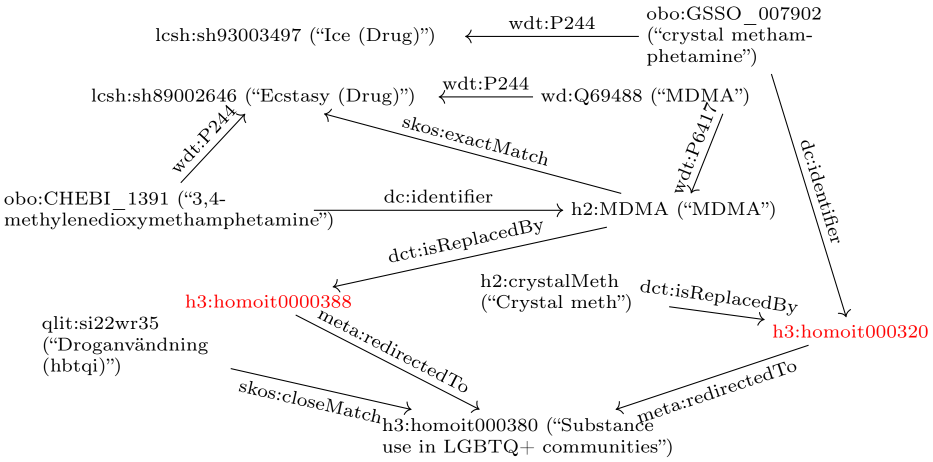
lcsh:sh93003497 ("Ice (Drug)")
wdt:P244
obo:GSSO_007902 ("crystal methamphetamine")
lcsh:sh89002646 ("Ecstasy (Drug)")
wdt:P244
wd:Q69488 ("MDMA")
wdt:P244
skos:exactMatch
wdt:P6417
dc:identifier
obo:CHEBI_1391 ("3,4-methylenedioxymethamphetamine")
dc:identifier
h2:MDMA ("MDMA")
dct:isReplacedBy
h3:homoit0000388
h2:crystalMeth ("Crystal meth")
dct:isReplacedBy
h3:homoit000320
qlit:si22wr35 ("Droganvändning (hbtqi)")
meta:redirectedTo
skos:closeMatch
h3:homoit000380 ("Substance use in LGBTQ+ communities")
meta:redirectedTo

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.

and is the main component of the popular party drug ecstasy. Crystal methamphetamine (a.k.a. Ice) is a different drug. Figure 4 illustrates the complexity and ambiguity when considering your identity by taking into account all related entities and their interconnected links. Moreover, when updating outdated links in conceptual models relying on translated Homosaurus terms, if such information wwere used, the subsequent conceptual model would inherit this problem. For example, in the case above, `h3:homoit0000380` has a `skos:closeMatch` link to the QLIT term "Droganvändning (HBTQI)" ("Drug use (LGBTQ+)" in English) with the identifier `qlit:si22wr35`. Should a connection be established between this Swedish term in QLIT and GSSO through Homosaurus, the established link would be problematic and inaccurate, and following these links would result in confusion and incorrect labels.

Next, we show an example of the change in scope of concepts using Figure 5. In GSSO, the term "being in love" with identifier `obo:GSSO_007692` has note "The state in which a person is when they are in love.". Its scope is not limited to the LGBTQ+ community. It was linked to `h2:beingInLove` ("Being in love"), which is a broader (`skos:broader`) term than "LGBTQ Love". Moreover, this entity in GSSO was linked to `h3:homoit0000107`, which was redirected to `h3:homoit0000894` ("LGBTQ+ love"). Meanwhile, it was noticed that `v2:LGBTQLove` was replaced by `h3:homoit0000894`. This convergence leads to a change in scope.

4.3 Scenario 3: Reusing Multilingual Information

Although Homosaurus is only available in English and Dutch, the demand for reliable multilingual conceptual models remains. As illustrated above, entities in

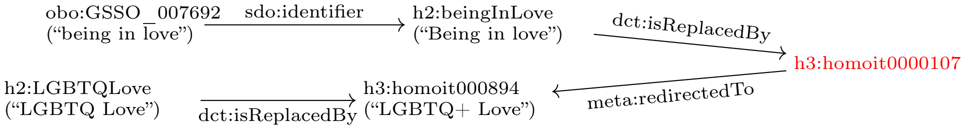

Fig. 5. Convergence of "Being in love" and "LGBTQ love" to "LGBTQ+ love". Following the links could lead to a change in scope. Highlighted in red is an entity no longer maintained in Homosaurus v3. Not all entities and links in the WCC were illustrated.

large WCCs involving multiple entities from the same source can face problems such as concept drift and ambiguity. Here, we study the reuse of multilingual labels for entities not involved in WCCs where there is a "one-to-one mapping" to entities of Homosaurus v3. More specifically, for GSSO, we study how much information can be reused regarding a total of nine relations on labels, synonyms, and alternative labels (see Section 3.2). We compare it with Wikidata where multilingual information is being provided by `skos:altLabel` and `rdfs:label`.

There are 1,779 GSSO entities in the integrated graph. 1,681 are the unique entity of GSSO in the WCC. When further restricting to exactly one entity of Homosaurus v3 in the WCC (with consideration of redirection and replacement), only 48 entities remain. The three languages with the most labels are English (356 labels for 48 entities), Danish (26 labels for 9 entities), and French (24 labels for 9 entities). The corresponding labels per entity are 7.42, 2.89, and 2.67 for English, Danish, and French, respectively.

As for Wikidata, there are 5,769 entities in the integrated graph, among which 4,939 are unique in their WCCs. When further restricting to the correspondence of exactly one entity in Homosaurus v3, only 429 entities remain. The four languages with the most labels are English (1,692 labels for 429 entities), Spanish (951 labels for 333 entities), Chinese (893 labels for 287 entities), and Portuguese (881 labels for 299 entities). The corresponding labels per entity are 3,94, 2.86, 3.11, and 2.95 for English, Spanish, Chinese, and Portuguese, respectively.

It was noticed that the number of entities that has a one-to-one mapping between GSSO and Homosaurus v3 is significantly smaller than that of Wikidata. Despite that the entities take more relations into consideration when retrieving multilingual labels from GSSO, the number of multilingual labels is remarkably larger for Wikidata (with the exception that GSSO can provide more labels in English per entity in this setting). The average number of labels is 2.56 per entity for the top 20 languages with the most labels, with an average of 268.7 entities per language for Wikidata. This indicates that Wikidata can be a better choice when considering reusing its multilingual labels to enrich Homosaurus with manual examination. Nevertheless, the these multilingual labels as suggestions cannot be directly used but remain to be assessed after manual revisions by experts for each language. Take `h3:homoit0000295` ("Coming out") for example, it has label "sortie du placard" using `rdfs:label` in GSSO. Moreover, "sortir du placard" and "coming out" are synonyms for relation `oboInOwl:hasRelatedSynonym` and `oboInOwl:hasSynonym`, with three labels using `owl:annotatedTarget`: "sortir du placard", "coming out", and "sortie du placard". The labels retrieved as suggestion from Wikidata are similar: there is a label of "coming out" using `rdfs:label`; some more labels using `skos:altLabel`: "coming-out", "sortie du placard", "sortie de placard", and "sortir du placard". This shows that GSSO and Wikidata have overlaps in the labels they provide. Ultimately, it is the responsibility of the developers of conceptual models to determine the preferred label, the alternative label(s), and recognize incorrect ones.

Similarly, we can extract labels from Wikidata as suggestions for QLIT. There are 914 entities with one prefLabel each but only a total of a total of 480 altLabels. Using the method above, we extracted 775 labels in Wikidata (524 prefLabels and 251 altLabels) for 524 entities. It was noticed that, in many cases, the difference is minor, either in the upper/lower case of the first character or the upper/lower case of 'hbtqi' (the Swedish word for LGBTQI). It remains a question if Wikidata has taken advantage of QLIT for its entries, or these terms are likely to be free from concept change and ambiguity due to the restriction of exactly one WCC.

5 Discussion

In Section 4.1, we demonstrated how we can find missing links using the WCCs. A further manual examination found some correspondence of terms between QLIT and Homosaurus. For example, a link between `qlit:tm80vg73` ("Pappor till homosexuella") and `h3:homoit0000427` ("Fathers of queer people") could be included. Similarly, `qlit:iq08ee58` ("Masters (hbtqi)") could be linked to `h3:homoit0000999` ("LGBTQ+ dominants"). These missing links can lead to discrepancies between conceptual models if not fixed before new versions of QLIT are released. Our proposed WCC-based method could reduce the effort of manually finding links between conceptual models. However, our approach is limited to the entities that are linked to at least one other entity. Moreover, assessing these links requires considerable manual effort. All the links in Section 4.1 have been submitted to the corresponding communities for manual revision by experts. For this reason, at the time of submission, the quality of newly found links in our approach remains unknown. This is also the case for our use case 3. Given that drift and change in the concept are mixed with ambiguity and errors, it is also difficult to evaluate the output in Section 4.2 where a significant amount of work is required for coordination between subcommunities.

In our study, `dct:replaces` and `dct:isReplacedBy` were included in the integrated graph, despite not necessarily implying equivalence relations. The values lies in the study of the dynamics and evolution of entities in Homosaurus and the impact on other linked entities. It was noticed that the concept change in Homosaurus is partially reflected by such links. Concept drift and change can result in potential duplicate terms (see Homosaurus terms in Figure 3 for example) that could violate the Unique Name Assumption [20]. This requires further manual examination for each concept model. Using the relation `owl:differentFrom` for different entities can ease future automated examination.

Section 4.3 showed that GSSO cannot suggest as many labels as Wikidata. This could be due to the restriction considering WCC. A further experiment with a relaxed condition only considering redirection and replacement shows that there are suggesting labels for 115 entities for Danish and 47 for French.

6 Conclusion and Future Work

In this paper, we studied the properties of LGBTQ+-related concepts and their links in the semantic web. With the links extracted, we constructed an integrated graph and evaluated its use in three scenarios. We illustrated how our data can be used to find missing links and outdated links. We addressed the issue of concept drift and demonstrated how WCCs can assist the community in easily locating entities with potential issues. Finally, we showcased the reuse of multilingual information for Homosaurus. Our findings indicate that Wikidata offers a substantially greater number of multilingual labels than GSSO.

In practice, experts can overlook the implications of diverging or converging concepts. This paper demonstrates how we can provide such insight to the specialists. If any results require the intervention of others, discussion in the community would be advantageous. Our code could be reused for automatic detection of outdated links in the future. Given the significant amount of manual work, an interface that supports manual revision of the WCCs could be helpful.

Heterogeneous use of relations between entities, especially that for mappings was observed (see Table 1). It could benefit the community, especially for interoperability, if they consider adapting a common FAIR Implementation Profile, a structured representation of the community's decision on knowledge representation languages, semantic models, metadata, etc [19].

Mistakes identified in GSSO should be corrected in upcoming versions (see Section 3.4). The 63 redirection links in version 3 presented in Section 3.3 remain to be checked to ensure that they maintain the original semantics [12]. If correct, they can be included (e.g. using the replacement relation) in the future release of Homosaurus. Some other conceptual models in the semantic web, such as LCDGT [5] and DBpedia [2], have been reported to contain some LGBTQ+ terms and could be included in some follow-up research together with the mappings from QLIT to some Swedish library thesauri [11]. Other links such as `skos:narrower`, `skos:broader`, and `skos:relatedMatch` can be explored.

Given the nature of the field, there is no perfect conceptual model [5]. As addressed in Section 2, there is no systematic report on the quality of links, the drift of concepts, and the potential harm of outdated links. Our data could serve to ease the manual work for this task. A good starting point is a list of historical terms (see `h3:homoit0000878`) and reclaimed terms (see `h3:homoit0001559`) in Homosaurus. Additional resources can be used as reference, for example, a list of selected terms about transgender and diversity in LCSH [1].

Acknowledgments. The authors used TeXGPT to rephrase some sentences through Writefull on Overleaf. The LaTeX code of Figure 1 was generated with the help of ChatGPT. Neither was used to generate complete sentences or paragraphs.

References

1. Trans & Gender Diverse LCSH (2024), `https://translcsh.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., Humelsjö, 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, `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., DuVall, 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., Kriström, O.: Building and serving the queerlit thesaurus as linked open data. Digital Humanities in the Nordic and Baltic Countries Publications **5**(1), 29–39 (2023)
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/lgtbq-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/queerlcsh`
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., Krötzsch, 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.K., Kuhn, T.: Fair implementation profiles for social science. In: Preceeding of 17th International Conference on Metadata and Semantics Research. Springer (2023)
20. Wang, S., Raad, J., Bloem, P., van Harmelen, F.: Refining large integrated identity graphs using the unique name assumption. In: European Semantic Web Conference. pp. 55–71. Springer (2023)
21. Watson, B.M.: "there was sex but no sexuality*" critical cataloging and the classification of asexuality in lcsh. Cataloging & Classification Quarterly **58**(6), 547–565 (2020)