Using controlled vocabularies to organise keywords in the SSH

Discover
Connect
Collaborate

Tomasz Umerle, Nikodem Wołczuk | **IBL PAN** | 2023-03-27 | Berlin, Germany

Triple

Organising keywords in the SSH

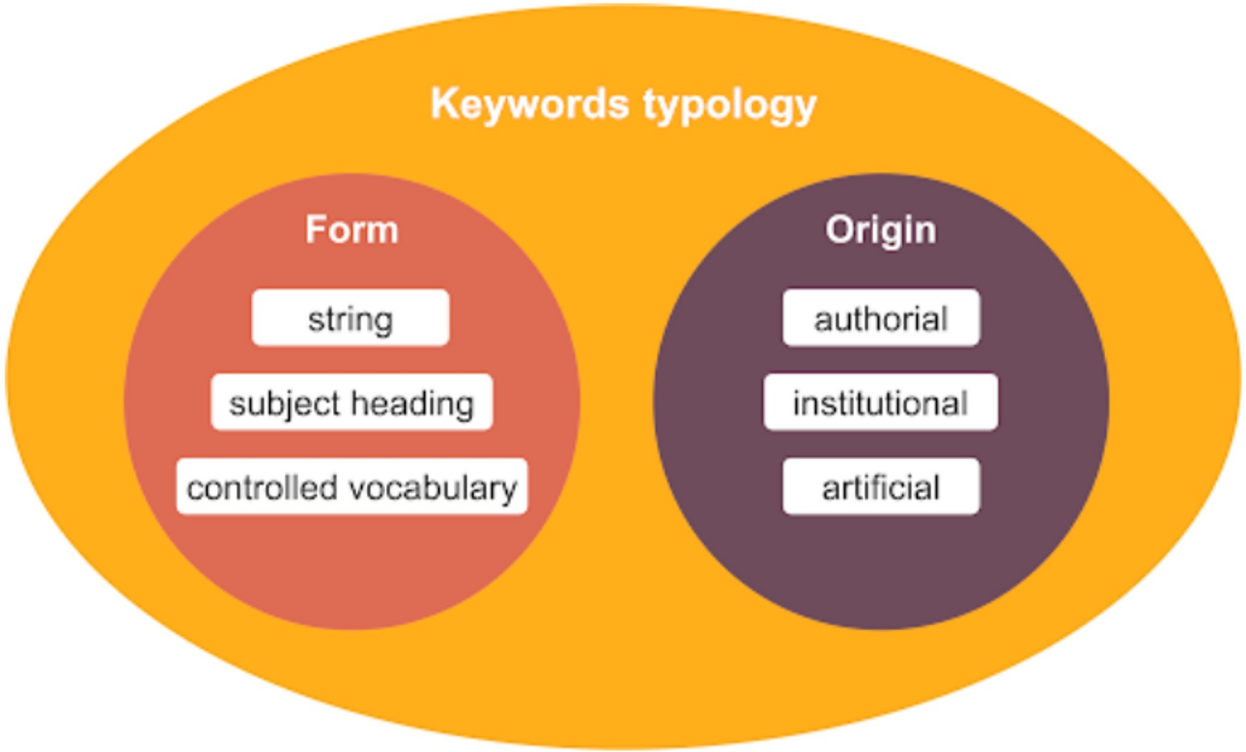
Subject description in SSH academic repositories, metadata aggregators or digital libraries is not necessarily based on attribution of entries from controlled vocabularies or this attribution is implemented improperly.

Although this hinders semantic interoperability, we should still treat existing subject descriptions as invaluable resources and aim to improve them.

Why *keywords*?

We use *keywords* as a broad term for short phrases describing topic of a document

Here we discuss the keywords which are uncontrolled strings (*keywords-strings*).



TRIPLE project contributed to these works which are now realised through DARIAH-PL's dariah.lab (ending in Dec 2023).

Organisational context

TRIPLE project contributed to these works, which are now realised through DARIAH-PL's dariah.lab (ending in Dec 2023).

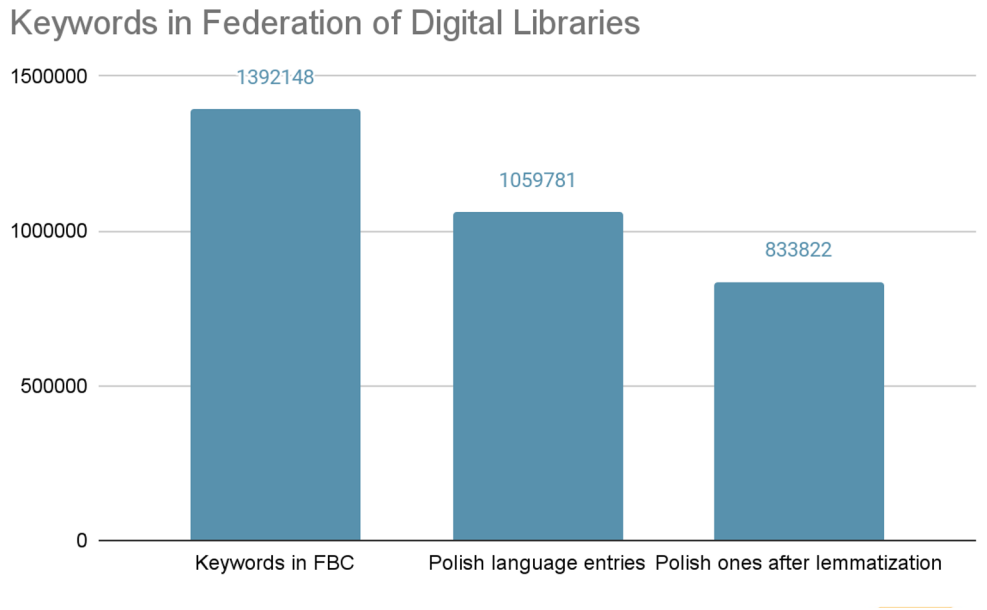
Discover Connect Collaborate

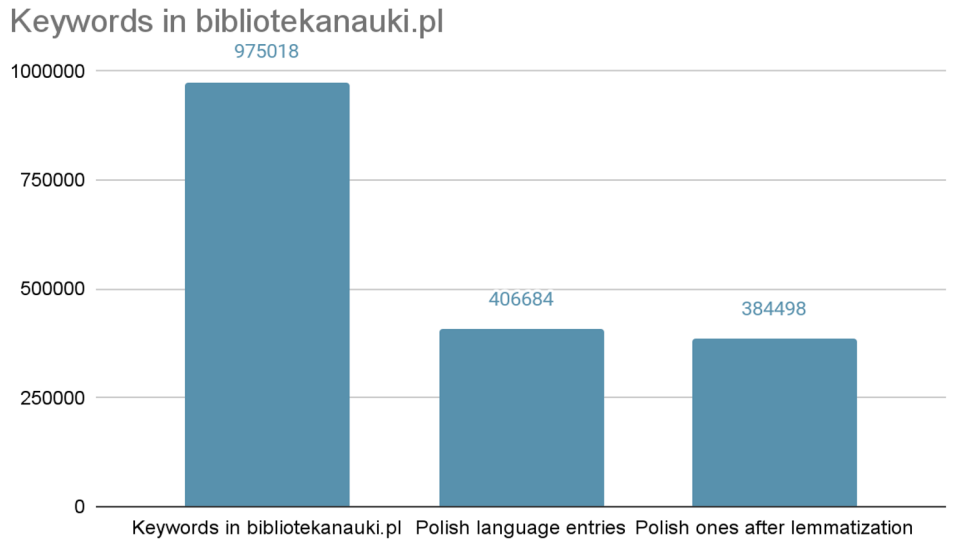
Keywords-strings are not FAIR

Main Polish SSH data providers:

Federation of Digital Libraries – more than 6,6 mln documents (cultural heritage documents)

Library of Science (bibliotekanauki.pl) – more than 0,5 mln documents (mostly scientific articles)





Why providers struggle with FAIR compliance?

- Limitations of software (e.g. local software does not support proper attribution of controlled vocabularies)
- Lack of know-how (it is not a priority to use controlled vocabularies)
- Limitations of data standards (e.g. DC-related challenges)

How to make keywords-strings more FAIR compliant?

Computations

Our mission in projects such as dariah.lab, TRIPLE or in planned future project is

expression of keywords-strings in the form of entries from controlled vocabularies

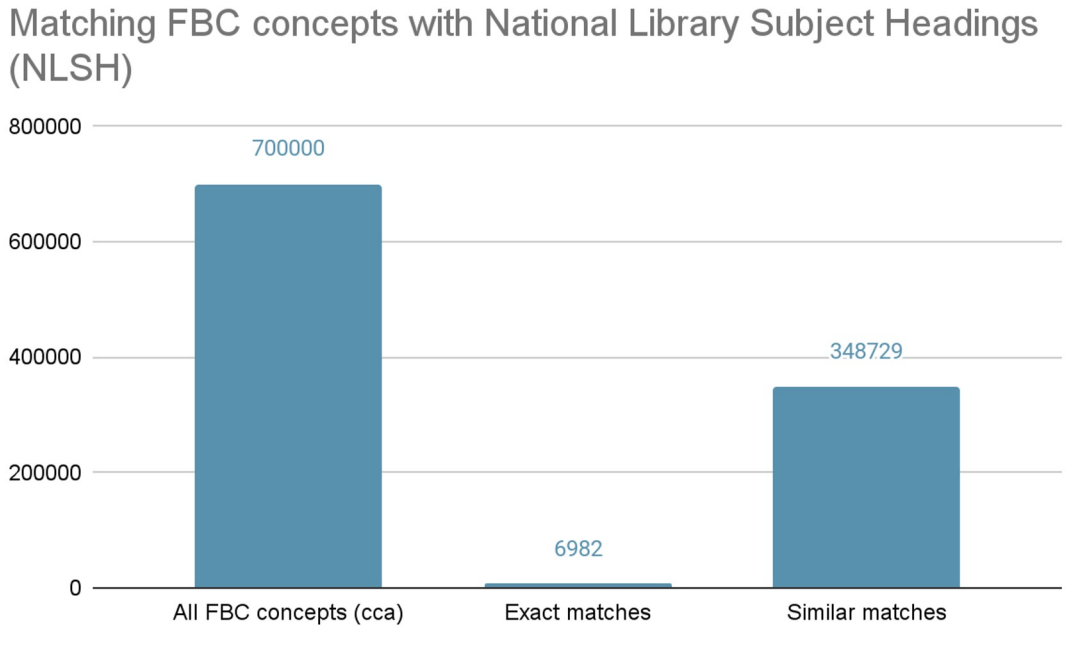
to foster semantic interoperability, data harmonisation and improvement of data quality.

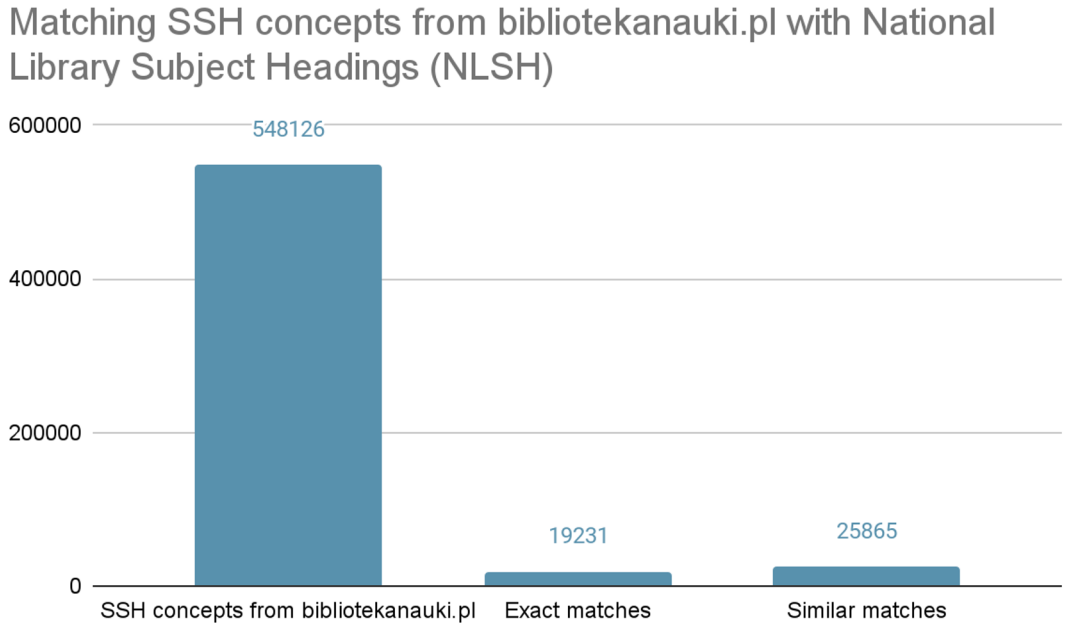
Discover Connect Collaborate

Triple

Hidden entries from controlled vocabularies

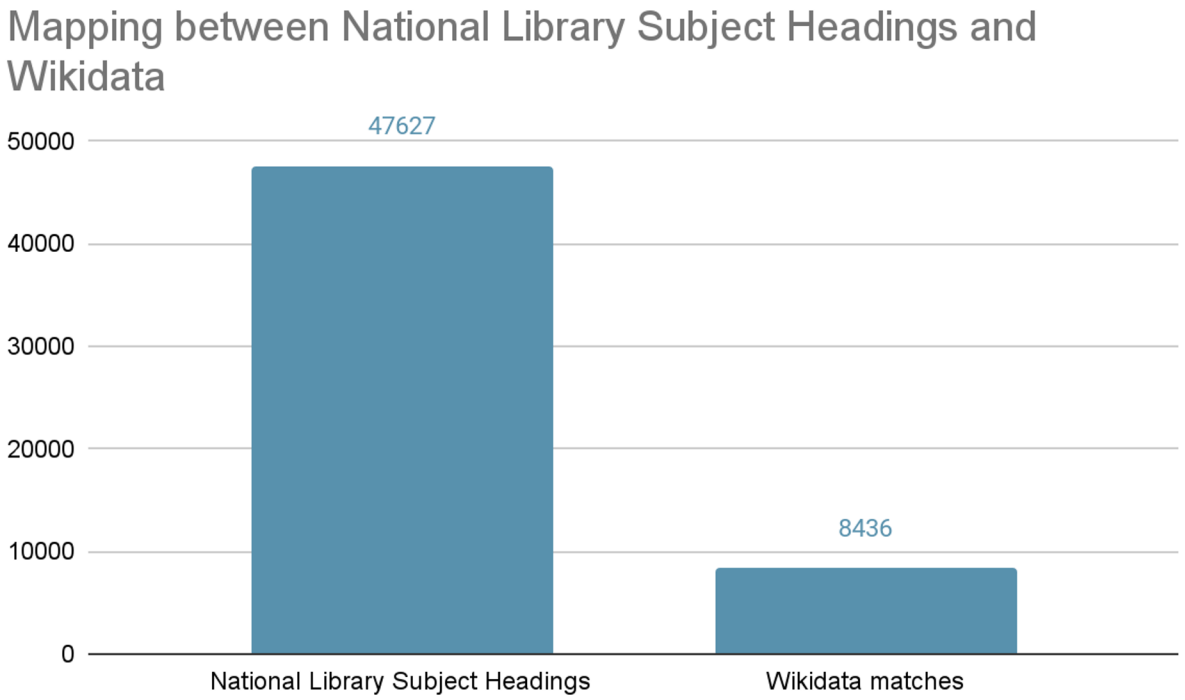
Our main finding – data providers are in fact using entries from controlled vocabularies, but in a non-FAIR compliant way.





National Library Subject Headings and Wikidata

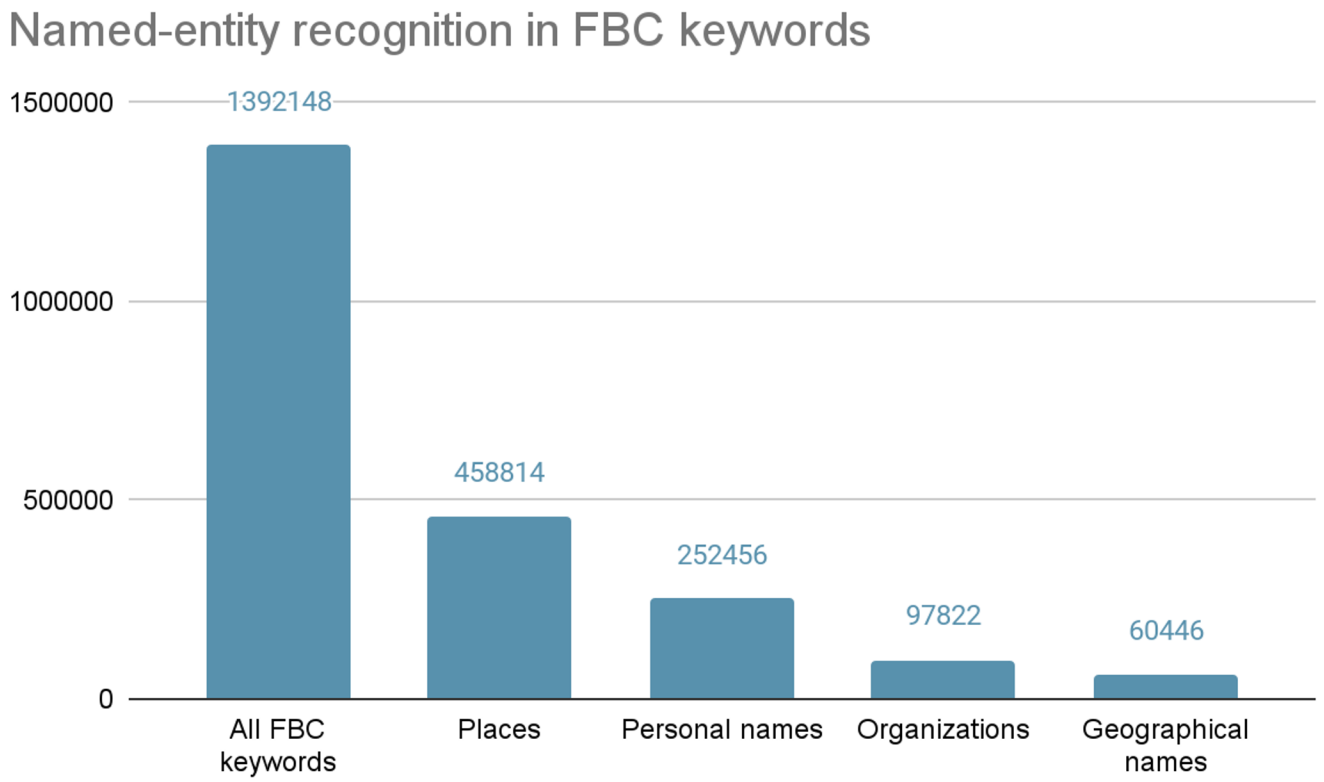
Mapping keywords onto NLSH fosters semantic interoperability as at least 18% of NLSH's entries could be mapped onto Wikidata.



Hidden entries from controlled vocabularies

Computations

Keywords include not only concepts, but also named entities which adds to harmonisation opportunities:



Discover Connect Collaborate

Triple

Expressing keywords-strings as entries in controlled vocabularies

Computations

We are metadata-focused, so our goal is not to assign a topic to a publication, but we aim to assign a discipline to the keyword (and ideally a more granular entry).

In short it means **using controlled vocabulary to harmonise keywords**.

Discover Connect Collaborate

Triple

Beyond uncovering hidden vocabularies

Bibliotekanauki.pl's journals have assigned disciplines.

However, these disciplines poorly correlate with keywords-strings' disciplines.

E.g. for random 3500 keywords for literary articles an average semantic similarity index with NLSH for literary discipline is 0,18 (according to spaCy).

Historia Slavorum Occidentis
2022/4(35)

CITATION EXPORT TO PBN

Pages: **11-32**

Main language of publication
Polish

Published **2022-12-31**

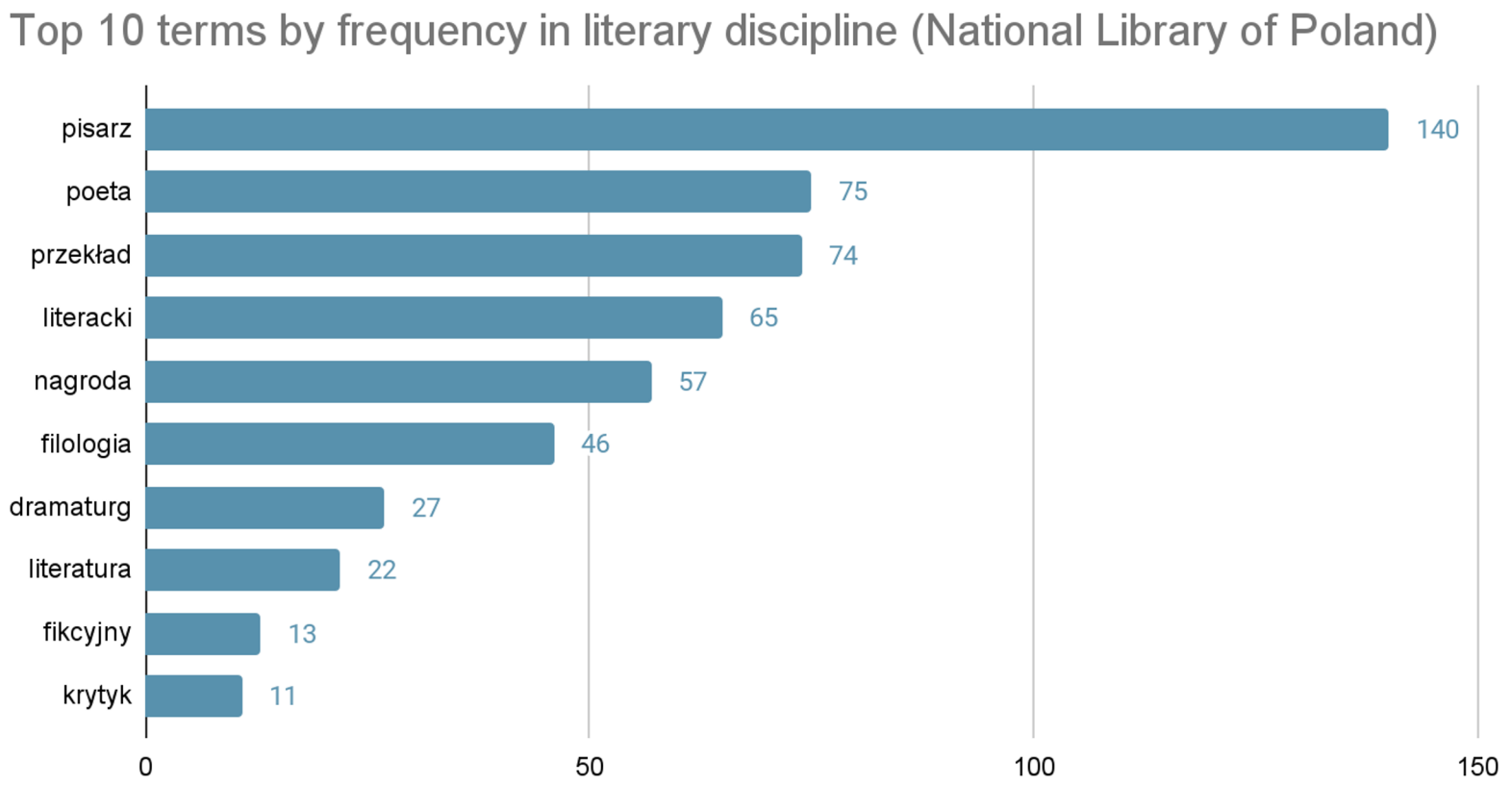
DOI: **10.15804/hso220401**

Humanities
literature

Vocabulary-based method

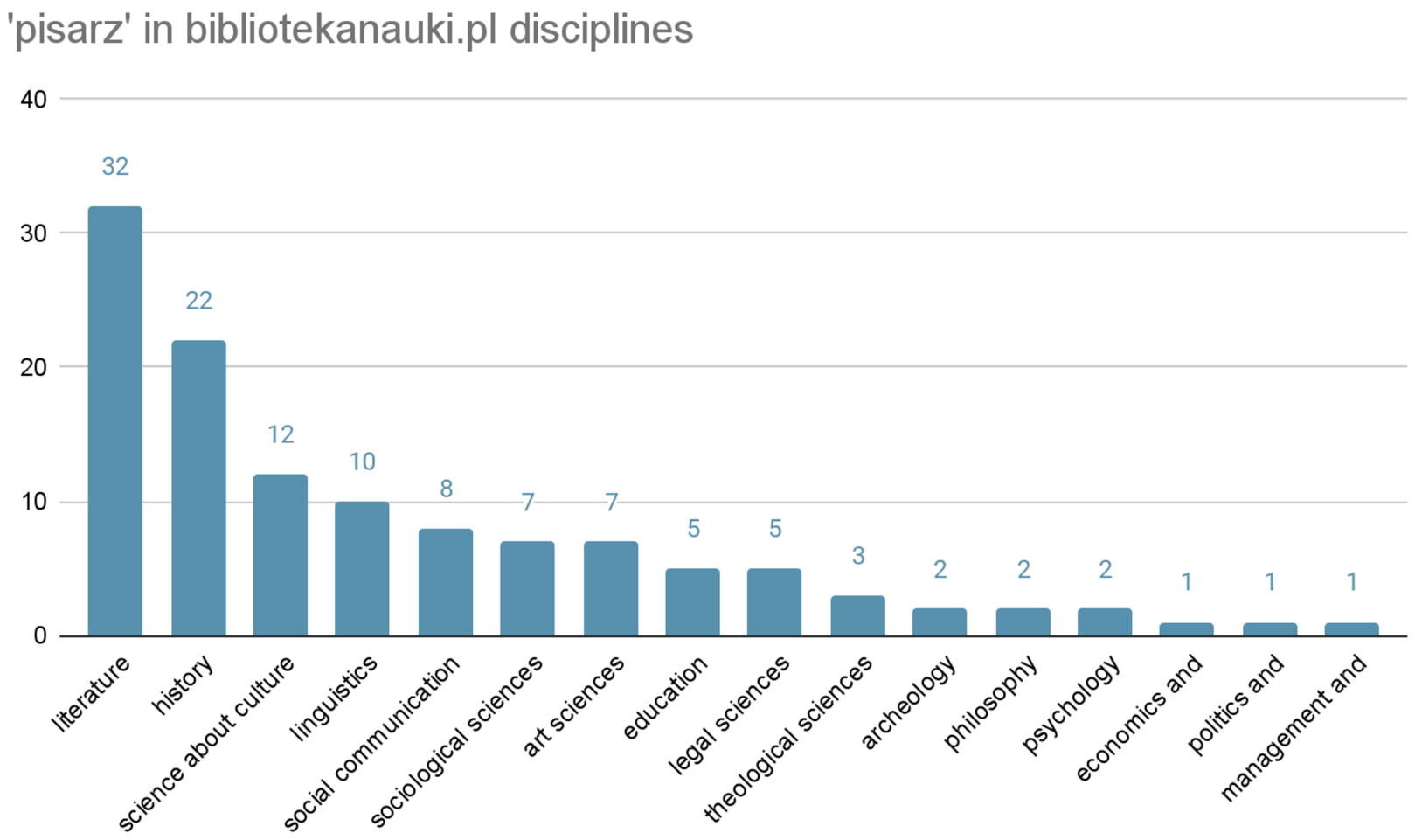
Vocabulary-based method for expressing keywords-strings in the form of entries from a controlled vocabulary relies partly on extracting most frequently used terms in the NLSH and mapping them onto keywords-strings.

Frequent literary terms in NLSH



Discover Connect Collaborate

Frequent literary terms in NLSH



Computations

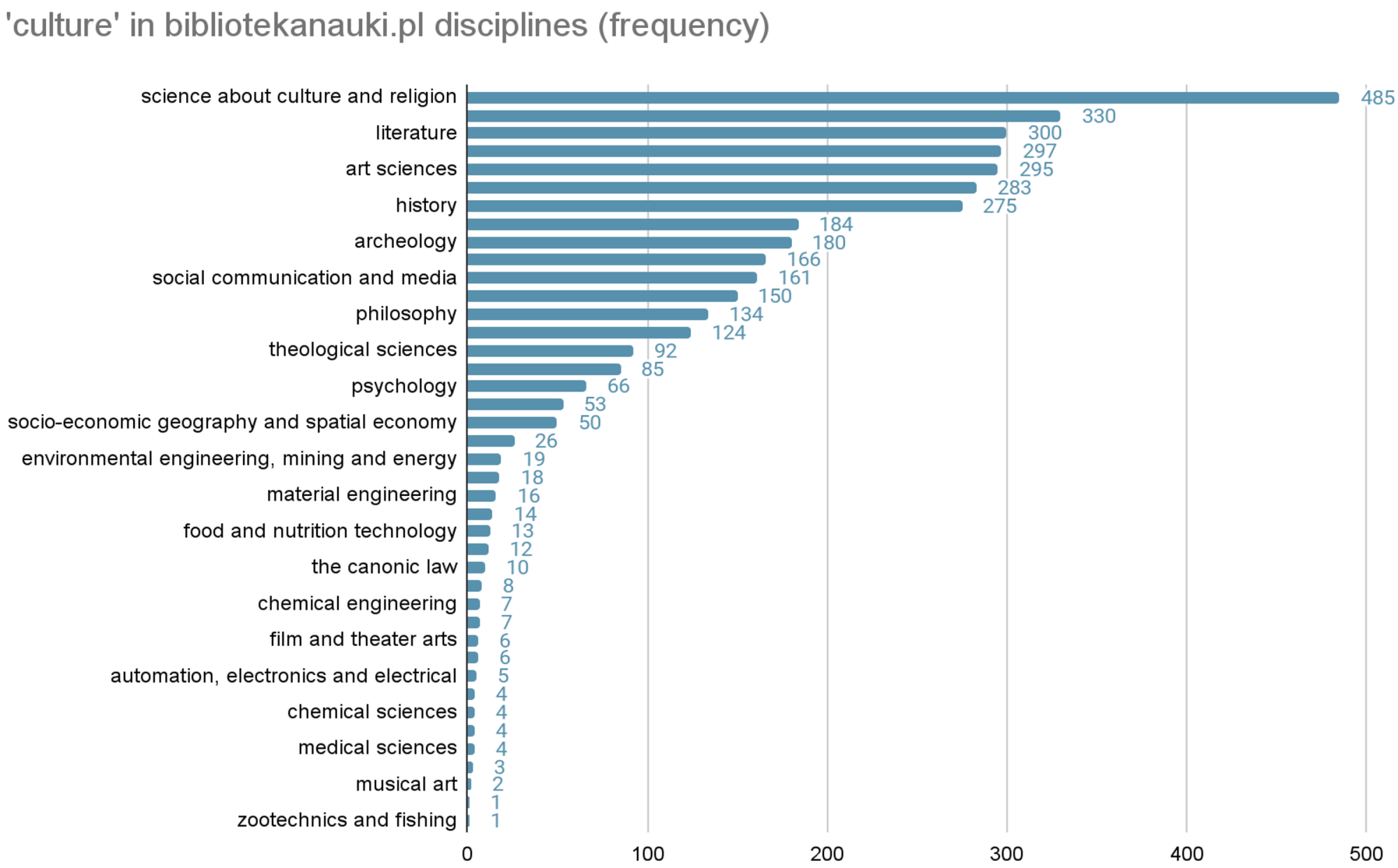
Discover Connect Collaborate

Triple

Bias of controlled vocabularies

Discover Connect Collaborate

Bias of controlled vocabularies



Discover Connect Collaborate

Using vocabularies to organise keywords

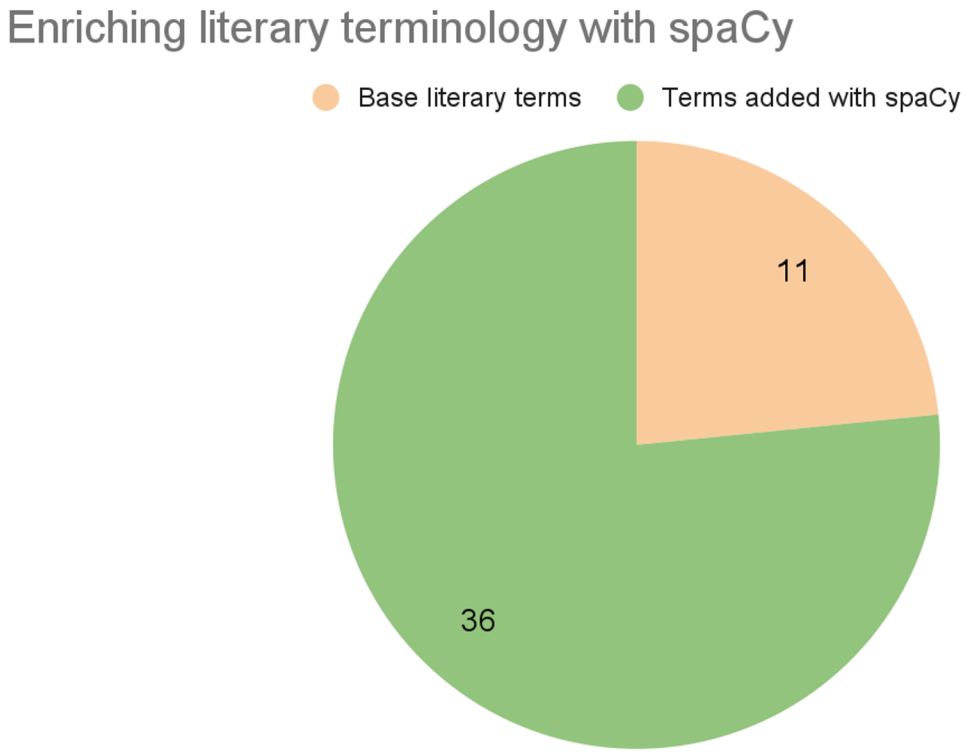
Computations

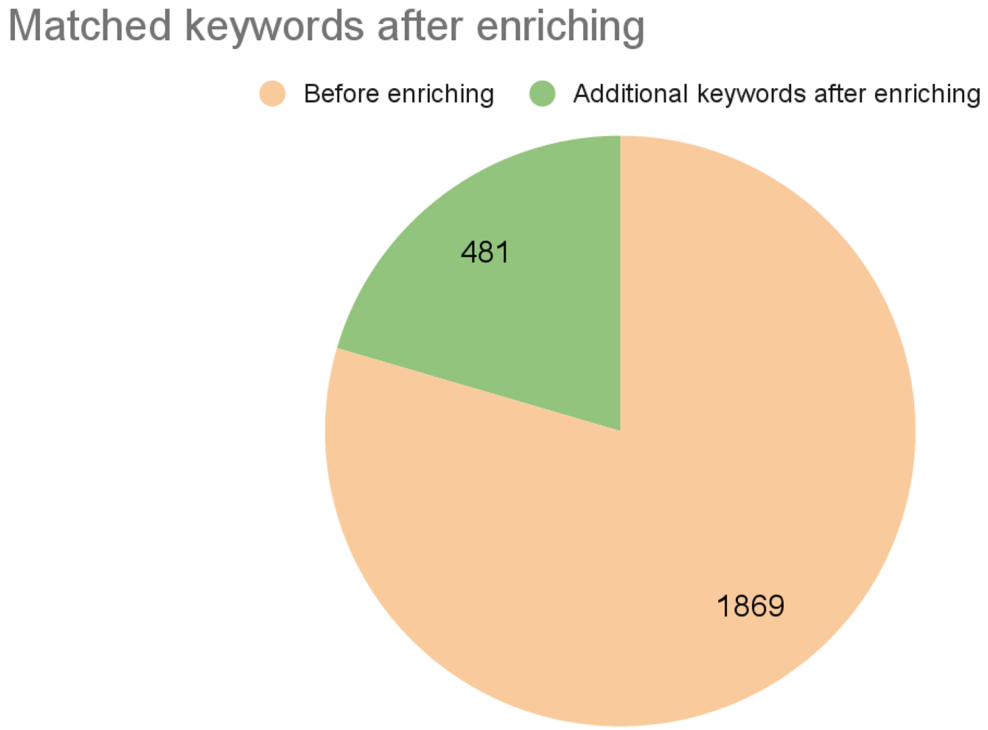
Vocabularies that are actually in frequent use (such as NLSH) – hence are attractive to be used for organising keywords-strings – could offer **limited semantic understanding** of domains or **might represent certain biased**.

Expanding concept understanding through language models

One of the ways to tackle vocabularies' limitations is through the use of language models to expand our understanding of domains.

Expanding concept understanding through language models





Conclusions

1. A lot of SSH content is *not* (properly) described through controlled vocabularies.

2. These descriptions are, however, valuable and improving their FAIR compliance aligns well with EIF and EU Data Strategy. We improve semantic interoperability while respecting contributions of other actors in the data “value chain”.

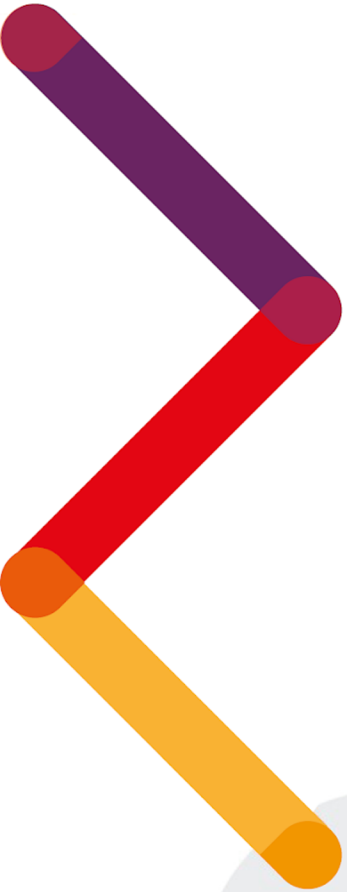
Conclusions

3. Using controlled vocabularies to make keywords-strings more FAIR is worthwhile especially if we use controlled vocabularies which are in frequent use (this adds to existing semantic interoperability).

4. Semantic interoperability in the field of improving quality of subject description demands collaboration between metadata experts and NLP experts. Many challenges are similar, but not exactly the same!

Challenges

5. More exciting opportunities are ahead, including new services development (semantic searching and fuzzy queries support), wider opportunities for international research and facilitating wider SSH data reuse for societal benefit.



Triple

Tomasz Umerle – tomasz.umerle@ibl.waw.pl

Nikodem Wołczuk – nikodem.wolczuk@ibl.waw.pl

visit us at: https://computations.eu

*The **GoTriple** platform will be the Discovery Service of the OPERAS Research Infrastructure.*

OPERAS
open access in the european research area through scholarly communication

The TRIPLE project has received funding from the European Union's Horizon 2020 Research & Innovation programme under grant agreement number 863420.