How to create a social sciences and humanities (SSH) vocabulary: The GoTriple Hackathon example

Discover
Connect
Collaborate

Iraklis Katsaloulis | **EKT**
Cezary Rosiński | **IBL-PAN**
23.11.2021 | 1st TRIPLE International Conference

Triple

Agenda

1. The need for a reliable SSH Vocabulary
2. Task 2.4 as a way to construct a SSH Vocabulary
3. The aims of the GoTriple Hackathon
4. Presentations of Hackathon's groups
5. An overview of the GoTriple Hackathon (difficulties, struggles, etc.)
6. Integration of the Hackathon to the GoTriple Vocabulary
7. Discussion

What is a Controlled Vocabulary?

- A controlled vocabulary is an organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching. It typically includes preferred and variant terms and has a defined scope or describes a specific domain.
- The purpose of controlled vocabularies is to organize information and to provide terminology to catalog and retrieve information.

The need for a reliable SSH Vocabulary

- SSH Researchers struggle for better access to resources relevant to their research, publications and data.
- In this effort reliable SSH vocabularies are essential:
  - They are necessary at the indexing phase because without them catalogers will not consistently use the same term to refer to the same concept, person, place, or thing.

The need for a reliable SSH Vocabulary

- In the retrieval process, various end users may use different synonyms or more generic terms to refer to a given concept. End users are often not specialists and thus need to be guided because they may not know the correct term.
- For SSH, multilingualism is an essential feature, so we are in a great need for multilingual vocabularies.
- The Triple Project will meet the needs of the SSH community in the respect of the SSH vocabulary.

The GoTriple Vocabulary

The GoTriple Vocabulary is going to be a vocabulary of subjects in SSH to be used by the annotation service of the GoTriple platform. In order for the annotation mechanism to be effective for publications of all the 9 languages that will be supported, the Vocabulary must contain a sufficient number of concepts and, at the same time, these concepts must have labels in as many of these languages as possible.

Classification/Annotation

- The aim of Task 2.3 of the TRIPLE project is to facilitate a strict classification of the resources of the GoTriple platform according to Morres categories.
- The aim of Task 2.4 is to develop a multilingual SSH vocabulary (covering nine languages) to be used for annotating (tagging) the publications that will be hosted in the GoTRIPLE platform.
  - annotating a publication in one of these nine languages (item or metadata) makes it more searchable for audiences that speak any of the other languages
  - the publication is enriched with links that help in the disambiguation of the term and increases its semantic interoperability.

Task 2.4 as a way to construct a SSH Vocabulary

- In this task, existing classifications and controlled vocabularies were gathered and compared.
- This work served as a basis for the creation of new vocabularies required for the description in new languages.
- During the initial phase of the work it was decided that the **Library of Congress Subject Headings** System will be used as a basic resource in English for the construction of the TRIPLE Vocabulary.
- The Library of Congress Subject Headings is probably the most widely adopted subject indexing system in the world and has been translated into many languages.

Work that has been done so far

- 14 basic concepts were identified from the Frascati taxonomy under Social Sciences and Humanities.
- Based on these 37 broad concepts from LCSH were identified.
- For each of these we used the Linked Data API of the Library of Congress and we retrieved their SKOS representation.
- For each of these we followed the skos:narrower property and we extracted their children in SKOS.
- We ended up with **2513** concepts.

Work that has been done so far

- We increased the multilingualism of the Vocabulary by three ways: 1) following links to LCSH 2) following links to wikidata and by extracting labels in various languages and 3) by ingesting existing mappings from national vocabularies (French and Italian National Libraries) to LCSH.
- The multilingualism of the vocabulary was increased further by using an automatic translation service to produce missing labels which then were validated and/or curated by partners.

Work that has been done so far

8 spreadsheets where all the automatically generated translations were validated and curated by partners

		Validated so far:				
Hierarchy	en	translation_1	validated_1	translation_2	validated_2	Final (read-only)
▶ Anthropology ▶ Ethnology ▶ Folklore	Communication in folklore	Communication dans le folklore	☑		☐	☑
▶ Anthropology ▶ Ethnology ▶ Folklore	Ethnic folklore	Folklore ethnique	☑		☐	☑
▶ Anthropology ▶ Ethnology ▶ Folklore	Intercultural communication in folklore	Communication interculturelle dans le folklore	☑		☐	☑
▶ Anthropology ▶ Ethnology ▶ Folklore	Mass media and folklore	Médias de masse et folklore	☑	Les médias et le folklore	☐	☑
▶ Anthropology ▶ Ethnology ▶ Folklore	Music and folklore	Musique et folklore	☑		☐	☑
▶ Anthropology ▶ Ethnology ▶ Folklore	National socialism and folklore	National-socialisme et folklore	☑	Le national-socialisme et le folklore	☐	☑
▶ Anthropology ▶ Ethnology ▶ Folklore	Public folklore	Folklore public	☑		☐	☑
▶ Anthropology ▶ Ethnology ▶ Folklore	Radio in folklore	Radio dans le folklore	☑		☐	☑
▶ Anthropology ▶ Ethnology ▶ Folklore	Symbolism in folklore	Symbolisme dans le folklore	☑		☐	☑
▶ Anthropology ▶ Ethnology ▶ Forest people	Rain forest people	Les gens de la forêt tropicale	☑	Peuples de la forêt tropicale	☑	☑
▶ Civilization	Aeronautics and civilization	Aéronautique et civilisation	☑		☐	☑
▶ Civilization	Climate and civilization	Climat et civilisation	☑	Le climat et la civilisation	☐	☑
▶ Civilization	Comparative civilization	Civilisation comparée	☑	civilisation comparative	☐	☑
▶ Civilization	Fishes and civilization	Poissons et civilisation	☑	Les poissons et la civilisation	☐	☑
▶ Civilization	Nature and civilization	Nature et civilisation	☑		☐	☑
▶ Civilization	Ocean and civilization	Océan et civilisation	☑		☐	☑
▶ Civilization	Soil and civilization	Sol et civilisation	☑	Le sol et la civilisation	☐	☑
▶ Civilization ▶ Education	After-school programs	Programmes parascolaires	☑	Parascolaire	☐	☑
▶ Civilization ▶ Education	Architecture in education	L'architecture dans l'éducation	☐	Architecture dans l'éducation	☑	☑
▶ Civilization ▶ Education	Archives and education	Archives et éducation	☑	Archives et de l'éducation	☐	☑

Work that has been done so far

		October 2020		June 2021	
		# of Concepts with labels in language (out of 2565)	language coverage percentage	# of Concepts with labels in language (out of 2565)	language coverage percentage
Greek	el	248	9.67%	2276	88.73%
French	fr	1624	63.31%	2219	86.51%
Polish	pl	336	13.10%	2561	99.84%
German	de	1028	40.08%	2561	99.84%
Italian	it	686	26.74%	2560	99.81%
Portuguese	pt	347	13.53%	2560	99.81%
Spanish	es	406	15.83%	2560	99.81%
Croatian	hr	153	5.96%	2228	86.86%

How to fill in the gaps?

- How to make the GoTriple Vocabulary even better suited to the needs of the SSH community?
- The creation of a vocabulary is a task that needs the cooperation of disciplines and experts such as programmers, librarians, SSH researchers etc. **So here comes the hackathon**.
- A hackathon is an event that brings together programmers with experts from other disciplines, for a short period of time in order to jointly work on a solution to a specific IT problem.
- Hackathon is a format that the SSH community is not familiar with, so the event was also experimental.

Organizational Issues and workflow

- The 1st hackathon was hosted by TRIPLE partner EKT (National Documentation Centre) and took place, virtually, during the following dates: 8-10 Nov. 2021 and 15-16 Nov. 2021.
- There was an open call inviting people to participate in this event.
- 84 people registered but 30 people actively participated.
- On November 3rd a pre-hackathon meeting took place. Participants were informed about the aim of the hackathon and the resources that would work with.

Hackathon’s Task Force

- Iraklis Katsaloulis: National Documentation Centre of Greece
- Cezary Rosiński: IBL PAN
- Irena Vipavc Brvar: CESSDA/UL-ADP
- Ana Inkret: CESSDA/UL-ADP

Opening questions

- What it was all about?
- Why Library of Congress Subject Headings (LCSH)?
- What languages and why them?
- Introduction to the resources
- What was done?
- Expected and received outcomes

What it was all about?

- Multilingualism
- Social Sciences and Humanities
- Reuse of SSH data
- International exchange of knowledge
- Mapping between Library of Congress Subject Headings and national data resources

What it was all about?

Linked Open Data is a mix of Linked Data and Open Data: it is both linked and uses open sources.

LOD helps to build bridges between various formats and allows to connect various interoperable sources of information. As a result, data integration and browsing through complex data become easier and much more efficient.

Why Library of Congress Subject Headings?

During the TRIPLE project Library of Congress Subject Headings service was identified as the **most appropriate resource for SSH** content.

For this decision, we took into account that LCSH is one of the **most popular** authoritative vocabularies, it contains a very large number of concepts from which we **could select** the most common SSH ones, there are **existing mappings** from national vocabularies to LCSH and, finally, the current **annotation mechanism of the ISIDOR platform** already uses LCSH.

What languages and why?

GoTriple is a **multilingual service**.

9 languages (French, Spanish, English, German, Greek, Portuguese, Croatian, Italian, Polish) are part of the platform.

We decided to focus on **languages less present** in the international exchange of knowledge and to confront the difficulties in **non-English speaking circuits**.

We used **openly available** resources.

Introduction to the resources

- **13 SSH resources**
- **Disciplines:**
  - Anthropology
  - Architecture, sculpture, painting and landscape
  - Art and archeology
  - Education; Information and communication; Politics, law and economics; Culture
  - Folklore
  - Linguistics
  - Literature and Literary Research
  - Philosophy
  - Psychology
  - Sociology

What was done?

The challenge for the groups that took part in the Hackathon was to find **solutions** to the problem of **mapping** between vocabularies in the languages that are represented in the Triple Project (Croatian, English, French, German, Greek, Italian, Polish, Portuguese, Spanish), and **LCSH**.

We were able to form groups for five languages.

Expected and received outcomes

- Case scenarios with data analysis, including data quality, structure, specificity
- Sets of good practices
- Code snippets with documentation of usage
- Tentative workflows
- Description of the broader landscape of issues related to the mapping of resources to authority databases
- Documentation with ideas, possibilities, dead ends

Presentations of Groups

- Greek team
  - Kostis Karozos and Agathi Papanoti
- Polish team
  - Patryk Hubar
- Portuguese-German team
  - Nelson H. Ferreira

Triple Hackathon

Mapping Greek SSH terms to LCSH

Team Greek-6:
- Kostis Karozos
- Agathi Papanoti
- Katerina Bartzi
- Vassiliki Apostolopoulou
- Haris Georgiadis

Discover Connect Collaborate

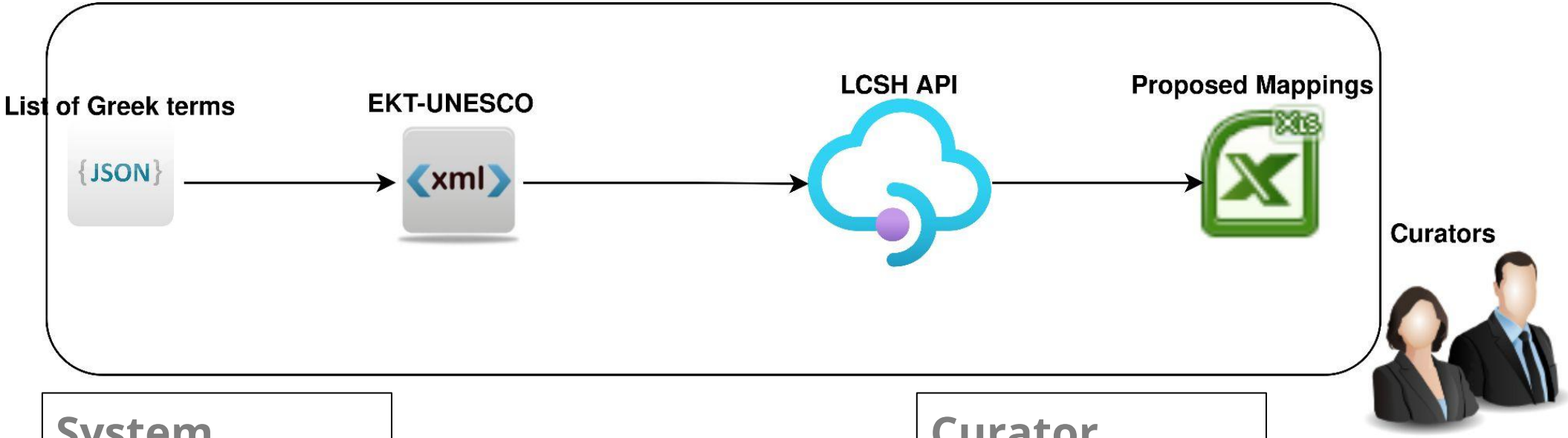
Triple

Available tools

- Given list in JSON with 1208 greek terms

- [UNESCO Thesaurus (EKT version)]
- LCSH search API

Methodology



System

For every LCSH search response:

- Get the first 5 mapping results
  - For each one of them get:
    - URI
    - Path

Curator

For every LCSH search response:

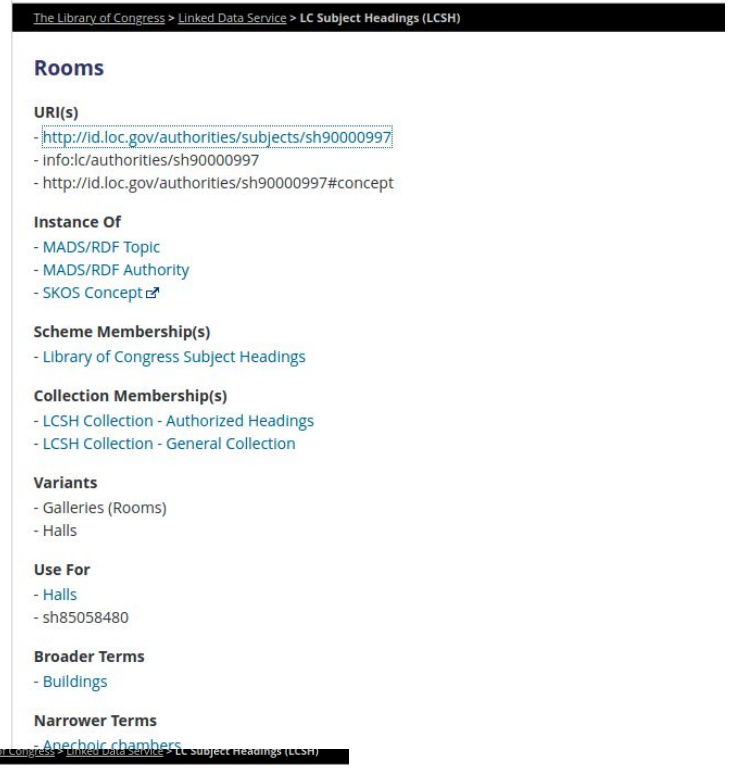
- Filter the SSH terms
- Approve the correct mappings
- If no mapping result found…

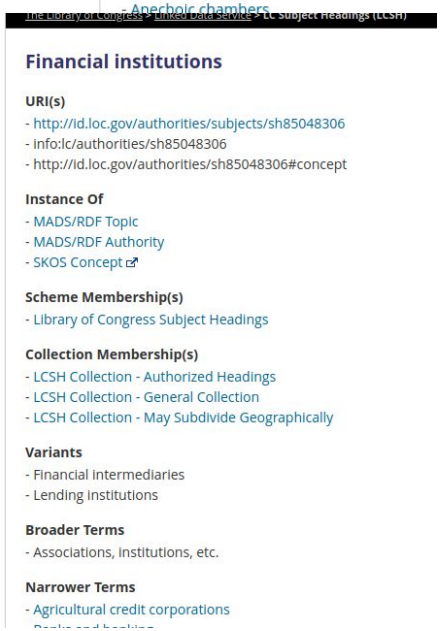
While sending requests to the LCSH search API…

We bumped into some inconsistencies…

- **‘Buildings’ has itself as a parent**

- **‘Associations, institutions, etc.’ might show up without its URI**





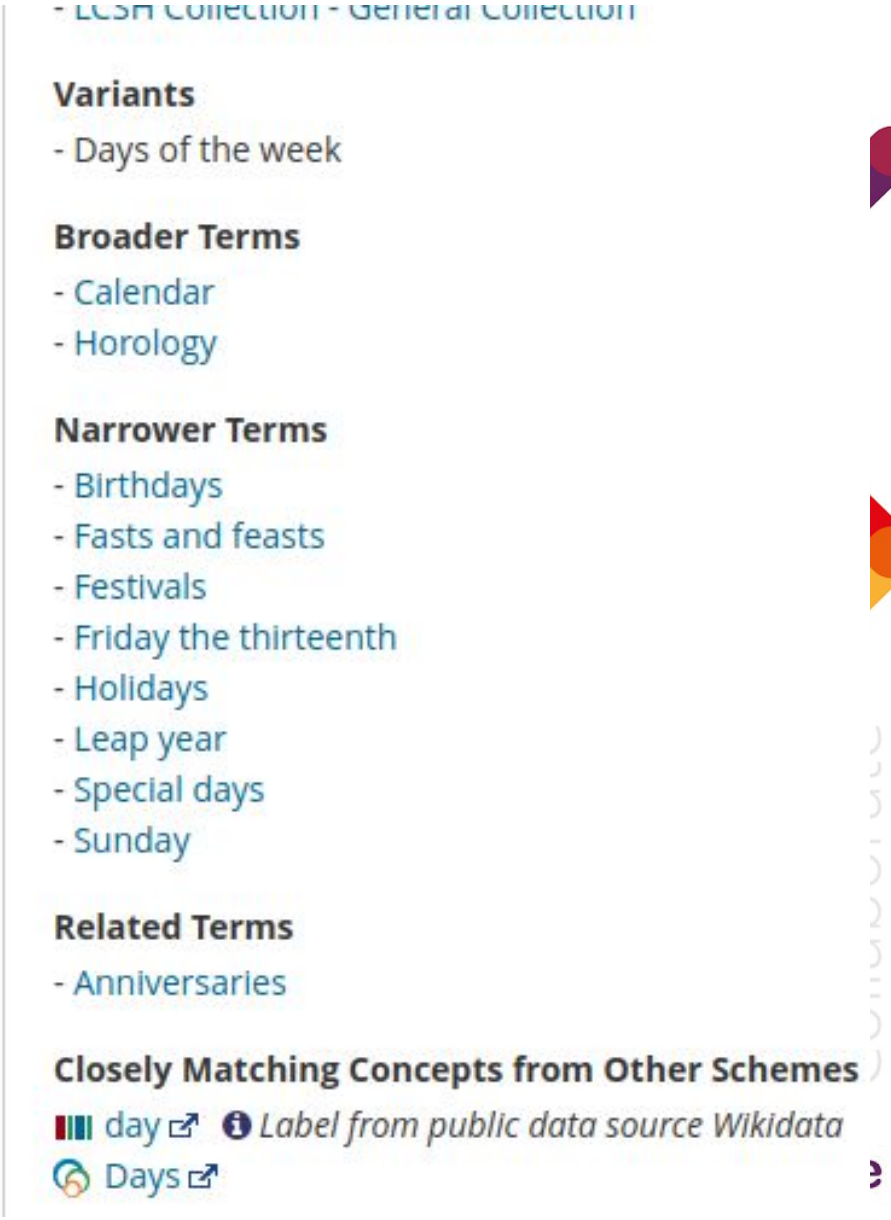
Discover Connect Collaborate

Triple

Semantic mistake (?)

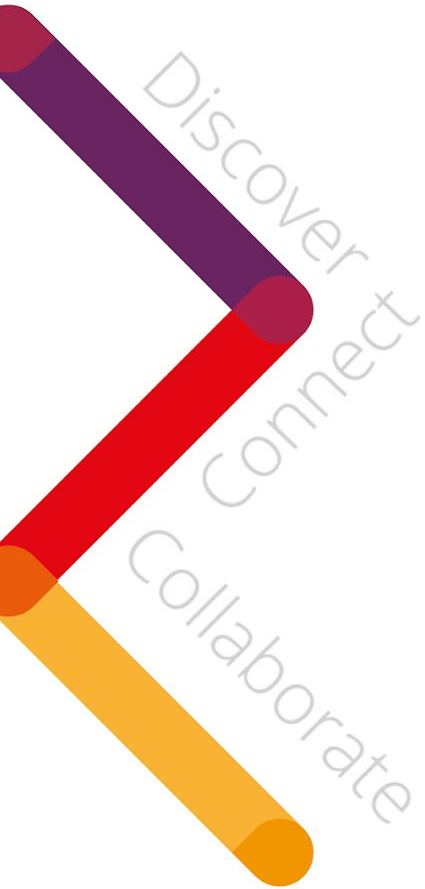
Misusage of <skos:broader>

Science / Physical sciences / Physics / Mathematical Physics / Physical measurements / Time measurements / Horology / Days



Results

All Terms	1208
Terms with at least 1 proposed mapping from LC	1018
Terms with a confirmed (exact match or close match) mapping	733
Terms with a confirmed (exact match) mapping	671
Terms with a confirmed mapping edited by a curator	71



Code:
https://github.com/EKT/GoTriple_Hack21_Greek-6

Mappings:
https://docs.google.com/spreadsheets/d/1sjZIFvhsC2DbkVN-XwtL5OzecZCI7MdMRB3myOD_nxg/edit?usp=sharing

www.
gotriple.eu

Thank u...

Triple

GOTRIPLE HACKATHON POLAND TEAM

Discover
Connect
Collaborate

Triple

What is PBL?

PBL
POLSKA BIBLIOGRAFIA LITERACKA

- database created in 1998 (in print from 1948)
- contains cca 0,7 mln for 1989–2003
- metadata of literary texts, studies, materials on theater, film, radio, and television
- files of people, objects, journals
- faithful to the printed original – internal data format, non-transferable, non-verifiable data
- 2015-2018 – funding obtained for new service and database (pbl.ibl.waw.pl)

Discover
Connect
Collaborate

Triple

Hierarchical structure of the PBL database

Hierarchical structure of PBL data does not correspond with the LCSH data structure

```
Array
(
    [lp] => 13
    [Polish Heading] => Aforyzm
    [Polish Path] => Aforyzm -> Genologia -> Teoria dzieła literackiego -> Teoria literatury
    [English Heading] => Aphorism
    [English Path] => Aphorism -> Genology -> Theory of literary works -> Theory of literature
)
```

	Label	Vocabulary	Concept	Subdivision	Identifier
1.	Aphorisms and apothegms Ana ; Apothegms ; Gnomes (Maxims) ; Sayings	LC Subject Headings (LCSH)	Topic		sh85005966
2.	Aphorisms and apothegms in art	LC Subject Headings (LCSH)	Topic		sh2006006839
3.	Aphorisms and apothegms in literature	LC Subject Headings (LCSH)	Topic		sh93008405

Discover
Connect
Collaborate

Triple

Chronology

- in Polish literary studies, literary movements carry chronological information
- in the PBL data, the chronology of literary movements/periods is specific to Polish literary studies, e.g. Romanticism in Poland began around 1820, coinciding with the publication of Adam Mickiewicz's first poems in 1822
- annual chronology in PBL vs centenary chronology in LCSH

Discover
Connect
Collaborate

Triple

Data cleaning

- Adapting the PBL chronology to the LCSH chronological terms, e.g. "po 1989" (after 1989) → 20th century, 21st century


```
{
    "heading": "Literatura polska po 1989",
    "path": "Literatura polska po 1989"
},
{
    "heading": "Ogólne (po 1989)",
    "path": "Ogólne (po 1989) -> Literatura polska po 1989"
},
{
    "heading": "Dramat (po 1989)",
    "path": "Dramat (po 1989) -> Literatura polska po 1989"
},
{
    "heading": "Krytyka (po 1989)",
    "path": "Krytyka (po 1989) -> Literatura polska po 1989"
},
{
    "heading": "Poezja (po 1989)",
    "path": "Poezja (po 1989) -> Literatura polska po 1989"
},
{
    "heading": "Proza (po 1989)",
    "path": "Proza (po 1989) -> Literatura polska po 1989"
```


```
daty = {
        'Od 1945': ['20th century', '21st century'],
        'Do roku 1939': ['medieval', 'to 1500', '1450-1600', '16th century',
                         '17th century', '18th century', '19th century', '20th centur
        'w l. 1939-1945': ['20th century'],
        'do r. 1939': ['medieval', 'to 1500', '1450-1600', '16th century',
                       '17th century', '18th century', '19th century', '20th century'
        'lat 1939-1945': ['20th century'],
        '1945-1989': ['20th century'],
        'po roku 1989': ['20th century', '21st century'],
        '1945-': ['20th century', '21st century'],
        'Od roku 1945': ['20th century', '21st century'],
        'wieku XIX i początku wieku XX (do 1918)': ['19th century', '20th century'],
        'od 1945': ['20th century', '21st century'],
        '1765-1831': ['18th century', '19th century'],
        '1831-1890': ['19th century'],
        '1890-1914': ['19th century', '20th century'],
        '1914-1939': ['20th century'],
        '1939-1945': ['20th century'],
        'od 1945 roku': ['20th century', '21st century'],
        'od 1992': ['20th century', '21st century'],
        'do 1996': ['medieval', 'to 1500', '1450-1600', '16th century',
                    '17th century', '18th century', '19th century', '20th century'],
        'średniowiecze': ['medieval', 'to 1500'],
```

Machine translation

google-trans-new 1.1.9

Poezja (1945-1989) -> Literatura polska 1945-1989	Poetry (1945-1989) -> Polish literature 1945-1989

```
for record in tqdm(data):
    heading=record['heading']
    path=record['path']
    headingtotranslate.append(heading)
    pathtotranslate.append(path)

    result1 = translator.translate(heading,lang_tgt='en')

    headingtranlsated.append(result1)
    result2 = translator.translate(path,lang_tgt='en')
    pathtranslated.append(result2)
```

Discover
Connect
Collaborate

Triple

Data cleaning 2.0

- Basic text cleaning: converting all characters to lowercase, removing punctuation

```
tematy_ang_all_clean = []
mapping = { '(':'', ')':'', '-':'', ',':'', '.':'', '"':'', '>':''}

for elem in tematy_ang_all_dates:
  elem = "".join([mapping[c] if c in mapping else c for c in elem])
  elem = elem.strip().lower()
  tematy_ang_all_clean.append(elem)
```

Poetry (20th century)	poetry 20th century

Library of Congress response

https://id.loc.gov/search/?q=poetry%2020th%20century%20scheme:http://id.loc.gov/authorities/subjects&start=160&format=json


```
[
"atom:entry",
{
    "xmlns:atom" : "http://www.w3.org/2005/Atom"
}
,
[
"atom:title",
{
    "xmlns:atom" : "http://www.w3.org/2005/Atom"
}
,
    "Poetry, Modern--20th century"
],
[
"atom:link",
{
    "xmlns:atom" : "http://www.w3.org/2005/Atom",
    "rel" : "alternate",
    "href" : "http://id.loc.gov/authorities/subjects/sh85103727"
}

],
[
"atom:link",
{
    "xmlns:atom" : "http://www.w3.org/2005/Atom",
    "rel" : "alternate",
    "type" : "application/rdf+xml",
    "href" : "http://id.loc.gov/authorities/subjects/sh85103727.rdf"
}

],
],
[
"atom:link",
{
    "xmlns:atom" : "http://www.w3.org/2005/Atom",
    "rel" : "alternate",
    "type" : "application/json",
    "href" : "http://id.loc.gov/authorities/subjects/sh85103727.json"
}

],
[
"atom:id",
{
    "xmlns:atom" : "http://www.w3.org/2005/Atom"
}
,
    "info:lc/authorities/subjects/sh85103727"
],
```


Discover Connect Collaborate

Triple

String similarity

```
x_list = []
for key, values in test.items():
  for value in values:
    seq = SequenceMatcher(None, key, value).ratio()
    print(f"{key} i {value} = {seq} similarity")
    x_list.append([seq, key, value])
```

- Terminology recognition with difflib SequenceMatcher (Python)

0.9375	cuban literature	Cuban literature
0.9333333333333333	thai literature	Thai literature
0.9333333333333333	science fiction	Science fiction
0.9333333333333333	cultural policy	Cultural policy
0.9333333333333333	cultural policy	Cultural policy
0.9285714285714286	special issues	Special issues
0.9285714285714286	neutralization	Neutralization
0.9285714285714286	film criticism	Film criticism
0.9230769230769231	visual poetry	Visual poetry

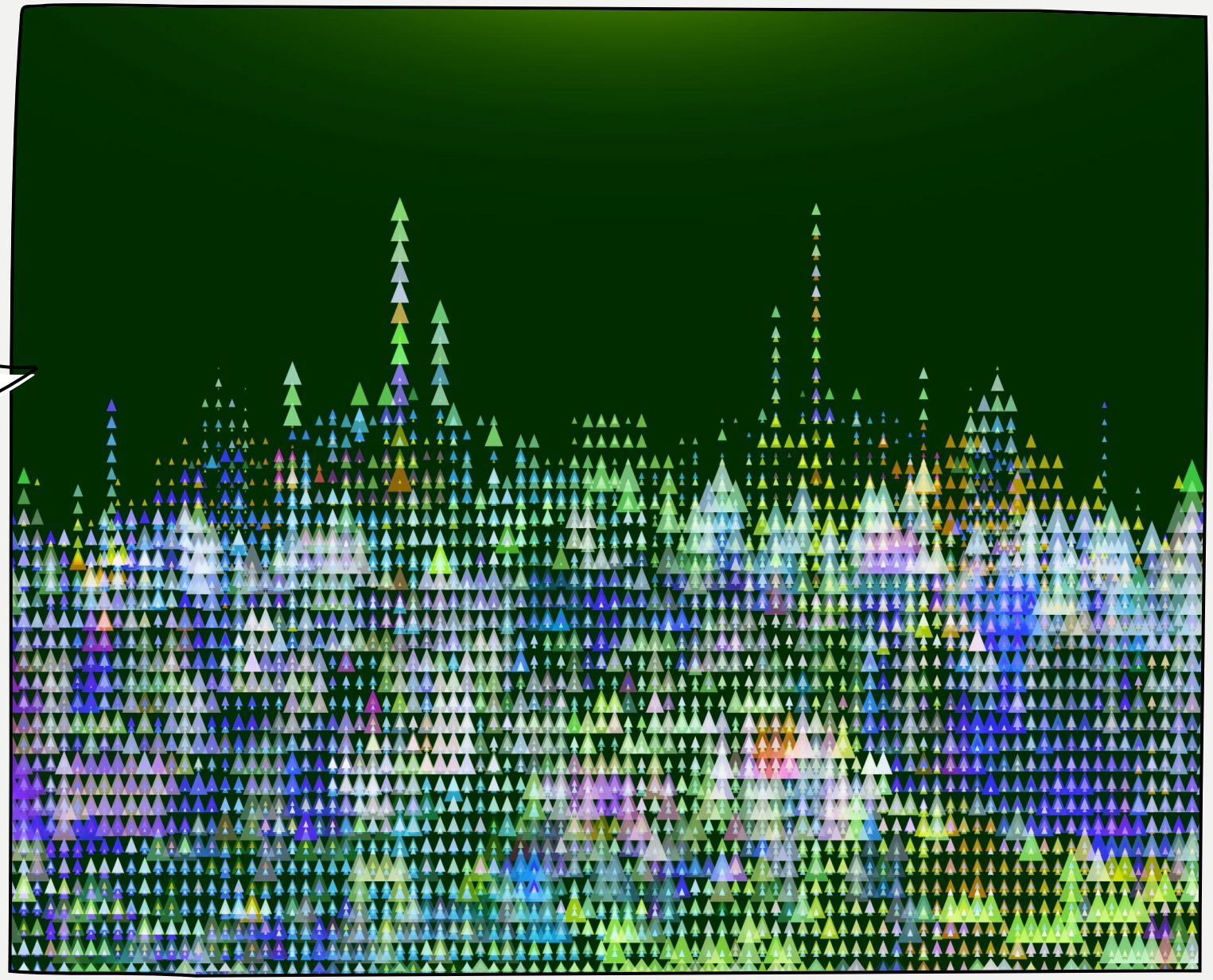
Our conclusions

- unreliability of automatic translation
- different data structures PBL vs. LCSH
- fundamental discrepancies (e.g. timing issues)
- need for manual corrections

Discover
Connect
Collaborate

Triple

Team for Portuguese language



Oleg

Isabel

Rodrigo

Iphigenia

Nelson

Methodology for testing LEXIKON IN LCSH

General Approach considering terminology and vocabulary management and matching

- Two list of vocabularies in Portuguese were considered
  - List 1. ‘Art and humanities’
  - List 2. ‘Social sciences and Folklore’
- List 1. – a sample of 90 terms was researched
- List 2. – first a sample of 30 terms, later a more extended one covering a great part of the list

Lack of terminology matching	terms, terms that need to fit within a hierarchy and do not exist as a	Issues with translations
• Consequences for research: difficult to find a match for closed concepts and composed terms.	• Consequence for research: hard to perform relations between concepts and dive in exploratory research through ramification and bridges between terms	• Consequence for research: wrong translations lead to extra effort for finding the correct term.

Main issues identified

Suggestions

- Management of terminology with the support of national subject headings databases (e.g. https://opendata.bnportugal.gov.pt/).

  Better terminology assessment will improve automatic translation to connect to the LCSH.

- Bilingual lexicons (databases) would benefit matches between universal concepts and avoid ambiguities
- Linguistic aspects of words like gender and number should be consider in vocabularies for amplifying the spectrum of research and reduce ambiguities or ‘not matches’

- databases considered to follow a kind of ‘standardized approach’ on terminology should be higher in hierarchy for word matching

What was the result of the resource analysis (content, suitability for SSH, data structure)?

- Some inefficiencies on lexicon matching
- Vocabulary lists are rich but databases for cross-reference may not be aligned
- There is no conclusions on data structure

What issues arose during the mapping between the resource and LCSH?

(mentioned regarding research procedures)

- In what concerns mapping we hadn’t de necessary time to analyse and consider it.

What is the result of your workflow in terms of code and documentation?

No results

What are your thoughts on creating a workflow for mapping between national SSH resources and authority databases (chain of activities, difficulties, engaged parties)?

At first

- approaches we struggle with understanding the task in its full dimension, resources available and objectives
- There was a lack of expertise on IT approaches to databases management
- Difficulties with aligning schedules because of patterns' professional commitments

After some adjustment and knowing partners

- We could applied a research accordingly to our resources
- The tasks could be completed and results compared

How do you build sufficient SSH vocabulary automatically or semi-automatically for a discovery service such as GoTriple platform?

Engagement with national authorities responsible for building vocabularies and national specialists in library databases sciences is crucial for reaching the more suitable sources for vocabulary.

Bilingual lists of matches could be helpful for generating standardization and disambiguation

Adding a tool pointer that present other possible words:

e.g. ‘do you mean this instead?’

How can the organizers support the work better in the next hackathon?

- **The teams could be assigned more in advance so they can know each other**
- **Before team to be build, general tasks that could be made by different levels of expertise should be suggested.**
- **MS team, despite being a good tool for this kind of workflows, did not work so well at the beginning, due to accounts issues of many participants**

Oleg

Isabel

Rodrigo

Iphigenia

Nelson

Thank you!

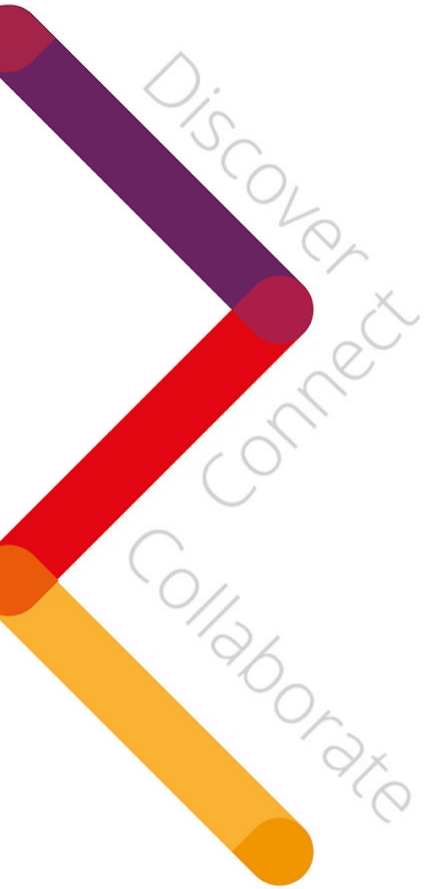
Is a hackathon the right tool for the SSH community?

- A hackathon is a format that is used mainly by programmers. Can it be useful for the SSH community?
- An obstacle is that SSH researchers are not familiar with the idea of a hackathon.
- In an era in which digital humanities expand continuously, effective collaboration between SSH researchers, librarians and IT experts is needed.
- A hackathon is an ideal place for this to happen.

“ We all tried to learn and understand the language that others speak”.

After the GoTriple Hackathon

- Preliminary observations:
  - truly accessible and updated resources
  - hierarchical data structures
  - bilingual control vocabularies
- The implementation of the results of the Hackathon into the Triple project workflow by enriching the existing concepts in T2.4.
- “We need a robust infrastructure for development, and maintenance of SSH multilingual vocabularies preferably based on OPERAS national nodes to provide language expert support.”



project.
gotriple.eu

Follow us on:

@TripleEU

TRIPLE Project