Interactive Visualization of Book Publication History

Yanzhe Li
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
y.li-7@student.utwente.nl

ABSTRACT

The book publication history is valuable for understanding how past cultures evolved over time and how knowledge spread cross cultures. To construct the publication history of a book, we need to identify relevant versions (e.g., the original, editions and translations) and relevant contributors (e.g., authors, editors and translators), as well as important attributes (e.g. year and location of the publication). Such information is traditionally catalogued in different bibliographic records, many of which are recently transformed into linked data or knowledge graphs (often in the form of RDF triples). Despite being friendly for automated processes, these RDF triples do not necessarily construct the publication history easily. In this research, we propose a data model for book publication history, based on which we collect data from the underlying knowledge graph and construct the publication history. We develop a prototype using zoomable icicle graphs to visually present the book history that uses can interactively explore. A preliminary user evaluation shows that this interactive visualisation tool helps to locate the right information faster and improves the user experience in general.

Keywords

book publication history, data model, interactive visualization

1. INTRODUCTION

Books, especially classic ones, often have different editions and translations in the long course of history. For example, *The Dream of the Red Chamber*, one of China's Four Great Classical Novels, has been translated into 23 different languages and has more than 160 different versions, since it was originally published nearly 300 years ago. Even within the same language, the original book or its different versions could be translated different translator. During this creation process, the translators usually integrated his/her own understanding of the original work into his/her translation, which might have further influences on the spread of the original knowledge or ideas. Therefore, the book publication history is valuable for historians or other researchers as it helps understanding how

past cultures evolved over time and how knowledge spread cross cultures.

However, constructing a book publication history is not trivial. As describe above, there are many entities related to book history, and the relationship among these entities are also complicated. we need to identify relevant versions (e.g., the original, editions and translations) and relevant contributors (e.g., authors, editors and translators), as well as important attributes (e.g. year and location of the publication). Such information is traditionally catalogued in different bibliographic records, many of which are recently transformed into linked data or knowledge graphs (often in the form of RDF triples). Within large aggregations systems, such as WorldCat[1] and Wikidata[2], simple gathering these RDF triples does not necessarily construct the publication history. Furthermore, browsing the RDF triples is not an intuitive way to study the complex relationships that exist in book histories. An interactive visualisation is therefore desirable to address this issue.

In this research, we address two research questions:

- How to model the important entities and their relationships for constructing book publication history?

- Once constructed, how can we visualise the book publication history interactively?

Inspired by the FRBR model [18], we propose our data model for representing the book publication history, as illustrated in Figure 1. This model incorporates a time line with the traditional FRBR model, which allows for the expression of historical events, such as the publication of new editions or translations. Based on a data set that consists the RDF triples extracted from WorldCat and Wikidata about five classic books, we design specific SPARQL queries to collect data that are relevant for constructing book publication history according to the data model. Once the publication history is constructed, we use a zoomable icicle graph to visually present the book history that users can explore interactively.

The rest of the paper is organised as the following: after the related work is introduced in Section 2, we describe the details of the data model in Section 3. Visualisation requirements are listed in Section 4 and implementation details are given in Section 5. We report our verification process and results in Section 6, before we conclude in Section 7.
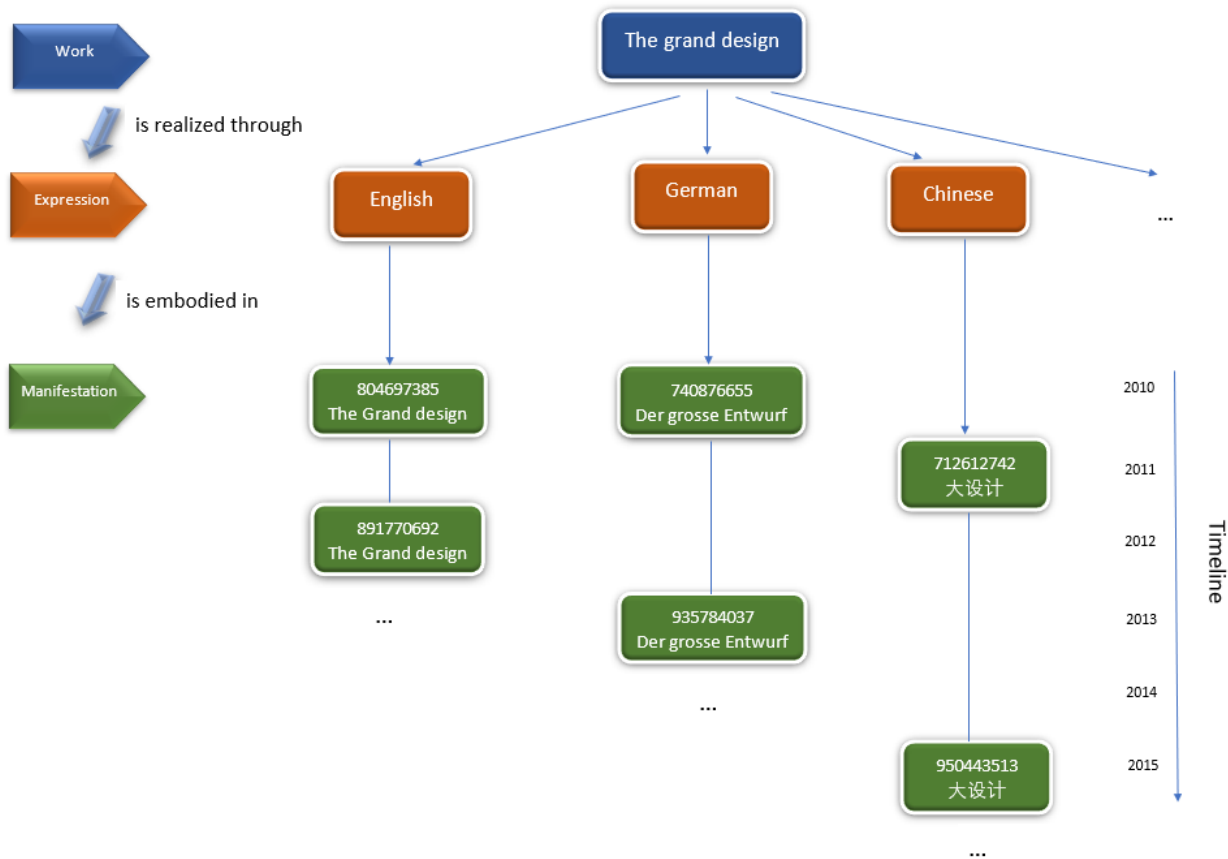
2. RELATED WORK

[1] https://www.worldcat.org/
[2] https://www.wikidata.org/

Figure 1. Data model of book publication history

2.1 Book history

Stephanie [1] has conducted research on the book history of Europe. besides introducing some research on the history of books, there is a tool to teach the expansion of European printing technology in a socioeconomic context, which is a visualization tool called Atlas of Early Printing [3]. At the same time, there are also other researches on the history of books, such as the researches on the journals [11], poetry [25] and linguistics [16]. Meanwhile, some organizations are working on it. For instance, an organization named "book history research network" was established and managed by the University of London [2], which aims to bring together current and former graduate students, teachers, and independent researchers working in any field or period of "History of this Book". How to automatically construct book publicatin histories based on available structured data (e.g. RDF triples) remains a challenge.

2.2 Data model visualization

A data model is an abstract model used to organize data elements and standardize their relationships and attributes with real-world entities [9]. Meanwhile , as the main method of knowledge representation for complete machine understanding, ontology is widely used in different application fields [8]. Where the ontology is the way to display the attributes of the subject area and how they are related by defining a set of concepts and categories that represent the subject [20]. The most related work of data modelling

---

visualization is ontology visualization. Normally, the data in the ontology format of the data model is used for machines, one of the difficulties is that the full machine understanding data is hard to be easily understood by a human [8]. That is also the reason why there are a large number of approaches and visualization tools. For example, FLEXViz [10] was created to improve the exploration of bio-ontologies. Olsvis [22] use a folder-interface to explore hierarchies.

Recently, knowledge graphs become a popular data model that is used to store interlinked entities [23]. The most relevant work that could be found is a visualization of a musical knowledge graph [4], which was founded by Cambridge University enterprise accelerator. It details how to make a music knowledge map. It could help the music expert to easily understand the evolution of music through the knowledge graph. Furthermore, there are some different knowledge graph achievements in other fields, such as tabular data [6], Hubway data [21], etc. Meanwhile, the knowledge graph is also applicable to the field of film [14], smart grids [12], geoscience [24] and etc.

Although visualizations are applied in some fields, visualization of book publication history is completely new. This is the contribution of this research.

3. DATA MODEL

The data model was inspired by the work of the IFLA Study Group, they developed a conceptual entity-relationship model called Functional Requirements for Bibliographic Records (FRBR) [18]. "This model consists of four enti-

```
1   <http://www.worldcat.org/oclc/804697385> <http://schema.org/inLanguage> "en" <http://www.worldcat.org/
        oclc/804697385.nt> .
2   <http://www.worldcat.org/oclc/804697385> <http://purl.org/library/oclcnum> "804697385" <http://www.
        worldcat.org/oclc/804697385.nt> .
3   <http://www.worldcat.org/oclc/804697385> <cdf:workId> "312338454" <http://www.worldcat.org/oclc
        /804697385.nt> .
4   <http://www.worldcat.org/oclc/804697385> <http://schema.org/name> "The Grand design" <http://www.
        worldcat.org/oclc/804697385.nt> .
5   <http://www.worldcat.org/oclc/804697385> <http://schema.org/publication> <http://experiment.worldcat.
        org/entity/work/data/312338454#PublicationEvent/london_bantam_press_2010> <http://www.worldcat.org/
        oclc/804697385.nt> .
6   <http://experiment.worldcat.org/entity/work/data/312338454#PublicationEvent/london_bantam_press_2010> <
        http://schema.org/startDate> "2010" <http://www.worldcat.org/oclc/804697385.nt> .
7   <http://www.worldcat.org/oclc/804697385> <cdf:StdRT2> "PrintBook" <http://www.worldcat.org/oclc
        /804697385.nt> .
8   <http://www.worldcat.org/oclc/804697385> <http://schema.org/creator> <http://viaf.org/viaf/102304634> <
        http://www.worldcat.org/oclc/804697385.nt> .
9   <http://www.worldcat.org/oclc/804697385> <http://purl.org/library/oclcnum> "804697385" <http://www.
        worldcat.org/oclc/804697385.nt> .
10  <http://www.worldcat.org/oclc/804697385> <http://schema.org/datePublished> "2010" <http://www.worldcat.
        org/oclc/804697385.nt> .
```

Listing 1. Example data in RDF format

ties: The entities defined as work (a distinct intellectual or artistic creation) and expression (the intellectual or artistic realization of a work) reflect intellectual or artistic content. The entities defined as manifestation (the physical embodiment of an expression of a work) and item (a single exemplar of a manifestation), on the other hand, reflect physical form." IFLA Study Group (2009) states that these entities represent the different aspects of user interests in the products of intellectual.

The data model for this research is based on this FRBR model, which has some custom designs for book history (As depicted in figure 1). In order to explain the data model clearly, some example data was used to illustrate how the data model was constructed. The entities of the book history are stored in the RDF format, there are some example data in listing 1. The data model will be explained by these example data.

The first level of the data model is the work ID of the work entity, which is the ontology representation of all the versions of a book, and the example data is the third line in listing 1. The second level is the language of work entity, which is the expression entity, and the example data is the first line in listing 1. The relation between the work entity and expression entity is the work is realized through the expression. The third level contains the OCLC control numbers, each representing one of the manifestations of the same work. The OCLC control number [5] is the identifier for a unique bibliographic record in OCLC's WorldCat and the example data is the ninth line in listing 1. The relation between the manifestation entity and the expression entity is that the expression entity is embodied in the manifestation entity. At the same time, the third level is grouped by the year, as the timeline is important in constructing book histories. Since the item entity is a single exemplar of a manifestation, it could be a book stored in the library around the world, so it will not be considered in this data model.

In conclusion, the data model is an idealized model. However, in the real world, data may be unbalanced or unclear, which brings challenges to visualization.

4. VISUALIZATION REQUIREMENTS

4.1 The results should be clear and reliable.

The purpose of this project is to provide an interactive visualization of book publishing history. If the visualization data is not reliable, it is meaningless for users. Thus, the requirement of this part is that the result should be the same as the users' demand at the ontology level.

4.2 The visualization should be efficient and easy to use

The most important part of the requirement is the visualization, which directly affects the efficiency and user experience. First of all, users could get a clear overview of the publication history of the book. This overview helps users take the next step. Second, the process of find target information should be easy for novice users. For example, there is a task to find the English version of a book published in 1990, users should complete this task in a few steps. Additionally, it does not need to put a lot of effort to learn it. Finally, the data distribution is uneven, so the system should properly handle these unbalanced data.

5. IMPLEMENTATION

5.1 Data collection

5.1.1 Dataset

In this research, we work with a data set, which was made availabe during the "Visualizing Digital Humanities" workshop held at the Lorentz Center at the University of Leiden in 2017 [3]. This data set consists of the RDF triples extracted from Wikidata and WorldCat. These triples describes different editions and translations of five classic books, including *The Grand Design* used in Figure 1. The data is stored in a Apache Jena Fuseki server and accessible via SPARQL queries, as shown in Figure 2.

5.1.2 Query

The next step is to query the correct information from the dataset. First, the target entities need to be determined. Base on the data model built before, it is capable to find that the book title, author, publication year, version and translator are the important entities related to book history. Consequently, the logical chain of the query process is clear. Step one is to get the ontology of target book from Wikidata. Step two is to find all the OCLC control number of this book, which represent every single version of

Figure 2. Framework of prototype

```
1  PREFIX wdt: <http://www.wikidata.org/prop/direct
      />
2  PREFIX library: <http://purl.org/library/>
3  PREFIX schema: <http://schema.org/>
4  PREFIX wd: <http://www.wikidata.org/entity/>
5  SELECT DISTINCT ?date ?inLanguage ?ocn ?
      translatorName ?book
6  WHERE {
7    ?wd wdt:P243 ?ocn.
8    ?wc library:oclcnum ?ocn;
9        schema:name ?book;
10       schema:datePublished ?date;
11       schema:inLanguage ?inLanguage.
12
13   VALUES ?wd {wd:Q93773 # ''The Grand Design"}
14   OPTIONAL{
15       ?wc
16       schema:translator ?translator.
17     ?translator schema:name ?translatorName.
18   }
19 }
20 ORDER BY ?date
```

Listing 2. SPARQL code of querying information of "The Grand Design"

this book. Step three is to pass these OCLC control numbers to the WorldCat database. According to the OCLC control numbers, it can find all other entities required by the project. The final step is to filter out the erroneous data in the client-server. For example, a book was first published in 2010, so that all the result with publish year before 2010 should be removed. The SPARQL [7] query code can be found in listing 2.

5.2 Design choice

Base on the format of the query result, it is convenient to construct a knowledge graph from the data. The reason is that the knowledge graph is considered as a tool for describing entities and their relationships [17]. Furthermore, knowledge graphs have a great ability in providing semantically structured information [26]. For the design choice of the knowledge graph, the first idea is a tree graph. It clearly introduces the affiliation between versions. The tree has five levels: title, year, language, version and translator from left to right. However, unbalanced data will destroy the graph, as it may happen that one of the nodes may be incredibly huge and another one contains only one child. For example, if a book has 800 versions published in 2000, these versions all displayed in the tree graph. When user looking for a book which published in 2011, they need to scroll down through all the books published in 2000. This will make the graph looks

awful and difficult to use. Whereas a zoomable treemap is a better way to visualize it. The visualization will be represented by a zoomable icicle graph, which is a subclass of a zoomable treemap, and its shape is similar to figure 3. The zoomable icicle graph can fold the unbalanced data, as well as display the proportion of the data. These attributes are very helpful for a better overview.

A color map is used to enhance the performance of presenting the information. MacDonald [15] demonstrated that color provides an important dimension in visual communication, it can greatly enhance the effectiveness of information if used well. Kenneth [13] also stated that the process of visualization is providing a visual representation of data. One of the most fundamental and import aspects is mapping the information to colors. For the color schema, Rogowitz [19] believed that human perception is less sensitive to changes in saturation or hue than changes in luminance. Thus, a color map with a brightness gradient is used for year layer. Meanwhile, a strong contrasting color map was used for the language layer, which makes it easier for users to distinguish the language represented by different colors.

In short, by using these two methods in combination, it provides users with an impressive overview when using the visualization.

5.3 Prototype

5.3.1 Framework

The prototype consists of three primary parts (As depicted in figure 2). The first part is the basis of the prototype. It is an Apache Jena Fuseki server with the prefabricated database, which simulates the real servers of Wikidata and Worldcat. The next part is a client-server written in python, which queries the data from the Apache server, and then parses and reformats the data to meet the format requirement of visualization. The last part is the interactive visualization, which is a zoomable icicle graph. The year and language are mapped to variable colors to bring a better overview. At the same time, some books have a long history, so an extra layer of year groups may be added to maintain usability. Another highlight is that the total number of nodes is labelled after the node name, and the graph is distributed proportionally based on these numbers. Finally, the prototype enables the users to filter the language and change to another book, these functions are completed by the two drop-down menus in the upper left corner.

5.3.2 Workflow

This prototype has 6 layers, namely the book title, year

group, single year, language, OCLC control number with book title of this version, and translator (if any). Figure 3 is the visual homepage of a book, including the first three layers. When clicking the "1966-1976", it will go to figure 4, which shows the language distribution of every year. Then clicking the "1966" to go to figure 5, where the OCLC control number and book title are displayed after the language layer. If a translator of this book recorded in this database, it could be found when clicking the OCLC control number layer.

6. VERIFICATION

The goal of this section is to validate whether the requirements are met. There should be a complicated experiment to compare the new solution with the existing solution. However, because of the time limitation and the outbreak of COVID-19 pandemic, the scale of the experiment was limited and the participant's group must be free of COVID-19 risks. As a result, the experiment only tested with 3 participants who are the roommates of the researcher. Such participants may lead to bias in results. For example, participants may simply choose the researcher's product is better because they are friends of the researcher. Although the results may be biased, the positive results prove that this interactive visualization is effective to a certain extent. The experiment tasks and the questionnaire are attached in the appendix. The results reveal that the time consumption of each task is greatly reduced compared to existing solutions (As depicted in table 1).

Table 1. Average time to complete the task

Task number	New visualization(s)	WorldCat(s)
1	6.5	32.7
2	7.3	19.9
3	13.2	36.1

Additionally, according to the results of the questionnaire survey (As depicted in table 2), all participants believed that this tool was better than the existing tools. Nevertheless, the participants still have some question about this tool. For example, "why there are only two books? what happened if I want to check other books?", "Why are some language filters not working?" etc. These questions are all about the incomplete parts of the prototype. This problem can be fixed in the actual product, but it is not necessary to spend time on these repetitive work since it is just a prototype. Additionally, there is a productive suggestion. It suggests that there should be a search box where the user can make inputs, and if the input content is the same as the string stored in this system, the visualization tool should directly go to that part. This can bring a better experience to users.

Table 2. Survey result

QN	Participant 1	Participant 2	Participant 3
1	A	A	A
2	A	A	A
3	A	A	A

In conclusion, although there are still some problems with the prototype, the verification experiment indicates that this interactive visualization of book publication history is effective to a certain extent.

7. CONCLUSION AND FUTURE WORK

This research proposes a novel interactive visualization of book publishing history. It not only provides a better overview of book history but also reduces the query time. This research especially contributes to the knowledge graph visualization of book publishing history. Furthermore, a practical tool has been built for digital humanities researchers or layman users who are interested in book history [2].

A data model was established for book publication history in this paper, which is based on the Functional Requirements for Bibliographic Records model. It consists of three entities. They are the work entity, expression entity and the manifestation entity. There are two relations between the entities. One is the expression entity is the intellectual realization of the work entity, and the other one is the manifestation entity is the physical embodiment of an expression of a work.

For the design of the visualization, a zoomable icicle graph is used to increase the usability and handle unbalanced data. Additionally, a color map is used for the language and year, together with the proportional graph to provide a better overview. Users are able to clearly acquire the total number of every section. Meanwhile, the steps of querying a book are reduced.

Future research can look into showing more entities in the visualization so that the user could get more information from that. The other aspect is to build visualizations from scratch. An existing library was used in this research because of time limitation. Hence, some functions cannot be flexibly adapted to this research. Therefore, the visualization established from scratch brings benefits for future research. Finally, the search box is also a good choice as suggested by one of the participants.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] Book history: A perspectives quick study | perspectives on history | aha. `https://www.historians.org/publications-and-directories/perspectives-on-history/october-2016/book-history-a-perspectives-quick-study` Accessed on 17 June 2020.

[2] Book history research network — cphc. `https://www.cphc.org.uk/book-history-research-network` Accessed on 19 June 2020.

[3] Center for scientific workshops in all disciplines - home. `https://www.lorentzcenter.nl/` Accessed on 17 June 2020.

[4] How to visualize a (musical) knowledge graph. `https://cambridge-intelligence.com/visualizing-a-knowledge-graph/` Accessed on 19 June 2020.

[5] Oclc control number. `https://www.wikidata.org/wiki/Property:P243` Accessed on 17 June 2020.

[6] Semantic web challenge on tabular data to knowledge graph matching. `https://www.cs.ox.ac.uk/isg/challenges/sem-tab/` Accessed on 15 June 2020.

[7] SPARQL Query Language for RDF. `https://www.w3.org/TR/rdf-sparql-query/`Accessed on 20 June 2020.

[8] N. Achich, B. Bouaziz, A. Algergawy, and F. Gargouri. Ontology visualization: An overview. In A. Abraham, P. K. Muhuri, A. K. Muda, and N. Gandhi, editors, *Intelligent systems design and applications*, Advances in intelligent systems and computing, 2194-5357, pages 880–891, Cham, Switzerland, 2018. Springer.

[9] A. Borgida, R. J. Brachman, D. L. McGuinness, and L. A. Resnick. Classic: a structural data model for objects. *ACM SIGMOD Record*, 18(2):58–67, 1989.

[10] S. M. Falconer, C. Callendar, and M.-a. Storey. Flexviz: visualizing biomedical ontologies on the web. In *International Conference on Biomedical Ontology, Software Demonstration*. Citeseer, 2009.

[11] J. K. Ha and C. H. Yun. The 30-year publication history of asian-australasian journal of animal sciences. *Science Editing*, 6(1):10–18, 2019.

[12] H. Huang, Y. Chen, B. Lou, Z. Hongzhou, J. Wu, and K. Yan. Constructing knowledge graph from big data of smart grids. In *Tenth International Conference on Information Technology in Medicine and Education*, Los Alamitos, California and Washington and Tokyo, 2019. Conference Publishing Services, IEEE Computer Society.

[13] Kenneth Moreland. Diverging color maps for scientific visualization. pages 92–103. Springer, Berlin, Heidelberg, 2009.

[14] Q. Lei and Y. Liu. Constructing movie domain knowledge graph based on lod. In *2019 Twelfth International Conference on Ubi-Media Computing (Ubi-Media)*. IEEE, 2019.

[15] L. W. MacDonald. Using color effectively in computer graphics. *IEEE Computer Graphics and Applications*, 19(4):20–35, 1999.

[16] F. J. Newmeyer. Martin joos's readings in linguistics. *Historiographia Linguistica*, 46(3):73–130, 2019.

[17] H. L. Nguyen, D. T. Vu, and J. J. Jung. Knowledge graph fusion for smart systems: A survey. *Information Fusion*, 61:56–70, 2020.

[18] I. S. G. o. t. F. R. f. B. Records and S. C. o. t. I. S. o. Cataloguing, editors. *Functional Requirements for Bibliographic Records: Final Report*, volume 19 of *UBCIM Publications. New Series*. K.G. Saur, Berlin and Boston, reprint 2013 edition, 2013.

[19] B. E. Rogowitz, L. A. Treinish, and S. Bryson. How not to lie with visualization. *Computers in Physics*, 10(3):268, 1996.

[20] B. Smith. Ontology. In G. Hurtado and O. Nudler, editors, *The furniture of the world*, Rodopi philosophical studies, pages 47–68. Rodopi, Amsterdam, 2012.

[21] N. Staff. Using (spring data) neo4j for the hubway data challenge. *Neo4j*, 2012/10/11.

[22] S. Vercruysse, A. Venkatesan, and M. Kuiper. Olsvis: an animated, interactive visual browser for bio-ontologies. *BMC Bioinformatics*, 13(1):116, 2012.

[23] G. Vetere, J. M. Gomez-Perez, and H. Wu. *Exploiting linked data and knowledge graphs in large organisations*. Springer Berlin Heidelberg, New York NY, 2016.

[24] C. Wang, X. Ma, J. Chen, and J. Chen. Information extraction and knowledge graph construction from geoscience literature. *Computers & Geosciences*, 112:112–120, 2018.

[25] S. S. Yudkoff. Monuments of poetry: On the publication history of avrom sutzkever's sibir. *Shofar: An Interdisciplinary Journal of Jewish Studies*, 38(1):218–258, 2020.

[26] X. Zou. A survey on application of knowledge graph. *Journal of Physics: Conference Series*, 1487(1):012016, 2020.
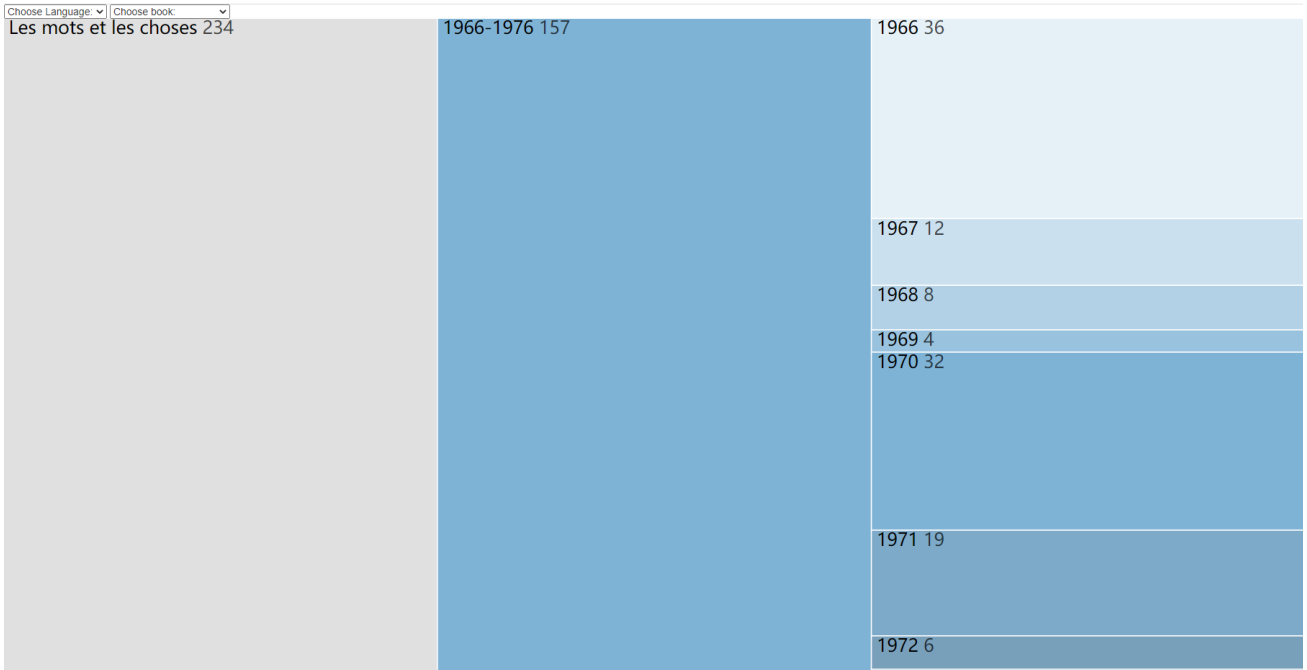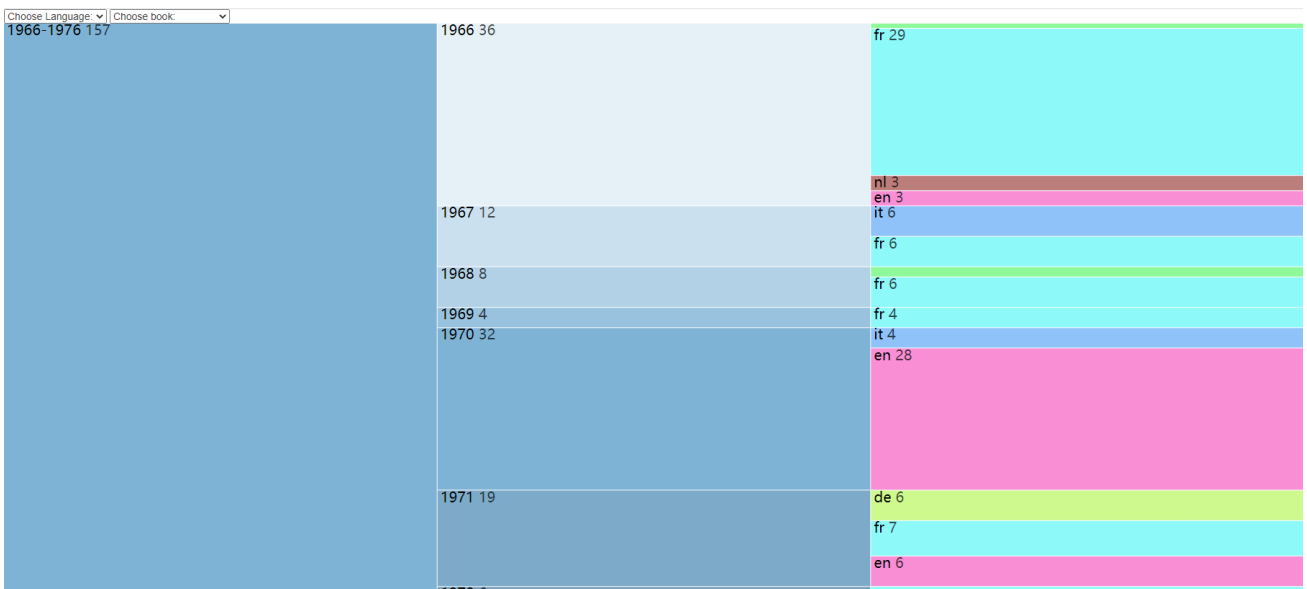
Figure 3. Workflow of prototype 1

Figure 4. Workflow of prototype 2

Figure 5. Workflow of prototype 3

APPENDIX

A. USER TEST AND QUESTIONNAIRE

A.1 Task

In principle, the participants had to learn SPARQL language to finish this test, consider no one would like to spend more than five days for that, the test will compare the performance between Worldcat website and the project. Test book is the grand design. The method is to measure the time-consuming to do the task in both websites.

1. Find the OCLC control number of the Polish (pl) version published in 2015

2. Find the title of the Portuguese(pt) version published in 2011.

3. Find the translator of the Finnish(fi) version published in 2011.

A.2 Questionnaire

IV: the new interactive visualization.

LB: the old list based visualization.

1. Which tool do you think can save more time?

   A IV   B LB

2. Which tool do you think is more convenient to use?

   A IV   B LB

3. Which tool do you think can provide a better overview?

   A IV   B LB

4. Do you have any suggestion for IV?