Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 813–816
Marseille, 11–16 May 2020

WeDH - a Friendly Tool for Building Literary Corpora Enriched with Encyclopedic Metadata

Mattia Egloff, Davide Picca
University of Lausanne
Switzerland
{ mattia.egloff, davide.picca}@unil.ch

Abstract

In recent years the interest in the use of repositories of literary works has been successful. While many efforts related to Linked Open Data go in the right direction, the use of these repositories for the creation of text corpora enriched with metadata remains difficult and cumbersome. In fact, many of these repositories can be useful to the community not only for the automatic creation of textual corpora but also for retrieving crucial meta-information about texts. In particular, the use of metadata provides the reader with a wealth of information that is often not identifiable in the texts themselves. Our project aims to fill both the access to the textual resources available on the web and the possibility of combining these resources with sources of metadata that can enrich the texts with useful information lengthening the life and maintenance of the data itself. We introduce here a user-friendly web interface of the Digital Humanities toolkit named WeDH with which the user can leverage the encyclopedic knowledge provided by DBpedia, wikidata and VIAF in order to enrich the corpora with bibliographical and exegetical knowledge. WeDH is a collaborative project and we invite anyone who has ideas or suggestions regarding this procedure to reach out to us.

Keywords: Corpora, Digital Humanities, LinkedOpenData

1. Introduction

In recent years the interest in the use of repositories of literary works has been successful. There is a large number of tools and software packages providing access to data repositories such as NLTK (Loper and Bird, 2002) or Spacy[1]. However, many of these resources are not powerful enough to exploit this data to their full extent.

While many efforts related to Linked Open Data go in the right direction, the use of these repositories for the creation of text corpora remains difficult and cumbersome. Many of them are not easily accessible since they barely provide the user with API or graphical interfaces. If we take the Gutenberg.org project[2] as an example, some attempts have been made to make this resource available in a systematic and structured way. There have been some attempts in order to make this resource accessible and linked to other data resources in the LOD Cloud (Auer et al., 2007) but at the moment, unfortunately, Bizer's work[3] does not seem to have been active since 2007 and, to our knowledge, there are no other projects aimed at systematically connecting the Gutenberg resource to the LOD resources for metadata retrieval. In fact, many of these repositories can be useful to the community not only for the automatic creation of textual corpora but also for retrieving crucial meta-information about texts. In particular, the use of meta-information provides the reader with a wealth of information that is often not identifiable in the texts themselves. Such information is often crucial for scholars in the literary or philological field because it conveys a considerable amount of information that decisively guides the interpretation and study of the texts themselves. The loss of such a patrimony of information necessarily implies a degradation in the interpretative finesse of exegetical investigations.

If we consider the study of exegetical comments in texts as an example, information such as the first date of publication or the first publisher is of vital importance and often such information is totally absent from these repositories. Moreover, the existing tools do not provide the user with metadata retrieval capabilities, limiting the extent of the study's field. A good integration of data and metadata not only has a positive impact on the investigative methods but also on the maintenance of the data itself. In fact, metadata help in data maintenance at least at two levels. On the one hand, they extend data longevity. As we have seen before, missing or unavailable relevant metadata renders texts useless for research purposes. On the other hand, metadata facilitates data reuse and sharing. Metadata is the cornerstone of guaranteeing that high-detail data is more easily interpreted, analyzed and processed by others in an easier way.

Thus, in the current state of affairs, we have two types of gaps to fill: firstly, the access to the textual resources available on the web, and secondly, the possibility of combining these resources with sources of metadata that can enrich texts with useful information lengthening the life and maintenance of the data itself.

Our project aims to fill both gaps by offering at the same time a tool for the automatic construction of corpora enriched with a meta-informative catalog. We introduce here a user-friendly web interface of the Digital Humanities ToolKit (DHTK) (Picca and Egloff, 2017) named WeDH[4] with which users can leverage the encyclopedic knowledge provided by Dpedia, wikidata and VIAF in order to enrich the corpora with bibliographical and exegetical knowledge.

[1] https://spacy.io

[2] https://www.gutenberg.org/

[3] https://www.lod-cloud.net/dataset/fu-berlin-project-gutenberg

[4] https://dhtk.unil.ch/WeDH/ For testing purpose as demo account has been set using the following credentials: *User*: demo@dhtk.unil.ch, *Password*: DemoWeDH

WeDH was conceived with the intent to offer a tool which is able to meet the needs of both communities. Built on the DHTK library (Picca and Egloff, 2017), WeDH is written in Python and proposes similar objectives which are detailed in section 3.. In particular, DHTK's main purpose is to provide the human scientist with a tool that leverages on the main semantic repositories as DBpedia (Auer et al., 2007) to complete annotation and search for metadata (e.g., the year of the first edition, main characters, book categories, etc.). In this paper we describe WeDH, outlining its architecture, its modules and some case studies to better highlight its potential and its use.

2. Related works

Although there are several tools in the field of the digital humanities, WeDH differs particularly in its vocation to combine the creation of literary corpora enriched with exegetical and bibliographic data. While Bibliopedia[5] (Cenkl and Widner, 2013) searches resources including JSTOR and Library of Congress for metadata about scholarly articles and books, it focuses exclusively on the works mentioning the famed medieval travel narrative The Travels of Sir John Mandevilly. Many other tools such as Voyant[6], Philologic[7] or GutenTag[8] (Brooke et al., 2015) are more centered on performing text analysis and TEI-codification of texts neglecting the metadata enrichment which is, instead, the main focus of WeDH. Although there are some technical features that overlap, in particular with the GutenTag tool, our software retains as its main objective the automatic construction of corpora featured by knowledge-enhanced encyclopedia such as DBpedia or VIAF. Moreover, other tools such as TAPoR[9] (Carlin, 2005) present numerous useful and user-friendly options for literary scholars, but their focus is on individual texts or small groups of texts instead of providing the user with the possibility of retrieving massive knowledge-enriched corpora. With respect to the other available tools, WeDH mainly focuses on gathering data and metadata in a structured manner in order to improve the exploitability of literary texts along with their metadata freely available on the LOD.

3. WeDH - a friendly tool for building literary corpora

WeDH is mainly the user interface for the DHTK library conceived and proposed by Picca and Egloff (2017) Thus, if the main purpose of the library is to facilitate the exploitation of textual repositories such as Gutenberg.org along with of LOD resources such as DBpedia, wikidata and VIAF, the Web interface has been conceived in order to be exploited by students and practitioners in the human science field with no or few coding skills. In fact, WeDH is conceived as a high-end user interface that provides a graphical access to textual and metadata repositories that can be easily accessed. Given the simplicity of WeDH, it could be adopted by universities as a tool for Digital Humanities courses specifically designed for humanists. The user interface has been built in accordance with some key principles that serve as a guideline for the entire development of the tool. The main features are:

- Easily create and manage corpora.
- Retrieve author's and books' metadata.
- Download the texts and metadata of a corpus.
- Enhance the metadata adding LOD links.

We will examine the general architecture of the tool, dwelling on each module that composes it showing how each of the points mentioned above is performed by WeDH.

3.1. General architecture

WeDH is designed to constantly grow and improve. In fact, some additional modules have already been planned for implementation in the near future, such as the addition of other textual and encyclopedic repositories like Europeana (Valtysson, 2012). Hence, WeDH is a work-in-progress project and, for the time being, the user can search for authors or works in the RDF Gutenberg Catalog as well as the Gitenberg Repository[10]. Since the RDF catalog provided by Gutenberg is not sufficiently reliable, metadata are retrieved using several other metadata repositories as shown in Figure 2. As shown in Table 2, the links to DBpedia directly available in Gutenberg are more contained than those caught by the searching algorithm implemented in DHTK. WeDH guarantees full compatibility with major text processing packages such as NLTK, Spacy or GutenTag thanks to the universal outputs encoded in json and txt formats. In Figure 1, the main workflow performed by the DHTK library is shown. The workflow functions in the following way: the user queries against different criteria such as literary periods or genres in one or more books and, after the search, WeDH automatically retrieves the metadata present in DBpedia or VIAF. WeDH builds the final corpus leaving the last step to the final user's choice who can decide whether to further process the files with any text processor.

As WeDH is a web service, the general architecture is built on a structure with a front-end on one side and a back-end on the other. The back-end is mainly built on two main modules: the *Text Resource* module and the *Metadata Retriever* module. The former manages the construction of corpora, while the latter manages the retrieval of encyclopedic meta-information from Dbpedia, WikiData and VIAF. We will describe both of them in further detail in the following section.

3.2. Text Resources module

This module is designed to contain the textual resources available on the Gutenberg Repository. We initially focused on Project Gutenberg because of its relevance in the humanist community and we found a lack of automatic tools available in order to exploit this specific resource. The main

[5] http://sul-cidr.github.io/Bibliopedia/
[6] https://voyant-tools.org/
[7] https://sites.google.com/site/philologic3/
[8] https://gutentag.sdsu.edu/
[9] http://tapor.ca/home
[10] https://www.gitenberg.org/

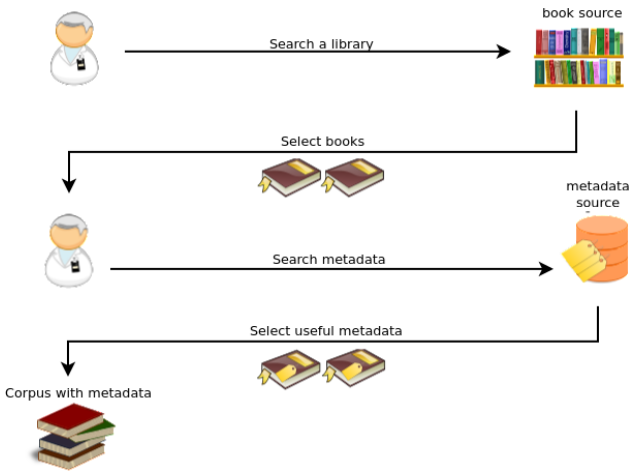

Figure 1: The workflow schema behind the WeDH interface

purpose of this module is to facilitate access to texts. At the time of writing this article, in Table 1 we report some basic statistics on data obtained from the results automatically crawled by DHTK. This module searches for texts

# books	# authors	# bookshelves	# categories
59137	20903	337	17867

Table 1: Basic statistics on Gutenberg Project

in Gutenberg to be linked to resources available in DBpedia, Wikidata, or VIAF. Such a linkage is provided by the Metadata Search module as is described hereunder.

3.3. Metadata search

The main purpose of this module is to ensure the connection between the textual resource and its metadata. As previously mentioned, one of the most important aspects for a humanist is not just the availability of texts but also the metadata coming with texts. Unfortunately, not all repositories have complete information such as the date of the first publication or the first publisher. For example, Project Gutenberg does not store the original publishing date of its books, which can represent a problem if a corpus needs to be delimited to a decade of published literature. In order to overcome this problem, we rely on some external resources available through the LinkedOpenData such as DBpedia or VIAF. DBpedia, thanks to its encyclopedic nature, helps to rethink these elements. The metadata search module on DBpedia allows to complete and expand the missing information. This task is performed by the *DBpediaMetada* Class. Since the covering provided by Gutenberg and DHTK are still not fully satisfactory as shown in Table 2, we introduced both the VIAF and Gitenberg resources since they have the feature of being manually checked, thereby resulting in a wider range of support. Moreover, Gitenberg relies on multilingual Wikipedia resources while Gutenberg contains only the English resource. Nonetheless, since Gitenberg does not provide a direct link to the DBpedia resource, a transformation of the multilingual Wikipedia links has been manually done beforehand for each available language.

Source	Links towards	authors	books
Gutenberg rdf	wikipedia.org	46119	-
DHTK	dbpedia.org	7344	2553
VIAF	viaf.org/viaf	19508	-
Gitenberg	github.com/GITenberg/	-	59061
Gitenberg	dbpedia.org	64587	14933
Gitenberg	wikipedia.org	30022	1130

Table 2: External Metadata links

In addition, the user always has the possibility to manually edit the links. Every correction made manually in the metadata is used to improve the quality of the data by leveraging on the collective intelligence provided by the WeDH users' community.

4. Client side

The WeDH application is written in Python using the Flask web application framework. Flask is modern, widely used and provides a high level of support for the development of web applications that are not based on extraordinarily complex but rather highly versatile data models. The choice of deployment environment was based on the widespread use of this environment and the consequent support for hosting these applications as well as the ease of finding developers with the appropriate skills to maintain and extend the software. Moreover, any web browser can be used as a client ensuring both portability and intuitive access to the system. The web interface has been designed to allow users to access different lexical resources available in a uniform manner, hiding the format of the data and their implementation on the server side.

4.1. Corpus access and creation

The client application mainly provides two interfaces: a detailed view for searching for literary texts and one for editing texts and collecting metadata.
The first screen enables the user to easily retrieve literary texts from the Gutenberg repositories (See Figure 2). Once logged in, the user directly accesses the main text retrieval interface from where the person can search according to several criteria such as: search by author, by title, by literary subject or even by Gutenberg category (called 'Bookshelves') (see the top of the Figure 2). At the bottom of the screen, in addition to the search results, the titles of the books selected to be part of the corpus are visible. Each selected book can also be removed from the corpus at any time (see the bottom of the Figure 2).

4.2. Metadata editing

Once the corpus with the collection of titles has been created, it is possible not only to retrieve the metadata already available for each book according to the numbers shown in the Table 1 but it is always possible to add a DBpedia, VIAF or wikidata resource to your corpus, if the system does not provide the appropriate link. An interesting aspect of this feature is the possibility to manually add the metadata links bound to the corresponding text resource, thus becoming

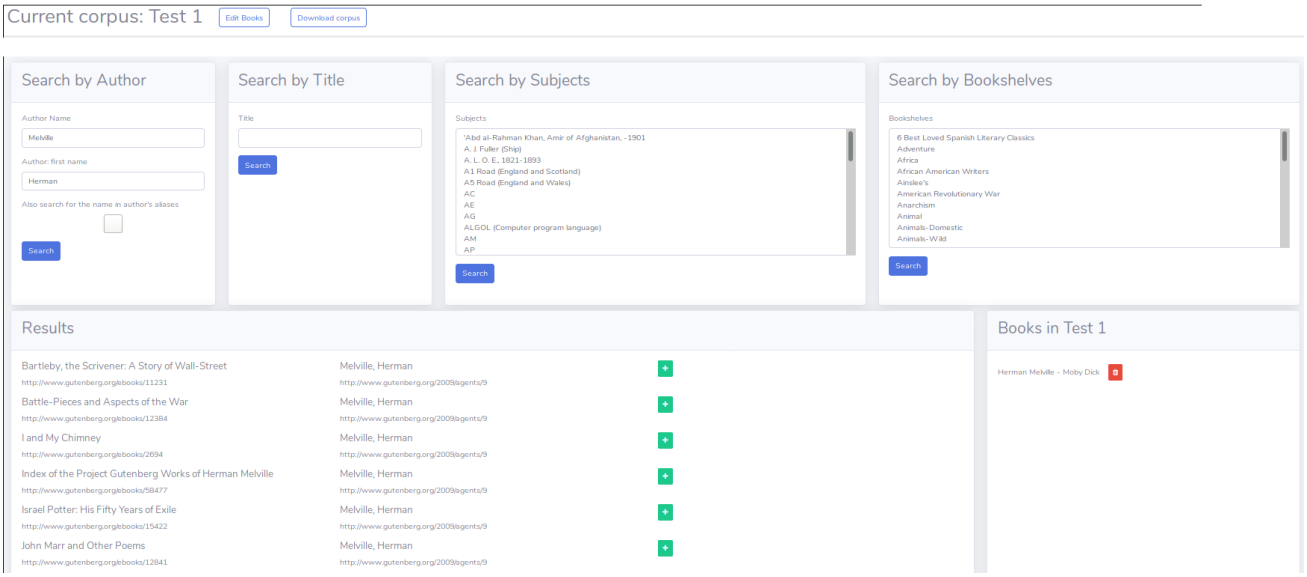

Figure 2: A simple way to search and add books to a corpus.

immediately available for later use. Once the desired metadata have been selected, the user can download the set of data (texts in txt format) and metadata in a single action in order to have available for each file containing the text, a file in json format with all the corresponding metadata.

5. Future work and conclusions

In this paper we have described WeDH, a work-in-progress web tool built on the DHTK (Picca and Egloff, 2017) library aiming to provide non-computer science specialists with a tool to access textual resources and metadata available on LOD. Being a work-in-progress, only Gutenberg as data source and DBpedia, VIAF and wikidata as metadata sources are available for the moment. The user interface is constantly improving and some features are already planned. In particular, the option to select specific metadata for download as well as the option to download texts in TEI format, are already in production. Another feature under development is the possibility to download metadata about the author. In addition, we are working to integrate other repositories like DBpedia, wikidata and VIAF not only as metadata sources but also as books' information retrieval.

In the future, DHTK has the ambition to go beyond textual resources to integrate other human resources such as images (e.g., paintings, comics, etc.) or sounds (e.g., music, transcripts, etc.). Since WeDH relies on DHTK, it is vital to improve this library in order to enhance WeDH's features. WeDH is also currently used in teaching for Digital Humanities courses. One of the objectives for the future, is to assess the pedagogical impact of this tool so as to extend its use not only to the areas of research but also to teaching.

As said above, we have been constantly improving DHTK and, since it is a collaborative project, we invite anyone who has ideas or suggestions regarding this project to reach out to us.

6. Bibliographical References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. G. (2007). Dbpedia: A nucleus for a web of open data. In *In proceedings of 6th International Semantic Web Conference (ISWC 2007)*, pages 722 – 735.

Brooke, J., Hammond, A., and Hirst, G. (2015). Gutentag: an nlp-driven tool for digital humanities research in the project gutenberg corpus. In *CLfL@NAACL-HLT*.

Carlin, C. (2005). Drawing knowledge from information: Early modern texts and images on the tapor platform. *Digital Studies / Le champ numerique*.

Cenkl, P. T. and Widner, M. (2013). Bibliopedia, linked open data, and the web of scholarly citations. In *Digital Humanities 2013, DH 2013, Conference Abstracts, University of Nebraska-Lincoln, Lincoln, NE, USA, July 16-19, 2013*, page 145.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP 2002, pages 63 – 70. Association for Computational Linguistics.

Picca, D. and Egloff, M. (2017). Dhtk: The digital humanities toolkit. In *Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II) co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017.*, pages 81 – 86.

Valtysson, B. (2012). EUROPEANA : The digital construction of Europe's collective memory. *Information, Communication and Society*, 15(2):151–170.