Scholarly Digital Editions: APIs and Reuse Scenarios

Spadini, Elena
elena.spadini@unibas.ch
Universität Basel, Switzerland

Losada Palenzuela, José Luis
joseluis.losadapalenzuela@unibas.ch
Uniwersytet Wrocławski, Poland

In this paper, we study data reuse in scholarly editing, providing insights into the current panorama and imagining future developments. We will focus on the reuse of data, leaving aside the reuse of code and models, which would require a separate enquiry; and will concentrate on machine-actionable reuse, as opposed to human consumption. Although the reusability of editions' data should be considered within the framework of research data management, many related issues, such as licensing, will be left out from this paper in order to limit its scope.

Obtaining data

We first address technological means to access the editions data for reuse, including web scraping, data dumps and APIs. A comparative analysis of these approaches highlights their advantages and limitations, with special attention to the presence of metadata and documentation.

We then focus on APIs for editions. A first distinction can be made between generic APIs (e.g. DTS, TEI Publisher) and project-specific ones (e.g. *Carl Maria von Weber Gesamtausgabe*, *The Proceedings of the Old Bailey*). We will address the advantages of using a standardised interface, namely following the OpenAPI Specification (https://spec.openapis.org/oas/v3.1.0, used by TEI Publisher and by *Carl Maria von Weber Gesamtausgabe*), and of providing query examples (as in the case of the *Registres de la Comédie-Française*and their Postman query collection). We will consider the scope of the query within a textual document, distinguishing between the APIs that query the XML structure and anchors (e.g., the first) and those querying the content of data and metadata (e.g., documents by date). We believe that the combination of these two approaches would be beneficial. Examples of API for retrieving content (not querying the XML tree directly) are those implemented in *edition humboldt digital*, *Registres de la Comédie-Française*, *Sandrart.net*, *The Folger Shakespeare*, *The Proceedings of the Old Bailey*.

We will also consider available parameters to refine the queries and the support for internal and external IDs, such as GND, Viaf, Geonames, Wikidata.

Eventually, we reflect on the suitability of standard APIs for editions and their possible features, building on top of the important contribution of DTS (Distributed Text Services).

Reusing data

In the second part of the paper, we describe concrete (real and fictitious) cases to exemplify reuse practices. Five reuse scenarios will be presented:

- Search multiple datasets with one authority record. Persistent identifiers and authority records are key to retrieve data for reuse. In this example, we retrieve multiple non-overlapping data from *Carl-Maria-von-Weber-Gesamtausgabe. Digitale Edition* and from *CorrespSearch* using GND authority records.
- Editions data in dictionaries. While dictionaries are sometimes integrated into (or linked to from) editions, the contrary is rare. We will explore some of the few cases in which an electronic dictionary references an electronic edition, namely the *Dictionnaire Étymologique de l'Ancien Français* and the *Dictionary of Old Norse Prose*.
- Detecting text reuse. Using the example of the Spanish Golden Age theatre, we will demonstrate how for certain distant reading practices it is important to take into account the type of sources to be included in the corpus, and the value of the curated text of scholarly editions in this context.
- Enriching editions with gazetteers. We will analyse the enrichment of data from the *Perseus Digital Library* with geolocations from the *Pleiades* gazetteer in the project *ToposText*.
- Internal reuse. Reuse can also occur within the edition project or in a follow-up, when data is used for replication and maintenance, or for purposes other than the edition itself. The two cases of data visualisation and of rebuilding an obsolete web application will be addressed.

Conclusions

The results of this study are presented in the form of simple, but still not widely implemented, suggestions to editors: provide multiple access to the data and in multiple formats; when this is not possible, make at least available a data dump in a standard format, such as XML/TEI or plain text; implement internal persistent identifiers and enrich data with external persistent identifiers; offer documentation. Additional remarks are presented on how the access to the data influences the type of reuse.

Some open questions will also be explored. The first concerns versioning and replicability, which apparently clash with our suggestion of multiplying the access points to data. Two of the projects mentioned in the first section, *The Proceedings of the Old Bailey*and *Registres de la Comédie-Française*, exemplify this tension, since both signal to the user that data retrieved through the API and from the data dumps “might represent slightly different versions”. The problem, however, is not that the data are exposed in different ways, but that the open-endedness of digital editions comes with certain disadvantages, among which the "perpetual beta status" (Gengnagel 2017). Moreover, the different versions of the data, of the software (including the API, if any) and sometimes even of the model, seem to make a complete replicability over time impossible to achieve.

The second open question concerns the reasons why the reuse of data from editions is still rare. Here, the nature of the data itself might play a role, as Humanities data are always the result of an interpretation, built according to scientific choices that might not be shared by others. The complexity of obtaining the data for a not

technologically savvy user can also be an obstacle, as well as the fact that reuse is at odds with the solitary way of working of many humanists up to the current days.

Bibliography

Complete Works of Carl Maria von Weber. Digital Edition (Version 4.7.0 of February 19, 2023). http://weber-gesamtausgabe.de/A070002 [1.05.2023]

Dumont, Stefan / **Grabsch, Sascha** / **Müller-Laackman, Jonas** (2021): *CorrespSearch – Briefeditionen Vernetzen (2.0.0)* [Webservice]. https://correspSearch.net [1.05.2023].

Emsley, Clive / **Hitchcock, Tim** / **Shoemaker, Robert** (2018): "Digital Projects Using Old Bailey Online Data", in: *Old Bailey Proceedings Online*, Version 8.0 | March 2018. https://www.oldbaileyonline.org/static/Projects.jsp [1.05.2023].

Franzini, Greta (2019): "Towards Connecting Scholarly Editions to Corpora in the LiLa (Linking Latin) Knowledge Base of Linguistic Resources". DOI: 10.5281/zenodo.3613371.

Franzini, Greta / **Terras, Melissa** / **Mahony, Simon** (2019): "Digital Editions of Text: Surveying User Requirements in the Digital Humanities", in: *Journal on Computing and Cultural Heritage* 12, 1: 1–23. DOI: 10.1145/3230671.

Glessgen, Martin / **Dallas, Marguerite** (2019): "L'intégration du vocabulaire des Documents linguistiques galloromans dans le DEAF électronique", in: *Lexicographica* 35: 235–67. DOI: 10.1515/lex-2019-0009.

Losada Palenzuela, José Luis (2022): "Análisis Cuantitativo de La Reutilización Textual En Las Comedias Colaboradas (Moreto)". DOI: 10.5281/zenodo.7271369.

Spadini, Elena (2022): "Reuse of Digital Editions. Exposing Data", in: *Éditions critiques numériques et multilinguisme*, Université de Montréal. DOI: 10.5281/zenodo.7271447.

Tittel, Sabine (2018): "Historical Corpus and Historical Dictionary: Merging Two Ongoing Projects of Old French by Integrating Their Editing Systems", in: #ibej, Jaka / Gorjanc, Vojko / Kosem, Iztok / Krek, Simon (eds.): *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.* Ljubljana: Ljubljana University Press 453–65.