exhibitions, artworks series, literary works and agents in their main life events, roles, places, and dates.

WRITE data model documentation[6]and some representative case studies (one artwork for each collection) are available to foster the understanding and reusability of the model. The contribution of the dataset is twofold: first, it supports scholars in investigating the WRITE domain by recording and classifying its data to allow a systematic analysis of its resources to help the domain experts answer their main research questions (e.g. the definition of a spectrum of forms and practices ranging from traditional calligraphy to modern/contemporary artistic expression: can those forms/practices still be defined as calligraphy?) and, hopefully, discover new knowledge; second, it is a challenge in the modelling of different sources from GLAM by categorizing and connecting the visual dimension of iconographic elements (calligraphy as visual artwork) with the respective textual elements (lexical features, text transcription and translation, literary sources of text) to attribute an actual meaning to the new concept of "calli-writing units" through the analysis of the hybridization of multimedia resources.

Bibliography

Bachi, V., Fresa, A. and Veselič, M. (2021). PAGODE – Europeana China. In Ioannides, M., Fink, E., Cantoni, L. and Champion, E. (eds), *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*. (Lecture Notes in Computer Science). Cham: Springer International Publishing, pp. 265–77 doi: 10.1007/978-3-030-73043-7_22.

Bekiari, C., Doerr, M., Le Bœuf, P. and Riva, P. (2015). *FRBR Object-Oriented Definition and Mapping from FRBRER, FRAD and FRSAD (Version 2.4)*. Final Report International working group on FRBR and CIDOC CRM harmonisation http://www.cidoc-crm.org/frbroo/fm_releases.

Charles, V. and Isaac, A. (2015). Enhancing the Europeana data model (EDM). *EDM WHITE PAPER*.

Dijkshoorn, C., Jongma, L., Aroyo, L., Ossenbruggen, J. van, Schreiber, G., Weele, W. ter and Wielemaker, J. (2018). The Rijksmuseum collection as Linked Data. *Semantic Web*, **9**(2). IOS Press: 221–30 doi: 10.3233/SW-170257.

Doerr, M., Ore, C.-E. and Stead, S. (2007). The CIDOC Conceptual Reference Model - A New Standard for Knowledge Sharing ER2007 Tutorial. , **83**: 51–56 doi: https://doi.org/10.13140/2.1.1420.6400.

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J. and Vrandečić, D. (2014). Introducing Wikidata to the Linked Data Web. In Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K. and Goble, C. (eds), *The Semantic Web – ISWC 2014*. (Lecture Notes in Computer Science). Cham: Springer International Publishing, pp. 50–65 doi: 10.1007/978-3-319-11964-9_4.

Iezzi, A. (2013). Contemporary Chinese Calligraphy Between Tradition and Innovation. *Journal of Literature and Art Studies*, **3**: 158–79 doi: 10.17265/2159-5836.

Iezzi, A. (2015). What is 'Chinese Modern Calligraphy'? An Exploration of the Critical Debate on Modern Calligraphy in Contemporary China. *Journal of Literature and Art Studies*, **5**: 206–16 doi: 10.17265/2159-5836/2015.03.007.

Li, J. (2020). Omeka Classic vs. Omeka.net. *Emerging Library & Information Perspectives*, **3**(1): 232–36 doi: 10.5206/elip.v3i1.8625.

Notes

1. See https://writecalligraphyproject.eu/
2. The project WRITE - *New Forms of Calligraphy in China: A Contemporary Culture Mirror* is an European Research Council (ERC) Starting Grant funded project based in the Department of Interpreting and Translation of the Alma Mater Studiorum – University of Bologna (GA n. 949645).
3. See https://www.britishmuseum.org/search?search_api_fulltext=chinese+calligraphy
4. See https://www.mediawiki.org/wiki/Wikibase/DataModel
5. OmekaS is a web publishing platform to collaboratively create collections with a shared pool of online resources.
6. See https://write-dataset.github.io/documentation/

Extracting and providing online access to annotated and semantically enriched historical data.The AGODA project

Puren, Marie Anna
marie.puren@epitech.eu
Epitech, MNSHS, France; Centre Jean-Mabillon, Ecole nationale des chartes, France

Vernus, Pierre
pierre.vernus@msh-lse.fr
LARHRA, France; Université Lyon 2, France

Pellet, Aurélien
aurelien.pellet@epitech.eu
Epitech, MNSHS, France

Bourgeois, Nicolas
nicolas.bourgeois@epitech.eu
Epitech, MNSHS, France

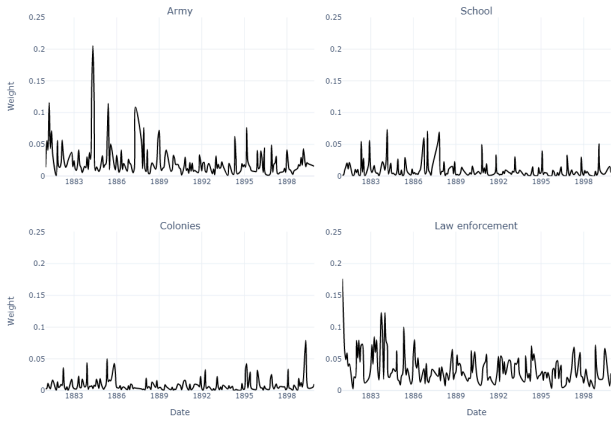
The AGODA project [1] (Puren and Vernus, 2021) is one of five pilot projects supported by the DataLab of the Bibliothèque nationale de France. It aims to create an online platform facilitating the exploration and use of the parliamentary debates of the Chamber of Deputies published in the *Journal officiel* from 1881 to 1940. In the framework of the DataLab, we are working on a test sub-corpus, namely the parliamentary cycle from 1889 to 1893, to test our hypotheses on a smaller dataset.

Over the past sixty years, a great deal of work has been done on parliamentary debates (Chester and Bowring, 1962; Franklin and Norton, 1993). It is indeed a valuable source for historians (Marnot, 2000; Ouellet and Roussel-Beaulieu, 2003; Ihalainen, 2016; Lemercier, 2021), political scientists (Van Dijk, 2010), sociologists (Cheng, 2015) or linguists (de Galembert et al., 2013; Hirst et al., 2014; Rheault et al., 2016). Access to digitised and ocerised debates thus seems to have a positive effect on the number of historical works using these documents (Mela et al., 2022). The same effect can be observed for other disciplines using contemporary debates (Fišer et al., 2018; Fišer et al., 2020). AGODA is thus part of a wider movement to facilitate the use and analysis of parliamentary data, following the example of ParlaClarin (Fišer and Lenardič, 2018) and ParlaMint (Erjavec et al., 2022a; Erjavec et al., 2022b), which propose to produce comparable and multilingual Parliamentary Proceedings Corpora according to the XML-TEI standard. Naomi Truan has also produced a corpus of parliamentary debates encoded in XML-TEI (Truan, 2016; Truan and Romary, 2021). The production of this type of resource facilitates the publication of works exploiting this data to better understand French political discourse (Diwersy et al., 2018; Blaette et al., 2020; Diwersy and Luxardo, 2020).

Between 1881 and 1899, 2596 issues of the *Journal Officiel* were published (50791 JPG images). The debates are also in TXT format but put online without extensive post-correction: the quality of the OCR is not sufficient to provide a satisfactory online browsing experience, and it could have a negative impact on the analyses performed on these texts (van Strien, 2020). Therefore, we chose to ocerise the text, to obtain a better-quality result. We use the PERO OCR (Kodym and Hradiš, 2021; Kohút and Hradiš, 2021; Kišš et al., 2021) based solution developed by the SODUCO project [2]. This tool, still in private alpha version, has been used to prepare the data in (Abadie et al., 2022) that will be accessible via Zenodo.

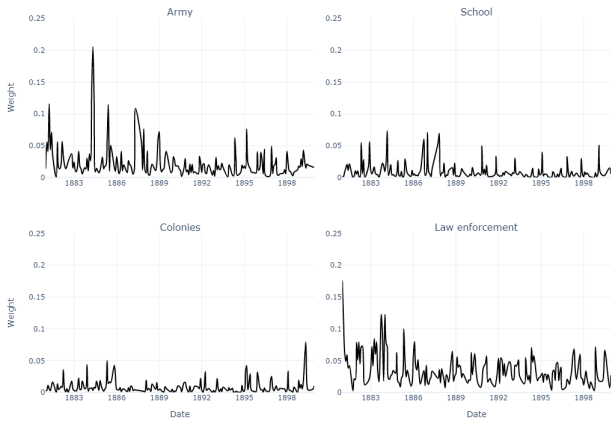
Ocerised texts are obtained in JSON format; we are developing Python scripts to convert this output into an XML file corresponding to the chosen TEI model. This model is formalised with an adapted XML schema, created using an ODD (Rahtz and Burnard, 2013). We chose to use the ODD created by ParlaClarin (Erjavec and Pančur, 2021) which can be easily adapted to annotate historical parliamentary debates. In the case of France, the rules for transcribing debates were set in the 19th century; thus, the recordings of today's debates are very similar to those produced during the Third Republic. The TEI-encoded corpus will be stored in an eXist-db database, and it will be visualised using the TEI Publisher application, which can transform the source data into HTML web pages. The parliamentary debates will thus be made available to online users as a digital edition and integrated into an application context.

We will also present the first analyses we have carried out on this corpus with "bag-of-words" techniques - these being not too sensitive to the quality of the OCR. We first used topic modelling, an unsupervised learning method that allows us to discover the latent semantic structures of a corpus of texts, without using semantic and lexical resources (Blei et al., 2003). This method is well suited to study parliamentary debates (Bourgeois et al., 2022).



Distribution of four different topics over time

Alternatively, we can use word embeddings to reduce the dimension of the original space from several tens of thousands of forms to a hundred axes, and then apply classical data science tools such as clustering or correlation analysis on the reduced space (Mikolov et al., 2013). Word embedding has thus shown its interest in the study of parliamentary debates (Rheault and Cochrane, 2020). We used a continuous bag-of-words model for dimension reduction and an unsupervised classification algorithm - in this case DBSCAN - to group words into clusters.



t-SNE projection of the centroïds of the clusters

Bibliography

Abadie, N., Carlinet, E., Chazalon, J., Dumenieu, B. (2022). A Benchmark of Named Entity Recognition Approaches in Historical Documents. Application to 19th Century French Directories. DAS 2022 15th IAPR International Workshop on Document Analysis Systems. La Rochelle, France. May 22-25, 2022.

Blaette, A., Gehlhar, S. and Leonhardt, C. (2020). The Europeanization of Parliamentary Debates on Migration in Austria, France, Germany, and the Netherlands. Proceedings of the Second ParlaCLARIN Workshop. Marseille, France: European Language Resources Association, pp. 66–74 https://aclanthology.org/2020.parlaclarin-1.12 (accessed 21 April 2022).

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3: 993–1022.

Cheng, J. E. (2015). Islamophobia, Muslimophobia or racism? Parliamentary discourses on Islam and Muslims in debates on the minaret ban in Switzerland. Discourse & Society, **26**(5). SAGE Publications.

Chester, D. N. and Bowring, N. (1962). Questions in Parliament. Oxford: Clarendon Press.

Diwersy, S., Frontini, F. and Luxardo, G. (2018). The Parliamentary Debates as a Resource for the Textometric Study of the French Political Discourse. Proceedings of the ParlaCLARIN@LREC2018 Workshop. Miyazaki, Japan https://hal.archives-ouvertes.fr/hal-01832649 (accessed 21 April 2022).

Diwersy, S. and Luxardo, G. (2020). Querying a large annotated corpus of parliamentary debates. LREC, ParlaCLARIN Workshop. (Proceedings of the Second ParlaCLARIN Workshop). Marseille, France https://hal.archives-ouvertes.fr/hal-03317717 (accessed 21 April 2022).

Erjavec, T. and Pančur A. (2021) Parla-CLARIN: A TEI Schema for Corpora of Parliamentary Proceedings https://clarin-eric.github.io/parla-clarin/ (accessed 21 April 2022).

Erjavec, T., Pančur A. and Kopp M. (2022a). ParlaMint: Comparable Parliamentary Corpora. GLSL CLARIN ERIC https://github.com/clarin-eric/ParlaMint (accessed 21 April 2022).

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., et al. (2022b). The ParlaMint corpora of parliamentary proceedings. Language Resources and Evaluation doi:10.1007/s10579-021-09574-0.

Fišer, D., Eskevich, M. and Jong, F. de (eds). (2018). Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Paris: European Language Resources Association (ELRA).

Fišer, D., Eskevich, M. and Jong, F. de (eds). (2020). Proceedings of the Second ParlaCLARIN Workshop. Marseille: European Language Resources Association (ELRA).

Fišer, D. and Lenardič, J. (2018). CLARIN resources for parliamentary discourse research. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), pp. 2–7.

Galembert, C. de, Rozenberg, O., Vigour, C. (eds) (2013). Faire parler le parlement: méthodes et enjeux de l'analyse des débats parlementaires pour les sciences sociales. Paris: LGDL-Lextenso.

Hirst, G., Feng, V., Cochrane, C. and Naderi, N. (2014). Argumentation, Ideology, and Issue Framing in Parliamentary Discourse. In Cabrio E, Villata S. and Wyner A. S. (eds), Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing, Forlì-Cesena, Italy, July 21-25, 2014, CEUR-WS.org, http://ceur-ws.org/Vol-1341/paper6.pdf (accessed 26 April 2022).

Ihalainen, P., Ilie, C. and Palonen, K. (2018). Parliament and Parliamentarism: A Comparative History of a European Concept, New York, Oxford: Berghahn.

Ilie, C. (2010). European Parliaments under Scrutiny: Discourse Strategies and Interaction Practices. Amsterdam; Philadelphia: John Benjamins.

Kišš, M., Beneš, K. and Hradiš, M. (2021). AT-ST: Self-training Adaptation Strategy for OCR in Domains with Limited Transcriptions. In Lladós, J., Lopresti, D., Uchida, S. (eds) Document Analysis and Recognition – ICDAR 2021. ICDAR 2021. Lecture Notes in Computer Science, vol 12824. Cham: Springer, https://doi.org/10.1007/978-3-030-86337-1_31 (accessed 26 April 2022).

Kodym, O. and Hradiš, M. (2021). Page Layout Analysis System for Unconstrained Historic Documents.

Document Analysis and Recognition – ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II. Berlin, Heidelberg: Springer-Verlag, pp. 492–506.

Kohút, J. and Hradiš, M. (2021). TS-Net: OCR Trained to Switch Between Text Transcription Styles. In Lladós, J., Lopresti, D., Uchida, S. (eds) Document Analysis and Recognition – ICDAR 2021. ICDAR 2021. Lecture Notes in Computer Science, vol 12824. Cham: Springer, https://doi.org/10.1007/978-3-030-86337-1_32 (accessed 26 April 2022).

La Mela, M., Norén, F., and Hyvönen, E. (2022). Digital parliamentary data in action (DiPaDA 2022), workshop co-located with the 6th Digital Humanities in the Nordic and Baltic countries conference (DHNB 2022), https://dhnb.eu/conferences/dhnb2022/workshops/dipada/ (accessed 26 April 2022).

Lemercier, C. (2021). Un catholique libéral dans le débat parlementaire sur le travail des enfants dans l'industrie (1840). Parlement[s], Revue d'histoire politique, **33**(1): 195–206.

Marnot, B. (2000). Les ingénieurs au Parlement sous la IIIe République. Paris: CNRS Editions.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [Cs] http://arxiv.org/abs/1301.3781 (accessed 26 April 2022).

Ouellet, J. and Roussel-Beaulieu, F. (2003). Les débats parlementaires au service de l'histoire politique. Bulletin d'histoire politique, **11**(3). Bulletin d'histoire politique: 23–40 doi:10.7202/1060736ar.

Puren, M. and Vernus, P. (2021). AGODA : Analyse sémantique et Graphes relationnels pour l'Ouverture et l'étude des Débats à l'Assemblée nationale. Inauguration Du BnF DataLab. Paris, France https://hal.archives-ouvertes.fr/hal-03382765 (accessed 26 April 2022).

Rahtz, S. and Burnard, L. (2013). Reviewing the TEI ODD system. Proceedings of the 2013 ACM Symposium on Document Engineering. (DocEng '13). New York, NY, USA: Association for Computing Machinery, pp. 193–96.

Rheault, L., Beelen, K., Cochrane, C. and Hirst, G. (2016). Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. PLOS ONE, **11**(12). Public Library of Science: e0168843 doi:10.1371/journal.pone.0168843.

Rheault, L. and Cochrane, C. (2020). Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. Political Analysis, **28**(1). Cambridge University Press: 112–33 doi:10.1017/pan.2019.26.

Strien, D. A. van, Beelen, K., Ardanuy, M. C., Hosseini, K., McGillivray, B. and Colavizza, G. (2020). Assessing the Impact of OCR Quality on Downstream NLP Tasks. ICAART doi:10.5220/0009169004840496.

Study of parliament group (GB), F., Mark N. and Norton, P. (1993). Parliamentary Questions. Oxford: Clarendon Press.

Truan, N. (2019). Débats parlementaires sur l'Europe à l'Assemblée nationale (2002-2012) [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v1.1, https://hdl.handle.net/11403/fr-parl/v1.1 (accessed 21 April 2022c).

Truan, N. and Romary, L. (2021). Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A Cross-Linguistic Account. Journal of the Text Encoding Initiative https://halshs.archives-ouvertes.fr/halshs-03097333 (accessed 21 April 2022).

Notes

1. https://github.com/mpuren/agoda
2. https://soduco.github.io/

Traven between the impostors. Preliminary considerations on an authorship verification case

Rebora, Simone
simone.rebora@univr.it
University of Verona, Italy

Salgaro, Massimo
massimo.salgaro@univr.it
University of Verona, Italy

This paper sets up the groundwork for an authorship verification project dedicated to the German novelist B. Traven, author of novels such as *The Death Ship*(1926) and *The Treasure of the Sierra Madre*(1927), whose real identity is still a mystery. Among the different theories, the most established is the one that sees B. Traven as the pseudonym of Otto Feige, author of a series of political pamphlets for the metal workers' union in Gelsenkirchen, who then changed his name into Ret Marut (publisher of the anarchist periodical Der Ziegelbrenner), before moving to Mexico and acquiring his final, world-famous pseudonym (Goldwasser, 1993; Hauschild, 2012).

From the point of view of stylometry, that of Feige/Matut/Traven is a typical authorship verification problem, where the goal is not that of attributing an anonymous text to a candidate author, but that of verifying if two (or more) texts were written by the same author. Extensive research