Journal of open archaeology data

Data From "A Biocodicological Analysis of the Medieval Library and Archive From Orval Abbey, Belgium"

DATA PAPER

JORSUA HERRERA BETHENCOURT
ISMAEL RODRÍGUEZ PALOMO
SIMON HICKINBOTHAM
BHARATH NAIR
SILVIA SONCIN
MARC DIEU
MATTHEW COLLINS
OLIVIER DEPARIS

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The dataset contains the first-ever comprehensive biocodicological analysis of medieval library books and charters using Zooarchaeological Mass Spectrometry (ZooMS). Here, we analyze 68 codices and 59 charters (1490+59 samples in total) from one single monastic institution, namely the Cistercian abbey of Orval in present-day Belgium. The data entails both peptide mass fingerprinting (using MALDI-ToF) and peptide sequencing (using LC-MS/MS) analysis of almost the entire library and all the preserved single leaf charters from the monastery. MALDI-ToF data is stored in Zenodo – a multidisciplinary open access repository while the LC-MS/MS data is deposited in ProteomeXchange Consortium via PRIDE – a publicly available repository for MS-based proteomics data. Mass spectrometric data generated from an entire monastic library and archive is of immense value to integrate with multiple case studies aiming at understanding parchment production and use in medieval Europe.

Paper linked with data:

Ruffini-Ronzani, N., Nieus, J.F., Soncin, S., Hickinbotham, S., Dieu, M., Bouhy, J., Charles, C., Ruzzier, C., Falmagne, T., Hermand, X., Collins, M.J. and Deparis, O 2021. A biocodicological analysis of the medieval library and archive from Orval Abbey, Belgium. *Royal Society Open Science*, *8*(6), p.210210.

CORRESPONDING AUTHOR:

Jorsua Herrera Bethencourt

Section for Evolutionary Genomics, Globe Institute, University of Copenhagen, DK

jherrera@palaeome.org

KEYWORDS:

Manuscripts; open science; charters; collagen; parchment; mass spectrometry; open dataset; open access; liquid chromatography

TO CITE THIS ARTICLE:

Herrera Bethencourt J, Rodríguez Palomo I, Hickinbotham S, Nair B, Soncin S, Dieu M, Collins M, Deparis, O 2022 Data From "A Biocodicological Analysis of the Medieval Library and Archive From Orval Abbey, Belgium". *Journal of Open Archaeology Data,* 10: 1, pp. 1–7. DOI: *https://doi.org/10.5334/joad.89*

(1) OVERVIEW

CONTEXT

Zooarchaeological Mass Spectrometry (ZooMS) has been widely applied to archaeological contexts and materials of cultural interest [1, 2, 4, 5]. In particular, the application of ZooMS on parchment, being the most common writing support in Medieval times, has been successfully used to investigate questions about its production, local economies, the use of the manuscripts and their conservation [3, 5, 6, 11, 12]. The study of parchment from a biological perspective has become so widespread that the name "Biocodicology" has been recently proposed to refer to the discipline [7]. Moreover, while fulfilling compliance with conservation standards in terms of non-invasiveness, the rapidity of sampling, thanks to triboelectric extraction [5], but also of sample preparation and analysis, often allows researchers to collect a large amount of samples, with the consequent development of new bioinformatic methods to also speed up the data analysis [8].

The article "*A biocodicological analysis of the medieval library and archive from Orval Abbey, Belgium*" presents the results obtained from the analysis of almost a complete corpus of books and charters from the library of the Cistercian abbey of Orval, Belgium, mostly over a period ranging from 12th to 13th centuries. In total, 1490 folia were analysed by ZooMS and 86 by peptide sequencing. The dataset we provide here is the largest proteomic dataset generated from a single scriptorium to date (***Figure 1***). For completeness we include all the analysis, thus there are replicates of a number of samples. These include instrument performance (repeat of a whole MALDI-ToF plate) or sample failure (replicate analysis of specific sample).

This large dataset allowed us to explore differences in the selection of animal skins in and outside the Orval scriptorium and the possible reasons behind the choice of sheep/goat or calf skin in the production [10].

This dataset with consistent metadata defined by a single codicological team, and analysed using MS1 (folia) and MS2 (charters) can be used to explore alternative tools for discriminating closely related species, and for refining tools for high throughput analysis of MALDI-ToF (ZooMS) data.

Spatial coverage

Description: The dataset in this study was generated from book and charter samples from one single monastic institution, namely the Cistercian abbey of Orval in present-day Belgium. Sixty-eight codices, representing 118 codicological units in total, were classified in the Orval manuscripts catalogues and belong to the collection of the National Library of Luxembourg, and 59 charters belong to the Belgian State Archives in Arlon.

Temporal coverage

The majority of the data were generated from samples of manuscripts and charters from between the ninth to seventeenth century, though mostly twelfth to thirteenth century.

(2) METHODS

SAMPLING STRATEGY

Sampling of parchment was conducted by non-invasive triboelectric extraction of collagen following a previously described method [5]. Briefly sampling procedure of parchment entailed gentle rubbing of non-written areas of parchment surface with an eraser followed by collection of the eraser crumbs (typically 10 to 50 mg PVC) in a 1.5 mL Eppendorf tube. We advise that nitrile gloves be used, and a freshly cut piece of PVC eraser is used for each sample. The eraser crumbs are transferred to clean Eppendorf tubes using acid free paper prior to analysis. Samples were stored in 4 °C or room temperature. As far as codices were concerned, samples were taken systematically on the recto of the first folium of each quire composing the manuscripts. For the charters, one sample was taken from the single leaf composing each charter. Subsequently the samples were subjected to peptide mass fingerprinting using MALDI-ToF and peptide sequencing using LC-MS/MS (***Figure 1***).

Steps

Peptide mass fingerprinting: All samples except manuscript 22 (1463 samples in total) were analysed by MALDI-ToF at BioArch laboratory at the University of York using a Bruker Ultraflex III mass spectrometer at the Center of Excellence in mass spectrometry. Scripts for processing the data were written in R. These scripts used the bacollite package for R, which resolves species ID on the basis of alignment with theoretical MALDI peaks that are generated from peptide data [8]. Briefly each peptide is aligned to each replicate spectra using cross-correlation, yielding a correlation value by exploring the number of matches to each of the test species at each of a number of threshold levels. In order to come up with a score for each taxa, the number of hits are summed for each correlation threshold so that for a given threshold, only the highest scoring taxa will be added, while the rest add 0. Therefore when a taxon has consistently more hits than the rest, it will get a higher score, whereas it will be lower in a more ambiguous case. The resulting automation has an associated confidence value that can be used to compare the automated and expert-based classifications. The comparison demonstrated that the automated approach to classification is identical to the expert-based classification where the confidence score is above 10 (for this peptide set). The benefit of this

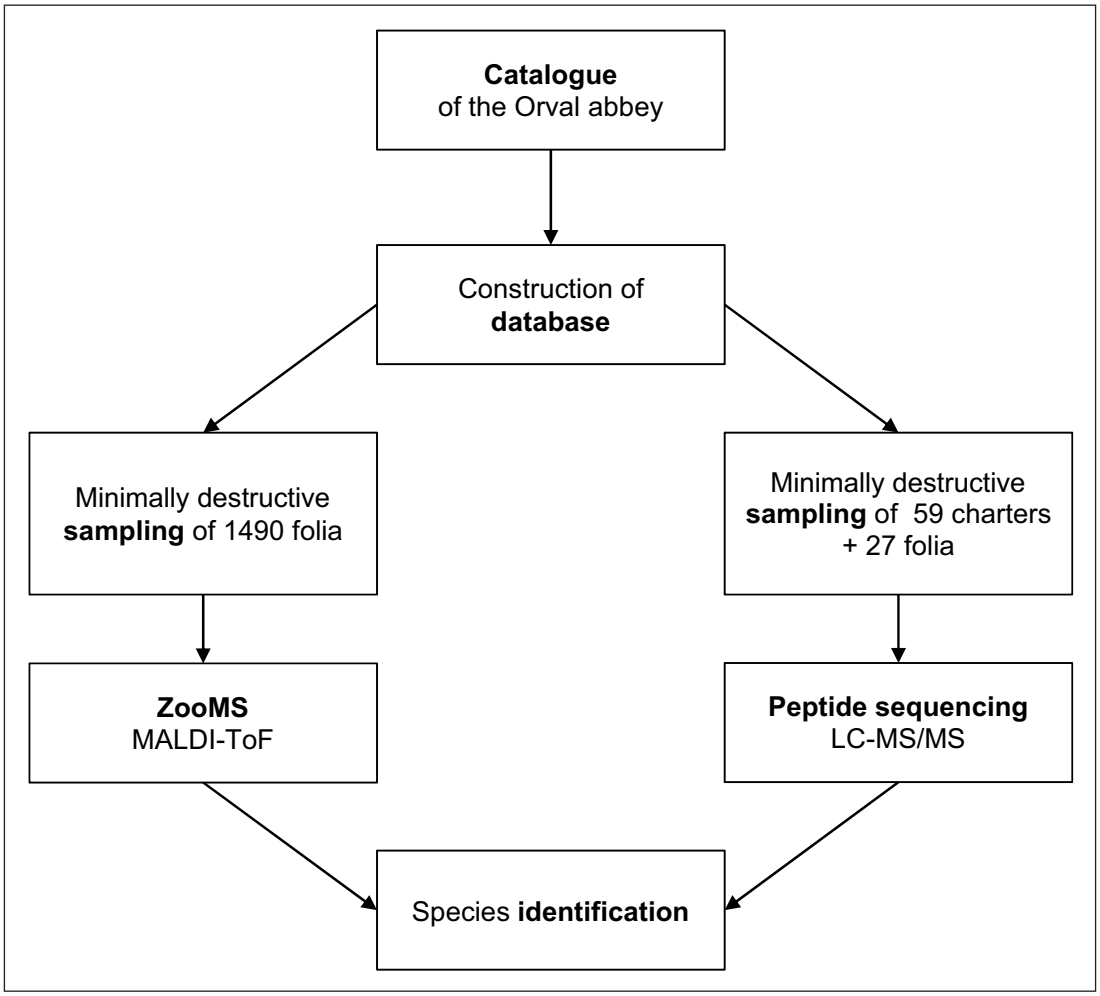

Figure 1 Workflow describing the steps followed for the generation of this dataset.

approach is that the expert analysis can be directed towards the interpretation of less clear spectra in a principled manner. The scripts used for the analysis are provided in appendix I below.

Peptide sequencing by LC-MS/MS: Samples from all charters (59 samples) and manuscript 22 (27 samples) were analysed by liquid chromatography (UltiMate 3000, ThermoFisher) coupled to electrospray ionisation tandem mass spectrometry (MaXis Impact UHR-, Bruker) at MaSUN mass spectrometry platform at the University of Namur, Belgium. Compass HyStar 3.2 (Bruker) was used to control the instruments. Raw files (.mgf format) were generated using the software DataAnalysis 4.0 (Bruker) for downstream data analysis using Mascot 2.4 (Matrix Science). Scaffold (version Scaffold_4.8, Proteome Software Inc., Portland) was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 90.0% probability by the Peptide Prophet algorithm.

Protein identifications were accepted if they could be established at greater than 95.0% probability and contained at least 2 identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm [9].

Quality Control

1. Each sampling was performed with new eraser and nitrile gloves.
2. To avoid cross contamination the table was cleaned with Isopropanol.
3. Proteomic analyses were repeated if it was not possible to attribute to a taxon.

Constraints

Not all bifolia in each codicological unit (CU) were analysed.

(3) DATASET DESCRIPTION

OBJECT NAME

In the present work the dataset is available in the following format:

1. *metadata.csv* – of each codicological unit.
2. *metadata.xlsx* – Excel version of the metadata.
3. *species_id.csv* – Species identification of each quire within the codicological unit.
4. *species_id.xlsx* – Excel version of the species identification.

5. *Orval_Bacollite_example_for_datapaper.Rms*. R markdown notebook with the script to generate the automatic species identification with bacollite.
6. *Orval_Bacollite_example_for_datapaper.Rmd*. Knitted html version of the notebook.
7. **.mzML files of MALDI-ToF data* containing spectra from each sample as mass and intensity vectors. The name of these files is encoded as ms_cu_q_r.mzML where *ms* stands for the manuscript number, *cu* is the codicological unit within the manuscript as a roman numeral, and *q* is the quire number within the codicological unit. Some quires were analyzed several times, this occurred if the sample size of erase crumbs was too small or due to an issue with the instrumental analysis (MALDI-ToF), and these replicate analyses are indicated by adding a -bis, -ter, ... suffix, or a number preceded by two underscores, as in "__1", "__2, ... In the case of sample size, this sample would fail. If the issue was with the instrument, typically a whole plate of samples would fail. Each sample is analyzed in triplicates within the same MALDI plate. Each replicate is identified by a final number preceded by an underscore, "_1", "_2" and "_3".
6. *LC-MS/MS data* was uploaded to the PRIDE server as .baf for raw files and .mgf format for peak list, and .dat files for Mascot result. Scaffold files were also included (.sf3).

Below is a detailed list of fields of the metadata file in which we have included the codicological information analyzed for each of the manuscripts. The column name and units or range, where relevant, is in parenthesis.

- *Manuscript number (ms):* Manuscript books from a single medieval Cistercian monastery (Orval Abbey, Belgium). The manuscripts are classified in the Orval manuscript catalogue and belong to the collection of the National Library of Luxembourg. The manuscript number refers to its number in the catalogue: Falmagne T. 2017 Die Orvaler Handschriften bis zum Jahr 1628 in den Beständen der Bibliothèque Nationale de Luxembourg und des Grand Séminaire de Luxembourg. Wiesbaden, Germany: Harrassowitz.
- *CU label (cu):* For this field we define each codicological unit (CUs) as a volume, a part of a volume, or a set of volumes whose production may be considered a single operation, prepared in one place, at one time using the same available resources.
- *Number of quires (n_quires):* Number of quires is a collection of leaves of parchment or paper, folded one within the other, in a manuscript.
- *Orval local production:* This field contains information regarding where they were produced. 26 of the 118 CUs were produced according to an analysis of codicological elements performed by T. Falmagne. 26 of the 118 were produced locally by the Orval scriptorium according to this analysis.
- *Production period (prod_period):* This field contains information about the historical period in which the library books and charters were produced. They were produced between the ninth to seventeenth century, though mostly twelfth to thirteenth century.
- *Origin (origin):* In this field is displayed the geographical place where the manuscripts and charters were produced. For this study we consider codicological units produced by the Orval scriptorium and outside.
- *Height (mm):* height of the book as reported in the catalogue.
- *Width (mm):* width of the book as reported in the catalogue.
- *Nbre of folia (n_folia):* This field encloses information regarding the number of folia contained in a codicological unit.
- *Typology (typology):* This field is based on the manuscripts topic. For the present study, eight types of texts were defined: bible, liturgy, grammar and rhetoric, sciences, narrative texts, law, preaching and theology. Also, we have included a category 'Other' which was composed of normative texts and letter collections.
- *Thickness index (thinckness_idx, 1–4):* Thickness is based on the number of folia a given CU contains. CUs with less than 10 folia were considered as 'very thin', (index 1). Those containing between 11 and 100 folia as 'thin' (index 2), 'medium' (index 3), when they counted between 101 and 200 folia, and finally they were regarded as 'thick', (index 4) when they counted more than 200 folia.
- *Quality index (quality_idx, 0–10):* This field contains information based on layout, scribe's skill (calligraphy) and decoration of the texts. Based on these features, we have set up a score ranging from 0 to 10: 'very low quality' (score lower than 3), 'low quality' (score higher than or equal to 3 and lower than 5), 'medium quality' (score higher than or equal to 5 and lower than 6.5), 'good quality' (score higher than 8) and 'superior quality' (score higher than 8).

Data type

Primary and secondary data.

Format names and versions

*.XLSX, *.CSV, *.mzML, *.baf, *.mgf, *.dat, *.sf3

Creation dates

Dataset created between November 2017 to August 2019.

Dataset Creators

The codicology assessment, identification of codicological units and measurement of bifolia was conducted by Thomas Falmagne, (Universitätsbibliothek Frankfurt am Main, t.falmagne@ub.uni-frankfurt.de). Nicolas Ruffini-Ronzani (nicolas.ruffini@unamur.be) with the support of Chiara Ruzzier (chiara.ruzzier@unamur.be, Université de Namur) created the database based on Falmagne's work, Catherine Charles (catherine.charles@unamur.be, Université de Namur) defined the categories based on the measurements from the original catalogue and measured the thickness of the charters.

Catherine Charles with Julie Bouhy (julie.bouhy@unamur.be, Université de Namur) and Olivier Deparis (olivier.deparis@unamur.be, Université de Namur) were responsible for sampling.

Silvia Soncin, (silvia.soncin@york.ac.uk, University of York) conducted ZooMS analysis (MALDI-ToF) and with Simon Hickinbotham (simon.hickinbotham@york.ac.uk, York) and Matthew Collins, Ismael Rodríguez Palomo and Bharath Nair (Copenhagen University) were responsible for data generation and checking conversion. Marc DIEU (marc.dieu@unamur.be,Université de Namur) generated LC-MS/MS data with the support of Julie Bouhy.

Olivier Deparis (olivier.deparis@unamur.be, Université de Namur) was in charge of the project and coordinated data collection.

Language

English

License

Repository location

Zenodo last version repository
DOI: *https://doi.org/10.5281/zenodo.5583377*
PRIDE: PXD029196

Publication date

02 June 2021

(4) REUSE POTENTIAL

The data generated and described here is the first of its kind, a complete analysis of monastic library and archive described by a single codicological team, which can be used as a reference dataset for multiple coherent case studies with an intent to understand parchment production and use in medieval Europe. Consequently, aggregating datasets from the analysis of new manuscripts and charters from all around Europe is necessary for a large-scale study that can provide insights into medieval literacy and intellectual life. The dataset can be used for validating downstream bioinformatic analyses, e.g., aiming at assessing the quality of parchment production from the MALDI-ToF spectra relating it to the metadata and potentially answering novel questions.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Appendix I:** Orval data analysis using Bacollite. DOI: *https://doi.org/10.5334/joad.89.s1*

ACKNOWLEDGEMENTS

The authors acknowledge funding from the Fondation Roi Baudouin, Belgium (Fonds Jean-Jacques Comhaire). They are grateful to the National Library of Luxembourg and the Belgian State Archives for granting access to their collections. O.D. thanks Luke Spindler (BioArCh, University of York) for ZooMS analyses on very first samples. O.D., J.-F.N., N.R.-R. thank J. Vnouček for enlightening discussion on parchment fabrication and visual inspection.

FUNDING INFORMATION

O.D. acknowledges funding from the Fondation Roi Baudouin, Belgium (Fonds Jean-Jacques Comhaire, Grant Agreement No. 2016-P2813310-206264). M.J.C. acknowledges funding from Beasts to Craft (ERC Horizon 2020 Grant Agreement No. 787282) as well as Danish National Research Foundation (DNRF128).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Jorsua Herrera Bethencourt led and helped write the paper.
Ismael Rodríguez Palomo compiled and processed the data and metadata, and also helped write the paper.
Simon Hickinbotham prepared the R scripts and processed the data.
Bharath Nair compiled and processed the data and metadata, and also helped write the paper.
Silvia Soncin performed the MALDI-ToF analysis and helped write the paper.

Marc Dieu performed the LC-MSMS analysis of ms22 and charters, processed the data for the PRIDE repository.
Matthew Collins helped write the paper.
Olivier Deparis helped write the paper.

AUTHOR AFFILIATIONS

Jorsua Herrera Bethencourt *orcid.org/0000-0001-7341-7868*
Section for Evolutionary Genomics, Globe Institute, University of Copenhagen, DK

Ismael Rodríguez Palomo *orcid.org/0000-0001-5313-9709*
Section for Evolutionary Genomics, Globe Institute, University of Copenhagen, DK

Simon Hickinbotham *orcid.org/0000-0003-0880-4460*
Department of Computer Science, University of York, York, UK

Bharath Nair *orcid.org/0000-0002-1897-4132*
Section for Evolutionary Genomics, Globe Institute, University of Copenhagen, DK

Silvia Soncin *orcid.org/0000-0002-9700-5259*
BioArCh, Department of Archaeology, University of York, York, UK

Marc Dieu *orcid.org/0000-0002-3902-542X*
University of Namur, MaSUN – Mass Spectrometry Facility, University of Namur, 61, Rue de Bruxelles, 5000, Namur, BE

Matthew Collins *orcid.org/0000-0003-4226-5501*
Section for Evolutionary Genomics, GLOBE Institute, Faculty of Health and Medical Science, University of Copenhagen, Øster Farimagsgade 5, 1353 Copenhagen, DK; McDonald Institute for Archaeological Research, University of Cambridge, West Tower, Downing St, CB2 3ER Cambridge, UK

Olivier Deparis *orcid.org/0000-0002-2161-7208*
Department of Physics, and Heritage, Transmissions and Inheritances Institute, University of Namur, 61 rue de Bruxelles, Namur 5000, BE

REFERENCES

1. **Brandt LØ, Mannering, U.** Taxonomic identification of Danish Viking Age shoes and skin objects by ZooMS (Zooarchaeology by mass spectrometry). *Journal of proteomics*, 2021; 231, (Jan. 2021): 104038. DOI: *https://doi.org/10.1016/j.jprot.2020.104038*
2. **Buckley M, Collins M, Thomas-Oates J, Wilson JC.** Species identification by analysis of bone collagen using matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry*, 2009; 23, 23 (2009): 3843–3854. DOI: *https://doi.org/10.1002/rcm.4316*
3. **Doherty SP, Henderson S, Fiddyment S, Finch J, Collins MJ.** Scratching the surface: the use of sheepskin parchment to deter textual erasure in early modern legal deeds. *Heritage Science*, 2021; 9, 1 (Mar. 2021): 29. DOI: *https://doi.org/10.1186/s40494-021-00503-6*
4. **Ebsen JA, Haase K, Larsen R, Sommer DVP, Brandt LØ.** Identifying archaeological leather--discussing the potential of grain pattern analysis and zooarchaeology by mass spectrometry (ZooMS) through a case study involving medieval shoe parts from Denmark. *Journal of cultural heritage*, 2019; 39, (2019): 21–31. DOI: *https://doi.org/10.1016/j.culher.2019.04.008*
5. **Fiddyment S,** et al. Animal origin of 13th-century uterine vellum revealed using noninvasive peptide fingerprinting. *Proceedings of the National Academy of Sciences of the United States of America*, 2015: 112, 49 (Dec. 2015): 15066–15071. DOI: *https://doi.org/10.1073/pnas.1512264112*
6. **Fiddyment S, Goodison NJ, Brenner E, Signorello S, Price K, Collins MJ.** Girding the loins? Direct evidence of the use of a medieval English parchment birthing girdle from biomolecular analysis. *Royal Society open science*, 2021; 8, 3 (Mar. 2021): 202055. DOI: *https://doi.org/10.1101/2020.10.21.348698*
7. **Fiddyment S, Teasdale MD, Vnouček J, Lévêque É, Binois A, Collins MJ.** So you want to do biocodicology? A field guide to the biological analysis of parchment. *Heritage Science*, 2019; 7, 1 (Jun. 2019): 35. DOI: *https://doi.org/10.1186/s40494-019-0278-6*
8. **Hickinbotham S, Fiddyment S, Stinson TL, Collins MJ.** How to Get Your Goat: Automated Identification of Species from MALDI-ToF Spectra. *Bioinformatics*, 2020; (Mar. 2020). DOI: *https://doi.org/10.1093/bioinformatics/btaa181*
9. **Nesvizhskii AI, Keller A, Kolker E, Aebersold R.** A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry*, 2003; 75, 17 (Sep. 2003): 4646–4658. DOI: *https://doi.org/10.1021/ac0341261*
10. **Ruffini-Ronzani N, Nieus J-F, Soncin S, Hickinbotham S, Dieu M, Bouhy J, Charles C, Ruzzier C, Falmagne T, Hermand X, Collins MJ, Deparis O.** A biocodicological analysis of the medieval library and archive from Orval Abbey, Belgium. *Royal Society open science*, 2021; 8, 6 (Jun. 2021): 210210. DOI: *https://doi.org/10.1098/rsos.210210*
11. **Teasdale MD, Fiddyment S, Vnouček J, Mattiangeli V, Speller C, Binois A, Carver M, Dand C, Newfield TP, Webb CC, Bradley DG, Collins MJ.** The York Gospels: a 1000-year biological palimpsest. *Royal Society open science*, 2017; 4, 10 (Oct. 2017): 170988. DOI: *https://doi.org/10.1101/146324*
12. **Vnouček J, Fiddyment S, Quandt A, Rabitsch S, Collins M, Hofmann C.** The parchment of the Vienna Genesis: characteristics and manufacture. *The Vienna Genesis Material analysis and conservation of a Late Antique illuminated manuscript on purple parchment*, 2020. Boehlau Verlag GmbH & Co. KG. 35–70. DOI: *https://doi.org/10.7767/9783205210580.35*

TO CITE THIS ARTICLE:
Herrera Bethencourt J, Rodríguez Palomo I, Hickinbotham S, Nair B, Soncin S, Dieu M, Collins M, Deparis, O 2022 Data From "A Biocodicological Analysis of the Medieval Library and Archive From Orval Abbey, Belgium". *Journal of Open Archaeology Data,* 10: 1, pp. 1–7. DOI: *https://doi.org/10.5334/joad.89*

Published: 22 February 2022