# PLAN OVERVIEW

*A Data Management Plan created using DMPonline.be*

**Title:** Paternity unveiled: genetic genealogy on premarital offspring in 19th century Western Europe

**Author:** Brecht Liefsoens, Maarten H.D. Larmuseau

**Principal Investigator:** Maarten H.D. Larmuseau

**Data Manager:** Brecht Liefsoens

**Project Administrator:** Brecht Liefsoens, Maarten H.D. Larmuseau

**Affiliation:** KU Leuven (KUL)

**Template:** KU Leuven BOF-IOF

**Project abstract:**
In an era of evolving family structures and assisted reproduction, the multifaceted nature of fatherhood, encompassing legal, social, and biological dimensions, is profoundly resonant in today’s academic and societal discussions. This research project delves into the historical context of biological fatherhood, focusing on 19th-century illegitimacy in Western Europe, where ca. 7-8% of children were born to unwed mothers and mostly later legitimized by their mothers' husbands. While microdata suggests that these husbands were often the biological fathers, the extent and nature of these relationships are unexplored. Our genetic genealogical study employs a citizen science approach, analyzing 500 ‘genealogical pairs’ of male relatives with a case of a premarital child born in 19th-century Belgium. Using Y-chromosome comparison and archival microdata, we aim to unravel the scale and motivations behind men's decisions to legitimize children in the 19th century, particularly in cases where they were not the biological fathers.

**ID:** 212618

**Start date:** 01-10-2024

**End date:** 30-09-2028

**Last modified:** 03-03-2025

# PATERNITY UNVEILED: GENETIC GENEALOGY ON PREMARITAL OFFSPRING IN 19TH CENTURY WESTERN EUROPE

**RESEARCH DATA SUMMARY**

**List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.**

| Dataset name / ID | Description | New or reuse | Digital or Physical data | Data Type | File format | Data volume | Physical volume |
|---|---|---|---|---|---|---|---|
| 1: Personal data citizen scientists | Database with personal data of citizen scientists (name, contactinfo, experience level) | N | D | T | CSV | < 1GB | |
| 2: Cit. Sci.-Y-genealogies | Collection of genealogies in paternal line, produced by citizen scientists | N | D | T | PDF | < 100GB | |
| 3: Cit. Sci.-context premarital children | Collection of contextinformation around premarital children, produced by citizen scientists | N | D | T | PDF | < 100GB | |
| 4: Saliva kits | Saliva samples from selected citizen scientists (ca. 1.000) | N | P | Other: biological sample | | | Small |
| 5: Y-reads | Collection of all reads collected during sequencing of the DNA from the saliva samples | N | D | T | FASTQ | initially: < 20 TB<br><br>after mapping, filtering: < 2 TB | |
| 6: Y-genotypes | Database of sequenced Y-genotypes from saliva samples | N | D | T | CSV | < 100GB | |
| 7: Pedigrees | Database of the genealogies of identified premarital children, containing the modern-day pairs for genetic analyses and context information | N | D | T | CSV | < 1GB | |
| 8: Modelling data | Database used for evolutionary modelling | N | D | T | CSV | < 1GB | |
| 9: Cit. Sci.-surveys | Results from periodical surveys of citizen scientists | N | D | T | CSV | < 1GB | |

**If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:**

Not applicable

**Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.**

- Yes, human subject data (Provide SMEC or EC approval number below)

Datasets no. 4, 5 and 6 are based on personal genetic data and thus require ethical approval by a federally acknowledged ethical committee.
We are in the process of preparing a request for ethical appraisal at the Ethics Committee Research UZ / KU Leuven (EC Research).

**Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).**

- Yes (Provide PRET G-number or EC S-number below)

Dataset no. 1 will contain personal data and will serve as the key file for the pseudonymisation of other datasets. Each of the participants will receive a persistent identifier (PID) that will serve to connect the different datasets. Upon receiving datasets no. 2 & 3, all personal information relating to the citizen scientist and to living relatives will be replaced by the PIDs from dataset no.1.
The identifiers of the samples from dataset no. 4 will also be added to dataset no. 1, as to maintain a single PID for each person implicated in the study.
Datasets no. 5, 6, 7 and 8 are derivatives of the already mentioned pseudonymised datasets, therefore they will be pseudonymised from their conception.
Dataset no. 9 will be the result of an anonymous survey among participants, so no personal information will be recorded intentionally.
Thus, the most vulnerable personal data will only be available in dataset no. 1, containing risks of data breaches to a single document. Appropriate measures to ensure the safety of this file will be taken (encryption and safe storage location).
We are in the process of preparing PRET appraisal (G-2025-9139).

**Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.**

- No

**Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or Data transfer agreements, Research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.**

- No

**Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.**

- No

**DOCUMENTATION AND METADATA**

**Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g. in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, codebook.tsv etc. where this information is recorded).**

Each dataset will be accompanied by a README file outlining its conception context, general form, contents, and what projects and publications it was used for.

**Will a metadata standard be used to make it easier to find and reuse the data?**
**If so, please specify which metadata standard will be used.**

**If not, please specify which metadata will be created to make the data easier to find and reuse.**

- Yes

The data we intend to share will be published at KU Leuven Research Data Repository, hence DataCite will be used as metadata standard.

**DATA STORAGE & BACK-UP DURING THE RESEARCH PROJECT**

**Where will the data be stored?**

- Sharepoint online
- Other (specify below)

(Pseudonymised) datasets 2, 3, and 6 through 9 will be stored at our labs' SharePoint. Dataset no. 1 containing personal information and the key to pseudonymisation will be kept on the projectmanagers' OneDrive provided by KU Leuven, during the course of the project. Afterwards it will pass on to the PI's OneDrive.
Dataset 5 will be stored on the Cloud Storage service provided by Genomics Core.

**How will the data be backed up?**

- Standard back-up provided by KU Leuven ICTS for my storage solution

**Is there currently sufficient storage & backup capacity during the project?**

**If no or insufficient storage or backup capacities are available, explain how this will be taken care of.**

- Yes

As of this moment, only this project will be using the lab's SharePoint is only reserved for this specific project. Should this change in the future, other options will be explored, such as the KU Leuven network drives.
The current Gold Tier membership Cloud Storage provided by Genomics Core is sufficient for this project.

**How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?**

The SharePoint Team site is only accessible for members of our laboratory. Moreover, the datasets stored there will not contain sufficient information to link it back to specific participants. The pseudonymisation key will be kept separate from these datasets at all times during and after the project.

**What are the expected costs for data storage and backup during the research project? How will these costs be covered?**

The costs associated with the Cloud Storage provided by Genomics Core are accounted for in the research project budget. For all other datasets in this research project, KU Leuven currently provides sufficient free storage.

## DATA PRESERVATION AFTER THE END OF THE RESEARCH PROJECT

**Which data will be retained for 10 years (or longer, in agreement with other retention policies that are applicable) after the end of the project?**

**In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).**

- All data will be preserved for 10 years according to KU Leuven RDM policy

The saliva kits from dataset no. 4 will be destroyed once they have lost their purpose (namely: after sequencing).
Datasets no. 1, 2, 3, 5 and 9 will be kept for at least 10 years to ensure reproducibility, verification and the potential for future reuse or re-evaluation.
Datasets no. 6 and 7 will be kept indefinitely because of their potential uses for future research and their low risk of data breaches.

**Where will these data be archived (stored and curated for the long-term)?**

- KU Leuven RDR
- Other (specify below)

Datasets 2, 3, and 6 through 9 datasets will be archived at our SharePoint Team site for future projects. Dataset 5 will be stored at the Genomics Core Cloud Storage for as long as our laboratory is a Gold Tier member. Should storage prove insufficient due to other projects competing for space at SharePoint or should we lose our Gold Tier membership at Genomics Core, KU Leuven shared network drives will be a second option.
A pseudonymised, non-retraceable version of dataset no. 3 and 8 will be stored at RDR as well, mostly for data sharing purposes.

**What are the expected costs for data preservation during the expected retention period? How will these costs be covered?**

As SharePoint Team sites and RDR are free at KU Leuven, we do not expect costs. The Genomics Core Gold Tier membership is covered through internal laboratory funds. Should storage space prove insufficient or Genomics Core Gold Tier membership too expensive, and KU Leuven shared network drives are needed, internal laboratory funds will be designated for this purpose.

**DATA SHARING AND REUSE**

**Will the data (or part of the data) be made available for reuse after/during the project?**
**Please explain per dataset or data type which data will be made available.**

- Yes, as restricted data (upon approval, or institutional access only)
- No (closed access)
- Yes, as open data

Because of sensitivities around both genetic and personal data, the majority of our datasets **will not be shared** outside our laboratory, in particular datasets no. 1, 2, 5, 6 and 7 (see below). Dataset no. 9 is intended for internal review of the project, hence sharing is not deemed scientifically useful. Since dataset no. 4 will be destroyed before the end of the project, sharing is not an option.
Dataset no. 3 will be available as **restricted data** for future scholars. Names of participants will have been pseudonymised and the key file will not be provided, ensuring participants anonymity for future researchers.
A pseudonymised, non-retraceable version of dataset no. 8 will be made available as **open data**. This dataset will serve as the basis of future publications, so the data is meant for verification of scientific validity.

**If access is restricted, please specify who will be able to access the data and under what conditions.**

The restricted data will be accessible for future scholars in the field, seeking to further our understanding of the topic. Because of cultural and possible familial sensitivities around premarital children, this data will only be shared with reputable scholars with a record in the field and who are fully aware of these sensitivities.

**Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?**

**Please explain per dataset or data type where appropriate.**

- Yes, other
- Yes, privacy aspects
- Yes, ethical aspects

Because of sensitivities around both genetic and personal data, the majority of our datasets will not be shared outside our laboratory, in particular datasets no. 1, 2, 4, 5, 6 and 7.
As mentioned before, sharing of dataset no. 1 is impossible for privacy reasons because it contains highly personal data. Datasets 2 and 7, even when pseudonymised, contain a lot of information on family relationships and kinship that is less than a century old. In line with the way the Belgian State Archives deal with this issue, we intend not to share such inimate data out of privacy concerns.
Because genetic data is highly personal, and ethics discussions around it are still an evolving topic in bio-ethics, we do not intend to share dataset 5. This is also why dataset no. 4 will be destroyed after the sequencing analyses, rendering the discussion of sharing this dataset obsolete.

**Where will the data be made available?**

**If already known, please provide a repository per dataset or data type.**

- Other (specify below)
- KU Leuven RDR (Research Data Repository)

Dataset no. 3 and no. 8 will be made available, in pseudonymised, non-retraceable form, and as respectively restricted and open data, through KU Leuven RDR. As part of publications, variations or selections of dataset no. 8 might also be published as Supplementary data, depending on journal requirements.

**When will the data be made available?**

- Upon publication of research results

**Which data usage licenses are you going to provide?**

**If none, please explain why.**

- CC-BY 4.0 (data)
- Data Transfer Agreement (restricted data)

**Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.**

- Yes, a PID will be added upon deposit in a data repository

**What are the expected costs for data sharing? How will these costs be covered?**

No costs are expected.

**RESPONSIBILITIES**

**Who will manage data documentation and metadata during the research project?**

Brecht Liefsoens

**Who will manage data storage and backup during the research project?**

Brecht Liefsoens

**Who will manage data preservation and sharing?**

Maarten H. D. Larmuseau

**Who will update and implement this DMP?**

Brecht Liefsoens (during project), Maarten H. D. Larmuseau (after project)