Project Number:

839730

Project Acronym:

SemioMaths

Project title:

Towards a theory of mathematical signs
based on the automatic treatment of mathematical corpora

DATA MANAGEMENT PLAN

v2.0 – 20.02.2023

Abstract of the project:

The SemioMaths project aims at proposing a formal model for the automatic analysis of mathematical corpora, based on semiological theories and on recent advances in formal and computational linguistics, in order to provide a broad and solid framework for the philosophy and the history of mathematics especially concerned with the study of mathematical practices. In order to achieve this objective, the action includes advanced training in the technical fields concerned by the project, the study of various existing mathematical corpora, the development of concrete implementations of linguistic computational models on mathematical corpora, and the elaboration of the corresponding results in the form of both scientific articles and a software package. Scientific training will be achieved by following suitable courses and exchanging with the research community of the host institution; the study of corpora will be done in close collaboration with the research team of the host chair; models will be implemented using the institution's advanced computing infrastructure and IT support; finally, a teaching activity will help to systematize the conceptual grounds of the proposed approach. As a result of its inherent interdisciplinarity, other than in the specific fields of philosophy and history of mathematics, an impact is expected from the connections to be established with the other fields with which this project is concerned (computational linguistics, computer science, logic), and between the corresponding scientific communities. In this way, the SemioMaths project will improve our comprehension of the relations between mathematical knowledge and empirical textual practices. This shall result in a better understanding of the singularity of mathematics as a human practice and of its role in society, thus ultimately helping to build new and original bridges between formal and natural sciences on one side, and the humanities on the other.

1. Data Summary

1.1. Purpose of data collection and generation

The SemioMaths project intends to propose a formal model for the automatic processing of mathematical corpora based on a critical analysis of existing approaches to Natural Language Processing (NLP). Data will be then mainly generated for the purpose of evaluating the capabilities of the intended model with respect to different kinds of corpora. Data corresponding both to models and to corpora will be used and produced for assessing state-of-the-art NLP models (aim 1 of the project), comparing the latter to the capabilities of the proposed model (aim 2), building training and test sets of mathematical corpora (aim 3) and applying and evaluating the intended model to the latter (aim 4).

1.2. Types and formats

Accordingly, four main types of data will be generated or collected: 1) textual corpora (both linguistic and mathematical); 2) computational implementations of formal models of analysis; 3) examples of the application of models to corpora; 4) analysis of models and results. As for the respective formats, corpora in (1) is expected to be mainly in text format (.txt, UTF-8 encoding), in particular for natural language corpora, as well as LaTeX and MathML for mathematical corpora, and PDF-A for possible diagrammatical content. The main programming language for models in (2) will be Python (.py). Intermediate and final results in (3) usually take the form of databases and hierarchical data, which is expected to be produced in .csv and .xml. Finally, data resulting from the analysis in (4) will also be produced mostly as .csv databases, as well as .pdf for diagrams, produced by available open source python libraries (matplotlib, plotly, seaborn, etc.). The use of literate programming (in a programming language such as noweb) will be tried as a side experiment.

1.3. Re-use of existing data

The project will rely upon existing data concerning linguistic corpora (BNC, COCA, multilingual Wikipedia dumps, Common Crawl, etc.) as well as specific mathematical corpora (EuDML, ALGO, Cuneiform, Project Euclid, ELibM, Numdam, etc.). New mathematical corpora will be built on the basis of existing works in the history of mathematics. The analysis of state-of-the-art linguistic models will also require the use of available models. In the case of existing corpora (both linguistic and mathematical), a pre-treatment will be performed in order to normalize the data, in particular to produce “raw” corpora (erasing annotations and normalizing the elementary characters in accordance with the requirements of the conceptual framework and the intended model). As for the new mathematical corpora, this normalization process will require theoretical decisions resulting from the development of the project.

1.4. Origin of data

The vast majority of the data is available online under open access license. Existing corpora are accessible through the websites and repositories of the different projects that produced them (see examples above). Mathematical documents needed for the construction of new corpora are freely available in different digital libraries (Internet Archive, Gallica, Google Books, etc.). Linguistic models are also available in open access repositories.

1.5. Size

The size of the data is expected not to exceed 3TB

1.6. Data Utility

The data used and produced within this project might be useful for researchers in the field of philosophy and history of mathematics and of science in general. Researchers in Natural Language Processing might also find this data profitable. The data related to the behaviour of the intended model can also be of interest to researchers in Machine Learning, Artificial Intelligence and Theoretical Computer Science.

2. FAIR data

2.1. Making data findable

2.1.1. Discoverability and Identification

Curated datasets (i.e. produced within the project as a result of the work related to aim 3) and software will be published at the end of the project on Zenodo, which makes them openly accessible and discoverable. In addition, the same data will also be made available on the ETH Research Collection repository. We will follow the Metadata standards of those repositories, which are all compatible with OpenAIRE recommendations (OAI-PMH and OAI-DC metadata schema). The software produced will also be available throughout the whole project in a GitHub repository (https://github.com/Gianni-G/semiolog), where the entire version history of the code will be openly accessible. All the information about the data will also be available in the dedicated website for the project (www.semiomaths.com). The data will be indexed using the EU Open Data Portal (https://data.europa.eu/euodp/en). All published datasets will receive a DOI that will be referred to in any scientific publication that made use of this data set.

2.1.2. Naming conventions

Following an established practice in software development, the directory structure for the intended software package is as follows:

```
/semiolog                  root folder named after the intended software package
      /docs                documentation of the software
      /lib                 custom functions and third-party libraries
      /models              selected examples of the application to different corpora
      /scripts             sample scripts corresponding to the different steps of the procedure
      /semiolog            source code
            /classifier    source code for classifier task
            /paradigmatic  source code for paradigmatic task
            /syntagmatic   source code for syntagmatic task
            /typing        source code for typing task
      /test                tests of correctness
```

Within the directory `/models`, each folder will correspond to a different model, named after its identifier. Folder names within each model directory will reflect the objects produced by the steps of the analytical procedure: corpus, vocabulary, syntagmas, paradigms, etc. The name of the sample scripts will conform to a sequence-based scheme for sub-steps, followed by a short descriptive expression. Ex.:

```
01_create_empty_project.py
02_build_corpus.py
03_build_vocabulary.py
04_build_syntagmas.py
05_build_paradigmatizer.py
...
```

The entire sequence of those files will constitute the elementary form of the analytic procedure, applicable to all the examples provided. Each example of analysis will be stored in a separate directory within `/models`, all following the same internal structure. Ex:

```
/models
    /en_bnc
        /corpus
            /original
            dev.txt
            test.txt
            train.txt
        /paradigms
            /checkponts
            tf_model.hs
            config.json
            history.json
        /syntagmas
            /tokenized
                /dev
                /test
                /train
        /vocabulary
            /checkpoints
            /ngrams
            alpha.json
            freq.json
            vocab.json
            merges.txt
        config.json
        script.py
```

Due to potential large sizes, models will be stored in the GitHub repository using the Git LFS protocol.

Filenames requiring dates will be prefixed by the schema `YYYY-MM-DD`. Lowercase unaccentuated letters will be prioritized, as well as the underscore character (_) in the place of spaces.

All non-self-explanatory codes will be documented in README files.

2.1.3. Metadata and Keywords

Each dataset will be tagged with the keywords: “SemioMaths”, “MSCA”, “H2020”. Furthermore, we will systematically use the keywords “Philosophy of mathematical practices”, “Structuralist semiology” and “Mathematical Language Processing”, defining the disciplinary region of the project. Finally, we will also use keywords to refer to the specific techniques relevant in each case (ex: “Byte Pair Encoding”, “Biorthogonal typing”, etc.)

Other metadata will be assigned according to the requirements of the repositories used and keywords will be added. The code will be adequately documented in a README file.

At the moment, no creation of further metadata seems necessary. Should this become imperative, the present document will be updated accordingly.

2.1.4. Versioning

Version numbering of the software will follow a sequence-based software versioning scheme. In particular, we will use semantic versioning, with a sequence of three digits (`[Major].[Minor].[Patch]`), together with an optional pre-release tag (`-alpha`, `-beta`) and optional build meta tag. Major number of zero (`0.y.z`), will be used to indicate a work-in-progress.

2.2. Making data openly accessible

All the data used and produced in the project will be made openly available by default with the end of the project. Data leading to specific results will be made available as soon as those results are ready to be published, even before the end of the project.

As already mentioned, data will be made accessible by deposition in 2 repositories: Zenodo and ETH Research Collection including metadata according to each repository’s standard. Publications will also be available at the dedicated website of the project (www.semiomaths.com). The code will be available in a GitHub. The data will be indexed using the EU Open Data Portal (https://data.europa.eu/euodp/en).

Datasets will be produced in standard open formats (txt, csv, xml, pdf, zip, etc.), which do not impose any specific requirement for their access other than the one established by the respective public specifications. Should we experience the need of producing dataset under a specific format for the software to be developed in the project, the specification of the format and its link to the software release will be clearly provided.

The intended software will be written in Python 3 and will rely on widespread python packages (ex: numpy, networkx, graphviz, etc.), which are accessible in open source distributions. All details of the versions of formats, dataset and software releases and packages used will be provided and the relevant documentation will be clearly referenced. Documentation about the software to be produced will be made available, as a result of a continuous documentation of the development process. Experimenting with literate programming can provide a documentation which is the direct result of the execution of the code, in such a way that both documentation and software evolve simultaneously as part of the same procedure.

Contact with members of the ETH Library, responsible for the Research Data Management and Digital Curation, has already been established.

There will be no restrictions of use and access, and no track of personal identity is envisaged at this stage.

2.3. Making data interoperable

The data, metadata, and documentation follow open standards and file formats. The basic reference standard for this project will be Dublin Core. Mathematical corpora will be indexed using the EuDML metadata schema (version 2.0). For all other types of textual data, the project will follow the TEI guidelines.

All other decisions concerning standards and meta-data schema will be documented in the final version of the present Data Management Plan. For specific aspects for which no decision has been made to this point, preference will always be given to existing standards.

Although Zenodo does not provide subject-specific metadata, a README file will be provided, describing the standards used in the project, thus helping others to identify and classify this work.

2.4. Increase data re-use (through clarifying licenses)

2.4.1. License

The software will be licensed under Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication (see https://creativecommons.org/publicdomain/zero/1.0/).

CREATIVE COMMONS CORPORATION IS NOT A LAW FIRM AND DOES NOT PROVIDE LEGAL SERVICES. DISTRIBUTION OF THIS DOCUMENT DOES NOT CREATE AN ATTORNEY-CLIENT RELATIONSHIP. CREATIVE COMMONS PROVIDES THIS INFORMATION ON AN "AS-IS" BASIS. CREATIVE COMMONS MAKES NO WARRANTIES REGARDING THE USE OF THIS DOCUMENT OR THE INFORMATION OR WORKS PROVIDED HEREUNDER, AND DISCLAIMS LIABILITY FOR DAMAGES RESULTING FROM THE USE OF THIS DOCUMENT OR THE INFORMATION OR WORKS PROVIDED HEREUNDER.

Statement of Purpose

The laws of most jurisdictions throughout the world automatically confer exclusive Copyright and Related Rights (defined below) upon the creator and subsequent owner(s) (each and all, an "owner") of an original work of authorship and/or a database (each, a "Work").

Certain owners wish to permanently relinquish those rights to a Work for the purpose of contributing to a commons of creative, cultural and scientific works ("Commons") that the public can reliably and without fear of later claims of infringement build upon, modify, incorporate in other works, reuse and redistribute as freely as possible in any form whatsoever and for any purposes, including without limitation commercial purposes. These owners may contribute to the Commons to promote the ideal of a free culture and the further production of creative, cultural and scientific works, or to gain reputation or greater distribution for their Work in part through the use and efforts of others.

For these and/or other purposes and motivations, and without any expectation of additional consideration or compensation, the person associating CC0 with a Work (the "Affirmer"), to the extent that he or she is an owner of Copyright and Related Rights in the Work, voluntarily elects to apply CC0 to the Work and publicly distribute the Work under its terms, with knowledge of his or her Copyright and Related Rights in the Work and the meaning and intended legal effect of CC0 on those rights.

1. Copyright and Related Rights. A Work made available under CC0 may be protected by copyright and related or neighboring rights ("Copyright and Related Rights"). Copyright and Related Rights include, but are not limited to, the following:

i. the right to reproduce, adapt, distribute, perform, display, communicate, and translate a Work;

ii. moral rights retained by the original author(s) and/or performer(s);

iii. publicity and privacy rights pertaining to a person's image or likeness depicted in a Work;

iv. rights protecting against unfair competition in regards to a Work, subject to the limitations in paragraph 4(a), below;

v. rights protecting the extraction, dissemination, use and reuse of data in a Work;

vi. database rights (such as those arising under Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, and under any national implementation thereof, including any amended or successor version of such directive); and

vii. other similar, equivalent or corresponding rights throughout the world based on applicable law or treaty, and any national implementations thereof.

2. Waiver. To the greatest extent permitted by, but not in contravention of, applicable law, Affirmer hereby overtly, fully, permanently, irrevocably and unconditionally waives, abandons, and surrenders all of Affirmer's Copyright and Related Rights and associated claims and causes of action, whether now known or unknown (including existing as well as future claims and causes of action), in the Work (i) in all territories worldwide, (ii) for the maximum duration provided by applicable law or treaty (including future time extensions), (iii) in any current or future medium and for any number of copies, and (iv) for any purpose whatsoever, including without limitation commercial, advertising or promotional purposes (the "Waiver"). Affirmer makes the Waiver for the benefit of each member of the public at large and to the detriment of Affirmer's heirs and successors, fully intending that such Waiver shall not be subject to revocation, rescission, cancellation, termination, or any other legal or equitable action to disrupt the quiet enjoyment of the Work by the public as contemplated by Affirmer's express Statement of Purpose.

3. Public License Fallback. Should any part of the Waiver for any reason be judged legally invalid or ineffective under applicable law, then the Waiver shall be preserved to the maximum extent permitted taking into account Affirmer's express Statement of Purpose. In addition, to the extent the Waiver is so judged Affirmer hereby grants to each affected person a royalty-free, non transferable, non sublicensable, non exclusive, irrevocable and unconditional license to exercise Affirmer's Copyright and Related Rights in the Work (i) in all territories worldwide, (ii) for the maximum duration provided by applicable law or treaty (including future time extensions), (iii) in any current or future medium and for any number of copies, and (iv) for any purpose whatsoever, including without limitation commercial, advertising or promotional purposes (the "License"). The License shall be deemed effective as of the date CC0 was applied by Affirmer to the Work. Should any part of the License for any reason be judged legally invalid or ineffective under applicable law, such partial invalidity or ineffectiveness shall not invalidate the remainder of the License, and in such case Affirmer hereby affirms that he or she will not (i) exercise any of his or her remaining Copyright and Related Rights in the Work or (ii) assert any associated claims and causes of action with respect to the Work, in either case contrary to Affirmer's express Statement of Purpose.

4. Limitations and Disclaimers.

a. No trademark or patent rights held by Affirmer are waived, abandoned, surrendered, licensed or otherwise affected by this document.

b. Affirmer offers the Work as-is and makes no representations or warranties of any kind concerning the Work, express, implied, statutory or otherwise, including without limitation warranties of title, merchantability, fitness for a particular purpose, non infringement, or the absence of latent or other defects, accuracy, or the present or absence of errors, whether or not discoverable, all to the greatest extent permissible under applicable law.

c. Affirmer disclaims responsibility for clearing rights of other persons that may apply to the Work or any use thereof, including without limitation any person's Copyright and Related Rights in the Work. Further, Affirmer disclaims responsibility for obtaining any necessary consents, permissions or other rights required for any use of the Work.

d. Affirmer understands and acknowledges that Creative Commons is not a party to this document and has no duty or obligation with respect to this CC0 or use of the Work.

Datasets will be licensed under Creative Commons CC BY-NC 4.0
(see https://creativecommons.org/licenses/by-nc/4.0/).

Users are free to:

Share — copy and redistribute the material in any medium or format
Adapt — remix, transform, and build upon the material

The licensor cannot revoke these freedoms as long as you follow the license terms.
Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

Non-Commercial — You may not use the material for commercial purposes.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

2.4.2. Timing

Data will at latest be made available with the publication of an accompanying scientific publication that references the data sets. All data will at latest be available with the project end, even those datasets, which are not referenced in scientific publications.

2.4.3. Re-use

All the data used and produced can be used by other scientists in the different fields concerned by the project.

The data will remain usable as long as the repository guarantees its availability, and in addition, data in the Research Collection is preserved in the ETH Data Archive.

2.4.4. Quality assurance

Datasets, metadata and versions of software packages will be released only after being reviewed by at least by one peer. The author and the reviewer are named in the metadata.

3. Allocation of resources

3.1. Cost Estimates

The human resources needed for complying with FAIR standards both for the software and the datasets are comprised in the project as part of WP 2 and 3. As such, the costs of these efforts are included in the project's budget and covered by the H2020 grant.

The chosen repositories are free of charge. Storage at ETH is covered by the hosting institution (chair of History and Philosophy of Mathematical Sciences). Should the storage needs exceed the one provided by the hosting institution, the exceeding costs will be covered by the grant budget. The costs of domain names and web hosting for the dedicated website will be also covered by the H2020 grant.

3.2. Data Management Responsibilities

The principal investigator of SemioMaths will be responsible for the data management of the project. He will rely upon the expertise of the members of the SemioLog team, to which the project is associated, as well as the assistance of the ETH Library, and ETH IT Services.

By the end of the project, it will be decided which data should be preserved on a long-term basis and which could be discarded without any impact on the re-usability or the replicability of the results. The decision will be discussed with the members of the SemioLog team, as well as with the project supervisor and peers in the scientific field, in constant dialogue with the members of the technical services involved. A specific budget from the H2020 will be planned in order to cover any cost arising from potential needs of long-term preservation. Data in the ETH Research Collection are hosted free of charge and will in addition be preserved in the ETH Data Archive funded by ETH Zurich.

4. Data security

All data will be backed up at the internal drive of the ETH chair of History and Philosophy of Mathematical Sciences. Furthermore, a copy of released data will be kept in the repositories already mentioned. All these services are intended for long-term storage of scientific research data. In addition, data from the Research Collection are also preserved in the ETH Data Archive. Finally, a copy of the released data will be kept on the Zenodo platform. All of these services are intended for long-term storage of scientific research data

Upon unintentional loss of data (misuse of the collaborative workspace, accidental removal), Juan Luis Gastaldi (ETH Zürich) needs to be contacted via email to juan.luis.gastaldi@gess.ethz.ch. He will interact with ETH's IT services to restore the latest known copy.

5. Ethical aspects

As stated in the project proposal, the SemioMaths project does not contain any dimension that requires special treatment concerning ethical issues or informed consent.

6. Other issues

Contact juan.luis.gastaldi@gess.ethz.ch by e-mail for any questions concerning the data sets and their management in the scope of the SemioMaths H2020 project.

This document has been produced with the help of the Research Data Management and Digital Curation Service of the ETH Library. Other DMPs from the MSCA H2020 program were used as models, in particular the DMP of EASITrain (Grant Agreement No 764879).