DMP title

Project Name My plan (FWO DMP) - DMP title
Grant Title 1281222N
Principal Investigator / Researcher Stefano De Pascale
Institution KU Leuven

1. General Information

Name applicant

Stefano De Pascale

FWO Project Number & Title

1281222N, "Meaning change in token space: a token-based computational approach to diachronic prototype semantics"

Affiliation

- KU Leuven
- Vrije Universiteit Brussel

2. Data description

Will you generate/collect new data and/or make use of existing data?

- Generate new data
- Reuse existing data

Describe in detail the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a table (see example) or as a data flow and per WP or objective of the project. If you reuse existing data, specify the source of these data. Distinguish data types (the kind of content) from data formats (the technical format).

Type of data	**Format**	**Volume**	**How created**
primary input data: historical text corpus (Delpher)	ALTO-XML	~ 1.5 TB (corpus files from 1830 to 1999; ~ 10 GB of files for every year)	access via API-key
secondary input data: database with semantic annotation of tokens extracted from primary input data	.csv	unknown	semantic annotation task gathered with a web-based tool (in collaboration with Qualtrics)
Natural Language Processing scripts: ? Python Scripts (for creation of token-based models) ? R-scripts (for cluster analyses and other numerical/statistical manipulation of the token-based models)	.R, and .py	several MB	Jupyter notebooks and Rstudio
primary output data: ? frequency lists, ? collocation ? distance matrices	.txt, .json, .npy	from several MB (.txt-files and .json-files), to 1 GB (.npy files)	Jupyter notebooks
secondary output data: interactive webpages (for visualization and sharing of results)	Java/HTML or as ShinyApps	several MB	Rstudio, Rmarkdown

3. Legal and ethical issues

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to your file in KU Leuven's Register of Data Processing for Research and Public Service Purposes (PRET application). Be aware that registering the fact that you process personal data is a legal obligation.

- Yes

Privacy Registry Reference: G-2021-4120

Prior to the semantic annotation task. very limited personal data of the recruited particiapants will be surveyed, namely gender, age and education. Also, each annotator will be given a "unique identifier" that will distinguish their annotated dataset from others, without the UI being linked to or revealing anything about the annotator's personal identity.

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, add the reference to the formal approval by the relevant ethical review committee(s)

- No

Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

- Yes

The creation of annotation datasets will allow the improvement of computational systems for historical semantic change detection.

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?

- No

4. Documentation and metadata

What documentation will be provided to enable reuse of the data collected/generated in this project?

1. For the storage of the primary input data (i.e. the corpus), a word document with the folder structure of the corpus and the explanation of the file name protocol will be created

2. For the secundary input data (i.e. the semantic annotation), another document will be drafted with the precise instructions on how the tokens for annotation have been selected; the annotation task itself will follow a specialized publicly available tool, Semann2.

3. The R- and Python scripts will contain internal, intra-code comments to explain every part of the code

3. For the output data (the matrices, concordances, frequency lists etc.) we will refer to the extensive documentation that has already been created for a project that has developed the computational methods used in this project.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- No

5. Data storage and backup during the FWO project

Where will the data be stored?

-
  - primary input data: texts after 1879 cannot be shared without agreement with Koninklijke Bibliotheek (van Nederland) and their use is at the moment only possible after signing a non-disclosure agreement with the Koninklijke Bibliotheek
  - secondary input data, NLP-scripts & secondary output data: on personal GitHub page (best practice in NLP)
  - primary output data: probably too large to share on Github, but stored on the research group’s private server (IAIN)

How is backup of the data provided?

stored in the research group’s private server (IAIN)

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely. If no or insufficient storage or backup capacities are available then explain how this will be taken care of.

- Yes

The server contains both personal folders and shared folders (in total 25 TB)

What are the expected costs for data storage and back up during the project? How will these costs be covered?

The costs are covered by the research group

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The semantic annotation data will be anonymised, and stored on the research group's private server

6. Data preservation after the FWO project

Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).

The (human) semantic annotation and the NLP-scripts to generate the output data.

Where will the data be archived (= stored for the longer term)?

- primary output data: can still be stored in the research group’s private server
- secondary input data, NLP-scripts & secondary output data: on personal GitHub page (best practice in NLP)
- primary output data: probably too large to share on Github, but stored on the research group’s private server (IAIN)

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?

The costs are covered by the research group

7. Data sharing and reuse

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

- No

Which data will be made available after the end of the project?

Everything, except the largest files (i.e. matrices) and the corpus files.

Where/how will the data be made available for reuse?

- In an Open Access repository

On Github

When will the data be made available?

- Upon publication of the research results

Who will be able to access the data and under what conditions?

Everybody will be able to access the files that are made publicly available, without any restrictions

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

No costs.

8. Responsibilities

Who will be responsible for data documentation & metadata?

The postdoc researcher

Who will be responsible for data storage & back up during the project?

The postdoc researcher and the supervisor

Who will be responsible for ensuring data preservation and reuse ?

The postdoc researcher

Who bears the end responsibility for updating & implementing this DMP?

The PI bears the end responsibility of updating & implementing this DMP.