FWO DMP Template - Flemish Standard Data Management Plan

Project supervisors (from application round 2018 onwards) and fellows (from application round 2020 onwards) will, upon being awarded their project or fellowship, be invited to develop their answers to the data management related questions into a DMP. The FWO expects a **completed DMP no later than 6 months after the official start date** of the project or fellowship. The DMP should not be submitted to FWO but to the research co-ordination office of the host institute; FWO may request the DMP in a random check.

At the end of the project, the **final version of the DMP** has to be added to the final report of the project; this should be submitted to FWO by the supervisor-spokesperson through FWO's e-portal. This DMP may of course have been updated since its first version. The DMP is an element in the final evaluation of the project by the relevant expert panel. Both the DMP submitted within the first 6 months after the start date and the final DMP may use this template.

The DMP template used by the Research Foundation Flanders (FWO) corresponds with the Flemish Standard Data Management Plan. This Flemish Standard DMP was developed by the Flemish Research Data Network (FRDN) Task Force DMP which comprises representatives of all Flemish funders and research institutions. This is a standardized DMP template based on the previous FWO template that contains the core requirements for data management planning. To increase understanding and facilitate completion of the DMP, a standardized **glossary** of definitions and abbreviations is available via the following link.

1. General Project Information	
Name Grant Holder & ORCID	Christof Koolen – ORCID Identifier: 0000-0002-2210-6821
Contributor name(s) (+ ORCID) & roles	N/A
Project number[1] & title	Project 3H230312 – The AI Legal Researcher: Using Natural Language Processing (NLP) for Legal Research
Funder(s) GrantID[2]	KU Leuven Postdoctorale Mandaten (PDM) PDMT2/23/005
Affiliation(s)	☒ KU Leuven ☐ Universiteit Antwerpen ☐ Universiteit Gent ☐ Universiteit Hasselt ☐ Vrije Universiteit Brussel ☐ Other: Provide ROR[3] identifier when possible: https://ror.org/05f950310

[1] "Project number" refers to the institutional project number. This question is optional since not every institution has an internal project number different from the GrantID. Applicants can only provide one project number.

[2] Funder(s) GrantID refers to the number of the DMP at the funder(s), here one can specify multiple GrantIDs if multiple funding sources were used.

[3] Research Organization Registry Community. https://ror.org/

Please provide a short project description	With legal textual information being omnipresent, Natural Language Processing (NLP) techniques have risen sharply in popularity. Fundamentally, NLP is a subdomain of artificial intelligence that deals with the interactions between computers and humans through natural language. When NLP techniques are used in conjunction with traditional legal research methods, they can change the way how researchers interact with scholarly sources, for instance by quickly dissecting cases, identifying key phrases, and extracting underlying themes within a corpus of legal documents. As a result, NLP techniques hold great promise for legal academic research, as they can help scholars interpret and analyse legal texts, uncovering insights and information that would otherwise be difficult or impossible to obtain. This postdoctoral project aims to critically examine how NLP techniques can be applied in legal scholarship, with the specific aim of integrating NLP into the methodological toolkit of legal scholars.

2. Research Data Summary

				ONLY FOR DIGITAL DATA	*ONLY FOR DIGITAL DATA*	*ONLY FOR DIGITAL DATA*	*ONLY FOR PHYSICAL DATA*
Dataset Name	Description	New or Reused	Digital or Physical	Digital Data Type	Digital Data Format	Digital Data Volume (MB, GB, TB)	Physical Volume
Legal document corpus	A collection of legal documents (e.g., legal doctrine, case law, legislation) intended to be used for further analysis	☐ Generate new data ☒ Reuse existing data	☒ Digital ☐ Physical	☐ Observational ☐ Experimental ☒ Compiled/ aggregated data ☐ Simulation data ☐ Software ☐ Other ☐ NA	☐ .por ☐ .xml ☐ .tab ☐ .csv ☒ .pdf ☒ .txt ☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☐ other: ☐ NA	☐ < 100 MB ☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	N/A
Annotated legal corpus	Legal documents annotated with metadata for NLP training and testing	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☒ Observational ☐ Experimental ☐ Compiled/ aggregated data ☐ Simulation data ☐ Software ☐ Other ☐ NA	☐ .por ☒ .xml ☐ .tab ☒ .csv ☐ .pdf ☒ .txt ☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☒ other: .json ☐ NA	☐ < 100 MB ☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	N/A

NLP tool outputs	Outputs generated by NLP tools (e.g., entity recognition, topic modelling results)	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Observational ☐ Experimental ☐ Compiled/ aggregated data ☒ Simulation data ☐ Software ☐ Other ☐ NA	☐ .por ☐ .xml ☐ .tab ☒ .csv ☐ .pdf ☐ .txt ☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☒ other: .json and .pkl ☐ NA	☐ < 100 MB ☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	N/A
Legal research software	Custom-developed scripts for processing and analysing legal texts	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Observational ☐ Experimental ☐ Compiled/ aggregated data ☐ Simulation data ☒ Software ☐ Other ☐ NA	☐ .por ☐ .xml ☐ .tab ☐ .csv ☐ .pdf ☐ .txt ☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☒ other: .py and .ipynb ☐ NA	☐ < 100 MB ☒ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	N/A

GUIDANCE: *DATA CAN BE DIGITAL OR PHYSICAL (FOR EXAMPLE BIOBANK, BIOLOGICAL SAMPLES, …). DATA TYPE: DATA ARE OFTEN GROUPED BY TYPE (OBSERVATIONAL, EXPERIMENTAL ETC.), FORMAT AND/OR COLLECTION/GENERATION METHOD.* *EXAMPLES OF DATA TYPES: OBSERVATIONAL (E.G. SURVEY RESULTS, SENSOR READINGS, SENSORY OBSERVATIONS); EXPERIMENTAL (E.G. MICROSCOPY, SPECTROSCOPY, CHROMATOGRAMS, GENE SEQUENCES); COMPILED/AGGREGATED DATA[4] (E.G. TEXT & DATA MINING, DERIVED VARIABLES, 3D MODELLING); SIMULATION DATA (E.G. CLIMATE MODELS); SOFTWARE, ETC.* *EXAMPLES OF DATA FORMATS: TABULAR DATA (.POR,. SPSS, STRUCTURED TEXT OR MARK-UP FILE XML, .TAB, .CSV), TEXTUAL DATA (.RTF, .XML, .TXT), GEOSPATIAL DATA (.DWG,. GML, ..), IMAGE DATA, AUDIO DATA, VIDEO DATA, DOCUMENTATION & COMPUTATIONAL SCRIPT.* *DIGITAL DATA VOLUME: PLEASE ESTIMATE THE UPPER LIMIT OF THE VOLUME OF THE DATA PER DATASET OR DATA TYPE.* *PHYSICAL VOLUME: PLEASE ESTIMATE THE PHYSICAL VOLUME OF THE RESEARCH MATERIALS (FOR EXAMPLE THE NUMBER OF RELEVANT BIOLOGICAL SAMPLES THAT NEED TO BE STORED AND PRESERVED DURING THE PROJECT AND/OR AFTER).*

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.	- EUR-Lex dataset: https://huggingface.co/datasets/eurlex
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, please describe these issues further and refer to specific datasets or data types when appropriate.	☐ Yes, human subject data ☐ Yes, animal data ☐ Yes, dual use ☒ No If yes, please describe:

[4] These data are generated by combining multiple existing datasets.

Will you process personal data[5]? If so, briefly describe the kind of personal data you will use. Please refer to specific datasets or data types when appropriate. If available, add the reference to your file in your host institution's privacy register.	☐ Yes ☒ No If yes: - Short description of the kind of personal data that will be used: - Privacy Registry Reference:
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.	☐ Yes ☒ No If yes, please comment:
Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements, research collaboration agreements)? If so, please explain to what data they relate and what restrictions are in place.	☐ Yes ☒ No If yes, please explain:
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 to what data they relate and which restrictions will be asserted.	☐ Yes ☒ No If yes, please explain:

3. Documentation and Metadata

[5] See Glossary Flemish Standard Data Management Plan

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).	**- Project overview document**: A high-level document will outline the project's goals, methodologies, and expected outcomes in order to serve as an interpretative guide. **- README files**: all datasets will be accompanied with a README file that contains a detailed description of the dataset, the structure of the data, an overview of any preprocessing steps applied, and information on how to access and use the data. **- Code documentation**: For any scripts developed as part of the research, source code comments will be added in order to explain the logic and functionality of the code. **- Version control**: A version control system (i.e., Git) will be used to keep track of code changes over time, thereby allowing for rollbacks to earlier versions and understanding the evolution of the project.
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. *REPOSITORIES COULD ASK TO DELIVER METADATA IN A CERTAIN FORMAT, WITH SPECIFIED ONTOLOGIES AND VOCABULARIES, I.E. STANDARD LISTS WITH UNIQUE IDENTIFIERS.*	☐ Yes ☒ No If yes, please specify (where appropriate per dataset or data type) which metadata standard will be used: If no, please specify (where appropriate per dataset or data type) which metadata will be created: A custom metadata scheme will be created to tailored to the unique aspects of legal documents and NLP research outputs. Relevant metadata elements include: keywords, creator, date published, version, format, language, legal domain, access information.

4. Data Storage & Back-up during the Research Project	
Where will the data be stored?	- KU Leuven OneDrive - Google Drive

How will the data be backed up? *What storage and backup procedures will be in place to prevent data loss? Describe the locations, storage media and procedures that will be used for storing and backing up digital and non-digital data during research.*[6] *Refer to institution-specific policies regarding backup procedures when appropriate.*	☒ Standard back-up provided by KU Leuven ICTS for my storage solution ☒ Personal back-ups I make (specify) ☐ Other (specify) Data also will be backed up on my personal password protected Google Drive storage.
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 ☐ No If yes, please specify concisely: There is enough space to store the data on the KU Leuven OneDrive and to make back-ups on my personal Google Drive. If no, please specify:
How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons? *Clearly describe the measures (in terms of physical security, network security, and security of computer systems and files) that will be taken to ensure that stored and transferred data are safe.* [6]	The normal access control measures will be applied, including encrypted passwords and two-factor authentication. Data backups will be performed regularly to prevent data loss, with the backup process also adhering to the same access protocols.
What are the expected costs for data storage and backup during the research project? How will these costs be covered?	Expected data storage costs are minimal as the KU Leuven OneDrive is provided at no additional cost for academic purposes. My personal Google Drive offers free storage up to 15GB, with affordable options for additional space if necessary.

[6] Source: Ghent University Generic DMP Evaluation Rubric: https://osf.io/2z5g3/

5. Data Preservation after the end of the Research Project	
Which data will be retained for at least five 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 the KU Leuven Research Data Management policy
Where will these data be archived (stored and curated for the long-term)?	The KU Leuven's Research Data Repository (https://rdr.kuleuven.be/)
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	Expected long-term data storage costs are minimal as the KU Leuven's Research Data Repository is provided at no additional cost for academic purposes.

6. 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. *NOTE THAT 'AVAILABLE' DOES NOT NECESSARILY MEAN THAT THE DATA SET BECOMES OPENLY AVAILABLE, CONDITIONS FOR ACCESS AND USE MAY APPLY. AVAILABILITY IN THIS QUESTION THUS ENTAILS BOTH OPEN & RESTRICTED ACCESS. FOR MORE INFORMATION: HTTPS://WIKI.SURFNET.NL/DISPLAY/STANDARDS/INFO-EU-REPO/#INFOEUREPO-ACCESSRIGHTS*	☐ Yes, in an Open Access repository ☒ Yes, in a restricted access repository (after approval, institutional access only, …) ☐ No (closed access) ☐ Other, please specify:
If access is restricted, please specify who will be able to access the data and under what conditions.	Data will be under a two-year embargo to enable the pursuit of follow-on research. Once the relevant publications have been released, the data will transition to open access for broader use.
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, privacy aspects ☐ Yes, intellectual property rights ☐ Yes, ethical aspects ☐ Yes, aspects of dual use ☐ Yes, other ☒ No If yes, please specify:
Where will the data be made available? If already known, please provide a repository per dataset or data type.	The KU Leuven's Research Data Repository (https://rdr.kuleuven.be/)

When will the data be made available? *This could be a specific date (dd/mm/yyyy) or an indication such as 'upon publication of research results'.*	Upon publication of the relevant research results
Which data usage licenses are you going to provide? If none, please explain why. *A data usage license indicates whether the data can be reused or not and under what conditions. If no licence is granted, the data are in a grey zone and cannot be legally reused. Do note that you may only release data under a licence chosen by yourself if it does not already fall under another licence that might prohibit that.* *Example Answer: E.g. "Data from the project that can be shared will be made available under a Creative Commons Attribution license (CC-BY 4.0), so that users have to give credit to the original data creators."* [7]	- **Data**: CC-BY 4.0 - **Code**: GNU GPLv3
Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, please provide it here. *Indicate whether you intend to add a persistent and unique identifier in order to identify and retrieve the data.*	☒ Yes ☐ No If yes: a DOI will be added upon deposit in the data repository.
What are the expected costs for data sharing? How will these costs be covered?	Data sharing costs are minimal as access to the KU Leuven's Research Data Repository is provided at no additional cost for academic purposes.

[7] Source: Ghent University Generic DMP Evaluation Rubric: https://osf.io/2z5g3/

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	Christof Koolen (the researcher)
Who will manage data storage and backup during the research project?	Christof Koolen (the researcher)
Who will manage data preservation and sharing?	Christof Koolen (the researcher)
Who will update and implement this DMP?	Christof Koolen (the researcher)