Leiden University Data Management Plan[i]

The Research Data Management Regulations Leiden University requires researchers to write a data management plan at the start of a long-term research project[ii]. Please contact the Centre for Digital Scholarship at the University Libraries Leiden if you need help: datamanagement@library.leidenuniv.nl

Name and contact details	*Anne Dirkson : a.r.dirkson@liacs.leidenuniv.nl*
Name of project and group	*Open Knowledge Discovery in Patient Forum Data* *Group: Prof. dr. Kraaij & Dr. Verberne*
Description of your research	*I aim to extract new clinical hypotheses from patient forum data using text mining techniques. These hypotheses could be used to drive future clinical research and would be aggregated from the personal experiences of the patients.*
Project duration	Start: *15-3-2018* End: *15-3-2022*
Names of people and their responsibilities for data management	*Anne Dirkson: responsible for managing the data during the PhD project (march 2018- march 2022).* *Prof Kraaij & Dr. Verberne: responsible for managing the data after march 2022.*
Funding body(ies)	*SIDN (Stichting Internet Domeinnaamregistratie)*
Grant number	-
Partner organisations	-

About this Data Management Plan

Date written	*25-05-2018*
Date last update	*28-06-2018*
Version	*2.0*

Changes in this version of the Data Management Plan

Component	**Progress / Execution** *Please describe briefly what progress you have made, any questions or issues you have encountered and want to discuss, etc.*
1. Data collection	*Data does contain personal names. I need to pseudonymize it and dispose safely of the raw data.*
2. Data storage and back-up	-
3. Data documentation	-
4. Data access, sharing and reuse	*As FAIR as possible –thus share the RDF of the linked concepts and relations that result from my pipeline. The pseudonymized data is still personal data. It contains medical information, which is searchable on the web.*
5. Data preservation and archiving	*I can archive the RDF file of the linked concepts and relations.* *The pseudonymized data can be granted upon permission if*

1. Data collection

Describing the data you will be creating/collecting

1.1 **Will the project use existing or third party data ?**

☐ No
☒ Own / group previous research
☐ Academic collaborators
☒ Commercial collaborators
☐ Publicly available database / archive
☐ Specialist commercial data provider
☐ Other (please specify)

Describe briefly provenance, type and format of this data. Are there any restrictions or requirements for use of third party data such as licensing conditions?

The data was scraped from a public Facebook group: GIST Support International and consists of posts and metadata in a JSON format. There are no restrictions for the use of this data for us, as my supervisors were official academic collaborators on the prior project during collection. This data was received and collected prior to the start of my project.

1.2 **What type()s of data will you collect or create, in what file format(s)?**[iii]

Prior data was collected in a JSON format and consists of the posted text and meta-data of the posts including the names of the users. I will pseudonymize the data by removing the user names and remove all meta-data that we will not use. Future data will be collected similarly. I will create processed data in .csv format (in which I will also replace proper nouns with –name--). Hereafter, I will create aggregated findings in a csv format and RDF.

1.3 **How will you collect and/or create your data?**

Prior data was collected using an in-house scraping script. Future data will be collected using scraping scripts combined with existing APIs

1.4 **What tools, instruments, equipment, hardware or software will you use to capture, produce, collect or create the data?**

Collect: Scraping scripts in Python, API (e.g. Facebook API, Reddit API, Wikipedia API).
Create/Analyse: I will use Python 3.0 to create pseudonymized, processed and analysed data sets (open source).

1.5 **What is the estimated size of the data?**

Please describe briefly. Stages to be adopted if relevant.

Data stage	Specification of type of research data	Software choice and file format	Data size now	Data size when project is finished
Pseudonymized data	*Text messages with user names removed -*	*Python, JSON*	*29,1 MB*	*Unknown*
Processed data	*Pre-processed text messages (incl. proper nouns replaced with –name -); Linked aggregated network of entities and their relations.*	*Python, csv*	*-*	*Unknown*
Results	*Linked aggregated network of entities and their relations.*	*Python, csv*	*-*	*Unknown*
Other…				

……

2. Data storage and security Ensuring that all research data are stored securely and backed up or copied regularly during your research	
2.1	**Where will you store your data?** *Please describe how safe storage is guaranteed. Specify your method if your data is collected and / or transported in different locations / countries.*
	☐ On university departmental network storage (J:) ☒ On university personal network storage (P:) ☐ In a Virtual Research Environment (Sharepoint) ☐ Physical storage (e.g. USB, external hard drive) ☒ Cloud service (e.g. SURFdrive) ☒ Other, namely: Personal drive of the Data Science servers
	….
2.2	**Will your data be backed up?** *Yes, the data will be backed up.* *My university personal storage > Automatic back-ups* *Personal drive of the Data Science servers > Automatic back-ups* *Surfdrive > Automatic back-up*
	….
2.3	**Are there any commercialisation, ethical or confidentiality restrictions about handling your data?** *Please specify briefly.*
	☐ Contractual obligations ☒ Requirements by law : protection of personal data (e.g. privacy law) : specify in 4.1 ☐ Requirements by law : copyright, intellectual property : specify in 4.1 ☐ Ethical restrictions (e.g. ethical review) : specify in 4.1 ☐ Commercial considerations (e.g. patentability) ☐ Formal security standards ☐ No requirements ☐ Other, namely: ……
	According to the GDPR, my data is sensitive information.
2.4	**How will access to the data be managed during the project?** *Please specify for each storage device, from different locations / countries.*
	Data will only be accessible be me from my personal university drive, SurfDrive and my personal drive of the Data Science servers. My supervisors can grant access to the data.
2.5	**What are the main risks to data security?** *Please list risks, e.g. accidental deletion, falling into the wrong hands.* *Please describe what would happen if the data get lost or become unusable.*
	The main risks to data security is data leakage to third parties. If the data would fall into the wrong hands, privacy of the patients could be breached.
2.6	**What measures do you take to comply with the security requirements and to mitigate the risks?** *Describe how you can restore your data in the event of data loss and who is responsible.* *If applicable, please describe procedures to ensure personal data are handled confidentially and who is responsible.*
	☒ Access restrictions ☐ Encryptions ☒ Data processing ☒ De-identification / Anonymisation ☒ Regular back-ups ☒ Master copy stored on university network storage ☒ Master copy stored elsewhere ☐ Other, namely: …
	……

2.7	**How do you differentiate between raw and processed data?** *Please explain briefly why you (do not) differentiate.*
	☐ I will not differentiate ☒ I will create a new file for processed data ☐ I will create a new file for processed data and I will lock raw data ☒ Other, namely: I will dispose of the raw data with user names and store only the raw data without
	……
2.8	**Is there any non-digital data or outputs that the project will generate? Where will these outputs be stored?** *Please specify briefly **and describe who is responsible for storage of these outputs**.*
	Not applicable
2.9	***Do you expect to have any supplementary costs for storage not covered by the project budget?*** *Please specify*
	No

3. Data documentation Documenting your data to help future users to understand and reuse it	
3.1	**How will files be named?** *Please describe briefly.*
	File names will contain my name, the project title, description of content and the date as follows: *Date - DIRKSONAR – project title – description – version* *For instance: 20180628 - DIRKSONAR – GISTFBdata – preprocessed_data – v1.0*
3.2	**How will folders be named and structured?** *You are invited to draw a folder structure and describe it briefly.*
	Data *Project (e.g. GIST_hypothesis_generation)* *Part of project (if applicable)* *File name*
3.3	**How do you handle version control to maintain all changes that are made to the data?** *Please explain your choice briefly. Remember to also document any deletion of data, if applicable.*
	☐ No version control (e.g. original files are overwritten) ☐ Version control software, namely: … ☒ Data/version number in filename/folder ☐ 'Track changes' feature in software ☒ By saving the script with which I process my data ☐ Other, namely: …
	……
3.4	**What metadata standard will be used, if any?**[iv] *Please explain why you use this standard (most used in my discipline, required by the data archive where I will deposit my data). Please outline how the metadata will be created (read me file, spreadsheet, in the data). If no standard exist, please specify which metadata is needed to understand the data.*
	☐ No metadata standard is used ☐ Generic metadata standard (e.g. Dublin Core) ☐ Standard automatic Windows metadata (e.g. from Word, Excel) ☐ Specialised metadata standard, namely: … ☒ Other metadata standard, namely: README files (standard in computer science)

	README file is the standard in my field
3.5	**What supporting information / documentation will you create to enhance understanding of the data ?** *Please describe briefly how peers should be able to understand the data. Examples are a readme.txt, lab journals, a codebook, survey questions etc. Is there a standard for documentation in your field? Describe at what moment in your research process you will add the documentation necessary to make sure the data is understandable for peers.*
	README file with descriptions of projects and included data. Data is also discussed in paper journal peer reviewed articles.

4. Data access, sharing and reuse

Managing access and security, sharing your data

4.1 Are there any restrictions placed on sharing / reuse of some / all of your data?

Please account for not sharing your data. Reasons may be ethical, commercial, security-related, protection of personal data rules, intellectual property, copyright,

Yes, due to protection of personal data rules the data is not fully reusable or shareable. A pseudonymized version with proper nouns removed can be shared upon request. The RDF / csv file of the concepts and their relations can be shared.

4.2 With whom will you share your data at which stage in your research? You can use the table below.

Please state any sharing requirements, e.g. funder data sharing policy. Please describe briefly how you will share your data: on request, pro-actively, etc.. Please specify how your data can be accessed.

	Would not share with anyone	Would share with my immediate collaborators	Would share with others in my research centre or at my institution	Would share with scientists in my field	Would share with scientists outside of my field	Would share with anyone
Immediately after the data has been generated		x	Not appropriate in terms of privacy of personal data	Not appropriate in terms of privacy of personal data	Not appropriate in terms of privacy of personal data	Not appropriate in terms of privacy of personal data
After the data has been normalized and/or corrected for errors		x	Not appropriate in terms of privacy of personal data	Not appropriate in terms of privacy of personal data	Not appropriate in terms of privacy of personal data	Not appropriate in terms of privacy of personal data
After the data has been processed for analysis		x	x	Not appropriate in terms of privacy of personal data	Not appropriate in terms of privacy of personal data	Not appropriate in terms of privacy of personal data
After the data has been analysed		x	x	Not appropriate in terms of privacy of personal data	Not appropriate in terms of privacy of personal data	Not appropriate in terms of privacy of personal data
Immediately before publication		x	x	Not appropriate in terms of privacy of personal data	Not appropriate in terms of privacy of personal data	Not appropriate in terms of privacy of personal data
Immediately after the findings derived from this data		x	x	x (upon request)	Not appropriate in terms of privacy of personal data	Not appropriate in terms of privacy of personal data

	have been published						

Based on: Interview worksheet, Jake Carlson, Purdue University Libraries / Distributed Data Curation Center

	.....
4.3	**If intending to share any part of the data, do your participant consent forms include information about intentions for sharing, retention of data and steps taken to protect participants privacy and confidentiality?**
	☒ Not applicable. ☐ Yes. *Please specify the relevant formula in the consent form.*
	.....
4.4	**Who has authority to grant (additional) access to your data?** *Please describe briefly.*
	☐ Only you ☐ A colleague from the project, namely: … ☒ Supervisor ☐ Funder ☐ Collaborator / research partner organisation ☐ Other, namely: …
4.5	**How will you manage copyright and Intellectual Property Rights issues?** *Who owns the data? How will the data be licensed for reuse? Please describe briefly your choices and their consequences.*
	The University of Leiden owns IP rights to data produced in this project.
4.6	**What is the audience for reuse?** *Please list possible audiences and purposes. Consider who might use it now and who might use it later.*
	Academic researchers

5. Data preservation and archiving Preserving your data	
5.1	**Which criteria will you use to decide which data has to be archived?** *Please briefly describe your choices.*
	☒ Type of data (raw, processed) and how easy it is to reproduce it ☒ Relevance of content for others ☒ Usability of format for others ☒ Data underlying publications ☒ Verification of research ☐ Available time ☐ Available money ☒ Other, namely: Sensitivity of the information
	.....
5.2	**How long should your data be preserved? Are there any requirements regarding the disposal of data?** *State obligations you have by law, funder, university, etc. if any.* *Describe how you will dispose of the data, e.g. how you will get approval, what people and/or tools you need, etc.*
	The data will be stored for the duration of the PhD project – 4 years. At the end of this term, we will evaluate whether it needs to be stored for longer. The user names and original id numbers will be disposed of as we do not need this data for our research. There are currently no requirements as to how to dispose of the data.
5.3	**Which data repository is appropriate for archiving your data?** *Please describe briefly. Does this archive have a 'data seal of approval' or another form of certification?*

	☐ Discipline specific (international) repository, namely … ☐ 4TU.Centre for Research Data ☐ SurfSara ☐ DANS Easy ☐ Other (international) repository, namely : ☒ Other, namely: Due to personal information , data will not archived in a data repository
	……
5.4	**Does the archive have specific requirements concerning file formats, metadata etc.** *Provide relevant urls to the documentation on these requirements. Describe how you intend to meet those requirements, e.g. converting the file formats, providing supplementary documentation.* *Will there be extra costs to prepare your data for archiving? Please specify. See http://www.data-archive.ac.uk/media/247429/costingtool.pdf*
5.5	**What costs (if any) will your selected repository charge? Who pays?** *Please state the costs in euro's and the institution that pays for it.*
5.6	**Who is responsible for the data after the project ends?** *Please state a position and the current person in that position.*
	Prof dr. Kraaij (professor at Leiden University & promotor) and dr. Verberne (assistant professor at Leiden University & co-promotor)

[i] This template is based on the 3TU data management plan, the University of Bath data management plan and the Data Management Checklist of the University of Western Sydney.

ii http://regulations.leiden.edu/research/research-data-management-regulations-leiden-university.html

[iii] Data types can be : documents (text, MS Word), spreadsheets, field notebooks, diaries, questionnaires, transcripts, surveys, codebooks, audiotapes, videotapes, photographs, (transcribed) test responses, models, algorithms, measurements, simulations, observations, software source code, computational model output, etc. Think of the different stages (for instance : video recording, transcript, annotation, lists of typological features ….).

[iv] See http://www.dcc.ac.uk/resources/metadata-standards or http://en.wikipedia.org/wiki/Metadata_standards or the relevant repository.