Data management plan
CELSA/22/007
Humour and Conflict in the Public Sphere: An interdisciplinary analysis of humour controversies and contested freedoms in contemporary Europe

January 2023
KU Leuven, Tartu University, Jagiellonian University, 2022-2024.
Project reference nr: CELSA/22/007
SAP project code: 3H220340

Authors: Giselinde Kuipers (KU Leuven), Wladyslaw Chlopicki (Jagiellonian University), Anastasiya Fiadotava (Jagiellonian University, Estonian Literary Museum), Liisi Laineste (Tartu University, Estonian Literary Museum), in close consultation with Joonas Koivukoski (Helsinki University)

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.

This project will produce three different datasets, each organised in a separate database containing information about public humor events and public responses to this in news media and social media. These databases will be linked; we will consult with digital humanities specialists about the best possible format for this. The first two databases will be made publicly available after the project, and will not contain any sensitive data. The third database will contain raw data, including GDPR-sensitive data, notably links to social media utterances by non-public persons. This database will not be publicly shared.

The three databases will be:

1. General European Public Humor database (GEPH). This database will contain an overview of contested public humor (including humor scandals) in four countries (Belarus, Belgium, Estonia, Poland) during the research period, with date, duration, topic, domain, country(ies), and a short description of each 'humor event'. This will be an excel file (.xls), with only textual information. The data volume will probably not exceed 5MB (very rough estimate). Note: at some point in the project this database will be merged with the database produced for the related Hu-Sca project (identifier: UNA/22/003, KU Leuven administrative identifier ZKE2184**).**
2. Specific Humor Scandal database (consisting of humor scandal datasets coded as HuSca_VVV_XX_YYZZ with VVV a number linked to GEHS, XX a two-digit country code, YY for month and ZZ for year.) Although it is hard to say how many scandals we can expect, we will aim to describe in detail ten humor scandals per country. For each controversy, we will collect a sample of most widespread/shared humor items with all important information, as well as non-humorous responses, in news (or heritage) media and social media. Each dataset will be stored as an excel file (xls) with codes in text (e.g., actor, date, platform/source, norm violation, domain), as well as links to and/or screenshots of the main trigger of the scandal (that is, the original humorous item), and a sample of central other responses. For this dataset, we will work with a digital research specialist to develop a relational database that can be linked to both GEPH and RaHuSca. After the project, we aim to make this database available to researchers in excel (.xls) format. The total size of these databases will probably not exceed 10MB.
3. Raw data for each humor scandal (coded as RaHuSca_VVV_XX_YYZZ with VVV a number linked to GEPH, XX a two-digit country code, YY for month and ZZ for year.). This database will contain the raw data, that is: all data collected related to a humorous event, including links and screenshots, with time, data, and topic. This database will therefore be multimodal, very large (if we include videos, a separate dataset may in some cases exceed 1GB), and will contain privacy-

sensitive data. We are currently not sure yet about the format of these datasets. Possibly, we will store everything in Excel with separate files for videos or images, but a more ideal option would be to build a database that can manages such files. We are currently in communication with a computer scientist who may be able to help. (It is important to note that storage of digital data is a fast-changing and specialist domain, for which we have no specific expertise in our team. The best options may be too expensive for a relatively small grant such as this. However, since the main aim of CELSA is to generate further grants, this is one of the options we will explore for further grant applications.)

Importantly, data collection and coding will be done in Qualtrics, which provides an online user interface that can be used to securely input information on all locations, including archives, university offices or while working at home. We will use the KU Leuven license of Qualtrics, which means that all input will be automatically stored securely. The input will be converted to Excel, which will form the basis of the three databases.

Reuse of existing data

We will not reuse existing data.

Ethical issues

The project has been submitted for ethical review to the ethics review board (SMEC) at KU Leuven. It is currently under review, registration number G-2022-5930.

Personal data

The project will process some personal data. The plans for processing these data have been screened for compliance with GDPR regulations, and has been approved by the ethics review board (SMEC) at KU Leuven. The registration number is G-2022-5930.

Commercial valorization, commercial re-use, 3rd party agreements.

This project does not have potential for commercial valorization. There are no plans or possibilities for commercial re-use, and no relevant 3rd party agreements.

Other legal issues

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.

In some cases, the data we analyze may be copyright-protected, e.g., newspaper articles, images or videos with copyright. Such copyrighted materials will primarily be stored in the RaHuSca database, which will not be made public. If materials are stored in the publicly accessible databases, we will either make sure we comply with regulations for fair use, or use links, or ask for permission, depending on what seems feasible. Note that for social media content, copyright regulations are at times rather unclear, variable across countries, and changeable. In case of doubt we will consult legal specialists.

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).*

We will create detailed codebooks for each of the datasets. These codebooks will be used to code all information, but can also be used to make the data accessible after our (first) analysis is completed. These codebooks will specify all variables in the dataset (that is: codes used for the coding of the research materials), with all possible values or categories for each variable, and the specific rules for coding each of these variables. The codebooks will be made available in pdf format, with an accompanying document explaining the relation between the codebooks and the datasets.

The GEPH and HuSca databases will be created in Excel (.xls), and names and numbered as described above. The codebooks will also contain information about the links between GEPH, HuSca and RaHuSca. The RaHuSca cannot be made publicly available because of GDPR regulations and other ethical considerations, but it will be kept by the researchers of the team for at least ten years.

Will a metadata standard be used to make it easier to **find and reuse the data***?*

Yes, we will adopt the Dublin Core (DCMI) metadata standard to make the data findable for other researchers. Only GEPH and HuSca will also be made reusable.

Data Storage & Back-up during the Research Project

Where will the data be stored?

The data will be stored on the secure servers of the participating universities (KU Leuven, Jagiellonian University, Tartu University) as well as the Estonian Literary Museum (also located in Tartu). In addition, data from GEPH and HuSca may also be stored on the secure servers of other universities participating in this project, which at this point include Hu-Sca (UNA-Europa) partners Complutense University (Madrid), University of Bologna, Helsinki University, as well as Masaryk University (Brno, Czech Republic), Athens University, University of Opole (Poland).

How will the data be backed up?

We will store all datasets on the servers of at least three participating institutions.

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

Yes.

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

The data will be stored on secure university servers that will be only available to people working on the projects, that is: the PIs and co-PIs, the postdocs, and possible the research assistants. Note that the research assistants will only gain access to the databases for the duration of the project, and only will get access to the part of the data collection they are working on for specific tasks such as data cleaning. For input of the data, they will use Qualtrics (which means they can enter but not change the data). Integration of databases will be handled by the postdocs and PI.

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

We don't expect specific costs for data storage and backup, as this can be done on the servers provided by the participating institutions.

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?

The GEPH and HuSca databases will be preserved on RDR, the repository of KU Leuven, and possibly also on the repositories of partner institutions. They will be made publicly available after the end of the project, and will be made available indefinitely. The RaHuSca database will be kept on the university servers, and retained for at least ten years, with evaluation after that.

We expect no additional costs for storing and preserving these data.

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.

The GEPH and HuSca databases will be preserved on RDR, the repository of KU Leuven, and possibly also on the repositories of partner institutions. For more information on RDR, see https://www.kuleuven.be/rdm/en/rdr. These data will be made publicly available after the end of the project, and will be made available indefinitely. RDR will provide identifiers for the datasets, and will also allow us to link the datasets with documentation and metadata, and will cover the costs for storage.

The RaHuSca database will be kept on the university servers, and retained for at least ten years, with evaluation after that. The RaHuSca database will remain available to the PIs and the postdocs, who jointly may decide to provide access to other researchers. The sharing of the RaHuSca database is restricted because of GDPR, ethical and copyright issues.

Responsibilities

The responsibility for the research team will be held jointly by the PIs (Chlopicki, Kuipers, Laineste), and the postdocs for the project (Fiadotava and two others who are yet to be hired). We will jointly manage data documentation, storage and backup, and data preservation, and update the DMP. After the project, data management will remain the responsibility of the PIs.