NanoBubbles project: Data Management Plan

Frédérique Bordignon

NanoBubbles project

D1.1 – Data Management Plan

Version 1 – November 2021

Main contributor: Frédérique Bordignon, École des Ponts Paris Tech, France
Project manager: Zoé Touré, Université Sorbonne Paris-Nord, France
Host PI: Raphaël Lévy, Université Sorbonne Paris-Nord, France

DMP template: Science Europe
Funder: European Commission (ERC Synergy Grant)

The NanoBubbles project has received Synergy grant funding from the European Research Council (ERC), within the European Union's H2020 programme, grant agreement no. 951393.

Project summary:
Science relies on the correction of errors to advance, yet in practice scientists find it difficult to erase erroneous and exaggerated claims from the scientific record. Recent discussion of a "replication crisis" has impaired trust in science both among scientists and non-scientists; yet we know little about how non-replicated or even fraudulent claims can be removed from the scientific record. This project combines approaches from the natural, engineering, and social sciences and the humanities (Science and Technology Studies) to understand how error correction in science works and what obstacles it faces, and stages events for scientists to reflect on error and overpromising.

The project's focus is nanobiology, a highly interdisciplinary field founded around the year 2000 that has already seen multiple episodes of overpromising and promotion of erroneous claims. We examine three such "bubbles": the claim that nanoparticles can cross the blood-brain barrier; that nanoparticles can penetrate the cell membrane; and the promotion of the "protein corona" concept to describe ordinary adsorption of proteins on nanoparticles. Findings based on error (non)correction in nanobiology should be generalizable to other new, highly interdisciplinary fields such as synthetic biology and artificial intelligence.

We trace claims and corrections in various channels of scientific communication (journals, social media, advertisements, conference programs, etc.) via innovative digital methods. We examine error (non)correction practices in scientific conferences via ethnographic participant-observation. We follow the history of conferences, journals, and other sites of error (non)correction from the 1970s (before nanobio per se existed) to the present. And we attempt to replicate neurobiological claims and, in case of non-replication, document obstacles to correcting those claims. Finally, we will spark a dialogue within the nanobiology community by organizing workshops and events at conferences for practitioners. Through the study and practice of nanobiology, we will analyse how, when and why science fails to correct itself, and explore ways to improve the reliability and efficiency of the scientific process.

Team members' institutions:
Université Sorbonne Paris Nord, Université Grenoble Alpes, Radboud University, Maastricht University, University of Twente, CNRS, IRIT, Ecole des Ponts ParisTech

1. Data description and collection or re-use of existing data

For this first version of the DMP, the project members have identified the following research outputs to be considered as data:

- Recordings and transcripts of interviews
- Field notes
- Scientific literature
- Scholarly works metadata
- Digital ephemera
- Software codes
- Experimental data and protocols
- Physical objects

1.1. Recordings and transcripts of interviews

Ethnographic, historical and informal interviews will be made. The transcripts of the interviews are generated from recordings that were made with the interviewees. They come in three different formats: recording, raw transcript, and pseudonymized transcript. This pseudonymisation process means that personal data would no longer be able to be attributed to a specific person without the use of additional information kept separately.

The transcription is done manually or with a tool installed on laptops.

Recordings are mp4 files. Experience shows that about 50 interviews are conducted per project. This requires between 250 and 500MB of storage, for sound files of 80 min/50MB each.

1.2. Field notes

Observations and thoughts will result in raw and pseudonymized notes in text format (docx or txt files). Photos and screenshots will sometimes be taken; these could be large files that require more storage space.

Image files are storage consuming with an estimation of 800MB max per fieldwork per member, amounting to less than 3GB storage needed.

1.3. Scientific literature

- **Documents coming from structured databases**

We will build up corpora of scientific texts by querying several legal sources, via their APIs or by downloading.
These documents will mostly be journal articles, but will also include magazine and newspaper articles, monographs, edited volumes and conference proceedings, textbooks, policy documents, government reports, parliamentary testimony, patents, legislation...
We plan to query the Elsevier and PubMed databases, Open Access archives, the ISTEX database (French national database of text archives) and the downloadable corpus of the PLoS journal.

In all cases, we will abide by the technical limitations set by these content providers (sometimes via subscription-based services held by members' institutions). These mainly concern the volumes that can be downloaded simultaneously and the operating licences.
We also plan to explore existing corpora, which are not corpora of full-texts, but corpora already resulting from the mining of full-texts. These are notably corpora of citation contexts (more or less sentences) but also the one that a team member is building within another research project (Cita&Re[1]), and which will be shared in open access.

- **Heterogeneous documents**

In some cases, the documents collected to be read or mined are collected from interviewees, or are derived from photographs taken during visits to archival centres.
These documents include technical reports, advertisements and promotional literature, conference programs, funding proposals and reviewers' comments, draft publications and reviewers'/editors' comments, correspondence, lab notebooks, manuals, conference reports, budgets...

We do not consider the scientific literature consulted/read for the scientific progress of the project as research data and it will be collected and shared between members via Zotero.

Wherever possible, the preferred format is XML based on the TEI schema, JSON as well.
In some cases, it will be necessary to convert PDF files into XML-TEI, and even OCRise images to get texts.
Finally, for the corpus of citation contexts, the CSV format is preferred.

The processing of the XML files can be done on the secured servers of the University of Grenoble-Alpes or at the IRIT (University of Toulouse). An estimate of the volumes resulting from this exploitation will be made once the delineation of the corpora has been decided by the team. PDF files and scans or photos of documents will result in large files, about 5/6GB seem necessary.

1.4. Scholarly works metadata

In this project, the metadata describing documents that will constitute the corpora of scientific documents are themselves research data that will be analysed and processed. These metadata are often even acquired before the constitution of corpora of full-texts with the querying of bibliographic databases (Scopus, Dimensions, PubMed, Ulrichsweb...), specific databases (abstracts of funded proposals), or dedicated services (Elsevier ICSR lab).
They are very easily accessible. They include textual data (authors, title, abstract, keywords, affiliations, name of the journal, etc.), but also numerical data (dates, disciplinary field codes) and identifiers (DOIs, ISSN).

Depending on the use made of this metadata, it will be processed in XML format or in CSV. The underlying schema will depend on the source but will be Dublin Core in most cases. When documents are collected from interviewees or created from images, metadata will be created manually and will be the same as the mandatory items in Dublin Core.

1 Project funded by ISITE-FUTURE (Université Paris-Est), dedicated to the investigation of citations of retracted papers.

1.5. Digital ephemera

Digital ephemera include various types of data available online on the Internet, collected manually, via dedicated APIs or scrapped with specific codes.

- **Tweets**

Tweets (textual content, tweetos and publication date) can be easily retrieved via the API on the Twitter development platform and the special plan for researchers (see https://developer.twitter.com/en/products/twitter-api/academic-research). Tools like DMI-TCAT (that IRIT can make accessible to the project) will be used.

- **Comments**

We are particularly interested in comments posted on PubPeer, whose owners already agreed to collaborate and provide data. The PubPeer posts produced by NanoBubbles will be posted with a 'Nanobubbles' account. Posts will link to the project's website.

- **Annotations**

It is part of the project to see how annotations of scientific outputs could benefit the correction of science. The way this kind of data has to be collected and promoted has to be discussed.

- **Websites/blogs pages**

This includes corporate websites, websites of lab groups in universities or government agencies, conference programs and reports, including CV/resumes or LinkedIn profiles. Apart from manual copies, when possible, the pages will be scraped, but within the conditions of the site and in compliance with the law in force (in any case these data will not be redistributed). We will take snapshots of the webpages of interest using the Internet Archive Wayback Machine (https://archive.org/web/) so that future readers can consult the archived pages even if they are no longer online.

- **Promotional videos**

They will be collected from websites, and might be transcribed if needed.

- **Mails from mailing lists**

Those mails are textual data which can be retrieved from one's own mailing tool but more conveniently from online archives of listservs.

Except for videos (which might not be numerous), this data results in textual documents in JSON, CSV, TXT or XML format.
The project needs to move forward before we are able to assess the volume this represents as it depends on the delineation of the corpus.

1.6. Software codes

Software and codes will be used according to their licence and will be produced by computer scientists in order to retrieve data and to analyse it. A *private* gitlab repository hosted at the University of Grenoble-Alpes will store the latest version of the code and the entire version history. The code intended for public release will be made available from a *public* gitlab

repository (https://gricad-gitlab.univ-grenoble-alpes.fr) hosted at the University of Grenoble-Alpes.

1.7. Experimental data and protocols

Since replication projects will be carried out, this will involve writing the detailed experimental protocols and recording the experiments. Experimental data will originate from measurements performed in a laboratory environment, including images from microscopy experiments, graphs and table from chemical and biochemical analysis, etc

1.8. Physical objects

These objects are given by interviewees or souvenirs/goodies from conferences. If needed, they will be described with dedicated metadata in a readme file.

2. Documentation and data quality

2.1. Recordings and transcripts of interviews

Metadata describing a set of interviews will be stored in a separate readme file, mentioning the place where raw data is stored.

2.2. Field notes

Metadata describing a set of interviews will be stored in a separate readme file, mentioning the place where raw data is stored.

2.3. Scientific literature

The metadata format depends on the database providing them but they will all result in CSV files when it comes to mix different sources before storage and processing. The describing items of the metadata will be based on common standards like Dublin Core or RIS: authors, dates, title, source title, author affiliation, abstract, keywords, DOI, ORCID, number of citations, bibliometric indicators related to the source, ISSN/ISBN, Open Access status.
If enrichments come from the project and feed the metadata template, they will be described in a readme file.
The two research engineers will check the consistency of the subsets of metadata.

2.4. Scholarly works metadata

We will not generate metadata for this type of data, but we will nevertheless add a readme file with a short description of the set of metadata contained in a specific folder.

The two research engineers will check the consistency of the subsets of metadata.

2.5. Digital ephemera

The metadata will differ according to the provider of the data and the way we will retrieve data; we will have to decide what template we will use when it comes to providing metadata from scratch.

- **Tweets** come along with a tweetID and a userID, and information about when it has been tweeted, and how many times it has been retweeted or liked. A documentation is available for developers using the API service.
- **Comments** from PubPeer will come along with commentID and userID on Pubpeer, DOI/PubMed ID of the commentated paper and its retracted or correction notice if it exists, dates of publication on the platform.
- For **Annotations, Websites/blogs pages, Promotional videos** and **Mails from mailing lists**, the team members will decide in the following months what should be precisely described with the intention to be mixed with other metadata and what should be described as a subset with a simple readme file.

The two research engineers will check the consistency of the subsets of metadata.

2.6. Software codes

Codes will be documented and commented when shared.

2.7. Experimental data and protocols

The replication project is multisite in the sense that synthesis, characterisation and imaging/sensing will be repeated not just in Levy's lab, but also with partners (collaborators or sub-contractors). In Levy's lab, an open lab notebook and open data approach will be followed, documenting the research process in as much detail as possible and sharing all the data acquired. The partners will be encouraged to follow the same approach and will otherwise be requested to provide a minimum mutually agreed list of data defined at the beginning of the partnership. In that third step, the data are therefore composed of the open lab notebook, the experimental data both internal and from partners, and the communications with the partners.

Replication projects do not start at the beginning of the NanoBubbles project, therefore, decisions about which electronic lab notebook tool to use and how to open the protocols will be discussed soon. The DMP will be updated accordingly in the next version.

2.8. Physical objects

On the occasion of interviews and conferences, the project members collect or receive physical objects of a very varied nature, whether they are goodies or an unexpected object characterising the topic of the interview. These objects are kept by each member but if they become research objects that can be cited, they will be photographed. This image file will then be described with some simple metadata and indexed in the Zotero shared library.

3. Storage and backup during the research process

3.1. Storage spaces

There will be 3 different storage spaces: members' computers, cloud-based storage systems provided by the IT services at all the team members' institutions (data protection guarantee), and SURFdrive, the Dutch secure community cloud system where non-Dutch members will access thanks to guest accounts validated by the University of Maastricht.
The project folder will be our master copy location.
All members get a virtual drive with 500 GB of storage space. Therefore, all non-sensitive data needed to be shared between team members will be stored in the project folder on SURFdrive. Sensitive data will be stored on personal folders on SURFdrive and will not be shared with all team members. The interviewer will give permissions to read those files to selected members. Therefore, thanks to SURFdrive, folders that contain privacy sensitive information will be encrypted and access restricted. Backup and recovery of 30 days are guaranteed.

When it comes to interviews and field notes, data will first be stored on the password protected computer of the interviewer, then synchronized with his/her own encrypted cloud-based storage system (set by his/her institution). USB flash drives are banned. When needed and once anonymized, the files will be shared via SURFdrive.
Computer scientists and research engineers dealing with large files will download and compute them directly on servers at their institutions.

3.2. File-naming convention

We will apply a standardized file-naming convention. Our file names consist of the following items separated by an underscore:

- an abbreviation for the data type,
- a content description (generic to specific),
- the date of modification (international standard: YYYYMMDD),
- a version number (v0.0).

We do not use special characters, full stops or spaces.

Example: WORKS_ScopusExportSNA_20221023_v5.0.csv
Example: EPHEM_Tweets_20221023_v3.6.json
Example: INTERV_ACSConf_20221023_v1.docx

The abbreviations for data types are:

- INTERV for Recordings and transcripts of interviews
- NOTES for Field notes
- WORKS for Scientific literature
- METADATA for Scholarly works metadata
- EPHEM for Digital ephemera
- CODES for Software codes
- OBJ for Physical objects

4. Legal and ethical requirements, codes of conduct

4.1. Recordings and transcripts of interviews

The interviewer will gain informed consent from the interviewee for preservation of the data resulting from the interview. In order to comply with the European motto "as open as possible, as closed as necessary", the interviewers will pseudonymise data before sharing.

The pseudonymisation method will be described in a readme file accessible to those researchers who need to access it.

Audio files and raw transcripts will not be uploaded to the common NanoBubbles folder on SURFDrive but will be saved on a dedicated folder with permissions to the interviewer/creator and a minimum number of people who need to see.

4.2. Field notes

When field notes contain personal data, the observer will gain informed consent from the person observed for preservation of the data resulting from the observation. In order to comply with the European motto "as open as possible, as closed as necessary", the interviewers will pseudonymise data before sharing.

The pseudonymisation method will be described in a readme file.

Field notes with personal data will not be uploaded to the common NanoBubbles folder on SURFDrive but will be saved on a dedicated folder with permissions to the interviewer/creator and the people he has to work with.

4.3. Scientific literature

Intellectual property rights and ownership will of course be respected, although there is no risk of infringing them as the aim of the project is not at all to disseminate the corpus of collected documents, but to analyse them. This is permitted by the European "digital single market" directive of 2019, when text and data mining is for research purposes.
The outputs of the mining process will be data that would be shared.

4.4. Scholarly works metadata

Intellectual property rights and ownership will of course be respected, although there is no risk of infringing them as the aim of the project is not at all to disseminate the corpus of collected documents, but to analyse them. This is permitted by the European "digital single market" directive of 2019, when text and data mining is for research purposes.
The outputs of the mining process will be data that would be shared.

4.5. Digital ephemera

This kind of data is associated with personal data and we are in the process of checking with legal departments how GDPR applies to this.

4.6. Software codes

If existing codes or parts of codes are reused, we will of course respect the licenses associated with the original code.

4.7. Experimental data and protocols

When we have to use protocols already published by others, we will of course cite them.

5. Data sharing and long-term preservation

We will abide by the European motto "as open as possible, as closed as necessary". Therefore, we will make every effort to share the data we have as soon as possible in compliance with applicable laws.
We will disseminate data as the project progresses, we will not wait until the end of the 5 years to do so.
Data (including codes) will be shared on trusted data repositories like Zenodo or Mendeley data and institutional repositories (e.g.: DataverseNL or the future French data repository) to guarantee longevity, the compliance to the FAIR principles and the possibility to reuse and cite the datasets thanks to the DOI minted by the repositories we will choose.
Standard formats are privileged during the project so no specific tools would be needed to read the data.
For some important outputs, data papers will be published to describe how data were obtained and how to get the most of it.

6. Data management responsibilities and resources

The institutions as members of the consortium will all be responsible for the data. Each interviewer/observer will be in charge of the data files resulting from interviews and field observations. Apart from this type of data, several persons will be in charge of data collection, metadata production, data quality, storage and backup, data archiving, and data sharing:

- Frédérique Bordignon (researcher, academic librarian),
- Zoé Touré (project manager),
- 2 research engineers (with skills in data management) recruited and placed close to Cyril Labbé (PI).

Maria Vivas-Romero from University of Maastricht will help as a data steward.
WP5 hosts an IT platform whose participants will be partly dedicated to the curation and sharing of the data, and will support other team members when it comes to the technical part.