Data Management Plan (Y2BA - AI in Tourism)

Rens van den Berg, Stijn Heesters

Breda University of Applied Sciences, Breda, Netherlands

November 2024

1 General Information

1.1 Name of applicant and project number

Project leader name: Rens van den Berg, Stijn Heesters

Project number: Project 2A ADSAI (retake)

1.2 Name of data management support staff consulted during the preparation of this plan

Data management support staff consulted: Bram Heijligers

Date of consultation with support staff: November 15, 2024

2 What data will be collected or produced, and what existing data will be re-used?

2.1 Will you re-use existing data for this research?

No.

2.2 If new data will be produced: describe the data you expect your research will generate and the format and volumes to be collected or produced.

New data to be generated will primarily be in the form of CSV files. The data expected from this research includes, but is not limited to, numeric entries stored in databases and spreadsheets. Emphasis will be placed on using open and standard formats to facilitate sharing and long-term re-use of data.

2.3 How much data storage will your project require in total?

Data storage requirements for the project are estimated to be within the range of **10 – 100 GB**.

3 Metadata and Documentation

3.1 Documentation Accompanying the Data

Comprehensive documentation will be prepared to support the reuse of the data. This documentation will include detailed methodology on data collection processes, analytical and procedural information, definitions of variables, and units of measurement. The documentation will be accessible via various formats such as a database with links to each data item, 'readme' text files, file headers, codebooks, and lab notebooks. These measures ensure that future researchers will have all necessary context to understand and reuse the data effectively.

3.2 Metadata

Descriptive metadata will accompany all data sets to aid in their discovery and to ensure they are findable, accessible, interoperable, and reusable (FAIR). The following strategies will be used:

- Adherence to community metadata standards as advised by the Research Data Alliance, which maintains a comprehensive Directory of Metadata Standards.
- Depositing data in certified or trustworthy repositories, which involves detailing the data according to a metadata standard scheme, typically employing the Dublin Core or DataCite Metadata Schema.

Researchers are encouraged to consult the university library and institutional Research Data Management (RDM) support staff for further advice on appropriate metadata standards and practices to enhance the data's long-term value and accessibility.

4 Data and Metadata Storage and Backup

4.1 Data Storage Details

- **Initial Data Capture:** All data collected, whether quantitative (from surveys) or qualitative (from interviews), will initially be stored in raw format directly on secure servers via encrypted transmission protocols.
- **Storage Format:** Data will be stored in standard formats conducive to long-term preservation and accessibility; quantitative data in CSV format and qualitative data in audio or video formats, subsequently transcribed to text files.
- **Security Measures:** Access to the SURF Research Drive is restricted to authorized project personnel only, using two-factor authentication and encrypted access controls.
- **Data Integrity Checks:** Regular data integrity checks will be performed to ensure that no data corruption occurs. This includes checksum verifications and periodic reviews of data access logs.

4.2 Archiving and Long-term Preservation

Upon completion of the project, essential data will be transferred to a long-term archival service that meets the NWO guidelines for data preservation.

- **Selection of Data for Archiving:** Critical data that underpin the project's findings and publications will be selected for long-term preservation.
- **Archival Format:** Data will be converted into formats suitable for long-term preservation, such as PDF/A for documents and TIFF for images.
- **Metadata Documentation:** Comprehensive metadata will accompany the archived data to ensure that it remains discoverable and usable in the future.

- **Access to Archived Data:** Access to archived data will be controlled but will adhere to the principles of being as open as possible and as closed as necessary, depending on sensitivity and privacy requirements.

4.3 Storage and Backup During the Project

During the project, all raw and processed data, along with metadata, will be stored securely on the SURF Research Drive. This platform is specifically designed to meet the BUas Regulation pertaining to Research Data Management and provides robust, managed storage with automatic backup features. This ensures ample storage space and high data security, which are critical for handling the project's data needs.

4.3.1 Additional Storage Details

- The data and metadata storage will exclusively utilize the institution's networked research storage, SURF Research Drive, to ensure data integrity and security.
- For backup, regular snapshots of the data will be taken as part of the project management process. These backups are crucial for safeguarding the work and ensuring that data and progress are not lost due to unforeseen issues.

4.4 Data Security and Protection of Sensitive Data

The project does not involve sensitive data that requires special security measures beyond those provided by the institution's networked research storage. However, the following additional security measures will be implemented to ensure the protection of all project data:

- Enforcing strict password policies.
- Restricting access to the data to only those researchers designated by the project leader.
- Employing encryption for both data at rest and in transit.

- Conducting data processing activities such as anonymization to further safeguard data integrity and confidentiality.

Risk Management: Potential risks such as human error and vulnerabilities in the survey platform are acknowledged. These risks are managed by a combination of the security measures mentioned, along with continuous monitoring and regular evaluations of the security practices in place.

5 Data Collection Methods

5.1 Overview of Data Types

This project will collect and process both quantitative and qualitative data. Here is an overview of the types of data:

- **Quantitative Data:** This includes responses from surveys, which are structured to allow for statistical analysis and will be stored in CSV format.
- **Qualitative Data:** Comprises recorded interviews and analyzed text data, providing deeper insights into the survey findings.

5.2 Collection and Processing of Quantitative Data

Quantitative data will be collected through structured surveys conducted using Qualtrics, an advanced online survey tool that allows for sophisticated data collection and analysis:

- **Survey Instrument:** Surveys will be designed within Qualtrics to capture scalable responses on various aspects of the research topic. The platform's flexibility in designing complex survey structures will enable precise data gathering according to the research needs.
- **Data Processing:** Responses from the surveys will be initially recorded in raw form directly within the Qualtrics platform. Data will then be exported and cleaned, formatted into CSV files using data processing scripts to ensure accuracy and usability. This step will involve checking for inconsistencies, removing any incomplete responses, and standardizing data formats for subsequent analysis.

5.3 Collection and Processing of Qualitative Data

Qualitative data will be collected primarily through interviews:

- **Recording Interviews:** Interviews will be conducted using Microsoft Teams, which allows for both video and audio recording. The choice of recording will depend on the consent provided by the interviewees.

- **Storage:** Immediately after recording, the digital files will be securely transferred and stored in an encrypted format on the institution's secure network drive designated for project data. This ensures both security and accessibility for designated research team members.

- **Transcription and Analysis:** Recorded interviews will be transcribed verbatim. Transcripts will be analyzed to extract relevant qualitative data. The transcription process will adhere to ethical guidelines, ensuring confidentiality and anonymity of the participants.

- **Metadata Creation:** Metadata for each interview will include details about the interviewee (as permissible), date of the interview, and main topics covered, enhancing the findability and reusability of the data.

5.4 Ethical Considerations

All data collection methods, particularly those involving human subjects, will comply with institutional ethical standards. Participants will be informed about the purpose of the research and their rights, including their right to withdraw from the study at any time. All personal data will be handled in accordance with GDPR regulations to ensure privacy and data protection.

5.5 Data Validation and Quality Control

To ensure the reliability and validity of the data collected, the following measures will be implemented:

- **Pre-testing:** Survey instruments and interview protocols will be pre-tested with a small segment of the target population.

- **Regular Audits:** Data collection and processing methods will be regularly audited to identify and correct any issues or biases.

This comprehensive approach to data collection ensures that all data, both quantitative and qualitative, are collected in a manner that supports the integrity and objectives of the research project while adhering to all ethical and legal standards.

6 Handling of Personal Information and Intellectual Property Rights

6.1 Processing and Storing Personal Data

Will personal data be processed and/or stored during the project? Yes

To ensure compliance with data protection laws, including GDPR/Dutch AVG, the following measures will be implemented:

- Observance of legislation on personal data is guaranteed at BUas by the Executive Board and the Academy Director, respectively, following the mandate in the rules on jurisdiction, supervised by the Data Protection Officer.
- Specialized support staff from the university, such as Rouah Marwa, and institutional RDM support staff have been consulted to ensure all data protection measures are up to standards and risks are managed appropriately.

6.2 Ownership and Intellectual Property Rights of Data

The management of data ownership and intellectual property rights is structured as follows:

- **Ownership of Data:** The data collected and processed during the project will remain under the ownership of BUas. For multi-partner projects, rights to control access to data will be clearly defined in the consortium agreement to ensure all parties are aware of their rights and responsibilities.
- **Intellectual Property Rights:** Intellectual property rights will be managed in accordance with BUas guidelines and agreements with project partners. All relevant intellectual property rights related to the data collected, such as copyrights or patents, will be specified in the collaboration agreements to prevent any disputes over data usage.

- No specific intellectual property rights issues are expected to arise that would affect the data; however, any potential issues will be addressed as per the collaborative agreements and BUas regulations.

Additional Documentation and Agreements:

- A detailed collaboration agreement including intellectual property rights management will be part of the project documentation.
- Advice and review by the Ethics Committee and relevant legal advisors will be sought to ensure all data management practices comply with national and international laws.

This section ensures that all aspects of data protection and intellectual property rights are comprehensively covered, adhering to legal standards and institutional policies. Adjustments can be made as necessary depending on changes in legal requirements or project scope.

7 Data Sharing and Long-term Preservation

7.1 Selection of Data for Long-term Preservation

All data resulting from the project will be preserved for at least 10 years, adhering to NWO's expectation and any applicable legal, contractual, or regulatory requirements. Decisions on what data to preserve will consider the data's relevance to the research community and the potential for future use. The primary criteria for preservation will include data validity, integrity, and demand, alongside the nature and content of the data collected, which is mainly quantitative survey data in CSV format.

7.2 Restrictions on Data Access

Are there reasons to restrict access to the data once made publicly available? No

No legal, IP, privacy-related, or security-related reasons preclude making the data publicly available. All necessary measures, such as anonymization and summary statistics, have been applied to ensure that the data can be shared without compromising individual privacy or the integrity of the data.

7.3 Data Availability for Re-use

All data resulting from the project will be made available for re-use. The data to be shared will include all datasets underpinning research papers, especially those necessary to validate research findings. The guiding principle will be 'as open as possible, as closed as necessary,' considering ethical limitations, property rights, and commercial interests.

7.4 Timing and Duration of Data Availability

When will the data be available for re-use, and for how long? Data will be available as soon as the corresponding research articles are published.

Data will remain available for at least 10 years post-publication, ensuring ample time for further research and analysis. There are no anticipated embargo periods; however, any necessary delays in data sharing to protect intellectual property will be clearly communicated, following NWO guidelines.

7.5 Repository and Licensing

The data will be archived and made available for re-use in a certified repository that adheres to international standards for data preservation. The chosen repository will provide persistent identifiers and use metadata standards broadly accepted by the scientific community. Data will be licensed under a Creative Commons license, facilitating re-use while protecting intellectual property.

7.6 Publishing Analysis Software

The analysis software developed during this project, including specific scripts, codes, or algorithms, will be made available alongside the dataset. This software will be essential for accessing, interpreting, and re-using the data. The software will be published on platforms like GitHub, which supports the FAIR principles for software sustainability, including clear documentation, open-source licensing, and persistent identifiers.

Software Availability:

- **Tools and Software Required:** Users will need R and potentially VScode to utilize the data and software efficiently.
- **Sustainability:** The software will be maintained with community input and regular updates to ensure its long-term usability.

8 Data Management Costs

8.1 Resources Dedicated to Data Management and FAIR Principles

The project will allocate substantial resources, including financial and time, to ensure that all data management practices adhere to the FAIR (Findable, Accessible, Interoperable, Re-usable) principles. Detailed planning has been undertaken to ensure efficient data curation and preservation:

- **Time:** Significant staff hours have been dedicated to the data management tasks, including data curation, metadata creation, and quality control, to prepare data for sharing and long-term preservation.
- **Financial:** The budget for data management includes costs for data storage, backup services, and any fees associated with depositing data in certified repositories. This also covers potential charges for maintaining high standards of accessibility and interoperability.

Additional Resources for Data Preparation and Deposit:

- Additional resources will be needed to prepare the data for deposit, specifically for the creation of detailed metadata and the conversion of data into formats that are widely usable across different platforms and systems.
- The costs associated with these activities have been included in the budget outlined in the NWO grant application. Specific allocations have been made for:
  - Metadata development and standardization.
  - Data formatting and interoperability enhancements.
- If additional costs arise, these will be covered by contingency funds set aside within the project budget. The project management team will monitor expenditures closely and adjust plans as necessary to ensure that all data management activities are funded adequately.

This section outlines the comprehensive approach to budgeting for and managing the costs associated with data management. It ensures that the project's data will be prepared adequately for sharing and preservation, adhering to the FAIR principles and meeting the expectations of both the research team and funding agencies.

9 Adherence to FAIR Principles

9.1 Findability

- Each dataset is assigned a DOI (Digital Object Identifier) to ensure it has a unique and persistent identifier.
- Metadata is created using rich, descriptive terms that are indexed in an accessible search platform.

9.2 Accessibility

- Data is accessible through a standard protocol (e.g., HTTP, FTP), and accessibility conditions are clearly documented.
- Metadata remains accessible even if the data is removed or access is restricted.

9.3 Interoperability

- Metadata incorporates standards such as Dublin Core, ensuring it can be integrated with other datasets.
- Data and metadata are linked using standardized vocabularies endorsed by the research community.

9.4 Reusability

- Metadata details the data's provenance, including the methodology used to collect the data and any processing or transformations it has undergone.
- All datasets are released under a Creative Commons license, facilitating reuse under specified conditions.