Data Management Plan

HISTORY OF CHANGES			2
Version	Publication date	Changes	
1.0	28/10/2022	Initial version	

The information, facts and opinions set out in this report are those of the authors. They do not necessarily reflect the opinion of the European Union or the European Commission, which are not responsible for the use which may be made of the information contained in this report by anyone.

Contents

1. Introduction and summary of datasets produced by SAPEA

SAPEA is part of the European Commission's Scientific Advice Mechanism. Together with the Group of Chief Scientific Advisors, SAPEA works to provide independent scientific

advice to European Commissioners, in support of their decision-making. SAPEA also raises awareness of scientific advice and evidence in policymaking, and stimulates debate in Europe about these issues.

SAPEA's main purpose is to undertake comprehensive evidence analysis and synthesis to inform the Scientific Opinion of the Group of Chief Scientific Advisors. This work is done by interdisciplinary working groups of scientific experts, who produce Evidence Review Reports and other scientific outputs. The process is supported by systematic reviews of the existing literature.

The main types of datasets produced by SAPEA as a result of this work are set out below.

The SAPEA consortium supports open science practices as a means towards greater openness, transparency and inclusiveness. For example, adherence to best practices in open science ensures the rigour and reproducibility of systematic literature reviews, as well as the greater visibility, discoverability, and uptake of SAPEA's outputs. SAPEA follows the broad principle of '*as open as possible, as closed as necessary*'. This Data Management Plan sets out how SAPEA's approach to data management meets the principles of *findability, accessibility, interoperability and reusability* (FAIR).

Data overview

SAPEA's main scientific outputs that produce associated data are as follows:

- **Evidence reviews**, made up of Evidence Review Reports and other evidence-based outputs, such as literature reviews and reviews of the policy landscape. SAPEA's evidence reviews are based on secondary data i.e. existing public-domain information, principally peer-reviewed literature and grey literature (reports, government publications etc). SAPEA evidence reviews are extensively referenced to the sources cited, resulting in a comprehensive dataset of bibliographic references; typically, this is a dataset of several hundred references. SAPEA releases these data for free and unrestricted use by other researchers, or anyone with an interest. The value of these data is in explicitly stating the sources on which SAPEA evidence reviews are based. The datasets will be made available in commonly-used formats (see below for more details) that ensure maximum access and uptake.

- **Survey work and interview data**. This data is collected as part of SAPEA's activities in work packages 2, 5, 6 and 7, specifically:
  - WP 2 (Scientific topics): SAPEA may occasionally carry out structured interviews with experts or conduct expert elicitation.
  - WP5 (Interaction of academies and sharing best practices): Involves survey work with academies.
  - WP6 (Early and mid-career researchers): Involves survey work to identify hurdles facing early- and mid-career researchers to engage in science advice for policy activities.
  - WP7 (Strategic development): Involves interviews and possible survey work with a range of stakeholders.

The data from these work packages will be released whenever it is possible to do so, whilst also observing the need to preserve confidentiality and privacy in some limited cases (for example, some data within WP6 and WP7). At all times, SAPEA is observant of the requirements set out in the Grant Agreement on data confidentiality, security, data protection and ethics. Full details are be set out in the template, which must be completed for each piece of work where data is generated (see Annex 1).

Ownership and implementation

The Data Management Plan is owned and supported by the Academy Networks in the SAPEA consortium. Overall management is provided by the coordinating organisation, acatech.

The lead Academy Network responsible for an area of work where data is produced (for example, a work package or Evidence Review Report) should ensure that the template and checklist (see Annex 1) are completed satisfactorily and in a timely manner. At the start of the work, the lead Academy Network will convene a team meeting to complete the template, providing basic information about the datasets that will be produced. The template will be a living document that is kept on the shared platform (SharePoint) and updated as needed. Adherence to the Data Management Plan will be monitored by the Project Coordinator (acatech), which will issue a checklist that must be completed and returned at the end of each piece of work where data are produced. Annex 1 sets out the initial version of the
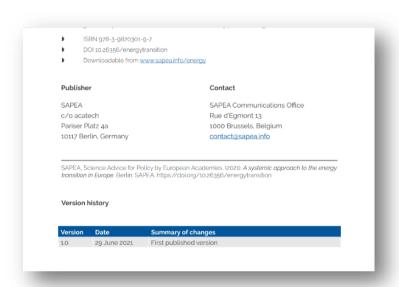
template and checklist; these will be refined over the course of the SAPEA project, in response to feedback.

Quality assurance

A short summary of this Data Management Plan is included in SAPEA's published Quality Guidelines.

2. FAIR data

2.1 Making data findable, including provisions for metadata

Persistent identifiers. SAPEA assigns persistent identifiers (PIDs) (for example, DOIs, and ISBNs where applicable) to all its main publications, such as Evidence Review Reports, Expert Workshop reports etc. As an aid to discoverability, this practice will be extended to more of SAPEA's published outputs, including systematic literature reviews, report chapters, executive summaries, the SAPEA Strategic Development Plan etc. These outputs will be made available as separate digital objects, each with their own rich metadata created by SAPEA.

Bibliographic datasets. SAPEA is meticulous in producing comprehensive bibliographic datasets for its Evidence Review Reports and other outputs, such as systematic literature reviews and reviews of the policy landscape. Metadata to cited references is generally harvested from Scopus and similar bibliographic sources. These bibliographic datasets will be given their own DOI and made available in machine-readable formats that can be easily harvested by others, such as researchers or anyone else with an interest in following up the sources. Each cited reference will include its own linked DOI, where this is known, so that access to the original source is both persistent and straightforward.

Survey and interview data. With the exception of potentially-sensitive data that should remain confidential, survey and interview data will be transcribed, anonymised and published as a text-based collection or dataset (depending on the format of the information), with a DOI and metadata that aids its discoverability and access.

Version control. SAPEA has a rigorous approach to version control. It records the version history of its publications, normally on the verso side of the title page. A procedure will also be established for version control of datasets. Versioning will be managed and conveyed through the associated metadata.

2.2 Making data accessible

 Trusted repository. All datasets arising from SAPEA will be uploaded to Zenodo, a trusted repository built and maintained by CERN. A DOI will be allocated to each dataset in Zenodo, which resolves to the relevant digital object (such as a bibliographic dataset, survey dataset etc).

Licensing. In line with most SAPEA publications, survey and interview data will normally be published under a Creative Commons CC-BY licence (currently 4.0 Intl), which is a permissive licence that allows extensive (re)use, as long as proper attribution is made to the author(s). In limited circumstances, some SAPEA data may be kept confidential and not published, due to its sensitivity and the possibility of it being traced back to an individual or organisation, owing to the nature of its content.

 Bibliographic datasets will be published under a Creative Commons CC0 licence, which provides data free of any restrictions.

2.3 Making data interoperable

Formats. Metadata will be published under a CC0 licence, free of all restrictions, and made available through Zenodo for downloading/harvesting. Bibliographic datasets will be published in accessible formats, such as .csv and .txt, as well as .ris (supported by a number of reference management services, such as EndNote, Zotero and Mendeley).

Standardisation and interoperability. SAPEA applies rich metadata (as a minimum, corporate author; publication title; date of publication; place of publication; source of funding; grant project name, acronym and number; licensing terms). SAPEA makes use of keywords and subject terms suggested by Zenodo, ensuring consistency in their use across relevant outputs published by SAPEA.

2.4 Increase data re-use

Method statements. As stated, most of SAPEA's data will be released under permissive licences, to encourage uptake and re-use by third parties. With relevant outputs (principally Evidence Review Reports, systematic literature reviews, associated bibliographic datasets, survey and interview data), SAPEA will include a detailed method statement, which will set out exactly how data have been gathered, analysed and interpreted.

- Every Evidence Review Report contains relevant information such as the background to the report, expert selection process, roles and responsibilities of participants in the process, the working structure and other quality assurance processes. Details of literature searches are summarised. They can include topic-related searches requested by Members of the Working Group, and a mapping of the policy landscape.
- In the case of bibliographic datasets, these are references cited within the main body of the Evidence Review Report or systematic literature review. The metadata to these references are harvested from public domain sources, such as Scopus and similar sources.

- Systematic literature reviews are carried out under published protocols, stating explicitly how the review is carried out (for example, the search strategy, inclusion and exclusion criteria, sources to be searched). This provides an important tool for their reproducibility.
- For primary data generated through SAPEA i.e. surveys and interviews, SAPEA will write up and publish the details of how the data was assembled, analysed and reported. In all cases, SAPEA will take rigorous steps to protect the identify of participants/respondents. The means of how this is done includes the use of generic descriptors and removing direct identifiers, such as name, organisation etc. Where necessary, SAPEA will take professional advice on best practice.

3. Other research outputs

There are no other data generated by SAPEA.

4. Allocation of resources

All work on gathering, preparing and publishing data will be undertaken by SAPEA staff, within the budget allocation for the SAPEA project; SAPEA does not need to rely on any third-party support.

5. Data security

Security and back-up. The primary copy of SAPEA outputs will be uploaded to Zenodo. The partner ALLEA, as work package leader for Communications (WP3), is responsible for the timely preparation and upload of the datasets. Back-up copies are kept and maintained securely by the project coordinator, acatech, on behalf of the Consortium, both during and after the project. The retention policy is not restricted; however, acatech will keep copies for at least 5 years after the end of the project. acatech keeps these copies on SharePoint, a cloud-based service hosted by Microsoft, using physical data storage in Germany. Access rights to SharePoint are managed by the project coordinator. The backup and restore features meet the highest security standards; daily backups are stored on a server in the T-systems vcloud. An additional virtual copy of the server is stored in the vCloud, to which only administrators have access. Veeam Backup for Microsoft 365 Licence is used. If data need to be restored, they are played back directly to the SharePoint Online.

Data privacy. SAPEA has a written, published policy on data privacy and confidentiality (see: Privacy and data protection policy – SAPEA). As mentioned already, a limited amount of data cannot be released openly, due to reasons of privacy and confidentiality. These data will be stored as an anonymised version by acatech, which will apply the appropriate access permissions to the data.

Professional support. When necessary, SAPEA is able to call on advice from Cardiff University (part of the SAPEA consortium), which has a dedicated professional team dealing with research data management issues.

6. Ethics

Personal data. SAPEA collects personal data on individuals as part of its survey work. For example, SAPEA may collect data on an individual's gender, country of work, etc. SAPEA will always comply with the requirements of GDPR. It will always obtain consent to gather, process and release data about individuals. Whenever necessary, SAPEA will protect

individuals involved in SAPEA work, by making no attribution of particular comments or views. All published survey/interview data will be anonymised, so that individuals and their organisations cannot be identified. Certain sensitive data will not be made public, due to the need to maintain the privacy of individuals and/or their organisations.

The Project Coordinator will inform SAPEA consortium partners regularly about the requirement to protect personal data and will monitor the actions taken. At the end of each reporting period, the partners will be reminded by the Project Coordinator to again check that they have adhered to all principles and rules, especially the data minimisation principle, for example, deleting all personal data which are not no longer required.

Professional support. The coordinating organisation, acatech, has a Data Protection Officer, who may be consulted if required.

7. Other issues

No other provisions are needed within this Data Management Plan.

ANNEX 1

Data Management Plan Template and Checklist

Introduction

The SAPEA Data Management Plan is a requirement of the Grant Agreement.

A Data Management Plan template and checklist <u>must be completed</u> by the SAPEA staff team involved in work packages where datasets are produced. These include:

- **Bibliographic datasets** for Evidence Review Reports
- Other evidence-based outputs, such as systematic **literature reviews** and **mappings of the policy landscape.**
- Work packages where **survey and/or interview data** is generated. These include WPs 5, 6 and 7, and possibly WP2.

Staff members of the team must follow the procedures set out in the SAPEA Data Management Plan [link]. The schedule is as follows:

- The **template** should be completed at the **start of the work** (such as an evidence review, survey or set of interviews). All datasets to be produced should be detailed. The completed template must be uploaded to the shared platform (SharePoint). Information in the template must be kept updated; for example, if new databases emerge over the course of the work.
- The **checklist** should be completed **within 2 weeks of submitting the dataset** for publication and/or storage, and sent to the Coordinator.

DATA MANAGEMENT TEMPLATE
To be completed at the start of all projects where datasets are to be generated and continuously updated throughout the project's life

Work package number:

Lead Academy Network:

Name(s) of person(s) completing the template:

Name of the scientific topic or brief description of the survey or interview(s) to be conducted:

1. **Summary of data** (*applicable to all datasets*)
List below the types of data to be generated by the project, their purpose, the format and estimated size

Dataset number and short name	Brief description of dataset	Types of data (e.g. bibliographic, survey, interview data)	Purpose of dataset (what is it for?)	Formats (e.g. possible file formats for output)	Estimated size (e.g. estimated number of records or files)	Comments
1.						
2.						
3.						

Add further lines, if needed

2. **Methods for collecting and processing the data** (*applicable to all datasets*)
Describe below the methods to be used to collect and process the data (add further lines if needed)

Dataset number and short name	Method to be used for data collection and processing (i.e. outline the approach to data collection and analysis, including details of any software to be used)
1.	
2.	
3.	

Add further lines, if needed

3. **Ethical, confidentiality and licensing issues** (*not applicable to bibliographic datasets, only survey and interview data*)
Describe below any issues around personal data and privacy.

Dataset number and short name (from previous table)	Does the dataset include private data about individuals? (Yes/no)	If yes, please give details and describe how private data will be protected and how compliance with GDPR will be met. Please ensure that a consent form has been signed and returned.
1.		
2.		
3.		

Add further lines, if needed

Describe below any issues around data confidentiality and protection of sensitive data

Dataset number and short name	Is there a need to preserve confidentiality around particularly sensitive survey or interview data? (Yes/no)	If yes, please give details about this and describe how the data will be managed
1.		
2.		
3.		

DATA MANAGEMENT CHECKLIST
To be completed within 2 weeks of publishing or uploading datasets

Work package number:

Lead Academy Network:

Name(s) of person(s) completing the checklist:

Name of the scientific topic or brief description of the survey or interview(s):

Name and brief description of dataset:

Area of work	Question	Completed? (Yes/no/not applicable)	Date of completion	Comments
Data description	Has a short description been written of what the dataset contains?			
Method description	Has a description been written of how the data were collected, analysed and the results written up?			
Metadata	Has high-quality and consistent metadata been assigned to the dataset (e.g. creator/title/keywords etc)?			
Identifiers	Have appropriate identifiers (e.g. DOIs) been assigned to the dataset?			
Accessible formats	Has the dataset been made available in the appropriate formats for accessibility and uptake?			
Version control	Has version control been implemented for the dataset?			
Licence	Has the appropriate Creative Commons licence been assigned to the dataset?			
Data protection	In the case of survey and interview data, have the necessary consent forms been obtained?			
Data privacy	In the case of survey and interview data, have the data been anonymised, whenever necessary?			
Sensitive data	In the case of survey and interview data, have sensitive data been identified and not released? Are the correct permissions in place for restricted access within SAPEA?			

Data accessibility	Has the dataset been uploaded to Zenodo?			
Data security	Has the dataset been given to acatech for secure storage and back-up?			