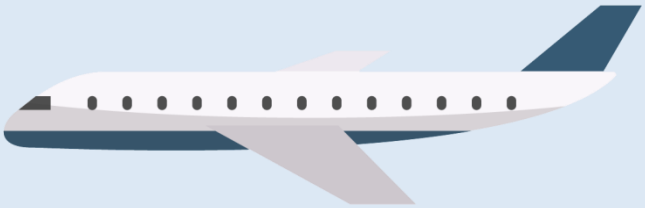


SAGE Data Deposit Manual

Version 1.1

PolicyWise
for Children & Families

Revision History

Version	Date	Produced by	Change
V 1.0	June 24 2016	Lucie Richard	New document
V 1.1	Aug 12 2016	Lucie Richard	Edits to reflect change in procedure, types of data accepted

Contents

Section I: Introduction

Overview

Secondary Analysis to Generate Evidence (SAGE) is a platform where research data, service delivery data, and (at a later stage) administrative data are catalogued and managed for secondary research and policy use. SAGE provides technical infrastructure and governance processes that protect participant/client privacy and ensure ethical re-use of data.

Why deposit data with SAGE?

Data is an incredibly valuable resource that often has significant re-use potential beyond answering the original question for which it was collected. Sharing data reduces the system-wide cost of conducting research and evaluating impact. It also decreases the overall burden on clients/participants. In addition to these high-level benefits, there are many immediate benefits to sharing your data:

- Your visibility and profile as a researcher and/or organization is enhanced;
- Your data becomes cited in a larger number of publications, white papers and other channels than otherwise possible;
- Data sharing connects you and catalyzes collaboration with researchers and knowledge users;
- Sharing enables transparency of your research findings, your impact reports, and permits third-party validation;
- It helps you meet the sharing requirements of an increasing number of funders and journals

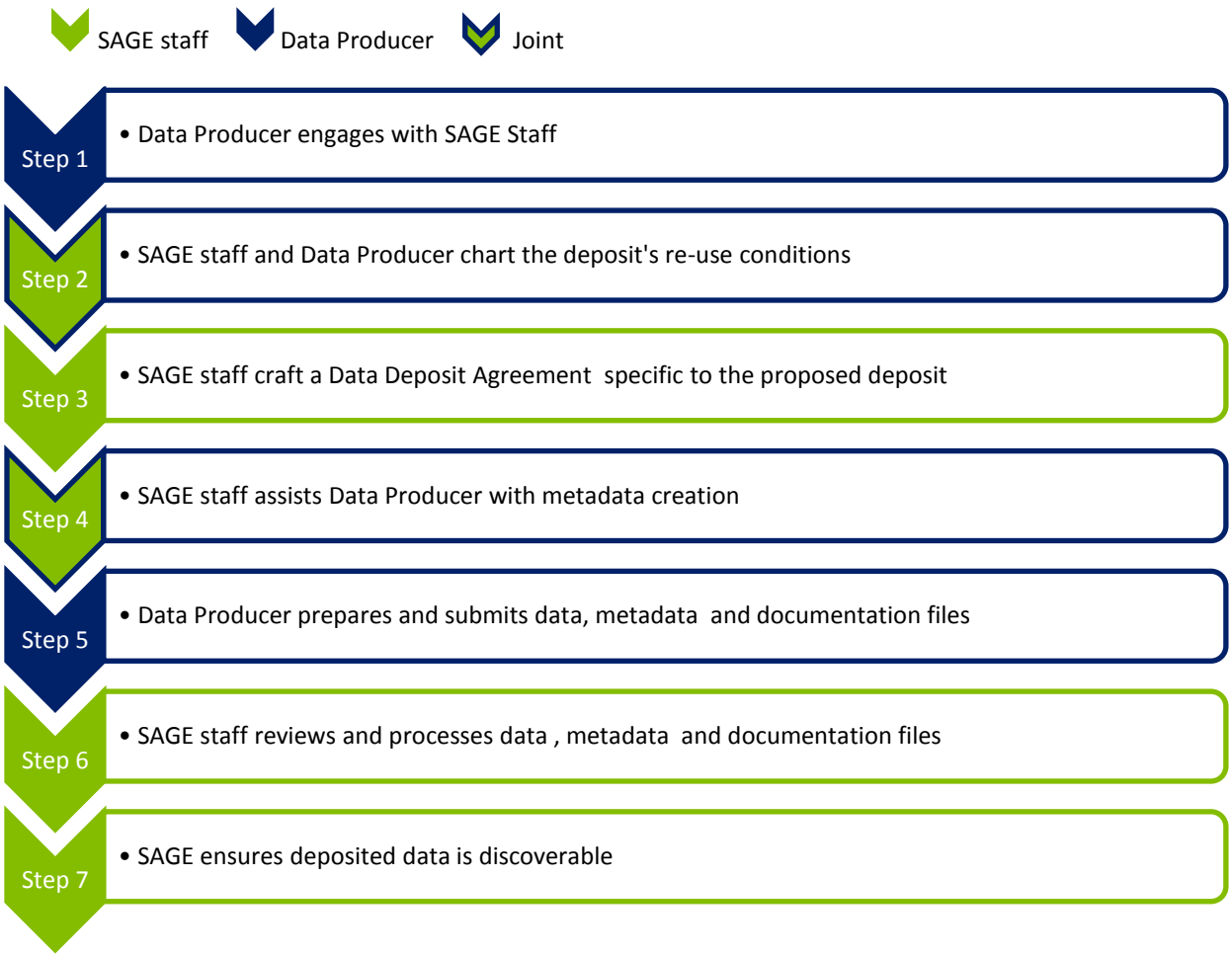
Data that does not require ethical review is, of course, much easier to share. However, a large proportion of data requiring ethical review can still be shared ethically and legally when the data is a) ethically cleared for re-use and b) properly handled, stored and controlled. SAGE is designed to help Data Producers (i.e. researchers or organizations who own or have stewardship of primary data with re-use potential) share data that is typically considered too sensitive or restricted for open access. We:

- Help navigate the ethical and legal hurdles to successfully making data re-usable;
- Provide a secure platform to manage the entire data access process, which:
    - eliminates the administrative burden of sharing data;
    - standardizes the access review process, ensuring re-use is high-quality and appropriate;
    - secures access to data such that risks to participant/client privacy (inherent in sharing) is minimized;
    - permits (where applicable) linkage to administrative and other datasets, greatly enhancing the potential re-use of your data!
- Provide advice and support for appropriate participant/client consent, study design, data management, metadata creation and strategies for de-identification;
- Promote your data for re-use;
- Track the results of your data's re-use

Section II: Data Deposit Process

SAGE works with the Data Producer the entire way to facilitate the deposit of data. In general, the process follows the trajectory in Figure 1, each step of which is further described in this section.

Figure 1 - Procedure Overview



Step 1: Data Producer engages with SAGE Staff

Data Producers should engage with SAGE staff at the earliest opportunity, even as early as during grant writing, study design or at the start of a new data collection strategy (in the case of a service organization). Engagement early in the data or research lifecycle allows Data Producers to receive advice about important data legacy and management issues, such as production of documentation, assessing disclosure risk, and receiving adequate client/participant consent for re-use. Early engagement generally simplifies the data depositing process and helps maximize the utility of your data.

Despite the above, Data Producers are encouraged to engage with SAGE staff at any stage of their data or research lifecycle. SAGE has the expertise and readiness to help retrofit existing data for re-use.

1.1 Eligible content

SAGE welcomes research, service delivery, and administrative data related to social, health, human ecology, environmental, community, learning and similar fields. Datasets should involve data collection on human participants or clients and have re-use potential independent of the actual or original intended use(s) by the original collectors. Data may be qualitative or quantitative.

Currently, SAGE does not accept data collected from non-human subjects, data with legal impediments to sharing (eg. Commercially-owned data), secondary research (eg. Literature reviews), genomic or imaging data, or data without re-use potential. If in doubt as to whether your data meets SAGE's eligibility criteria, please contact us.

Step 2: SAGE staff & Data Producer chart the deposit's re-use conditions

Every organization, research team and/or project has a unique history as well as unique data. The charting of your data's re-use requirements and conditions will vary depending on the nature of your proposed data deposit and data lifecycle stage.

2.1. Depositing for projects or studies at the consent stage or earlier

SAGE will help Data Producers develop grants, ethics applications and consent forms (if and where appropriate) that include language about data deposition and future re-use. SAGE may also, at the request of the Data Producer, assist with study design or data collection considerations to maximize their data's re-use value, data management planning and documentation standards.

2.2. Depositing for projects or studies already past consent stage

SAGE staff will assess existing ethics approval, participant or client consents and data collection protocols (formal or informal) to determine whether data depositing is feasible, and if so under what conditions. In the event that ethics approval and/or client or participant consent does not permit deposit for re-use, we will work with Data Producers to amend approval, waive consent and/or determine the appropriate re-consenting procedure.

Step 3: SAGE staff craft a Data Deposit Agreement specific to the proposed deposit

Prior to data deposit, SAGE and Data Producers must sign a Data Deposit Agreement that outlines the terms and conditions for the proposed data deposit and the conditions governing its secondary use. The Data Deposit Agreement must be signed by the data owner or steward (or consent from the data owner or steward be formally obtained), and all legal or ethical issues that might forbid or unduly limit the sharing of data must be resolved before the agreement is signed.

In general, while most terms and conditions have a standard format reflecting SAGE's standard access rules and procedures, most are quite customizable to account for the Data Producer's wishes or any unique characteristics of their data. Some examples of commonly customized sections include:

3.1 Access conditions

In addition to the access conditions SAGE imposes to protect the anonymity of research participants (for example, restrictions on direct or quasi-identifiers), Data Producers may impose custom access conditions to their proposed data deposit.

3.2 Embargoes

Data Producers can request that an embargo period be imposed on deposited data, whereby no access will be permitted to the data until after a specified date (please note: this date must be an actual date, there is no "to be determined" option). This period allows Data Producers to finish publishing from their data or giving a "cool-down" period prior to allowing access for others. At the end of the embargo period, secondary users may access data as per the access permissions outlined in the Data Deposit Agreement.

Metadata (See section 4) will still be available during the embargo period, to permit discovery and access request planning by Secondary Users.

3.3 Security Level

SAGE assigns a default security level to its data assets which govern how they are accessed. These levels are:

- Level 4: Online transfer.
  For "open access" – data which has been completely de-identified and is not considered sensitive
- Level 3: Secure online transfer.
  For data that has been completely de-identified but which is considered too sensitive for insecure online transfer.
- Level 2: Access using the Virtual Research Environment.
  This is SAGE's default security level, which permits secure, remote access to sensitive data that has been de-identified to the degree considered necessary to protect participant privacy in that particular access request.
- Level 1: Physical enclave only
  For data considered too sensitive even for remote access, Level 1 data is only accessible from special enclave terminals at SAGE's Calgary or Edmonton office.

Data Producers may request a level of security higher (ie. More restrictive) than would be considered appropriate and sufficient by SAGE staff. Data Producers may not, however, impose

a level of security lower (ie. Less restrictive) than recommended by SAGE staff unless transformations to the data deposit are made that alter its sensitivity/re-identification risk.

3.4 Vetting requirements

All outputs (eg. Tables, figures or syntax) from access requests involving Level 1 or Level 2 data must meet, at a minimum, SAGE's vetting requirements checklist. In addition to this checklist, Data Producers may impose other vetting requirements to outputs involving their data deposit.

3.5 Recognition

By default SAGE requires that all Secondary Users cite the data assets they use in all publications, reports and other presentation of results. We also expect, in cases where the Data Producer collaborates in a significant way with the Secondary User, that they receive co-authorship on publications, reports and other presentations of results.

Data Producers may request different recognition from this default. They may, for example, require acknowledgement of the research team and/or of funders in publications or citation of the Data Producer's publications.

Step 4: SAGE staff assists Data Producer with metadata creation

Metadata is distilled documentation regarding the contents and context of data. It dramatically improves data longevity, its discoverability and appropriate re-use potential. SAGE does not accept data for which the metadata is indecipherable or non-existent, as it makes appropriate re-use impossible.

That being said, SAGE staff engage significantly with Data Producers, particularly when approached at an early stage of the research lifecycle, to assist in developing the structure and procedures for recording metadata. If SAGE staff is engaged at a later stage, they will assist Data Producers in discovering, organizing and structuring metadata for their project retroactively.

Step 5: Data Producer prepares and submits data, metadata and documentation files

5.1. Data

The Data Deposit Agreement will specify any special procedures needed to prepare the data prior to deposit. In general, however, Data Producers will prepare data so that it meets the SAGE Data Quality Standards Checklist (available in the *SAGE Data Quality Guide*). When the Data Deposit Agreement is signed prior to data collection, these quality control measures can be implemented throughout the collection process, greatly simplifying preparation prior to deposit.

Some general considerations to contemplate when preparing data for submission:

- Naming Conventions: Use consistent and where possible meaningful names for variables;
- Labels: Provide descriptive variable labels;
- Category labels: Provide labels for categorized variable values;
- Missing data: Treat missing data consistently throughout your data. In particular, differentiate between types of missing data (eg. Don't know vs. Valid skip vs. Missing);
- Derived Variables: Remove dummy, temporary, or redundant variables which have little re-use value for Secondary Users.
- Data correctness: audit your own data for outliers and illogical/impossible values.

5.2. Data Documentation

Data Producers use SAGE online metadata forms to create distilled metadata for their project. In addition to this, Data Producers are invited to submit (where available) the following documentation:

- Ethics approval and blank copy of consent forms, where applicable
- Project summary, proposals, data collection manuals, standard operating procedures
- Blank copies of data collection instruments
- Dataset and Variable list
- Codebook and/or data dictionary, including extra documentation for:
    - Derived variables (what variables are they derived from? What procedure was applied to derive the new variable?);
    - Imputed variables: If missing values were imputed, document the method by which this was accomplished;
    - Items from a standardized or validated scale (what scale is the variable from, it is altered in any way);
    - Direct or indirect identifiers (eg. names, addresses);
- Programming code that is necessary to reproduce any derived variables
- List of publications pertaining to the data, including where possible a list of variables used in the analysis of each.

After preparation, Data Producers will transfer data, metadata and documentation files to SAGE via a secure file transfer protocol.

Step 6: SAGE staff reviews and processes data, metadata and documentation files

SAGE staff will review submitted data, metadata and documentation for consistency, clarity and quality. Where applicable staff will create a de-identified version of the dataset for re-use, keeping the identifiable version for any future linkage to other data permitted in the Data Deposit Agreement. SAGE staff will contact the Data Producer to clarify any ambiguities and resolve any issues with the data and metadata submissions.

After the data has been reviewed, SAGE staff will notify the Data Producer that their data has been accepted. SAGE staff will then organize and enhance the deposited data to optimize data discovery and secondary use. This enhancement includes creation of a metadata record and the assignment of DOI (digital object identifier – a persistent identifier used to uniquely identify your dataset) for the data.

Step 7: SAGE ensures deposited data is discoverable

As part of a comprehensive marketing plan, SAGE will promote the data asset for enhanced discoverability and re-use by secondary users, all the while respecting the conditions listed in the Data Deposit Agreement.

Contact Information

If you have any questions regarding SAGE's Deposit Manual, please contact SAGE staff at data@policywise.com