Best Practices in Data Management, Curation, and Use

This guide provides advice and examples of best practices in the management, curation, and use of data in the national ecosystem assessment process. The use of high-quality data by authors critically ensures the assessment is based on robust evidence, increasing the credibility of key findings. Plans for data management, curation, and use should be considered during the *scoping* stage of the assessment to ensure data accuracy, validation, and availability from the start of the process. It should also highlight any data gaps that can be addressed during the assessment.

Data management

Data management involves planning for all stages of the data lifecycle and is implemented throughout the assessment process. Good data management supports the effective use of data and can increase data access and use by both developers and users of the assessment, increase operational efficiency, reduce the likelihood of mistakes or the loss of data, and help keep data secure. Strong data management strategies for data storage, usage, disposal, and sharing are key for making data use processes secure and efficient.

A **data management plan** outlining agreed processes can be developed in the *scoping* stage of the assessment before data and resources are gathered or analysed. This can ensure consistency across an organization and reduce data related costs by increasing efficiency.

Consider your budget: Data management will have costs associated with it in terms of data storage, data security, and staff time. These costs should be considered at the beginning of the national ecosystem assessment process.

Important considerations for a data management plan:

Strong file naming and cataloguing conventions can be defined and implemented across the assessment team. This will allow authors and other team members to easily find, use, and update data. Ideally, both the naming and catalogue conventions are user friendly and future friendly. Tips for file naming best practices include:

- A folder hierarchy that fits the structure of the assessment.
- Use short but descriptive file names consistently.
- Use a consistent format to include dates in file names. YYYYMMDD is suggested.
- Include a version number.
- Use CamelCase and underscores instead of spaces.

Data management is everyone's responsibility. The technical support unit is usually responsible for the organization and storage of data used in the assessment process. However, all authors are responsible for sourcing data and metadata, and adhering to the data management plan. Data management roles and responsibilities could be clearly assigned to specific individuals and teams during the planning process of the assessment.

Record and update the metadata throughout the data lifecycle. Metadata is information about the data that is crucial for maintaining data quality. It includes information about data content, structure, and permission and helps ensure the data will be properly used and re-used. Ideally, the data management plans encourage the assessment team to adhere to FAIR principles[1] (Findable, Accessible, Interoperable, Reusable) for handling metadata. Agreement on metadata standards can also contribute to consistent use within the assessment team.

A data storage plan is necessary to ensure the security, accessibility, and availability of data both during and after the assessment process. The plan should address storage, backup, sharing, and preservation – all interconnected aspects of data storage.

Data storage: Data storage may include digital or non-digital data. The team could determine requirements ahead of time for what information is retained and what will be destroyed. Storage options include local storage (laptops, network drives, optical storage, and external storage devices) and cloud storage. Cloud storage, such as Dropbox, Sharepoint and Onedrive, has grown in popularity because it can reduce capital investments, physical storage space, and provide multiple levels of data security. Privacy issues, ease of accessibility based on location, and cost should be assessed when choosing storage options.

Best practices in digital data storage[2]:

- Use high-quality storage systems
- Use non-proprietary formats
- Migrate data files every two to five years to new storage
- Check stored data regularly to ensure nothing has been lost
- Use different forms of storage for the same data
- Label and organize stored files logically
- Consider encryption

Backup: Best practice highlights the need for data to be backed up both during and after (see "Preservation" below) the assessment processes. Data can be lost due to hardware failure, software issues, virus infection, power failure, human error, theft or loss, and hardware damage[3]. An ideal backup strategy could include protection against all these risks, but emphasis could be put into protecting against the most likely risks in the specific context. The backup plan could include what data gets backed up, how frequently this is carried out, where data gets backed up, who is responsible for performing backups, and whether full or incremental backups are needed. A recommended practice for backing up and storing data is the 3-2-1 Rule, which recommends having *three* copies of data, on *two* different pieces of storage media, *one* of which should be stored off-site[4]. Testing the backup system upon initial setup at and at regular intervals thereafter helps to ensure that backed up files can be retrieved and read.

Sharing: Data sharing needs are an important consideration when developing a data storage plan for a national ecosystem assessment. A well-though out data storage plan can simplify data sharing and collaboration with team members for current and future use. Many cloud storage options also serve as file sharing services, allowing users to share and collaborate on data, documents, and other files. It is important to determine who will need to access data and where they will need to access it from when devising the data management plan as this can be important for selecting how to best store and share data. A critical step is the development of data sharing agreements and potential restrictions on data usage, particularly before the publication has been shared publicly. If data is to be shared before the publication of the assessment, non-disclosures and permission

[1] FAIR Principles https://www.go-fair.org/fair-principles/

[2] British Ecological Society. 2018. Data Management. British Ecological Society. Available at: https://www.britishecologicalsociety.org/wp-content/uploads/2019/06/BES-Guide-Data-Management-2019.pdf?utm_source=web&utm_medium=web&utm_campaign=better_science

[3] British Ecological Society. 2018. Data Management. British Ecological Society. Available at: https://www.britishecologicalsociety.org/wp-content/uploads/2019/06/BES-Guide-Data-Management-2019.pdf?utm_source=web&utm_medium=web&utm_campaign=better_science

[4] Preston, W.C. 2020. For secure data backup, here's how to do the 3-2-1 rule right. Available at: https://www.networkworld.com/article/3527303/for-secure-data-backup-here-s-how-to-do-the-3-2-1-rule-right.html

documents will need to be signed. This is key to ensuring the protection of the data and its sources for the assessment.

Preservation: A data preservation plan is essential for ensuring data can be used beyond the assessment process. It involves the archiving of data to ensure future viability and accessibility. A plan for how and where data will be archived could be made during the planning process and be included in the data management plan. Archived data could include data files with descriptive file names in a well-organized file structure, robust metadata, and supplementary documentation and files that will assist future users use and understand the data[5]. Alternatively, data can be deposited in an archive or repository upon completion of the assessment to preserve it and facilitate its use for future decision-making processes, research or other activities.

Data security and protection is essential for ensuring the privacy, availability, and integrity of data. A robust security infrastructure could be planned and implemented to reduce the risk of data breaches. For example, antivirus software and antispyware tools could be installed on all workstations. Also, compliance with privacy regulations could be essential and it will need to be considered from the scoping stage of the assessment.

Identifying data sources, acquiring and curating data

Identify data needs

Before gathering data, it is important to consider the key policy questions that will be the focused on the existing need for information on the status and trends of biodiversity and ecosystem services, which accounts for their contributions to human well-being in decision-making across sectors. In order to identify the need for an assessment, it is important to understand the current national environmental, social and economic context, for example through the conceptual framework and related challenges faced by decision-makers. Identifying these gaps in the scoping stage helps guide authors on the specific data needs of the assessment.

Data quality and validation

Before gathering data, it is also important to consider the acceptable level of data quality. The use of high-quality data ensures the assessment is based on robust evidence, increasing the credibility of the key findings. Data quality metrics could be used to help decide if a dataset is of sufficient quality. Key data quality metrics to consider include the terms of use of the data, the accessibility of the data, frequency of update, the geographic coverage of the data, the scale of the data, the inclusion of data collection methodology and metadata, and the authenticity of the data provider.

Essential tips for gathering and curating data

- Keep a record of all data sources, including the contact information of individual data providers and websites where publicly available datasets were found. Keep in mind that data from data providers may need formal data sharing agreements.
- Discuss your intended use of the data with the data providers. This will clarify the rights you have to use the data.
- Use unique identifiers for each data point within a dataset. This will help track data over time, even if other descriptive information about the data point changes.
- Use versioning. If a dataset gets updated or changed over time, each new version could be saved as a new file with the same naming convention and new timestamp.
- A consistent coordinate system is recommended when dealing with spatial data.
- Consistent field headings and consistent values could be used when datasets are updated.

Using data effectively

[5] Axiom Data Science. 2017. Metadata and Documentation. Available at: https://www.axiomdatascience.com/best-practices/MetadataandDocumentation.html#metadata-metadata

All aspects of data management lead to data use and reuse. Once data sources have been identified, it's important to consider how to effectively use data in the analyses.

Author tips for using data effectively

- Pace yourself. It can be useful to follow the 70-20-10 rule of working with data. 70% of your time should be spent finding and cataloguing data, 20% could be spent cleaning and prepping data, and 10% should be spent on the actual analyses.
- Organization is crucial and having a plan for every step of the process could help to remain focused. Tips include a naming system for outputs and structuring outputs. You can also take detailed notes of every step of the analyses.
- Make your work re-traceable. This will be helpful for future use of the information and data collected as part of the assessment process. Github is a powerful and useful platform for sharing analysis code with internal and/or external partners
- It is essential that the terms of use of datasets are followed, and this information is recorded in metadata.
- Data use agreements such as intellectual property rights, permission information, and licenses, should be read and respected. Referencing software is recommended for citing sources.