Open Data Management in Agriculture and Nutrition

This e-learning course is the result of a collaboration between ***GODAN Action*** *partners, including* ***Wageningen Environmental Research*** *(WUR),* ***AgroKnow****,* ***AidData****,* ***the Food and Agriculture Organization of the United Nations*** *(FAO),* ***the Global Forum on Agricultural Research*** *(GFAR), and* ***the Institute of Development Studies*** *(IDS),* ***the Land Portal****,* ***the Open Data Institute*** *(ODI) and* ***the Technical Centre for Agriculture and Rural Cooperation*** *(CTA).*

GODAN ACTION
Global Open Data for Agriculture & Nutrition

UKaid
from the British people

GODAN Action is a three-year project UK's Department for International Development to enable data users, producers and intermediaries to engage effectively with open data and maximise its potential for impact in the agriculture and nutrition sectors. In particular we work to strengthen capacity, to promote common standards and best practice and to improve how we measure impact. [www.godan.info]

UNIT 3: MAKING DATA OPEN

LESSON 3.1: MANAGING DATASETS

Photo by Neil Palmer (CIAT) licensed under CC BY-SA 2.0

Aims and learning outcomes

This lesson aims to;

- explain basic principles behind data management
- introduce the process of preparing a data management plan
- explain concepts on data storage, versioning and documentation practices.

After studying this lesson, you should be able to:

- identify the steps to ensure that good scientific data management practices are followed
- prepare a data management plan
- understand the potential of using data storage and versioning good practices.

Contents

List of tables

1. Data management

Many datasets are inherently ephemeral: market data, production and consumption data, and weather data are good examples. These data are meaningless unless we know the timeframe for them and they are updated regularly. Soil data, for example, are similar, even if not updated as often as weather and market data. They change by location and during the course of – and between – seasons.

In order to build and maintain trust in open data such as these and others, it is necessary to have stable data management principles and practices in place. Good **data management principles** help to ensure that data produced or used are registered, stored, made accessible for use and reuse, managed over time and/or disposed of, according to legal, ethical, funder requirements and good practice. For open data consumers, trust depends on numerous factors:

- *Knowing the source*. Trust in data begins with knowledge of its source.
- *Trusting the source*. If you know that data comes from a trusted source, then you can rely on it, and on the conclusions you draw from it.
- *Timeliness of the data*. Even when from a trusted source, data is not useful if it is outdated.
- *Data quality*. Trusted data must accurately and precisely reflect what it measures.
- *Sustainability*. A trusted dataset must have some guarantee of availability.
- *Discoverability*. Like documents, data is only useful if it is straightforward to find. (More on discoverability in Unit 1, Lesson 2.1)
- *Documentation and support*. Consumers should be able to access support for data if needed.
- *Interaction*. Consumers should be able to provide feedback if there is a problem with data.

Data management therefore, is a process involving a broad range of activities from administrative to technical aspects of handling data[1] in a manner that addresses the factors listed above. A sound data management policy will define strategic long-term goals for data management across all aspects of a project or enterprise.

A data management policy is a set of high-level principles that establish a guiding framework for data management. A data management policy can be used to address strategic issues such as data access, relevant legal matters, data stewardship issues and custodial duties, data acquisition, and other issues. As it provides a high-level framework, the data management policy should be flexible and dynamic. This allows for it to readily adapt to unanticipated challenges, different types of projects and potentially

[1] A. Gordon (ed.) 2015. *Official (ISC)² Guide to the CISSP CBK* (4th edn) CRC Press, Boca Raton, FL, USA

opportunistic partnerships while still maintaining its guiding strategic focus. The data management policy will help inform and develop a **data management plan**, which will be discussed more in this lesson.

In order to meet data management goals and standards, all involved parties must understand their associated **roles and responsibilities**. The objectives of delineating data management roles and responsibilities are to:

- clearly define roles associated with functions
- establish data ownership throughout all phases of a project
- instil data accountability
- ensure that adequate, agreed-upon data quality and metadata metrics are maintained on a continuous basis.

Quality as applied to data has been defined as fitness for use or potential use. Many data quality principles apply when dealing with species data and with the spatial aspects of those data. These principles are involved at all stages of the data management process, beginning with data collection and capture. A loss of data quality at any one of these stages reduces the applicability and uses to which the data can be adequately put.

All of these affect the final quality or fitness for use of the data and apply to all aspects of the data. Data quality standards may be available for:

- accuracy
- precision
- resolution
- reliability
- repeatability
- reproducibility
- currency
- relevance
- ability to audit
- completeness
- timeliness.

Data quality is assessed by applying verification and validation procedures as part of the quality control process. Verification and validation are important components of data management that help ensure data is valid and reliable. This is fully elaborated in Unit 2, Lesson 2.2.

Some of the data management principles described in this unit also borrow from principles of research data management (RDM). Good RDM principles can be applied to open data initiatives with success, thereby ensuring:

- enhanced visibility and discovery of the datasets
- facilitation of longevity and the sharing and reuse of the datasets
- data users' understanding of the content context and the limitations of datasets
- reduction of the risk of data loss by keeping it safe and secure

- interoperability of datasets and data exchange
- compliance with funders' and/or institutional expectations and policies
- provision of opportunities for collaboration with others who might engage with the data.

1.1. The diversity of data

Data may be viewed as the lowest level of abstraction from which information and knowledge are derived. Data may also be viewed as the level at which measurements were originally collected, e.g. individual responses to a survey or census; hourly measures of temperature, wind speed and wind direction; number and price of shares traded in each stock buy/sell transaction; etc.

However the word data means different things to different people in different contexts. Different disciplines have and use discipline-specific language around the subject research data[2].

Table 1 Various types of data as encountered in different contexts and disciplines

Types of Data		
General	Social sciences	Physical/agricultural data
• images • video • mapping/GIS data • numerical measurements	• survey responses • focus group and individual interview transcripts • economic indicators • demographics • opinion polling	• measurements generated by sensors/laboratory instruments • computer modelling • simulations • observations and/or field studies • specimen

Data can therefore be generated for different purposes and through different processes in a multitude of digital formats. The following classification was compiled by the Research Information Network:

- **Observational:** data captured in real time, usually unique and irreplaceable, e.g. brain images, survey data
- **Experimental:** data from experimental results, e.g. from lab equipment, often reproducible, but can be expensive, e.g. chromatograms, microassays
- **Simulation:** data generated from test models where model and metadata may be more important than output data from the model, e.g. economic or climate models

[2] http://mantra.edina.ac.uk

- **Derived or compiled:** resulting from processing or combining 'raw' data, often reproducible but expensive, e.g. compiled databases, text mining, aggregate census data
- **Reference or canonical:** a (static or organic) conglomeration or collection of smaller (peer-reviewed) datasets, most probably published and curated, e.g. gene databanks, crystallographic databases.

In the next section we will describe the basic data management considerations to be made during the lifecycle of data i.e. from creation and initial storage to the time when it becomes obsolete and is deleted.

1.2. Metadata and documentation

All datasets should be identified and documented to facilitate their subsequent identification, proper management and effective use, and to avoid collecting the same data more than once. To provide an accurate list of datasets held by an organisation, a catalogue of data should be compiled. This is a collection of discovery-level metadata for each dataset, in a form suitable for users to reference. These metadata should provide information about the content, geographic extent, currency and accessibility of the data, together with contact details for further information about the data.

All datasets, once catalogued, should also be documented in a detailed form suitable for users to reference when using the data. These detailed metadata should describe the content, characteristics and use of the dataset, using a standard detailed metadata template.

Metadata

Metadata, or 'data about data' explains your dataset and allows you to document important information for:

- finding the data later
- understanding what the data is later
- sharing the data (both with collaborators and future secondary data users).

It should be considered an investment of time that will save you trouble later several-fold.

Examples

- Dublin Core
- Darwin Core
- FGDC (Federal Geographic Data Committee)
- DDI (Data Documentation Initiative)
- ABCD (Access to Biological Collections Data)
- CSDGM (Content Standard for Digital Geospatial Metadata).

A distinction that is often cited when dealing with data management is that of data vs. metadata (i.e. data about data). There are a number of specific distinctions that these might refer to:

- *Metadata as schema*. When we collect tabular data, we need to know what the 'columns' in the data refer to. Even in non-tabular data, some schema information is helpful for interpreting the data. Many data formats include ways to specify schema metadata, e.g. XSD for XML, RDFS for RDF, and DDL for databases.
- *Bibliographic metadata*. Librarians and library scientists have used metadata to describe documents (books, articles, pictures, etc) for centuries, and have determined structures for recording and searching this kind of metadata. This sort of metadata includes provenance information (authorship, publication data), dates, size of the publication (e.g. in page count), and is applicable to datasets as well (e.g. DCAT27 and VoID28).
- *Shared vocabulary*. Alignment of different datasets is a challenge in any distributed setting. A key tool in governing such datasets is the use of a shared vocabulary. The vocabulary is used in the content of the data, rather than describing the data per se. For example, the AGROVOC (for AGRiculture VOCabulary)[3] provides (amongst other things) terminology for talking about agricultural products, e.g. milk, milk byproducts, milk fat, etc. Agroportal[4] and the VEST registry[5] provide access to many vocabularies related to agriculture.

File Contents

In order for others to use the data, they must understand the contents of the dataset, including the parameter names, units of measure, formats, and definitions of coded values. At the top of the file, include several header rows containing descriptors that link the data file to the dataset (for example, the data file name, dataset title, author, today's date, the date the data within the file was last modified, and companion file names). Other header rows should describe the content of each column, including one row for parameter names and one for parameter units. For those datasets that are large and complex and may require a lot of descriptive information about dataset contents, that information may be provided in a separate linked document rather than as headers in the data file itself.

- *Parameters.* The parameters reported in datasets need to have names that describe their contents, and their units need to be defined so that others understand what is being reported. Use commonly accepted

[3] FAO. 2016. *AGROVOC Multilingual agricultural thesaurus*. Available at: http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus.
[4] IBC Agroportal. Available at: http://agroportal.lirmm.fr/
[5] VEST Directory, Agricultural Information Management Standards (AIMS). http://aims.fao.org/vest-registry

parameter names. A good name is short, unique (at least within a given dataset), and descriptive of the parameter contents. Column headings should be constructed for easy importing by various data systems. Use consistent capitalisation and use only letters, numerals, and underscores – no spaces or decimal characters – in the parameter name. Choose a consistent format for each parameter and use that format throughout the dataset. When possible, try to use standardised formats, such as those used for dates, times, and spatial coordinates.

All cells within each column should contain only one type of information (e.g., text, numbers, etc.). Common data types include text, numeric, date/time, Boolean (Yes/No or True/False), and comments, used for storing large quantities of text.

- *Coded Fields*. Coded fields, as opposed to free text fields, often have standardised lists of predefined values from which the data provider may choose. Data collectors may establish their own coded fields with defined values to be consistently used across several data files. Coded fields are more efficient for the storage and retrieval of data than free text fields.

- *Missing Values. There are several options for dealing with a missing value. One is to leave the value blank, but this poses a problem as some software does not differentiate a blank from a zero, or a user might wonder if the data provider accidentally skipped a column. Another option is to put a period where the number would go. This makes it clear that a value should be there, although it says nothing about why the data is missing. One more option is to use different codes to indicate different reasons why the data is missing.*

1.3. Security and storage

Effective data sharing depends on a strong network of trust between data providers and consumers. Infrastructure for data sharing will not be used if the parties who provide and use the data do not trust the infrastructure or one another. If sensitive data is to be shared, there must be provisions in the platform to ensure security of that data. Whether data is closed or shared with specific individuals or organisations, it will need to be hosted in a controlled way. Depending on the sensitivity of the data, this will include some guarantee of security, e.g. against hacking. In the most extreme cases, the security requirements for shared data in agriculture could be as severe as for shared data in the military. These principles are not unique to agricultural data, and have been studied in depth.

The basic concepts behind these principles are that services should be hard to compromise, that a compromise should be easy to detect, and that the impact of a compromise can be contained. For open data, this is much less of

a concern, but to build trust among data providers, some support for data security must be in place.

Some important physical dataset storage and archiving considerations for electronic/digital data include:

- *Server hardware and software*. What type of database will be needed for the data? Will any physical system infrastructure need to be set up or is the infrastructure already in place? Will a major database product be necessary? (See Lesson 3.3 on Creating Open Data Repositories for software platforms.) Who will oversee the administration of this system?

- *Network infrastructure*. For open data the database needs to be connected to the internet for accessibility. How much bandwidth is required to serve the target audience? What hours of the day does it need to be accessible?

- *Size and format of datasets*. The size of a dataset should be estimated so that storage space can properly be accounted for. The types and formats should be identified so that no surprises in the form of database capabilities and compatibility will arise.

- *Database maintenance and updating*. A database or dataset should have carefully defined procedures for updating. If a dataset is live or ongoing, this will include such things as additions, modifications, and deletions, as well as frequency of updates. Versioning will be extremely important when working in a multi-user environment.

- *Database backup and recovery requirements*. The requirements for the backing up or recovery of a database in case of user error, software/media failure, or disaster, should be clearly defined and agreed upon to ensure the longevity of a dataset. Mechanisms, schedules, frequency and types of backups, and appropriate recovery plans should be specified and planned. This can include types of storage media for onsite backups and whether off-site backing up is necessary.

1.4. Data access and dissemination

Data production is one thing, its dissemination is another. Open data is useful when it can be delivered into the right hands (or the right machine) and within a context where it can be most valuable. In some cases, this might be a laboratory researching the efficacy of a treatment (for example, the effectiveness of herbicide for treating some pest). Sometimes the data must be delivered into the field, so it can be used to help a smallholder make informed decisions on which crop varieties to grow or which treatments to apply. There must be a variety of data delivery channels, fine-tuned to each case for data delivery. The 'fine-tuning' of data delivery channels can become

a business opportunity for data intermediaries in the case where the data is fully 'open'. An intermediary can provide services to customise data delivery for the vast range of customers that might exist for the data. Open data creates the possibility of a marketplace, where alternative sources of relevant data are available.

To be made available, data has to be stored in a way that makes it accessible. Even in the modern era of cloud deployment, the data and applications are stored on some hardware somewhere, even if it is virtualised. A strategy for sharing data on a global scale must specify where it will be stored and what service level agreements (SLAs) will be maintained (up time, throughput, access controls, etc).

The following should strongly be considered when deciding how to disseminate data:

- access to the data should be provided in line with the organisation's data policy and the national laws/acts on access to information
- access to data should be granted without infringing the copyright or intellectual property rights of the data or any statutory/departmental obligations.

2. Data management plans

Funders are increasingly demanding the development of data management plans as a condition for funding.[6] The requirements are usually to allow for the sharing of outputs of projects or research, including data (and publications). For example, the EU has published new guidelines on data management in Horizon 2020 research projects as of December 2013: ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

The planning process for data management begins with a data planning checklist. A checklist later assists in the development of a data management plan. That checklist might include considering some or all of the following questions.

- What data will you collect or create, how will it be created, and for what purpose?
- How will you manage any ethical issues? How will you manage copyright and intellectual property rights issues?
- What file formats will be used? Are they non-proprietary, transparent and sustainable? What directory and file naming conventions will be used? Are there any formal standards that you will adopt? What documentation and metadata will accompany the data?

[6] https://ocw.mit.edu/resources/res-str-002-data-management-spring-2016/workshop-materials/MITRES_STR002S16_IntroDM.pdf

- How will the data be stored and backed up? How will you manage access and security? Who will be responsible for data management?
- Are there existing procedures that you will base your approach on? For example, are there institutional data protection or security policies to follow, department or group data management guidelines, defined by your institution or funder that must be considered?
- What is the long-term preservation plan for the dataset? For example, which data should be retained, shared, and/or preserved? How will you share the data, and are any restrictions on data sharing required?
- What resources will you require to deliver your plan? For example, are there tools or software needed to create, process or visualise the data?

More can be added to or subtracted from this suggested list depending on the nature of the project. A more detailed checklist is available from the Digital Curation Centre at www.dcc.ac.uk/sites/default/files/documents/resource/DMP_Checklist_2013.pdf

Important to note: Data management plans must be continuously maintained and kept up-to-date throughout the course of a project or research.

Table 2 Components of a data management plan

Administrative information	• Name and ID of the project • Project Description • Funding body/bodies • Project Data Contact • Related Policies • Date of First Version • Date of Last Update
Data collection	• Data description, including anticipated type, format and volume • Existing datasets to be re-used • Methods by which data will be collected or created • Structures, naming and versioning system for folders and files • Quality assurance processes

Documentation and metadata	● A list of information you expect will be needed for the data to be read and interpreted in the future ● How you plan to collect or create this documentation and metadata ● The metadata standards you will use **Some examples of data documentation**: ● Laboratory notebooks & experimental protocols ● Questionnaires, codebooks, data dictionaries ● Software syntax and output files ● Information about equipment settings & instrument calibration ● Database schema ● Methodology reports ● Provenance information about sources of derived data Add detailed descriptions for collections or files, e.g. what is in a file, where did it come from, how could it be retrieved if needed, any existing problems etc.
Ethics and legal compliance	**Ethics** ● Details of consent needed for data preservation and sharing ● Steps to be taken, if needed, to protect the identity of any participants ● Steps to be taken, if needed, to ensure sensitive data is stored and transferred securely **Copyright and Intellectual Property Rights** ● Name(s) of the owner(s) of the data ● Licence(s) for re-use which will be applied (e.g. one of the licences available from Creative Commons or Open Data Commons) ● Restrictions on third party use ● Any expected delay to data sharing e.g. pending a patent application or embargo related to publication in a journal

Storage and backup	• Where (physically) data will be stored • Backup provision • Person or team responsible for backup • Recovery procedures **Security** • Risks, and how they will be managed • Access arrangements • Any arrangements, if needed, for safe and secure transfer of data collected in the field
Selection and preservation	• Details of which data should be retained, shared and/or preserved, with particular reference to contractual, legal or regulatory requirements • Foreseeable research uses for the data • Length of time for which data will (or should) be kept beyond the life of the project • The repository or archive where the data will be held, and any associated charges • Time and effort needed to prepare data for preservation / data sharing
Data sharing responsibilities and resources	• Named person responsible for implementation of the Data Management Plan • Named person responsible for each data management activity • Hardware and software required (any that is additional to existing institutional provision) • Additional specialist expertise or training required • Charges to be applied by data repositories

3. Data organisation

Data files and folders need to be labelled and organised in a systematic way so that they are both identifiable and accessible for current and future users. The benefits of consistent data file labelling are:

- Data files are distinguishable from each other within their containing folder
- Data file naming prevents confusion when multiple people are working on shared files
- Data files are easier to locate and browse
- Data files can be retrieved not only by the creator but by other users
- Data files can be sorted in logical sequence
- Data files are not accidentally overwritten or deleted
- Different versions of data files can be identified

- If data files are moved to another storage platform their names will retain useful context.

There are three main criteria to consider regarding the naming and labelling data files, namely:

- *Organisation:* important for future access and retrieval, and needs to take into account the file naming constraints of the system where the file is located
- *Context*: this could include content specific or descriptive information, independent of where the data are stored
- *Consistency:* choose a naming convention and ensure that the rules are followed systematically by always including the same information (such as date and time) in the same order (e.g. YYYYMMDD).

Common elements of a file naming strategy

- Version number
- Date of creation
- Name of creator
- Description of content
- Name of team/department/unit associated with the data
- Publication date
- Project number.

It is important to also consider the following when naming files:

- The use of generic file names that may conflict when moved from one location to another. Ensure filenames are independent of location.
- File names should outlast the file creator who originally named the file.
- How scalable the file naming policy needs to be, e.g. if the project number is limited to two digits, you can only have ninety nine projects.

Version control

It is important to identify and distinguish versions of datasets consistently. This ensures that a clear audit trail exists for tracking the development of a dataset and identifying earlier versions when needed. Thus you will need to establish a method that makes sense to you that will indicate the version of your dataset.

- A common form for expressing data file versions is to use ordinal numbers (1, 2, 3, etc.) for major version changes and decimals for minor changes (e.g., v1, v1.1, v2.6)
- Confusing labels should be avoided e.g. revision, final, final2, definitive_copy, as you may find that these accumulate
- Record ALL changes (minor and major)
- Discard or delete obsolete versions (whilst retaining the original 'raw' copy)
- Use an auto-backup facility (if available) rather than saving or archiving multiple versions
- Turn on versioning or tracking in collaborative documents or storage utilities such as Wikis, GoogleDocs etc

- Consider using version control software e.g. Subversion, TortoiseSVN.

Some structured examples of maintaining version control [document name] [version number] [status: draft/final]:

- Jones_interview_July2010_V1_DRAFT
- Lipid-analysis-rate-V2_definitive
- 2001_01_28_ILB_CS3_V6_AB_edited

Summary

Good data management principles help to ensure that data produced or used are registered, stored, made accessible for use and reuse (if appropriate), managed over time and/or disposed of, according to legal, ethical, funder requirements and good practice.

Data management therefore is a process involving a broad range of activities from administrative to technical aspects of handling data in a manner that addresses the factors listed above. A sound data management policy will define strategic long-term goals for data management across all aspects of a project or enterprise.

A data management policy is a set of high-level principles that establish a guiding framework for data management. A data management policy can be used to address strategic issues such as data access, relevant legal matters, data stewardship issues and custodial duties, data acquisition, and other issues. Data management plans must be continuously maintained and kept up-to-date throughout the course of a project or research.

Components of a data management plan will include:

- administrative information
- data collection methods and quality assurance processes
- documentation and metadata
- ethics and legal compliance
- storage and backup
- selection and preservation
- data sharing responsibilities and resources.

Further Readings

- Arms, C. R., Fleischhauer, C. and Murray, K. (2013). Sustainability of digital formats: planning for Library of Congress collections. Library of Congress, Washington DC, USA. Available at www.digitalpreservation.gov/formats
- Beagrie, N. and Houghton, J. (2014). *The value and impact of data sharing and curation - synthesis of three recent UK studies*. Jisc. Available at: repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf
- Charles Beagrie Ltd (2013). *Keeping research data safe: cost / benefit studies, tools, and methodologies focussing on long-lived data*. Available at: http://www.beagrie.com/krds.php (accessed 4 August 2014)
- Digital Curation Centre (DCC). (2010). *Data management plans*. Available at: http://www.dcc.ac.uk/resources/data-management-plans (accessed 4 August 2014)
- Drummond, C.G. (2009). Replicability is not reproducibility: nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop*, 26th ICML, Montreal, Canada. Available at: http://cogprints.org/7691
- European Commission (EC). (2017). *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*. Available at: ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (Version 3.2, 21 March 2017)
- GODAN 2016 *A Global Data Ecosystem for Agriculture and Food*. GODAN, Wallingford, UK. Available at: http://www.godan.info/documents/data-ecosystem-agriculture-and-food
- Jones, S. (2011). *How to Develop a Data Management and Sharing Plan*. DCC How-to Guides, Digital Curation Centre, Edinburgh, UK. Available at: www.dcc.ac.uk/resources/how-guides/develop-data-plan
- UK Data Archive (UKDA). *Plan to Share*. Available at: www.data-archive.ac.uk/create-manage/planning-for-sharing (accessed 4 August 2014)