# 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 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 3.3: CREATING AND MANAGING OPEN DATA REPOSITORIES

Photo by German Climate Computing Center licensed under CC BY NC-ND 2.0

## Aims and learning outcomes

This lesson aims to;

- provide an overview of how to create an open data repository and explain the requirements for managing a repository
- describe the importance and the different components of data curation

After studying this lesson, you should be able to;

- identify the steps to consider when creating an open data repository
- select the appropriate software tool for setting up a repository according to her requirements
- understand the requirements to efficiently manage an open data repository
- understand the different components in data curation

# Contents

# List of figures

# List of tables

# 1. Introduction

In Lesson 3.2, Managing Dynamic Datasets, we looked at the various options for publishing dynamic data and we described automated processes to facilitate this. In this lesson we explore the various options available for hosting an open data repository. These range from commercially hosted to free and open source solutions.

# 2. Creating an open data repository

Generally, a data repository serves as the centrepiece of an open data effort and serves as a central location to find data, a venue for standardising practices, and a showpiece of the use of that data. In a practical sense, a repository serves as a central, searchable place for people to find data. Some repository software will automatically convert data from one format to others, so even though you can only provide data in one format (e.g., CSV), it will generate XML, JSON, Excel, etc.

Some software will visualise datasets in the browser, letting people map, sort, search, and combine datasets, without requiring any knowledge of how to program.

Some repository software allows syndication, permitting other organisations to automatically incorporate your own data (e.g., a state transportation agency could gather up all localities' transportation data and republish it).

Generally, repository software supports either uploading files to be stored in the repository, or pointing the repository at an existing website address where the file lives. The former works well for smaller organisations (localities, small agencies), while the latter works well for larger organisations or governments, for which centralising assets can be impractical.

To make a broad division, there are three types of hosting available: commercial hosting, self-hosting, and free hosting. Below in Table 1 is a list of available data repositories.

*Table 1 Overview of data repositories*

| Name | Type | Notes |
|---|---|---|
| ArcGIS Open Data | hosted | |
| CKAN | open source | |
| DataHub | free, hosted | CKAN-powered |
| DKAN | open source | |

| | | |
|---|---|---|
| GitHub | free, hosted | |
| JKAN | open source | |
| Junar | hosted | |
| NuData | hosted | |
| OpenData.city | free & hosted | CKAN-powered |
| OpenDataSoft | hosted | |
| Open Data Catalog | open source | |
| Socrata | hosted | |

## 2.1. Establishing user needs and requirements

Studies that investigated the causes of IT project failure have found that 'requirements definition' is the most frequently cited project stage that caused failure. Therefore understanding what the repository should do and making sure it meets the stakeholders' needs is critical to the success of the project.

***Identifying needs:***

- Understand as much as possible about the users, as well as their work and the context of their work.
- System under development should support users in achieving their goals.
- Identifying needs is crucial to the next step.

***Establishing requirements:***

- Building upon the needs identified, produce a set of requirements.
- Use a user-centered approach to development:
- Understanding what the product should do and making sure it meets the stakeholders' needs are absolutely critical to the success of the product.

***What are requirements?***

A requirement is a statement that specifies what an intended product should do, or how it should perform. Traditionally, there are two types of requirements:

- *functional* requirements which specify what the system should do
- *non-functional* requirements which specify what constraints there are on the system or its development.

Table 2 gives an expanded list of categories of requirements.

*Table 2 Overview of requirements*

| Categories of requirements | Description |
|---|---|
| Functional requirements | What the product should do. |
| Data requirements | The type, volatility, size/amount, persistence, accuracy and value of the amounts of the required data. |
| Environmental requirements | Or 'context of use' - circumstances in which the interactive product must operate |
| User requirements | Characteristics of the intended user group. |
| Usability requirements | The usability goals and associated measures. |

Based on the needs and services of the repository, institutions will then want to assess the available software platforms. There are three types of options available:

- **Open Source Software:** The software is free to download, but usually requires some level of expertise to implement and maintain. A central governing body manages the source code, but it is open for changes and enhancements from the development community (for example, CKAN, DKAN, JKAN).

- **Commercial Software:** You typically pay for the software and, optionally, any additional subscription or consulting fees. You own the use of the software and, with a subscription, get software upgrades. With a programming interface, or API, you can customise the software, but the software vendor owns, creates and maintains the source code.

- **Software Service Model:** A software vendor owns and distributes a software platform, or also hosts and manages your data for you. In this model, the software vendor provides additional services for a fee, and also controls and updates the software source code (examples are EPrints Services, Open Repository and bepress).

Implementers will want to choose the software that best matches their needs and available resources (budget and staffing). For example, institutions without significant technical expertise may want to look at some of the commercial services available. In terms of open source software platforms, each has its own unique strengths.

## 2.2. Selecting repository software

### *2.2.1. Open source software*

There are some excellent open source data repository programs that are solid options for technically savvy organisations, for organisations with a commitment to use the open source software, or for organisations with the budget to hire a consultant to deploy the software.

**CKAN:** CKAN[1] is nominally an initialism for 'Comprehensive Knowledge Archive Network,' but it is only ever referred to as CKAN. A creation of UK-based Open Knowledge[2], CKAN is the most commonly used open source data repository software. It is written in Python, and is the standard-bearer for repository software. Lamentably, it is also known for being difficult to install, although Docker images[3] have simplified this substantially.

Users of CKAN include Data.gov[4], and the National Oceanic and Atmospheric Administration[5], among many others. CKAN consultants include Open Knowledge, Ontodia[6], and Accela[7], in addition to many independent consultants. Paid CKAN hosts include Open Knowledge and Ontodia. Please check a CKAN demo site[8].

**DKAN:** DKAN[9] is a clone of CKAN, although it shares no code with CKAN – it has been rewritten in PHP, as a Drupal module. For an organisation that uses the Drupal content management system and also wants a data repository, DKAN is an especially good option. Users of DKAN include the U.S. Department of Agriculture[10], among others[11]. Please check a DKAN demo site[12].

**JKAN:** JKAN[13] is nominally based on CKAN, although it shares no code with it. JKAN was created by Tim Wisniewski[14], Philadelphia's Chief Data Officer, as a data catalog powered by Jekyll[15]. Note that JKAN is a data *catalog,* not a *repository,* which is to say that it stores links to data and metadata about that data, but not the data itself. The data could be hosted on an FTP server, in-place on agency websites, in Amazon S3, in Dropbox, or anywhere else one

[1] https://ckan.org
[2] https://okfn.org
[3] https://hub.docker.com/u/ckan/
[4] https://www.data.gov
[5] https://data.noaa.gov/dataset
[6] https://opengov.com/open-data
[7] https://www.accela.com
[8] https://demo.ckan.org/pt_BR/
[9] http://www.nucivic.com/dkan/
[10] https://data.nal.usda.gov
[11] https://github.com/NuCivic/dkan-sites
[12] http://demo.getdkan.com
[13] https://jkan.io
[14] https://usopendata.org/2016/03/28/jkan/
[15] https://jekyllrb.com

might store a file for public access. Setting up a site takes just a few minutes. Please check a JKAN demo site[16].

### 2.2.2. Commercial hosting

For some organisations, commercial hosting is going to be a viable option. Paying somebody to host your own data requires little to no technical knowledge on the part of your organisation, and the host will hold your hand through the process. Your organisation will not have to provide any technical infrastructure (e.g., servers) or know how to program. It is important however that you carefully consider the service level agreements.

**ArcGIS Open Data:** ArcGIS Open Data[17] is a new entrant in the field, having been released in late 2014. ArcGIS Open Data is included with an ArcGIS Online contract – because of the universality of that service among municipalities and states, it is effectively free for those existing customers. This makes it a very attractive option for governments with low levels of buy-in to an open data program, because it eliminates the cost of a data catalog. ArcGIS Open Data is only available as hosted software – it is not possible to run an instance of it on your own servers.

**Junar:** Junar[18] provides platforms and packages for businesses, governments, NGOs, and academia with a focus on data collection and analysis. Junar is bilingual, supporting English and Spanish audiences. Their pricing is targeted at small- to medium-sized organisations, starting at around US$10,000. Junar's demo site is available upon request.

**NuCivic Data:** NuCivic Data[19] is based on DKAN, which was created and is maintained by nücivic. They're a mid-range provider, in terms of pricing – their rates are much lower than socrata, but more expensive than, for example, Junar.

**CivicDashboards:** Open data consulting firm Ontodia provides hosted CKAN under the CivicDashboards[20] banner. They offer a free tier, for storing a small number of datasets. Their pricing is comparable with Junar's.

**OpenDataSoft:** OpenDataSoft[21] is a French company that has moved into the US market recently. They offer a free tier (up to 5 datasets, each of up to 20,000 records).

[16] https://demo.jkan.io
[17] https://hub.arcgis.com/pages/open-data
[18] http://www.junar.com
[19] https://getdkan.org
[20] http://www.civicdashboards.com
[21] https://www.opendatasoft.com

**Socrata Open Data:** Socrata[22] is the major vendor in the open data repository space, with their Socrata Open Data platform. Socrata only offers hosted options – there is no way to run Socrata's software on your own servers. It is both the most feature-rich and the most expensive option, with plans running into hundreds of thousands of dollars a year.

### 2.2.3. Free hosting

There are some options available for free hosting of open data repositories. (Note that the above listed open source options are also free, but require setup, a server, and maintenance time.) Generally, this is the lowest tier of service provided by paid hosts.

**DataHub:**The Open Knowledge Foundation provides DataHub[23], a free, CKAN-based data host. It is a large, collective repository – users do not get their own site, although it is possible to list only one's own data, and share a URL that only lists those datasets.

**GitHub:** GitHub[24] isn't really *meant* as a data repository, but it can serve as one. It has none of the niceties of proper repository software (conversion of formats, retrieving data from remote URLs, etc.), but it does offer previews of some types of data, publicly tracks changes, and is a reasonable place to store datasets.

It does offer one significant advantage, which is that GitHub – unlike any other repository software – provides a mechanism for people to propose changes to your datasets, which you can accept or decline, if they spot mistakes or areas for enhancement.

**JKAN on GitHub:** JKAN[25] is designed to be deployed onto GitHub, where the resulting data catalog can be hosted for free. In this way, GitHub can serve as a free host without sacrificing the niceties of a data catalog.

# 3. Managing an open data repository

## 3.1. Interoperability and data exchange

According to the FAIR principles, data needs to be 'Findable, Accessible, Interoperable, and Reusable'. The FAIR Data principles act as an international guideline for high-quality data stewardship. More in depth coverage on the exchange of data and best practices is covered in Unit 4: Exchanging Open Data.

[22] https://socrata.com

[23] http://datahub.io

[24] https://github.com

[25] https://how-to.usopendata.org/en/latest/The-Basics-of-Open-Data/Data-Repositories/#jkan

**To be findable:**

- (meta)data are assigned a globally unique and eternally persistent identifier
- data are described with rich metadata
- (meta)data are registered or indexed in a searchable resource
- metadata specify the data identifier.

**To be accessible:**

- (meta)data are retrievable by their identifier using a standardised communications protocol
- the protocol is open, free, and universally implementable
- the protocol allows for an authentication and authorisation procedure, where necessary
- metadata are accessible, even when the data are no longer available.

**To be interoperable:**

- (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
- (meta)data use vocabularies that follow FAIR principles
- (meta)data include qualified references to other (meta)data.

**To be re-usable:**

- meta(data) have a plurality of accurate and relevant attributes
- (meta)data are released with a clear and accessible data usage license
- (meta)data are associated with their provenance
- (meta)data meet domain-relevant community standards.

### *3.1.1. Publishing Metadata*

The requirement for enlightening, consistent, usable metadata is not new, but still is a challenge for many datasets. Having accurate metadata is vital not only for findability but also cataloguing – poor metadata can undermine the repository itself. Koesten *et al.*[26] explored the data search and sense-making needs of 20 data professionals including aspects that are directly relevant when deciding whether a dataset is relevant or not. They distinguish between three dimensions: *relevance* (is this the data I need?); *usability* (can I use it in practice?) and *quality* (how good is the data and how easy is it going to be do use it?). Data should be accompanied by descriptions of these aspects, either as structured metadata, but also in the form of comments, case studies, experience reports, examples of use, etc. See Table 3 below.

[26] Koesten, L, Kacprzak, E, Tennison, J and Simperl, E (2017) Trials and Tribulations of Working with Structured Data - a Study on Information Seeking Behaviour *CHI '17 Proceedings of the 2017 ACM SIGCHI Conference on Human Factors in Computing Systems* ACM, New York, USA. http://dx.doi.org/10.1145/3025453.3025838

*Table 3 Metadata users consider useful to make sense of data*

| Assess | Information needed about |
|---|---|
| Relevance | Context, coverage, original purpose, granularity, summary, time frame |
| Usability | Labelling, documentation, licence, access, machine-readability, language used, format, schema, ability to share |
| Quality | Collection methods, provenance, consistency of formatting/labelling, completeness, what has been excluded |

## 3.2. Governance considerations

Increased transparency and accountability of governance are key in the value proposition of open data for agriculture and nutrition. It is important to consider what data to open and how.  How accountable are data holders to both the demand side and policy makers? How do data producers and actors assure the quality of data? Who are the data stewards tasked to make the data open?[27]

Different countries or organisations will have different models to govern and administer their activities – i.e., different governance models. For example, you will find that some countries are more devolved in their decision making, while others have more centralised public administration. These governance models clearly impact how open data is governed – providing a broad patchwork of different open data governance across the world and making it difficult to identify who the open data decision makers and data gatekeepers or stewards are within a given country.

For example, if one wants to accelerate the opening up of agriculture data this may fall under the authority of sub-national government (such as states, provinces, territories or even cities), while in other countries agriculture is governed by central government or implemented through public–private partnership arrangements. Government data may be privatised, while in other cases it may be the responsibility of municipal or regional government. Responsibilities are therefore often distributed across administrative levels and agencies affecting how (open) data is produced, and published.

When considering the open data governance model for your open data initiative, map the open data governance process and ecosystem by identifying the following key stakeholders, their roles and responsibilities in the administration of open data, and seeking how they are connected:

- *Decision makers* – who leads/asserts decision authority on open data in meetings, procedures, conduct, debate, voting and other issues?

[27] Brandusescu, A., Lämmerhirt, D. AND Verhulst, S. **(**2017) Mapping open data governance models: Who makes decisions about government data and how? https://blog.okfn.org/2017/02/16/mapping-open-data-governance-models-who-decides-and-how/

- *Data holders* – which organisations/government bodies manage and administer agricultural data?
- *Data producers* – which organisations/government bodies produce what kind of agricultural and nutritional data?
- *Data quality assurance actors* – who are the actors ensuring that produced data adhere to certain quality standards?
- *Data gatekeepers/stewards* – who controls open data publication?

## 3.3. Curation and preservation

Data curation and preservation combines policies, strategies and actions to ensure the most accurate rendering possible of the data over time, regardless of the challenges of file corruption, media failure and technological change. Digital preservation applies to content that is born digital or converted to digital form.

Once the repository selection is made, data publication to the repository can be a challenge with several issues. The first issue to be addressed is selection of the data to publish. Does all data have value, or should only a selection be made available? Given that resources are finite, a form of selection for data must be made, but what criteria should be used to identify the data which are suitable for publication? All these considerations are articulated in policies and workflows which in reality are always evolving. These will be fully elaborated in Lesson 3.5. The objective is therefore not to set specific practices and responsibilities in stone, but rather to document new and existing policies with the intent of tracking changes as they occur.

People who have the primary role of managing or 'looking after' data come in a wide range of guises[28]. Their job titles include:

- data curators
- archivists
- librarians
- data librarians
- annotators.

As an example, the tasks of a data curator in the biosciences context can include ongoing data management, intensive data description, ensuring data quality, collaborative information infrastructure work, and metadata standards work. The full range of tasks and responsibilities encompassed by data curation might look something like this:

- developing and implementing policies and services
- analysing digital content to determine what services can be provided from it

[28] DCC – Digital Curation 101: Curate and Preserve. http://www.dcc.ac.uk/sites/default/files/documents/DC%20101%20Curation%20and%20Preservation.pdf

- providing advice to data creators and users/reusers
- ensuring submission of data to a repository
- negotiating agreements
- ensuring data quality
- ensuring that data are structured in the best way to provide access, rendering, storage and maintenance
- enabling the use and reuse of data
- enabling data discovery and retrieval
- preservation planning and implementation (for example, ensuring appropriate storage and backup routine, obsolescence monitoring)
- ensuring that policies and services are in place to make sure that data is viable, able to be rendered, understandable and authentic
- promoting interoperability.

The **DCC Curation Lifecycle Model**[29] (see Figure 1) provides a graphical high-level overview of the stages required for successful curation and preservation of data from initial conceptualisation or receipt. The model can be used to plan activities within an organisation or consortium to ensure that all necessary stages are undertaken, each in the correct sequence. The model enables granular functionality to be mapped against it; to define roles and responsibilities, and build a framework of standards and technologies to implement. It can help with the process of identifying additional steps which may be required, or actions which are not required by certain situations or disciplines, and ensuring that processes and policies are adequately documented.

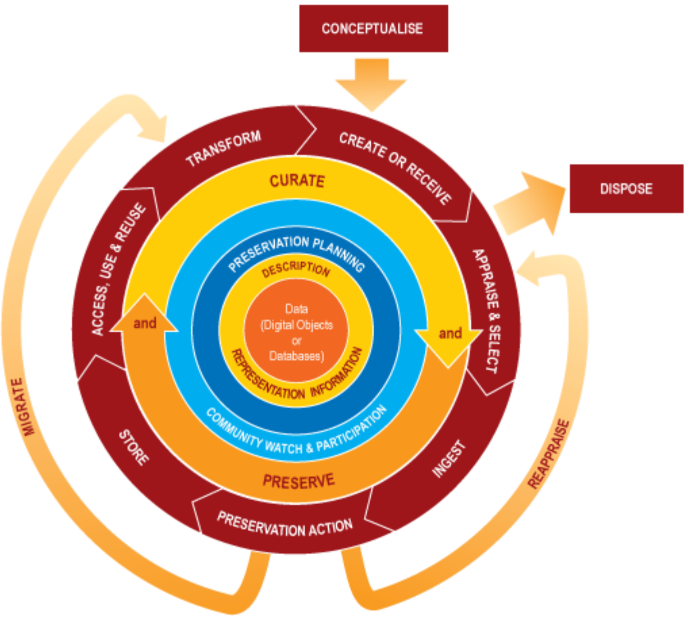

*Figure 1 The DCC Curation Lifecycle Model*

[29] http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf

*Table 4 Actions within the DCC Curation Lifecycle Model*

| **Full Lifecycle Actions** | |
|---|---|
| Description and Representation Information | Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand and render both the digital material and the associated metadata. |
| Preservation Planning | Plan for preservation throughout the curation lifecycle of digital material. This would include plans for management and administration of all curation lifecycle actions. |
| Community Watch and Participation | Maintain a watch on appropriate community activities, and participate in the development of shared standards, tools and suitable software. |
| Curate and Preserve | Be aware of, and undertake management and administrative actions planned to promote curation and preservation throughout the curation lifecycle. |
| **Sequential Actions** | |
| Conceptualise | Conceive and plan the creation of data, including capture method and storage options. |
| Create or Receive | Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation. Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centres, and if required assign appropriate metadata. |
| Appraise and Select | Evaluate data and select for long-term curation and preservation. Adhere to |

| | |
|---|---|
| | documented guidance, policies or legal requirements |
| Ingest | Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements. |
| Preservation Action | Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats. |
| Store | Store the data in a secure manner adhering to relevant standards |
| Access, Use and Reuse | Ensure that data is accessible to both designated users and reusers, on a day-to-day basis. This may be in the form of publicly available published information. Robust access controls and authentication procedures may be applicable. |
| Transform | Create new data from the original, e.g., by migration into a different format, by creating a subset, by selection or query, to create newly derived results, perhaps for publication. |
| **Occasional Actions** | |
| Dispose | Dispose of data, which has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements. Typically data may be transferred to another archive, repository, data centre or other custodian. In some instances data is destroyed. The data's nature may, for legal reasons, necessitate secure destruction. |

| | |
|---|---|
| Reappraise | Return data which fails validation procedures for further appraisal and reselection. |
| Migrate | Migrate data to a different format. This may be done to accord with the storage environment or to ensure the data's immunity from hardware or software obsolescence. |

## Summary

A data repository serves as the centerpiece of an open data effort. It serves as a central location to find data, a venue for standardising practices, and a showpiece of the use of that data. In a practical sense, a repository serves as a central, searchable place for people to find data.

Before selecting a solution for the repository, first consider the needs of your use and establish requirements:

- understand as much as possible about the users, as well as their work and the context of their work
- the system under development should support users in achieving their goals
- build upon the needs identified and produce a set of requirements
- use a user-centred approach to development
- understand what the repository should do and make sure it meets the stakeholders' needs.

When considering the **open data governance** model for your open data initiative, map the open data governance process and ecosystem by identifying the following key stakeholders, their roles and responsibilities in the administration of open data, and seeking how they are connected:

- decision makers
- data holders
- data producers
- data quality assurance actors
- data gatekeepers/stewards.

**Data curation and preservation** combines policies, strategies and actions to ensure the most accurate rendering possible of the data over time, regardless of the challenges of file corruption, media failure and technological change.

The **DCC Curation Lifecycle Model** provides a graphical high-level overview of the stages required for successful curation and preservation of data from initial conceptualisation or receipt. The model can be used to plan activities within an organisation or consortium to ensure that all necessary stages are undertaken, each in the correct sequence.