Documentation and metadata.

James Bukosia Wafula - KWTRP

June 25, 2019

Course content

- Documentation for reproducible research.
- Categories of documentation in RDM.
- When & what to document.
- Data management process flow documentation.
- Metadata: What is this?
- Functions of metadata.
- Types/categories of metadata.
- Metadata standards.
- Exposing metadata via harvesting.

Documentation

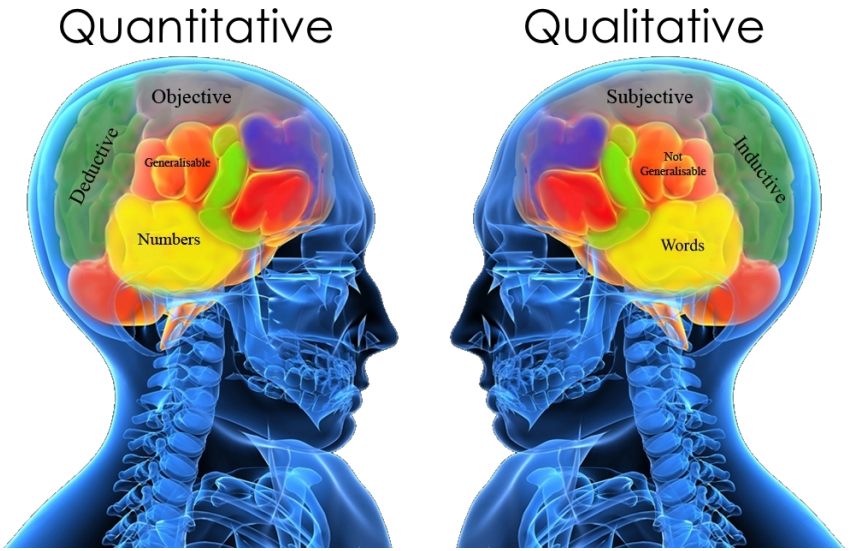
- Documentation ensures that anyone re-using the data can understand it and interpret it correctly.
- Explains how the data was created, context for the data, structure of the data and its contents, and any manipulations that have been done to the data. E.g: Segmentation manipulations on DICOM files, manifold learning techniques employed for NLP and other pattern recognition applications.
- Documentation ensures that data can be searched for and retrieved efficiently by users of data centres and repositories.

Documentation for reproducible research

- Documentation is not an option especially now that reproducible research is a key requirement from funders.
- Other researchers will be interested in reproducing results with minimal effort.
- Hence data documentation should be shared so that someone else can follow and replicate your results. E.g: Set random seeds for simulated data.

Documentation: Qualitative data

- Qualitative data is usually not generalizable.
- Documentation will make it easier for other qualitative researchers to understand your data.
- Makes research output visible and increases re-use potential, even when research results are not reproducible.



Categories of documentation in RDM

Documentation can be categorized into two:

- **Internal (embedded) documentation**: Which is written as program comments. It includes:
  - Code, field and label descriptions.
  - Descriptive headers or summaries.
  - Transcripts.
- **External (supporting) documentation:** Written for people who need to use the software, in a separate file. This can further be split into:
  - **Library documentation:** Which describes tools that a programmer can use.
  - **User documentation:** Which is intended for users of an application e.g. codebooks, questionnaires or interview guides.

Data documentation in RDM life cycle

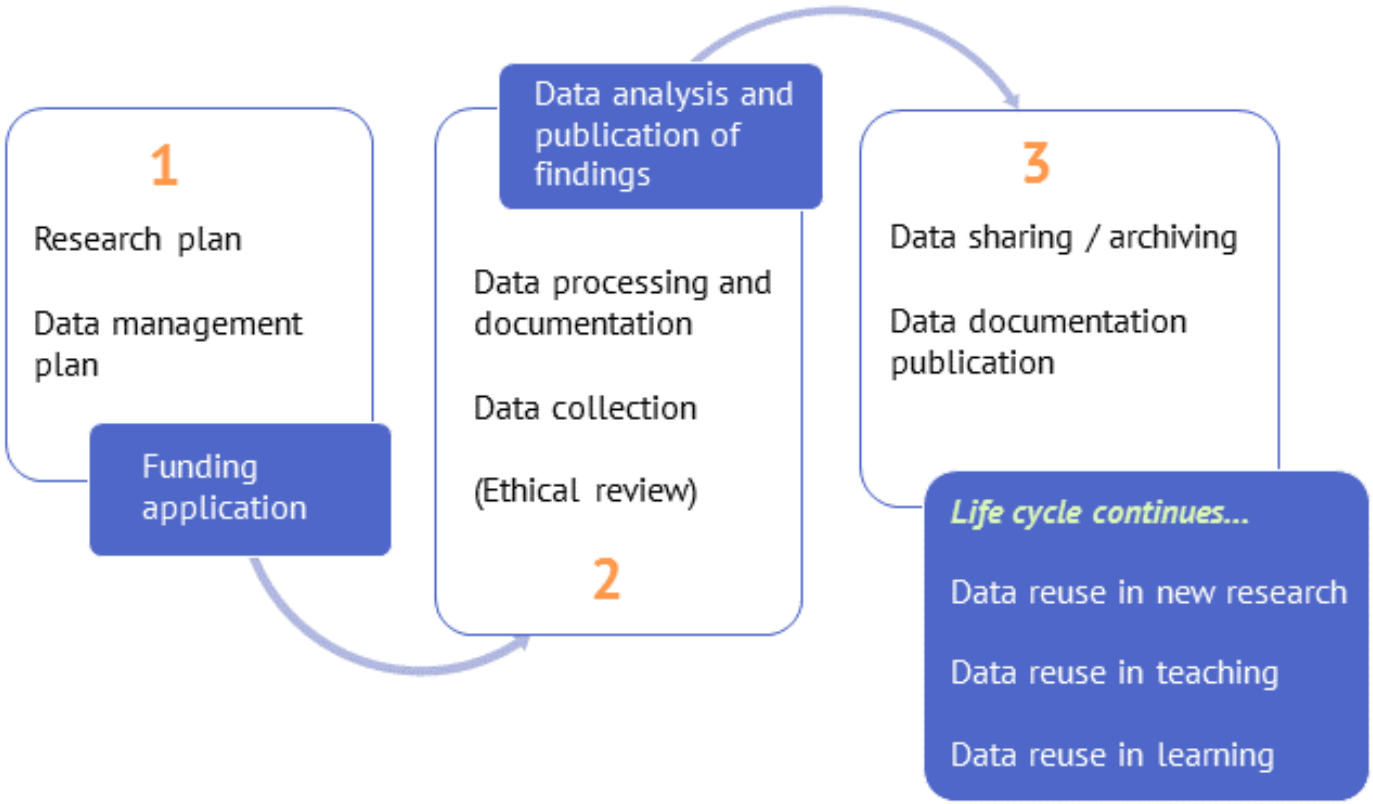


Figure: Data life cycle

What to document

It is vital to document the study for which the data has been collected and the data itself. Therefore, we can split documentation into:

- Project-level documentation.
- Data-level documentation.

Project-level documentation

- Purpose for which data was collected.
- Contents of the dataset.
- Data collection method employed.
- Who collected the data.
- Data processing techniques.
- Manipulations/ transformations done on data.
- Quality assurance.
- Data access.

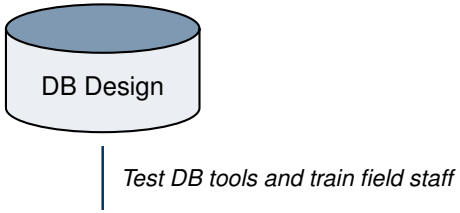
Data-level (object-level) documentation

- Provides information at the level of individual objects e.g. interview transcripts or variables in a database.
- Can be embedded in datafiles e.g. descriptive information about an interview at the begining of each file; or variable names embedded in each file for quantitative data.

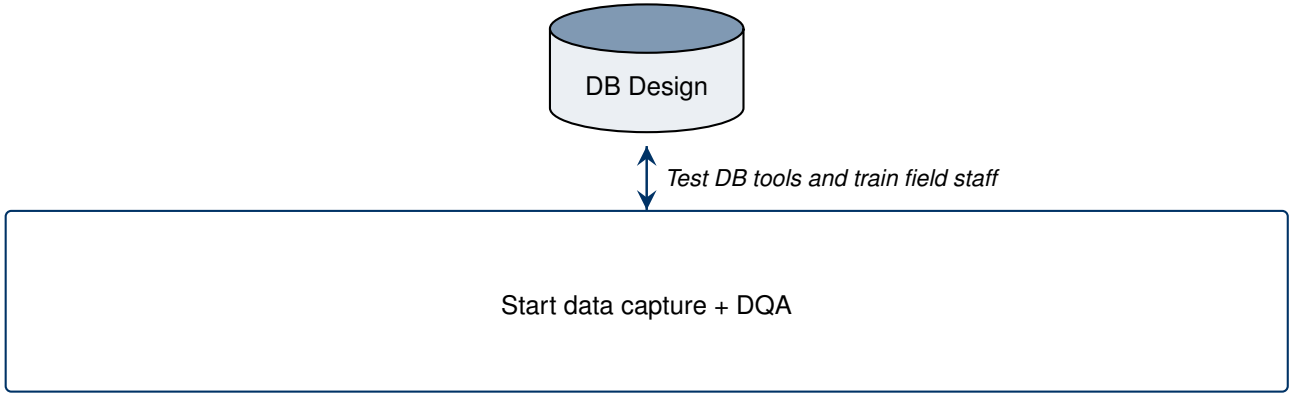
IBV_Codebook.pdf

Variable Name	Description	Value Label	Format	Type
age_months	Age in months		Numeric	Byte
age_years	Age in years		Numeric	Byte
sex	Gender of participants	Female Male	String	Character
facility	Study Health facility	Chasimba Jaribuni Junju Matsangoni Mavueni Mtondia Ngerenya Pingilikani Sokoke KCH	String	Character
fever	History of fever	No Yes	String	Character
cough	Cough	No Yes	String	Character
nasaldischarge	Nasal discharge	No Yes	String	Character
breathing_difficulty	Breathing difficulty	No Yes	String	Character
e229	MERS-Cov E229 ct value		Numeric	Double

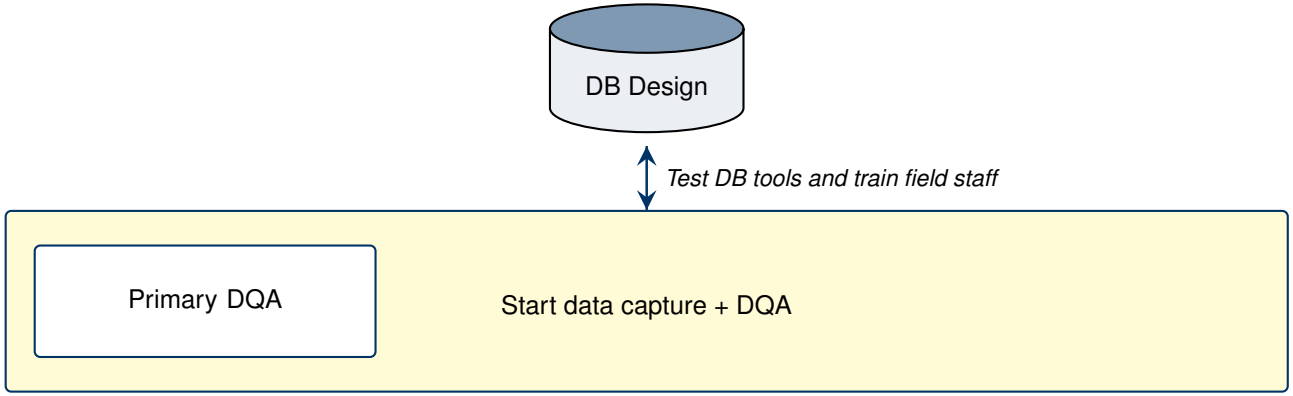
Data management process flow documentation example



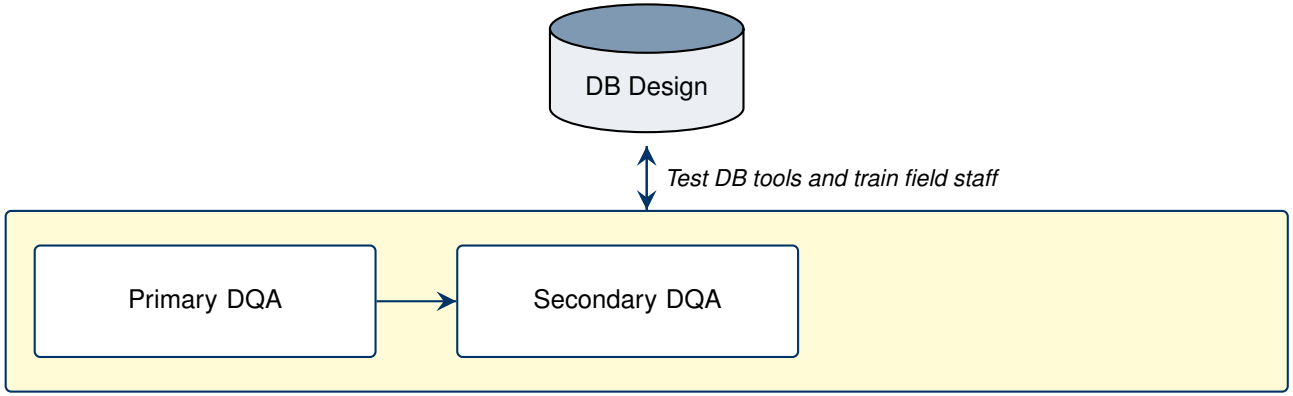
Data management process flow documentation example



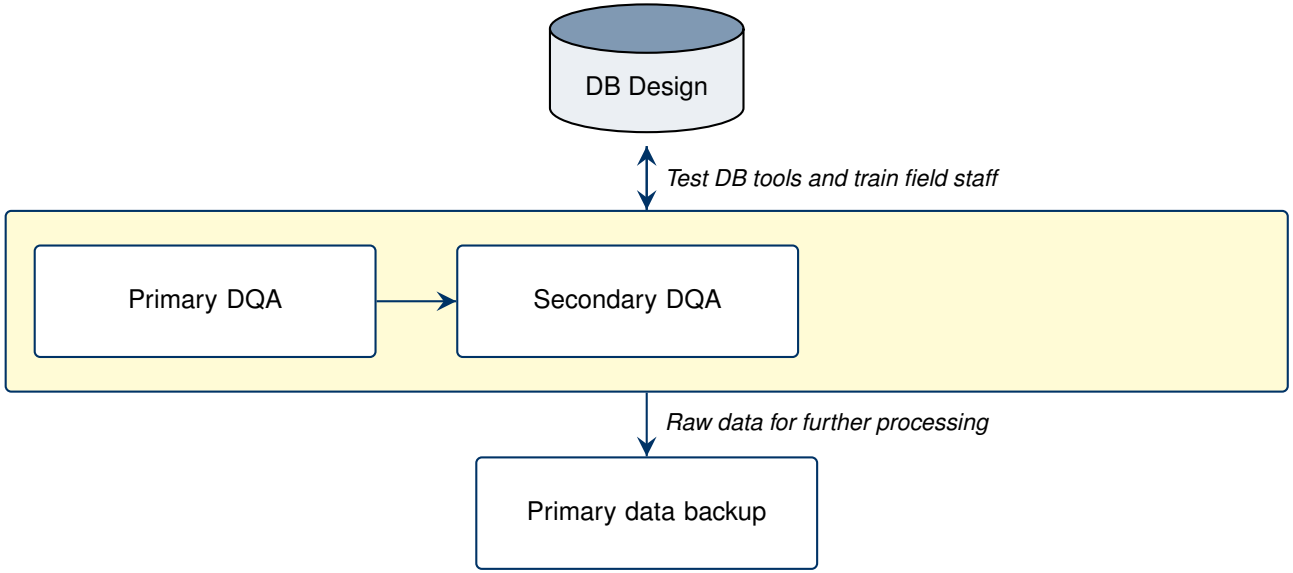
Data management process flow documentation example



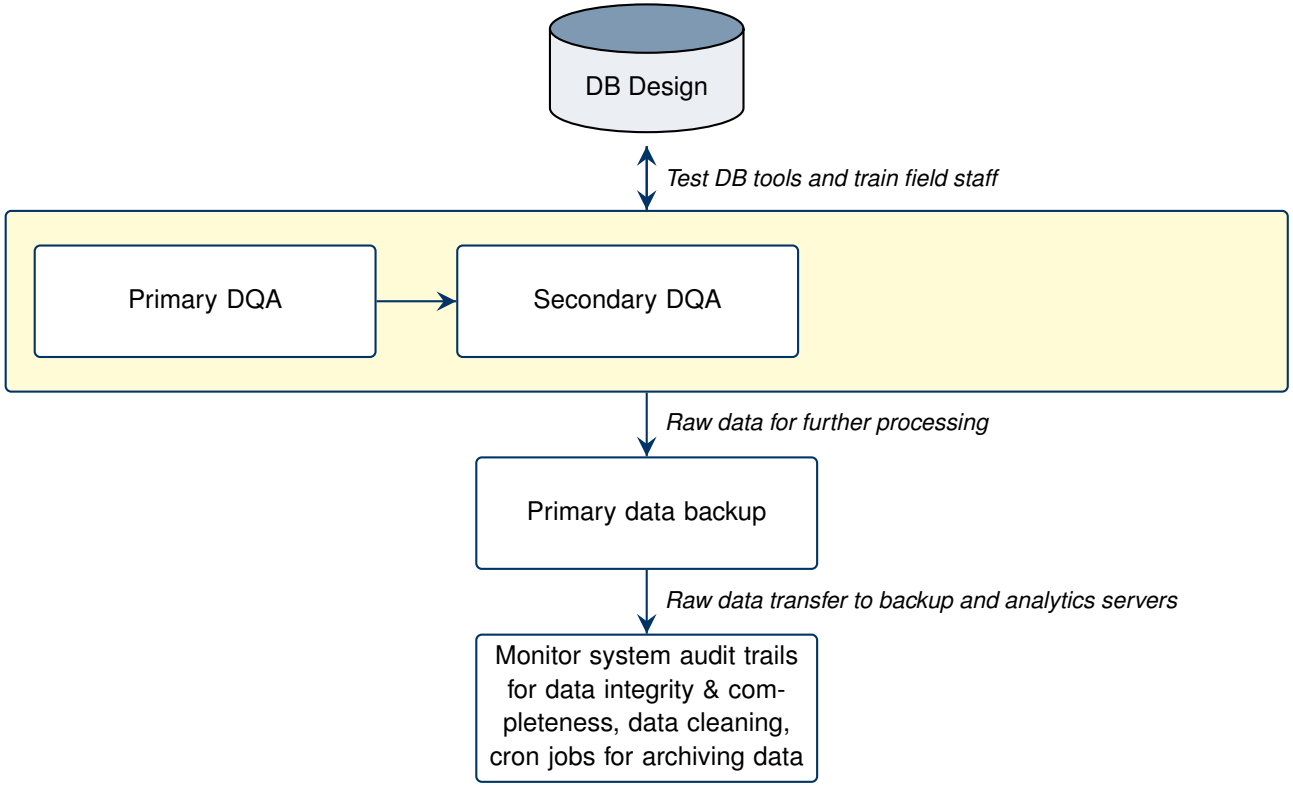
Data management process flow documentation example



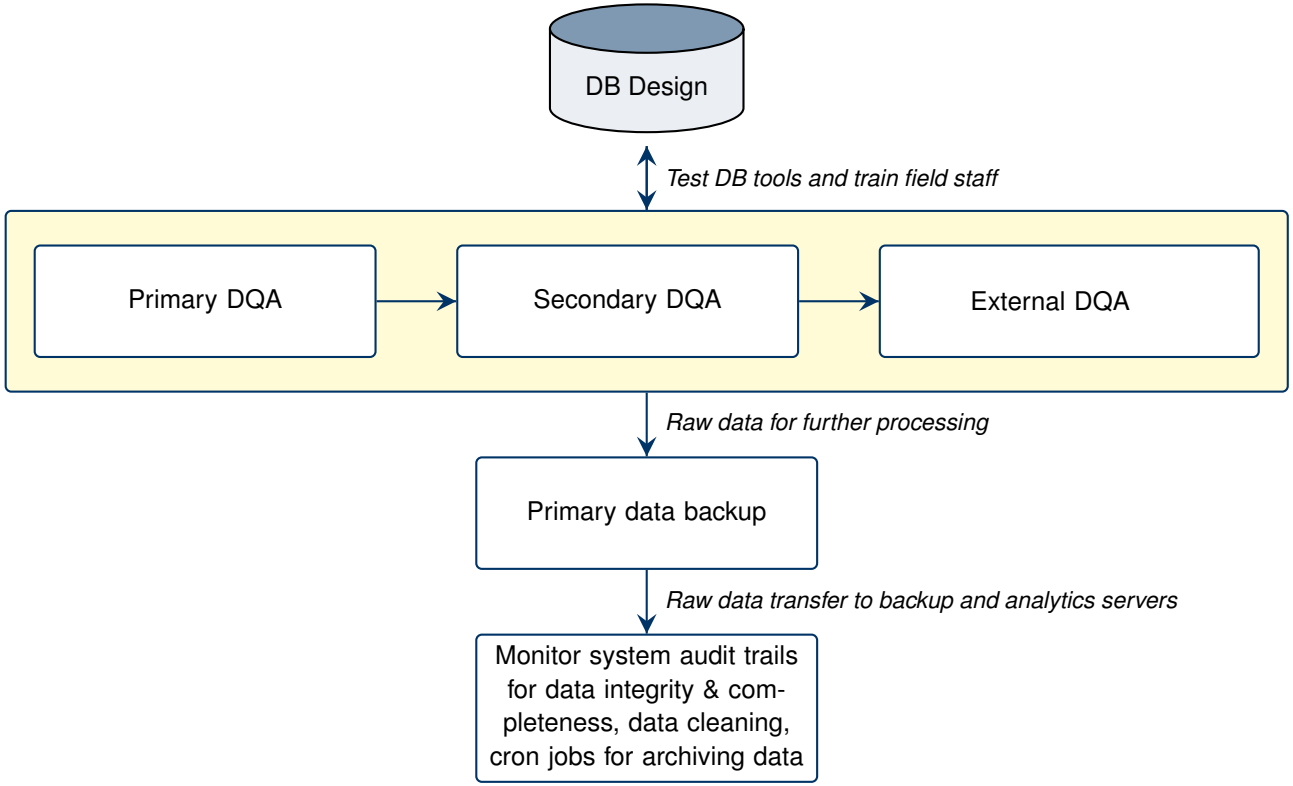
Data management process flow documentation example



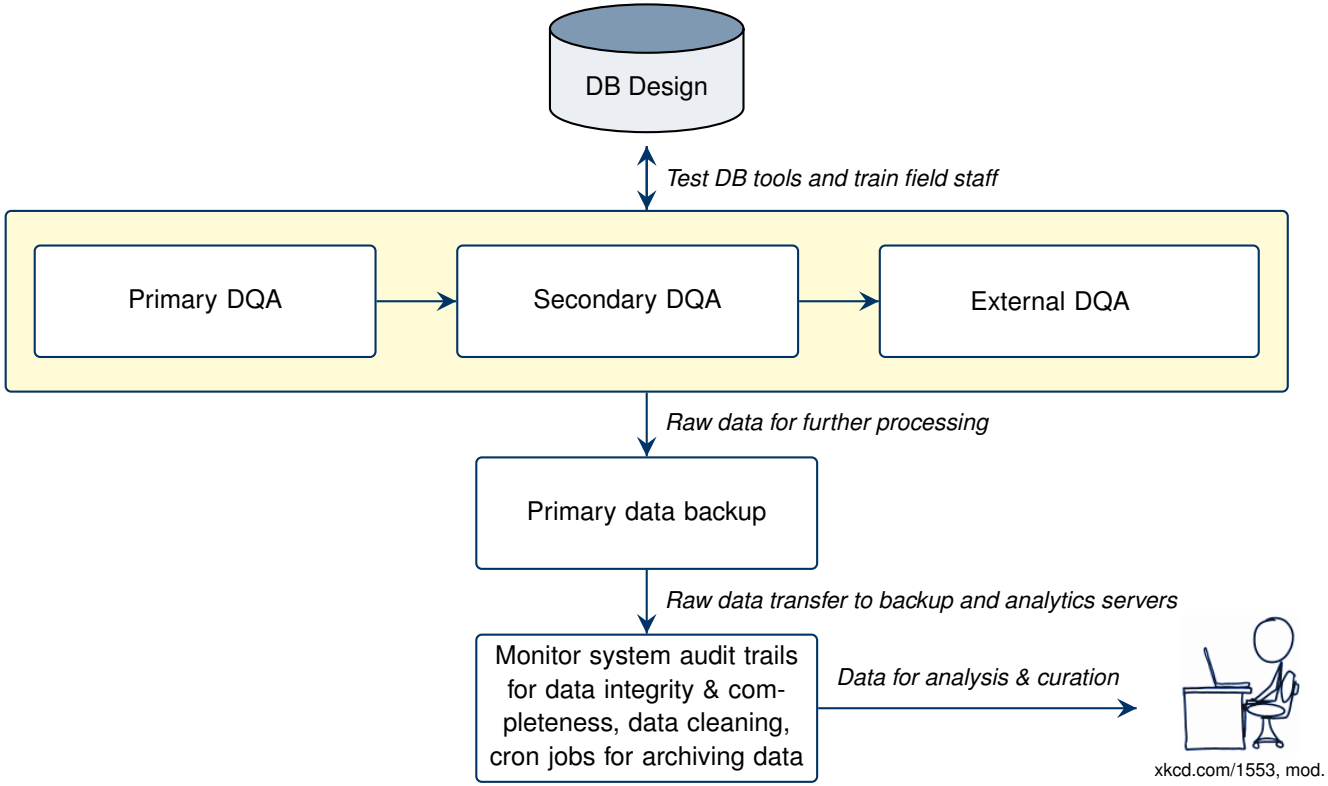
Data management process flow documentation example



Data management process flow documentation example



Data management process flow documentation example



Data capture documentation (SOP) example

Activities Image Viewer Sat Jun 22 14:18:09 74 %

106% Screenshot from 2019-06-19 08-51-28.png

BIO DATA

Information common to all patients

1. **Unique ID** –This is a seven (7) digit auto-generated and incremental number which acts as the unique identifier for this record. *For example; Kiambu the first record would read* ***51****00001*.
2. **Document source**-indicate which of the following forms is used: Paediatric Admission Record (PAR), New-born Admission Record (NAR) or admission notes in free text.
3. **Surgical or Burns diagnosis**: Check from the clinician's notes. Please indicate Yes if surgical or burns diagnosis are the ONLY diagnosis made. If no mention of a surgical or burn diagnosis OR there is other diagnosis in addition to the surgical/burns diagnosis fill NO. Surgical diagnosis include- hernia, acute abdomen, fracture (also written as #), joint dislocation, fall from a height, head injury, road traffic accident (RTA) and soft tissue injury.
4. **Admission date** – please indicate the date of admission as indicated on the ward clinical admission notes- it is the date when the patient is first seen by a clinician, if not available check in other sections of the file (e.g. cardex).

Figure: Eg.SOP

Example post-project (publication) documentation

See example documentation files @ https://doi.org/10.7910/DVN/PP6QRQ

- variables codebook.
- readme file.

Metadata

What is this?

☺ ☺ ☺

TOP DEFINITION

metadata

A fancy word for "information" invented by tech folks to make their jobs sound harder than they really are.

My metadata is flaring up today.

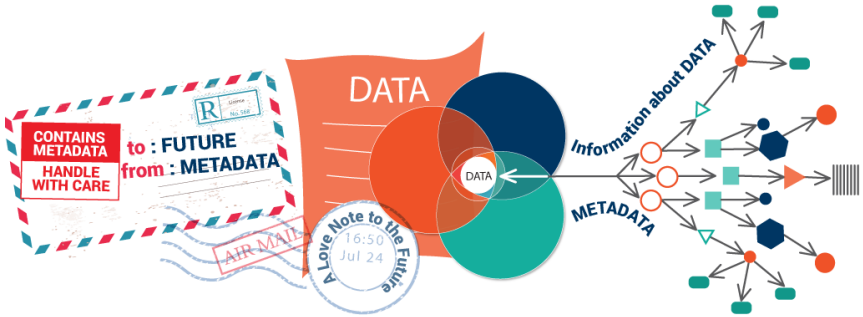
#information #data #words #seo #marketing

by clayton rattlesnake January 25, 2012

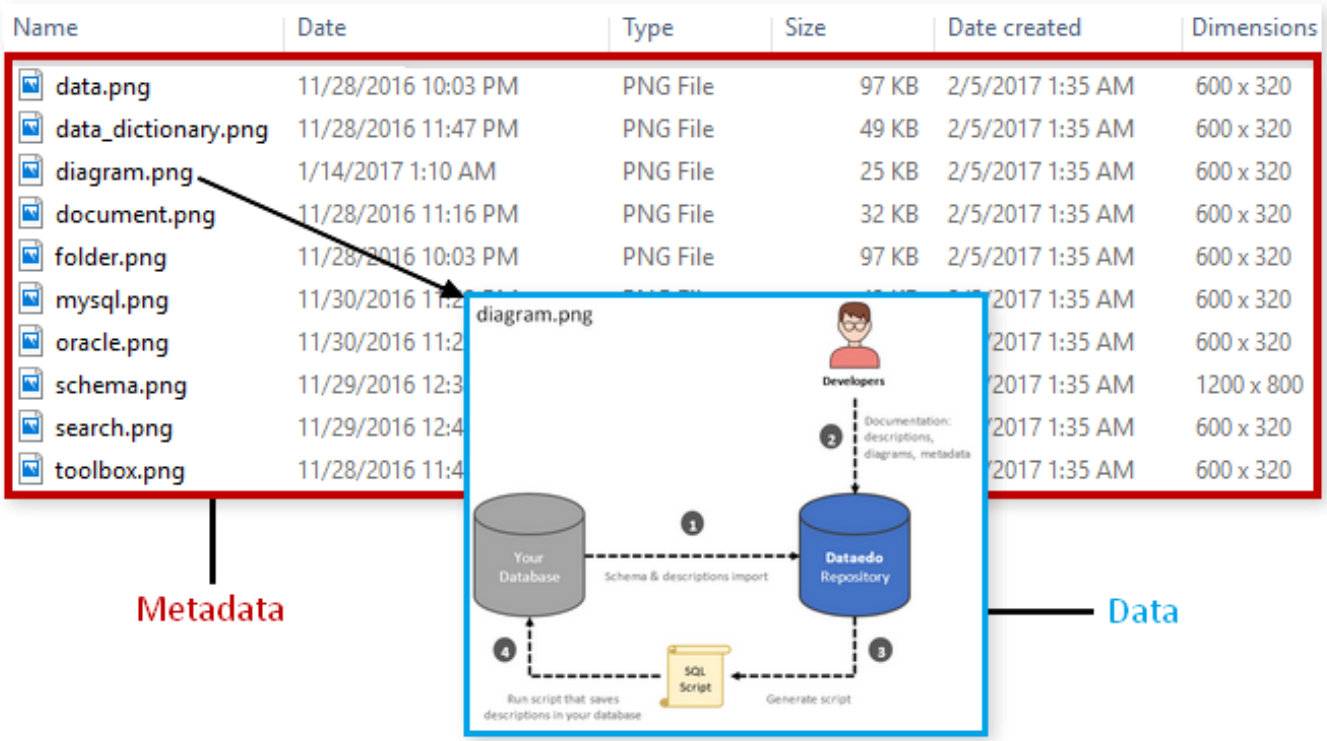
18 9

Metadata made simple

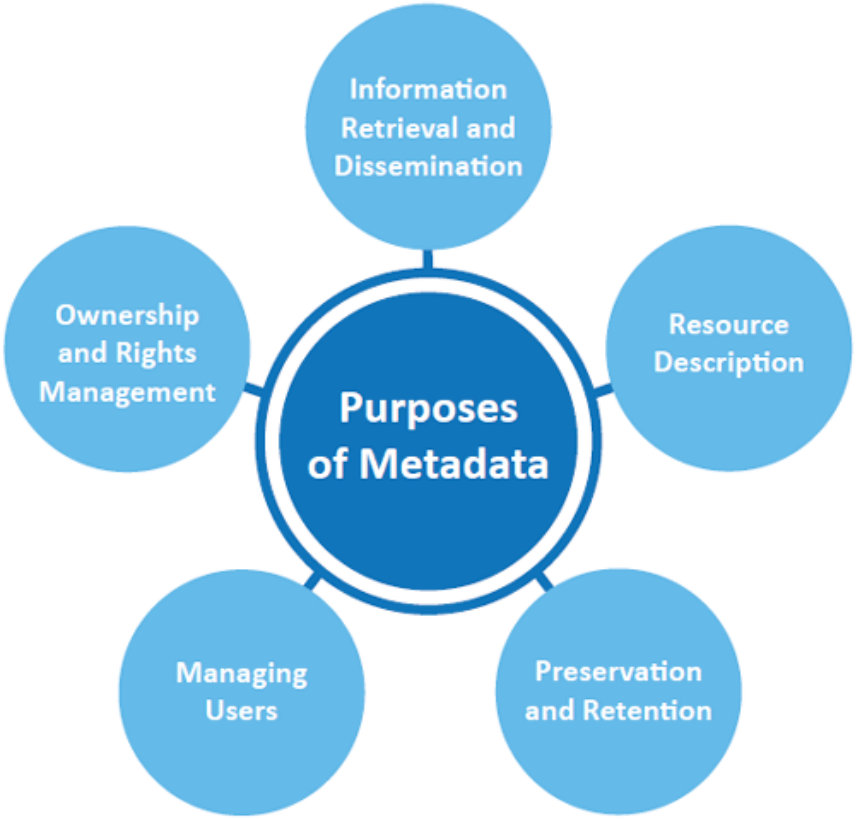
- Data about data.
- It explains the origin, purpose, time, geographic location, creator, access, and terms of use of the data.
- Important for purposes of retrieving and indexing data in a repository or archives, and for citation.



Quick example of everyday metadata



Functions of metadata



Functions of metadata

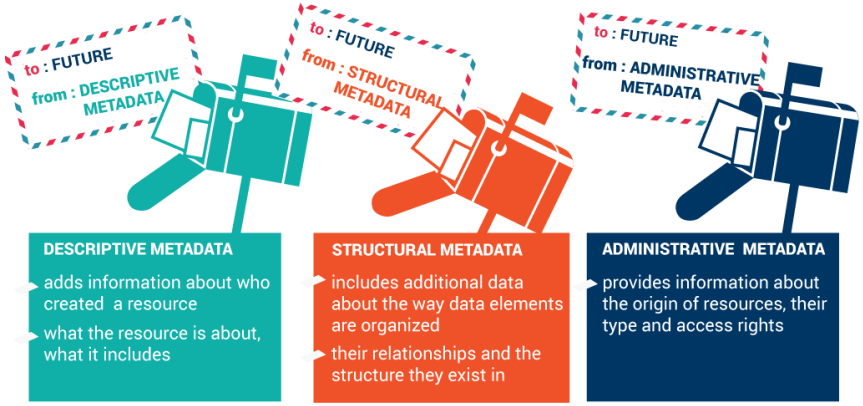
- **Resource discovery**: Allowing resources to be found by relevant criteria; Identifying resources; Bringing similar resources together; Distinguishing dissimilar resources; Giving location information.
- **Digital identification**: e.g. using digital object identifiers (DOI).
- **Organizing e-resources**: Organizing links to resources based on audience or topic; Building these pages dynamically from metadata stored in databases.

Functions of metadata

- **Facilitating interoperability**: Using defined metadata schemes, shared transfer protocols, and crosswalks between schemes, resources across the network can be searched more seamlessly. Cross-system search, e.g., using Z39.50 protocol; Metadata harvesting, e.g., OAI protocol.
- **Archiving and preservation**: requires tracking of the lineage of a digital object, and documentation of its behavior in order to emulate it in future technologies.

Types of metadata

- **Descriptive**: for discovery and identification.
- **Structural**: indicates how compound objects are put together e.g. how pages are ordered to form chapters.
- **Administrative**: Rights management and preservation.



Categories of metadata

- **Application Metadata:** Data created by the application specific to the electronically stored information (ESI) being addressed, embedded in the file, and moved with the file when copied. Copying may alter application metadata.
- **Document Metadata:** Properties about the file stored in the file, as opposed to stored in the document content. Often this data is not immediately viewable in the software application used to create/edit the document, but can generally be accessed via a "Properties" view. Examples include document author and company, and creation or revision dates.

Categories of metadata

- **File System Metadata:** Metadata generated by the system to track the demographics (name, size, location, usage, etc.) of the ESI that are stored externally from, rather than embedded within, the ESI.
- **Embedded Metadata:** Generally hidden, but an integral part of ESI, such as "track changes" or "comments" in a word processing file or "notes" in a presentation file. While some metadata is routinely extracted during processing and conversion for e-discovery, embedded data may not be. Therefore, it may only be available in the original, native file.

Metadata example 1

Applications Places Firefox Wed 07:26

CIN (Paediatric) Content - CIN Paediatrics - Mozilla Firefox

CIN (Paediatric) Content

10.0.10.199:8686/web/guest/site-set-up?p_p_id=110_INSTANCE_5CWzol0t50E8&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&p_p_col_id=column-1&p_p_col_count=1&_110_INSTANCE_

Welcome | Research Portal | Code Archives | CIN (Paediatric) Content

Download | Edit | Move | Checkout | Permissions | Move to the Recycle Bin

Tips On Fluids_Non 14_GI_GM.pptx

Uploaded by admin HSU, 2/18/16 8:13 AM

Average (0 Votes)

Comments

No comments yet. Be the first. Subscribe to Comments

Version 1.0

Last Updated by admin HSU

2/18/16 8:13 AM

Status: Approved

Download (16.8MB) Get URL or WebDAV URL.

Automatically Extracted Metadata

Modified
2015-08-19T04:42:43Z

Creator
George Mbevi

Created
2014-11-14T06:06:09Z

Title
Tips On Data Entry On Fluids

Content Type
application/vnd.openxmlformats-officedocument.presentationml.presentation

Last Modification
2015-08-19T04:42:43Z

Last Author
George Mbevi

Author
George Mbevi

Application Name
Microsoft Office PowerPoint

Revision Number
19

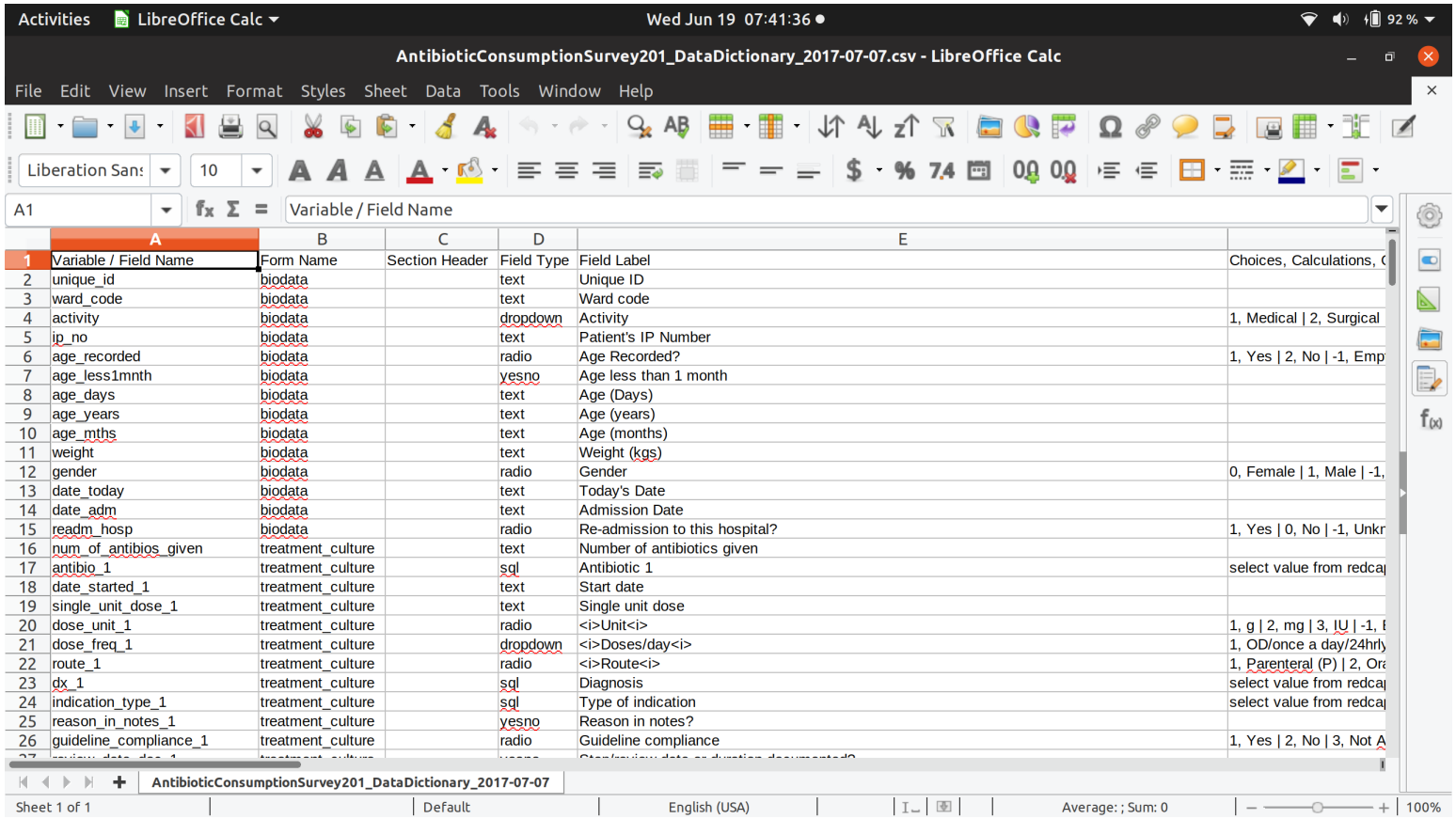
Template

CIN (Paediatric) Content - CIN Paedi...

1 / 4

Figure: Eg.1 Metadata

Metadata example 2



	A	B	C	D	E	
1	Variable / Field Name	Form Name	Section Header	Field Type	Field Label	Choices, Calculations, (
2	unique_id	biodata		text	Unique ID	
3	ward_code	biodata		text	Ward code	
4	activity	biodata		dropdown	Activity	1, Medical | 2, Surgical
5	ip_no	biodata		text	Patient's IP Number	
6	age_recorded	biodata		radio	Age Recorded?	1, Yes | 2, No | -1, Emp
7	age_less1mnth	biodata		yesno	Age less than 1 month	
8	age_days	biodata		text	Age (Days)	
9	age_years	biodata		text	Age (years)	
10	age_mths	biodata		text	Age (months)	
11	weight	biodata		text	Weight (kgs)	
12	gender	biodata		radio	Gender	0, Female | 1, Male | -1,
13	date_today	biodata		text	Today's Date	
14	date_adm	biodata		text	Admission Date	
15	readm_hosp	biodata		radio	Re-admission to this hospital?	1, Yes | 0, No | -1, Unkr
16	num_of_antibios_given	treatment_culture		text	Number of antibiotics given	
17	antibio_1	treatment_culture		sql	Antibiotic 1	select value from redca
18	date_started_1	treatment_culture		text	Start date	
19	single_unit_dose_1	treatment_culture		text	Single unit dose	
20	dose_unit_1	treatment_culture		radio	<i>Unit<i>	1, g | 2, mg | 3, IU | -1,
21	dose_freq_1	treatment_culture		dropdown	<i>Doses/day<i>	1, OD/once a day/24hrly
22	route_1	treatment_culture		radio	<i>Route<i>	1, Parenteral (P) | 2, Ora
23	dx_1	treatment_culture		sql	Diagnosis	select value from redca
24	indication_type_1	treatment_culture		sql	Type of indication	select value from redca
25	reason_in_notes_1	treatment_culture		yesno	Reason in notes?	
26	guideline_compliance_1	treatment_culture		radio	Guideline compliance	1, Yes | 2, No | 3, Not A

Figure: Eg.2 Metadata

Metadata example 3

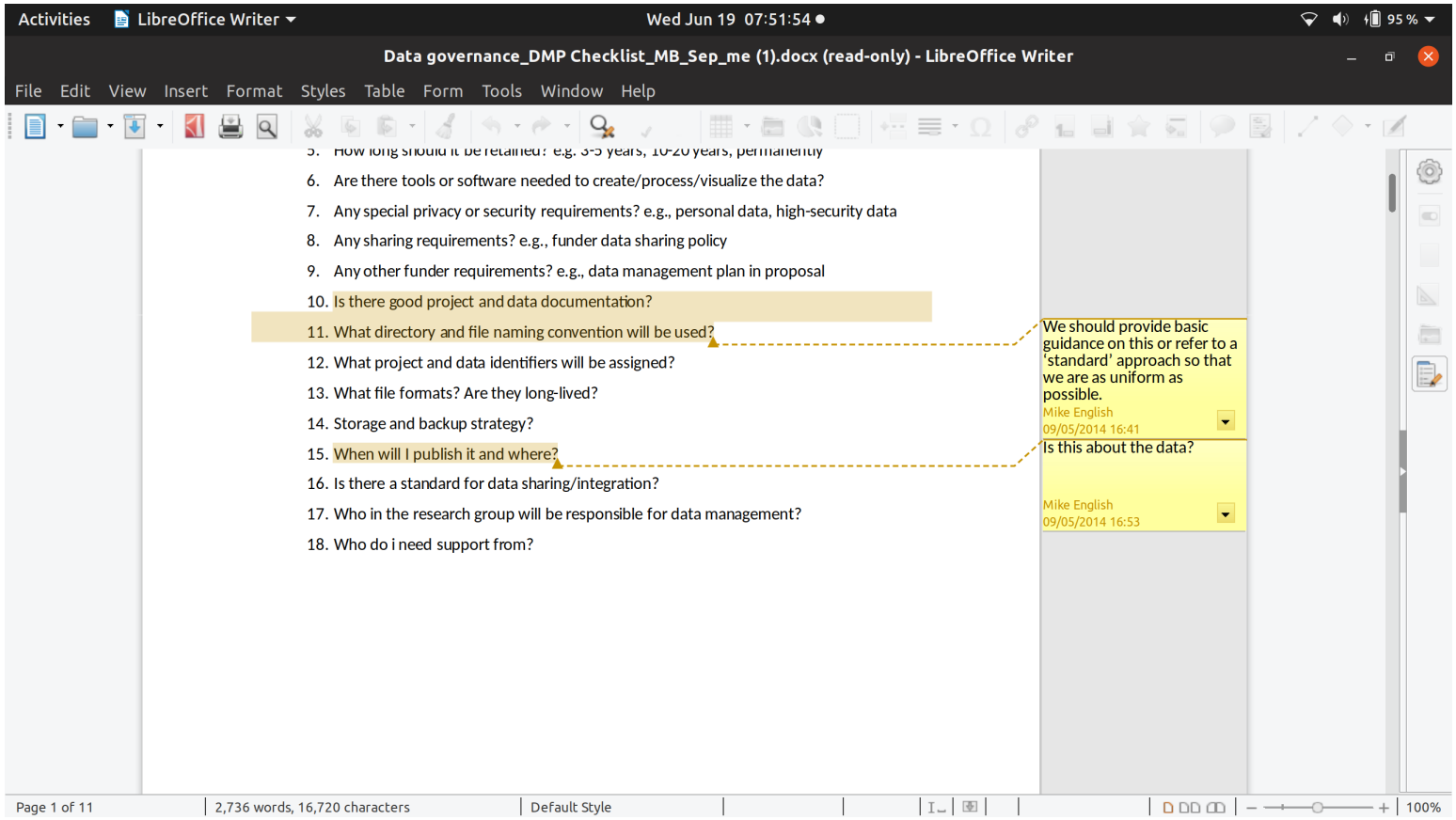
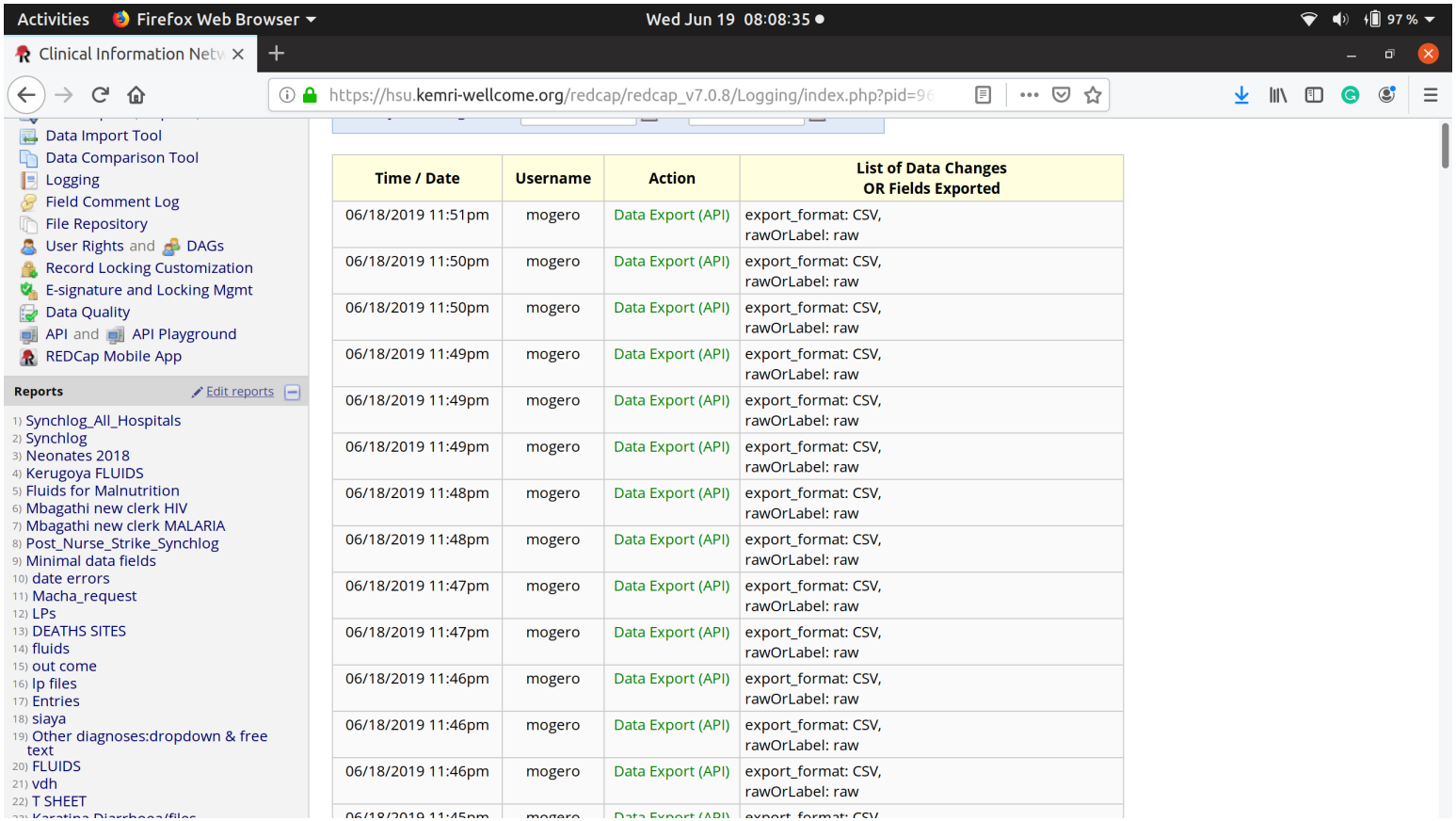


Figure: Eg.3 Metadata

Metadata example 4



Time / Date	Username	Action	List of Data Changes OR Fields Exported
06/18/2019 11:51pm	mogero	Data Export (API)	export_format: CSV, rawOrLabel: raw
06/18/2019 11:50pm	mogero	Data Export (API)	export_format: CSV, rawOrLabel: raw
06/18/2019 11:50pm	mogero	Data Export (API)	export_format: CSV, rawOrLabel: raw
06/18/2019 11:49pm	mogero	Data Export (API)	export_format: CSV, rawOrLabel: raw
06/18/2019 11:49pm	mogero	Data Export (API)	export_format: CSV, rawOrLabel: raw
06/18/2019 11:49pm	mogero	Data Export (API)	export_format: CSV, rawOrLabel: raw
06/18/2019 11:48pm	mogero	Data Export (API)	export_format: CSV, rawOrLabel: raw
06/18/2019 11:48pm	mogero	Data Export (API)	export_format: CSV, rawOrLabel: raw
06/18/2019 11:47pm	mogero	Data Export (API)	export_format: CSV, rawOrLabel: raw
06/18/2019 11:47pm	mogero	Data Export (API)	export_format: CSV, rawOrLabel: raw
06/18/2019 11:46pm	mogero	Data Export (API)	export_format: CSV, rawOrLabel: raw
06/18/2019 11:46pm	mogero	Data Export (API)	export_format: CSV, rawOrLabel: raw
06/18/2019 11:46pm	mogero	Data Export (API)	export_format: CSV, rawOrLabel: raw

Figure: Eg.4 Metadata

Metadata example 5

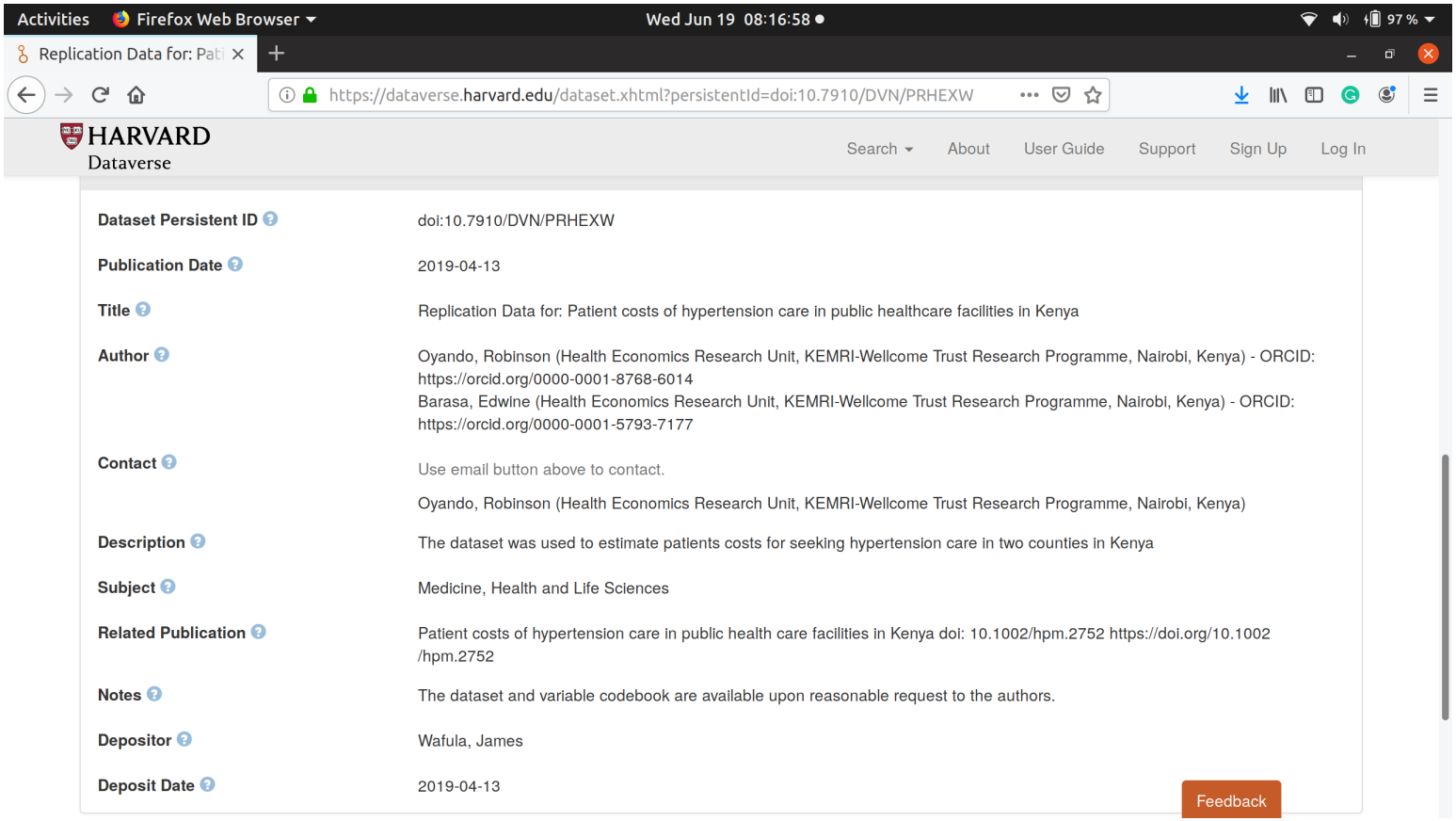


Figure: Eg.5 Metadata

Metadata example 6

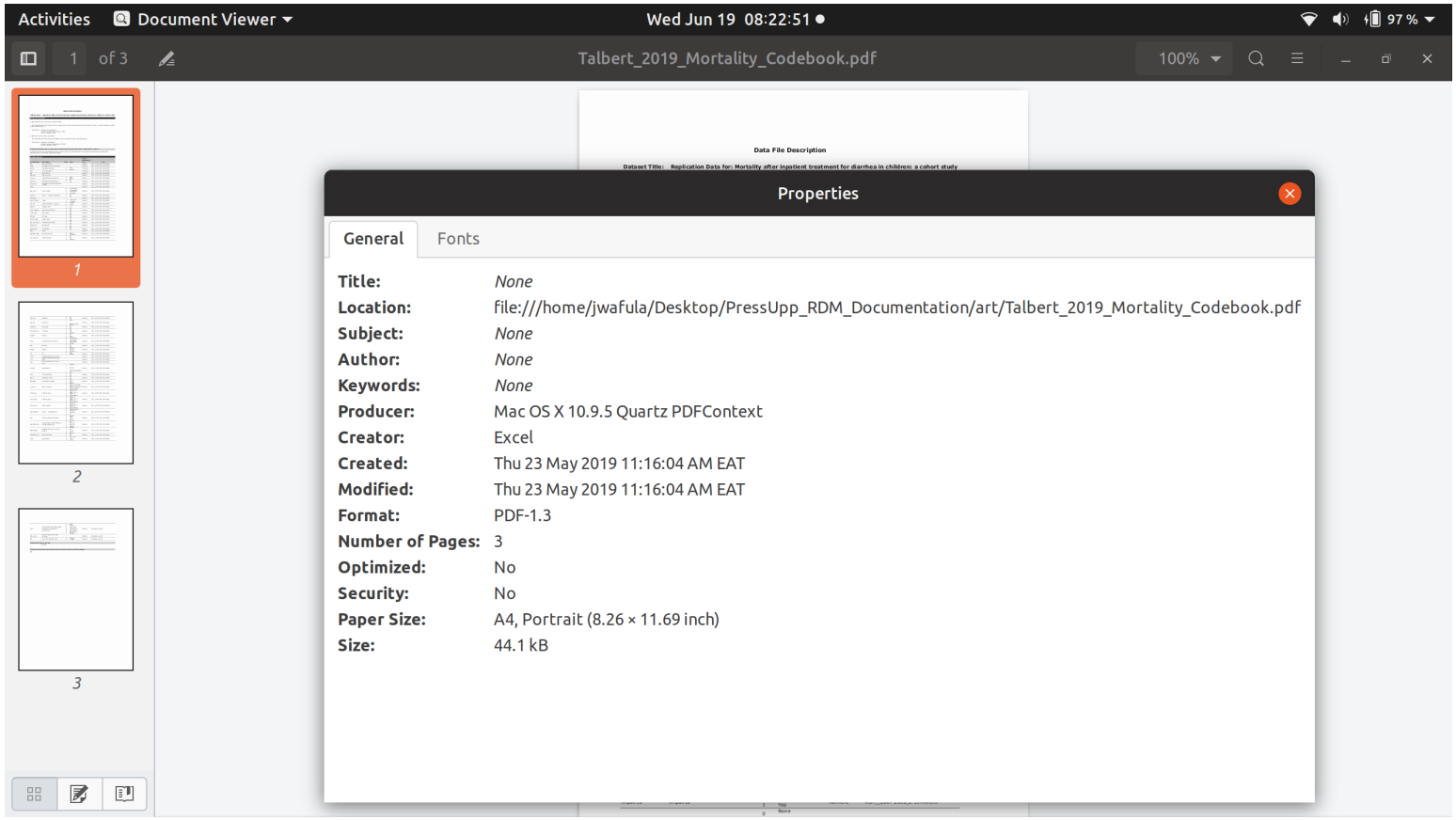


Figure: Eg.6 Metadata

Metadata standards

- Metadata takes the form of a structured set of elements which describe the information resource and assists in the identification, location and retrieval of it by users.
- Aim of standardisation is to make metadata support a number of defined functions.
- Often starts as schemas developed by a particular user community to enable the best possible description of a resource type for their needs.

Examples of metadata standards

Refer to:
http://www.dcc.ac.uk/resources/subject-areas/biology
http://rd-alliance.github.io/metadata-directory/standards/

- OAI Dublin Core - minimal requirement for all OAI providers.
- MARCXML - utilizes marcxml as the metadata schema name.
- e-GMS (e-Government Metadata Standard).
- ISO 19115: 2003(E) — Geographic Information: Metadata.
- PREMIS: Data Dictionary for Metadata Preservation.
- Metadata Object Description Schema (MODS): for library applications.

Metadata standards supported by Harvard Dataverse

- Citation Metadata: compliant with DDI Lite, DDI 2.5 Codebook, DataCite 3.1, and Dublin Core's DCMI Metadata Terms.
- Geospatial Metadata: compliant with DDI Lite, DDI 2.5 Codebook, DataCite, and Dublin Core.
- Social Science Humanities Metadata: compliant with DDI Lite, DDI 2.5 Codebook, and Dublin Core.
- Astronomy and Astrophysics Metadata : Based on Virtual Observatory (VO) Discovery and Provenance Metadata.
- Life Sciences Metadata: based on ISA-Tab Specification.

Exposing metadata via harvesting

- There is increasing need for open access to published work and reproducibility in research.
- Metadata harvesting is the automated collection of metadata descriptions from different sources to create useful aggregations of metadata.
- Metadata can be harvested for data sharing through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). OAI-PMH is a low-barrier mechanism for repository interoperability.
- OAI-PMH has two levels of participants:
  - Data providers - who administer the system.
  - Service providers - who use the metadata harvested to build their digital collection.

Exposing metadata via harvesting

- A harvester is used by a service provider as a way to collect metadata from a repository.
- A repository is a network-accessible server that is able to process OAI-PMH requests. Dataverse is an example of an OAI server.
- Only the published, unrestricted datasets in Dataverse can be made harvestable.
- A repo is managed by the data provider to allow harvesters access to its metadata.

Metadata harvesting

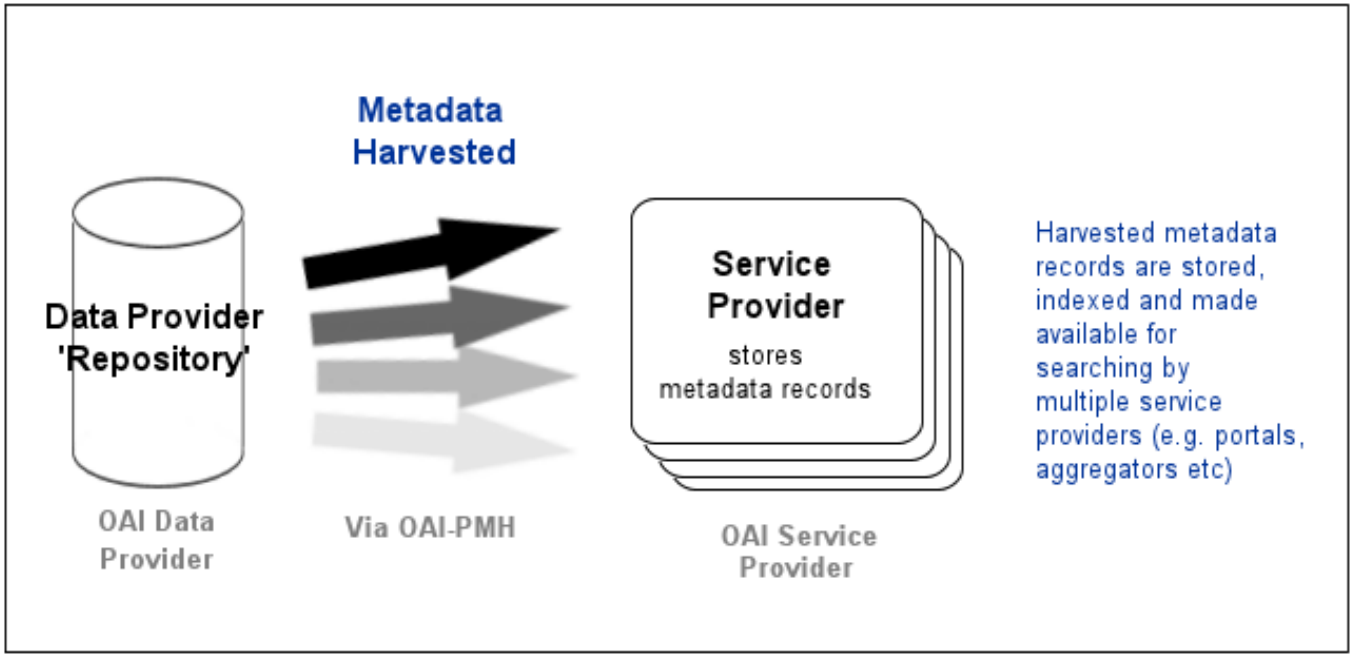


Figure 2. Exposing Metadata via Harvesting

Metadata harvesting

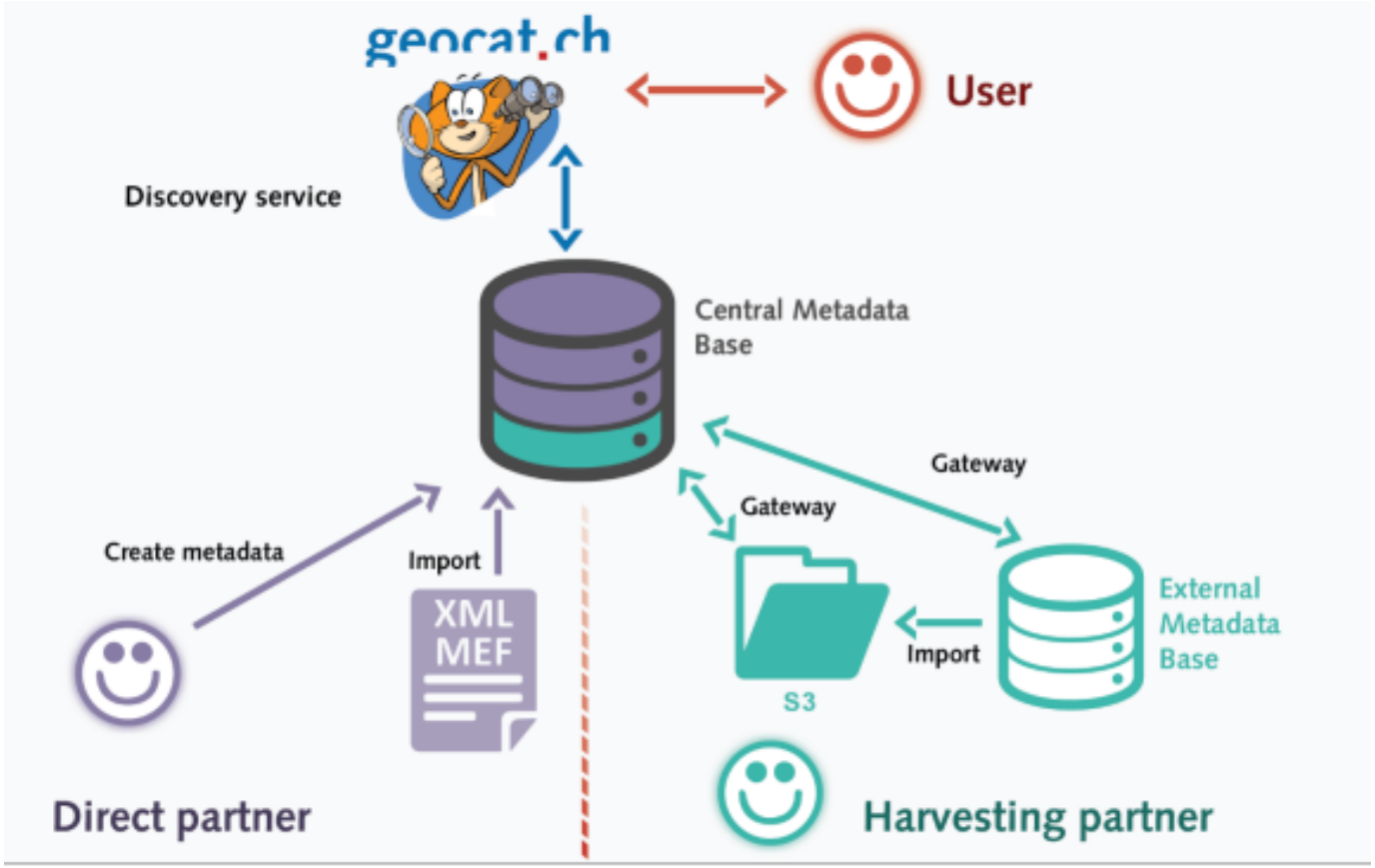


Figure: https://www.geocat.admin.ch/en/information/partner-model.html

Acknowledgements

Initiative to Develop African Research Leaders

tibaK

KEMRI | Wellcome Trust