Scientific Data Management
(case of *icipe*)

Dr Kennedy Senagi

www.icipe.org

icipe

1

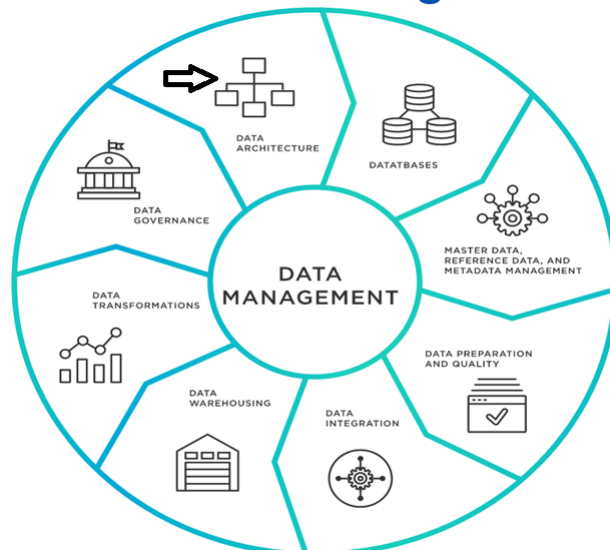
Key Aspects of Data Management

Elements
- Planning
- Collection
- Analysis
- Archiving
- Re-use

www.icipe.org

icipe

2

FAIR Data Management Principles

$F_{indable}$ $A_{ccessible}$ $I_{nteroperable}$ $R_{eusable}$

Findable
- Unique and persistent identifiers for (meta)data
- Rich set of metadata describing the datasets
- (Meta)data indexed within a centrally available catalog/server

Reduce time wasted searching for and finding relevant data

Accessible
- (Meta)data are retrievable through the unique identifiers (see findable)
- Authentication and authorization procedures ensure compliance
- Accessible ≠ open

Foster sharing of data within and across institutions, respecting patients and IP

Interoperable
- (Meta)data follow a common format/structure and terminology
- (Meta)data is represented by FAIR vocabularies
- Vocabularies are mapped against industry standards

Reduced time to integrate data across entities/ projects and create new insights

Reusable
- Provenance information is sufficiently available and linked to source data
- Domain experts are able to understand and work with data without additional explanation

Reduce spend in data generation

www.icipe.org

icipe

3

Open Science

OPEN SCIENCE LIFE CYCLE

Open Science: Innovation, Science and Economic Development Canada [2022]

www.icipe.org

icipe

4

Open Data
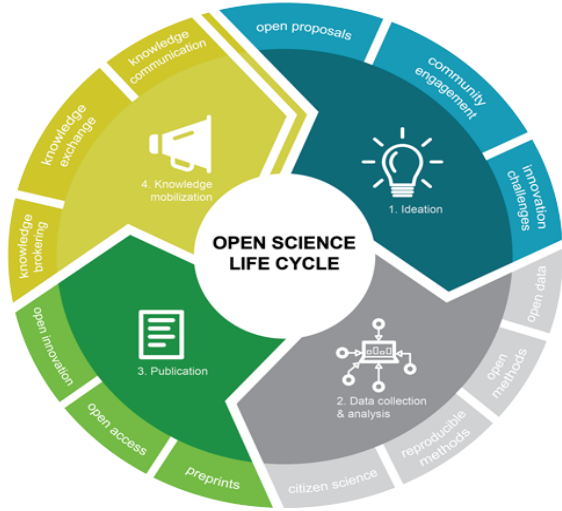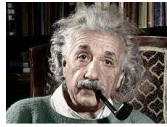

www.icipe.org

icipe
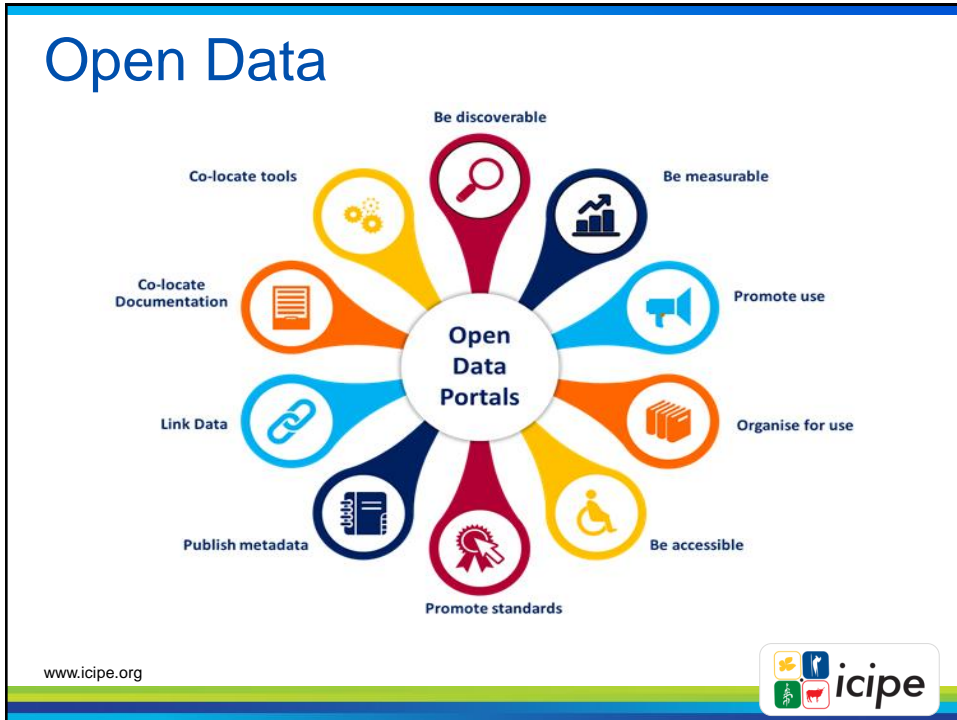
5

Commentary | Open Access | Published: 21 January 2022

Behind every great research project is great data management

Samantha Kanza ✉ & Nicola J. Knight

Abstract

Research data management (RDM) is the cornerstone of a successful research project, and yet it often remains an underappreciated art that gets overlooked in the hustle and bustle of everyday project management even when required by funding bodies. If researchers are to strive for reproducible science that adheres to the principles of FAIR, then they need to manage the data associated with their research projects effectively. It is imperative to plan your RDM strategies early on, and setup your project organisation before embarking on the work. There are several different factors to consider: data management plans, data organisation and storage, publishing and sharing your data, ensuring reproducibility and adhering to data standards. Additionally it is important to reflect upon the ethical implications that might need to be planned for, and adverse issues that may need a mitigation strategy. This short article discusses these different areas, noting some best practices and detailing how to incorporate these strategies into your work. Finally, the article ends with a set of top ten tips for effective research data management.

www.icipe.org

icipe

6

The Master Software and Data Engineering Pipelines

Key:
1. Common ontologies annotated to variables in the DMP, new data from collectors, legacy data, and data from cloud systems
2. Fetching readily available data from other sources e.g. Google engine, satellite, etc.
3. Loading raw data to memory for further cleaning
4. Extraction transformation and loading (ETL) processes: cleaning, profiling and annotation; batch/real-time
5. Load clean and annotated raw data to the data lake for machine learning/modelling/analysis/visualization
6. Data analysis, modelling and machine learning pipelines
7. Execute APIs to archive meta-data, dictionary, protocols, experiment plan, raw data and clean data
8. Execute APIs to summarize data and provide reports/visualization/dashboards
9. Links of external data sources related to *icipe* updated in the data repository

www.icipe.org

7

---

icipe Data Resources

icipe Main website - http://www.icipe.org/

DMMG data resources landing page- http://dmmg.icipe.org/

Links to Data Resources:
- ❑ Services Request - http://dmmg.icipe.org/service_request.html
- ❑ Common Ontology (CO) - http://dmmg-co.icipe.org
- ❑ Data Management Plan (DMP) - http://dmmg-dmp.icipe.org
- ❑ Data Collect (REDCap) - https://redcap.icipe.org/
- ❑ Data Collect (ODK) - https://odk-server.icipe.org/
- ❑ Project Manager Viewer (PMV) - http://dmmg-pmv.icipe.org
- ❑ KOBO Collect (KOBO) - http://kf.dmmg-apps.icipe.org/
- ❑ Data Lake / Common Data / Working Area - [work in progress]
- ❑ Data Warehouse (CKAN) - http://dmmg.icipe.org/dataportal
- ❑ Version Control (GitHub) - https://github.com/icipe-official

www.icipe.org

8

Key points - RDMA

o Data belongs to *icipe*
o FAIR (findable, accessible, interoperable, re-use) principles of data management, a guide to:
  ▪ Open science
  ▪ Reproducible science
o Principle investigators (PIs) are the primary data owners and should strive to complement the data management tools:
  ▪ Data management plan
  ▪ Common ontology
  ▪ Data collectors – ODK, REDCap and KOBO
  ▪ Data lake – [year marked]
  ▪ Data repository/warehouse
  ▪ Source code version control system – GitHub
o Legacy data – in the process of restoring and centralizing institution memory
  o Papers published and related data should at the data repo
o On-going projects – adopt the data management tools
o Cost implications: support development and maintenance

www.icipe.org

icipe

9

Way Forward

o Set up the remaining software components

o Reinforcement of the RDMA

o Collect and populate legacy

o System maintenance and updates

www.icipe.org

icipe

10

11

12