Brief Intro to ITCOOcean

Dr. TVS Udaya Bhaskar

Head, Ocean Data Management (ODM) Division,

INCOIS, MoES, Govt. of India

About ITCOocean

- Approved in Dec 2012 as a part of XII Five Year plan initiatives of INCOIS.
- First set of classes started in July 2013.
- Successfully conducted 62 courses, which includes both National and International.
    - 2 weeks - 5 no., 1 week - 19 no., Few days - 2 no
- Participants from 84 countries (mainly Indian Ocean Rim) participated in the courses.
- A total of 3819 students (2782 Indian, 1037 Foreign nationals) were accommodated in these courses.

Category 2 Centre (C2C)

- ITCOO was Proposed as a UNESCO Category 2 Centre in, 29thSession of IOC Assembly 20-29 June 2017.
- Formally approved in November, 2017 and approved by Cabinet, Government of India.
- After Iran, India is the only C2C centre in IOC category.

Academic Block

Hostel Blocks

Programme Schedule: Fundamentals of Ocean Data Management (August 23 - 27, 2021)

Time	Day 1 (23-08-2021)	Day 2 (24-08-2021)	Day 3 (25-08-2021)	Day 4 (26-08-2021)	Day 5 (27-08-2021)
11:00 to 12:00	About ITCOO and Introduction to Data Management for researchers **(TVS Udaya Bhaskar)**	Introduction to Meta data and standards and data formats **(R Venkat Shesu)**	Introduction to Quality control of ocean data **(TVS Udaya Bhaskar)**	Visualization of data, Generation of gridded products,Introduction to open source S/Ws **(Ravi Kumar Jha)**	Remote sensing data management and applications **(N Srinivasa Rao)**
12:00 to 13:00	Ocean Data Visualization: Overview, applications and web based ODV **(Linta Rose)**	Ocean Data Visualization: Basic plotting **(Linta Rose)**	Ocean Data Visualization: Importing different data types **(Linta Rose)**	Ocean Data Visualization: QC, Interpolation techniques and additional features **(Linta Rose)**	Ocean Data Visualization: Usage of own and example datasets **(Linta Rose)**

63RD

AND COUNTING

Introduction to Data Management for Researchers

Dr. TVS Udaya Bhaskar
Head, Ocean Data Management (ODM) Division,
INCOIS, MoES, Govt. of India

Definition of Data

There is a range of definitions of data:

The data, records, files or other evidence, irrespective of their content or form (e.g. in print, digital, physical or other forms), that comprise research observations, findings or outcomes, including primary materials and analysed data.

Ref: Australia National Data Service (ANDS)

"…determined by the community of interest through the process of peer review and project management." *Ref: National Science Foundation (NSF). 2010*

- Examples (from NSF)

Data	Physical collections
Publications	Software
Samples	Models

Data management

- Data management refers to activities and practices that support long term preservation, access, and use of data.
- Data management activities can include :

Planning

Storing

Documenting

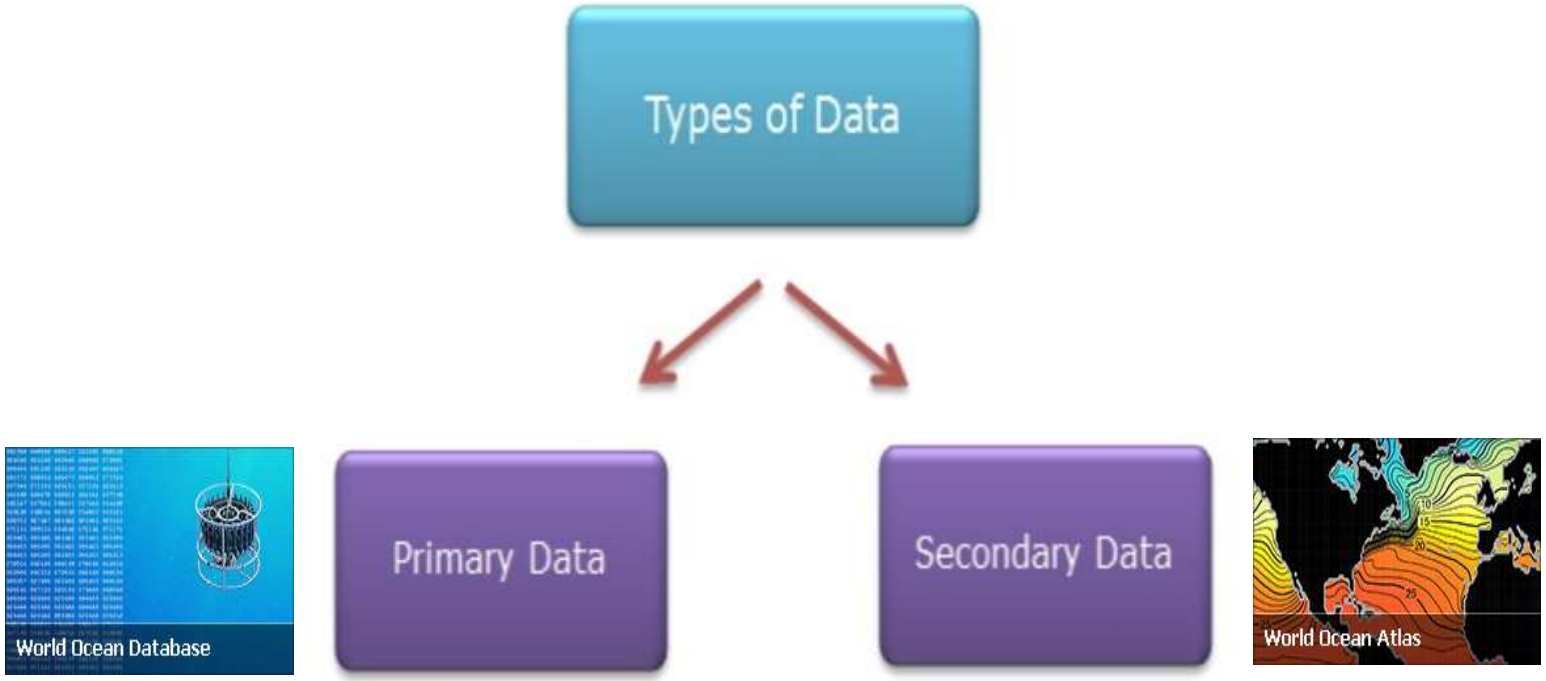
Anonymising

Formatting

Controlling access

Primary & Secondary data

- Primary data is factual and original
- Secondary data is the analysis and interpretation of the primary data
  - Can include reports, conference papers, books, web sites



Why manage data?

- Data management helps researchers do better research:
  - helps researchers optimize use of data during the active phase of the research project.
  - helps them collaborate with other researchers
  - ensures that data is preserved for future researchers to discover, interpret, and reuse.
  - sustains the value of data

Funding agencies

- A growing number of funding agencies, journal publishers, and research institutions require data management
  - Provides transparency in research projects
  - Provides increased return on investment by ensuring the data that is available for secondary analysis, replication, or reuse for future innovations

DST@ 1971-2021 DEPARTMENT OF SCIENCE & TECHNOLOGY

RESEARCH COUNCILS UK

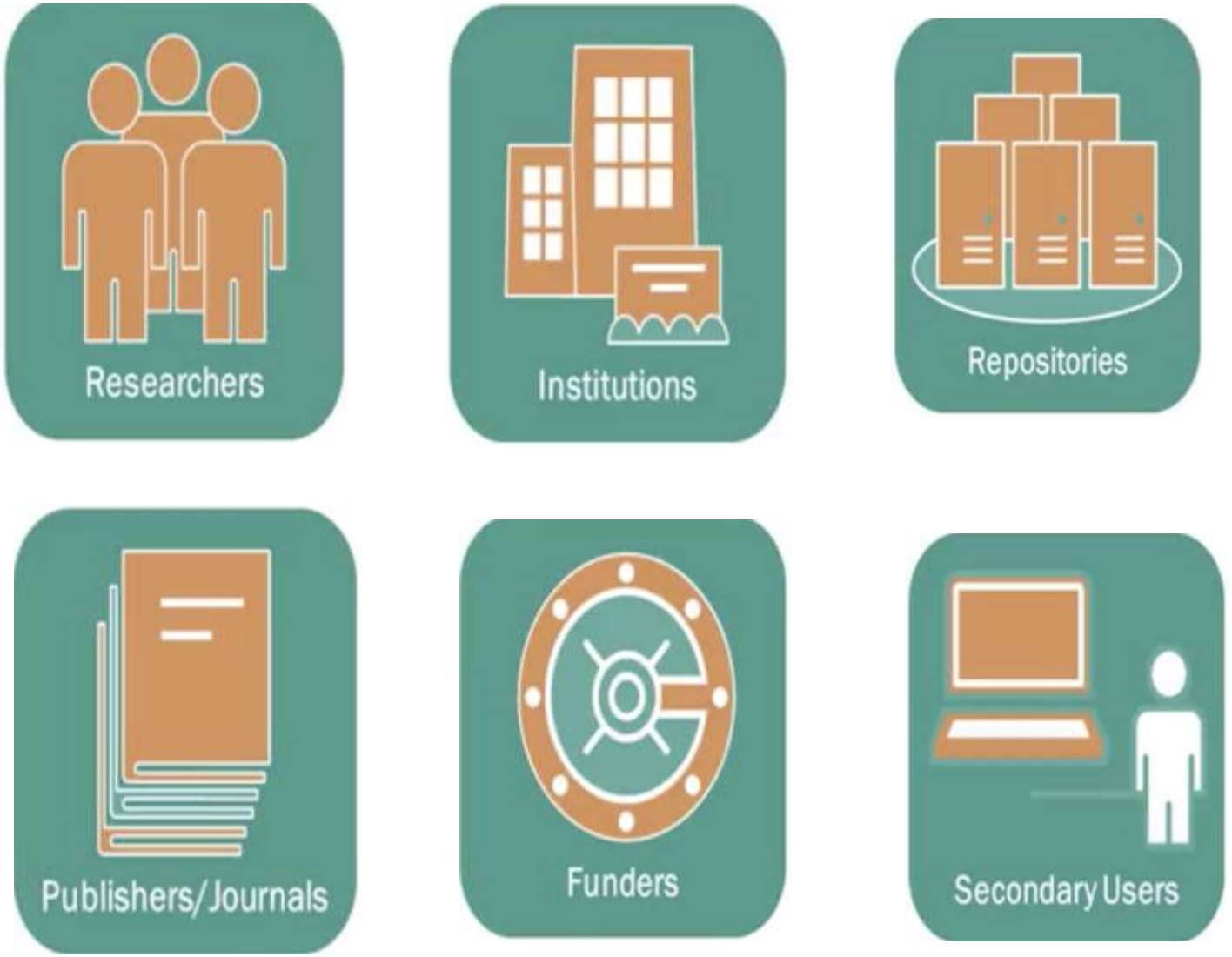
NSF

Australian Government
Australian Research Council

Ministry Of Earth Sciences
Government of India

erc European Research Council

Data management stakeholders



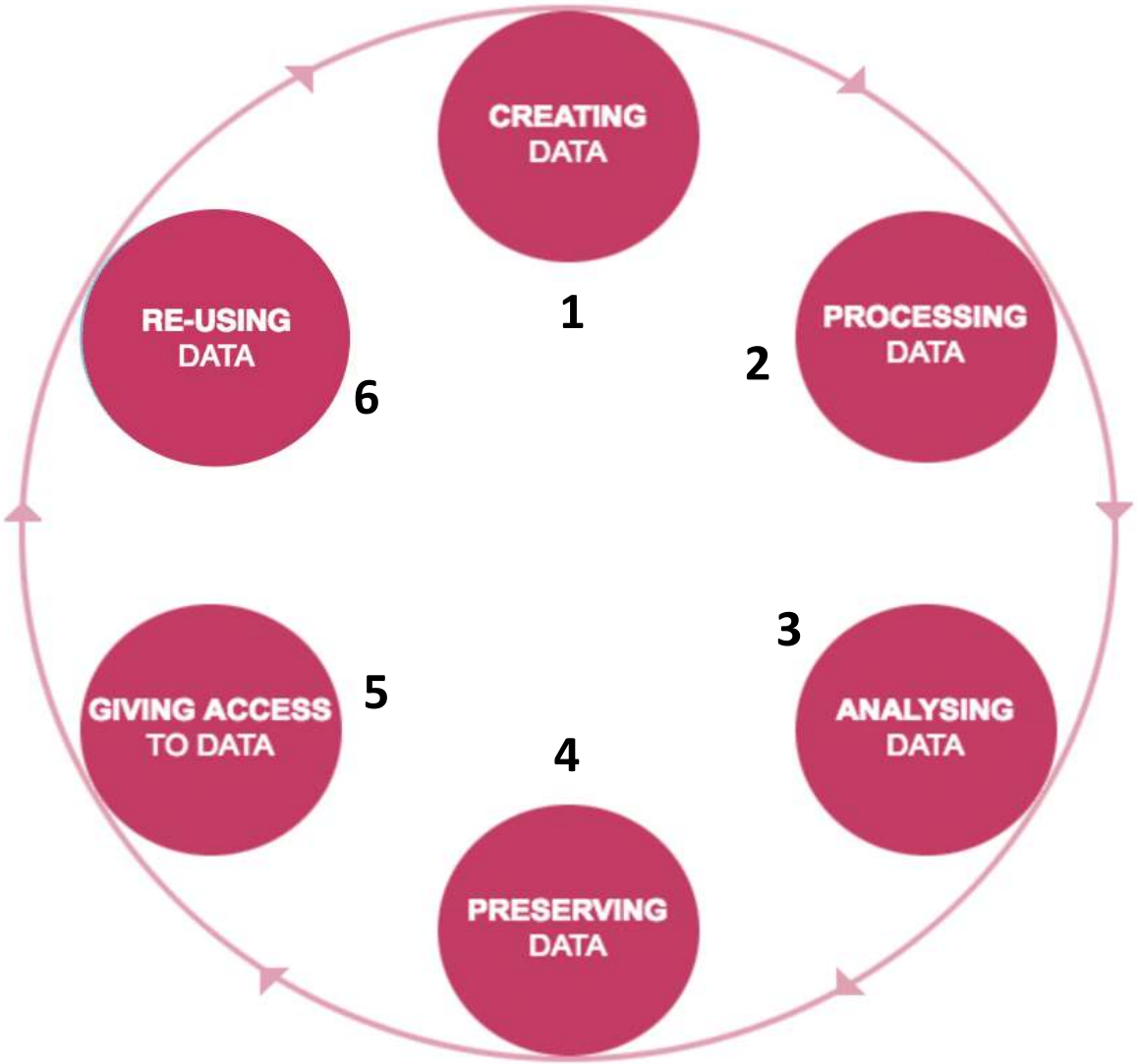
Data management stakeholders

1. Primary researchers, principal investigators
   - design the study
   - specify what data will be collected
   - determine how to analyze the data
   - draw conclusions from these analyses.
2. Institutions (such as universities and research institutes)
   - set data management policies
   - provide data management resources to researchers including:
     - data management training
     - support for researchers writing data management plans
     - services for archiving data
3. Data repositories
   - curate data
   - ensure the long-term preservation of data
   - provide access to data

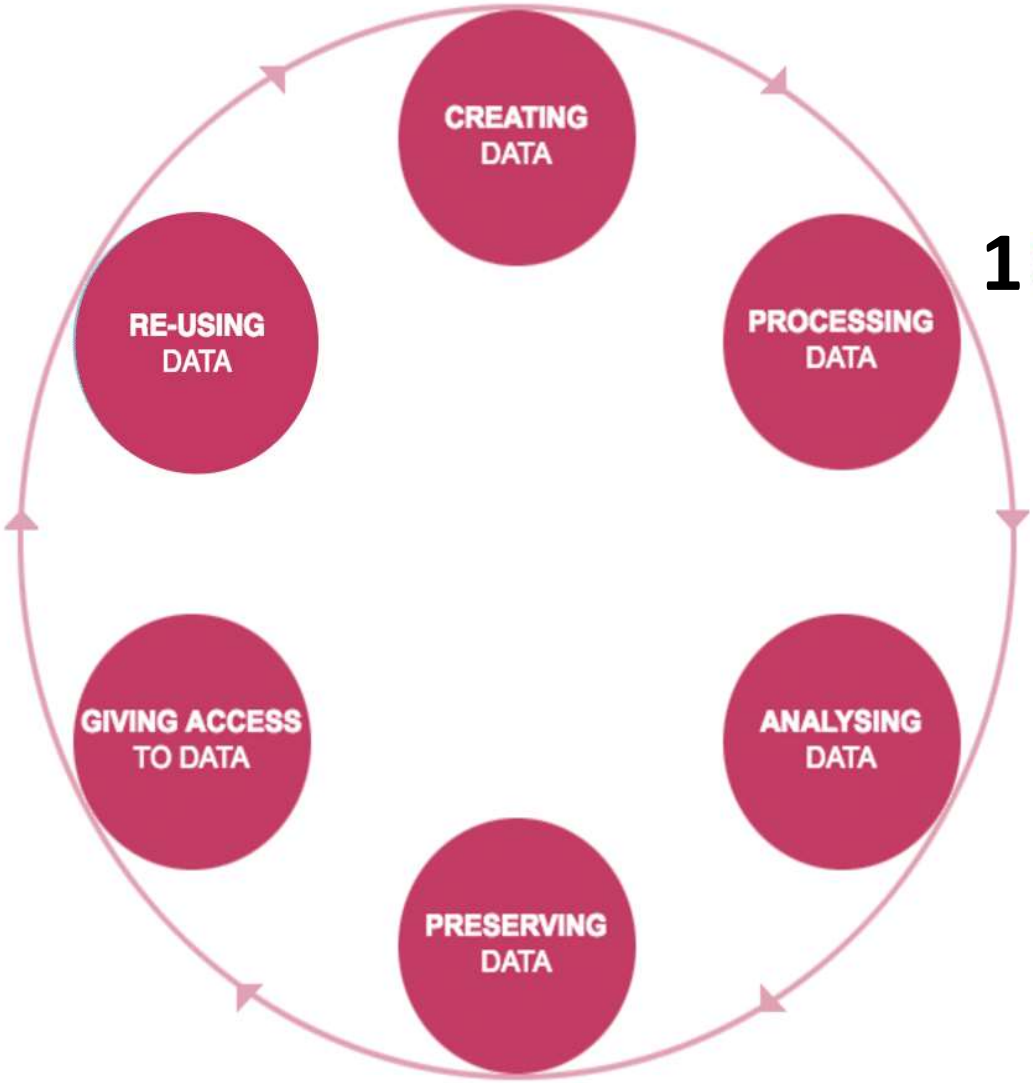
Data management stakeholders

4. Publishers and journals
   - disseminate scientific discovery and maintain the integrity of the scientific record
   - encourage researchers to cite data
   - issue data sharing policies requiring researchers to make data available
5. Funders
   - provide the money to support a research project
   - require researchers to actively and properly manage their data
   - require a data management plan for project proposals
6. Secondary users can use the data
   - to verify published results
   - for secondary analyse
   - for teaching

Research data lifecycle



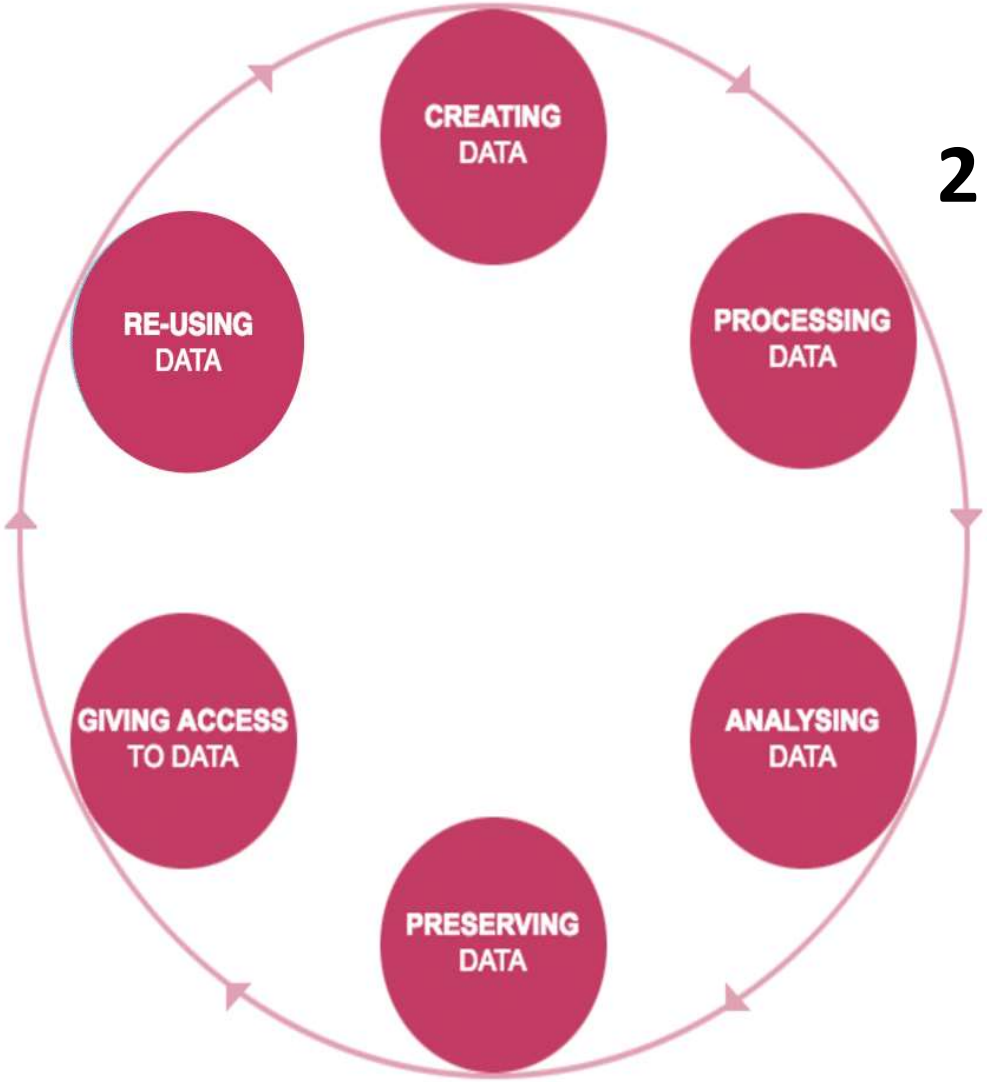
Research data lifecycle



1 CREATING DATA

- design research
- plan data management (formats, storage etc)
- plan consent for sharing
- locate existing data
- collect data (experiment, observe, measure, simulate)
- capture and create metadata

Research data lifecycle



2 PROCESSING DATA

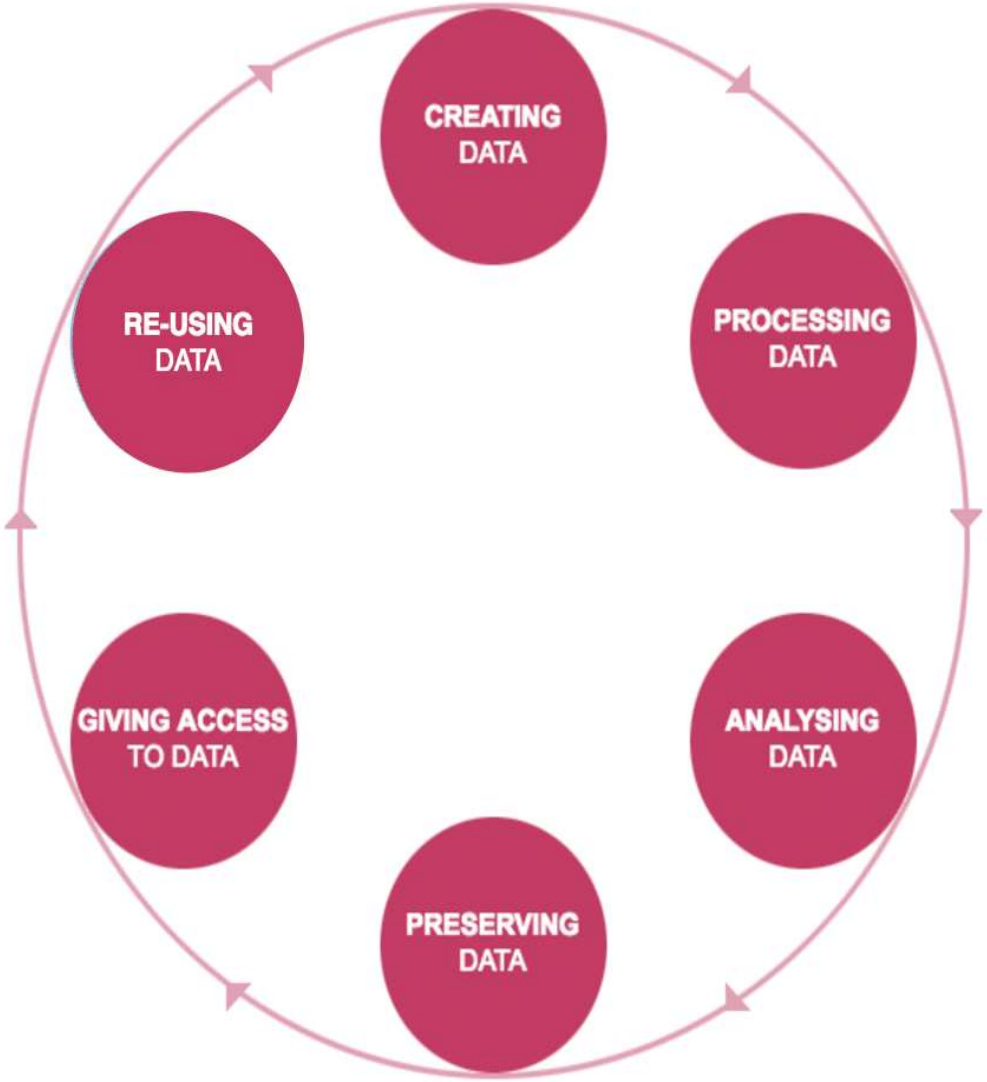
- enter data, digitize, transcribe, translate
- check, validate, clean data
- anonymize data where necessary
- describe data
- manage and store data

Research data lifecycle

3 ANALYZING DATA

- interpret data
- derive data
- produce research outputs
- author publications
- prepare data for preservation

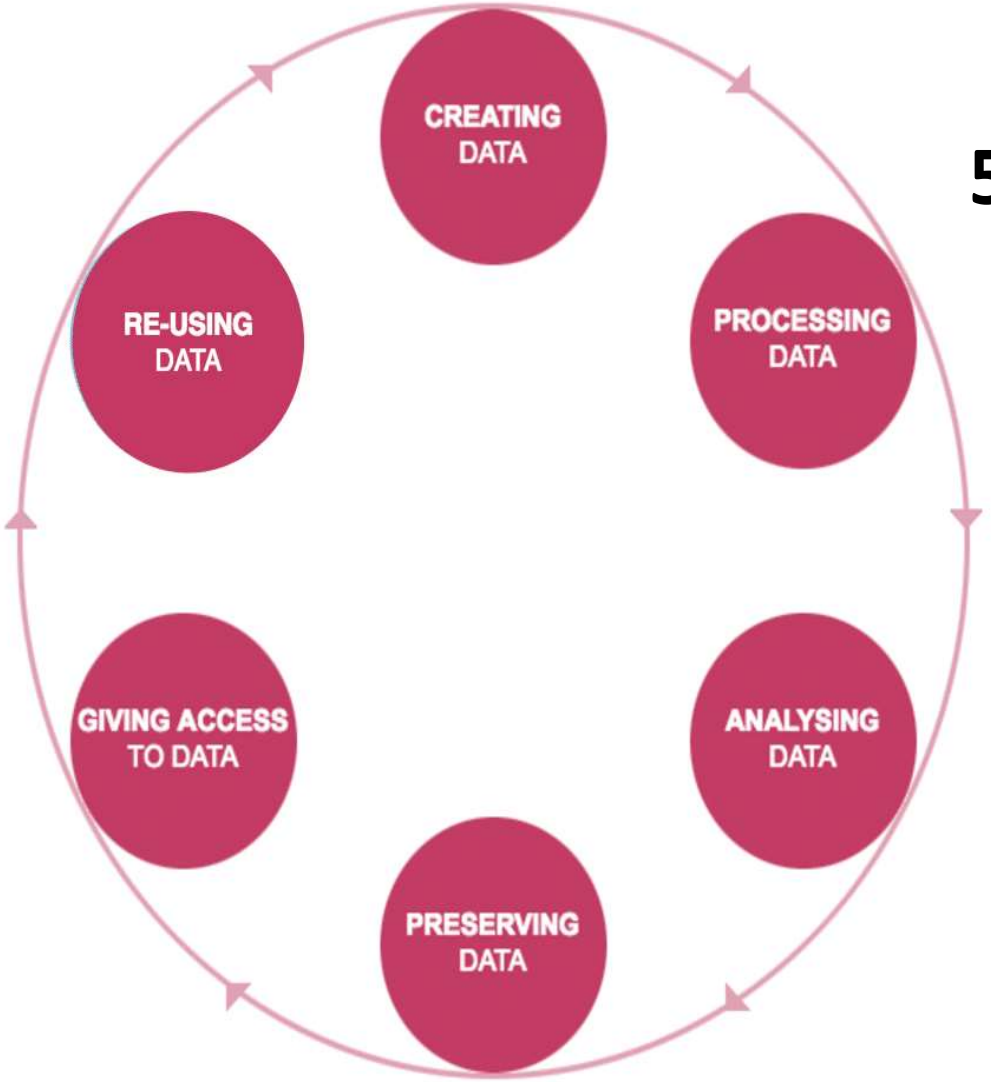
Research data lifecycle



4 PRESERVING DATA

- migrate data to best format
- migrate data to suitable medium
- back-up and store data
- create metadata and documentation
- archive data

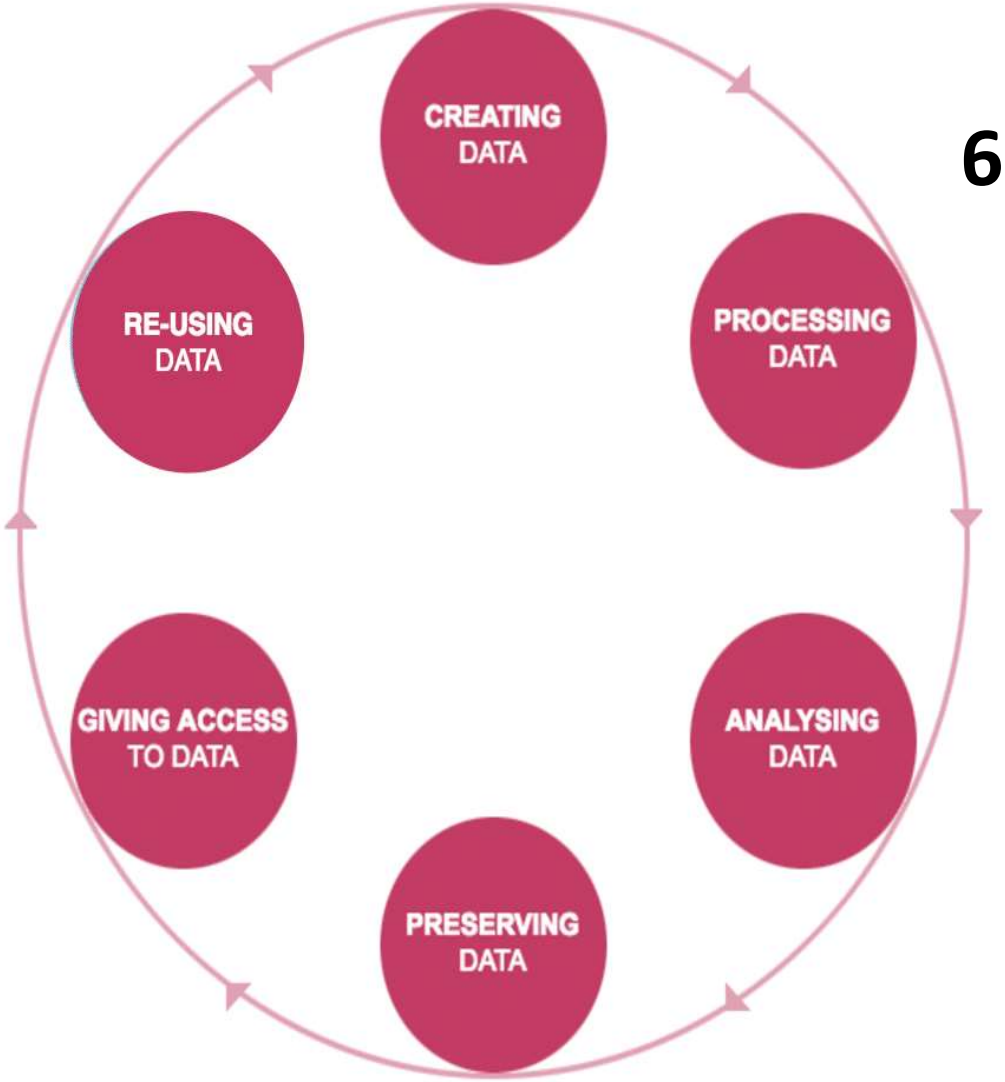
Research data lifecycle



5 GIVING ACCESS TO DATA

- distribute data
- share data
- control access
- establish copyright
- promote data

Research data lifecycle



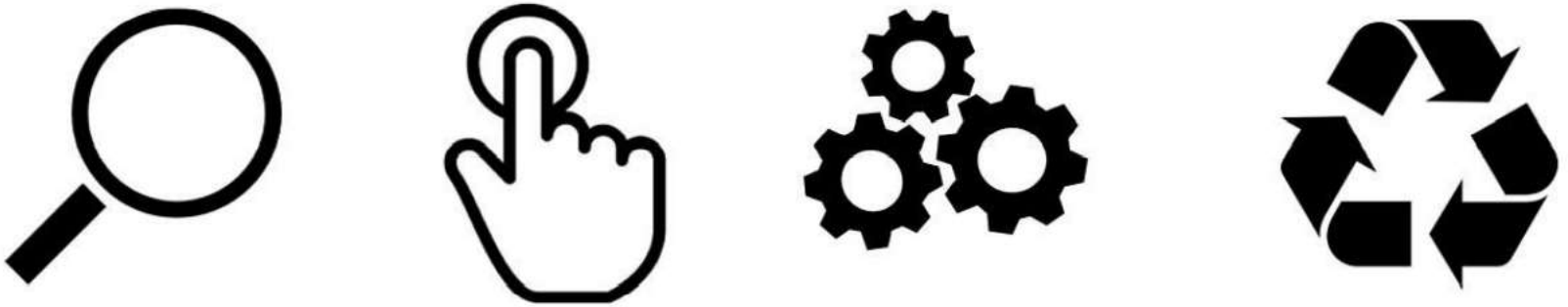
6 RE-USING DATA

- follow-up research
- new research
- undertake research reviews
- scrutinize findings
- teach and learn

FAIR data Services

Enhance the re-use of data: principles

Findable Accessible Interoperable Reusable



Ocean data Mgt Pyramid

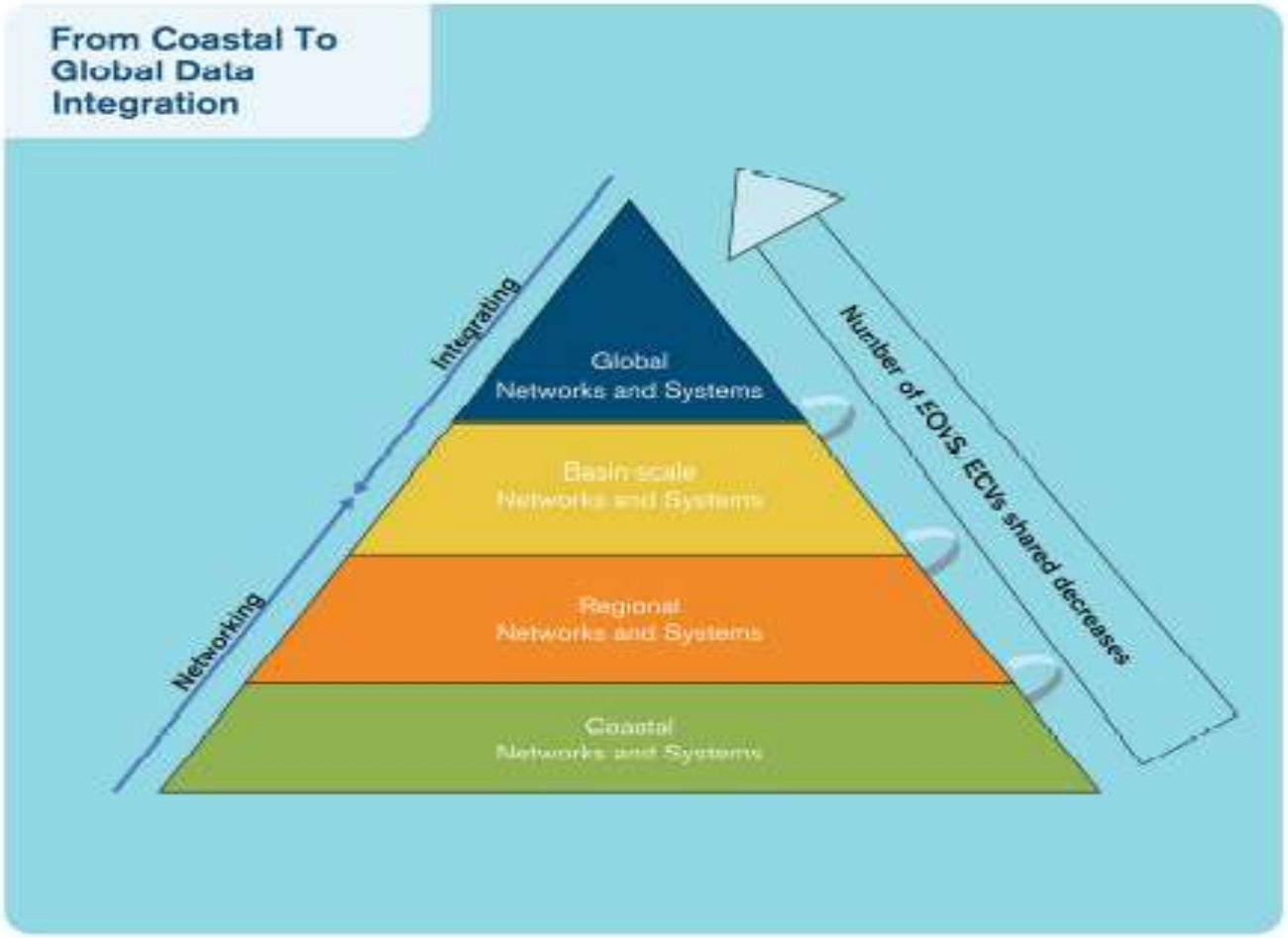
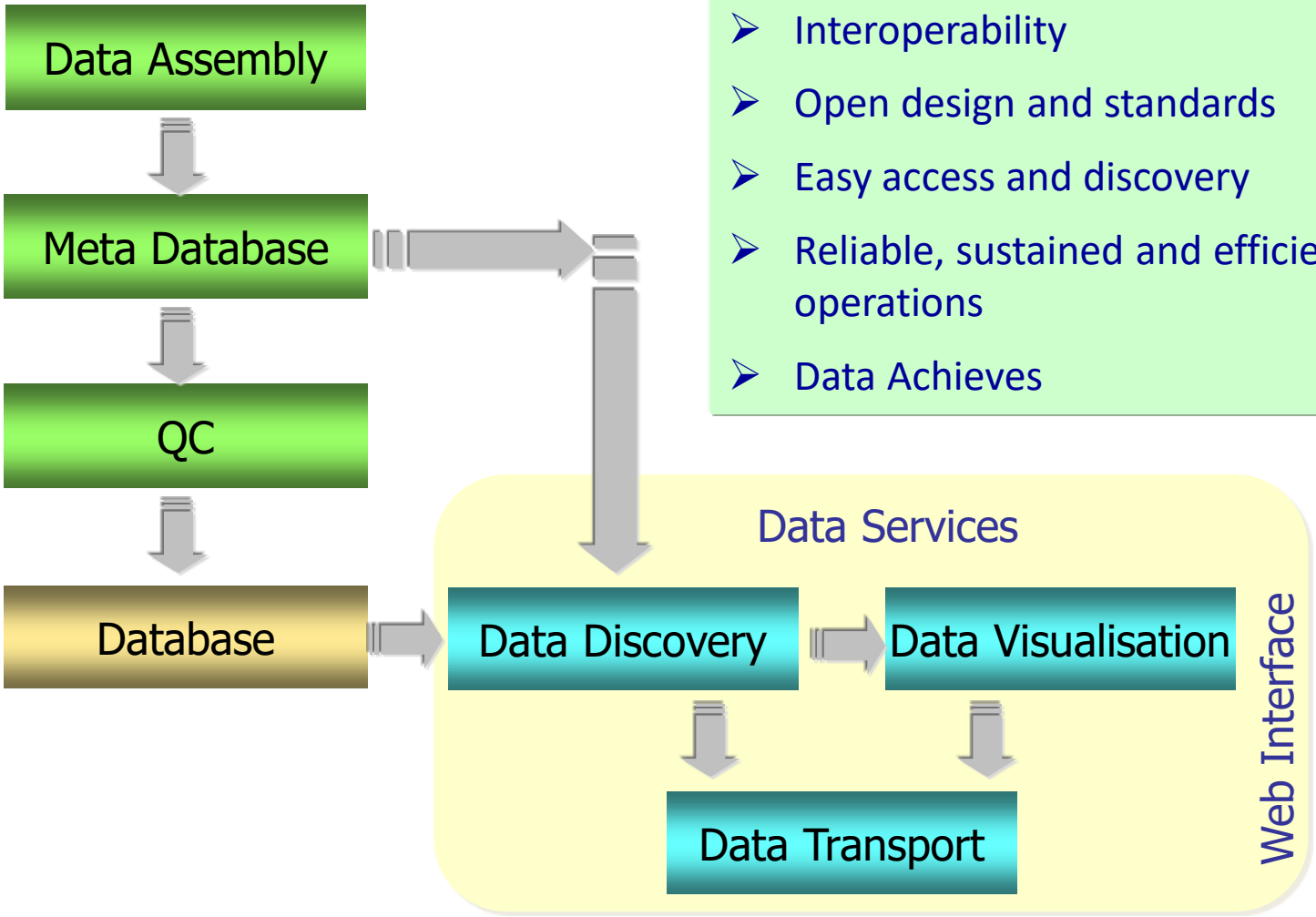


FIGURE 3 | A view of the data management systems ranging from coastal to global data integration.

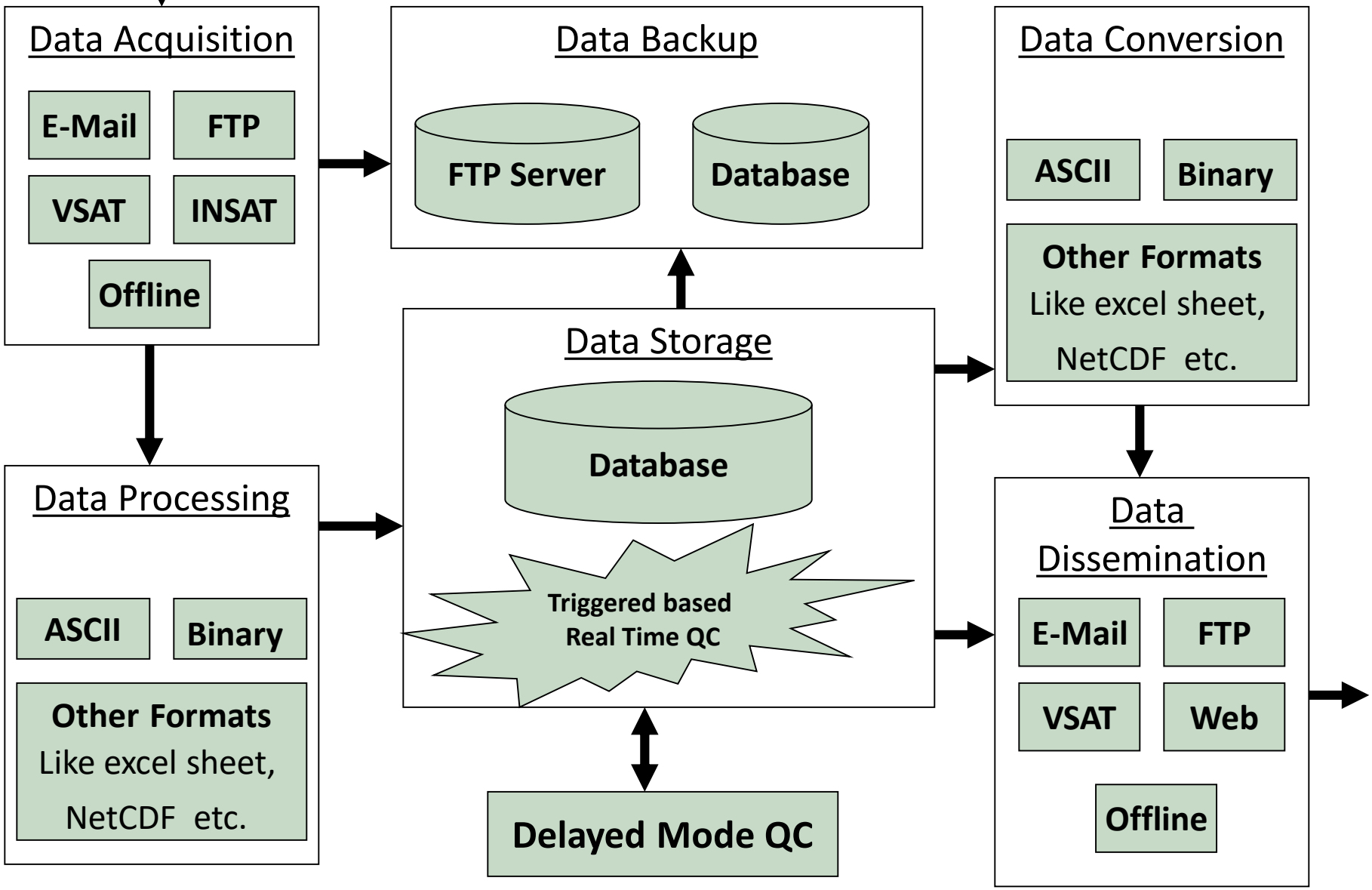
Tanhua T, Pouliquen S et., al. (2019) Ocean FAIR Data Services. Front. Mar. Sci. 6:440. doi: 10.3389/fmars.2019.00440

Components of Data Management



- Interoperability
- Open design and standards
- Easy access and discovery
- Reliable, sustained and efficient operations
- Data Achieves

Managing Data Flow



Data Organization and Standards

What is DOS: This is the best practices for organizing and managing research data, which includes:

- File formats
- File naming
- Folder structures
- Vocabularies

File Formats

Formats

- *Format* refers to the way in which the data is structured and made available for humans and machines
- File formats apply to datasets, documents, images, audio files and video files
- File formats should be considered and decided upon *before* the commencement of data collection
- Where possible, use open formats as they are more likely to be readable in the future and are easier to share with others
- It is usually safe to use a proprietary format if it is very widespread as free programs may exist to read these formats

Examples

Research data come in many different formats including:

- Text
  - Plain text files
  - Documents (e.g. MS Word)
  - Portable Document Format (PDF)
  - Rich Text Format (RTF)
  - Hyper-Text Markup Language (HTML)
  - Extensible Markup Language (XML)
- Numerical
  - MS Excel
  - Flat files: fixed field format files, delimited files
- Multimedia
  - JPEG, TIFF, GIF, PNG, MPEG, Quicktime
- Models
- Software

The Importance of Format Standards

- Standard file formats are essential for effective data sharing
- In many cases a research discipline will have a mandatory or preferred standard for saving and storing research data
- Using appropriate file formats ensures the sustainability of the data and its access and reuse
- Migrating data from an unsuitable format to a more sustainable option is always difficult and expensive
  - In some cases data may be lost or conversion impossible

File and folder naming

- File and folder naming is an important aspect of organizing your data
- A good file name makes it easy to identify, locate and retrieve your data
- With logical and clear file naming you ensure that your data will be understandable for your colleagues, or after a period of time also by yourself
- There is not one recommended way to name your files and folders, but the first rule of thumb is that you should name your files consistently
- If you work as part of a research group, you should decide on a file and folder naming system with your colleagues

Importance of File Naming

- Choosing a logical and consistent way to name and organise your files allows you and others to locate and use them
- Decide on a file naming convention at the start of your project. Useful file names are:
  - consistent
  - meaningful to you and your colleagues
  - allow you to find the file easily
- An agreed naming convention will help provide consistency, resulting in:
  - easier to find and correctly identified files
  - prevent version control problems when working on files collaboratively
  - prevent errors in research

Elements of a File name

- A research project should agree on the following elements of a file name:
  - **Vocabulary** – choose a standard vocabulary for file names, so that everyone uses a common language
  - **Punctuation** – decide on conventions for the use of punctuation symbols, capitals, hyphens, etc.
  - **Dates** – agree on a logical use of dates so that they display chronologically i.e. YYYY-MM-DD
  - **Order** - confirm which element should go first, so that files on the same theme are listed together and can therefore be found easily
  - **Numbers** – specify the amount of digits that will be used in numbering so that files are listed numerically e.g. 01, 002, etc.

File naming conventions

Consider the following examples:

Files names without a naming convention	Files with a naming convention
Test_data_2018	20180503_DOProject_DesignDocument_ODM_v2-01.docx
Project_Data	20180709_DOProject_MasterData_Uday_v1-00.xlsx
Design for project.doc	20180825_DOProject_Ex1Test1_Data_INCOIS_v3-03.xlsx
Lab_work	20180825_DOProject_Ex1Test1_Documentation_ITCOO_v3-03.xlsx
Second_test	20181002_DOProject_Ex1Test2_Data_Bhaskar_v1-01.xlsx
Meeting Notes Oct 23	20171023_DOProject_ProjectMeetingNotes_Kumar_v1-00.docx

Ref: Purdue Libraries. (2016). Data Management for Undergraduate Researchers: File Naming Conventions. Retrieved from http://guides.lib.purdue.edu/c.php?g=353013&p=2378293

Folder Structure for Data Files

- The structure of folders/directories for organizing research data files should have a clear, documented naming convention
- The top-level folder or directory should include the project title, unique identifier, and date (year)
- Folders within the sub-structure should be divided by a common theme
- There is generally “no right or wrong way” to create a folder structure
  - Must be stable, scalable and consistent

Example of sample cruise

- Sagar Kanya Cruise – 310
  - Argo floats
    - Metadata
    - Sensor information
    - Sample CTD collection
  - CTD
    - Metadata
    - Water samples
  - Core
    - Metadata
    - Core samples
  - Meteorological data
    - Metadata
    - Sensor details
  - XBT probes
    - Metadata
    - Sensor information

Summary

- Understand the importance of managing data as a part of responsible research
- Identifying the various data management tasks within the research lifecycle
- understanding data in the context of the research data lifecycle.
- File formats
  - Decided upon *before* the commencement of data collection
  - Where possible, use open formats
- Design how data files will be named and organised
  - This will avoid problems with duplicate names or identity confusion
  - Naming schemes should be descriptive, unique, and reflect the content