#CSIC

Ciclo de Vida de la Investigación

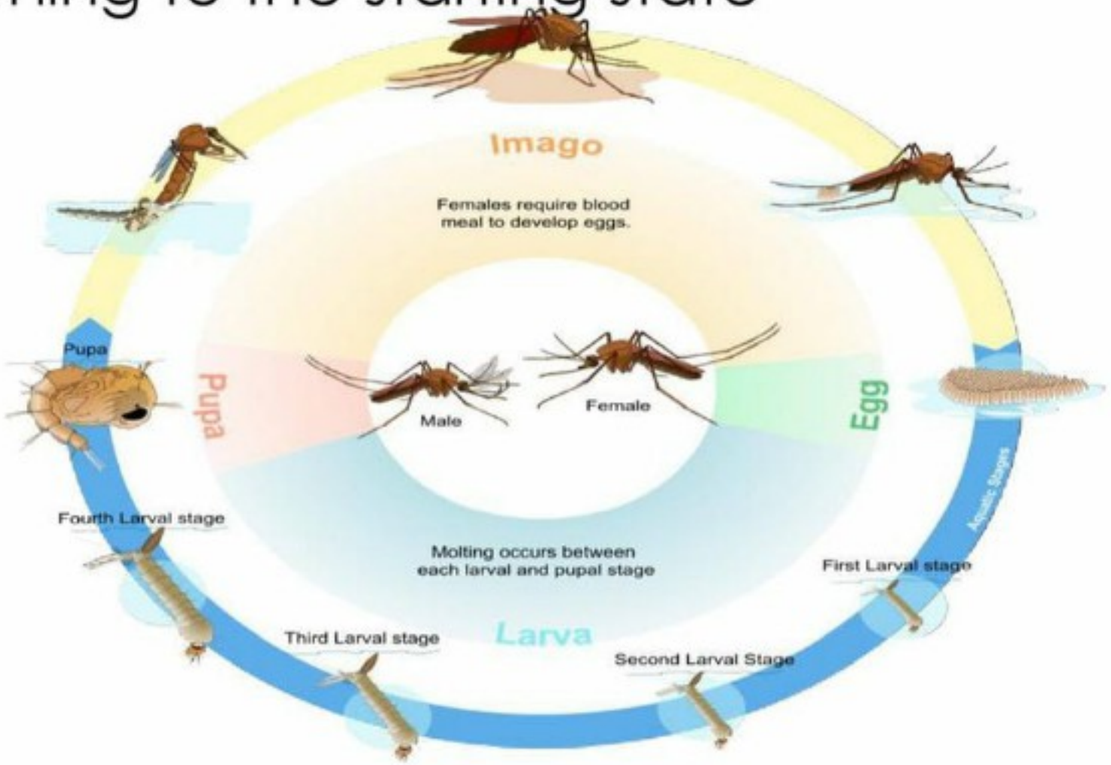
Fernando Aguilar
aguilarf@ifca.unican.es

CSIC
Consejo Superior de Investigaciones Científicas

Life cycle: a biological concept

#CSIC

"In biology, a life cycle is a series of changes in form that an organism undergoes, returning to the starting state"



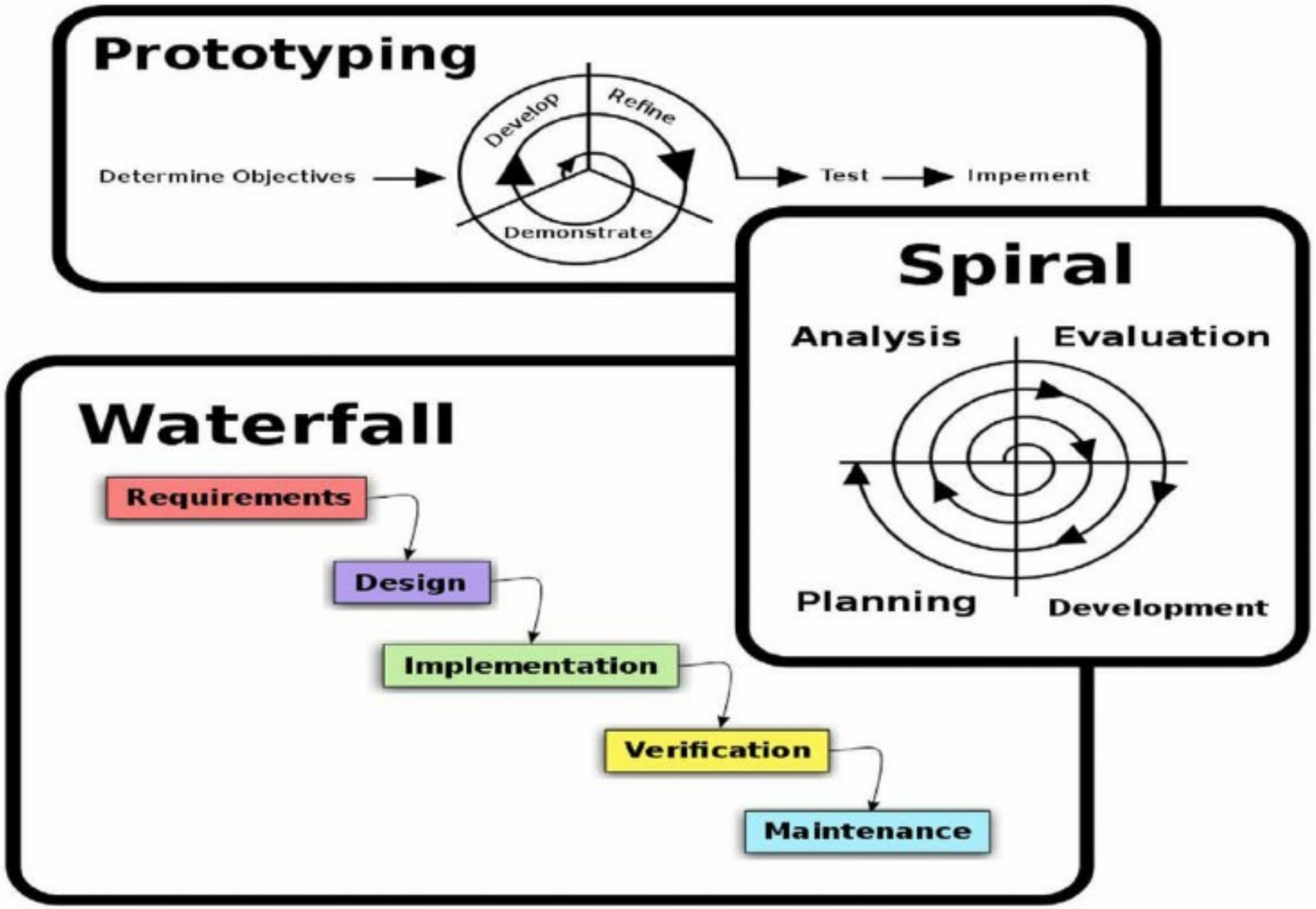
From https://en.wikipedia.org/wiki/Biological_life_cycle

Author: Mariana Ruiz Villarreal LadyofHats
Public domain

CSIC
Consejo Superior de Investigaciones Científicas

Life cycle: Software

#CSIC



A concept extended to products and data

#CSIC

From Merriam Webster:

- *"a series of stages through which something (such as individual, culture, or manufactured product) passes during his lifetime."*

From Cambridge dictionary:

- *"the length of time that something lasts or can be used."*

This applies to digital entities or data

- Evolution of the data description
- Need to change format as technology evolves
- Obsolescence of data

CSIC
Consejo Superior de Investigaciones Científicas

"Elementos" o "Componentes" en investigación

Investigación: ¿Ciclo?

#CSIC

Elementos en Investigación

- Problemas, retos, soluciones, teorías…
- Resultados: Informes, papers, libros
- Información: **datos**, análisis.
- Soporte: Software, recursos, etc.

Algunos elementos reutilizables:

- Papers, libros: punto de partida de nuevas investigaciones.
- Software: reutilización para el análisis.
- **Datos**: derivados, nuevos usos.

CSIC
Consejo Superior de Investigaciones Científicas

¿Cuál es tu "campo"?

Supporting curation with Data Life Cycle

#CSIC

- Concept of digital curation (Pennock, 2007)
  - Maintaining the authenticity and integrity of a trusted body of digital information for current and future use
  - Adding value to the body of digital information
- A lifecycle approach ensures to identify and plan the necessary data management stages (Higgins, 2008)
- Provide a structure for considering the many operations that will need to be performed on a data record throughout its life (Ball, 2012)

Pennock M. 2007 Digital curation: a life cycle approach to managing and preserving usable digital information.Library and Archives Journal, Issue 1
Higgins S. 2008 The DCC Curation Lifecycle Model, the International Journal of Digital Curation, Issue 1, Volume 3
Ball A. 2012 Review of Data Management Lifecycle Models. University of Bath (unpublished)

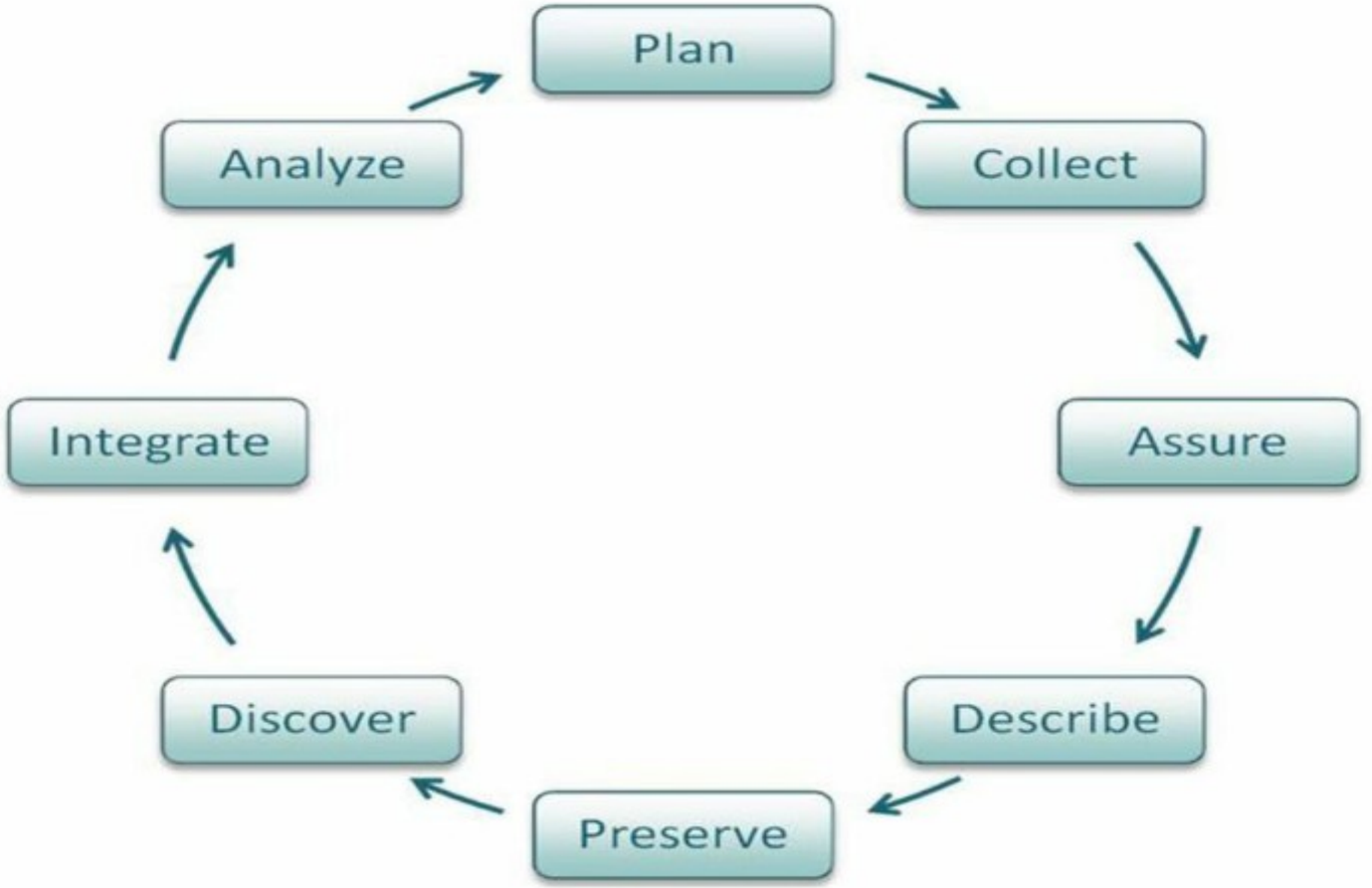
DLC: Approaches

#CSIC

- Data One DLC
- UK Data Archive
- Data Curation Centre
- INDIGO-DataCloud Project

#CSIC

The Data One DLC

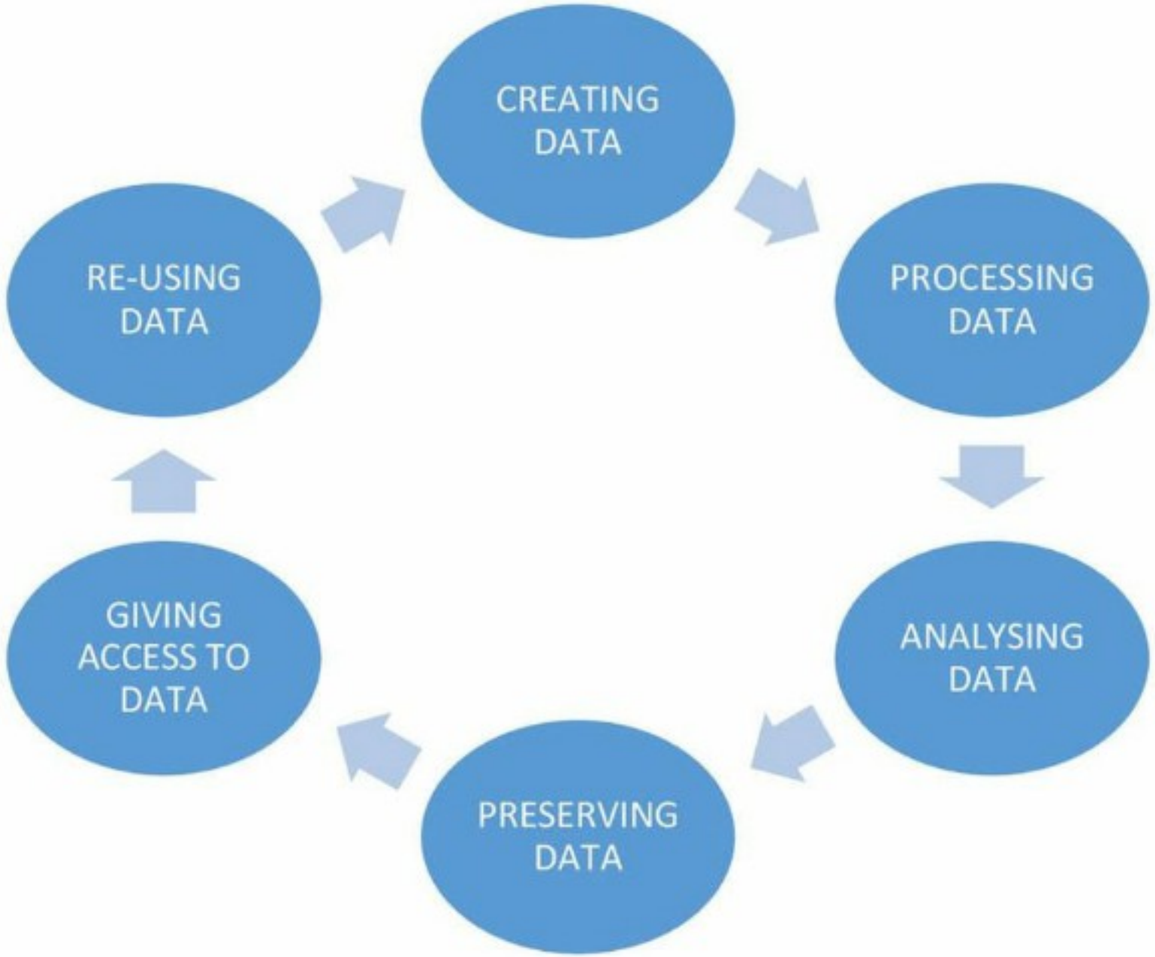


https://www.dataone.org/data-life-cycle

CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

UK Data Archive DLC

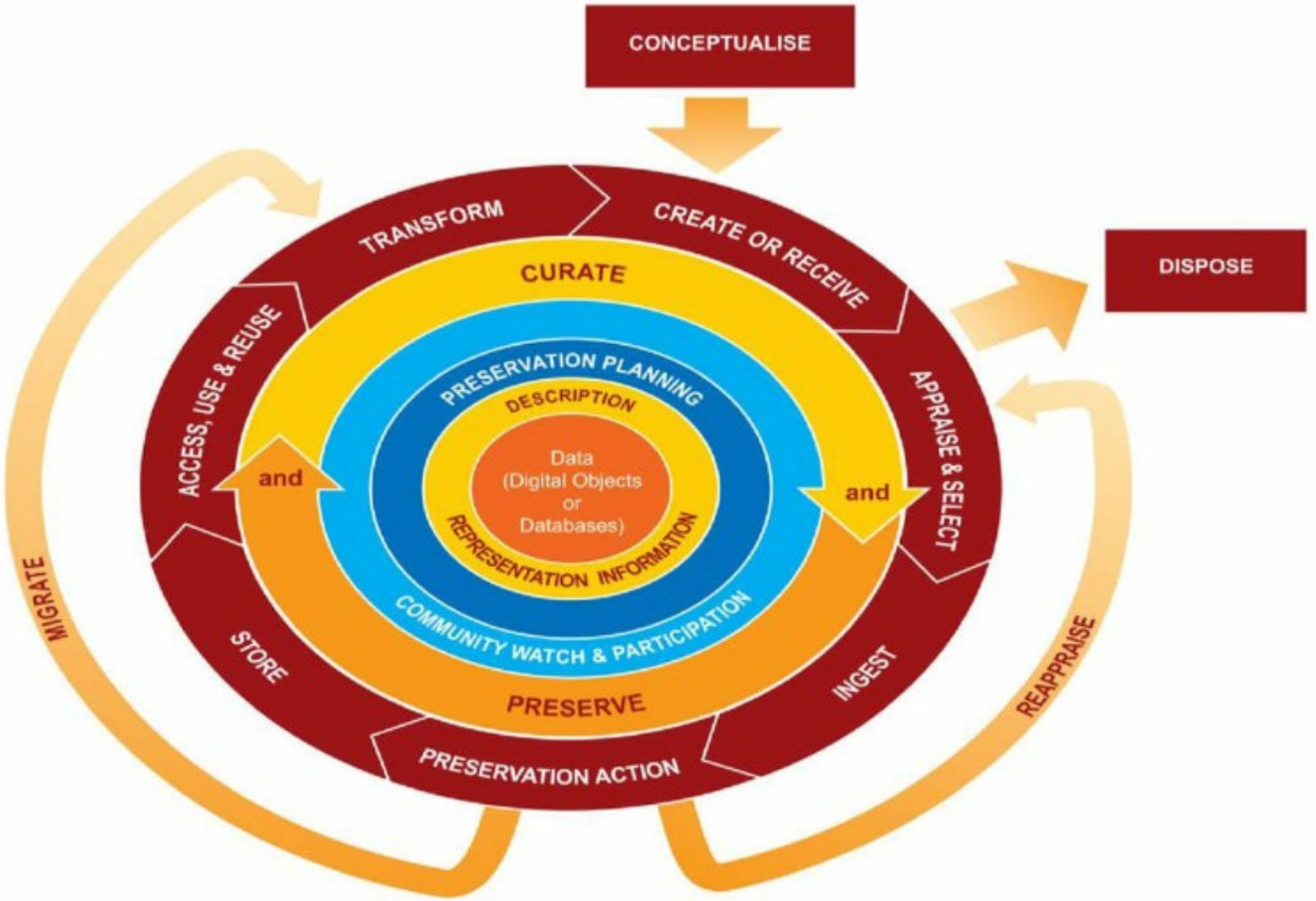


Ref: UK Data Archive: http://www.data-archive.ac.uk/create-manage/life-cycle

CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

Digital Curation Center DLC



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

CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

FULL LIFE CYCLE ACTIONS

DATA

Data, any information in binary digital form, is at the centre of the Curation Lifecycle.

#CSIC

FULL LIFE CYCLE ACTIONS

DATA

Data, any information in binary digital form, is at the centre of the Curation Lifecycle.

Digital Objects

Simple digital objects :
discrete digital items such as text files, image files or sound files, along with their related identifiers and metadata)

Complex digital objects :
Discrete digital objects made by combining a number of other digital objects, such as websites or datasets.

Databases

Structured collections of records or data stored in a computer system.

CSIC
Consejo Superior de Investigaciones Científicas

FULL LIFE CYCLE ACTIONS

#CSIC

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.

FULL LIFE CYCLE ACTIONS

#CSIC

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.

#CSIC

FULL LIFE CYCLE ACTIONS

Community Watch and Participation

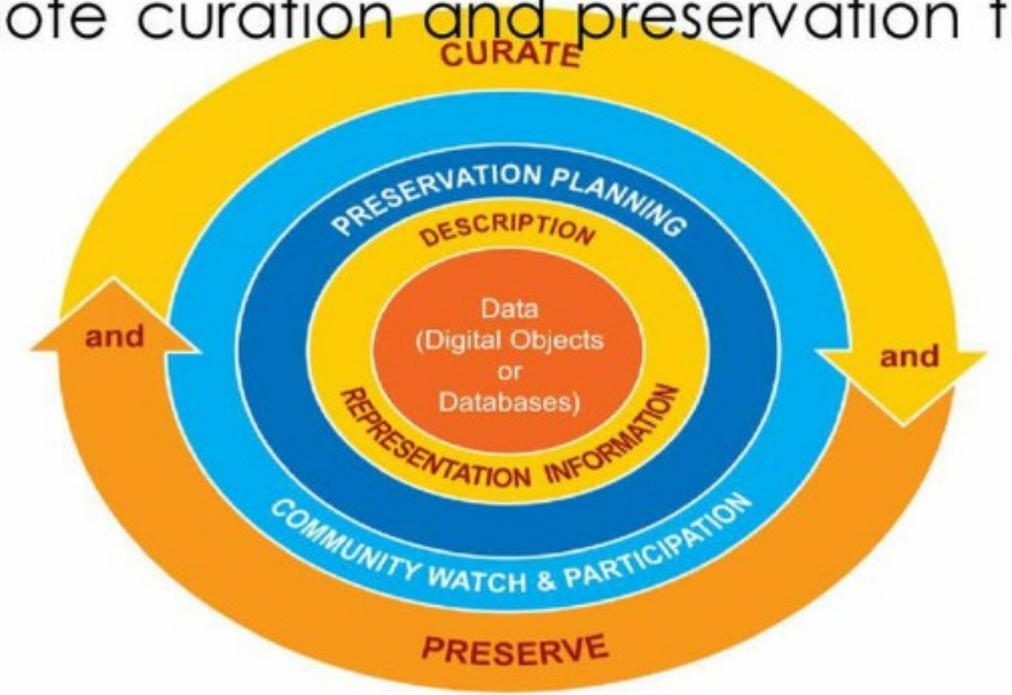
Maintain a watch on appropriate community activities, and participate in the development of shared standards, tools and suitable software.

FULL LIFE CYCLE ACTIONS

#CSIC

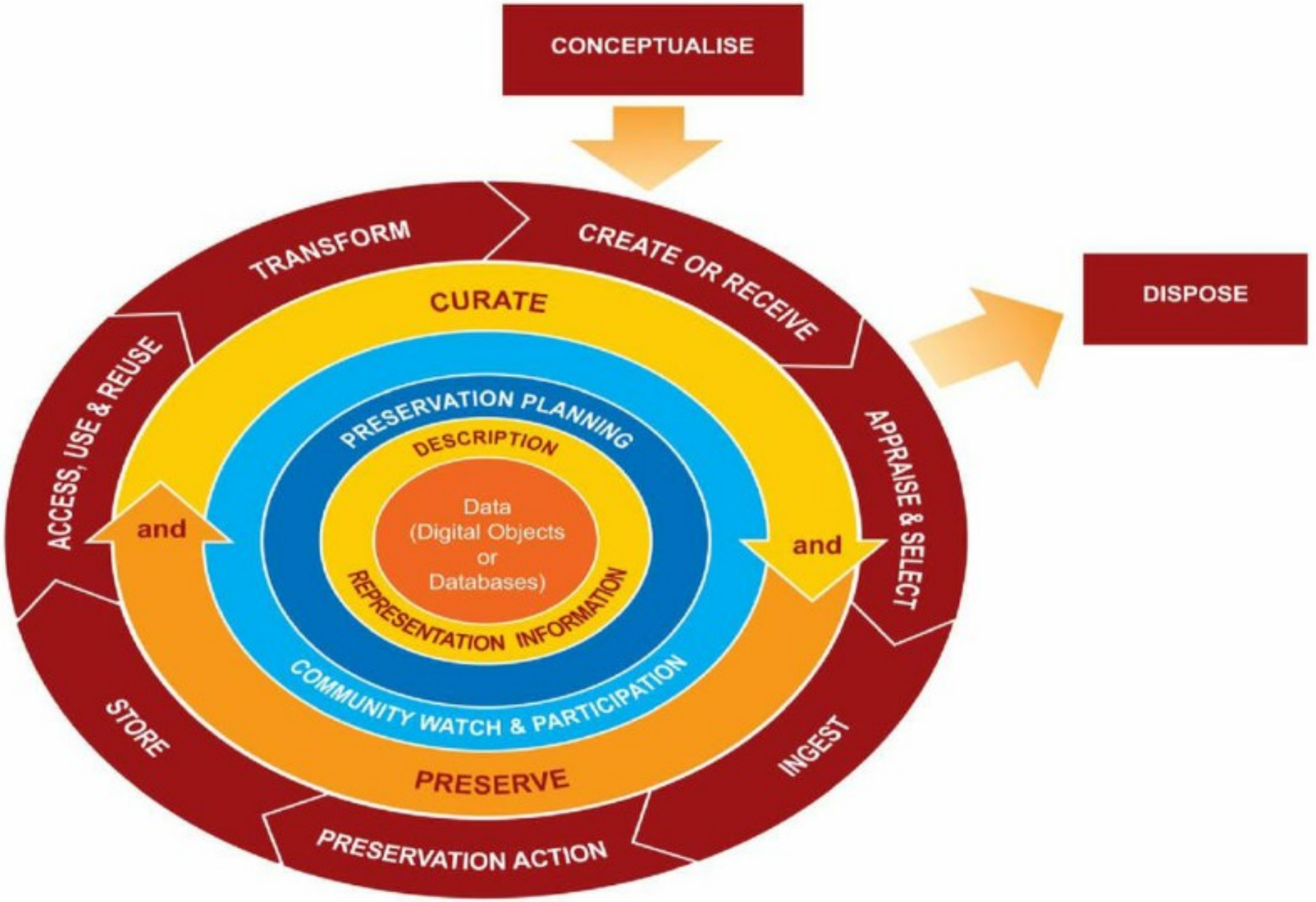
Curate and Preserve

Be aware of, and undertake management and administrative actions planned to promote curation and preservation throughout the curation lifecycle.



#CSIC

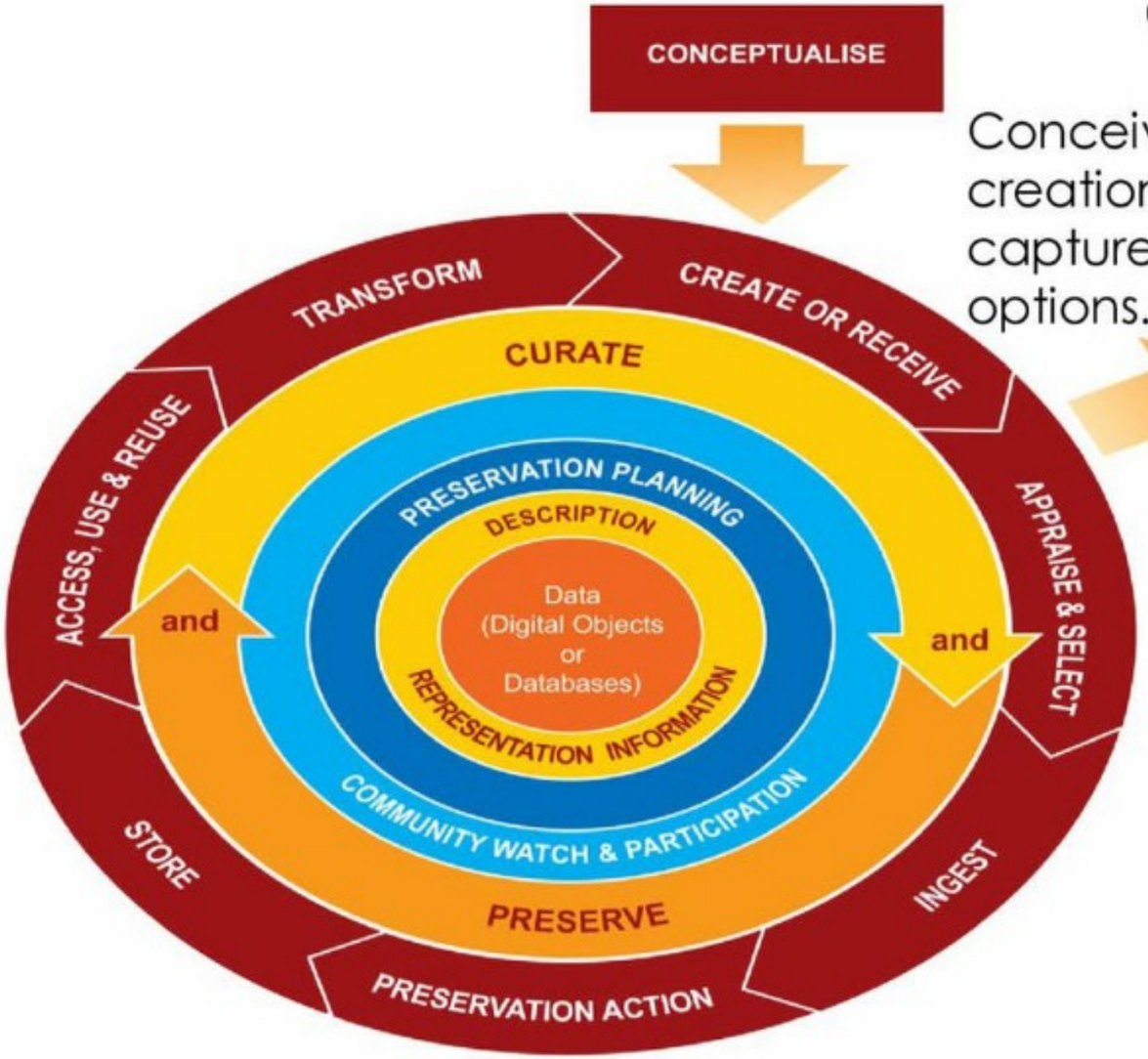
SEQUENTIAL ACTIONS



CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

SEQUENTIAL ACTIONS



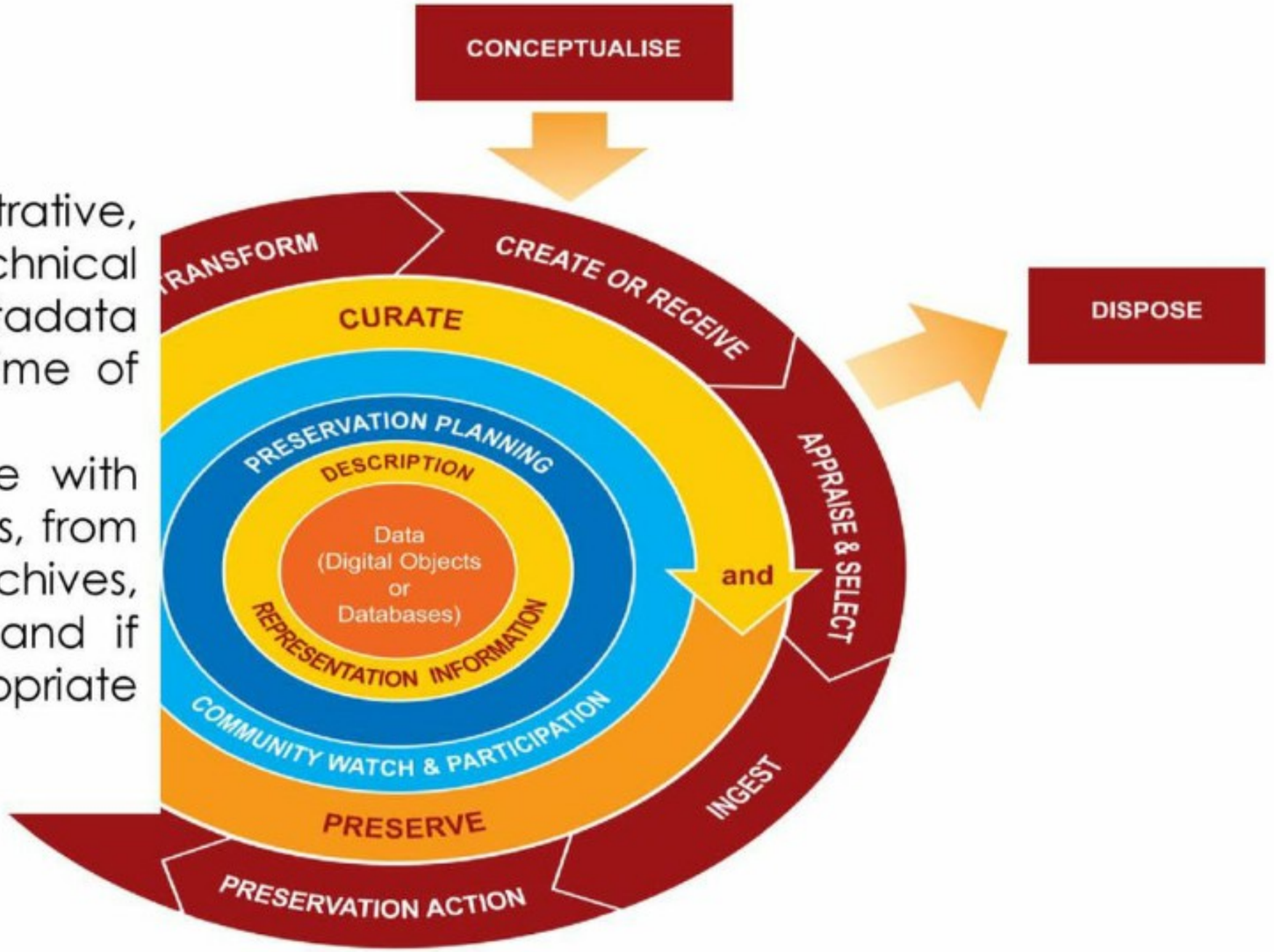
Conceptualise

Conceive and plan the creation of data, including capture method and storage options.

Create or Receive

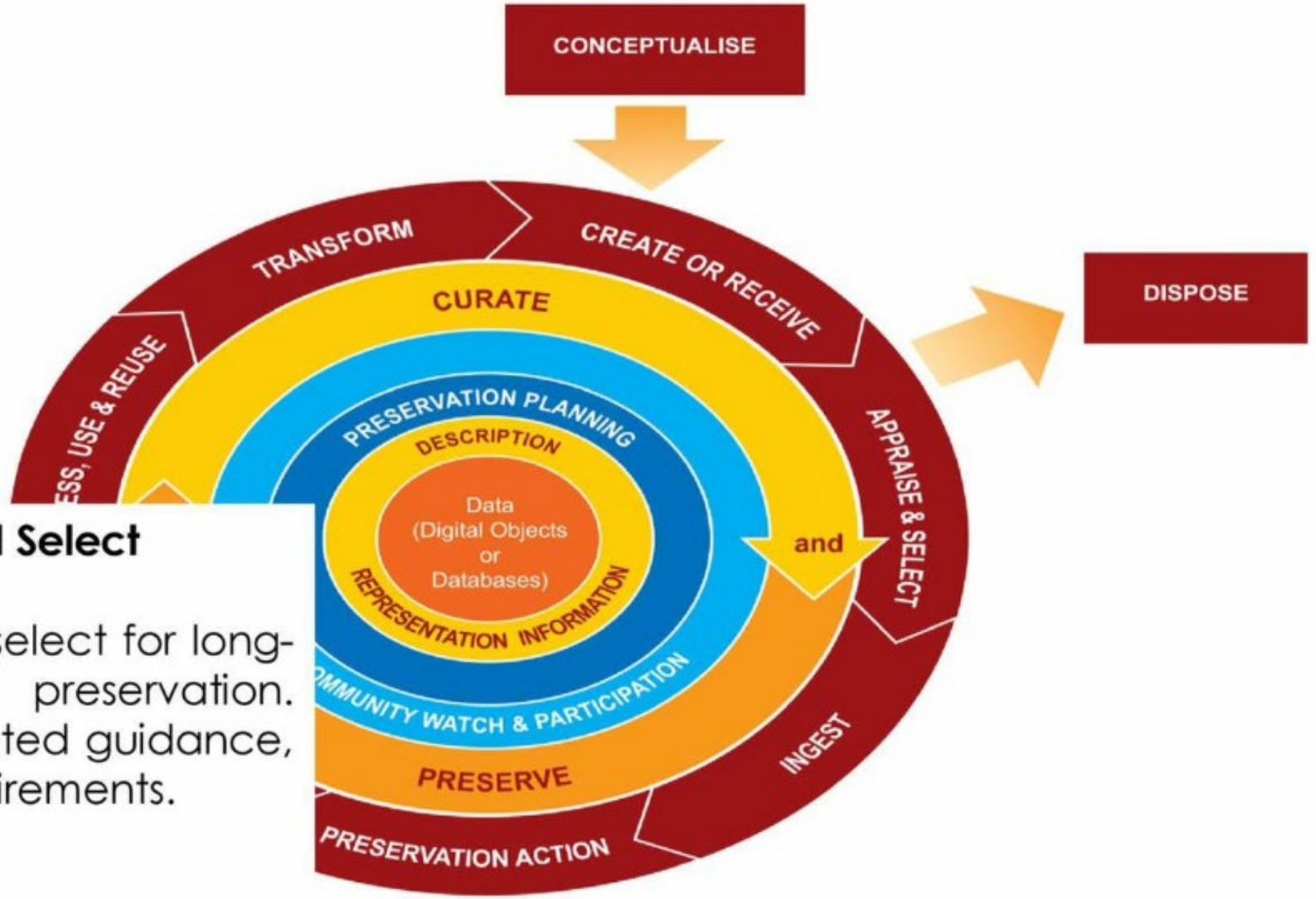
Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation.

Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centres, and if required assign appropriate metadata.



#CSIC

SEQUENTIAL ACTIONS



Appraise and Select

Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.

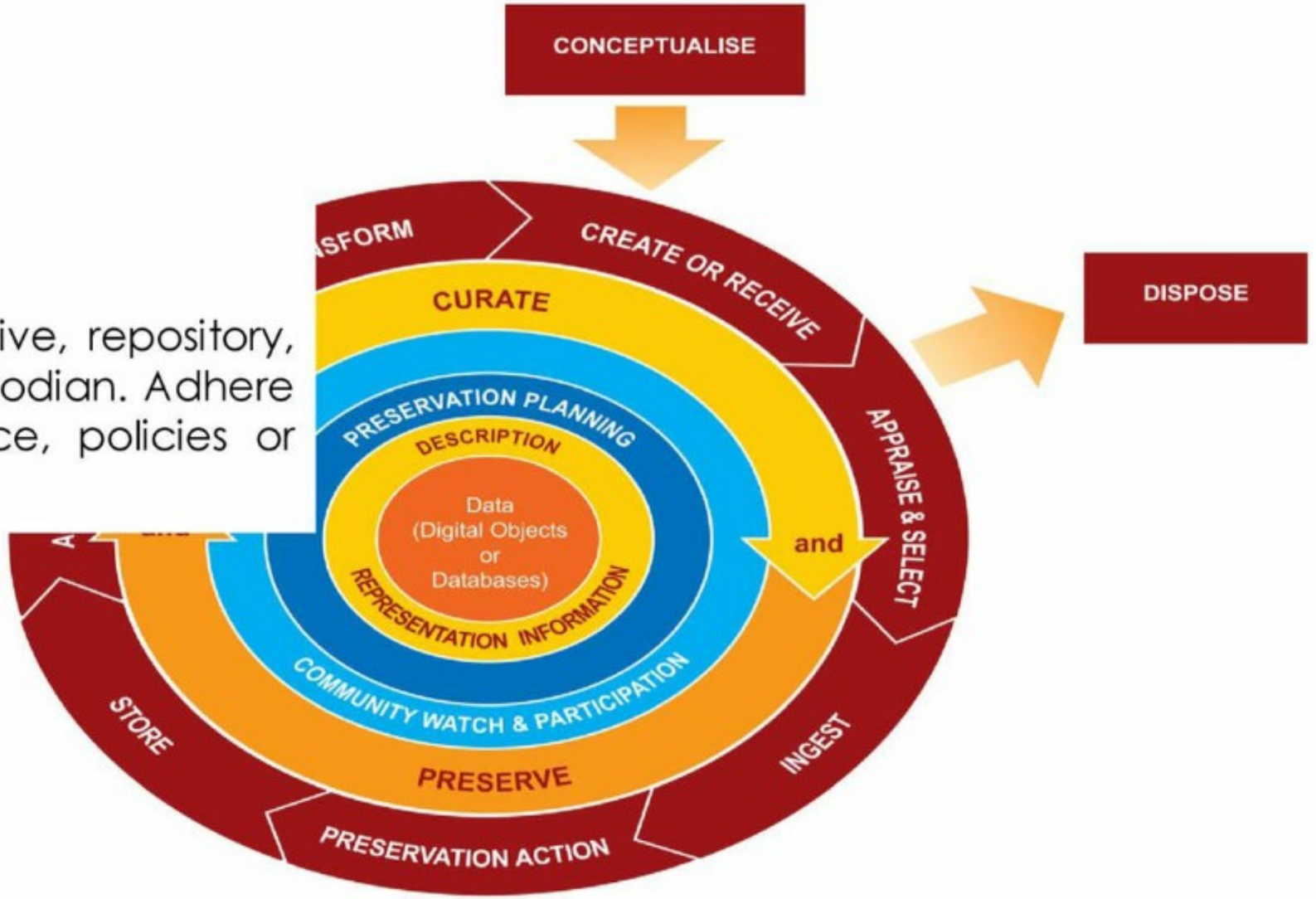
CSIC
Consejo Superior de Investigaciones Científicas

SEQUENTIAL ACTIONS

#CSIC

Ingest

Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.



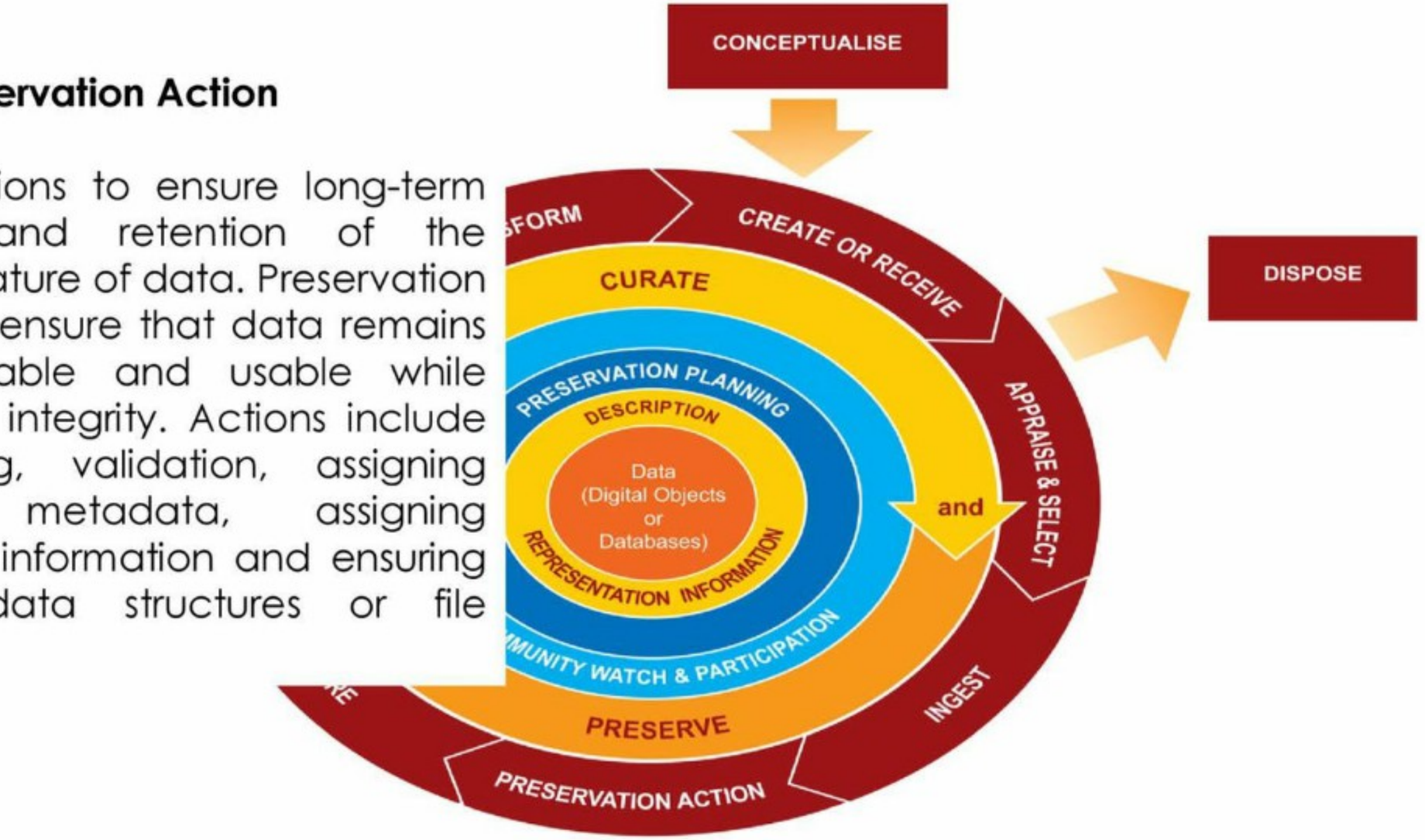
http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf

CSIC
Consejo Superior de Investigaciones Científicas

SEQUENTIAL ACTIONS

Preservation Action

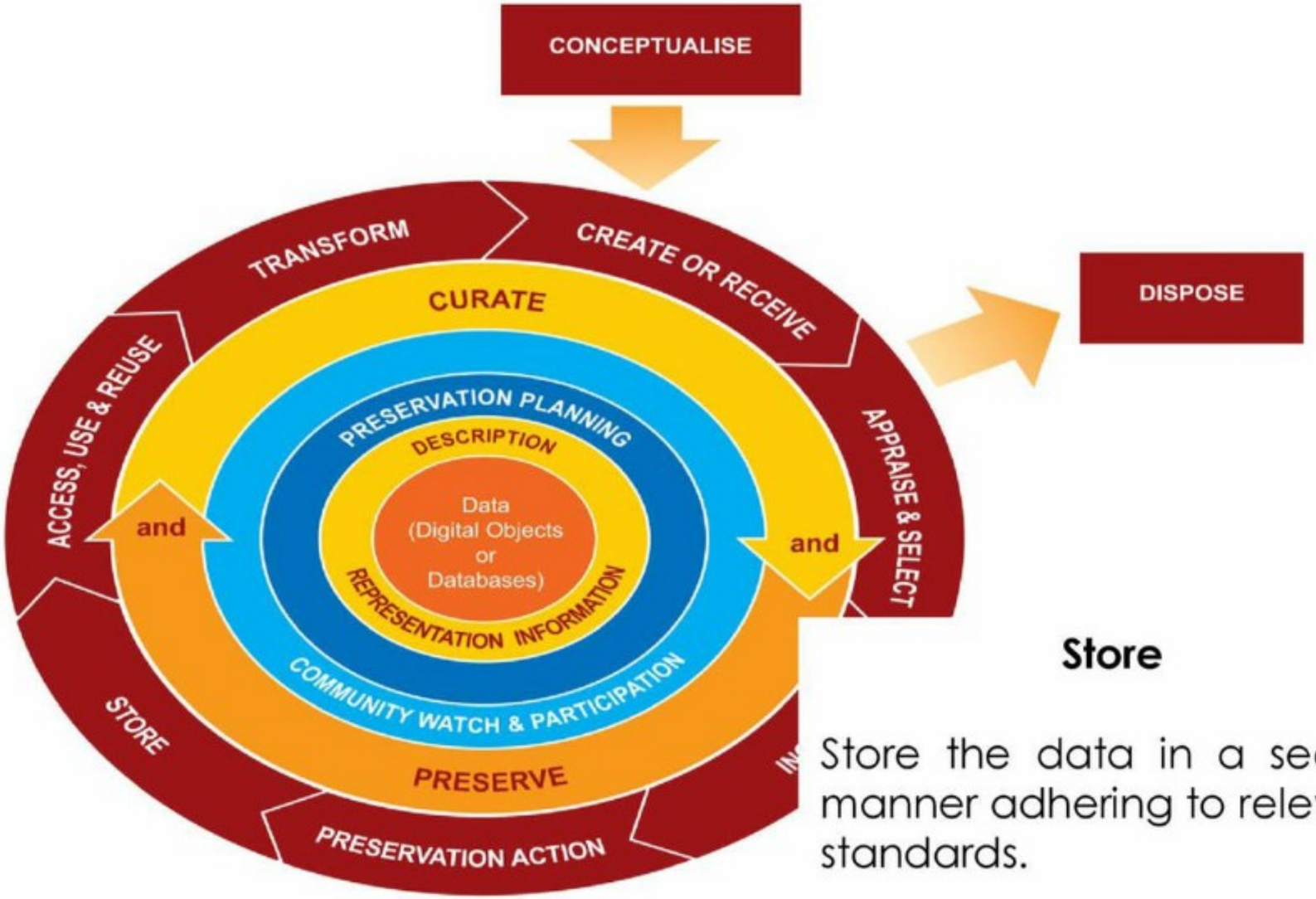
Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats.



CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

SEQUENTIAL ACTIONS



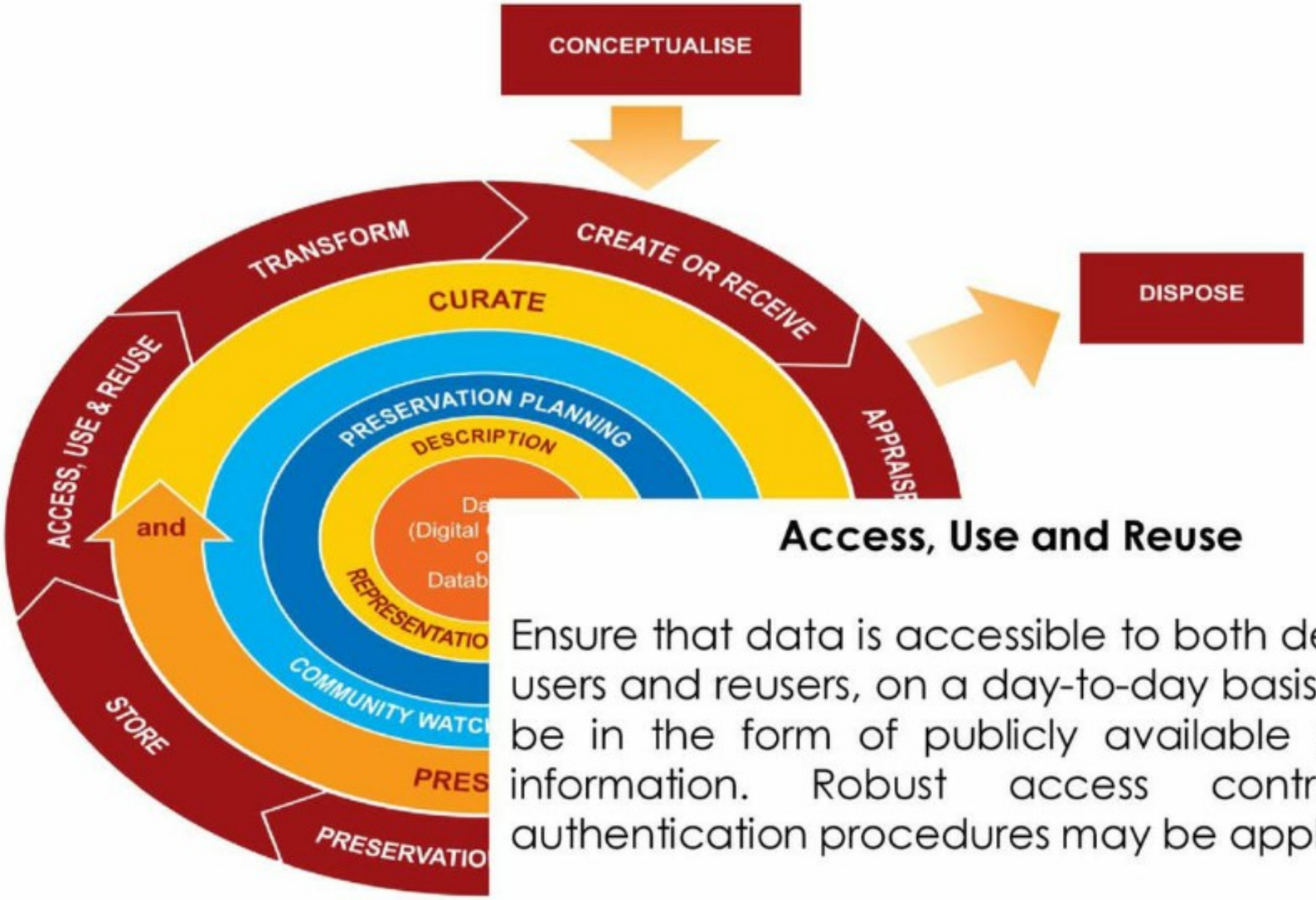
Store

Store the data in a secure manner adhering to relevant standards.

CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

SEQUENTIAL ACTIONS



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.

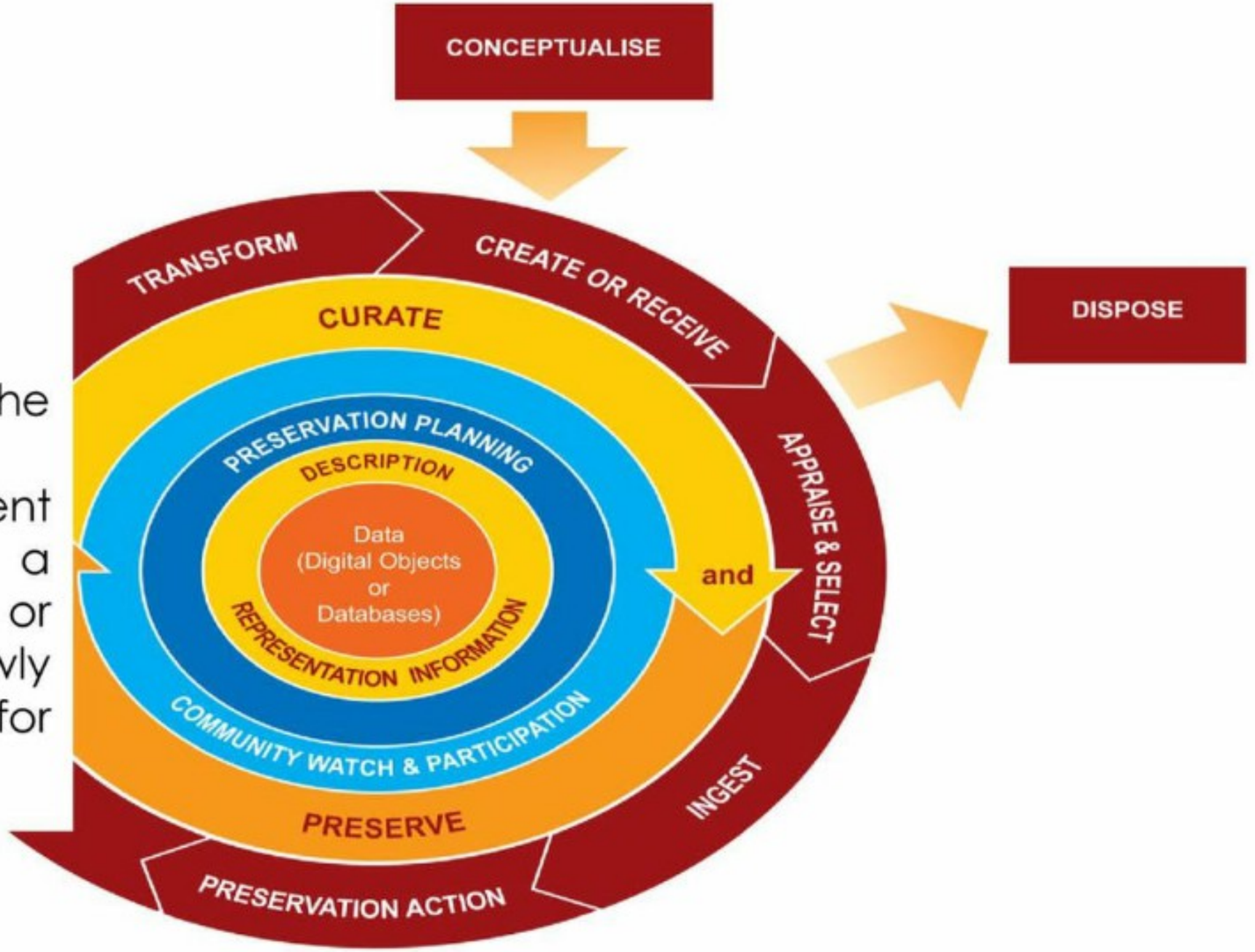
CSIC
Consejo Superior de Investigaciones Científicas

SEQUENTIAL ACTIONS

#CSIC

Transform

Create new data from the original, for example:
by migration into a different format, or by creating a subset, by selection or query, to create newly derived results, perhaps for publication



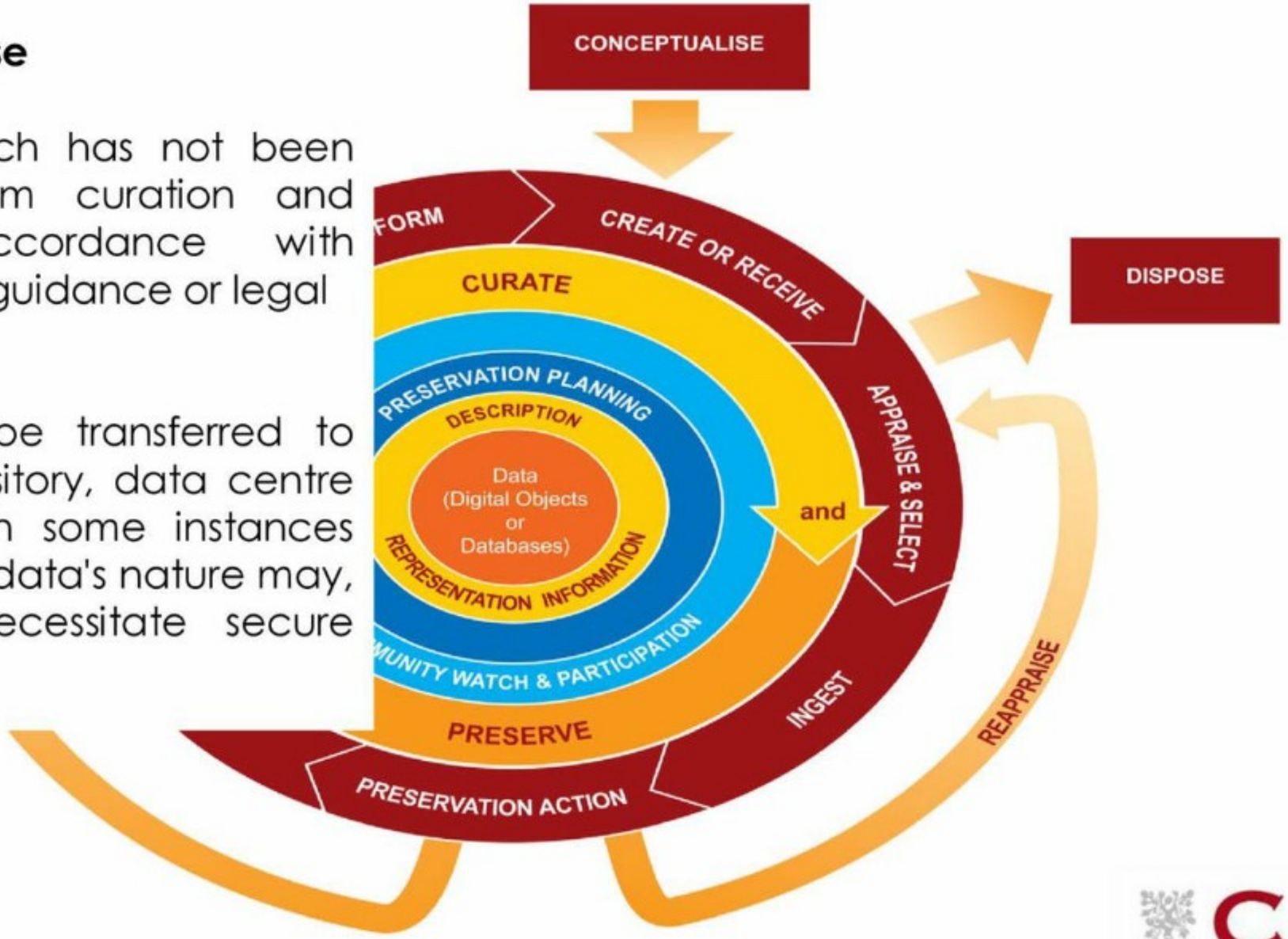
CSIC
Consejo Superior de Investigaciones Científicas

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.



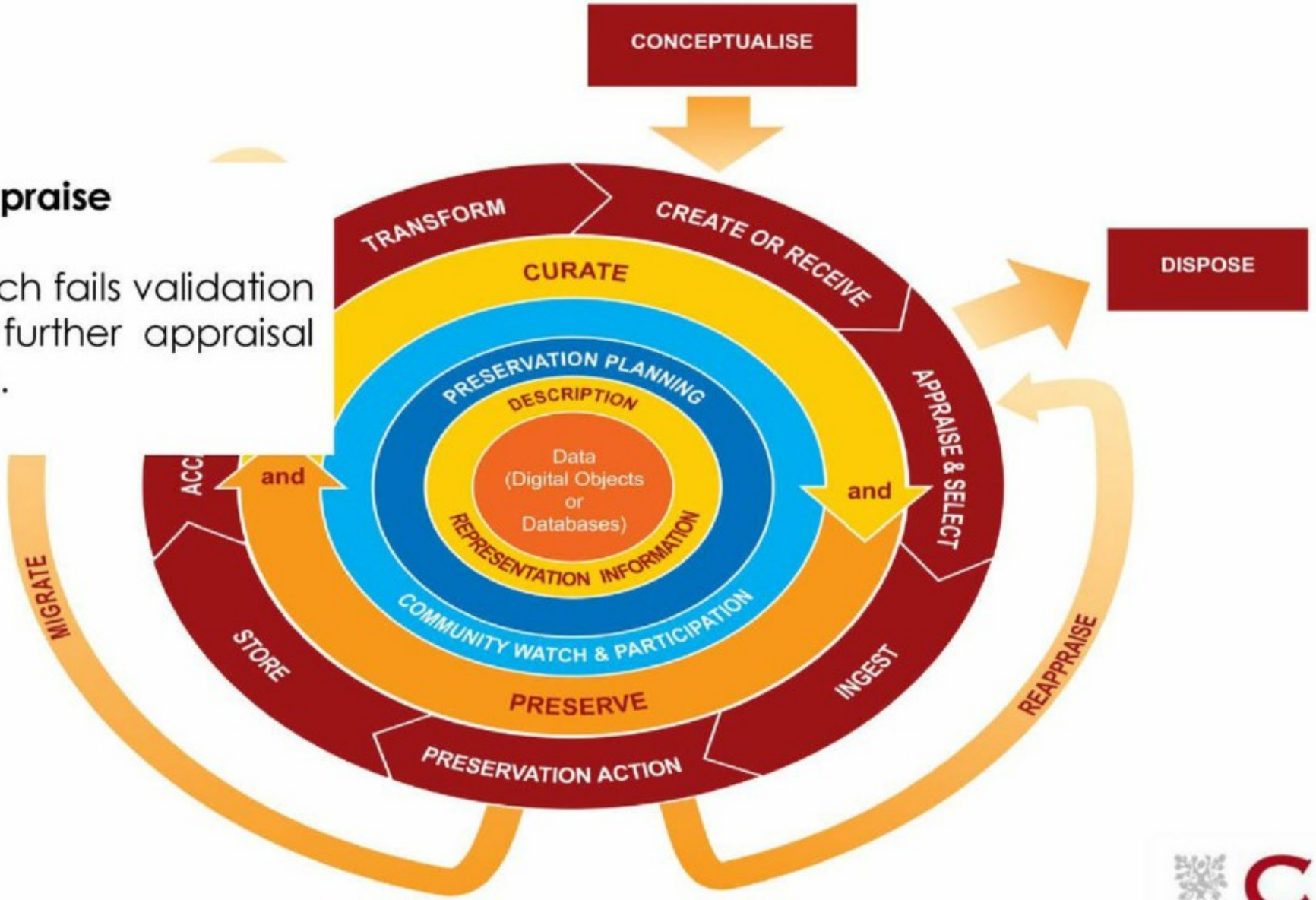
CSIC
Consejo Superior de Investigaciones Científicas

OCCASIONAL ACTIONS

#CSIC

Reappraise

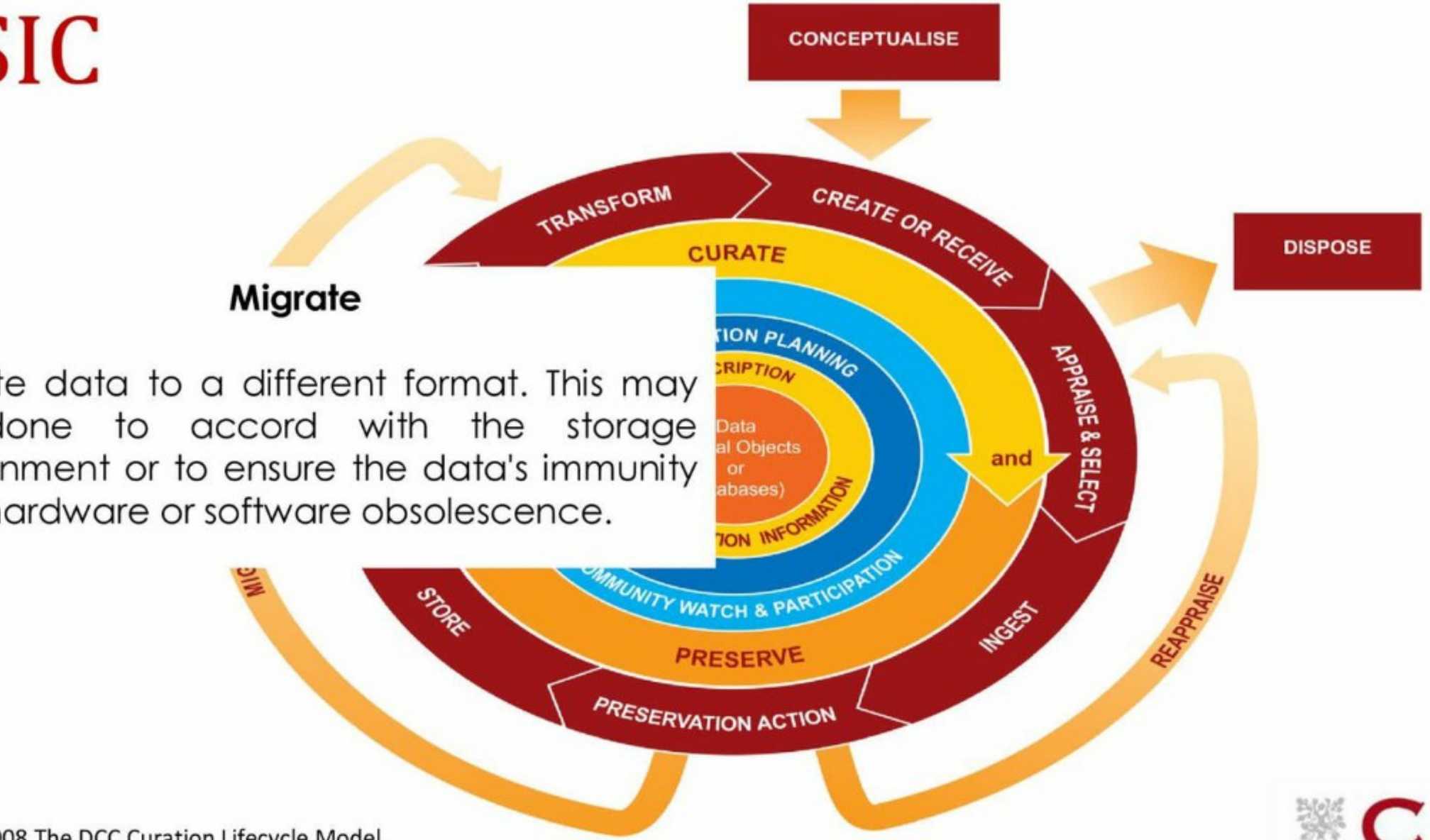
Return data which fails validation procedures for further appraisal and re-selection.



CSIC
Consejo Superior de Investigaciones Científicas

OCCASIONAL ACTIONS

#CSIC



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.

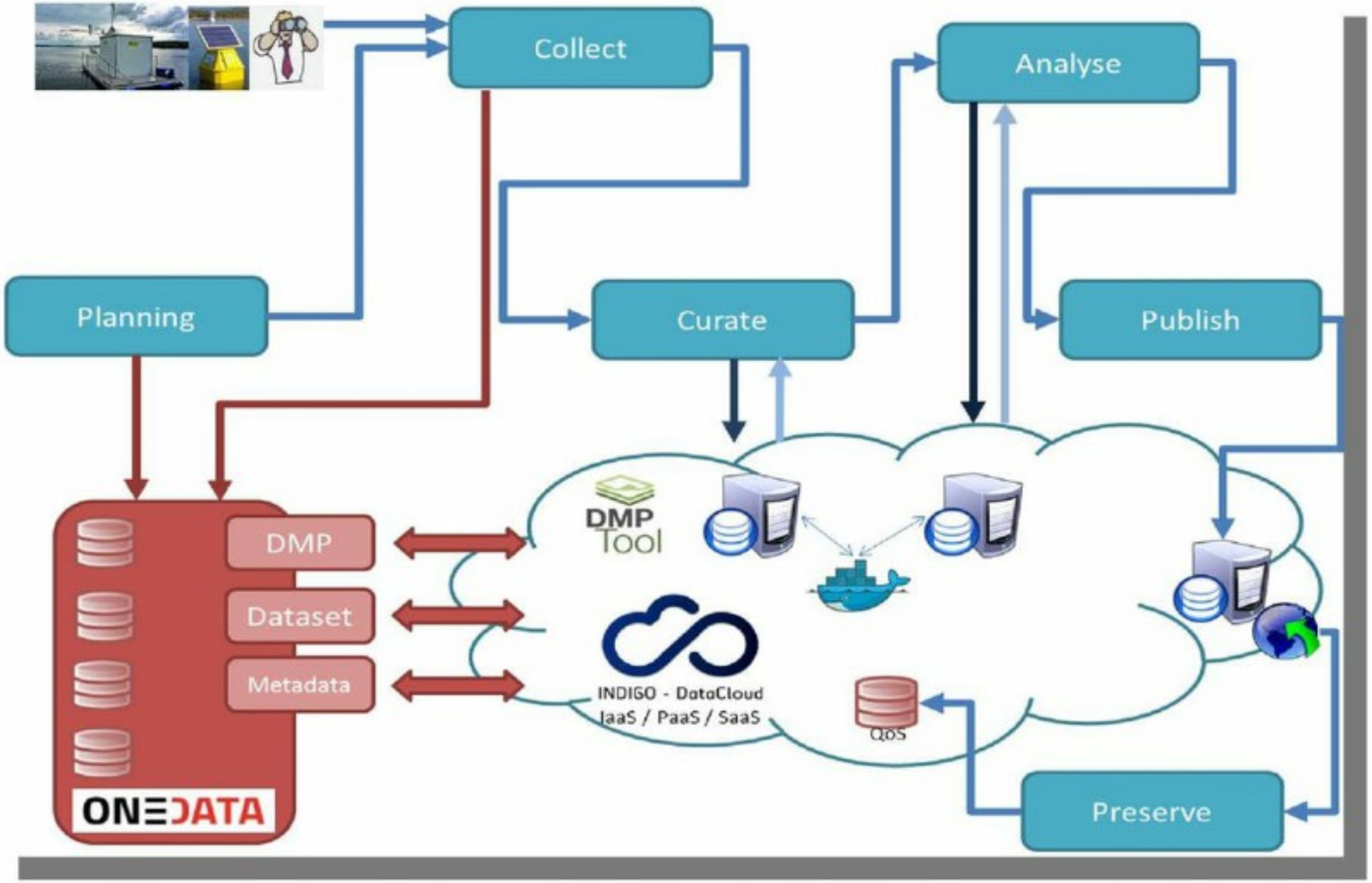
Higgins S. 2008 The DCC Curation Lifecycle Model, the International Journal of Digital Curation, Issue 1, Volume 3

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

CSIC

INDIGO Data Life Cycle ("6S") – Open Science

#CSIC

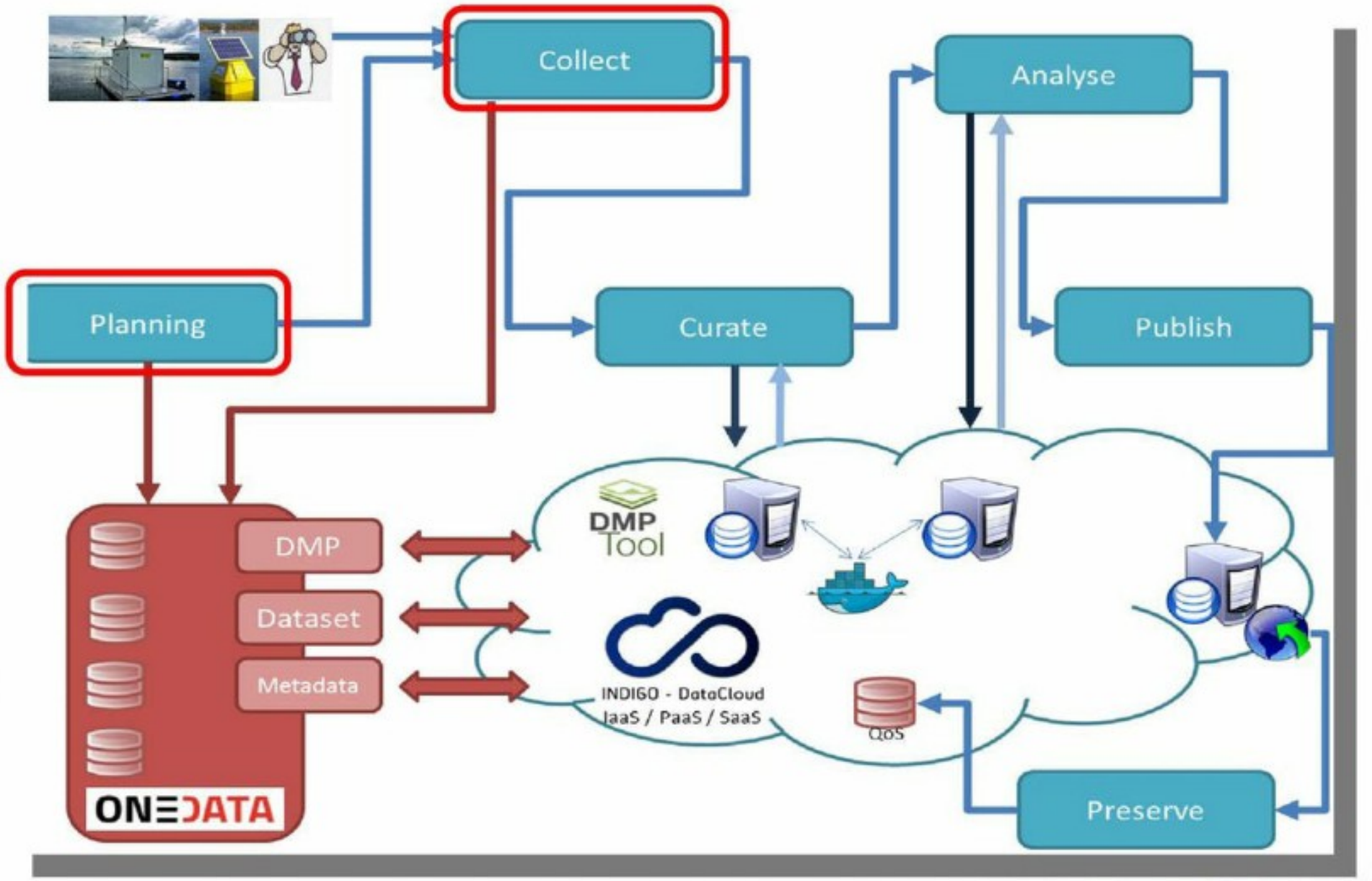


INDIGO Data Life Cycle ("6S")

#CSIC

Stage 1: Plan: prepare a Data Management Plan, including how data will be gathered, metadata definition, preservation plan, etc.

Stage 2: Collect: including both creation and acquisition, it is the process of getting data, in different ways. A storage service is needed as well.

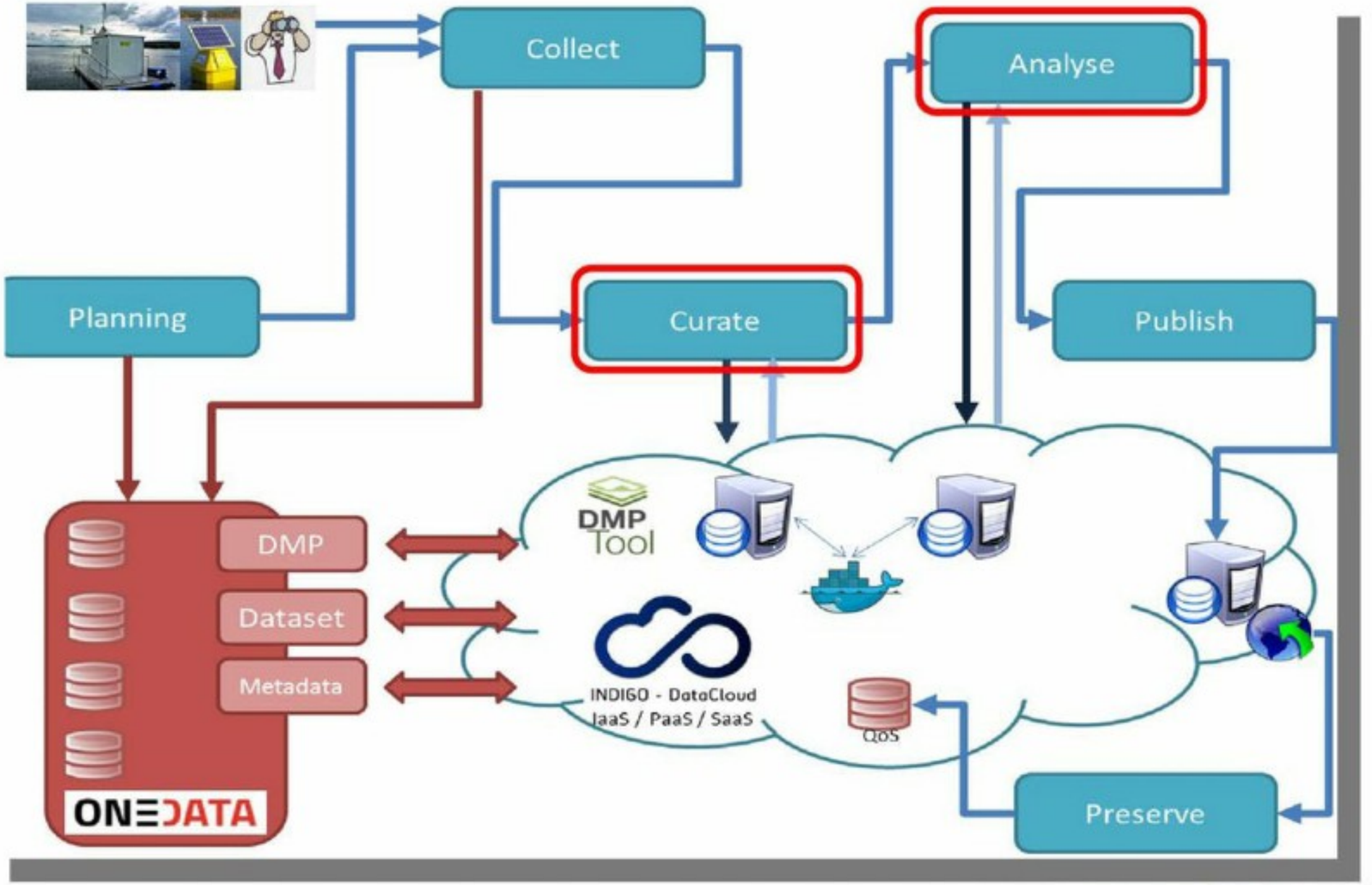


INDIGO Data Life Cycle ("6S")

#CSIC

Stage 3: Curate: also known as "Transform": using the raw data collected in the previous stage, manual or automatic actions are performed over the data, which is converted and also filtered.

Stage 4: Analyse: an optional step also called "Process", that implies performing different actions to give the data an added value and get new derived data.

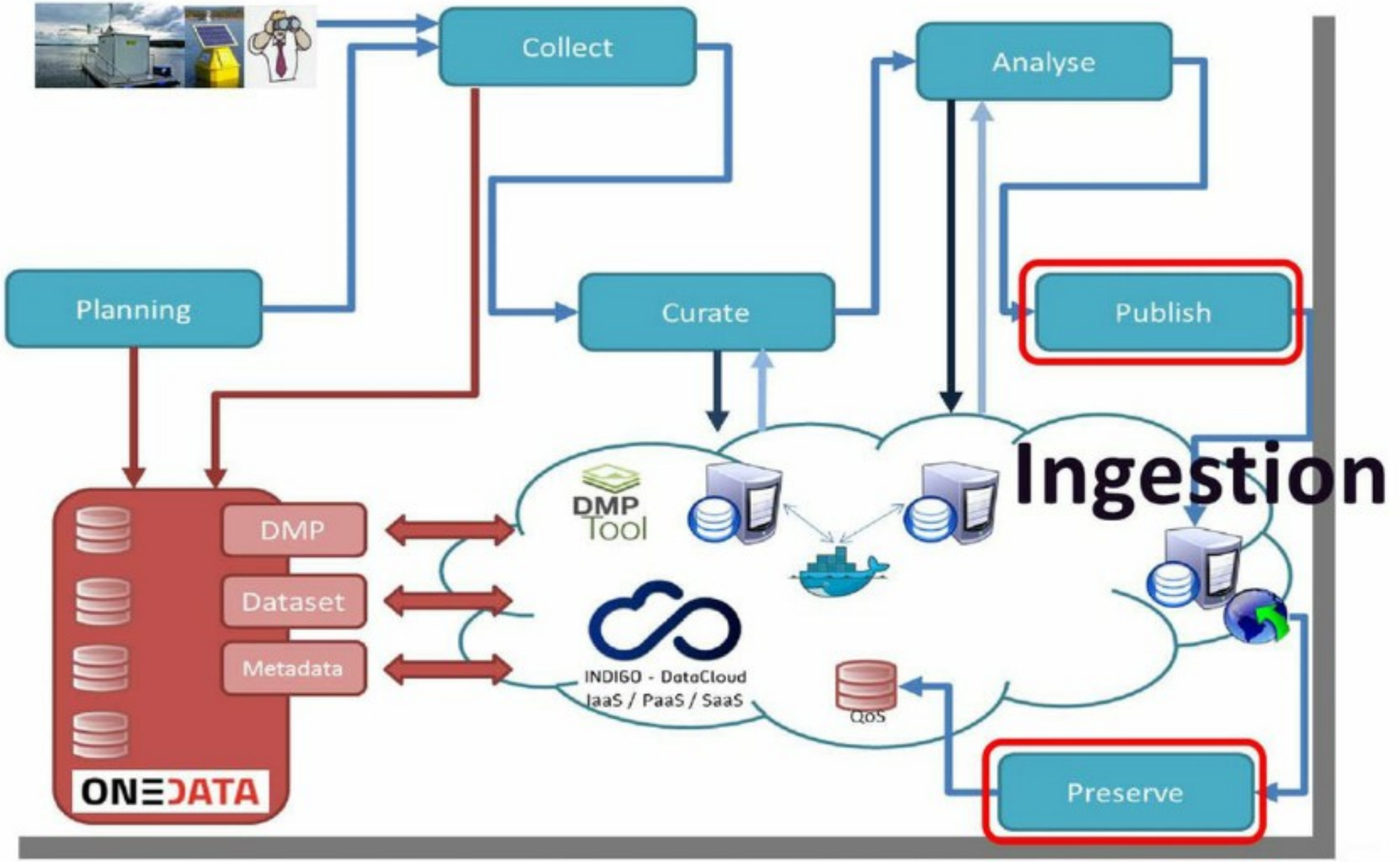


INDIGO Data Life Cycle ("6S")

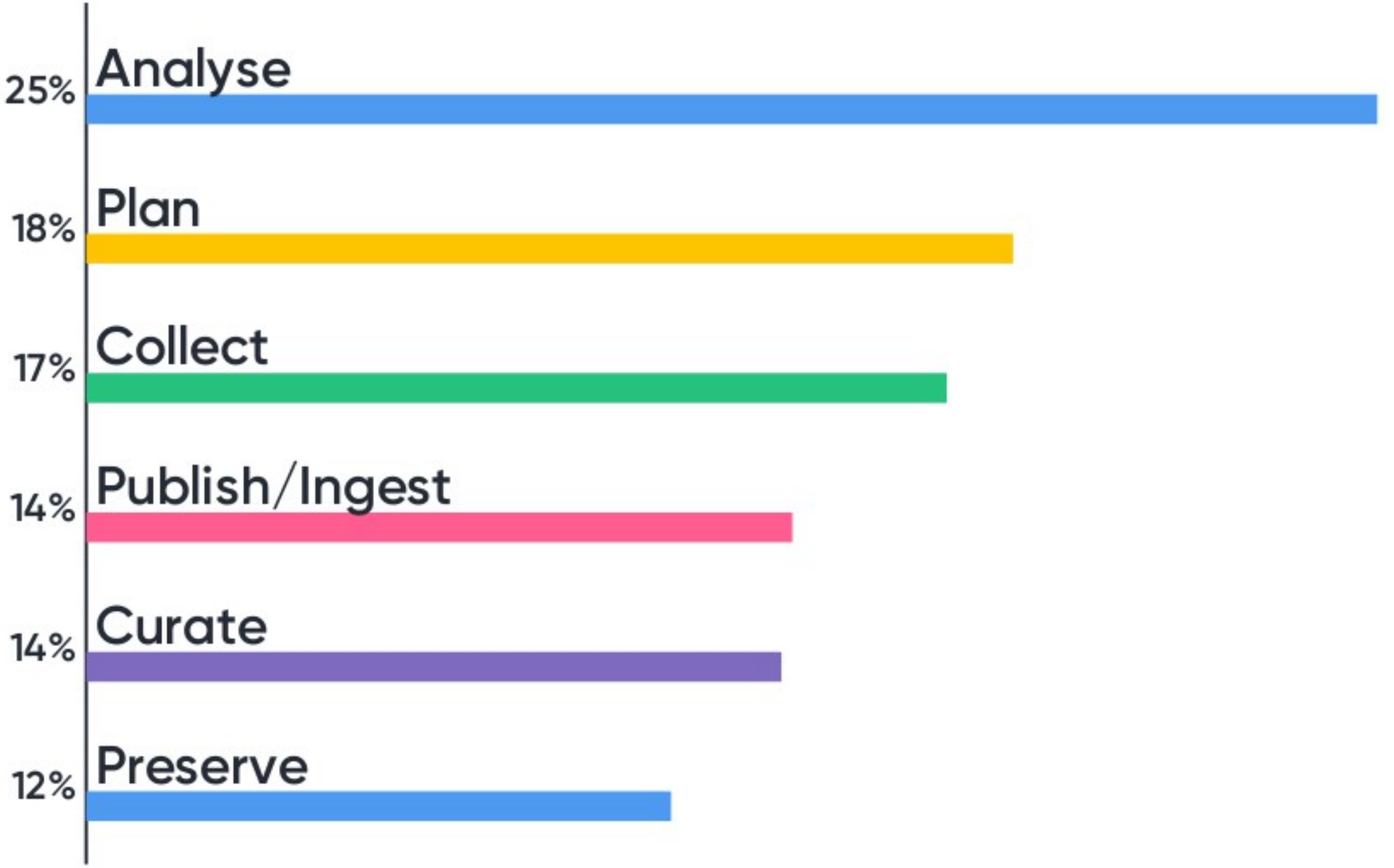
#CSIC

Stage 5: Ingest (& Publish): including other steps like "Access", "Use" or "Re-use", in this stage, data is normally associated to metadata, gets a persistent identifier (a DOI) and is published in an accessible repository or catalogue, under a format that makes it useful for further re-use.

Stage 6: Preserve: "store" both data and analysis for long-term. Licenses and methods need to be taken into account.



Ordena según su importancia



Data Management Plan

#CSIC

A DMP is a document (or similar) to define:

- How the data will be created.
- How it will be documented.
- Who will be able to access it (and how).
- Where (and how) will be stored.
- Who will back it up.
- Whether (and how) it will be shared & preserved.

DMPs are often submitted as part of grant applications, but they are useful whenever scientists are creating data.

G8, 2003, regarding funded projects:

"Open scientific research data should be easily discoverable, accessible, assessable, intelligible, usable, and wherever possible interoperable to specific quality standards".

CSIC
Consejo Superior de Investigaciones Científicas

Why DMPs?

#CSIC

Facilitate data reusing. DMPs contain a coherent set of sections describing how data life cycle is handled. Therefore, data can be tracked along its life, including mechanisms to ensure the provenance traceability.

Ensure Reproducibility. DMPs describe all the elements related to the data gathering, curation, and analysis, so the results of the research can be reproduced in the future.

Control costs. DMPs provide the funders a way to estimate and limit the costs associated with data collection. Data Management Plans must include a clear description of the purpose of the data gathering, including what is needed for the research project and the expected results or findings.

Think before act. The preparation of a DMP is the phase where all the elements that may influence the data life cycle can be integrated from a global perspective. This way, the resources can be optimized and no-sense or duplicated actions avoided.

Capture requirements along all the data life cycle phases. Although DMPs are sometimes considered as a static document, ideally it can be progressively updated taking into account the evolution along the data life cycle. DMPs should not be "closed" before the project starts, as unexpected issues will appear, nor "started" by the end of the project, when data may even have disappeared.

DMPs Approach

#CSIC

- As open as possible, as closed as necessary.
- DMP is a living document. Should be changed during the project.
- Different implementations available:
  - DMPonline
  - DMPtool
  - DMPRoadmap
- New developments towards machine-actionability.
- Oriented to create “FAIR” Data.

CSIC
Consejo Superior de Investigaciones Científicas

Concepts (To be clear…)

#CSIC

- **Metadata**: Data about data. Describes the context, the content and the structure of a dataset.
- **Machine-actionable** or **Machine-readable**: features that allows a software to automatize any action.
- **FAIR**: See next slide.
- **DOI** (Digital Object Identifier): implementation of a persistent identifier that can be assign to any digital object.
- **Ontology**: a formal naming and definition of the types, properties, and interrelationships of the entities that really exist in a particular domain.

- See European Commission H2020 DMP Guidelines:

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

CSIC
Consejo Superior de Investigaciones Científicas

What is FAIR Data?

#CSIC

FAIR Data aims to support existing communities in their attempts to enable valuable scientific data and knowledge to be published and utilized in a 'FAIR' manner.

Findable - (meta)data is uniquely and persistently identifiable. Should have basic machine readable descriptive metadata.

Accessible - data is reachable and accessible by humans and machines using standard formats and protocols.

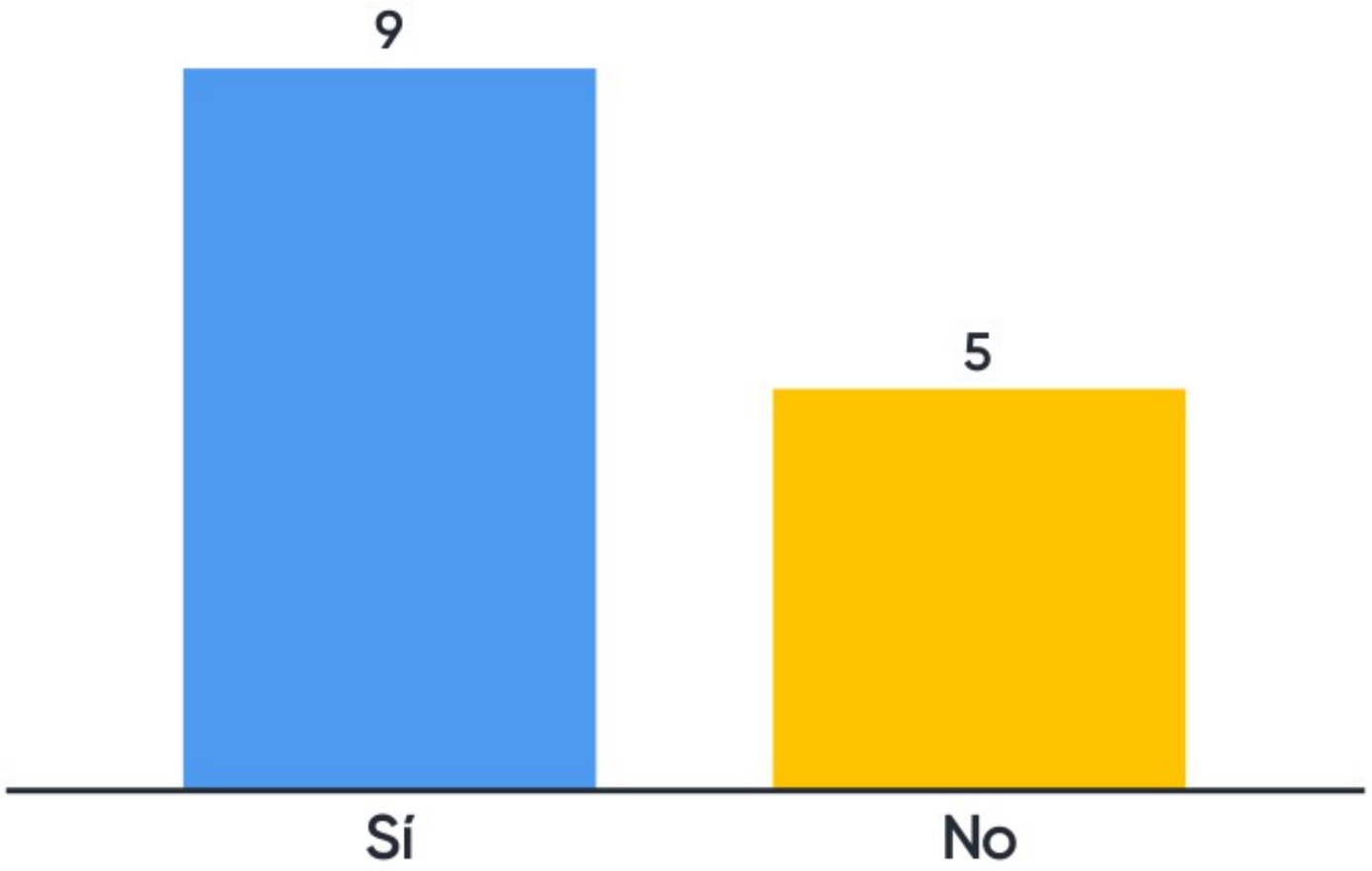
Interoperable - (meta)data is machine readable and annotated with resolvable vocabularies/ontologies.

Reusable - (meta)data is sufficiently well-described to allow (semi)automated integration with other compatible data sources.

Reproducible – Elements related to data are identified and relationships are well known (software, methods, related dataset, etc.).

CSIC
Consejo Superior de Investigaciones Científicas

Habías escuchado antes el término "FAIR data"?



DMPs in funding programs

#CSIC

NSF (National Science Foundation, US)

"Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants"

EC H2020: The European Commission DMP approach is oriented to make the data generated by a funded project "FAIR" and to ensure that the new data will be available, under certain conditions, for other researchers or even citizens aiming to use it if the security or ethical aspects allow doing so. Templates provided include in many cases the description of the software used during the data life cycle, aiming to ensure also data reproducibility.

Components of a General DMP

#CSIC

1. Information about data & data format
2. Metadata content and format
3. Policies for access, sharing and re-use
4. Long-term storage and data management
5. Budget

H2020 DMP

#CSIC

1. DATA SUMMARY

What is the purpose of the data collection/generation and its relation to the objective of the project?

What types and formats of data will the project generate/collect?

Will you re-use any existing data and how?

What is the origin of the data?

What is the expected size of the data?

To whom might it be useful ('data utility')?

H2020 DMP

#CSIC

2. FAIR DATA

Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

What naming conventions do you follow?

Will search keywords be provided that optimize possibilities for re-use?

Do you provide clear version numbers?

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

CSIC

Consejo Superior de Investigaciones Científicas

H2020 DMP

#CSIC

- **2. FAIR DATA**
- **Making data openly accessible**
- Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
- Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.
- How will the data be made accessible (e.g. by deposition in a repository)?
- What methods or software tools are needed to access the data?
- Is documentation about the software needed to access the data included?
- Is it possible to include the relevant software (e.g. in open source code)?
- Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.
- Have you explored appropriate arrangements with the identified repository?
- If there are restrictions on use, how will access be provided?
- Is there a need for a data access committee?
- Are there well described conditions for access (i.e. a machine readable license)?
- How will the identity of the person accessing the data be ascertained?

2. FAIR DATA

Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

H2020 DMP

#CSIC

2. FAIR DATA

Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

How long is it intended that the data remains re-usable?

Are data quality assurance processes described?

Further to the FAIR principles, DMPs should also address:

CSIC
Consejo Superior de Investigaciones Científicas

H2020 DMP

#CSIC

3. ALLOCATION OF RESOURCES

What are the costs for making data FAIR in your project?

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Who will be responsible for data management in your project?

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

4. DATA SECURITY

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Is the data safely stored in certified repositories for long term preservation and curation?

5. ETHICAL ASPECTS

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

CSIC

Consejo Superior de Investigaciones Científicas

H2020 DMP

#CSIC

6. OTHER ISSUES

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

7. FURTHER SUPPORT IN DEVELOPING YOUR DMP

The Research Data Alliance provides a Metadata Standards Directory that can be searched for discipline-specific standards and associated tools.

The EUDAT B2SHARE tool includes a built-in license wizard that facilitates the selection of an adequate license for research data.

Useful listings of repositories include:

Registry of Research Data Repositories

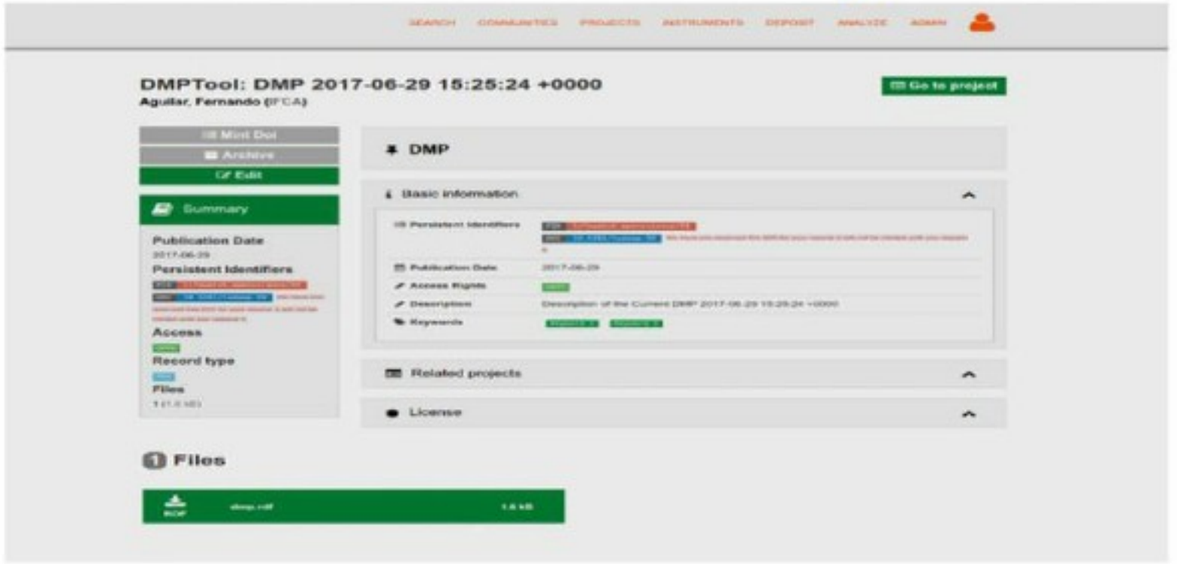
Some repositories like Zenodo, an OpenAIRE and CERN collaboration), allow researchers to deposit both publications and data, while providing tools to link them.

Other useful tools include DMP online and platforms for making individual scientific observations available such as ScienceMatters.

CSIC
Consejo Superior de Investigaciones Científicas

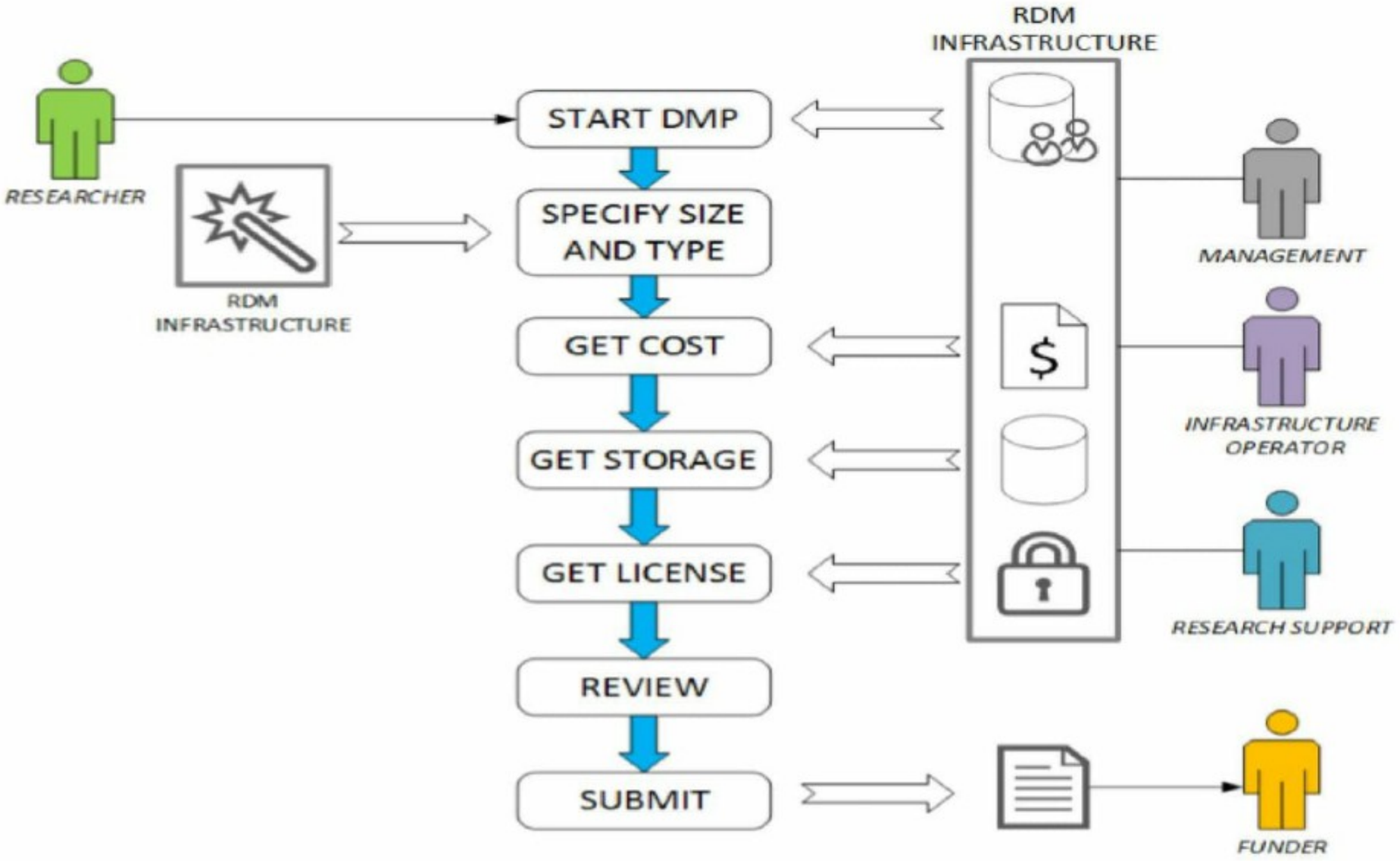
Regarding Machine-actionability

#CSIC



CSIC
Consejo Superior de Investigaciones Científicas

Automated Data Management Workflow

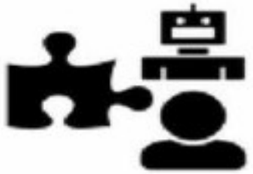


10 principles for maDMPs

1 Integrate DMPs with the workflows of all stakeholders in the research data ecosystem

2 Allow automated systems to act on behalf of stakeholders

3 Make policies (also) for machines, not just for people



4 Describe—for both machines and humans—the components of the data management ecosystem

5 Use PIDs and controlled vocabularies

Miksa, Tomasz, Simms, Stephanie, Mietchen, Daniel, & Jones, Sarah. (2018). Ten simple rules for machine-actionable data management plans (preprint). http://doi.org/10.5281/zenodo.1434938

10 principles for maDMPs

6 Follow a common data model for maDMPs

7 Make DMPs available for human and machine consumption

8 Support data management evaluation and monitoring

9 Make DMPs updatable, living, versioned documents

10 Make DMPs publicly available

Miksa, Tomasz, Simms, Stephanie, Mietchen, Daniel, & Jones, Sarah. (2018). Ten simple rules for machine-actionable data management plans (preprint). http://doi.org/10.5281/zenodo.1434938

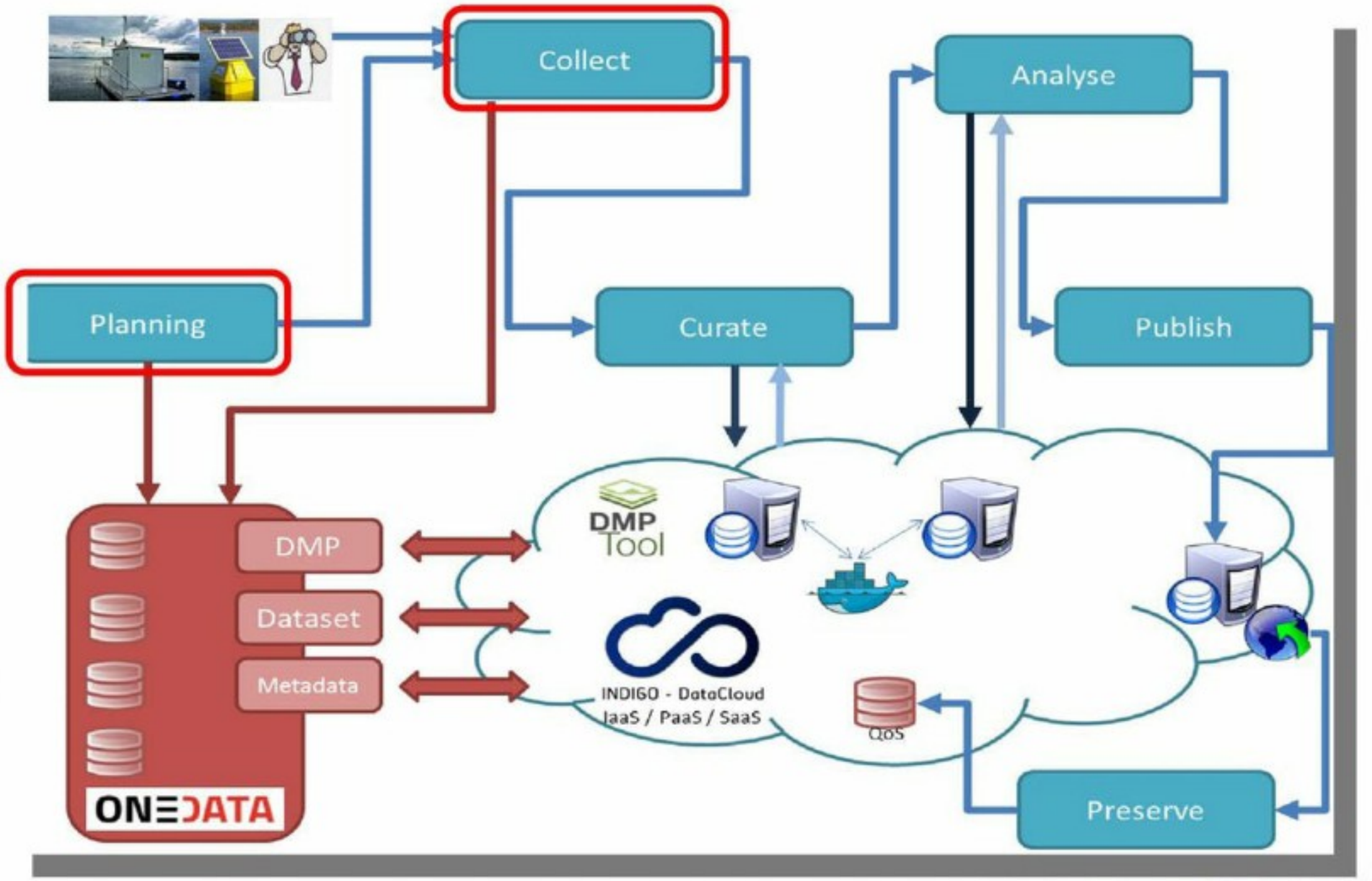
Experiencias con DMPs

INDIGO Data Life Cycle ("6S")

#CSIC

Stage 1: Plan: prepare a Data Management Plan, including how data will be gathered, metadata definition, preservation plan, etc.

Stage 2: Collect: including both creation and acquisition, it is the process of getting data, in different ways. A storage service is needed as well.



#CSIC

What is Metadata?

- *"Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information."*

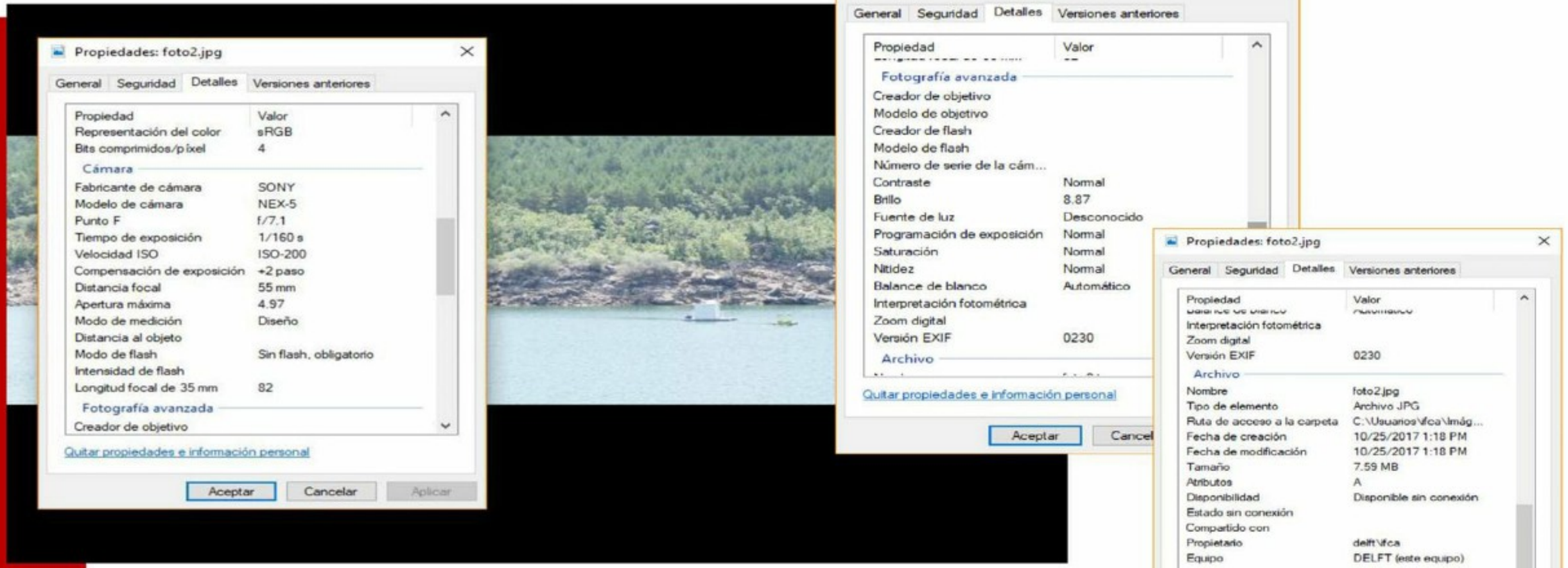
-- National Information Standards Organization
http://www.niso.org/publications/press/UnderstandingMetadata.pdf

- Metadata provides information enabling to make sense of data (e.g. documents, images, datasets), concepts (e.g. classification schemes) and real-world entities (e.g. people, organizations, places, paintings, products).
- "Data about Data"
- "A love note to the future"
- Makes your data **F**indable, **A**ccessible, **I**nteroperable and **R**e-usable

CSIC
Consejo Superior de Investigaciones Científicas

Metadata in our files

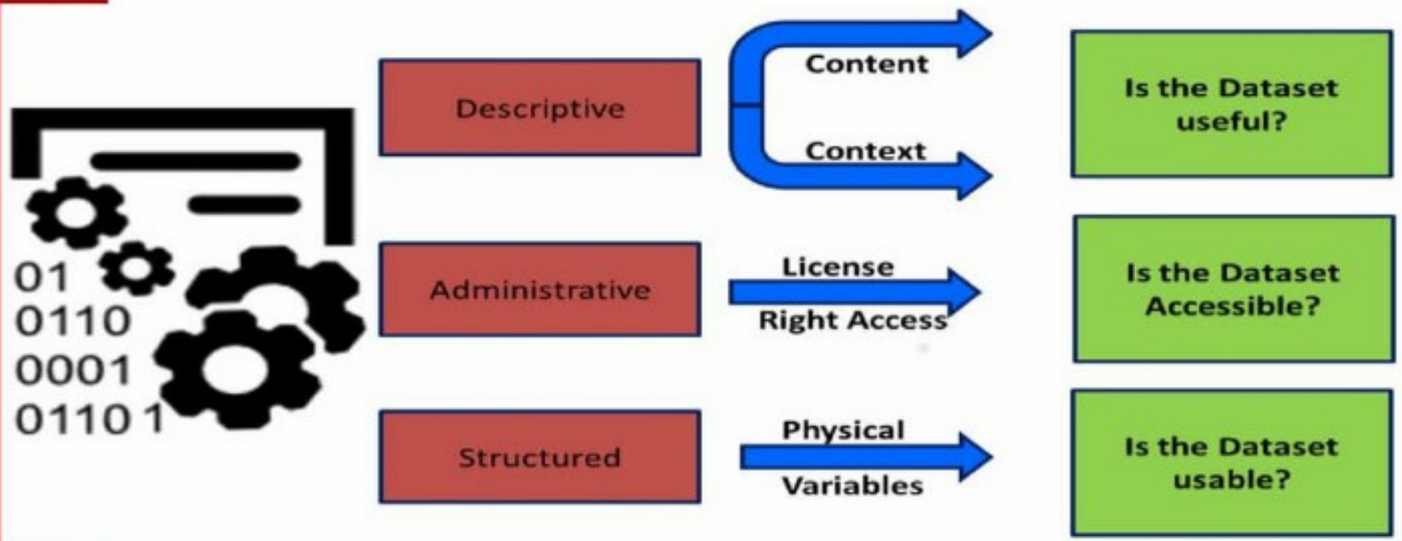
#CSIC



CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

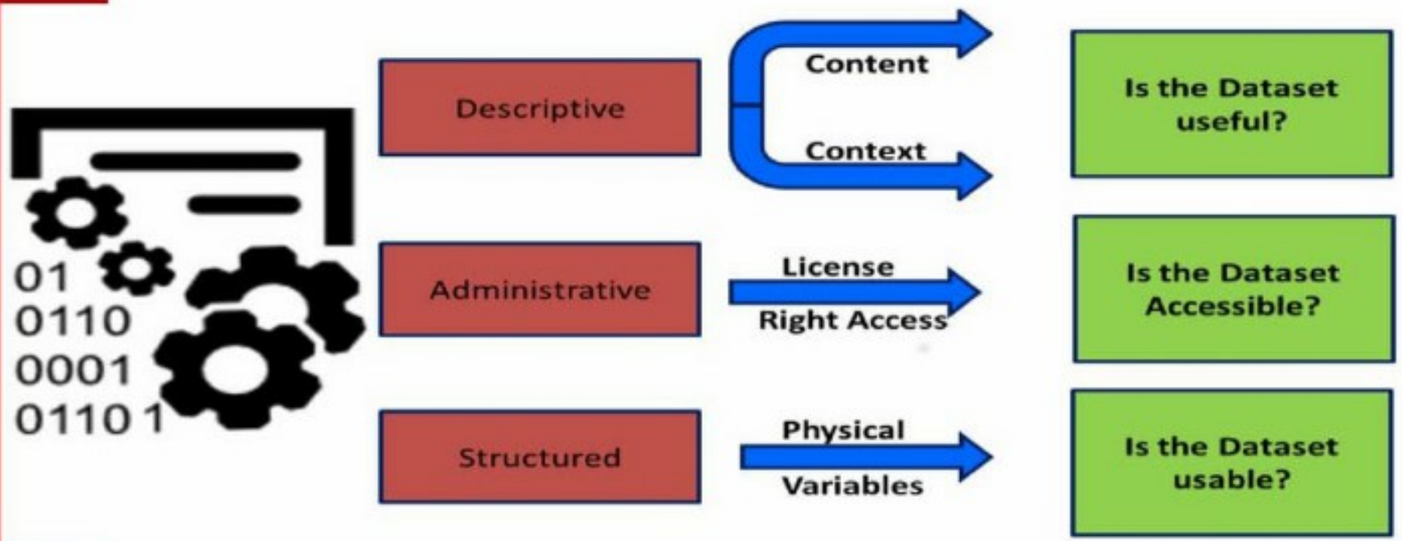
Metadata Types



Descriptive metadata defines and identify digital objects, and provide the set of elements that makes data findable. Those elements may include unique or persistent identifiers and other bibliographic attributes like title, author, creator, keywords, description, etc. **Dublin Core** is the standard and referent that is oriented to support this type of metadata. Exploiting this class, users or automatic algorithms can determine if the digital object is potentially useful.

#CSIC

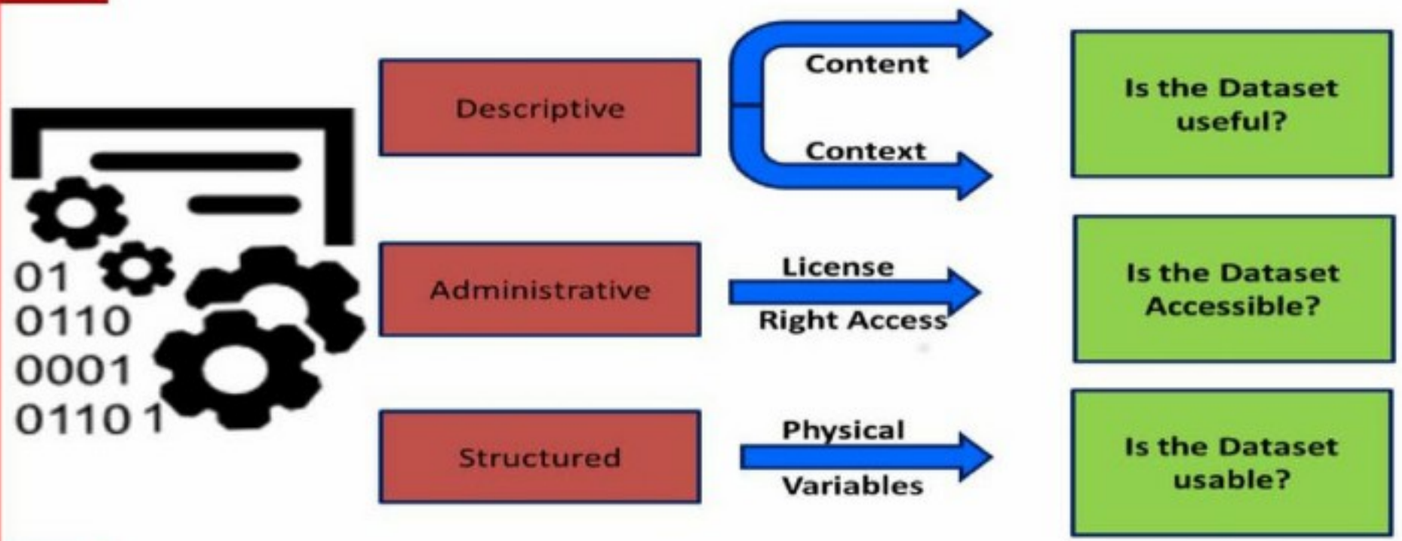
Metadata Types



Administrative metadata provides information about how the dataset has been created and also the details about the quality control and assurance. It also includes the license type, embargo period, access rights details and any other information related to how accessible the data is. The use of this metadata category should help to decide if the dataset or digital object is or not accessible by a potential user.

#CSIC

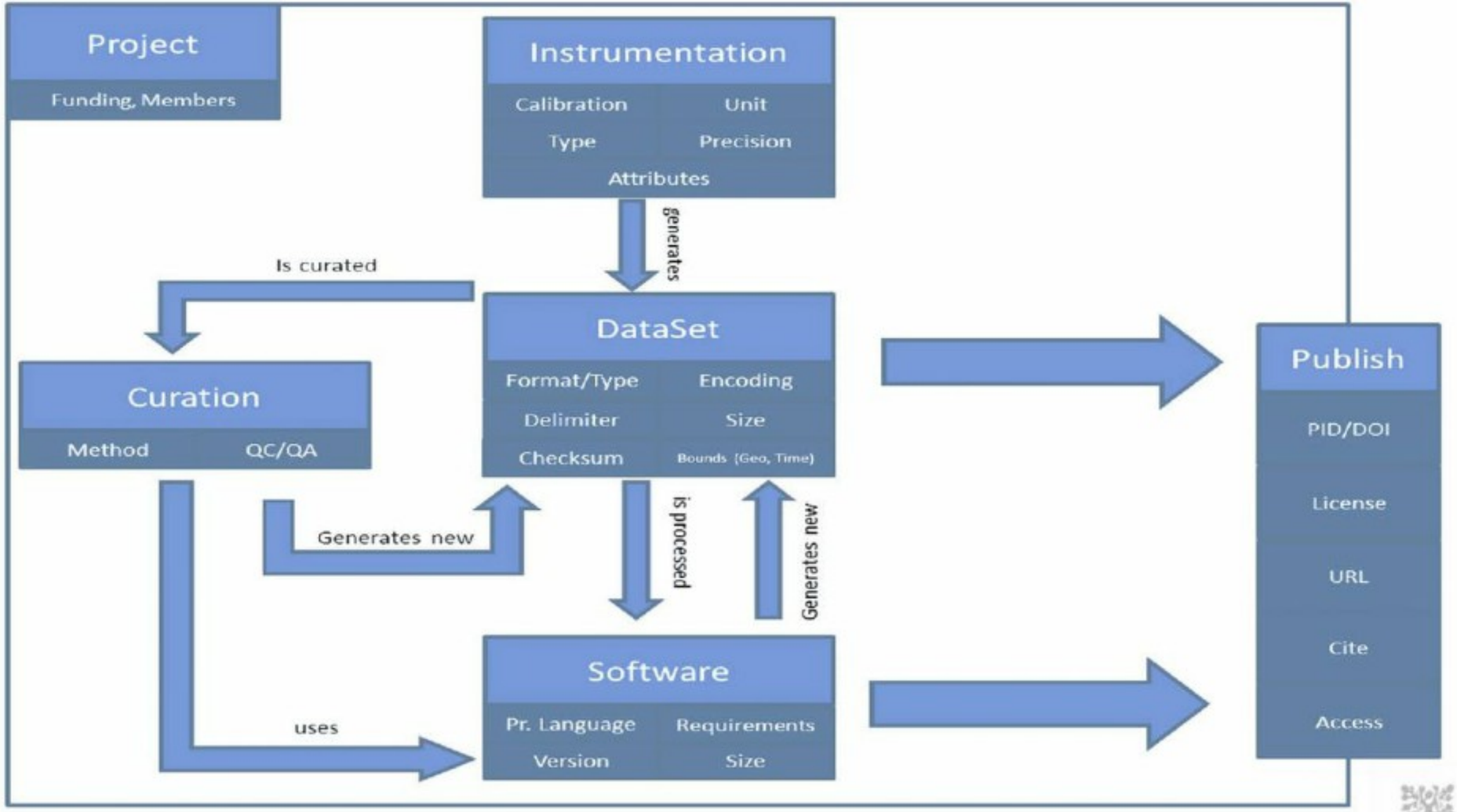
Metadata Types



Structural metadata provides information about the physical structure of the dataset or digital objects, such as its format or encoding and other logical details like the names of the variables, their units and the relationship with other objects or elements.
Complex metadata standard like the Ecological Metadata Language, integrate different metadata elements at a structural level. This type of metadata enables interoperability, since the integration of different described objects can be done automatically. It also provides enough information to decide if data can be exploited or if it is reusable.

#CSIC

Ecological Metadata Language



CSIC
Consejo Superior de Investigaciones Científicas

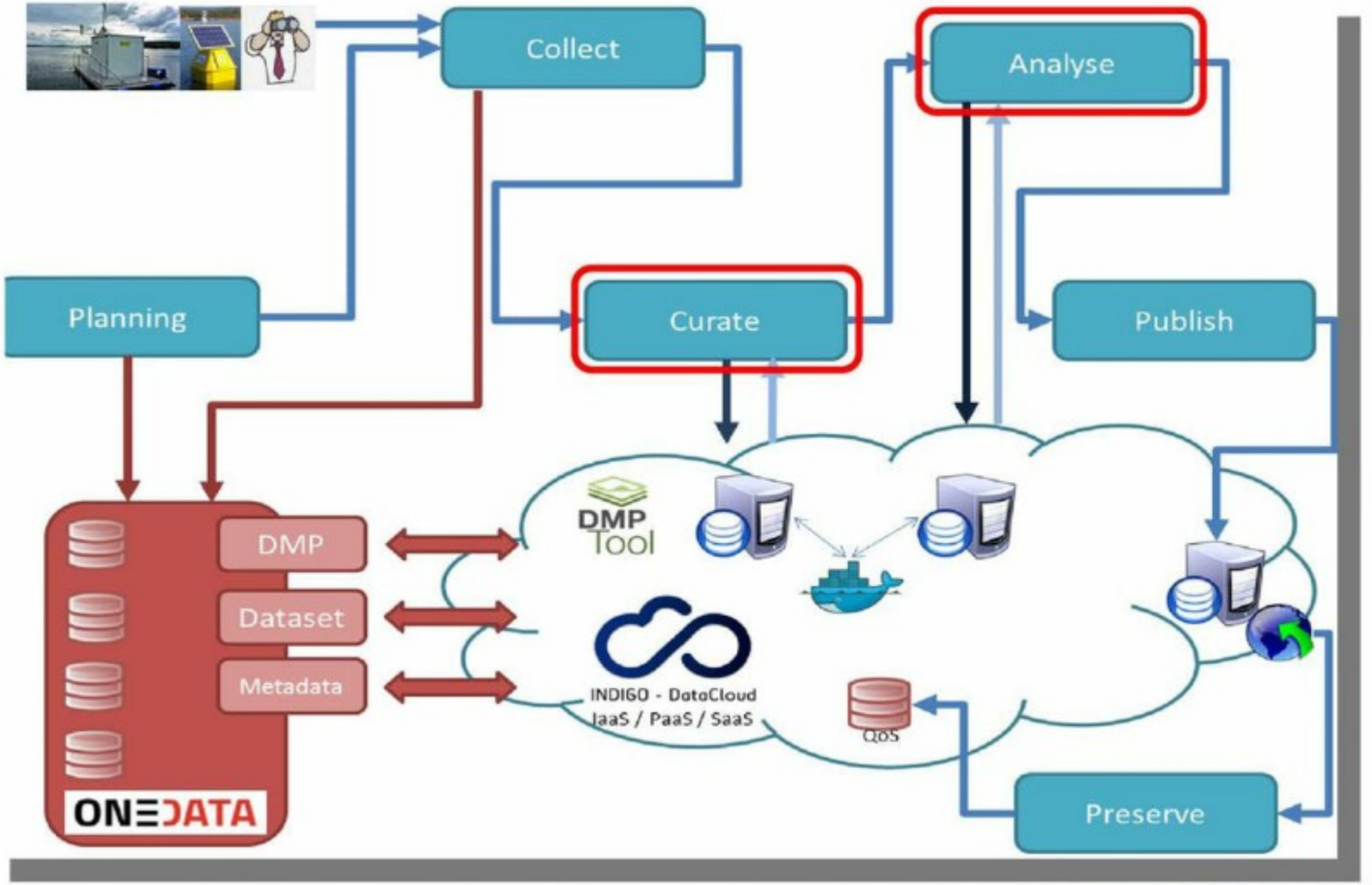
Metadatos para "Las Meninas"

INDIGO Data Life Cycle ("6S")

#CSIC

Stage 3: Curate: also known as "Transform": using the raw data collected in the previous stage, manual or automatic actions are performed over the data, which is converted and also filtered.

Stage 4: Analyse: an optional step also called "Process", that implies performing different actions to give the data an added value and get new derived data.



Data Curation

#CSIC

- The general definition for the term Curate: “the item selection among a large number of possibilities to become part of something”.
- The term “Data Curation” refers specifically to the process of organizing and integrating data from different sources, prepare them to be used in a coherent way, making them compatible.
- Other relevant term is “Data Cleaning”, referring to the process to detect corrupt or inexact records within a table, database or value list, and if possible, correct and complete them (missing values), or otherwise tag them as erroneous.
- It is usual in Data Science to use the term “Data Curation” to refer to both processes, as they are usually considered as a single phase in the Data Life Cycle, transforming the raw data gathered into data ready to be used.
- Curation techniques are so quite related to the different data gathering methods as all of them require curation processes to be applied to ensure their quality.

Digital Curation

#CSIC

- DCC definition:
  - "... maintaining and adding value to a trusted body of digital information for current and future use; specifically, we mean the active management data over the life-cycle of scholarly and scientific materials" (http://www.dcc.ac.uk/)
- Main themes:
  - Curation is seen as an ongoing process, e.g. the active management of data over time
  - It is also about adding-value through things like community annotation
  - Life-cycles are important, long-term stewardship not always necessary
  - Not identical to digital preservation

Why curate data?

#CSIC

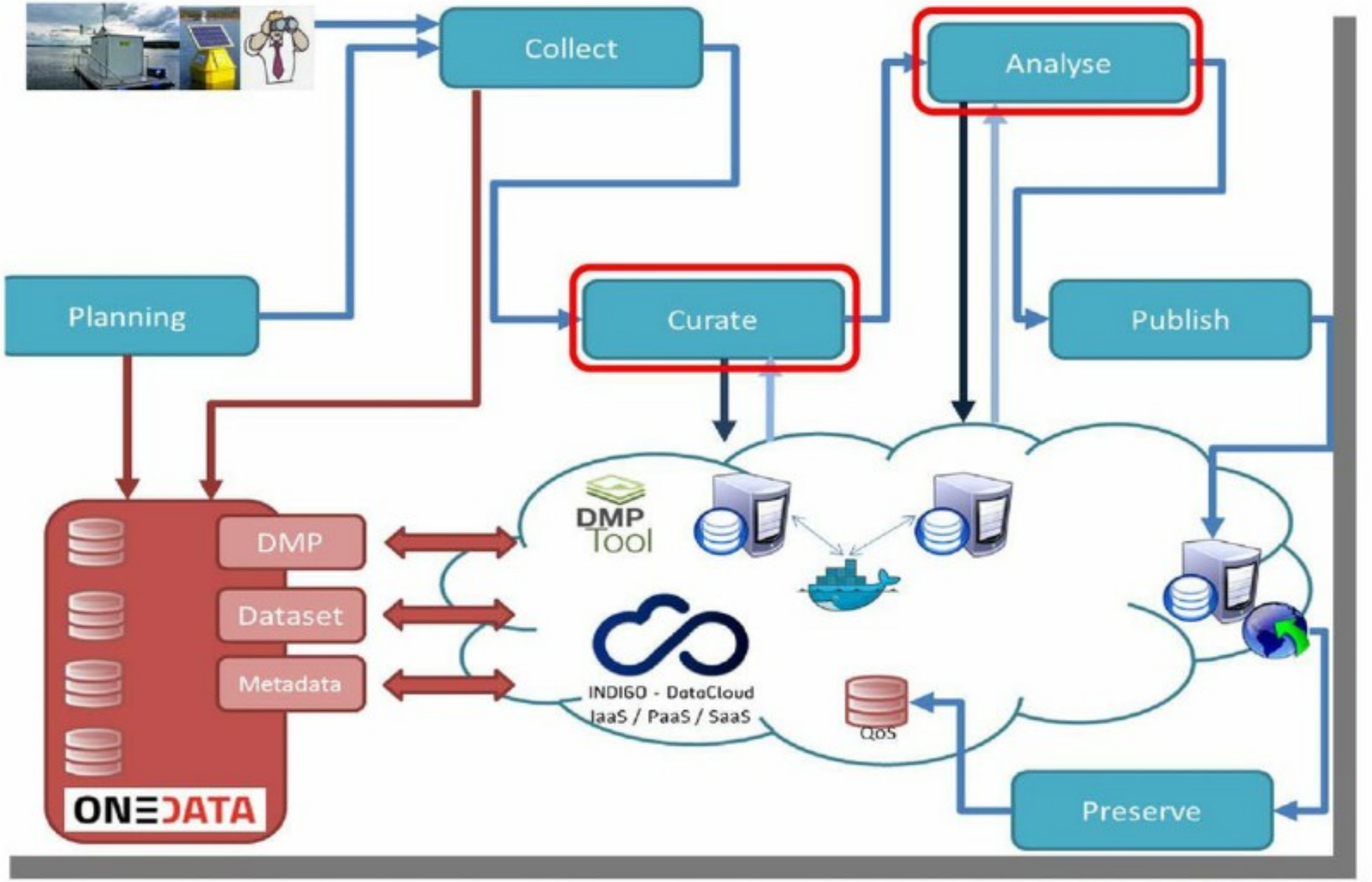
- The potential for creating 'new' knowledge from existing data:
  - Re-use, re-analysis, data mining
  - Annotation, e.g. in molecular biology astronomy
  - Combining datasets in innovative ways, e.g. mapping biodiversity data onto ecological GIS
  - “Science 2.0”
- It is increasingly a requirement of some research funding bodies
  - Some have quite mature data retention policies (not necessarily for permanent retention)
  - Increasing expectation of access to data from publicly-funded research
  - OECD Principles and guidelines for access to research data from public funding (2007)
  - FAIR principles

INDIGO Data Life Cycle ("6S")

#CSIC

Stage 3: Curate: also known as "Transform": using the raw data collected in the previous stage, manual or automatic actions are performed over the data, which is converted and also filtered.

Stage 4: Analyse: an optional step also called "Process", that implies performing different actions to give the data an added value and get new derived data.



Analysis

#CSIC

- Big Data in different fields
  - Physics: Particle physics, astrophysics…
  - Geospatial data: satellites, environmental data.
  - Genomics
  - Social media data
- Computing resources required:
  - HPC: Red Española de Supercomputación, PRACE
  - HTC: Grid, EGI.eu, IBERGRID
  - Cloud: Privados, públicos (EGI FedCloud).

CSIC
Consejo Superior de Investigaciones Científicas

Analysis

#CSIC

- Explore different Methods
  - GIS: Essential to analyze georeferenced data.
  - Data Analytics: Python, R. Jupyter/Rshiny as a user-friendly environment.
  - Modelling: weather forecasting, seismology, etc.
- WorkFlows
  - Kepler, Taverna, **TRUFA**
  - Integration method in the EOSC Framework.
- Big Data Analytics

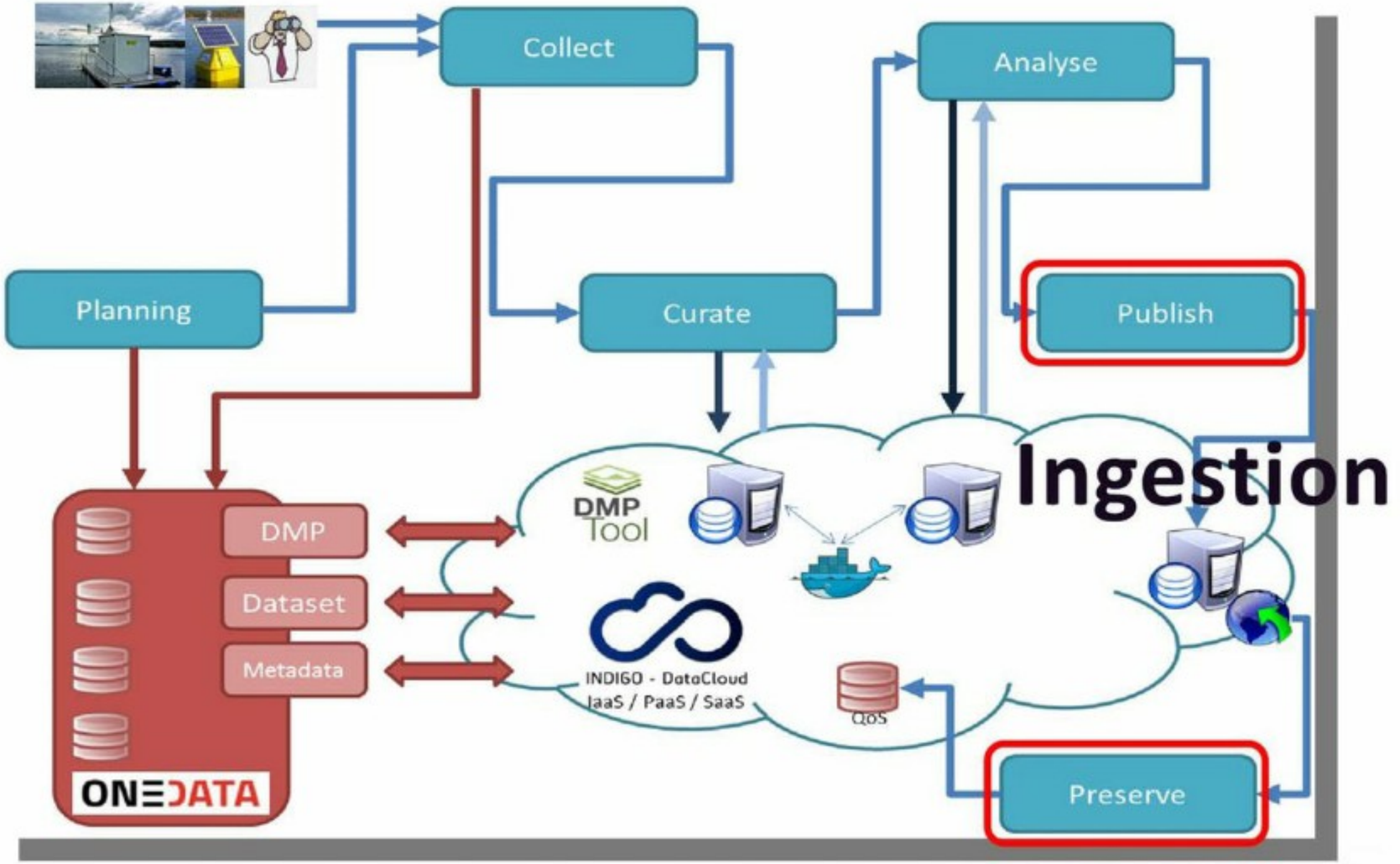
CSIC
Consejo Superior de Investigaciones Científicas

INDIGO Data Life Cycle ("6S")

#CSIC

Stage 5: Ingest (& Publish): including other steps like "Access", "Use" or "Re-use", in this stage, data is normally associated to metadata, gets a persistent identifier (a DOI) and is published in an accessible repository or catalogue, under a format that makes it useful for further re-use.

Stage 6: Preserve: "store" both data and analysis for long-term. Licenses and methods need to be taken into account.



Publish

#CSIC

- Not only papers: datasets, images, media…
- Links between papers and data
- Select proper repository: DIGITAL.CSIC, community repository.
- Metadata: Key! Dublin core at least.
- PIDs

CSIC
Consejo Superior de Investigaciones Científicas

PIDs – Why?

#CSIC

- Managing **increasing numbers of digital objects**
- **Sharing data from different sources** amongst researchers
- Data needs to be **(globally) identifiable and addressable** → reuse of data
- Citation
- Linking data from different sources
- Linking data with publications.

CSIC
Consejo Superior de Investigaciones Científicas

PIDs Systems (Technical Solutions)

#CSIC

The Uniform Resource Name (URN)

It was specified in 1997 and it is a solution that does require that the resource being referred is available, but it is not extended to resolve, and other actionable features are missing.

The Archival Resource Key (ARK)

- It is one of the most extended solutions for PIDs. It was released by the US National Library of Medicine and its URL-based structure includes information within the identifier like the name of the hosting authority, the name or resource ID and the qualifier.
- ARK is used to get three elements: the digital object itself, the metadata and the statement indicating the current provider.
- https://ezid.cdlib.org/search
- https://ezid.cdlib.org/id/ark:/87278/s6057dv2

Digital Object Identifier (DOI)

#CSIC

DOI 10.5281/zenodo.840080

- It was released in 1998, is currently one of the most extended and used solutions in the academic environment.
- The DOI is an indirect identifier for electronic documents based on Handle resolvers, that are a mechanism for permanent identification of digital content, which first tries to resolve the resource address and if it is not available, it returns the information or metadata attached to the identifier.
- Many journals and editorials use currently the DOI as a bibliographic identifier, complementing other standards like ISBN. Currently, about 133 million DOIs have been assigned by commercial and non-commercial providers that participate in the International DOI Foundation.
- The DOI format is composed of two sections: a numeric identifier that includes a prefix identifying the term as a DOI (10.) and a suffix identifying the publisher. The document is then identified with a separate code.

CSIC
Consejo Superior de Investigaciones Científicas

#CSIC ORCID

❐ Guidelines for the use and interpretation of assays for monitoring autophagy (3rd edition).
(PMID:26799652)

Abstract | Citations | BioEntities | Related Articles | External Links

Klionsky DJ . Abdelmohsen K . Abe A . Abedin MJ . Abeliovich H . Acevedo Arozena A . Adachi H . Adams CM . Adams PD . Adeli K . Adhihetty PJ . Adler SG . Agam G . Agarwal R . Aghi MK . Agnello M . Agostinis P . Aguilar PV . Aguirre-Ghiso J . Airoldi EM . Ait-Si-Ali S . Akematsu T . Akporiaye ET . Al-Rubeai M . Albaiceta GM . Albanese C . Albani D . Albert ML . Aldudo J . Algül H . Alirezaei M . Alloza I . Almasan A . Almonte-Beceril M . Alnemri ES . Alonso C . Altan-Bonnet N . Altieri DC . Alvarez S . Alvarez-Erviti L . Alves S . Amadoro G . Amano A . Amantini C . Ambrosio S . Amelio I . Amer AO . Amessou M . Amon A . An Z . Anania FA . Andersen SU . Andley UP . Andreadi CK . Andrieu-Abadie N . Anel A . Ann DK . Anoopkumar-Dukie S . Antonioli M . Aoki H . Apostolova N . Aquila S . Aquilano K . Araki K . Arama E . Aranda A . Araya J . Arcaro A . Arias E . Arimoto H . Ariosa AR . Armstrong JL . Arnould T . Arsov I . Asanuma K . Askanas V . Asselin E . Atarashi R . Atherton SS . Atkin JD . Attardi LD . Auberger P . Auburger G . Aurelian L . Autelli R . Avagliano L . Avantaggiati ML . Avrahami L . Awale S . Azad N . Bachetti T . Backer JM . Bae DH . Bae JS . Bae ON . Bae SH . Baehrecke EH . Baek SH . Baghdiguian S . Bagniewska-Zadworna A . Bai H . Bai J . Bai XY . Bailly Y . Balaji KN . Balduini W . Ballabio A . Balzan R . Banerjee R . Bánhegyi G . Bao H . Barbeau B . Barrachina MD . Barreiro E . Bartel B . Bartolomé A . Bassham DC . Bassi MT . Bast RC jr . Basu A . Batista MT . Batoko H . Battino M . Bauckman K . Baumgarner BL . Bayer KU . Beale R . Beaulieu JF . Beck GR jr . Becker C . Beckham JD . Bédard PA . Bednarski PJ . Begley TJ . Behl C . Behrends C . Behrens GM . Behrns KE . Bejarano E . Belaid A . Belleudi F . Bénard G . Berchem G . Bergamaschi D . Bergami M . Berkhout B . Berliocchi L . Bernard A . Bernard M . Bernassola F . Bertolotti A . Bess AS . Besteiro S . Bettuzzi S . Bhalla S . Bhattacharyya S . Bhutia SK . Biagosch C . Bianchi MW . Biard-Piechaczyk M . Billes V . Bincoletto C . Bingol B . Bird SW . Bitoun M . Bjedov I . Blackstone C . Blanc L . Blanco GA . Blomhoff HK . Boada-Romero E . Böckler S . Boes M . Boesze-Battaglia K . Boise LH . Bolino A . Boman A . Bonaldo P . Bordi M . Bosch J . Botana LM . Botti J . Bou G . Bouché M . Bouchecareilh M . Boucher MJ . Boulton ME . Bouret SG . Boya P . Boyer-Guittaut M . Bozhkov PV . Brady N . Braga VM . Brancolini C . Braus GH . Bravo-San Pedro JM . Brennan LA . Bresnick EH . Brest P . Bridges D . Bringer MA . Brini M . Brito GC . Brodin B . Brookes PS . Brown EJ . Brown K . Broxmeyer HE . Bruhat A . Brum PC .

Viola G . Viscomi MT . Vitale G . Vogl DT . Voitsekhovskaja OV . von Haefen C . von Schwarzenberg K . Voth DE . Vouret-Craviari V . Vuori K . Vyas JM . Waeber C . Walker CL . Walker MJ . Walter J . Wan L . Wan X . Wang B . Wang C . Wang CY . Wang C . Wang C . Wang C . Wang D . Wang F . Wang F . Wang G . Wang HJ . Wang H . Wang HG . Wang H . Wang HD . Wang J . Wang J . Wang M . Wang MQ . Wang PY . Wang P . Wang RC . Wang S . Wang TF . Wang X . Wang XJ . Wang XW . Wang X . Wang X . Wang Y . Wang Y . Wang Y . Wang YJ . Wang Y . Wang Y . Wang YT . Wang Y . Wang ZN . Wappner P . Ward C . Ward DM . Warnes G . Watada H . Watanabe Y . Watase K . Weaver TE . Weekes CD . Wei J . Weide T . Weihl CC . Weindl G . Weis SN . Wen L . Wen X . Wen Y . Westermann B . Weyand CM . White AR . White E . Whitton JL . Whitworth AJ . Wiels J . Wild F . Wildenberg ME . Wileman T . Wilkinson DS .

Authors: Same name, surname

Wrong reference

Difficult to measure some indexes

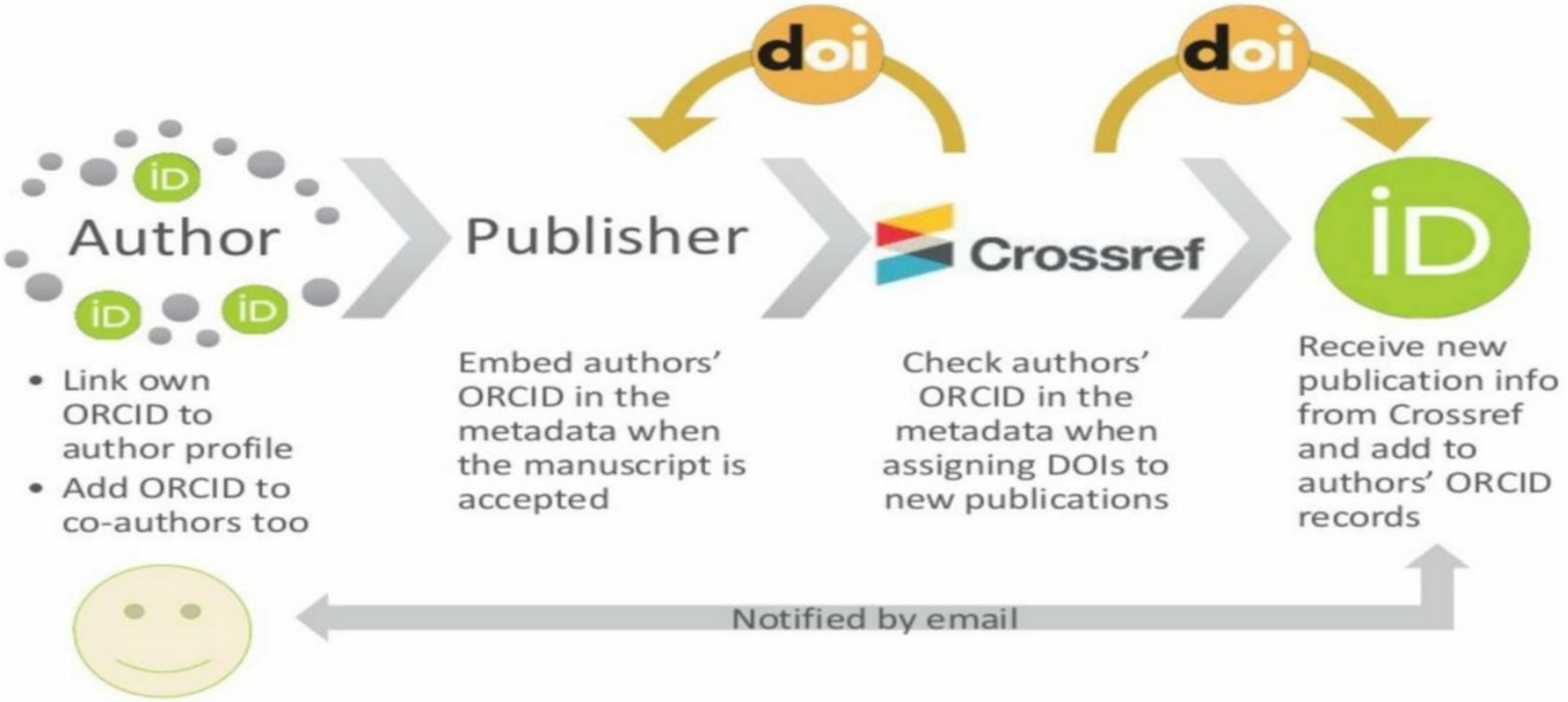
ORCID

CSIC
Consejo Superior de Investigaciones Científicas

#CSIC ORCID

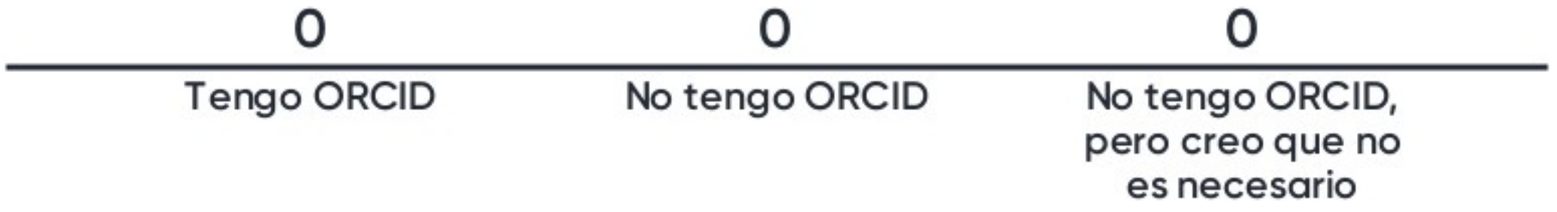

JOURNAL OF SPORTS SCIENCES, 2016
VOL. 34, NO. 16, 1564–1580
http://dx.doi.org/10.1080/02640414.2015.1123284
REVIEW ARTICLE
Numerical and experimental investigations of human swi
Hideki Takagi [a], Motomu Nakashima [b], Yohei Sato [c], Kazuo Matsuuchi [d] a
[a]Faculty of Health and Sport Sciences, University of Tsukuba, Tsukuba, Ibaraki, Japan; [b]Department of
Institute of Technology, Tokyo, Japan; [c]Nuclear Energy and Safety, Paul Scherrer Institute, Villigen, Sv
Tsukuba, Tsukuba, Ibaraki, Japan; [d]Exercise and Sport Science, The Univers ty of Sydney, Sydney, Aus
ABSTRACT
This paper reviews unsteady flow conditions in human swimming and identifies the lim
future potential of the current methods of analysing unsteady flow. The capability of co
fluid dynamics (CFD) has been extended from approaches assuming steady-state condit
sideration of unsteady/transient conditions associated with the body motion of a swimmer.
predict hydrodynamic
numerical methods ar
data reflecting local fl
limited, because at pr
could be improved if
swimming are expect
tant advances relatin
swimming: one is a m
to increased speed ar
ORCID
Hideki Takagi http://orcid.org/0000-0001-8797-7014
Motomu Nakashima http://orcid.org/0000-0002-1349-87
Yohei Sato http://orcid.org/0000-0001-5313-5767
Kazuo Matsuuchi http://orcid.org/0000-0002-7565-1858
Ross H. Sanders http://orcid.org/0000-0003-0489-3048
Search
ORCID
Connecting Research and Researchers
FOR RESEARCHERS
FOR ORGANIZATIONS
SIGN IN
REGISTER FOR AN ORCID ID
LEARN MORE
Hideki Takagi
ORCID ID
orcid.org/0000-0001-8797-7014
Country
Japan
Keywords
Swimming, Water Polo, Hydrodynamics, Sports Engineering
Websites
Hideki Takagi's Home Page
Other IDs
Scopus Author ID: 55127314200
Biography
Professor at Faculty of Health and Sport Sciences(Tsukuba
Education (3)
Mie University: Tsu, Mie, Japan
1992-04 to 2000-12-01
Doctor of Engineering (Graduate school of Mechanical Engi
Source: Hideki Takagi
University of Tsukuba: Tsukuba, Ibaraki, J
1985-04 to 1987-03-25
Master of Science (Graduate School of Comprehensive Hum
Source: Hideki Takagi
University of Tsukuba: Tsukuba, Ibaraki, J
1981-04 to 1985-03-25
Bachelor of Physical Education (School of Health and Physic
Source: Hideki Takagi
Employment (2)
University of Tsukuba: Tsukuba, Ibaraki, J
2001-09 to present
Professor (Faculty of Health and Sport Sciences)

ORCID



CSIC
Consejo Superior de Investigaciones Científicas

¿Tienes ORCID?



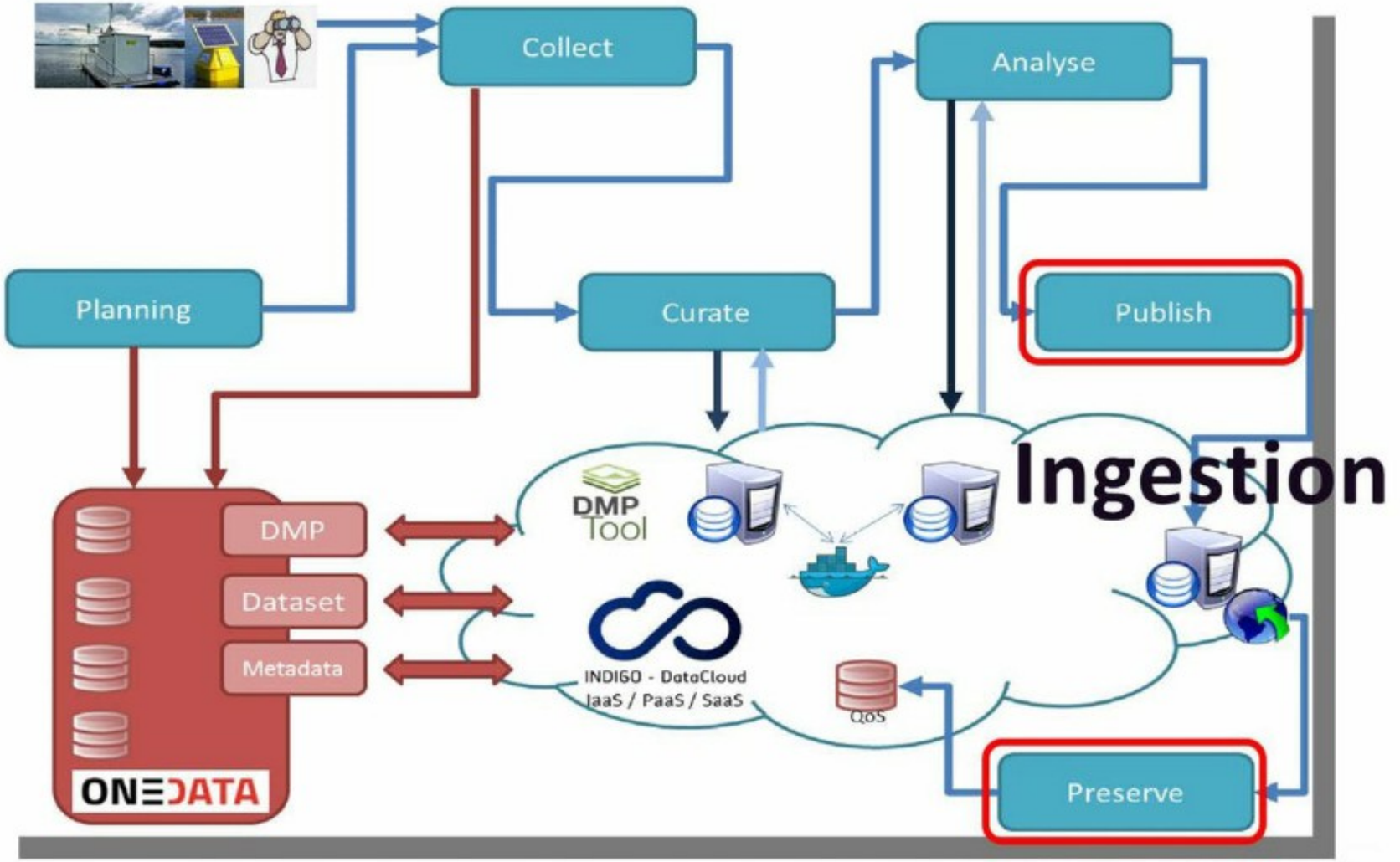
0

INDIGO Data Life Cycle ("6S")

#CSIC

Stage 5: Ingest (& Publish): including other steps like "Access", "Use" or "Re-use", in this stage, data is normally associated to metadata, gets a persistent identifier (a DOI) and is published in an accessible repository or catalogue, under a format that makes it useful for further re-use.

Stage 6: Preserve: "store" both data and analysis for long-term. Licenses and methods need to be taken into account.



Preservation

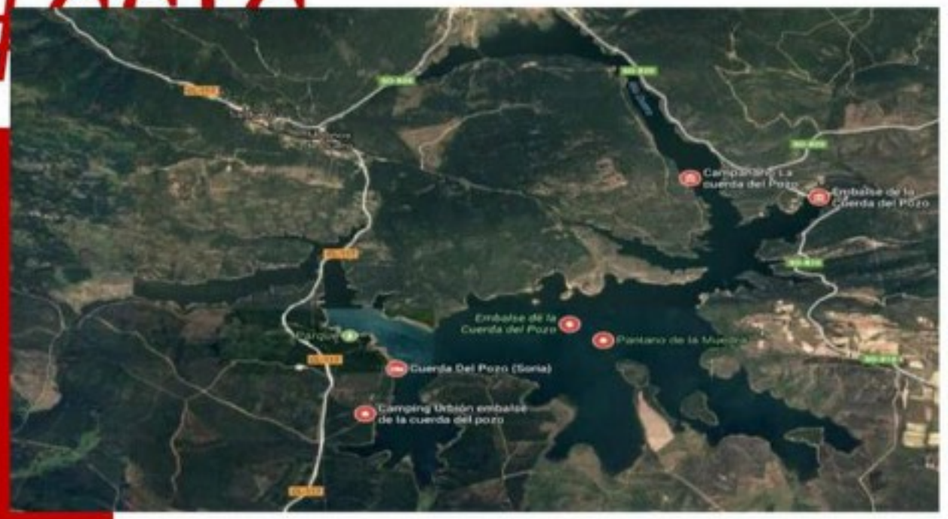
#CSIC

- Oriented to long-term use of data.
- Involves infrastructures, software, hardware, formats… Dealing with obsolescence.
- Must ensure the reproducibility.
- Virtual Machines (images, dockers) + software + data
- Use of Persistent identifiers:
  - PIDs, DOIs
- Backup, geographically distributed, hardware formats, well-documented, format migration if needed…

Ejemplos/Proyectos con ciclo de vida claro

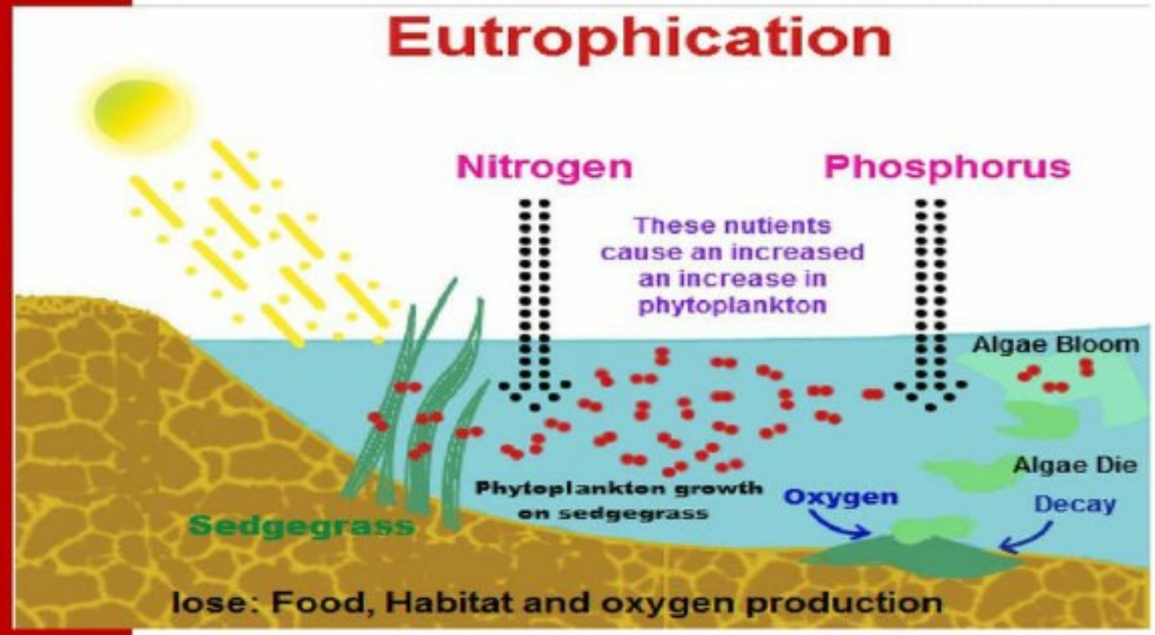
Example: CdP Project

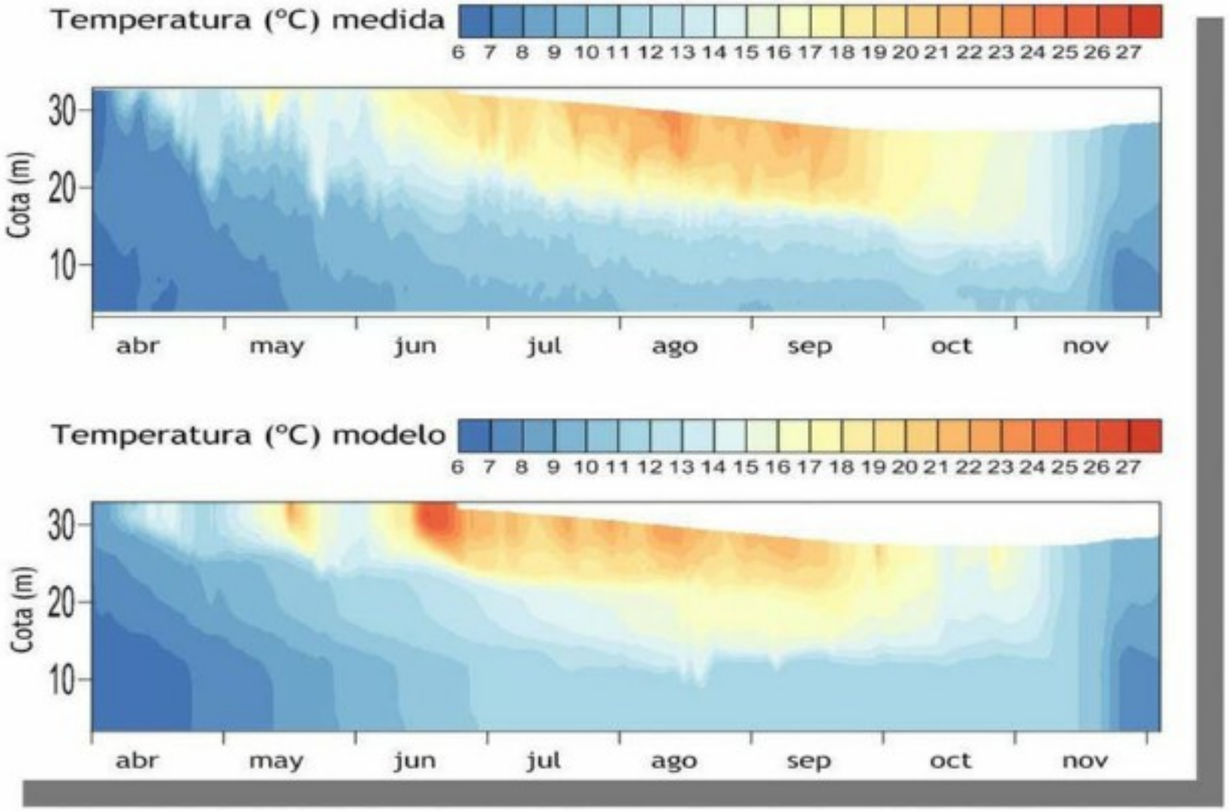
#CSIC



Example: CdP Project

#CSIC





CSIC
Consejo Superior de Investigaciones Científicas

Example: CdP Project

#CSIC

- Datos necesarios:
  - Temperatura de la columna del agua – ~~Ristra de sensores.~~ Jaula (sube y baja)
  - Irradiancia (visible) en la columna de agua – ~~Ristra de sensores.~~ Jaula
  - Turbidez - ~~Ristra de sensores.~~ Jaula
  - Carbono (orgánico total)/biodegradable?. **Medido en laboratorio.**
  - Nitrógeno (amonacial? Nitratos). Jaula
  - Fósforo. **Medido en laboratorio.**
  - Potential RedOx. Medida indirecta de los compuestos oxidantes del agua.
  - Presión.
  - Corrientes. **Problemático**.
  - Mapa de viento sobre la superficie del embalse. **Problemático** – Solución: Interpolación? Modelos de viento?
  - Humedad del aire
  - pH
  - Oxígeno Disuelto
  - Concentración de algas: clorofila, cianos. Sensores, jaula.
- Distribución Espacial/Temporal

Example: CdP Project

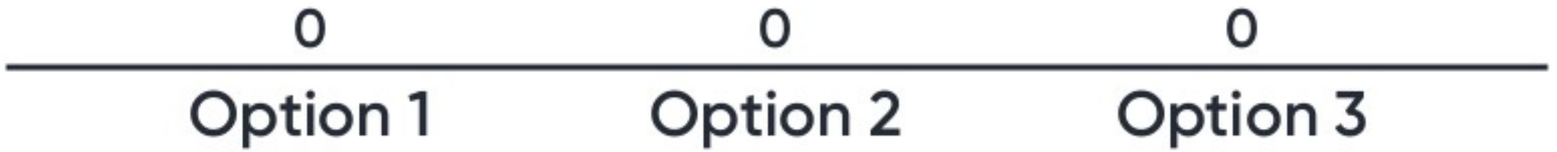
- Buscar fuentes de datos históricas:
  - algas? Probablemente no.
  - Problemas en el agua de Soria? Diario de Soria? Portal de transparencia ayuntamiento de Soria?
  - Datos de depuradora: Empresa privada?
- Calibración de sensores, tener en cuenta la degradación.
- Despliegue: Sensores, muestras, etc. Sensores: precisión, error, fiabilidad, etc.
- Sistema de adquisición: reloj (toma de medidas), decidir la frecuencia de las medidas de cada variable, memoria, duración del ciclo de vida.
- Sistema de comunicaciones
- Definir el modelo de datos. Modelo lógico, físico.
- Servidor de bases de datos, Disco.
- Descripción de los datos:
  - **Contexto**: propiedad de los datos, proyecto, instituciones…
  - **Administrativos**: licencia, copyright, Tipo de sensor (marca, modelo, limitaciones, garantía, fecha del sensor), modelo de información,
  - **Estructurales:** Variables, Unidades, Datos técnicos del sensor (Precisión), Rangos, Condiciones de medida, Localización, tipo de fichero/objeto digital.

Example: CdP Project

#CSIC

- Preprocesado:
  - Conversión de unidades, definir decimales.
  - Homogenización de formatos
  - Selección, filtrado.
  - **Assure/Curación**: limpieza de datos erróneos, picos extremos.
- Procesar los datos:
  - Buscar correlaciones entre datos de variables y aumento de algas.
  - Meter datos en un sistema deep learning para intentar predecir.
  - Procesar un modelo.
  - Construir un modelo sobre cada variable.
  - Analizar datos procesados: Salida de los modelos, validación de modelos, etc.
- Preservación
  - Decidir cómo de críticos son tus datos.
  - Réplica de la Base de Datos (en cloud externo)
  - Archivo de datos en cintas

Image Choice



0

#CSIC

Ciclo de Vida de la Investigación

Fernando Aguilar
aguilarf@ifca.unican.es

CSIC
Consejo Superior de Investigaciones Científicas