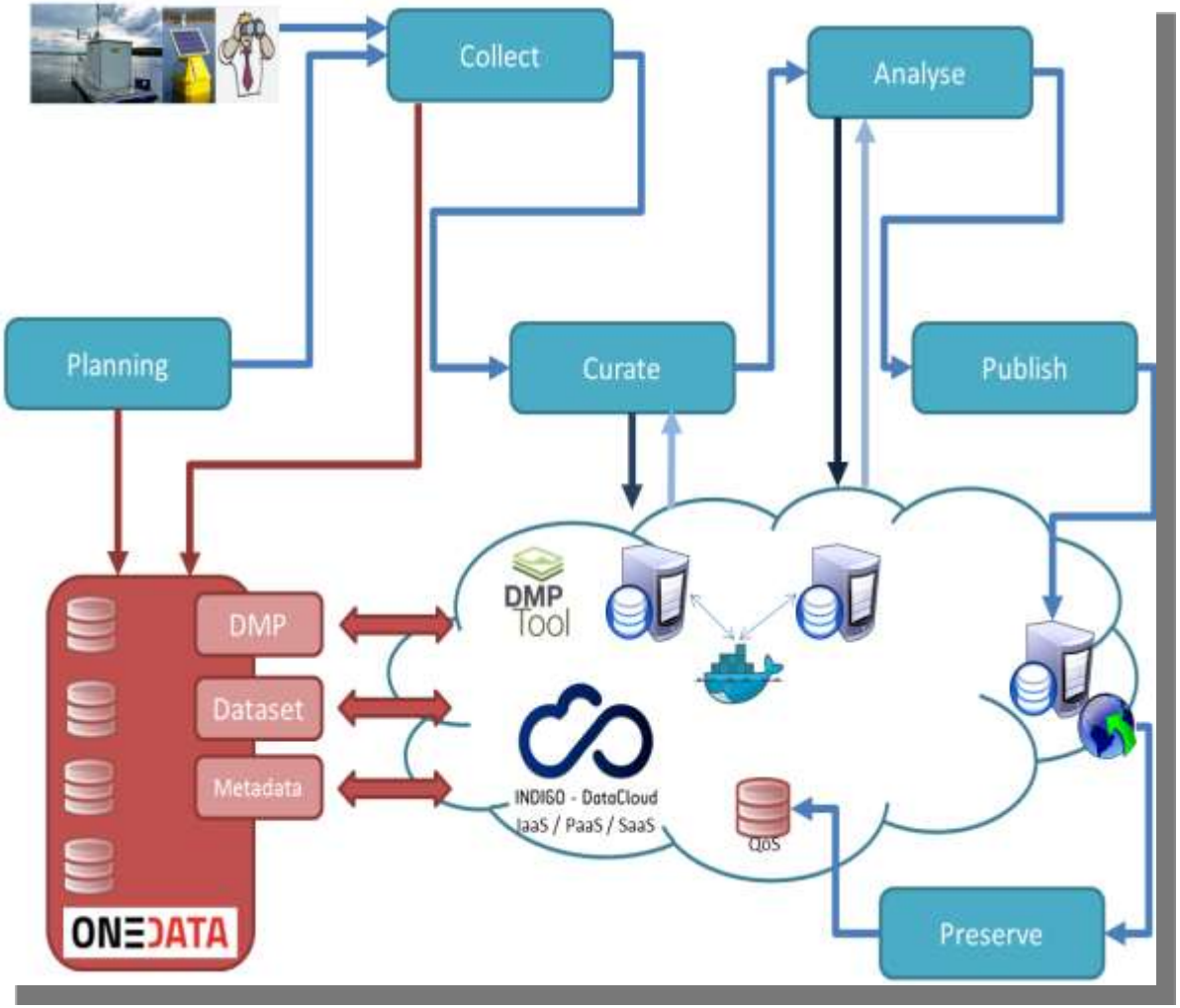
Data Management Planning

Fernando Aguilar

INDIGO Data Life Cycle (“6S”)

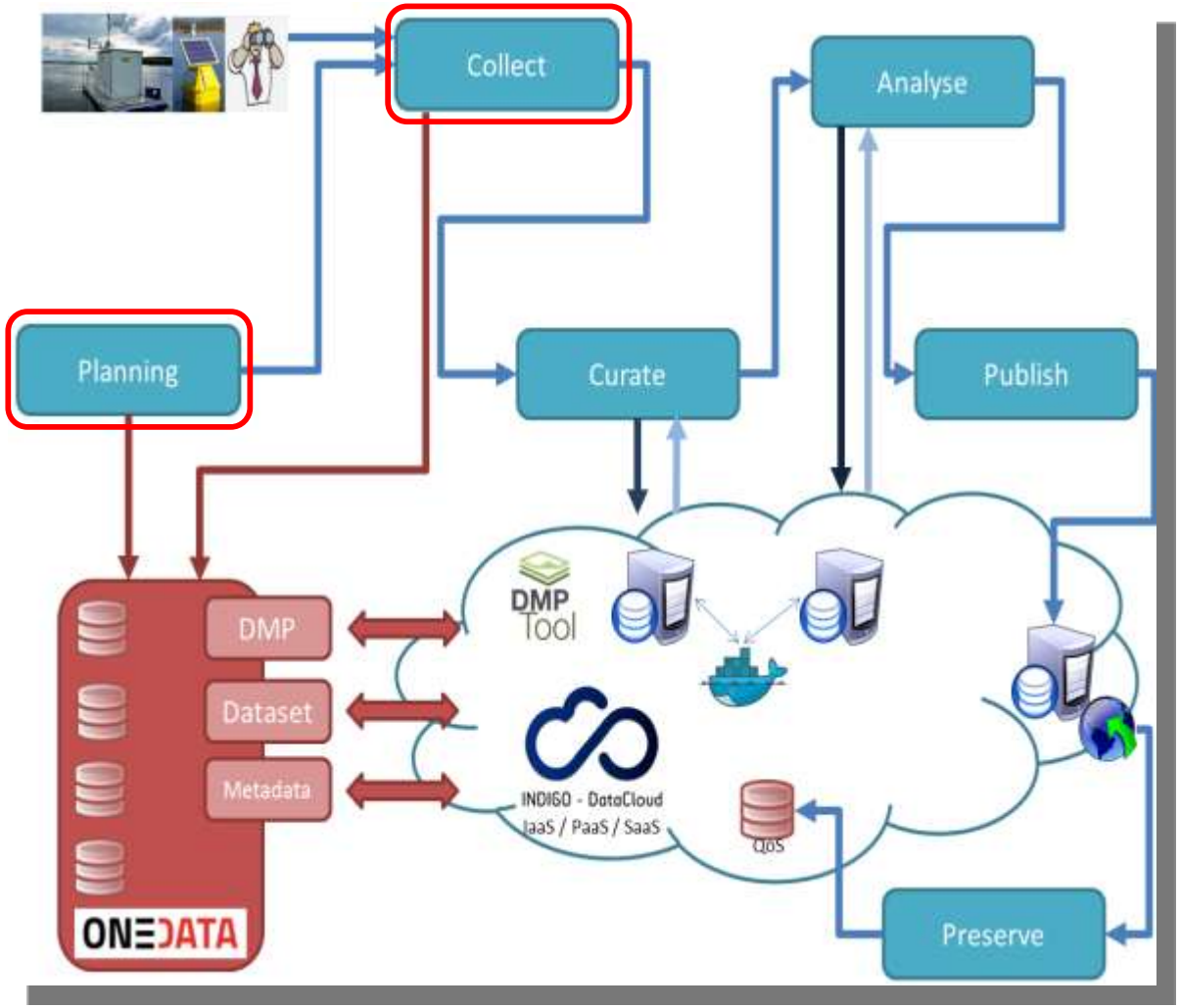


INDIGO Data Life Cycle ("6S")

Enfoque que tomamos en la asignatura para planificar un

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

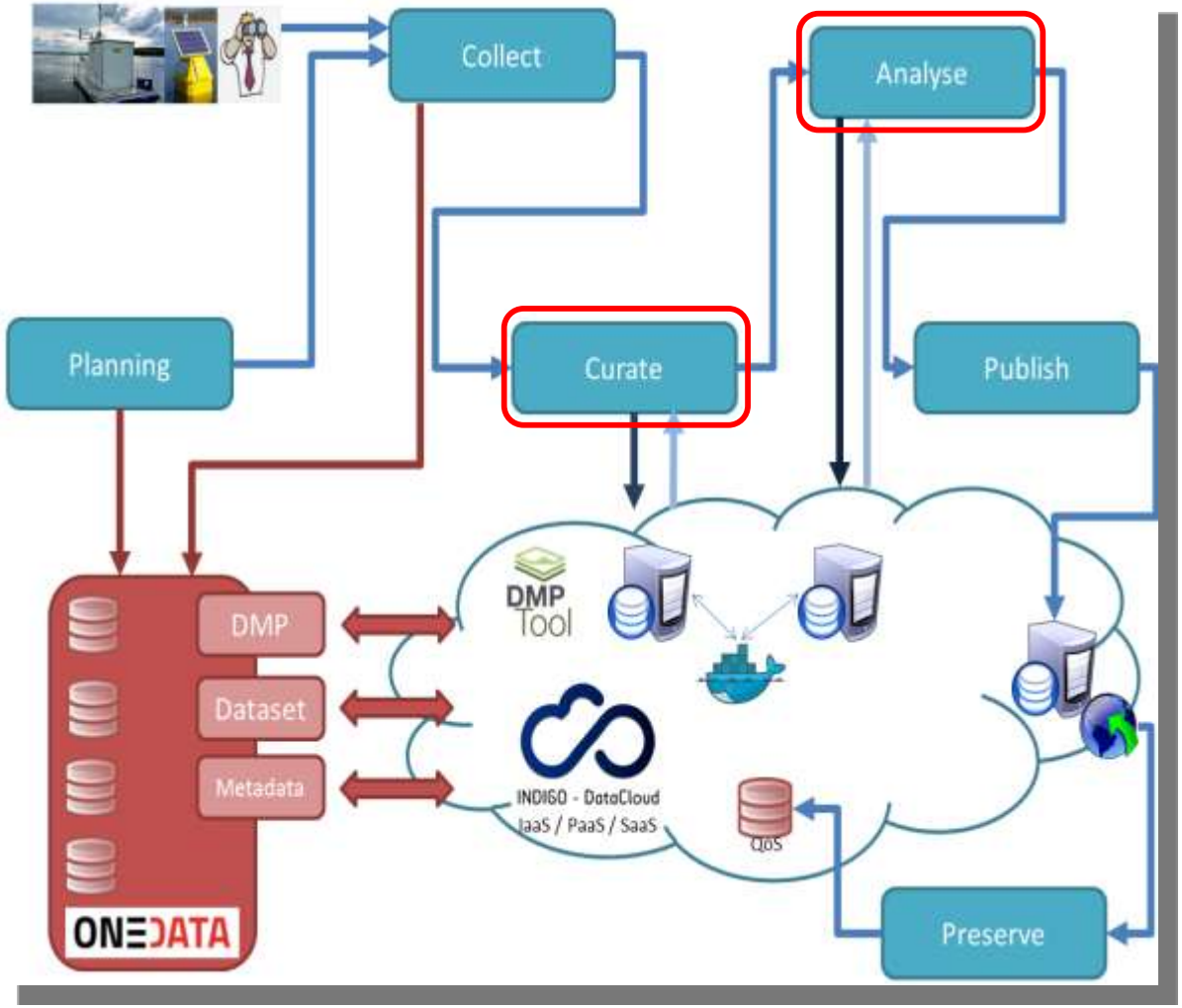
Stage 2: Collect: including both creation and acquisition, it is the process of getting data, in different ways. A storage service is needed as well.



INDIGO Data Life Cycle (“6S”)

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.

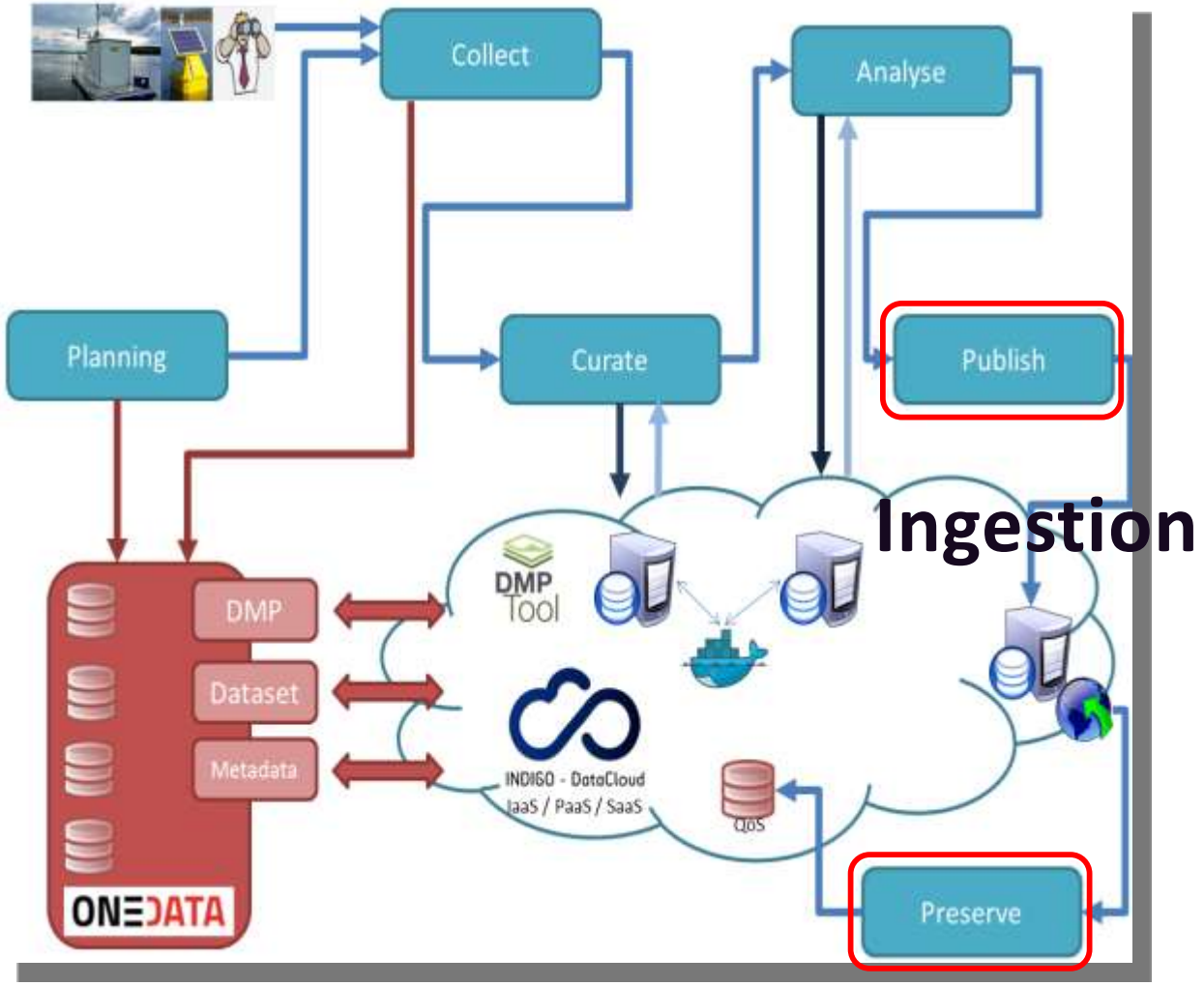
Stage 4: Analyse: an optional step also called “Process”, that implies performing different actions to give the data an added value and get new derived data.



INDIGO Data Life Cycle ("6S")

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.



Data Management Plan

Primera etapa del ciclo de datos

- 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".

Data Management Plan

- Formal Document (*Document?*) o una web que pueda organizar todo.O scripts
- Outlines what you will do with your data during and after you complete your research
- Ensure your data is safe for the present and the future

From University of Virginia Library

Why DMPs?

- **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

- 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
  - RDA Working Groups
- New developments towards machine-actionability.
- Oriented to create “FAIR” Data.

Concepts (To be clear…)

- **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. Objeto digital: Fichero, video, articulo, pag web, libro, etc.
- **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

What is FAIR Data?

- 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.).

Datos y metadatos

-> que sea explotable por una máquina

Interaccionable

Vocabulario tipo estándar. Ejemplo: json, xml

Implementation approach

- **Findable** - standards for describing the dataset with the relevant metadata;
- **Accessible** - standards for represent and access the data according to the defined usage license;
- **Interoperable** - standards for machine readable descriptions of the (meta)data and (semantic)annotation;
- **Reusable** - standards for semantic annotation of the (meta)data supporting machine reasoning, and standards for defining data provenance and support citation;

- The standards include technologies (e.g., RDF, nano pub, JSON, OWL, etc.) as well as protocols and APIs.

DMPs in funding programs

- 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

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

Forma general mundial sobre los DMP

1. Information About Data & Data Format

1.1 Description of data to be produced

- Experimental
- Observational
- Raw or derived (e.g. encuestas)
- Physical collections
- Models and their outputs
- Simulation outputs
- Curriculum materials
- Software
- Images
- Etc…

DataONE

1. Information About Data & Data Format

1.2 How data will be acquired

- When?
- Where?

1.3 How data will be processed

- Software used
- Algorithms
- Workflows

CC image by Ryan Sandridge on Flickr

1. Information About Data & Data Format

1.4 File formats

O base de datos

- Justification
- Naming conventions

Como llamas a los archivos

1.5 Quality assurance & control during sample collection, analysis, and processing

e.g. quiero almacenar temperaturas:PARAM_FECHA_LC

1. Information About Data & Data Format

1.6 Existing data e.g. imágenes por satélites

- If existing data are used, what are their origins?
- Will your data be combined with existing data?
- What is the relationship between your data and existing data?

1.7 How data will be managed in short-term

- Version control
- Backing up
- Security & protection
- Who will be responsible

2. Metadata Content & Format

Metadata defined:

- Documentation and reporting of data
- Contextual details: Critical information about the dataset
- Information important for using the data
- Descriptions of temporal and spatial details, instruments, parameters, units, files, etc.

2. Metadata Content & Format

2.1 What metadata are needed

- Any details that make data meaningful

2.2 How metadata will be created and/or captured

- Lab notebooks? GPS units?
- Auto-saved on instrument?

2.3 What format will be used for the metadata

- Standards for community
- Justification for format chosen

3. Policies for Access, Sharing, Reuse

3.1 Obligations for sharing
- Funding agency
- Institution
- Other organization
- Legal

3.2 Details of data sharing
- How long?
- When?
- How access can be gained?
- Data collector rights

3.2 Ethical/privacy issues with data sharing

3. Policies for Access, Sharing, Reuse

3.4 Intellectual property & copyright issues

- Who owns the copyright?
- Institutional policies
- Funding agency policies
- Embargos for political/commercial reasons

3.5 Intended future uses/users for data

3.6 Citation

- How should data be cited when used?
- Persistent citation?

CC image by buddawiggi on Flickr

4. Long-term Storage & Data Management

4.1 What data will be preserved

4.2 Where will it be archived

- Most appropriate archive for data
- Community standards

3.6 Data transformations/formats needed

- Consider archive policies

4.4 Who will be responsible

- Contact person for archive

5. Budget

5.1 Anticipated costs

- Time for data preparation & documentation
- Hardware/software for data preparation & documentation
- Personnel
- Archive costs

5.2 How costs will be paid

H2020 DMP

El template que pide la comunidad europea

1. DATA SUMMARY

What is the purpose of the data collection/generation and its relation to the objectives 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')?

Todos los puntos a realizar en el DMP del proyecto

H2020 DMP

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.

H2020 DMP

- **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?

H2020 DMP

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

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?

H2020 DMP

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?

H2020 DMP

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.

H2020 DMP

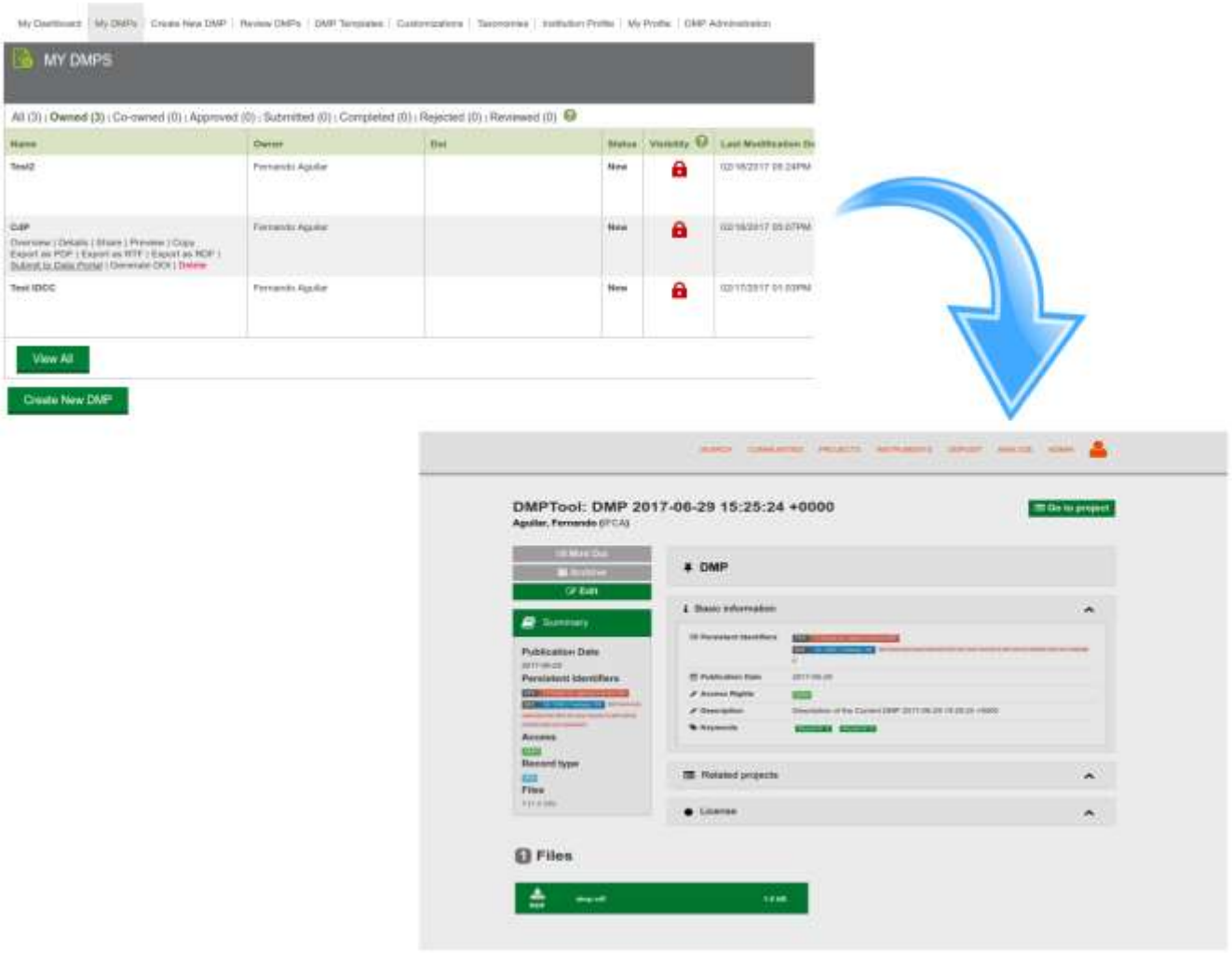
- DMP services. Example: DMPonline

https://dmponline.dcc.ac.uk/

DMPs - Proyectos

-

Regarding Machine-actionability



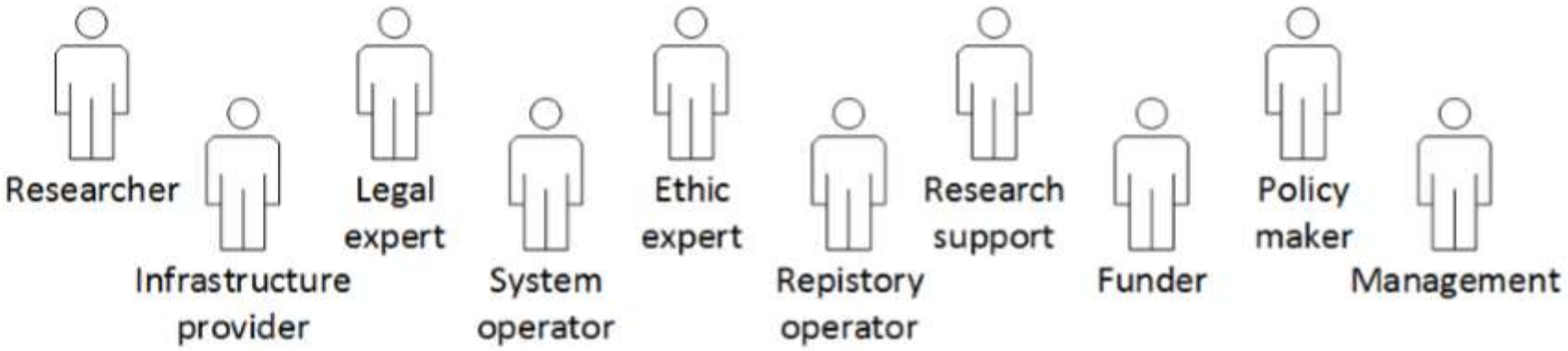
Regarding Machine-actionability

Semantic layer:

https://bioportal.bioontology.org/ontologies/SWEET

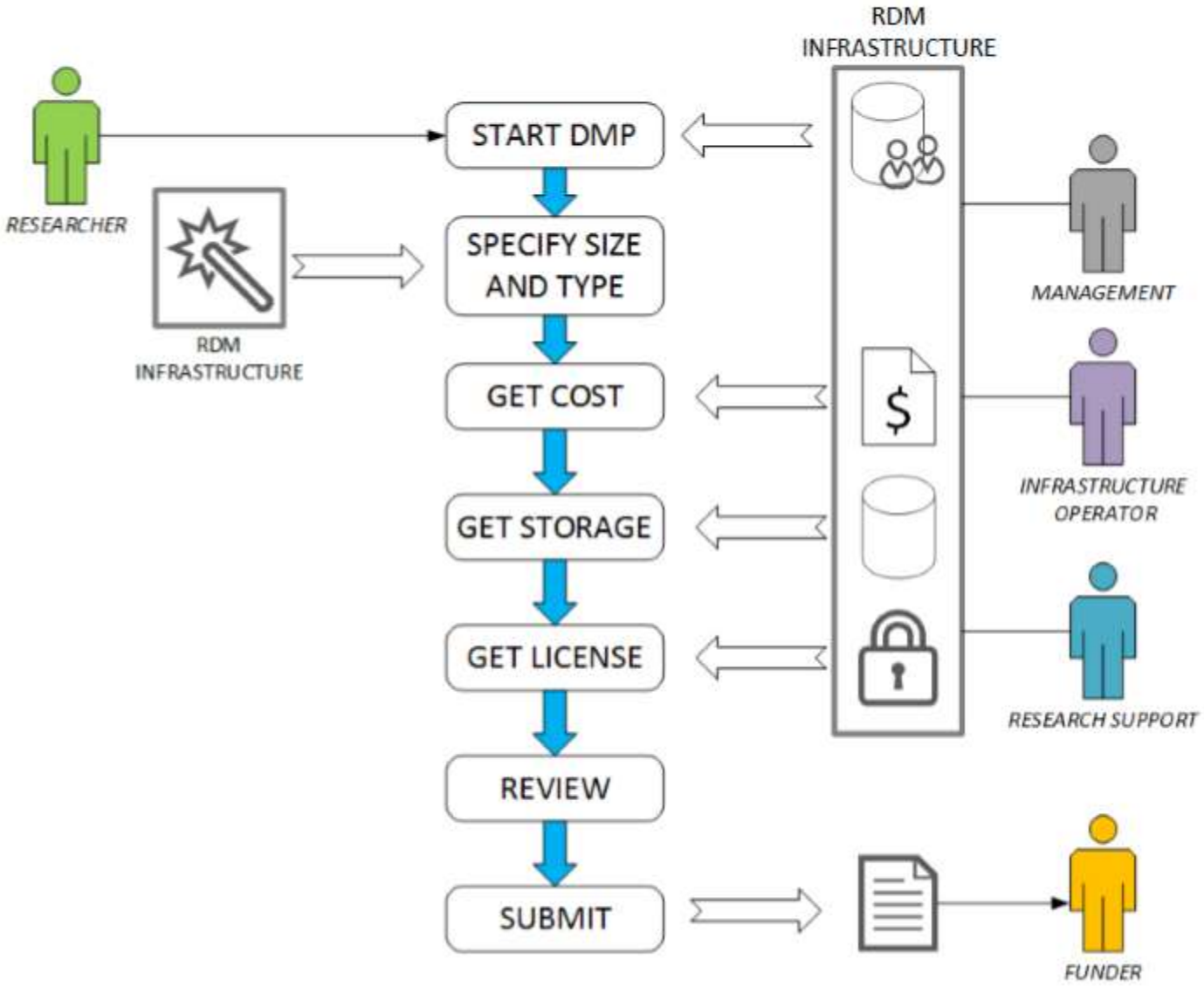
Research data lifecycle

- Stakeholders involved in research data management
  - require information at certain stages
  - can provide information if requested at a proper stage
- Many problems can be avoided when
  - timing is right
  - information flow is ensured



Tomasz Miksa, Paul Walk, Peter Neish

Automated Data Management Workflow



Example

- Current DMPs – model questionnaires

```
<administrative_data>
    <question>Who will be the Principle Investigator?</question>
    <answer>The PI will be John Smith from our university.</answer>
</administrative_data>
```

- Machine-actionable DMPs – model information

```
"dc:creator":[ {
            "foaf:name":"John Smith",
            "@id":"orcid.org/0000-1111-2222-3333",
            "foaf:mbox":"mailto:jsmith@tuwien.ac.at",
            "madmp:institution":" AT-Vienna-University-of-Technology"
 } ],
```

Example

- Currently available – not very useful

```
<administrative_data>
    <question>Who will be the Principle Investigator?</question>
    <answer>The PI will be John Smith from our university.</answer>
</administrative_data>
```

- Machine-actionable DMP

```
"dc:creator":[ {
        "foaf:name":"John Smith",
        "@id":"orcid.org/0000-1111-2222-3333",
        "foaf:mbox":"mailto:jsmith@tuwien.ac.at",
        "madmp:institution":"AT-Vienna-University-of-Technology"
 } ],
```

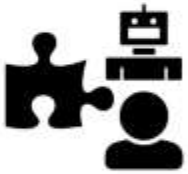
Develop own concepts and vocabularies only when needed

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



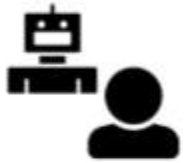
4 Describe—for both machines and humans—the components of the data management ecosystem

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

New developments

- https://github.com/oblassers/dmap
- https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard

Herramientas que permiten trabajar en los DMP de forma más automatizada