# 02 - Digital Preservation

### At what levels are digital objects threatened?

Digital objects are threatened on **three different levels**:

1. **Bit Rot - Physical Preservation**: How to keep 0's and 1's intact (bit loss on hard drives, magnet tape, …)
2. **Object Formats Logical Preservation:** How to remain able to open a file, run a program (Difficult to emulate old Atari games ae, software not supported by the new OS, ..)
3. **Authenticity, Interpretability - Semantic Preservation:** How to ensure that we can understand/read data correctly (measurement formats, …)

In order to preserve digital information, we need a solution for problems from all 3 layers!

### What are the time intervals at each level?

1. Bit-Level strategies mostly concern themselves with the life expectancy of the medium the data is persisted on (e.g. longevity of hard drives, magnet tapes, …)
2. Logical strategies mostly buy time-buffers to find better solutions that allow data to be safely moved to a time-safe standard format, like tech museums, migration, and/or emulation.
3. Semantic: Mostly add up to creating more metadata, that also has to be standardized and maintained

→ For the bit level threats, the time frame is mostly the life duration of a physical medium and/or hardware system. For the other two, nearly constant adaptation and improvement are required, as technology and standards constantly evolve over time.

### How can we identify objects at risk?

Every digital object is at risk to fall under one of the threats mentioned above, no matter which application or domain. Making oneself aware of the issues that can arise with the current format of the digital object allows to think of the appropriate risks.

### What can we do to mitigate the risk?

For all of the three leves, there are quite a few strategies to omit the risks:

1. **Bit Level:** Redundant storage, Controlled storage conditions (z.b. Datacenters with CO2 showers in case of fire), Regular maintenance like tape rewinding or disk spinning, Distribute the storage geographically, Maintain access devices to storage
2. **Logical Level:** Maintain costly Hardware-Museums of usable HW and SW, Migrate or Emulate data with risk of losing some information or changing the data, Use standardized and open formats for the data to ensure readability

and machine actionability, Extract information to abstract format like XML, Print to paper if all very traditional

3. **Semantic Level:** Semantic enrichment of data, Lots and lots of metadata, Documentation, Document intended meaning and interpretations of data

### How can we recover if mitigation fails / is missed?

Look at who is responsible, what can be done with current tech to recover and how will we manage such problems in the future → DMP should cover this in any case.

## 03 - Open Archival Information System (OAIS) Models

OAIS is a **reference model** and NOT a concrete implementation of a preservation system!

### Model View of an OAIS System

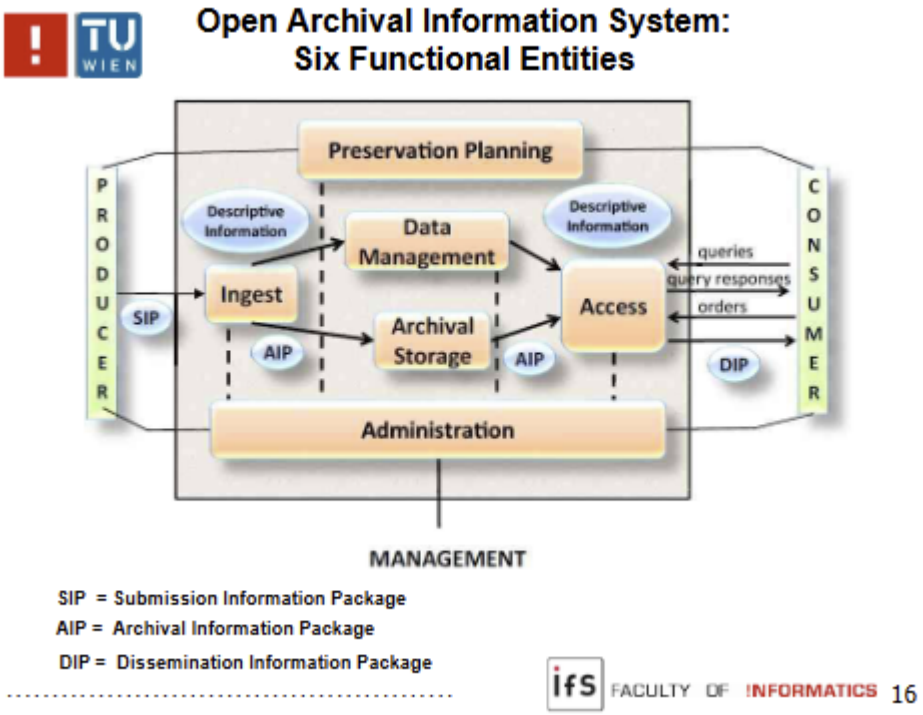
A **Producer** provides information to be preserved (e.g. the client systems or persons).
The **Management** role is played by those who set the OAIS policy.
**Consumers** are the people or client systems that interact with OAIS to find and acquire preserved data.

### OAIS Information Definition

**Information** is always represented by some type of data. **Data** interpreted using **Representation Information** yields Information. An **Information Object** provides clear identification and understanding of the **Data Object** and associated **Representation Information.**

An **Information Package** holds both Content Information and **Preservation Description Identifiers (PDI)**.

### Functional entities in an OAIS




- **Ingest:** Provides a service and functions to accept Submission Information Packages from producers. Prepares the content for storage (creates an AIP) and for management. Generates the descriptive Information for the AIP.
- **Archival Storage:** Provides services to store, maintain and retrieve AIP packages. Also does disaster recovery and error checking.
- **Data Management:** Provides services to populate and maintain and access both descriptive information that identifies and documents archive holdings and interal archive administrative data (e.g. a DB for all the metadata stuff).
- **Administration:** Manages the operation of the archival system (system config, submission agreements, standards and policies, audit of submissions, …)
- **Preservation Planning:** Monitor the environment of the OAIS and provide recommendations to ensure that information stored remains accessible for the long term even if the original computing environment becomes obsolete (Strategies, Monitoring of technological evolution HW and SW, Migration tools and opportunities,...)
- **Access:** Allows customers to determine the existence, description, location and availability of information persisted in the OAIS. Allows to request and receive data for consumers.

### Digital Migration

Motivators can be **Media Decay**, **Increased Cost-Effectiveness**, **New User Requirements** or **Proprietary Software Evolution.**

There exist four approaches to migrating information, in order of risk of information loss:

1. **Refreshment**: Media replacement with no bit changes
2. **Replication:** Noch change to the package of content information bits (copy file to a new location)
3. **Repackaging:** Some bit changes to packaging information (e.g. putting together files in a new folder structure)
4. **Transformation:** Change content information to a new format, either reversible or irreversible by an algorithm.

# 04 - Data Management Plans  (DMPs)

### Why and what for we need DMPs?

**Reproducibility of experiments** for research during and after the research has concluded. It ensures the **data is safe for the present and the future.** We need DMPs as an **awareness tool** that makes you **think about what happens to the data** (how you use and get it, what SW and infrastructure, how data sharing works,...). It helps to **organize oneself** better.
→ But also useful for other areas, not only research, overlaps with **FAIR principles** in large parts.

### How to improve your own data management?

→ By creating a **DMP** using either a **template** checklist or a **software tool**. It requires you to have **honesty with your workflow** to be efficient and useful. Requires to **think about things early** on and **plan your budget** → Decisions now alter what is possible later on.

### What is a DMP and what kind of information does it contain?

A DMP is a **living document**, usually generated by a template or software tool with a questionnaire. It concerns itself with the data used during and after research for **reproducibility**. It ensures **data is accessible** in the **present** and **future** (= for the long term). It consists of the following **five parts**:

1. **General Information:** Contains information about who is **responsible for the administration** of the data and other parts of the **data management responsibilities**. To this end, **unique ORCID** (=people identifiers) are used. These allow people to be uniquely identified even after moving, changing names, etc.

2. **Data Characteristics:** Contains a summary about what the data represents. It describes the **type of data**, **source** it was received from (measurement, simulated, experiments, …), the **volume of data** used and the **subsets** used

and a **decription of the wholedata** used along with the **formats** and **usage within the community**.

3. **Documentation and Data Quality:** Used to **encapture the Metadata** of the data. Helps to **understand and interpret the data**. Allows to read about details of the experimental setup. Uses community standards (**DublinCore, DataCite**, …) in compliance with the **FAIR principles**.

4. **Data Storage, Sharing, and Long-Term Preservation:** Concerns itself with **how data will be stored and managed** during and after research (convetions, versioning, backups, access restrictions, responsibilities,...). Also manages how the data will be available for sharing and how it will be archived. These problems can all be **addressed by using a data repos**itory that **archives the data** and **supports sharing** along the **FAIR principles**.
   → More on that in the mDMPs part.

5. **Legan and Ethical Aspects:** Which licenses are used, is sensitive data to be censored, and embargo periods of the data. Can the data even be published openly?

# 05 - Data Citation

### Why should we want to cite data?

- **Egoistic Motivation**: First of all, we cite data to **show a solid basis** for our research. We **build on others' work** so we do not have to prove things again that have already been done. This **speeds up the scientific process** and allows for the **basis of discourse**.

- **Enable Reproducibility and re-use**: Core to the scientific method. Reproducibility **requires Citation and unique identification.** Data citation **increases impact**, **builds trust,** and **fosters reuse**.

- **It is the right thing to do to give credit!**

There are **8 principles of data citation** that should encourage communities to develop practices that embody uniform data citation principles:

1. **Importance**: Data Citations are as important as those of publications.
2. **Credit and Attributions**: Give normative and legal attribution to data contributors.
3. **Evidence**: If claims rely upon data (= used as evidence for the claim), cite it.
4. **Unique Identification**: Data Citations should use a unique identifier that is **machine-actionable**.
5. **Access**: Data Citations should include infos for human and machine on how to access the data.
6. **Persistence:** Unique identifiers and Metadata should persist beyond the data's lifespan.

7. **Specificity and Verifiability:** Include infos about provenance and fixity that identify the exact granularity, timestamp, or timeslice of the data used.
8. **Interoperability and Flexibility**: Citation methods should be flexible to accommodate various practices in different communities.

**Benefits of Citation**: Identification, Documentation, Context, Impact, Transparency, Reproducibility, Reuse

### What identifier system should I use?

A identifier is a symbol used to **uniquely identify an object**. It references a **location**, **provides metadata**, and **can be resolved**.

To avoid link rot, one should **use persistent identifiers** (PIDs). The **delegation method** can be used to always resolve the unique identifier via a software portal. Systems are **Handle Systems**, **Digital Object Identifiers (DOIs),** as well as **PURL** and **ARK**.

→ The best ones for **digital objects** to use are **DOIs**, which are **Handle URNs enriched with Metadata**.
→ For **people, ORCID is used**, which works similarly to DOIs.
Using a handle system allows for global resolution of a handle (also DOI or ORCID).
A store used to **store metadata and identifiers and register DOIs** is for example **DataCite**.

**ARK is a free version of DOI**, which is **good for the early life cycle** of a project, but **can be deleted → may not be persistent forever**!

### What are the challenges in data identification and citation?

**Data is usually dynamic**, it can be **difficult to reproduce and track the exact granularity** and **version** of the data used in research. It is therefore also **difficult to use just a citation to uniquely identify the exact data** used in the experiments.

### How should we do it, according to the RDA WG?

A solution to this issue is to use a **query store.**
Researchers are provided **software with a query interface to the data**. The **store versions the data with timestamps** in the background.
If researchers want to **access the data**, they **define a query to the data** specifying **exactly what they need**:

- The **query store then stores this query** and **generates a PID** for it. → **Queries are unique** now and **can be identified** if people want to **gather the same data**.
- This **PID is enhanced with a timestamp** for **re-execution against the versioned DB**.
- It also **re-writes the query** using **normalization**, **sorting,** and **historical mapping**.
- It also provides a **hash of the result set** for **validation** and to **identify correctness** later on.

Query stores also provide **machine actionable interfaces via APIs** and can **handle stable migrations**. They are also **scalable** for data versioning and evolution.

**Benefits of using a query store**:

- Allows to **uniquely identify, retrieve and cite the exact data subsets** used in experiments with **minimal storage overhead**
- Allows **retrieving data as it was** AND using the **current view** as well.
- Can provide **valuable provenance data**
- Supports **verification of correctness** and **authenticity** using metadata such as checksums.
- Works for **all types of data.**

# 06 - FAIR principles

### Why we need FAIR principles

**FAIR** stands for **Findable**, **Accessible**, **Interoperable,** and **Reusable.** It concerns itself with moving **data** to a **searchable format**. In **response to FAIR, many solutions were developed** like **data repositories**, **data management plans,** and **persistent identifiers**.

FAIR allows to uniquely find, access, and identify datasets and reuse them → Basis for most other methods in this lecture!

### Differences between specific principles

1. **Findable:**
    a. (Meta)-Data has a **globally unique persistent identifier** (z.b. DOI, ORCID).
    b. Data is described in **rich metadata**.
    c. **Metadata clearly** describes and **includes the identifier** (z.b. DOI) of the data.
    d. (**Meta)-Data** are **registered or indexed** in a **searchable resource** (z.b. Zenodo).

2. **Accessible**:
    a. (Meta)-Data are **retrievable** by the **unique identifier** using a stan**dard communication protocol** → **Open**, **free**, **universally used** and supports **auth** and **aut** (z.b. OAuth2)
    b. **Metadata is accessible** even **after** the **original data are no longer available**.

3. **Interoperable**:
    a. (Meta)-Data uses a **formal, accessible, shared and broadly available language** for **knowledge representation** (**RDF**, **JSON, CSV + README**, …)
    b. (Meta)-Data uses **vocabularies that follow FAIR** (z.b. **DublinCore**).
    c. (Meta)-Data **include qualified references** to other (**meta**)-**data** (ae. X is derived from Y)

4. **Reusable**:
    a. (Meta)-Data are **richly described** with a **plurality of accurate and relevant attributes**. At least the following: **Usage License** (z.b. CC-BY), **detailed**

**provenance** (how data was derived, who manages it etc. → PROV-O as system) and must meet **community standards**.

### Relation between FAIR principles, machine-actionability, and open data

**Machine actionability is core** to all of the four FAIR principles → The more machine actionable data is, the better the data itself is! **Openness is however not required by FAIR**. There can also be closed data that is still fair, by specifying after finding the data what the access criteria are!

→ **FAIR is always machine-actionable and can be open**.

### How to apply FAIR principles in practice

You can use a **do-it-yourself FAIR assessment** via various online tools. This provides you with a checklist on what should be done if you have some differences between your current workflow and the FAIR standard. → There are also some **automated FAIR assessment tools**!

One could also use a tool such as **RO-Crate** to create a **packaged FAIR Digital Object**.

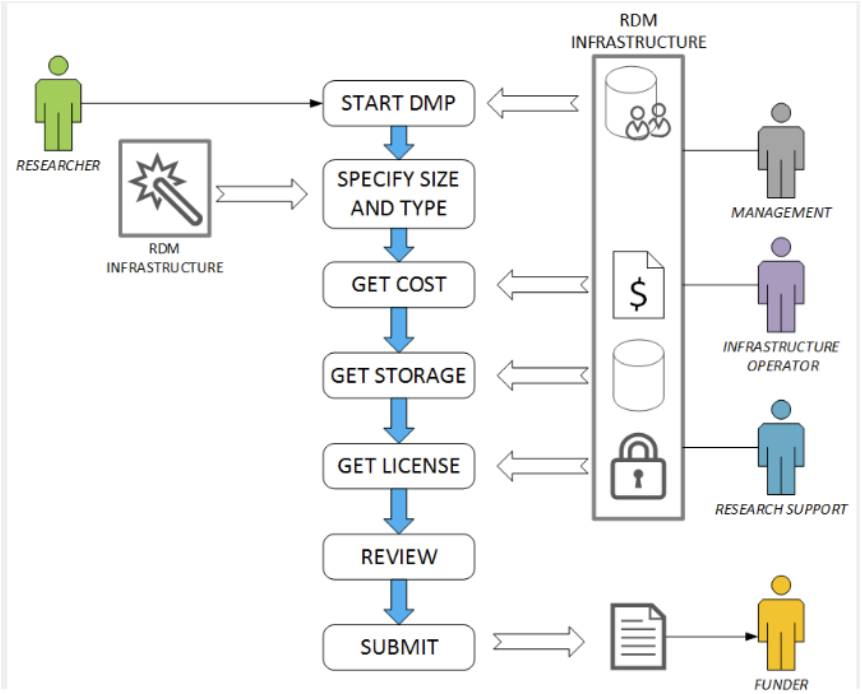
## 07 - Machine Actionable DMPs (maDMPs)

### Why traditional DMPs are not enough

**Normal DMPs** are **bureaucracies** and are usually **created too late and vague**. They also largely depend on the **human factor.** Humans can be scrupulous and need to be aware that something needs to be done.

### What is required to make DMPs machine-actionable?

→ Certain **stakeholders** in projects **need different information** at **different stages** → Many **problems can be avoided** if the **timing is right** and **communication is ensured!**




→ **Automated Data Management Workflow:** Requires a **common data model** to exchange information, a well-defined RDM workflow of **WHO, WHAT, WHEN, HOW,** and a **data management infrastructure.**

### Scoping maDMPs procedure (RDA)

The Research Data Alliance (RDA) defines scoping maDMPs by DMP Common Standards WG as the following steps:

1. **1st consultation**: The goal is to **identify the stakeholders at each lifecycle stage**
   ae. As <stakeholder>, I want <goal> so that <reason>
   ”As a researcher, I want to inform repository operator on the amount of data in the planning phase, so that they provide information on costs.”

2. **2nd consultation**: **Goes deep**. How are **specific requirements** modeled? Which **specific fields** are needed? **Which models exist** already?

3. **Proof of concept tools**: Provide **minimum input**. **Import from existing systems** as much as possible to **help in creating maDMPs**.

4. **BMPN processes**: **Identify systems** and **stakeholders** involved. → **What can be automated**?
   Define **processes** that **help** identify:
   a. **tasks to be performed by stakeholders** (ae. operators provide cost of storage)
   b. **systems that need to be put in place** (ae. maDMP repository)
   c. **concepts to be developed** or agreed on (ae. Cost model for the storage).

5. **Model development:** Develop maDMP model for use case and stakeholder requirements.

### How common standard for maDMPs works (RDA)

Builds a **common model for a DMP** in a **machine actionable format**, like **JSON**. Contains all things a DMP should contain in a common data mode standard.

**Assumptions of the standard**:

- **Relaxes constraints:** Model must be applicable in different settings.
  Restraints are relaxed inside of a model → DMP **can** relate to a project [0,*]
  Constrains introduced at the business level (eg. tool implementing the model) → DMP **must** relate to the project.
  → **BOTH** of these **DMP instances have to be compatible**!

- **Interoperability:** Model used to exchange information predominantly. **No 'meta-fields'** about the DMP, internal representations of information in a DMP may differ between tools.
  → **Should still work**!

- **Versioning:** The model **provides fields for the DMP version**. It d**oes not track connections between versions → Done by systems** using the model!

- **Evolving Information**: The model **expresses the certainty of provided information** (ae. embargo period for code or data).

### What are the related standards and how application profiles work

There are **standards that can be reused** to make building a standard for maDMPs easier. **Adding on top of one standard** and **introducing tighter constraints without breaking compliance** is called the **Application Profile**. For example, **maDMPs build on top of DCAT, which builds on DublinCore**.

### 10 principles for maDMPs

1. **Integrate DMPs** with the **workflows of all stakeholders** of the ecosystem.
2. Allow **automated systems to behave on behalf of stakeholders**.
3. Make **policies for machines and people**.
4. **Describe the components of the data management ecosystem** for **both humans and machines.**
5. **Use PIDs (DOI, ORCID**) and **controlled vocabularies (DublinCore, DCAT, …)**
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 living, updatebale documents**.
10. Make **DMPs publicly available**.

# 08 - Sensitive Data Repostories

## Sensitive data and data visiting

**Sensitive Data** = “ any information that is **protected against unwarranted disclosure** for **legal** reasons, **ethical** reasons, **personal privacy** or **proprietary considerations**”

**Digital repositories struggle to make sensitive data available** in their collection:

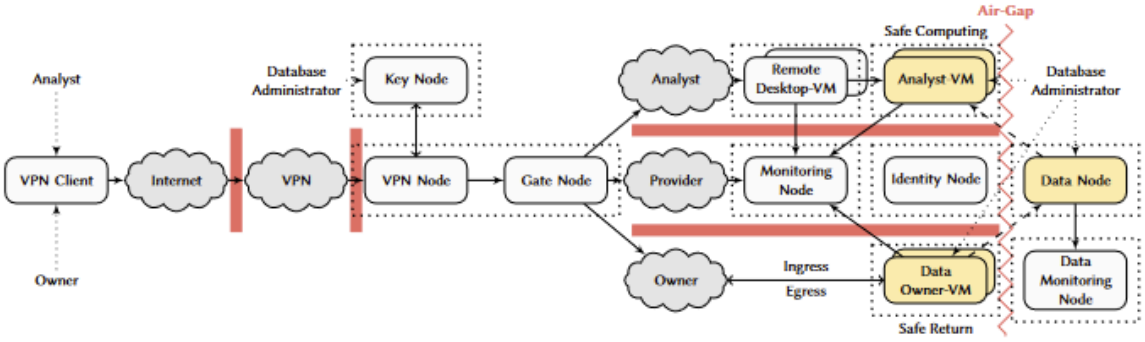
- **Control**: Maintain control over sensitive data
- **Open science**: Conflicts with allowing access to third parties when sharing data
- **Data sharing**: Control is gone once shared usually

Therefore, the concept of **data visiting** is introduced:

- **Keep complete control** over the data, control **when**, which **subset** and **by whom** data can be accessed.
- Allow **consumers** (human and machine) to **come to the data!**
- **Closely monitor data and accesses** + Prevention of breaches

Data visiting is **implemented via an Air-Gap Isolation system** (see below).

## Three aspects of sensitive data repositories




1. <u>Technical</u>**: Safeguarding data** against **technical threats** and **design flaws**, providing a
   **highly-controlled virtual research environment:**
   a. **Secure Enclaves** provide **CPU isolation** and **memory encryption**
   b. **AirGap System: Imports** only **over tools**, **exports** of **summary data only** → **5 safe dimensions:** projects, people, data, settings, outputs
   c. **Data Versioning** (below)
   d. Accessing sensitive data **machine-readable through APIs**: HTTP, JDBC, AMQP

2. <u>Organisational</u>**:**
   **a. Ingressing data:** Only over **GUI**, providing **data and metadata** (ae. for units of measurement very important)**.**
   Deposit **into data-owner VM** → **DB Admin copies** over and **restores Air-Gap.**

   b. **Requesting Access:** Data is **FAIR, but not open!**
      Access via proposal → Extracted **subset** gets **placed into analyst VM**.
      → **Persisted subsets** identified via **query store!**

3. **<u>Legal</u>:**
   a. **Sign legal agreements** for Analysts: **NDAs**, Agreement to **monitoring**, Download prohib.
   b. **Collect information** that is **usable in case of unauthorized leaks**: Personally identifiable information, **evidence of misuse** against agreements, **evidence via subset fingerprints**

### Data versioning

**Snapshots pollute the database** because they **need full copies**. If we only know the query, it is **difficult** to **get the right snapshot** again. We already had this problematic also with the data citation part. The **solution** is therefore to employ **fingerprinting and a query store**!

### Stored and persisted queries

The key difference is, that **stored queries only are saved in a database themselves**. A **persisted query** however additionally **saves** only the **metadata in** a **metadata store separately**! This makes the query FAIR, eg. by **making the metadata findable even after the data is not available** anymore..

### FAIR sensitive data

Data is **not open**, but **still FAIR**:

- **Findable**: Assign PID to each query, add metadata, link PID of data along with metadata, metadata searchable
- **Accessible**: Open and authentication via HTTPS, AMQP or JDBC, Metadata always available in metadatastore
- **Interoperable**: Metadata interlinked in metadata store, OWL/RDF concepts for units of measurement
- **Reusability:** Open source lincences, description of data

# 09 - Repositories External Visivbility

### How to make repository contents visible?

Only **having repository is not enough** → Contents **must be discoverable and FAIR!**
In order to make repository contents align with “Findable”, a solution is to **register the DOIs and metadata** of all data managed in a **DOI registration body like DataCite or CrossRef.**

### What options there are and how to choose the best one for you setting?

This can be achieved by using **Interoperability protocols.** These protocols allow metadata (and data) to be deposited in a data catalogue. There are two (three if sholix counts) choices

for protocol:

- **OAI-PMH** (Open Archives Initiave Protocol for Metadata Harvesting):
  Data remains in repo, **harvester aggregates metadata**. Uses **Dublin Core**. **Not for depositing**, **only metadata.** → **In the EU**, one can **use OpenAIRE** if an **OAI-PMH endpoint is available** that does the harvesting.

- **SWORD** (Simple web-service Offering repository deposit): **Deposit data to multiple repositories** at once, also deposit by third part systems like lab equipment.

- **Scholix**: **Link** between multiple **registration bodies and repositories, z.B. DataCite - OpenAIRE**

### how to describe data using discussed standards (DCAT, DataCite)

- **DCAT**: W3C vocabulary is an **RDF vocabulary** for **interoperability** between **data catalogs.**
  **Facilitates** federated **dataset search**. Has **relaxed constraints** (most fields optional). Can be embedded in SPARQL, HTML, RDF/XML, Turtle, … → Mostly **used by governmental repos**.

- **Schema.org**: Metadata embedding used by google, yahoo etc. (zb google search cards with infos):
- **DataCite**: uses JSON-LD

### How do repositories support FAIRness?

They mainly focus on the "**Findable**" principle of making sure "**Metadata is registered or indexed in a searchable resource**". The first step is to **mint a DOI** with a **registration body** (**DataCite, CrossRef**).

Repositories can then achieve being findable by implementing **interoperability protocols** like OAI-PMH to **register their metadata** with a **metadata registry** like **OpenAIRE** or **OpenDOAR.** The **registries use harvesters** to access the OAI endpoints of the repositories and **aggregate the metadata**.

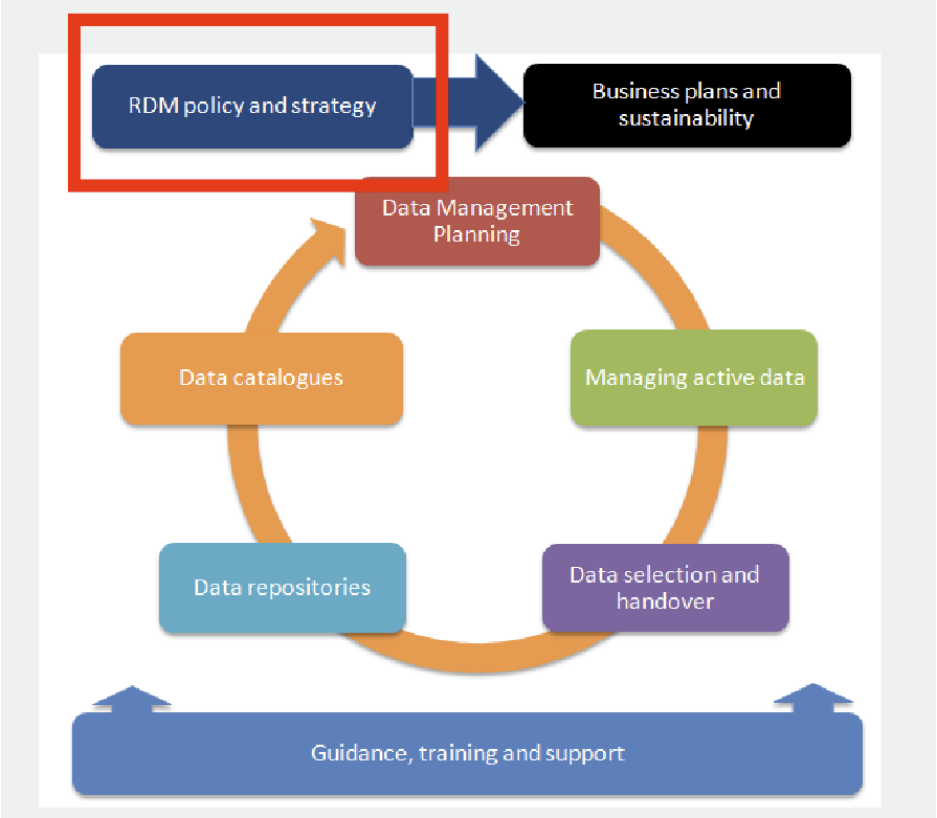
# 10 - RDM and Certification

### What are the lifecycle models and how to use them?

Research lifecycle models help to:
- Understand the **roles of stakeholders**
- Tailoring **services**, identifying **responsibilities**, defining **infrastructure**
- **NOT** used **by researchers** themselves

There are three major models, the **UK Data Archive**, **Digital Curation Centre** and the **University of Central Florida** model.

## What are the components of an RDM service infrastructure?




- **RDM policies and strategy**: Below
- **Business plans and sustainability**: Develop a **business plan** for 3,5,10 years. Consider **costs**, **sustainability,** and **long-term costs**. Also considers cost recoupment.
- **Guidance, training, and support**: **Single point of information** (Helpdesk, Website), **Trainings** and **consultation** hours → Helpdesk staff
- **Data Management Planning**: **Create a DMP**, at best machine-actionable (maDMP). Make a **policy** to **make DMP use obligatory**, and provide **support** and **awareness training**.
- **Managing active data**: **Review current infrastructure**. Make **case for investments** where needed. **Develop procedures** for **allocation** and **management** of data storage. **Provide** a flexible **system** for the **creation**, **management,** and **sharing** of data.
- **Data selection and handover**: **What** data **to keep**. **Integrate data management** into **existing infrastructure** → Develop deposit tools. **Support** researchers and **offer input** for decisions.

- **Data repositories**: Automate **repository recommendation**, **Reduce effort**, **Lower expertise** needed.

- **Data catalogues**: Define **metadata** to be **recorded in data catalog** → Lecture on **external visibility**.

### What is the scope of policies and how they drive DM activities and obligations?

Start with Taboos → Go over to related principles → Create policy → Create rules, legislations, regulations
A **policy** is a **course of action** that **offer a frame for the generalisation of rules** and is not in competition with other policies. As example, “Organisation X will preserve its research data using FAIR principles”.

**Helps** to **establish core RDM principles** for the organization under **local context** and based on **external drivers**. Closely related to strategy. The **strategy defines the course of action**, the **policy implements the framework** for the strategy to be established. **Requires broad consensus** and **support** from organization members.

### How do develop support services and who are data stewards?

Data Stewards **provide guidance**, **training,** and **support** in the RDM service infrastructure. They **analyze** the **data management needs** and **provide advice to researchers**. They **help the faculties comply** with the **policies** set in place (**roles, responsibilities**). They act as a **trusted point of contact** for **data management questions** and **train** and **inspire**.

### What to consider when implementing DMPs in an institution?

**Not only about technical solutions**. **Integration** into **existing infrastructure** is crucial. Develop a **vision** and **plan** (strategy). **Include** all of the **stakeholders** and roles. Make **incremental developments** and always offer **guidance**, **training** and **support** to faculty and researchers.

### What standards exist and why certification matters?

**Certification builds trust** that the data remains **useful** and **meaningful** in the **future**. It **improves** the **communication** within the repository (**roles, responsibilities**). Ensures **public transparency**. **Attracts funding** and **enables comparison** to other repositories. There are three main standards: **ISO 16363**, **DIN 31644,** and **CoreTrustSeal**.

### What are the certification criteria?

Depends on the certificate, based on self-assesment that is the audited by external reviewers multiple times.