EOSC-Life

FAIR hackathon 2021

The data life-cycle and providing data provenance

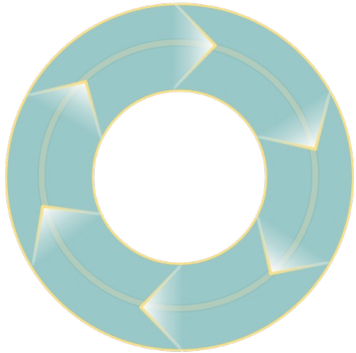
Katrina Exter and Rudolf Wittner (WP6)

Poll questions

- Have you ever written or sketched out the data life-cycle in your experimental/data-based work? *Y/N*
- Do you think you understand what “data provenance” means, especially when it comes to sharing and publishing your data? *Y/N*
- Do you know why provenance is important to FAIR data? *Y/N*
- In your experience, do you believe that providing provenance is done well enough in your field? *Y/N*
- Have you ever provided provenance information/data/files along with your published data (outside of the metadata you were required to add when you added your data to the online catalogue, archive, or portal where you published it)?

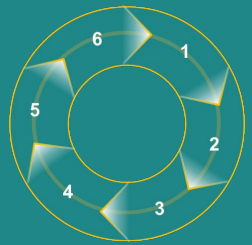
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 824087

What is a data life-cycle?



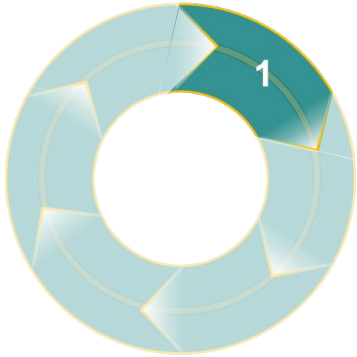
Data Life-Cycle:

- *covers the entire period of time over which data exists*
- *encompasses all the stages: **first Capture** → **data re-use***
  - → ***Experimental stage (physical material)***
  - → ***Creating and manipulating digital data***
  - → ***Publishing data and data re-use***



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

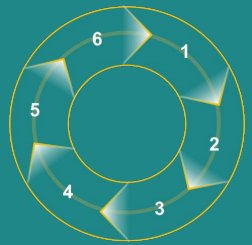
What is a data life-cycle?



Data Life-Cycle:

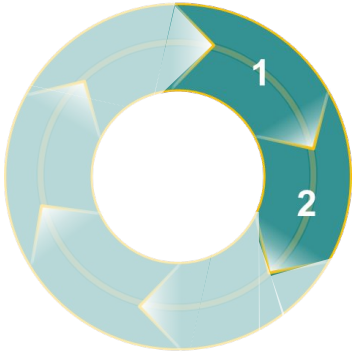
- *covers the entire period of time over which data exists*
- *encompasses all the stages: **first Capture** → **data re-use***

1. **Sample acquisition** -> raw data:
   sample preparation, experimental settings / parameters, raw data acquisition



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

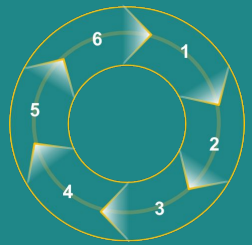
What is a data life-cycle?



Data Life-Cycle:

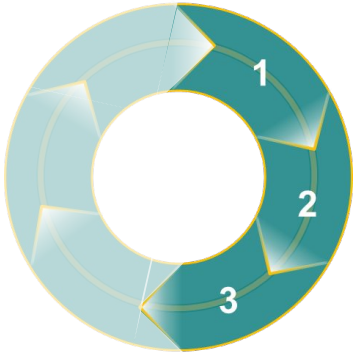
- *covers the entire period of time over which data exists*
- *encompasses all the stages: **first Capture** → **data re-use***

1. **Sample acquisition** -> raw data:
   sample preparation, experimental settings / parameters, raw data acquisition
2. Data **quality control**: checking and updating of collected data
   documentation of QC procedures



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

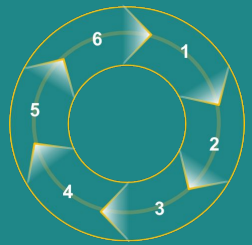
What is a data life-cycle?



Data Life-Cycle:

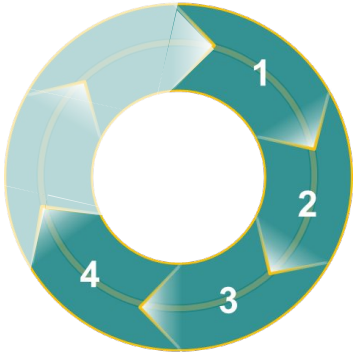
- *covers the entire period of time over which data exists*
- *encompasses all the stages: **first Capture** → **data re-use***

1. **Sample acquisition** -> raw data:
   sample preparation, experimental settings / parameters, raw data acquisition
2. Data **quality control**: checking and updating of collected data
   documentation of QC procedures
3. Data **processing** & **analysis**: guided by scientific question
   documentation of processing steps, analysis methodology



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

What is a data life-cycle?

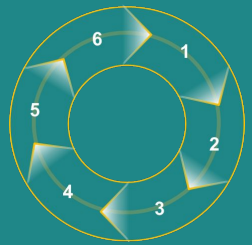


Data Life-Cycle:

- covers the entire period of time over which data exists

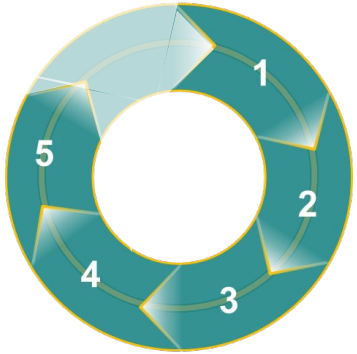
*- encompasses all the stages: **first Capture** → **data re-use***

1. **Sample acquisition** -> raw data:
   sample preparation, experimental settings / parameters, raw data acquisition
2. Data **quality control**: checking and updating of collected data
   documentation of QC procedures
3. Data **processing** & **analysis**: guided by scientific question
   documentation of processing steps, analysis methodology
4. **Archiving** & **publication**: data are placed in an online catalogue
   discovery metadata, provenance metadata, provenance files, references and links



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

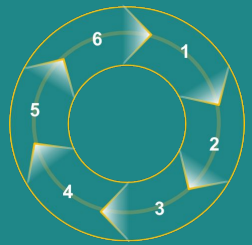
What is a data life-cycle?



Data Life-Cycle:

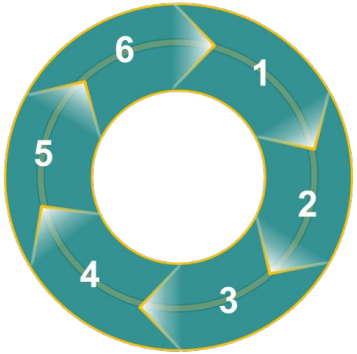
- *covers the entire period of time over which data exists*
- *encompasses all the stages: **first Capture** → **data re-use***

1. **Sample acquisition** -> raw data:
   sample preparation, experimental settings / parameters, raw data acquisition
2. Data **quality control**: checking and updating of collected data
   documentation of QC procedures
3. Data **processing** & **analysis**: guided by scientific question
   documentation of processing steps, analysis methodology
4. **Archiving** & **publication**: data are placed in an online catalogue
   discovery metadata, provenance metadata, provenance files, references and links
5. Data **dissemination** / **integration**: adding data to well known portals, brokers



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

What is a data life-cycle?

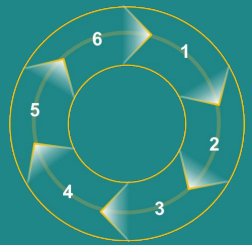


Data Life-Cycle:

- covers the entire period of time over which data exists

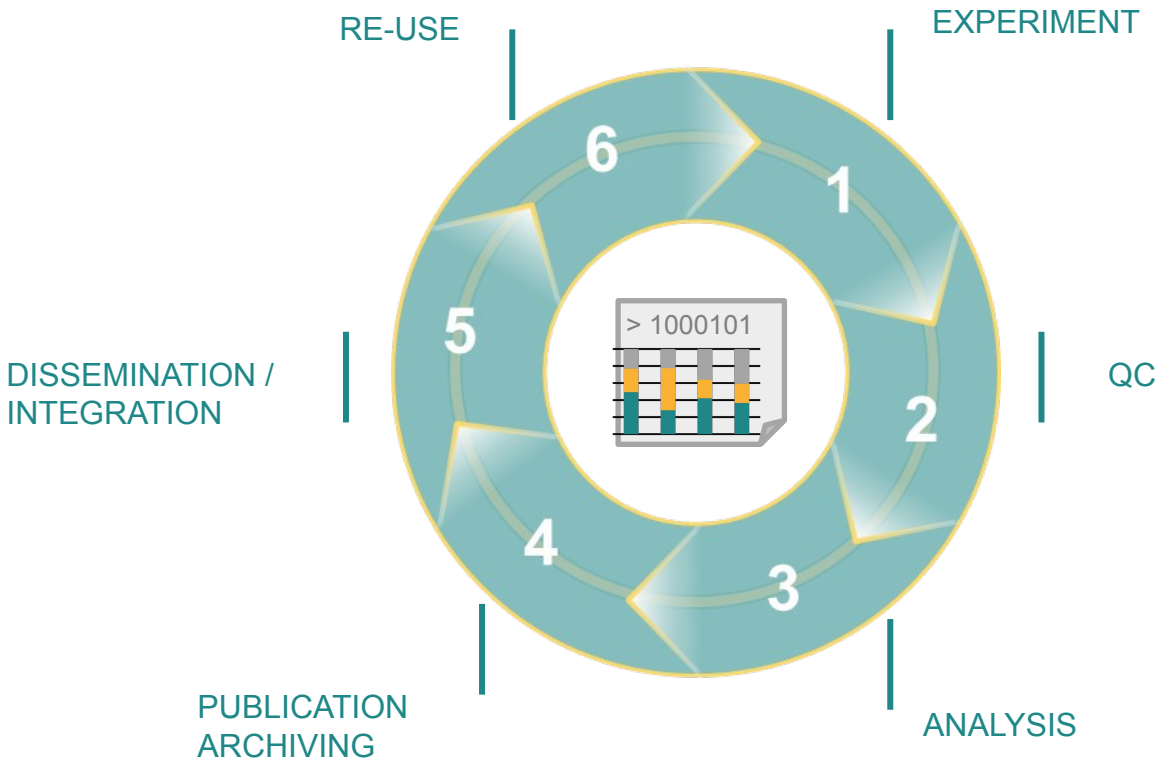
*- encompasses all the stages: **first Capture** → **data re-use***

1. **Sample acquisition** -> raw data:
   sample preparation, experimental settings / parameters, raw data acquisition
2. Data **quality control**: checking and updating of collected data
   documentation of QC procedures
3. Data **processing** & **analysis**: guided by scientific question
   documentation of processing steps, analysis methodology
4. **Archiving** & **publication**: data are placed in an online catalogue
   discovery metadata, provenance metadata, provenance files, references and links
5. Data **dissemination** / **integration**: adding data to well known portals, brokers
6. Data **reuse**: only possible with sufficient provenance information!



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

DATA LIFE-CYCLE

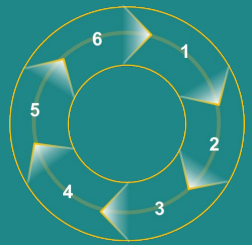


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

What is provenance?

Dictionary definition:

*"A **record of** the **ownership** of a **work of art** or antique, used as a **guide** to authenticity and quality "*

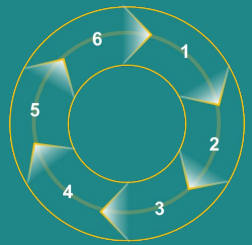


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

What is provenance?

Our definition:

*"A record of the **creation** and **manipulation** of **data** (from cradle to grave), (used as a **guide** to authenticity and quality")*



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

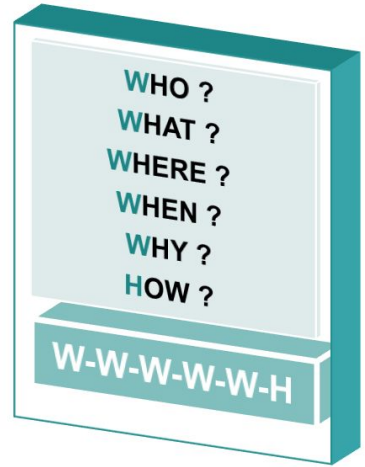
What is provenance?

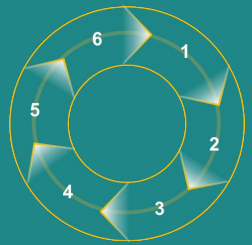
Our definition:

*"A record of the **creation** and **manipulation** of **data** (from cradle to grave), (used as a **guide** to authenticity and quality")*

Provenance is about providing

- **Traceability**:
  precursors, inputs from which your data/results were derived
- **Reusability**:
  others can re-use / integrate your data
- **Reproducibility**:
  others can repeat your results
- **Trustability**:
  others can trust your results



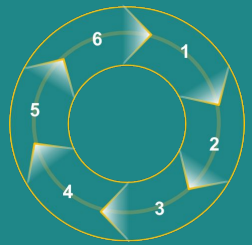


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

What provenance information should be provided for each step of the data life-cycle?

*Examples for: sample acquisition, and processing -> **Raw data***

- Dealing with **physical material**
  - temperature, storage conditions
  - exposure to forces, radiation
  - sample container(s)
  - additives, preservatives
  - changes in conditions
- **Procedures** and **protocols** followed (Links to files)
- **Instruments** and/or **software** used:
  - instrument calibration, operational qualification, maintenance info
  - software versions, references (e.g. DOIs), URLs (e.g. Github)
  - software logfiles, parameter input files

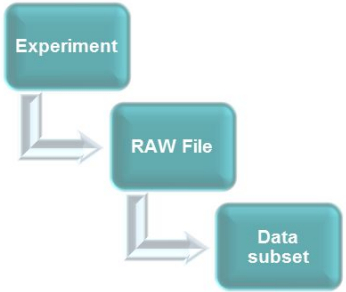


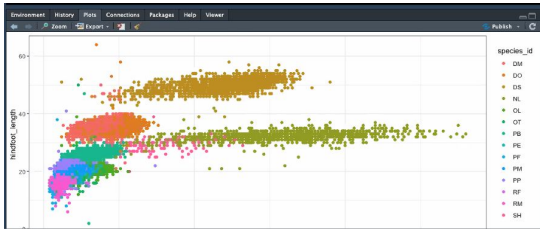
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

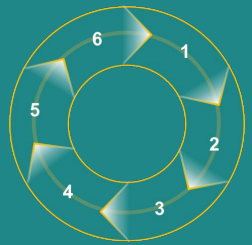
What provenance information should be provided for each step of the data life-cycle?

*Examples for: sample acquisition, and processing -> **Raw data***

- **Experimental** /**sampling**/other permit **information** to be made public (either as documents or their references)
- For **collated Data**
  - References to the original dataset locations
  - Details of which parts of the data in particular were used





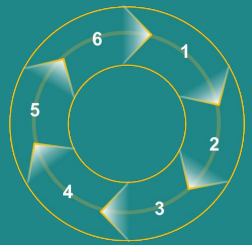


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

What provenance information should be provided for each step of the data life-cycle?

Examples for: ***Data quality control***

- **Documenting** the QC process
  - Parameter file / procedures & protocols - documents
- **QC software**:
  - Software name
  - Software version (s)
  - Software input settings (for acquiring data)
- When the data are made public, provide **links** between **Raw data** and **QC data**

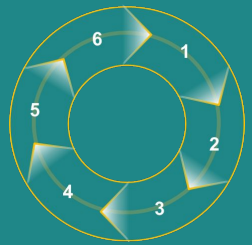


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

What provenance information should be provided for each step of the data life-cycle?

*Examples for: **Data analysis** and scientific **results***

- **Procedures** and **protocols** followed (actual files or links thereto)
- **Software** / **Workflows** / **Pipelines**
  - its names and references (e.g. DOI or github URL)
  - the versions used
  - input settings (parameter files, algorithms)
  - even the whole image of the computational environment (e.g. docker)
- When the data are made public, provide **links** between
  Raw data --> QC data --> Results data

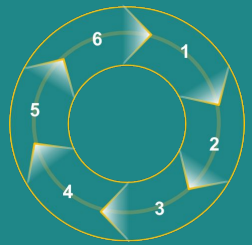


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

What provenance information should be provided for each step of the data life-cycle?

*Examples for: **Archiving** and **Publication***

- At a minimum, publish the **raw data** - first step of data life-cycle, and its provenance information
- If publishing also **processed data**, catalogue them **separately** (as the metadata descriptions will be different)
- Link **published data** **<-->** **published science** to each other
- As much **metadata** as the archive/catalogue allows, (supplement with additional files where that is not sufficient)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

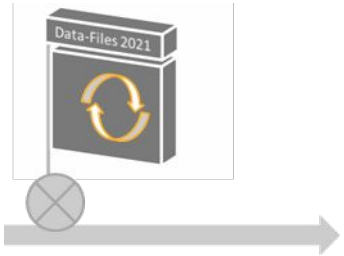
What provenance information should be provided for each step of the data life-cycle?

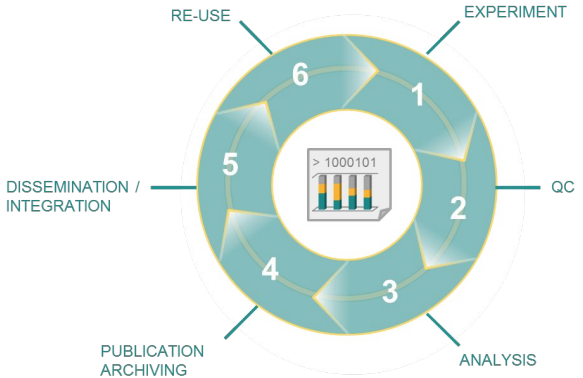
Data Dissemination

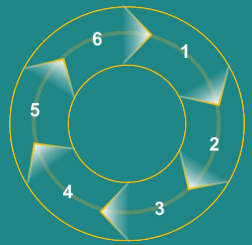
- If you have provided **sufficient provenance** (meta)data so far, there is nothing else to do here

Data Reuse

- Is only possible if the provenance for **all stages** of the **data life-cycle** is provided!
- Any data re-used, starts the life-cycle again!

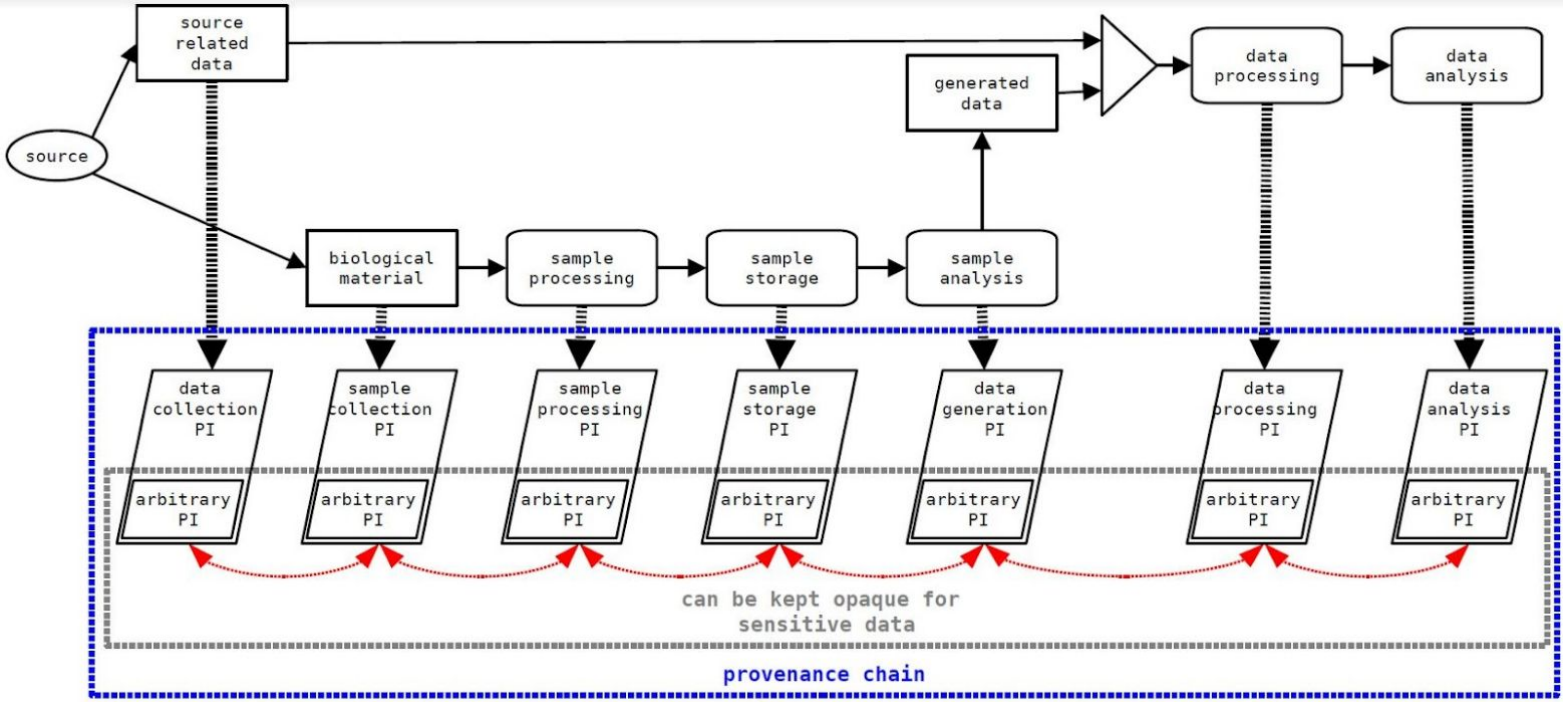


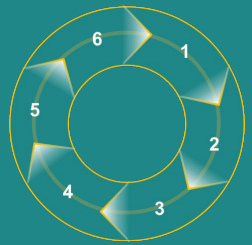




This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Distributed provenance information

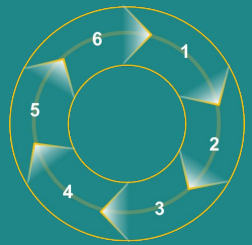




This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Desired properties of distributed provenance

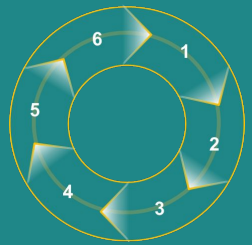
- **Interoperability & compatibility**
  - to enable integration of provenance coming from different sources
  - approaches to address further requirements should be compatible
- **Access control** & **support** for opaque provenance components
  - to hide sensitive information
  - to to ensure privacy
- **Integrity & non-repudiation**
  - to support trustworthiness and reliability
- **Versioning** mechanism
  - to prevent from disrupting integrity of distributed provenance
- **Identification** of **objects** in provenance
  - to prevent from collisions of identifiers, and to achieve their persistency



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Provenance representation
to support interoperability of provenance

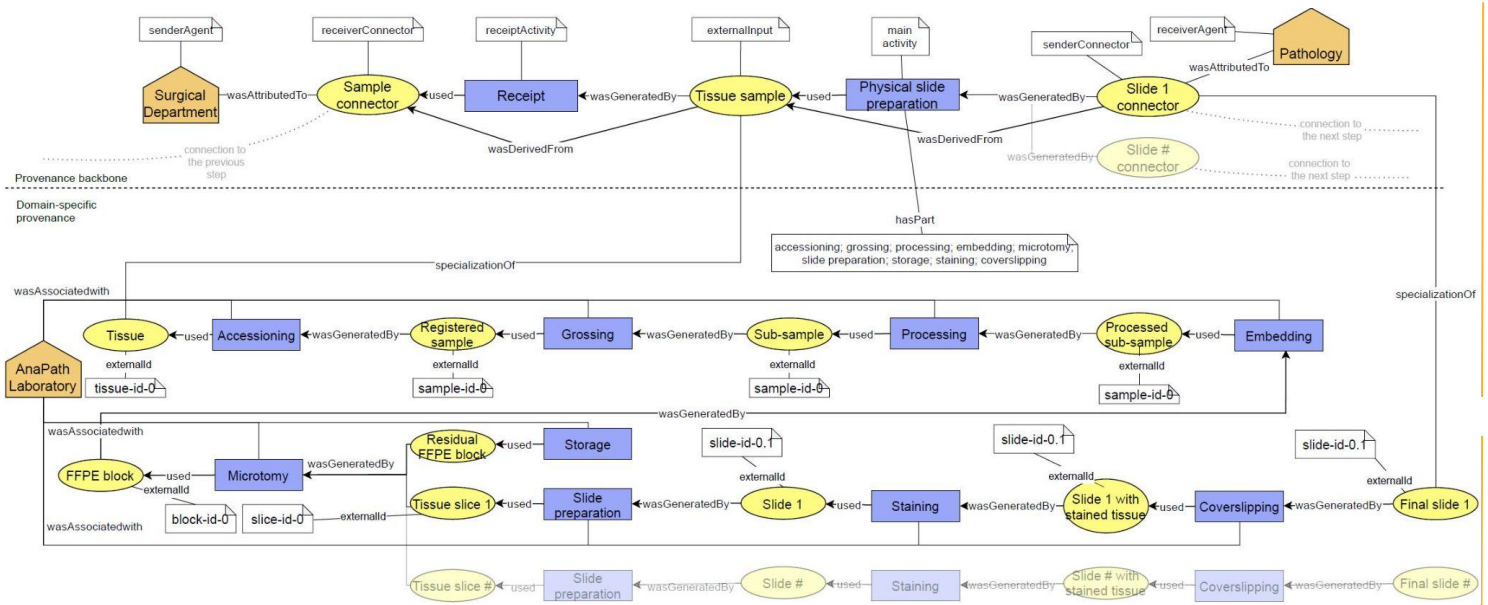
- **W3C PROV** - current community standard for provenance
  - to enable inter-operable interchange of provenance information in heterogeneous environments
- **PROV-DM** - core of the standard
  - conceptual data model for provenance
  - interaction of three structures: **entity**, **activity**, **agent**
  - predefined set of their mutual relations
- **Consists of several documents**
  - ontology PROV-O
  - normalization, equivalence & validation rules
  - serializations
  - …



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Provenance representation

example – pre-analytical process inside a laboratory

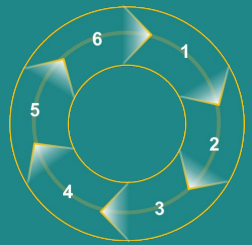


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Provenance representation

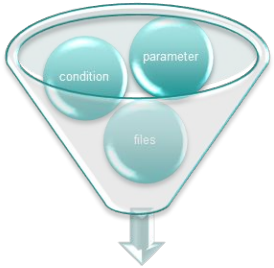
to support interoperability of provenance

- Common Provenance Model (CPM)
  - an extension of the PROV to capture the **common semantic**
  - being developed in EOSC Life WP6
- **Examples**
  - links between distributed provenance parts
  - derivation paths between inputs and outputs
  - hashes & digital signatures
  - versioning
  - responsibilities
  - timestamps, location, measurement units, …
  - ……..



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Provenance representation
to support interoperability of provenance



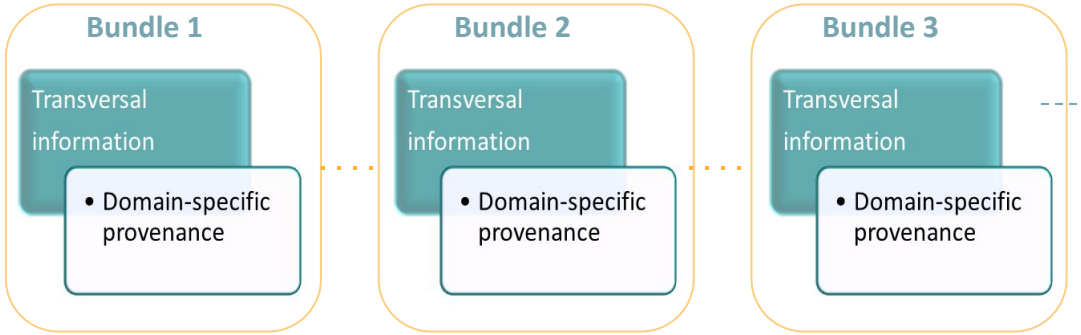
collecting provenance

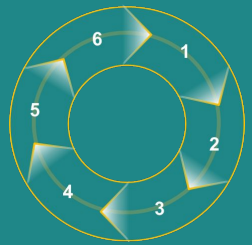
translating into PROV/CPM

In order to use it:

→ expression of **domain specific semantic** in terms of the data model

- General procedure
  - → purpose of provenance collection
  - → **collect provenance** information in an arbitrary format
  - → at some point - **a finalisation event**
    translate existing provenance into PROV/CPM
- Content of finalized provenance:





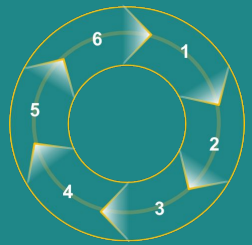
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Access control and support for opaque provenance components

- Provenance information can contain sensitive/non-public information
  - subject to authorization
  - e.g. person identifiers (responsible people, patients, donors, ...)
- Provenance information with opaque provenance components may be still useful for a non-authorised subject
- The main challenges
  - meaningfulness after hiding opaque parts
  - coarse vs. fine granularity of access control mechanisms

So what can I do?

- Do I need to store sensitive information in provenance?
- Who is authorized to see the sensitive information?
- Can I create an ordering on information categories?



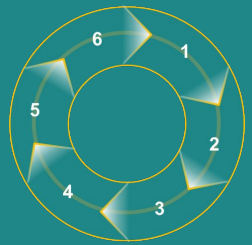
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Integrity, non-repudiation

- Integrity
  - ❑ provenance information not changed in an unauthorized way
- Non-repudiation of origin
  - ❑ creator of provenance can not falsely deny its creation
- Achieved by application of hashes, digital signatures and policies
- Involvement of a third party – a notary
- Required evidence for provenance non-repudiation generated during the finalisation event

Versioning

- Keeping history of versions of the finalized provenance bundles
- Related to distributed provenance chain resolution



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

ISO 23494 — Biotechnology — Provenance information model for biological material and data

Sample Acquisition, Processing, Transport, and Storage Provenance
→ Part 3

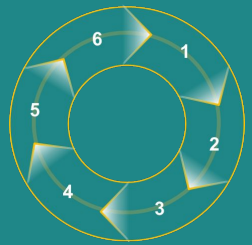
Data Generation Provenance (NGS, OME, …)
→ Part 4

Data Storage and Processing Provenance (CWL, …)
→ Part 5

Provenance Information Management Requirements
→ Part 1

Common Provenance Model
→ Part 2

Security Extensions
→ Part 6



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Self-assessment questions

Before starting and during each step of your work

1. Can I describe the stages of my data life-cycle e.g. in a bulleted list?
2. Do I know what software, protocols, instruments I will be using at each stage of that life-cycle,
3. …can I make a list of what provenance information I need to record for each of these softwares, protocols, and instruments?
4. Can I provide that information to external users, or are they only available on a piece of paper on my desk?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Self-assessment questions

<u>When you are ready to publish your data</u>

1. Have I gathered the provenance information at a fine enough granular level?
   a. → have I decided what granular level I need?
   b. → could someone recreate my results from my raw data using this provenance information?
2. Have I gathered that information together with the data they describe?
   a. → am I ready to submit this all to my data catalogue/archive?
3. How much provenance information can I provide as metadata in the data catalogue where I will put my data, or …
4. …do I need to provide provenance information as files and documents to be included with the data in the catalogue? Do I know how I will do that?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087