



AGATA Collaboration Council meeting, 10-12/11/2021 Legnaro

Open science, open data Data Management Plan for AGATA Phase 2

O.Stézowski

On behalf of the Data Processing Group

Work from dedicated DMP meetings March 2021 → June 2021

« OPEN » path / future of the AGATA Data



Conclusions

Impacts for the collaboration

What has been done so far ... phase1

What should, could or have to change for the Phase2

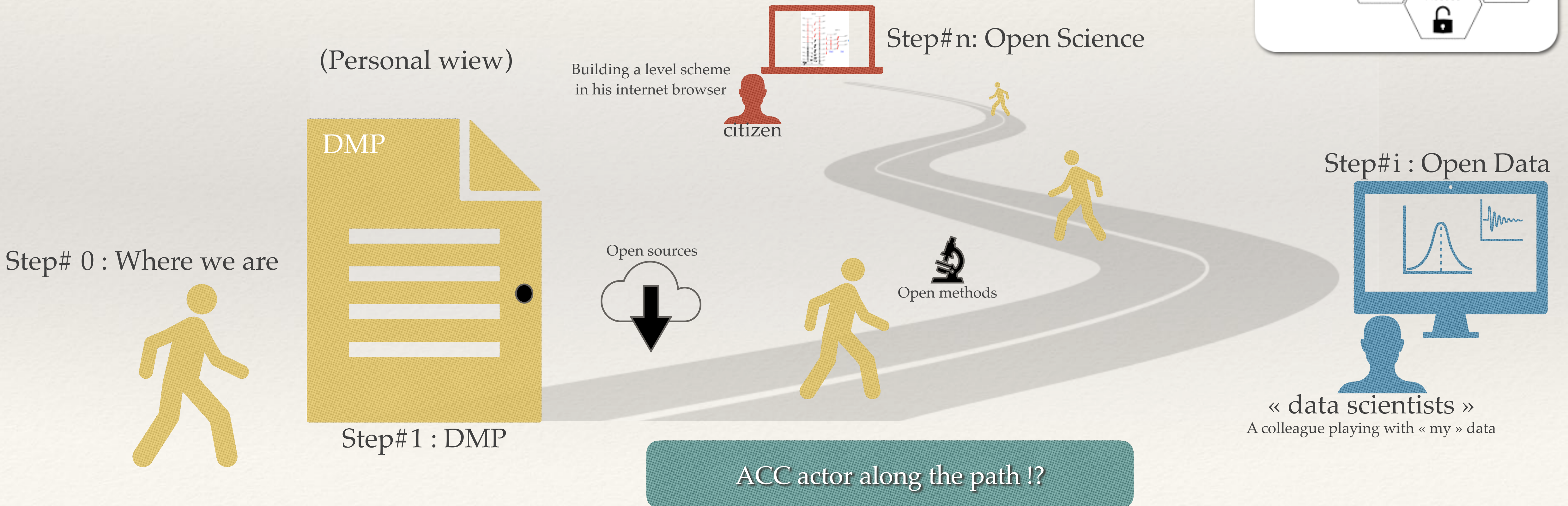
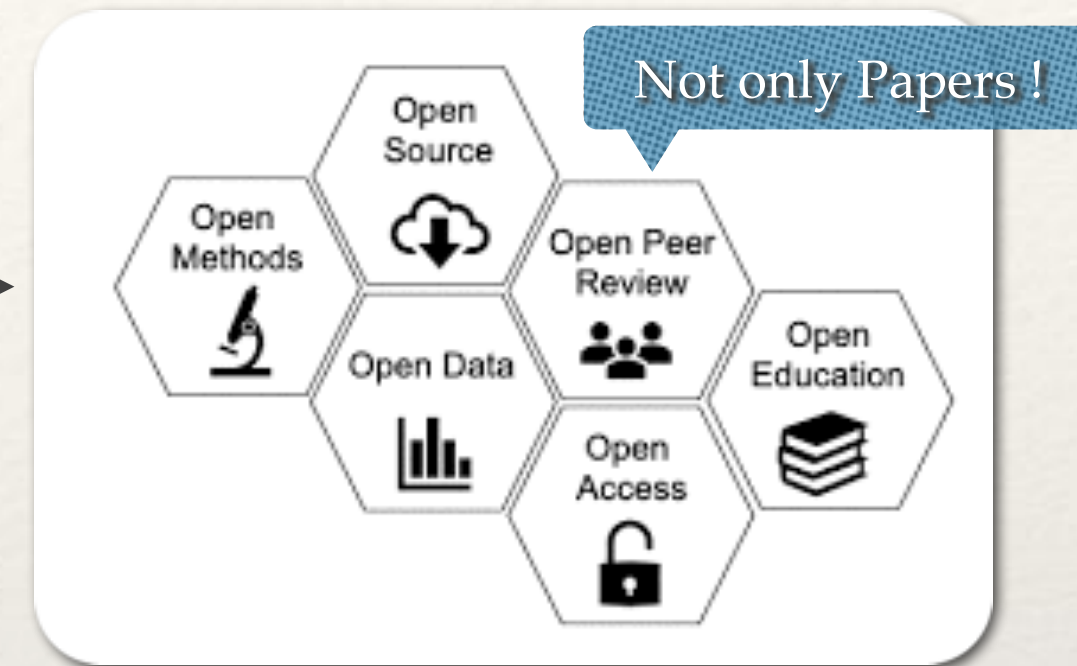
Contexts

- External context
- Internal context

External Context : what is behind the 'DMP' door ?

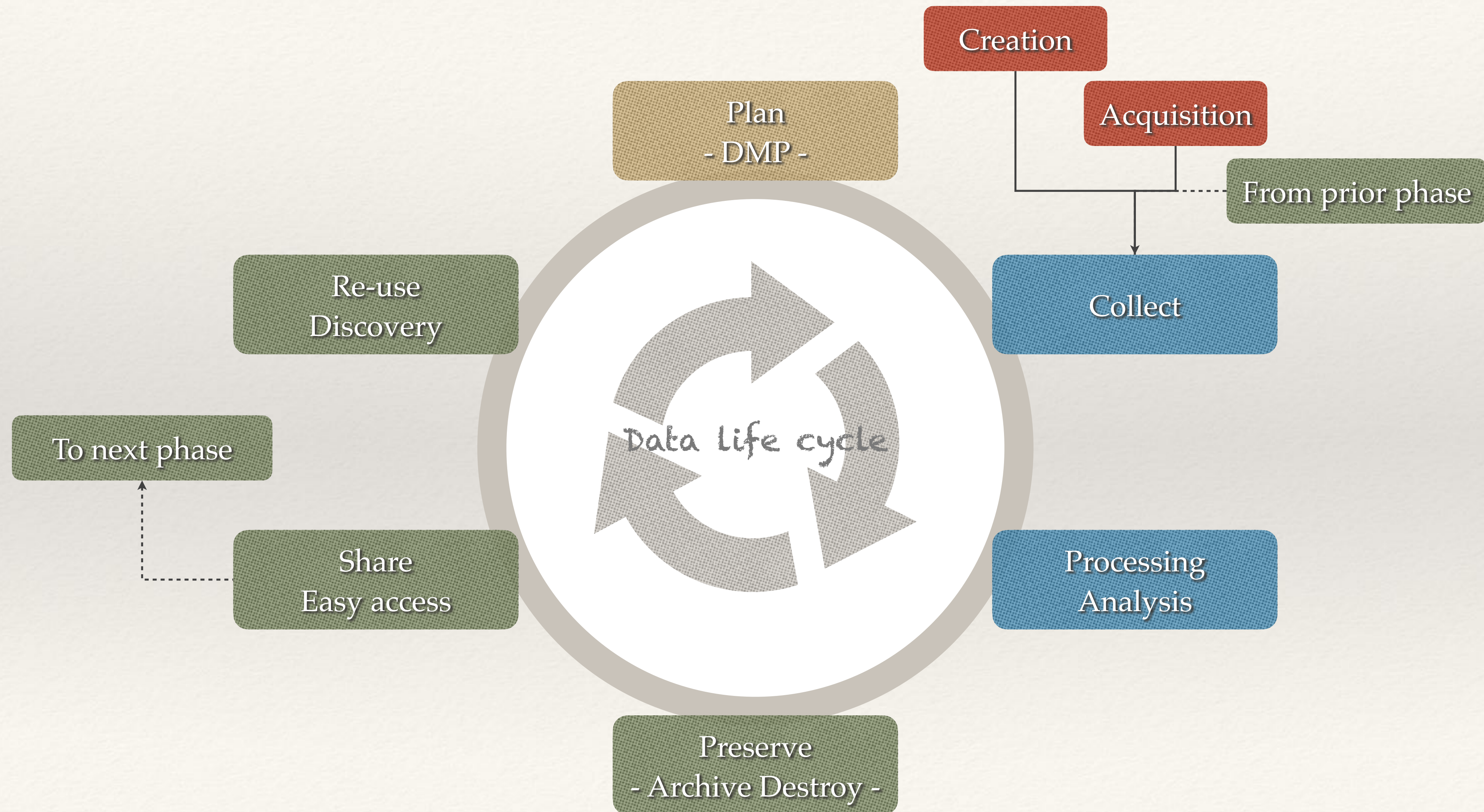
Our researches are almost fully founded by **public** agencies !

All the products from such funds have to be given back to the society →



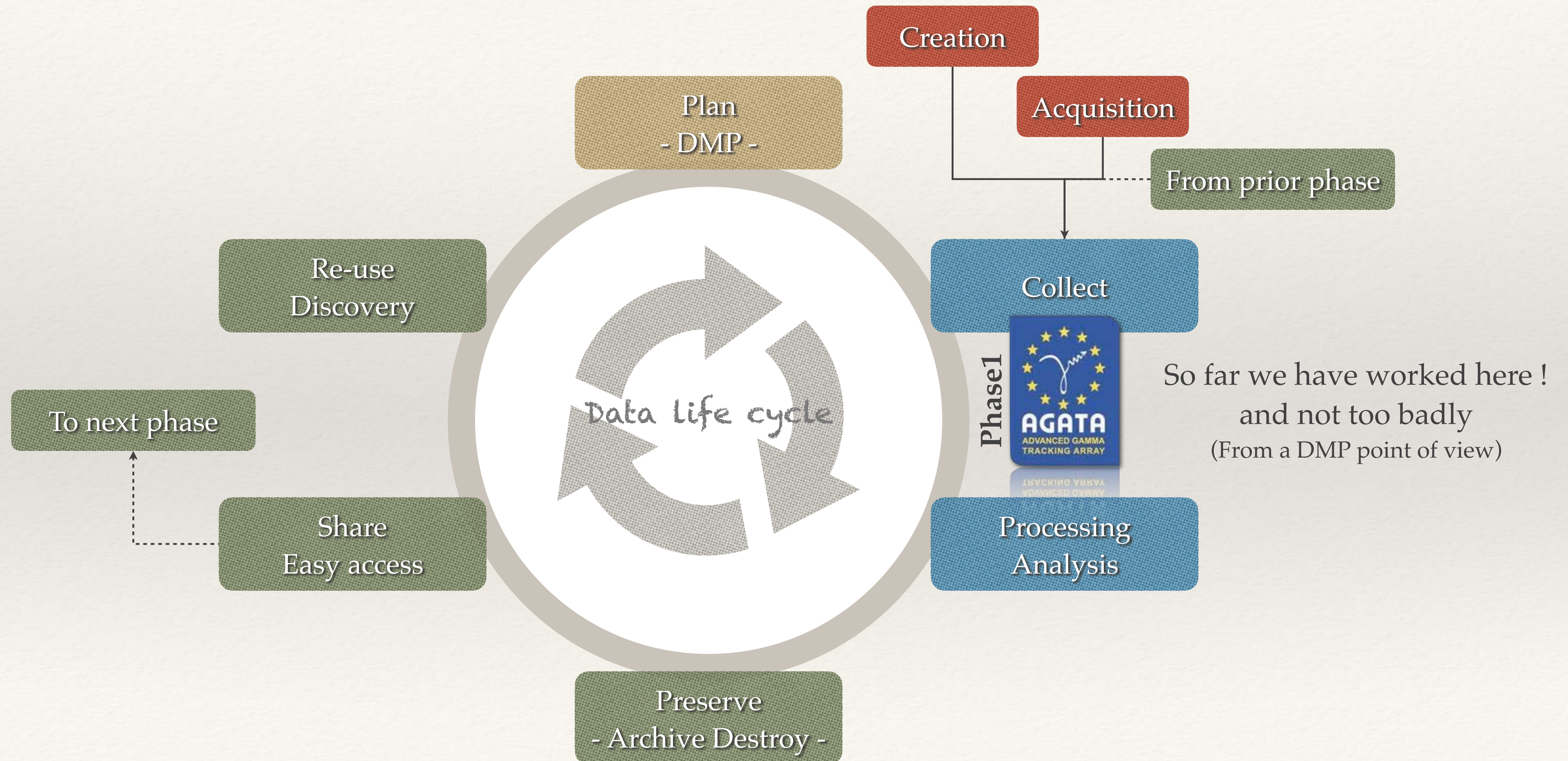
External context : what is a DMP ?

It means we should establish a plan to deal with the cycle of life of the data produced by AGATA ...



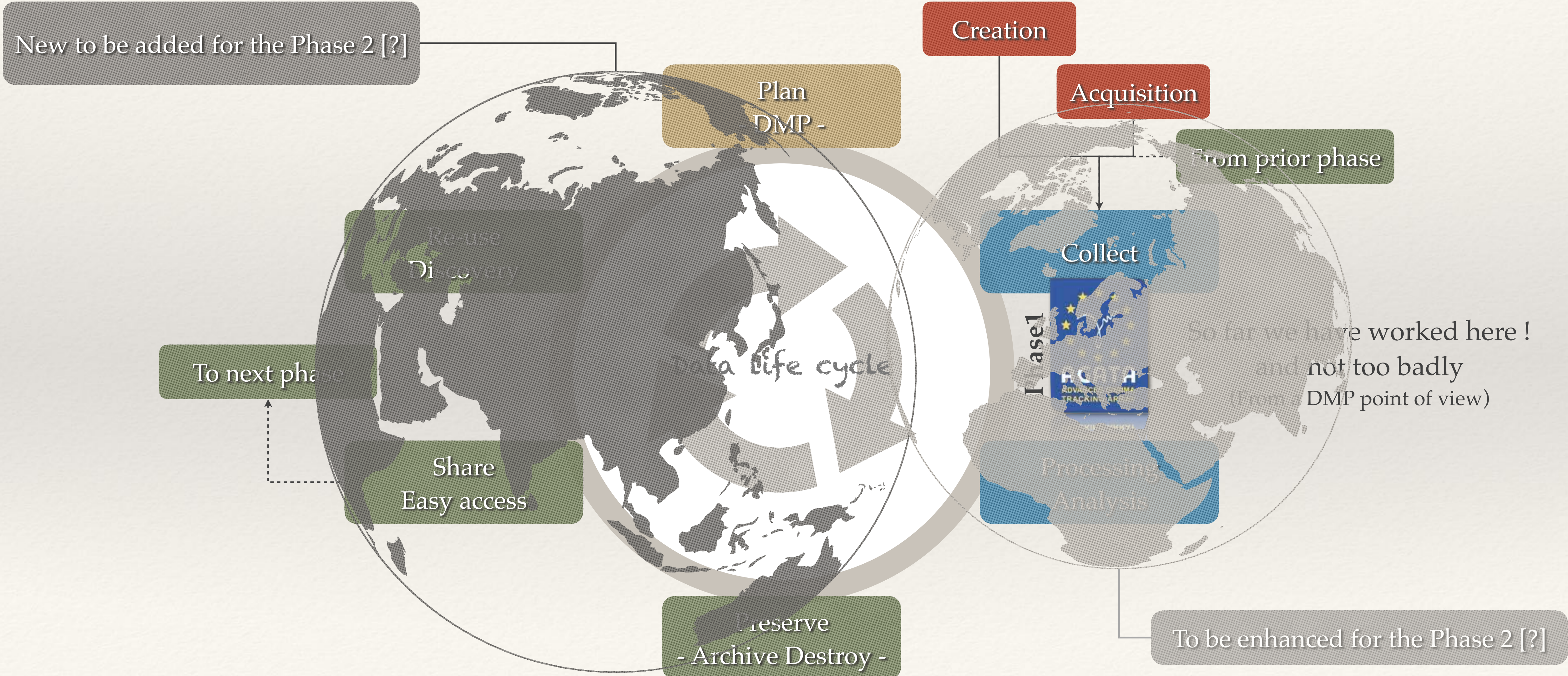
External context : what is a DMP ?

It means we should establish a plan to deal with the cycle of life of the data produced by AGATA ...



External context : what is a DMP ?

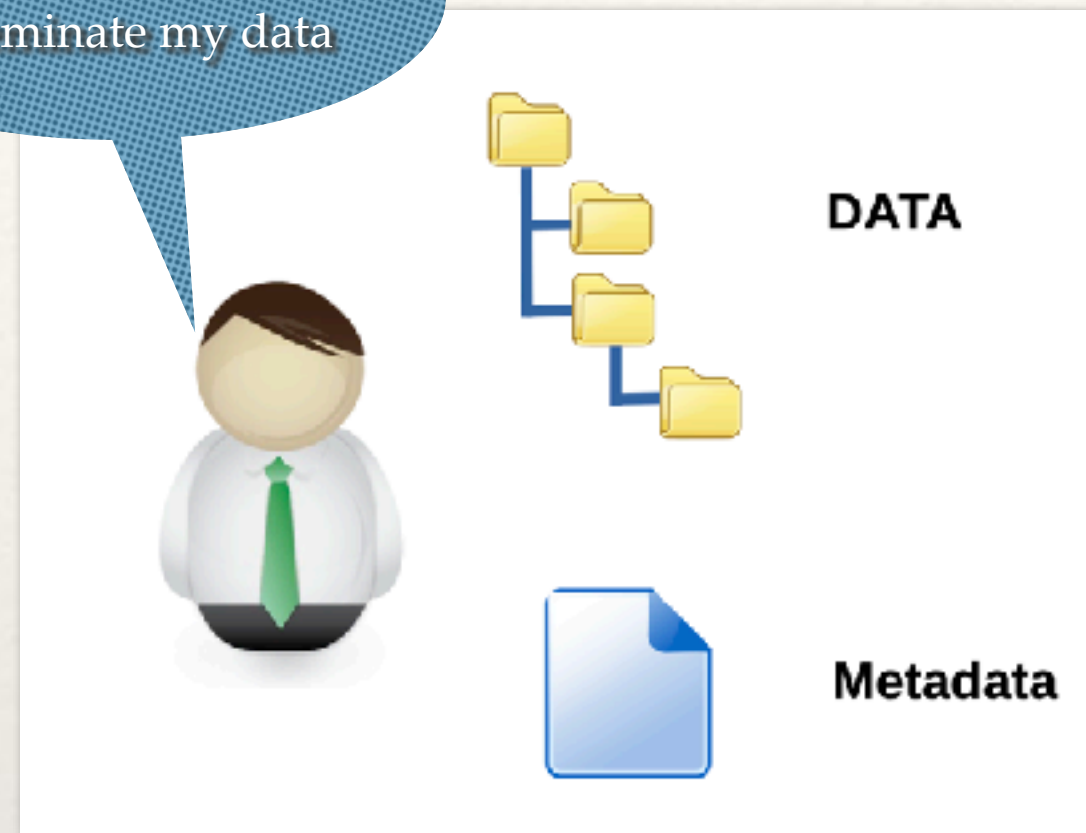
It means we should establish a plan to deal with the cycle of life of the data produced by AGATA ...



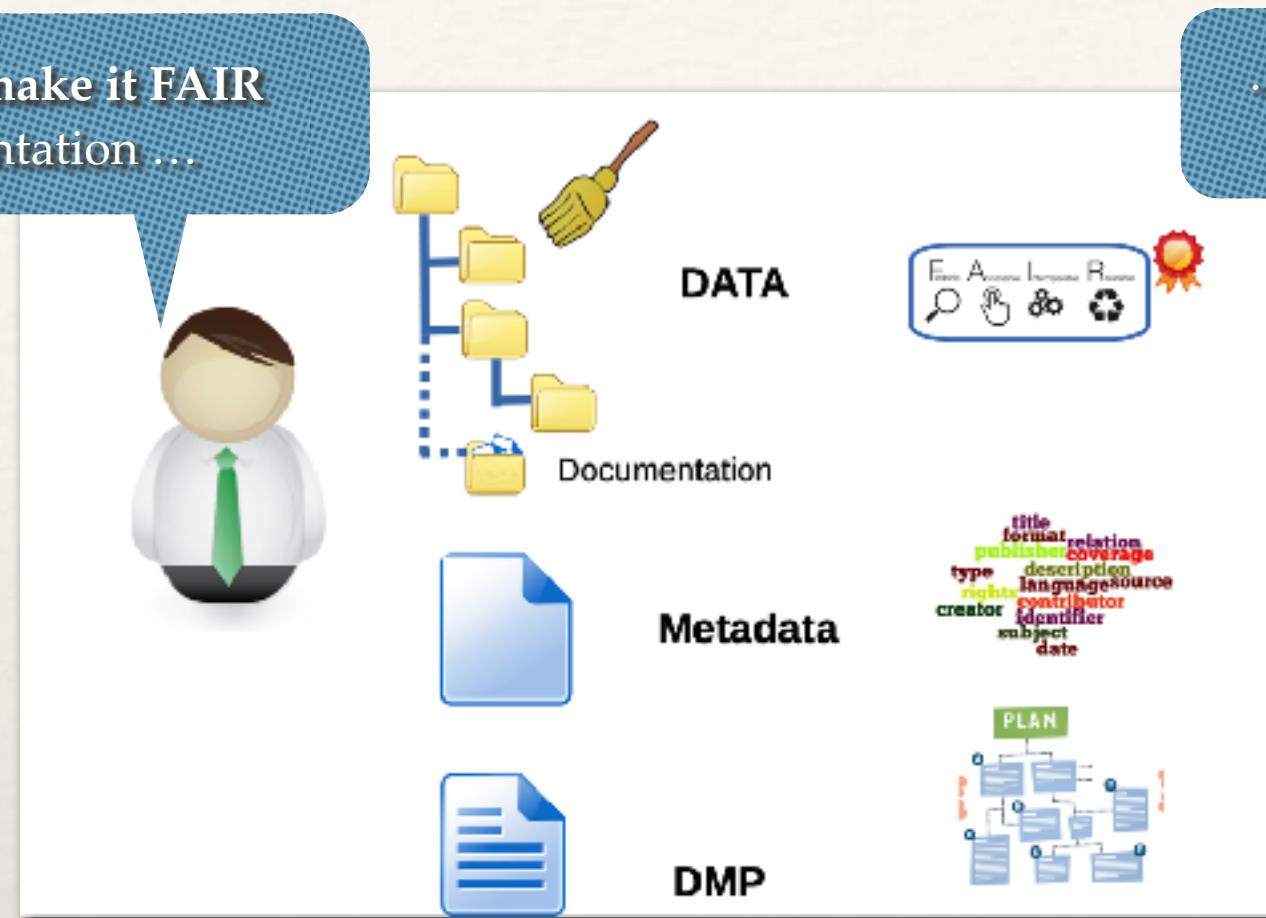
External context : what is a DMP ?

How to fulfil the cycle ?

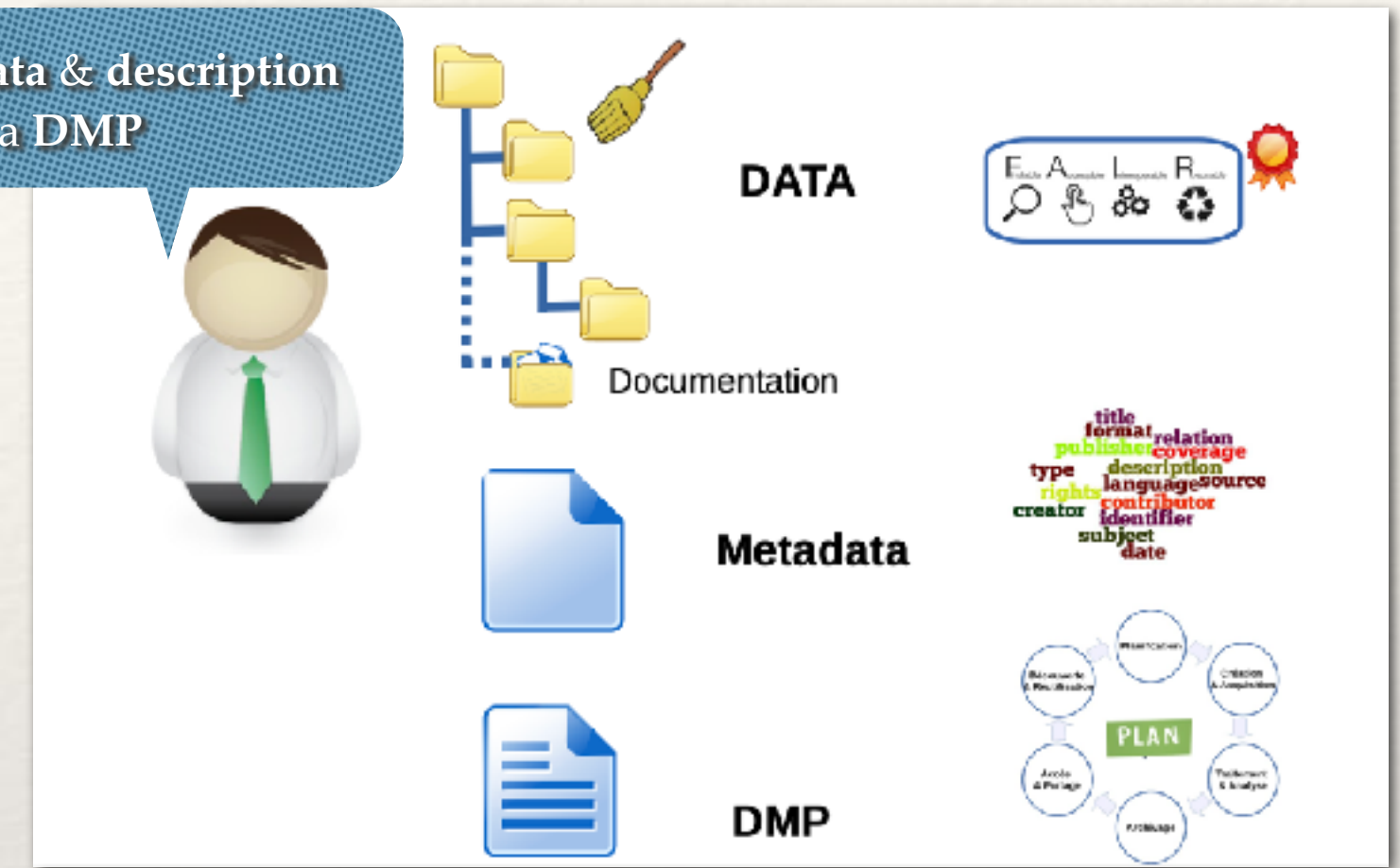
I want to archive and disseminate my data



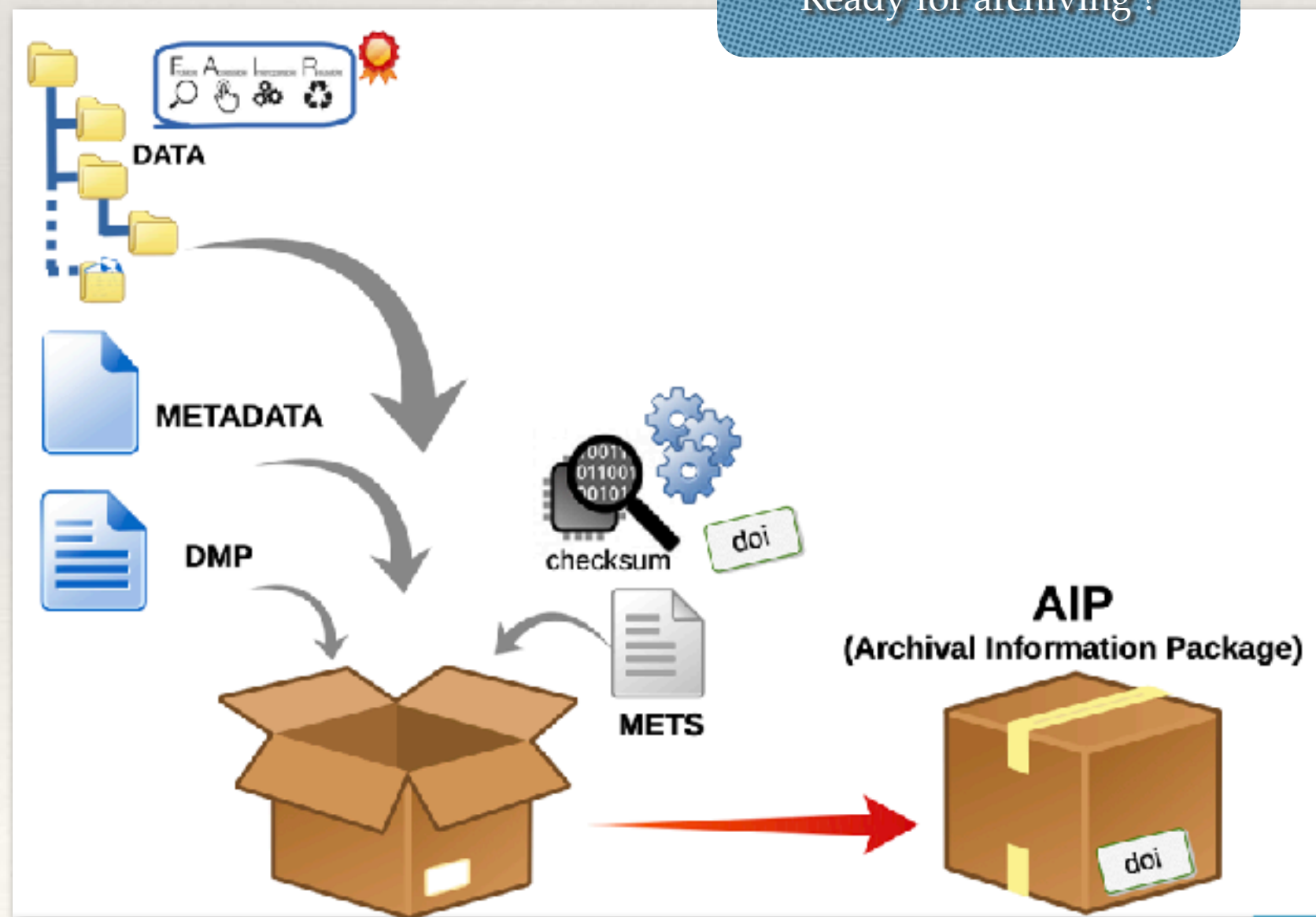
Clean the data, make it FAIR
Add documentation ...



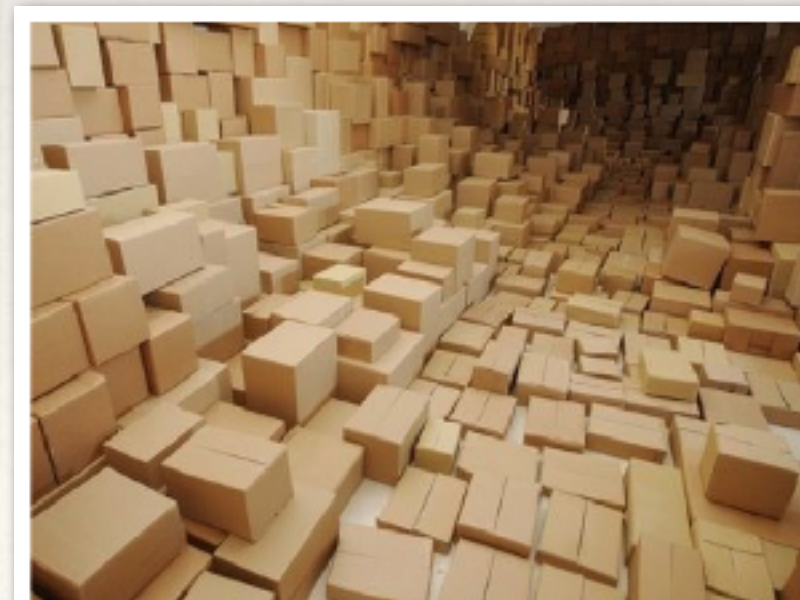
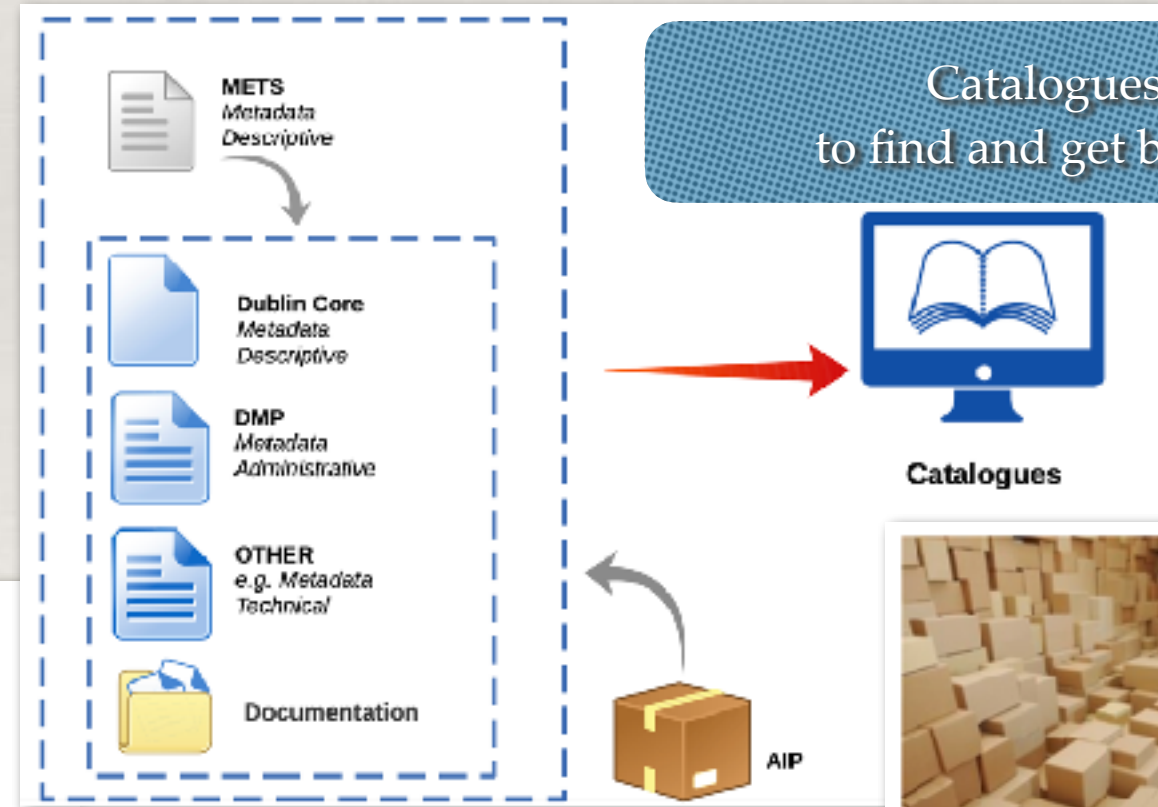
... Rich meta-data & description
Write a DMP



Ready for archiving ?



Catalogues required
to find and get back an archive

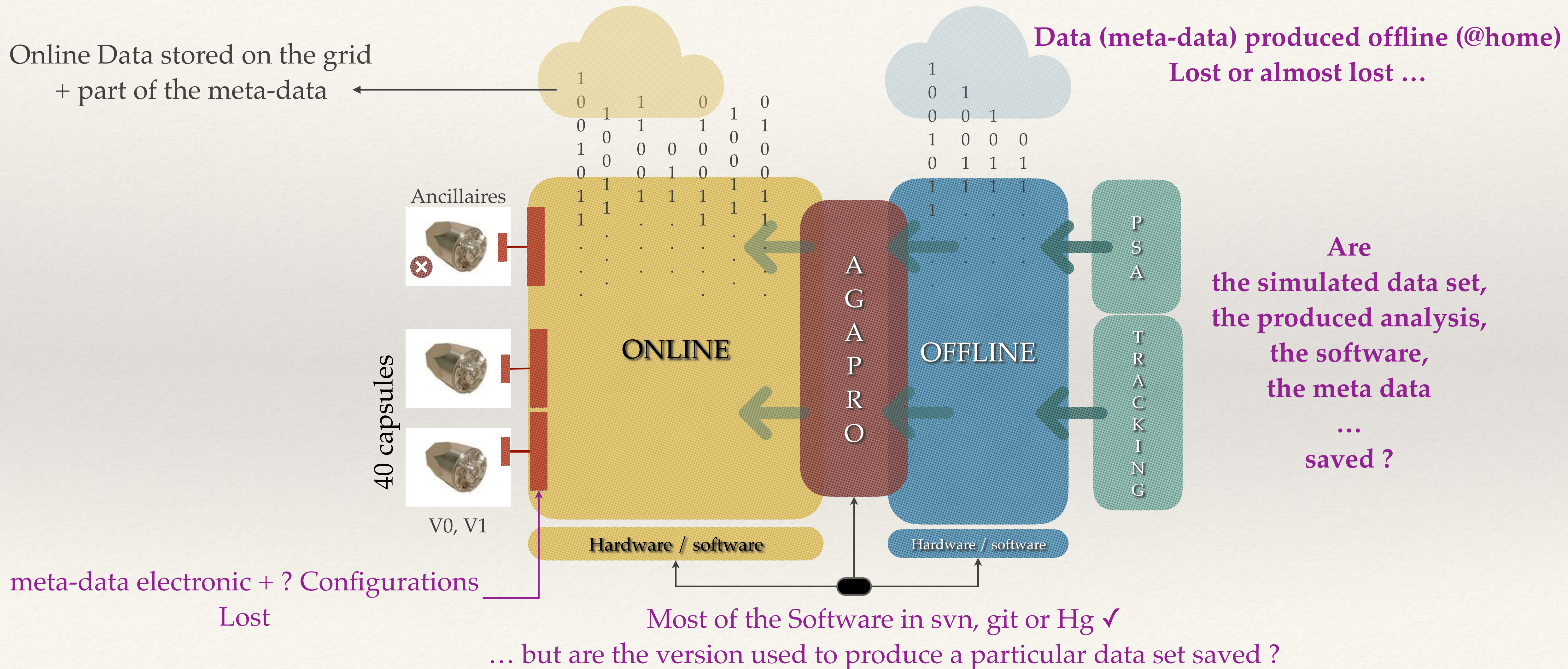


We don't want this ...



Internal Context : what AGATA produced as Data

Schematic view of the Data Production



Phase 1 : our practices so far ...

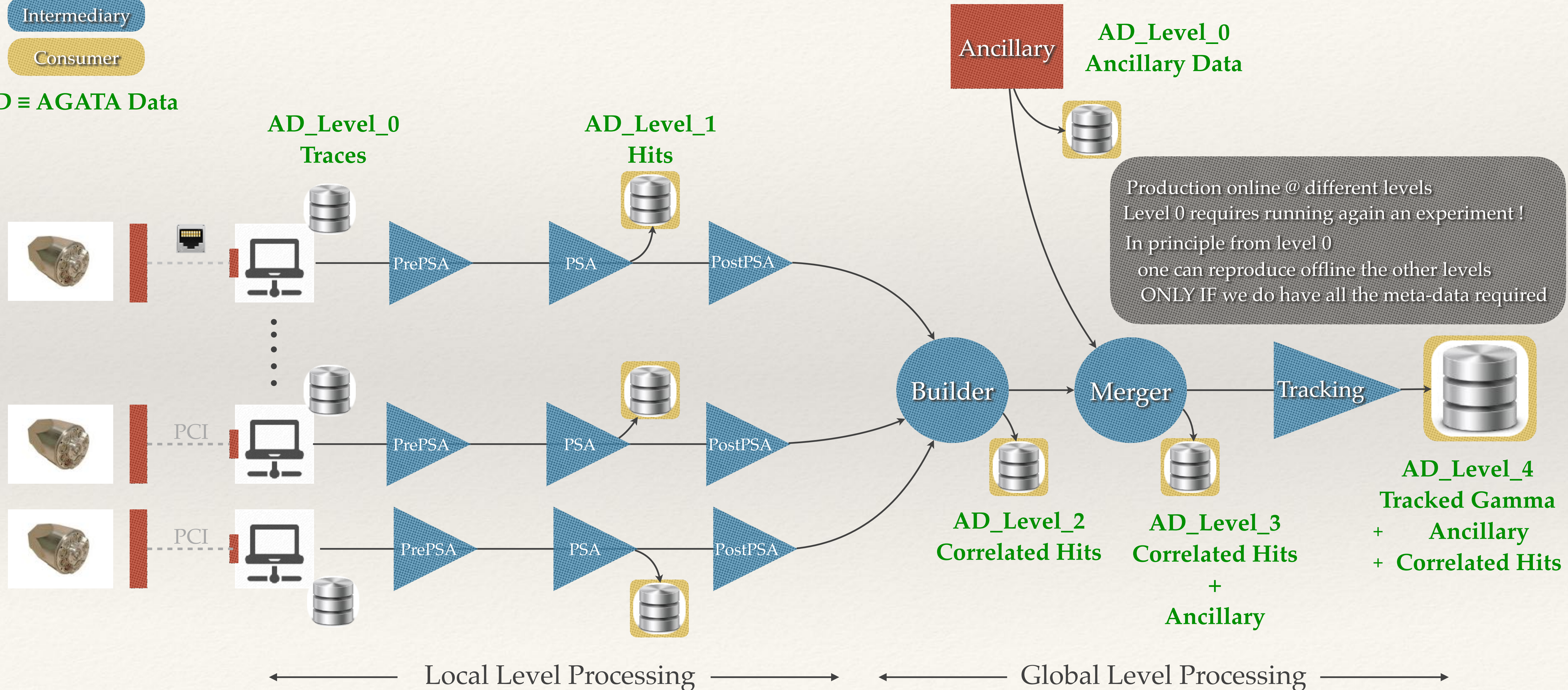
More in the document called ADP-Part1, to be sent to the ACC

Producer

Intermediary

Consumer

AD ≡ AGATA Data



Phase 1 : our practices so far ...

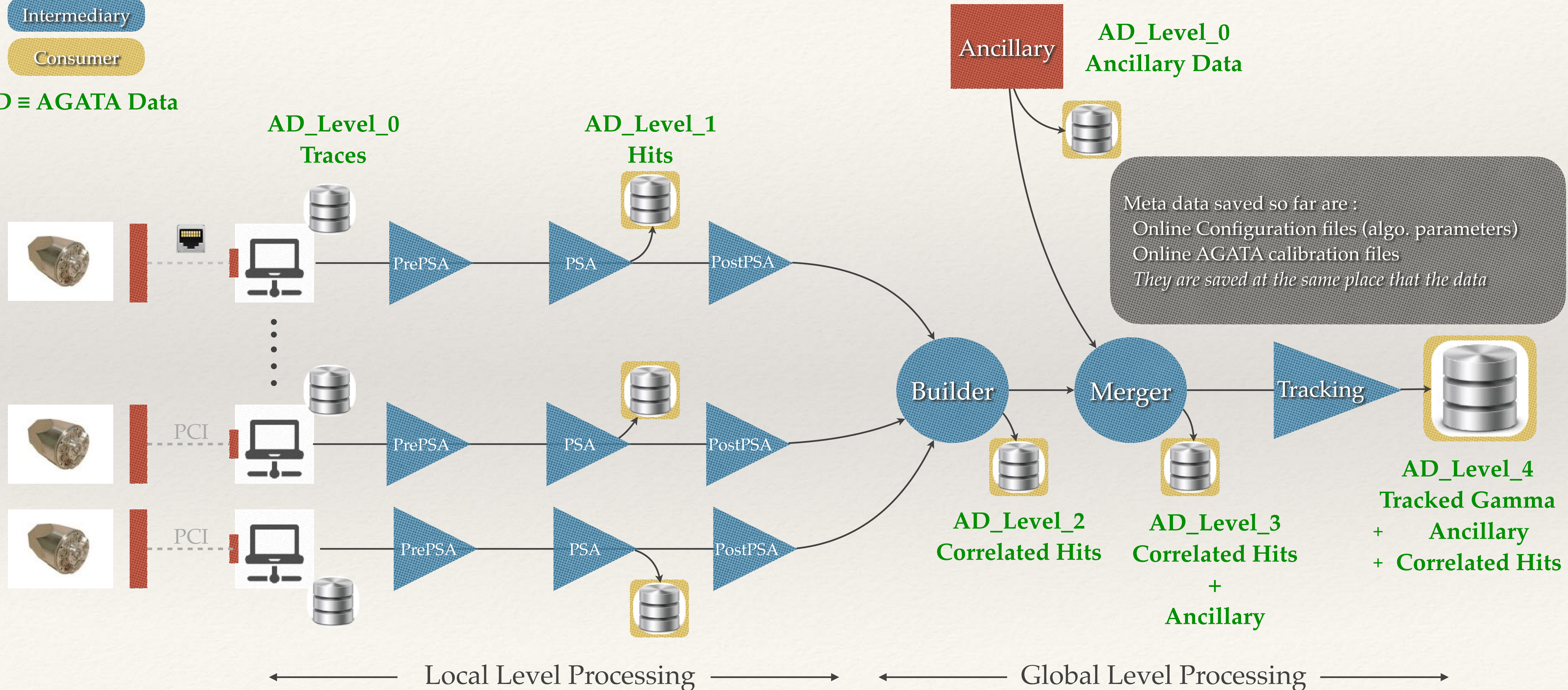
More in the document called ADP-Part1, to be sent to the ACC

Producer

Intermediary

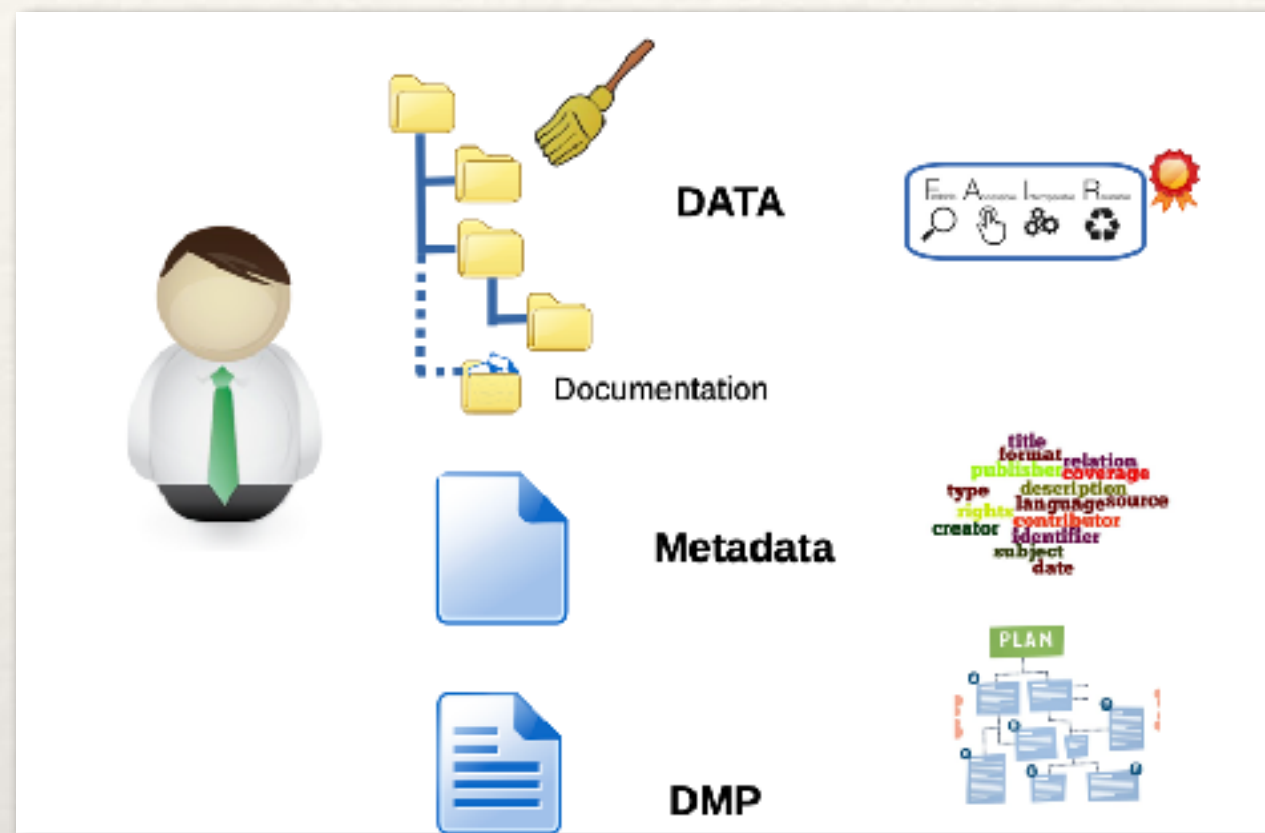
Consumer

AD ≡ AGATA Data



First step toward a DMP for the Phase 2

FAIRification of the data (see also the document called ADP-Part2, to be sent to the ACC)



Cleaning, documentation ... ok ... but what means FAIR ???

FAIR means **F**indability, **A**ccessibility, **I**nteroperability, **R**eusability

- ➔ FAIRification process, make sure the data (+meta) produced are FAIR
 - ➔ **likely to have an impact on the way we produce, store etc ... our data !**
- ➔ There are guidelines for that (see for instance <https://www.go-fair.org/fair-principles/>)
- ➔ let's have a look at some recommendations to be FAIR

Findable

- F1. Data (and meta-data) are assigned a globally unique and persistent identifier (PID)*
- F2. Data are described with rich meta data
- F3. Meta data clearly and explicitly include the identifier of the data they described
- F4. (Meta)data are registered or indexed in a searchable resource

* an example of PID is DOI. PID = web page stored in a repository
(See for instance zenodo)

F1. Obviously not the case

➔ *it might be good to start with at least a standard name for AGATA experiments*

F2. We have only a minimal amount of meta data **and** only for online data

F3. Our meta data are stored inside the data ...

➔ *metadata and data should be separated, see also A2*



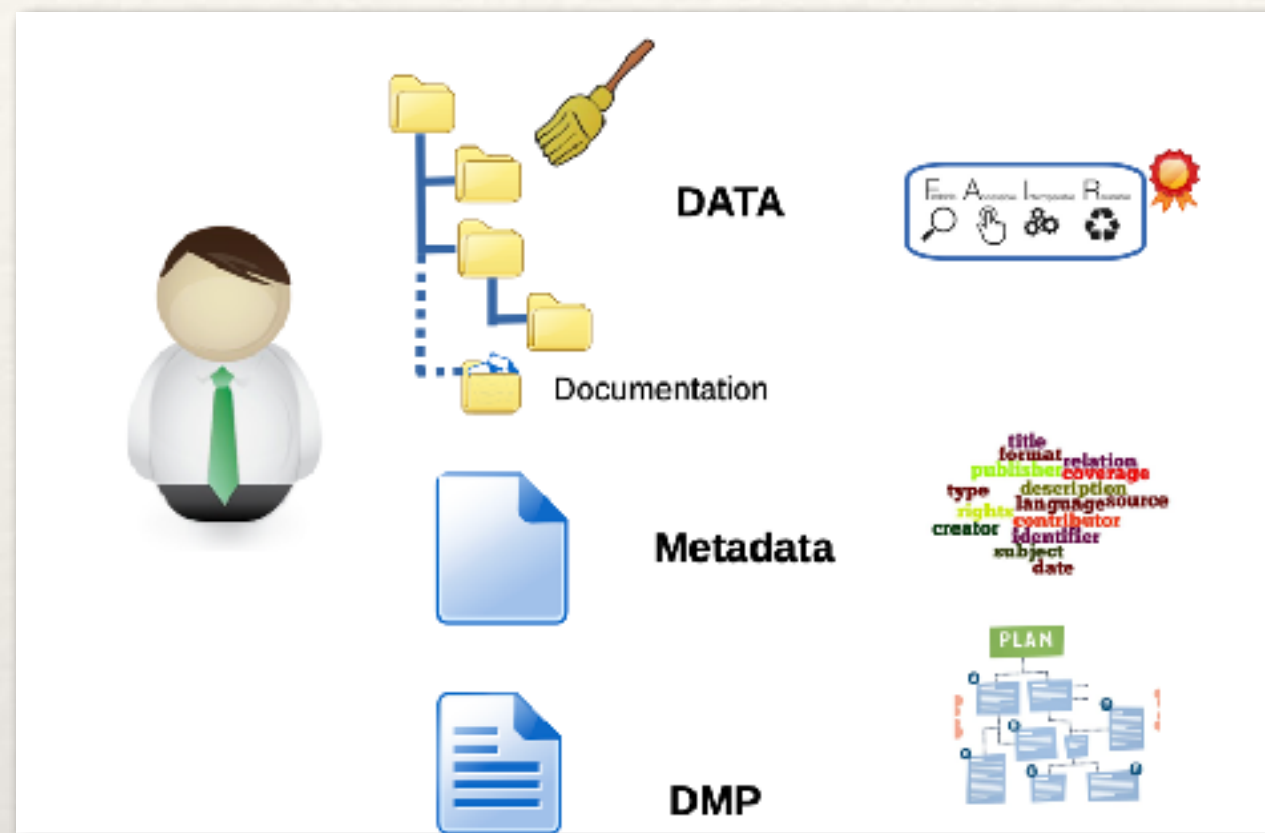
F4. Obviously not the case (see again zenodo)

➔ *searchable by humans and computers !*

➔ *Ex of a search: try and find all the data set produced at GANIL with NEDA ?*

First step toward a DMP for the Phase 2

FAIRification of the data (see also the document called ADP-Part2, to be sent to the ACC)



Cleaning, documentation ... ok ... but what means FAIR ???

FAIR means **F**indability, **A**ccessibility, **I**nteroperability, **R**eusable

- ➔ FAIRification process, make sure the data (+meta) produced are FAIR
 - ➔ **likely to have an impact on the way we produce, store etc ... our data !**
- ➔ There are guidelines for that (see for instance <https://www.go-fair.org/fair-principles/>)
- ➔ let's have a look at some recommendations to be FAIR

Accessible

- A1. (Meta)data are retrievable by their identifier using a standardised communication protocol
 - A1.1. The protocol is open, free and universally implementable
 - A1.2. The protocol, where necessary, allows for an authentication & authorisation procedure
- A2. Metadata are accessible, even when the data are no longer available

A1. Grid access

- ➔ *difficult to retrieve a particular data set without browsing all*

A1.1 Grid access

- ➔ *Not completely universal, heavy for the collaboration, time to simplify if possible*

A1.2 AGATA Virtual Organisation

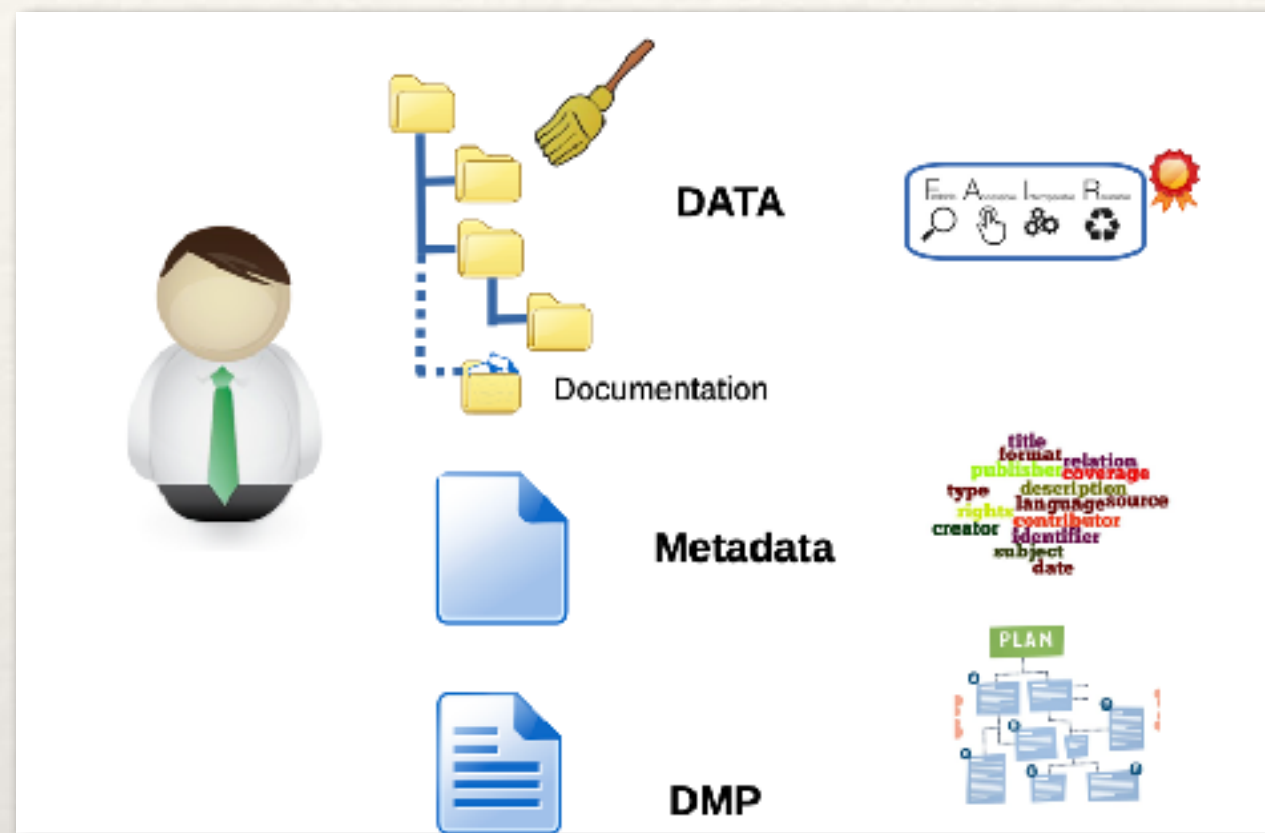
- ➔ *is this enough ?*

A2. Obviously not the case

- ➔ *AGAIN meta data should be separated from data*

First step toward a DMP for the Phase 2

FAIRification of the data (see also the document called ADP-Part2, to be sent to the ACC)



Cleaning, documentation ... ok ... but what means FAIR ???

FAIR means **F**indability, **A**ccessibility, **I**nteroperability, **R**eusable

- ➔ FAIRification process, make sure the data (+meta) produced are FAIR
 - ➔ **likely to have an impact on the way we produce, store etc ... our data !**
- ➔ There are guidelines for that (see for instance <https://www.go-fair.org/fair-principles/>)
- ➔ let's have a look at some recommendations to be FAIR

Interoperable

- I1. (meta)data use a normal, accessible, shared and broadly applicable language for knowledge representation
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. Meta-data qualified references to other (meta)data

This is for integration of AGATA data with other data ...

Almost nothing done so far to help in that path ...

Re Usable

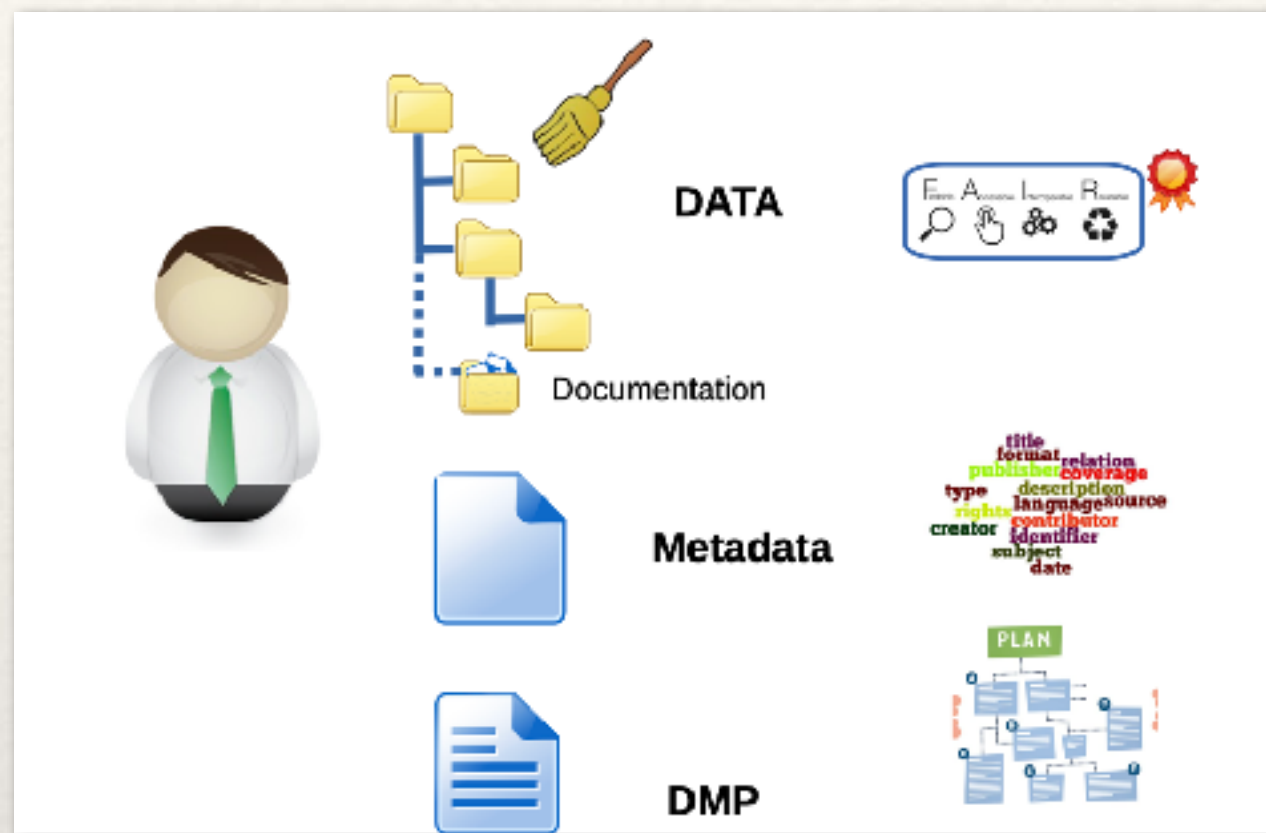
- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible usage licence
 - R1.2. (meta)data are associated with detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards

This is 'others' to play with AGATA data ...

Almost nothing done so far to help in that path ...

Writing a DMP for the Phase 2

First proposition (see also the document called ADP-Part2, to be sent to the ACC)



Many guidelines to really write a DMP
Here from the French ANR agency
All are quite similar
Moving from one to another one should be easy

1. Data description and collection, re-use of existing data

- A. How will new data be collected or produced and/or **how will existing data be re-used?**
- B. What data (for example the kinds, formats, and volumes) will be collected or produced?

2. Documentation and data quality

- A. **What metadata** and documentation (for example the methodology of data collection and way of organizing data) will accompany data?
- B. What data quality control measures will be used?

3. Storage and backup during research process

- A. **How will data and metadata be stored and backed up during the research process?**
- B. How will data security and protection of sensitive data be taken care of during the research?

4. Legal and ethical requirements, codes of conduct

- A. if personal data are processed, how will compliance with legislation on personal data and on data security be ensured?
- B. How will other legal issues, such as **intellectual property rights and ownership, be managed?** What legislation is applicable?
- C. How will possible ethical issues be taken into account, and codes of conduct followed?

5. Data sharing and long-term preservation

- A. **How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?**
- B. **How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)?**
- C. **What methods or software tools will be needed to access and use the data?**
- D. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

6. Data management responsibilities and resources

- A. **Who** (for example role, position, and institution) will be **responsible for data management** (i.e. the **data steward**)?
- B. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

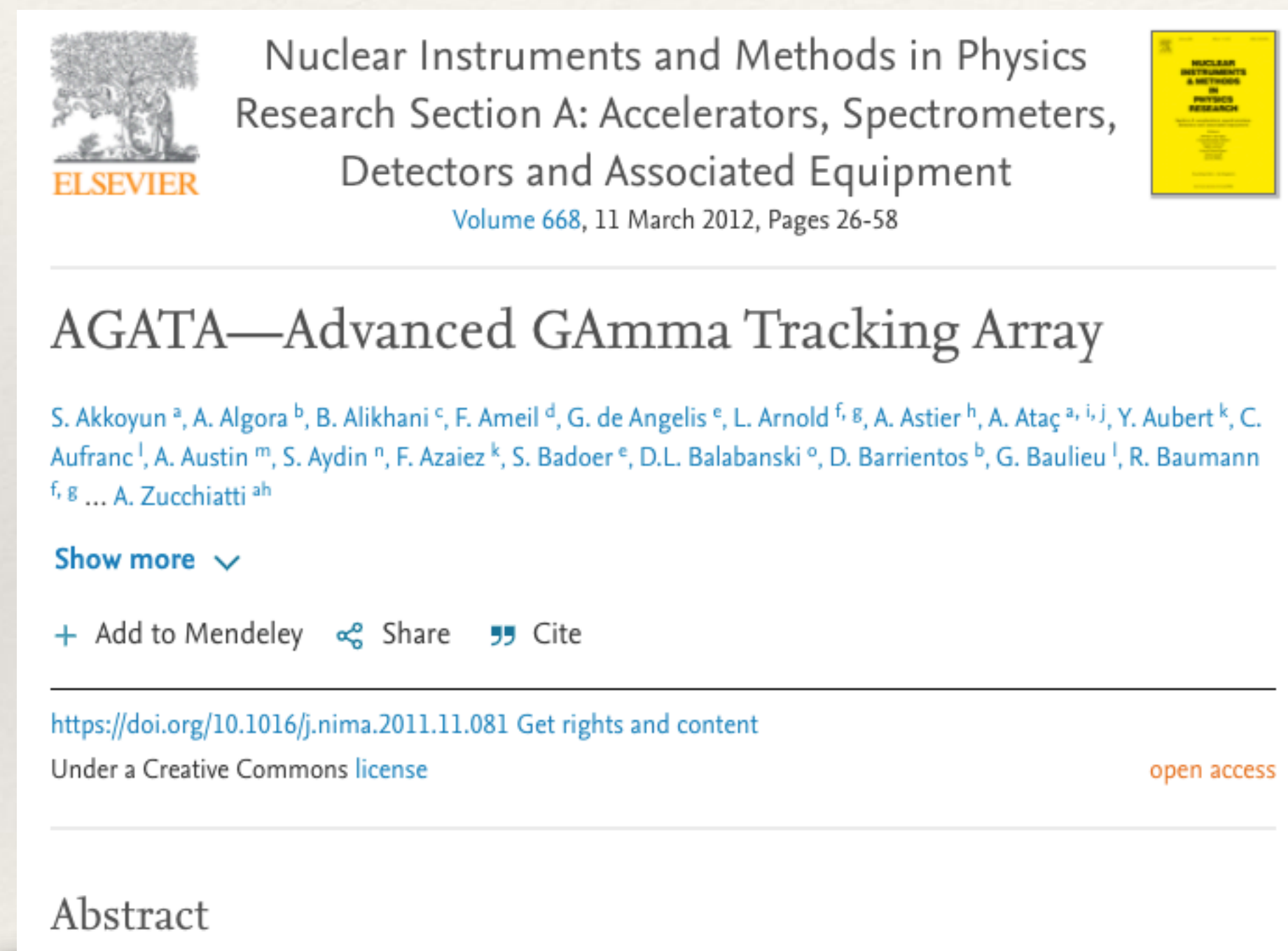


AGATA
Data Policy
for
Publications

Impacts on future AGATA users'

A concrete exemple:

let's assume in a near future, I would like to publish a paper using AGATA data



Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment
Volume 668, 11 March 2012, Pages 26-58

AGATA—Advanced GAMMA Tracking Array

S. Akkoyun ^a, A. Algora ^b, B. Alikhani ^c, F. Ameil ^d, G. de Angelis ^e, L. Arnold ^{f, g}, A. Astier ^h, A. Ataç ^{a, i, j}, Y. Aubert ^k, C. Aufranc ^l, A. Austin ^m, S. Aydin ⁿ, F. Azaiez ^k, S. Badoer ^e, D.L. Balabanski ^o, D. Barrientos ^b, G. Baulieu ^l, R. Baumann ^{f, g} ... A. Zucchiatti ^{ah}

<https://doi.org/10.1016/j.nima.2011.11.081> Get rights and content

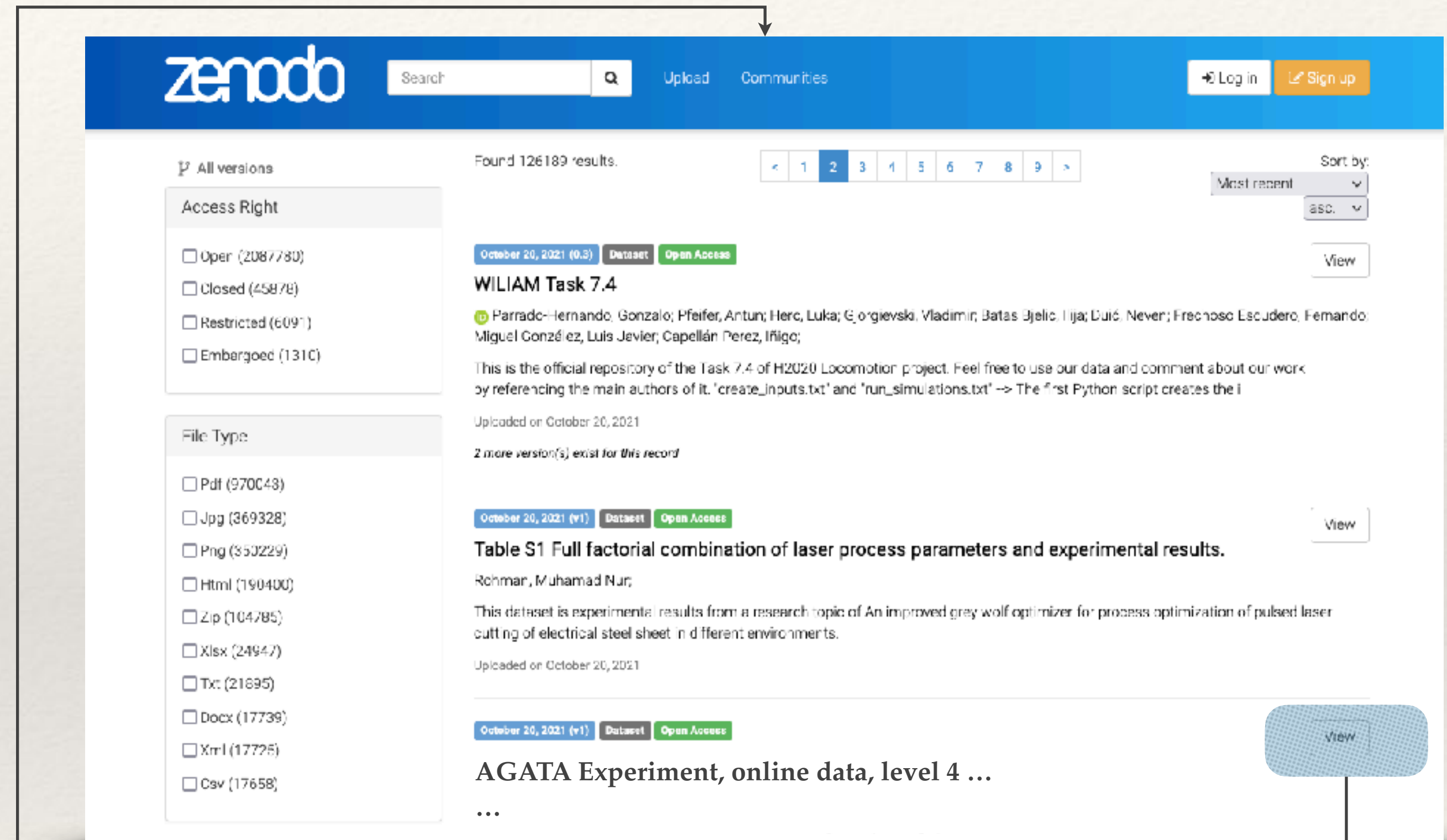
Under a Creative Commons license open access

Abstract

I have to add a reference (PID) to the data set used !!

References

- PID Dataset
- PID Software
- PID Methods



zenodo Search Upload Communities Log in Sign up

All versions Found 126189 results. Sort by: Most recent asc. View

Access Right

- Open (2087780)
- Closed (45878)
- Restricted (6091)
- Embargoed (1310)

File Type

- Pdf (970043)
- Jpg (369328)
- Png (350229)
- Html (190400)
- Zip (104785)
- Xlsx (24947)
- Txt (21695)
- Docx (17739)
- Xml (17725)
- Csv (17658)

October 20, 2021 (v1) Dataset Open Access View

WILIAM Task 7.4

Parrado-Hernando, Gonzalo; Pfeifer, Antun; Herc, Luka; Gorgievski, Vladimir; Batas Ujelic, Ilija; Duić, Neven; Frechoso Escudero, Fernando; Miguel González, Luis Jevier; Capellán Pérez, Iñigo;

This is the official repository of the Task 7.4 of H2020 Locomotion project. Feel free to use our data and comment about our work by referencing the main authors of it. 'create_inputs.txt' and 'run_simulations.txt' -> The first Python script creates the i

Uploaded on October 20, 2021

2 more version(s) exist for this record

October 20, 2021 (v1) Dataset Open Access View

Table S1 Full factorial combination of laser process parameters and experimental results.

Rohmar, Muhammad Nur;

This dataset is experimental results from a research topic of An improved gray wolf optimizer for process optimization of pulsed laser cutting of electrical steel sheet in different environments.

Uploaded on October 20, 2021

October 20, 2021 (v1) Dataset Open Access View

AGATA Experiment, online data, level 4 ...

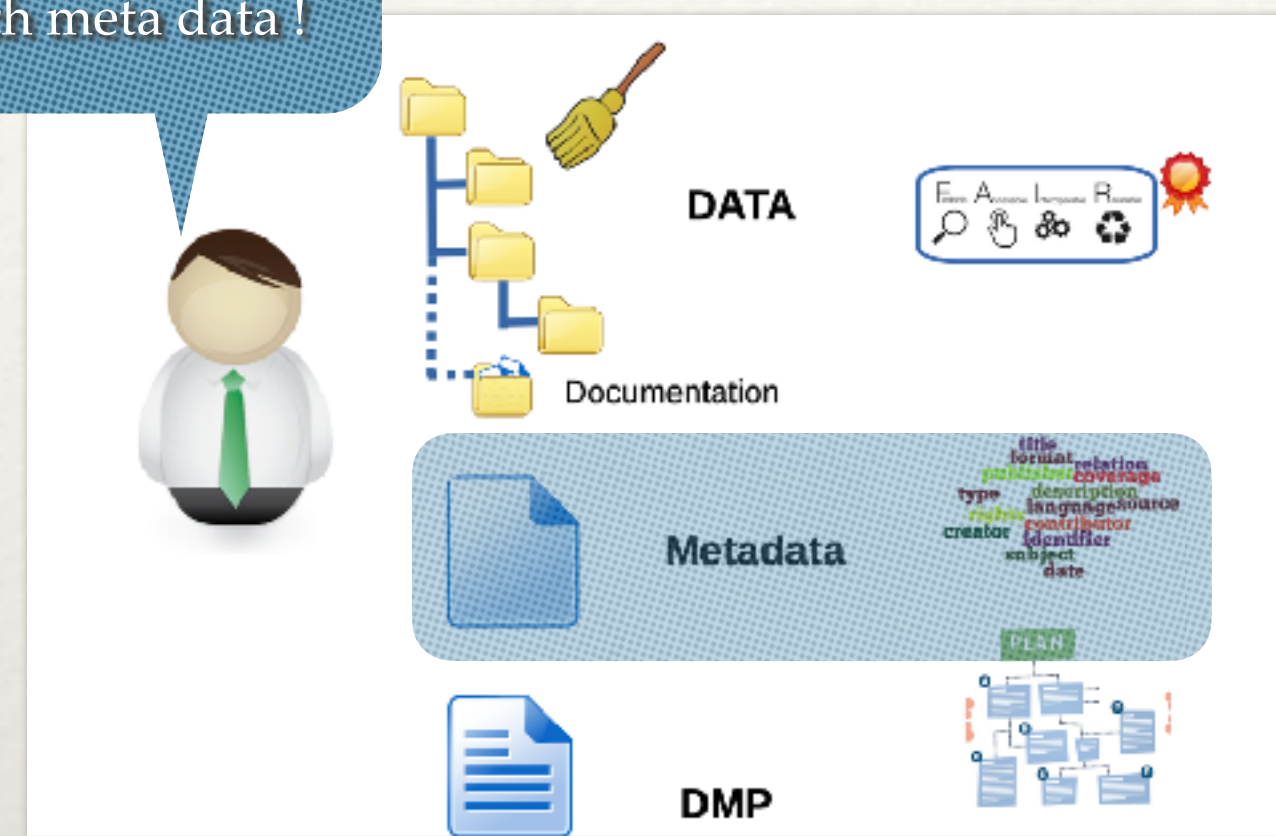
...

Many informations included

- Data set description, were to find it, produced when, by who, with what ...
- Additional detectors, beam used, beam current ...
- What tracking has been used with which parameters ...
- License attached to the data set ...

Metadata : some questions to the ACC

Remember :
rich meta data !



Description with metadata of a dataset, from the point of view of the collaboration
(+more technical from the Data Processing group)

Minimal
(Almost standard & mandatory)

- **Unique name***
Local name Ex : e780
Campaign Ex: NEDA DIAMANT
- **Context***
Short description*
Full description*
- **Ownership***
- **Date of creation**
- **Licence, policy at date**
- **Configuration AGATA**
of Ge crystal etc ...
- **link to elog**
- ...

Useful
(What you think it helps a lot)

- **Ex: Tags/keywords ?**
#Beam #target #energy
- **Ex: List of runs, # of events ?**
- **Software versions ?**
- ...

Ideal
(Your perfect world)

- **Ex: Beam current ?**
- **Ex : PID of beam related data set ?**
- ...

*Propositions / more details
(To be validated by ACC)

Unique name

AD_P2_EXP_XXX, AD_P2_COM_XXX, AD_P2_SOU_XXX,
AD_P2_TEST_XXX, AD_P2_SIMU_XXX ... what can be XXX ??

Context

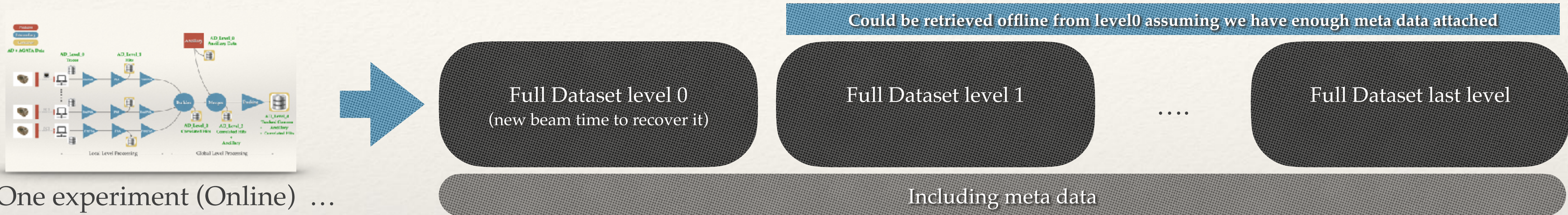
Short : a text giving the objectives of the experiment
Full : DOI pointing to the full proposal ??? OK ??

Ownership

All AGATA 'users' , spokesperson, what about ancillaries ?
period of retention in case it is protected ?
—> connected to the choice of licences !

Feedbacks from ACC now and all along the OPEN path ...

Data life cycle : some questions to the ACC



Q1 : What is to be kept, for how long ? Procedure to destroy data ?

Q2 : Ownership ? FAIR principle : “as open as possible, as closed as necessary”

Note : it has an impact on the licence to be attached

Q3 : What about AGATA (meta)data produced offline ???????

So far almost lost !

It may exist but we don't know where and how to get it !

It contains optimal calibration parameters, optimal algo. parameters !

Remember, a future paper may need a PID of the dataset used to publish !

As collaboration, do we ask/require to have back the (meta)data of at home reprocessings ?
 Could it be a requirement to be published (remember the core list procedure) ?

Propositions / more details
 (To be validated by ACC)

Online dataset

Q1: Level 0 X [∞] - Intermediate Y[2] - last Z [10] years ?

Q1: Level 0 X [∞] - Intermediate Y[first paper] - last Z [first paper] years ?

...

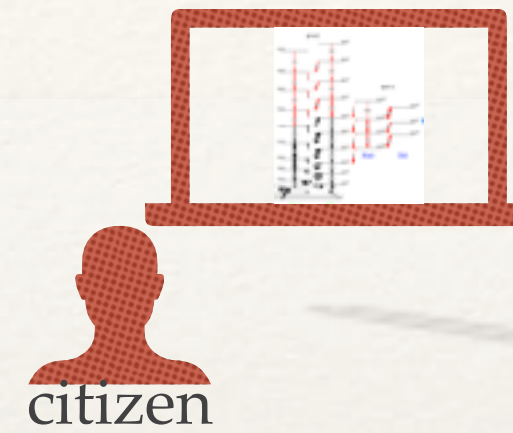
Q2: Actual is AGATA with a restriction 2-3 years for 'spokesperson'

→ 'gentleman agreement ... technically not done VO has access to all data

Q2: Per level restrictions ?

Q2: Restriction by steps ? Step1 AGATA X years then all Y years ?

Conclusions



⑤ Is it something reachable for « me » ?

④ Already some questions on the table
We also need to set a method (Work ADPG ↔ ACC)

③ ADPG* in charge of making + propositions
AGATA collaboration [ACC] to define

② We need to draw it using the DMP
A DMP is regularly modified
We do not start from 0 🖱️ progressive approach

① We are at the beginning of the 'open' path

*ADPG = AGATA Data Processing Group

Is all this useful for us (researchers)?

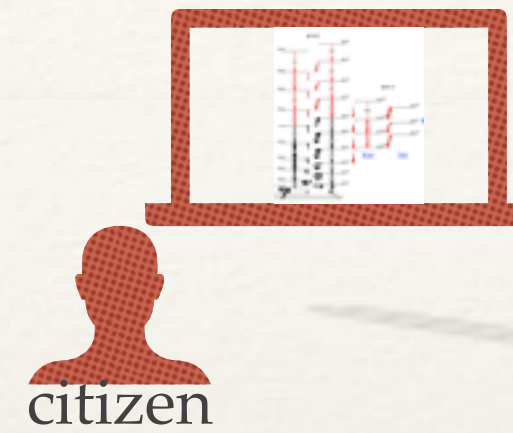
→ We may not have the choice ...

/

Imagine you would like to add / compare
a LNL experiment with a GANIL one ??

→ It should be easier with all this !

Conclusions



We already have foundations !

Ex : data analysis schools are open education

Probably many aspects missing , feel free to add !

Ex: data sharing between collaborations

④ Already some questions on the table

We also need to set a method (Work ADPG ↔ ACC)

③ ADPG* in charge of making + propositions
AGATA collaboration [ACC] to define

② We need to draw it using the DMP

A DMP is regularly modified

We do not start from 0 🖱️ progressive approach

① We are at the beginning of the 'open' path

*ADPG = AGATA Data Processing Group

Thank you for listening

Questions ?

Phase 1 : our practices so far ...

