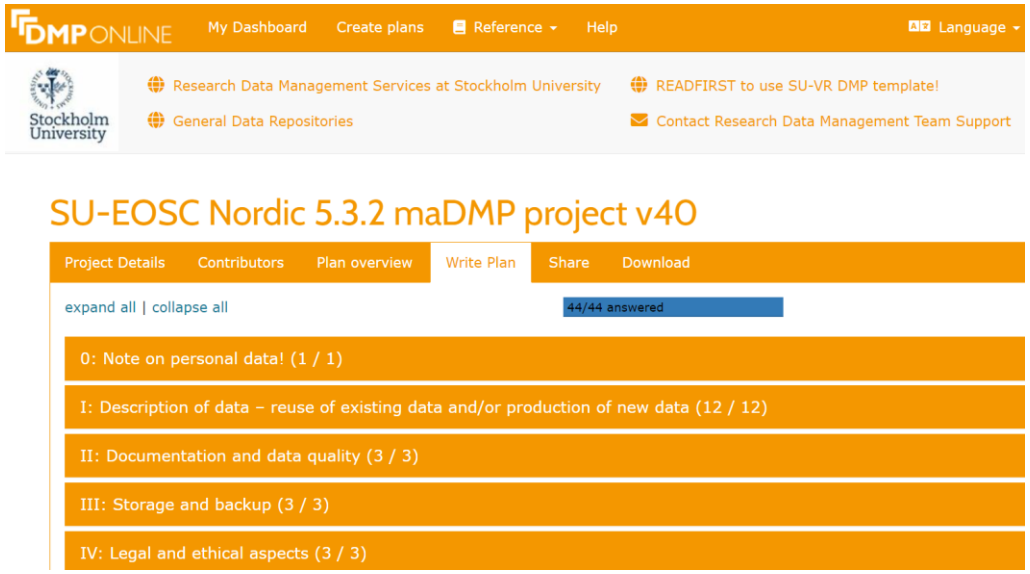
# Integrating Collection, Storage and Archiving of Research Information via machine-actionable DMPs





**PV2023 CERN, May 3, 2023**

joakim.philipson@su.se,
Stockholm University


Stockholm University

**Integration of research data and other types of research information objects** (funding applications, ethical review documents, research software, data management plans (DMPs), electronic lab notebooks (ELNs) etc.) -
**two basic stakeholder needs:**

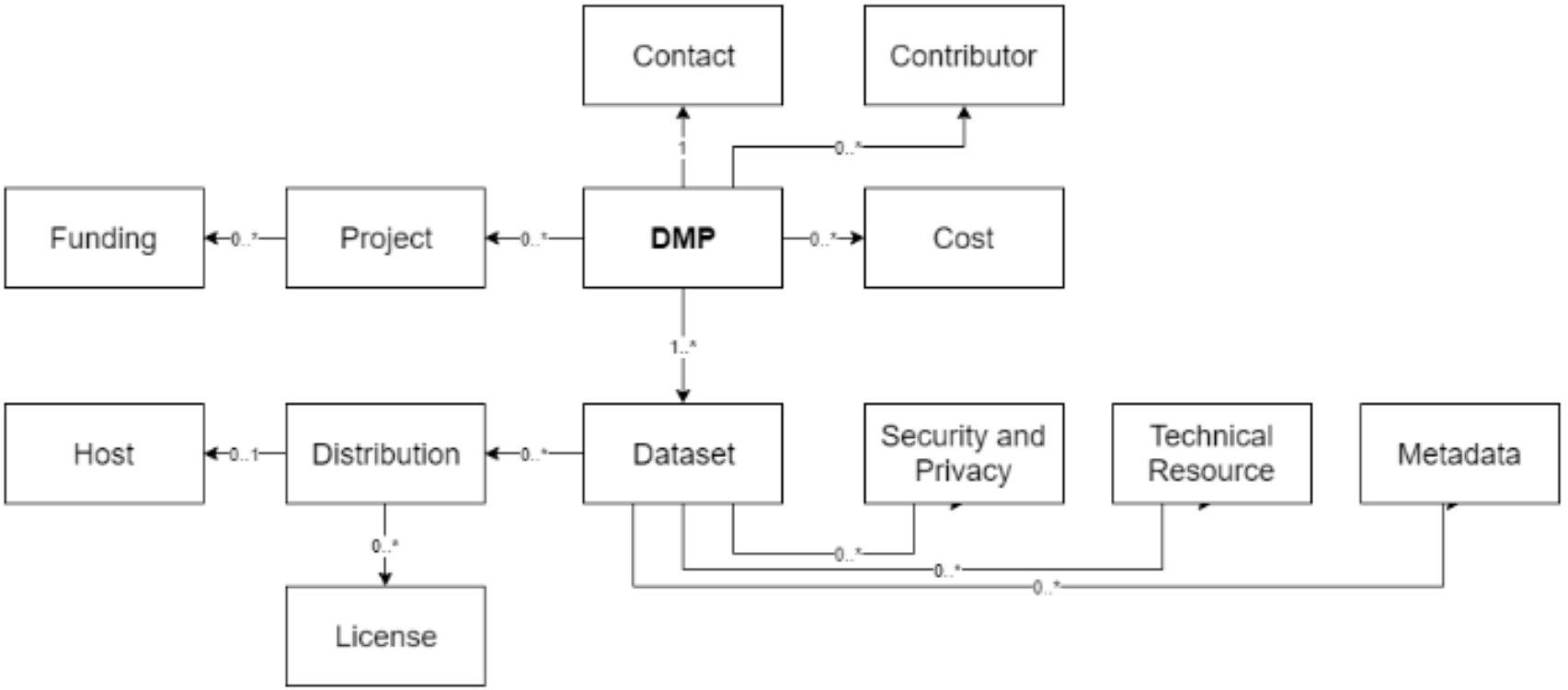
1) **Reducing the administrative burden on researchers** by means of an efficient re-use of information and metadata already present elsewhere, so that researchers will not have to fill out web-forms with the same metadata elements over-and-over again.

2) The growing ***scientometric* demand** on the part of research institution administrative staff and archivists **to monitor, describe and preserve the entire research output** of all affiliated scholars at their institution.

An efficient use of Data Management Plans (DMPs) and in particular **machine-actionable DMPs** (maDMPs) might serve both of these different stakeholder needs.

Stockholm University

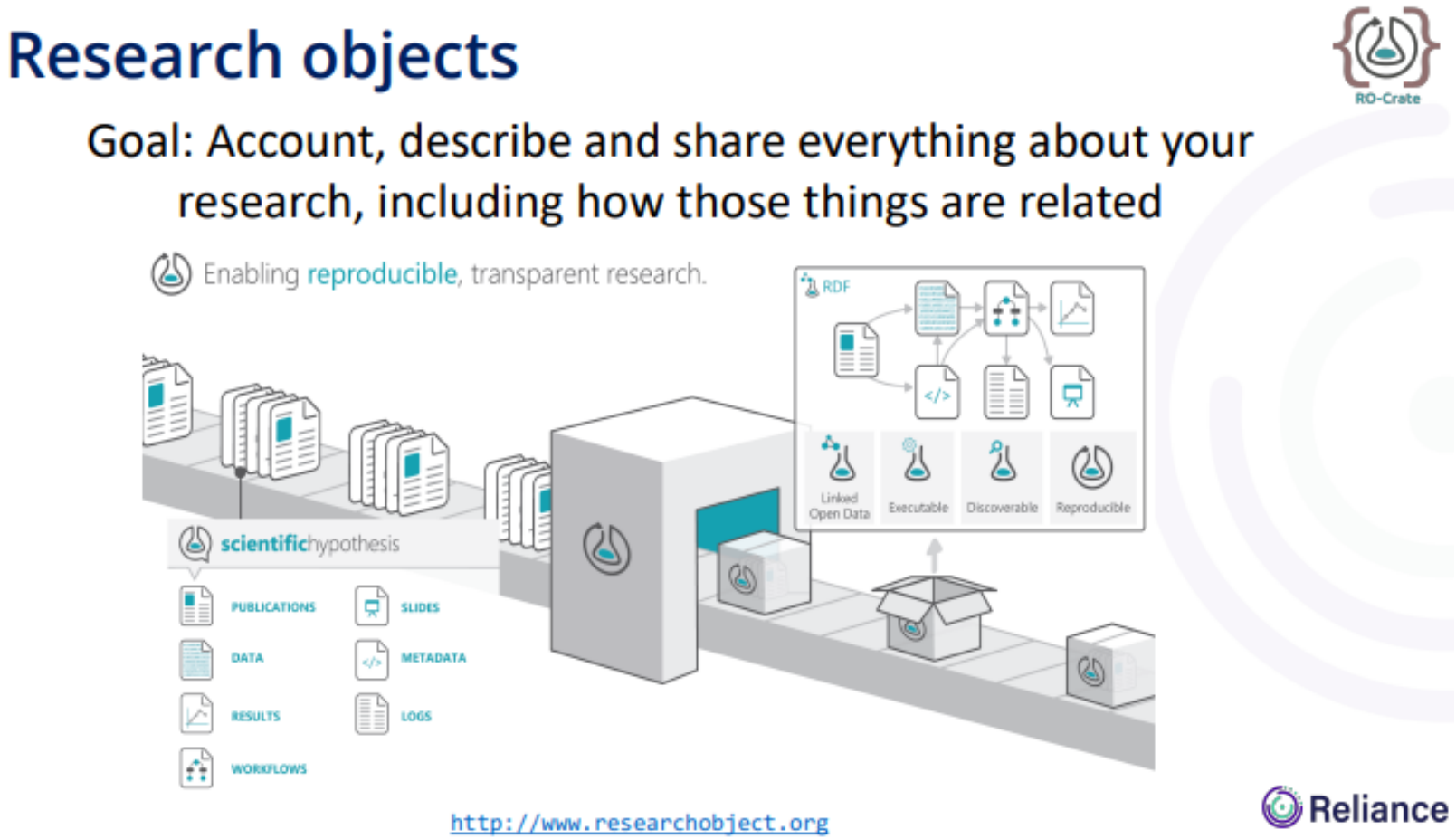
**Data Management Plans (DMPs)** and in particular **machine-actionable DMPs (maDMPs)** are described in the **RDA DMP Common Standard**
as a possible central node in a network for the integration of research information of different types




[https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard]

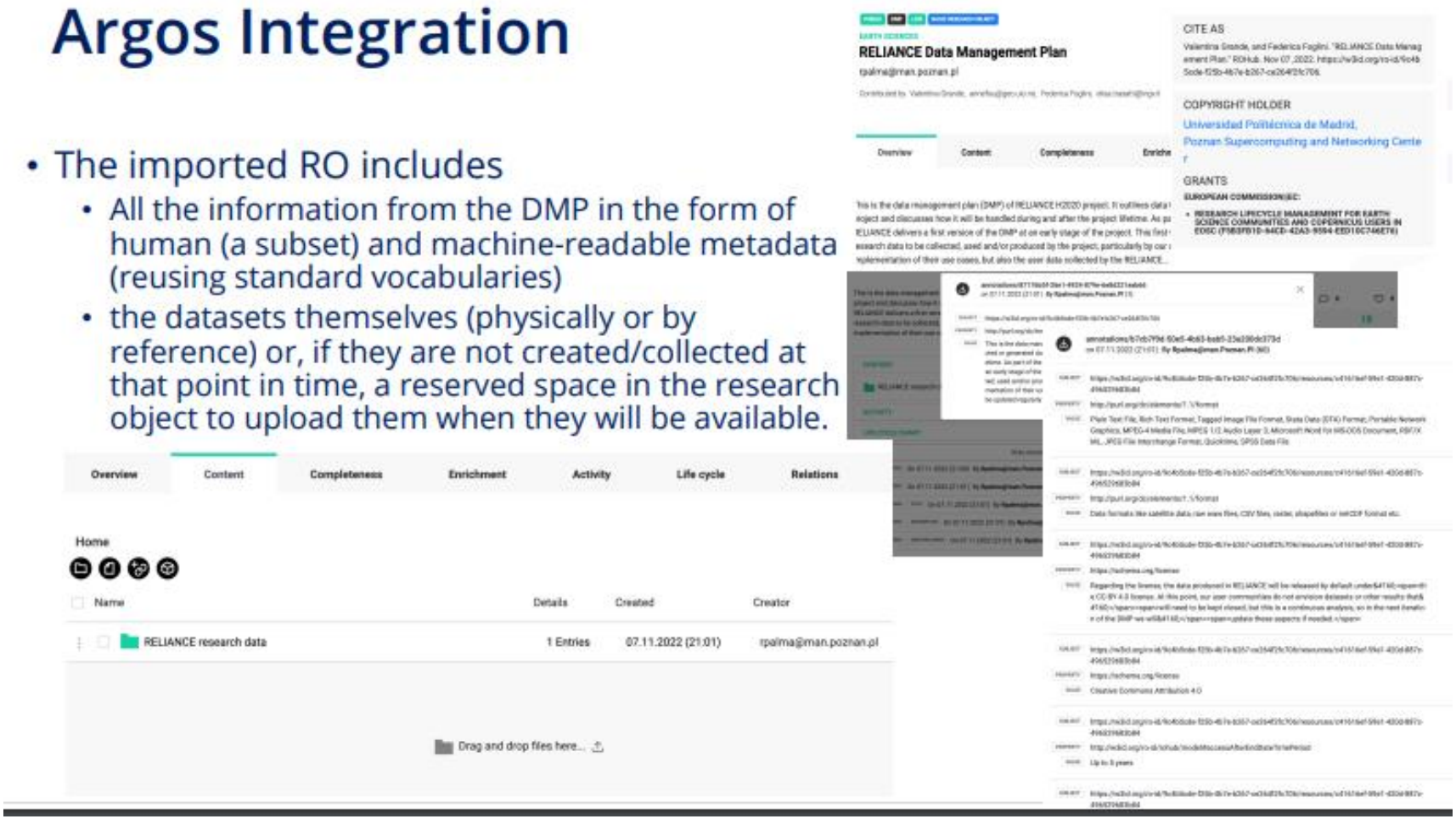
Stockholm University

Several attempts to implement the **RDA DMP Common Standard** integrative model.
**Argos** (OpenAIRE) teamed up with **ROHub** which can import Argos DMPs in XML (+JSON). Based on the imported DMP it then **generates a research object (RO)**




[https://doi.org/10.5281/zenodo.7669563]

Several attempts to implement the **RDA DMP Common Standard: Argos** (OpenAIRE) teamed up with **ROHub** which can import Argos DMPs in XML (+JSON). Based on the imported DMP it then **generates a research object (RO)**

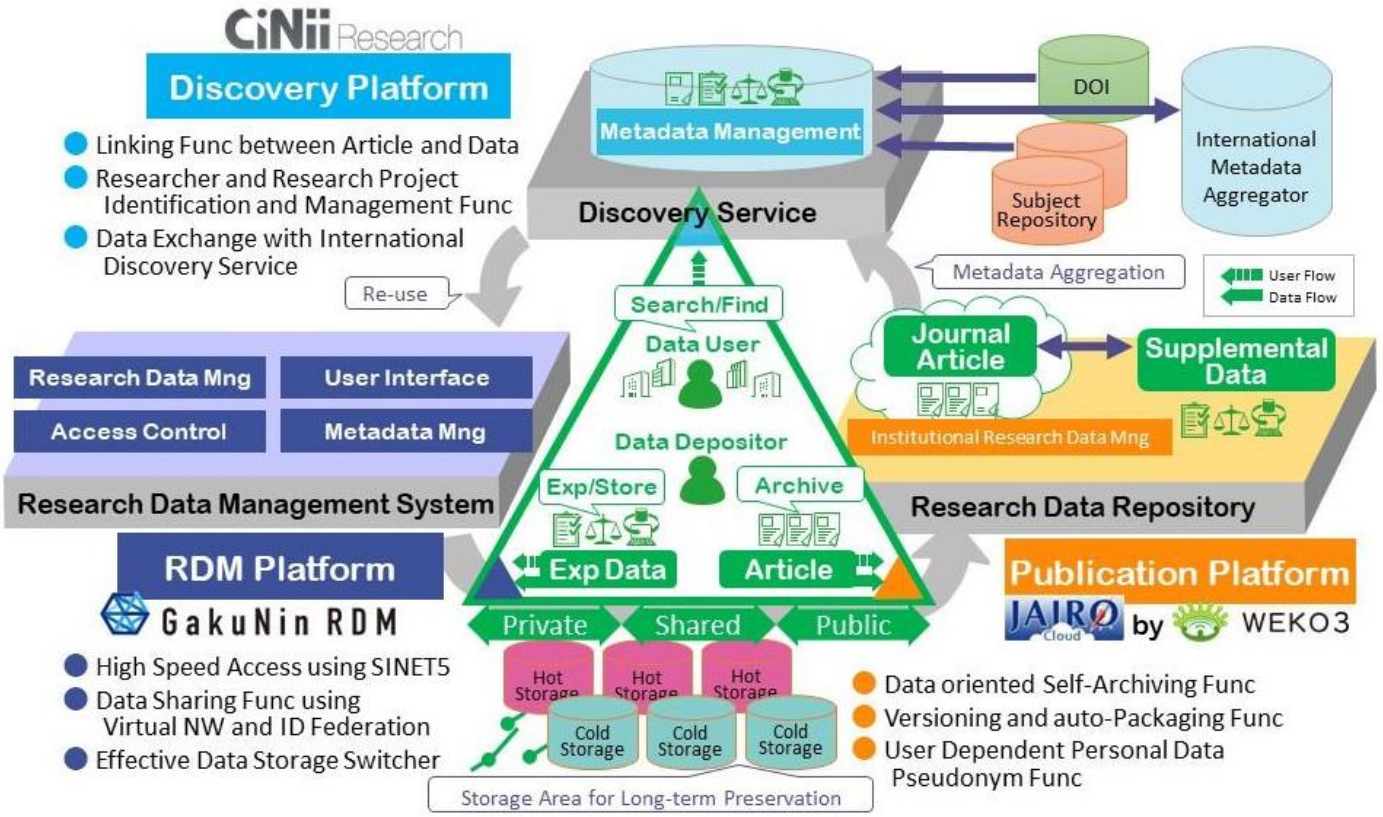



[https://doi.org/10.5281/zenodo.7669563]

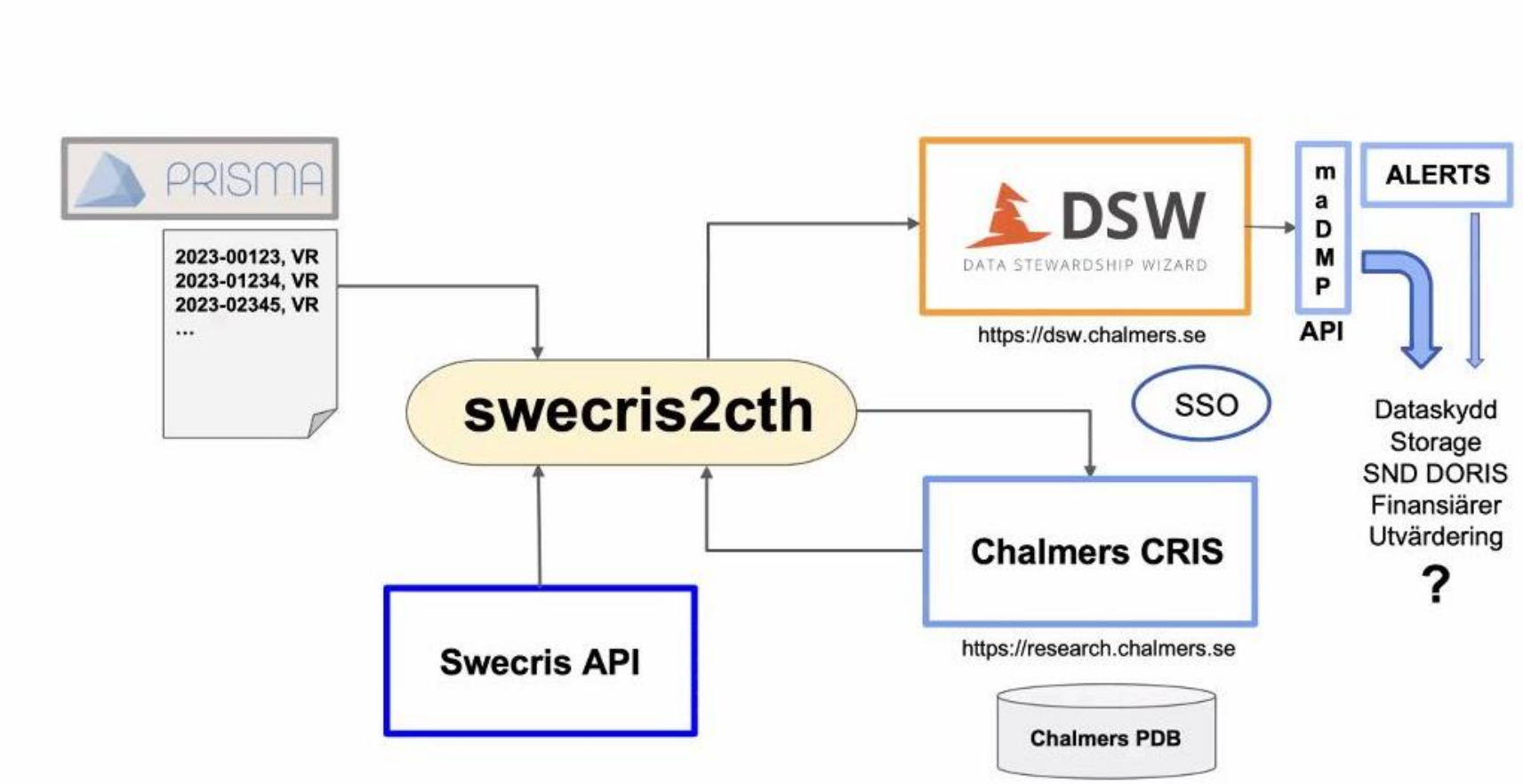
Another highly automated integrative model is that of the **NII RDC** (*National Institute of Informatics – Research Data Cloud*) in Japan.
Here **maDMPs** play a key role as "orchestrating" the NII RDC, since the automation of the data management process is based on DMPs.
[https://rcos.nii.ac.jp/en/service/]
The NII RDC has three main parts:

1. a data management platform (GakuNin RDM),
2. a publication platform (WEKO3), and
3. a discovery platform (CiNii Research)

There are also modules for secure storage and packaging together of data, program code and analysis environments that enables re-use. The NII RDC also holds both private and public, "hot" and "cold" storage areas for long-term preservation of all kinds of research objects



Another interesting attempt to implement the **RDA DMP Common Standard,** while **automating** to some extent the creation of DMPs: **Chalmers University of Technology** (Sweden) with **DSW**

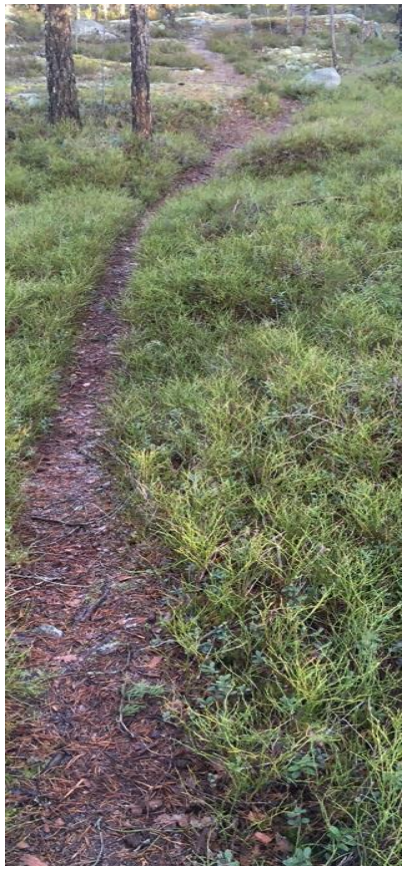


# Collection Events and Processing of Information 1:
## the Time Element and Path Dependence

- Knowledge graphs (SKG, Researchgraph.org) *connect* and describe *relations* between research information items, but little to say about actual *collection events* and *processing order*

- **Collection Events**: here *harvesting, extraction* or *other retrieval, transformation, integration* and *packaging* of (digital) research information

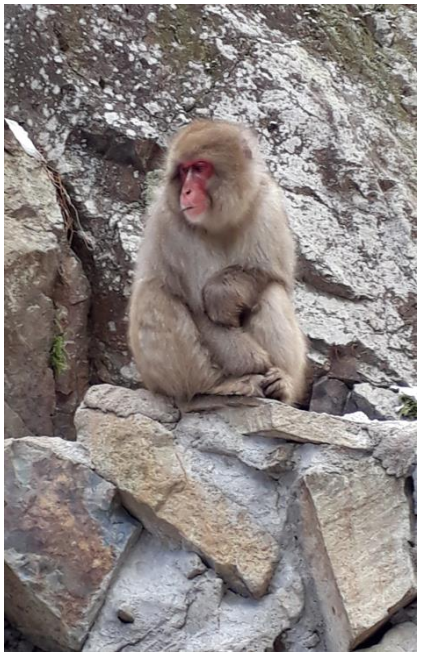
- How? When? By whom? From where?– impact on information *quality* (e.g. via metadata enrichment), data *source* selection and priorities.



# Collection Events and Processing of Information 2:

- **Collection Events**: here *harvesting, extraction* or *other retrieval, transformation, integration* and *packaging* of (digital) research information

- How? When? By whom? From where?– impact on information *quality* (e.g. by metadata enrichment), information *source* selection at the right moment in time.



# Metadata Enrichment and Quality Control Example

A DMP created in DMP Online may have on its cover page
**Funder**: Swedish Research Council (=VR) and **Grant Number**: 2022-01234

all information needed to retrieve metadata from the project application as a json-file by means of the SweCris API, provided by the Swedish Research Council.
Same funding information could also be extracted from the standard DMP Online json-export of the same DMP, or, if not found on the cover page, in the DMP Online API (v0) output of the entire DMP content, as created with our SU maDMP template.
Currently, in both cases this extraction and retrieval of the funding application information via the SweCris API is still made manually, but could naturally be automated through a processing script.
Conclusions:

- retrieval of information from an external source used for *quality control of metadata* - original start date for the project in DMP erroneous, *corrected by means of SweCris project application description*.
- enrich and *add value* to metadata by *prioritizing the most trusted source* for information types and metadata entries at each collection event.
- ultimate goal: *identify* at each stage in the workflow, *at each point in time*, *the most trusted data source*, offering the *highest information quality*

# Collection Events and Processing of Information 3:

- **Collection Events**: here *harvesting, extraction* or *other retrieval, transformation, integration* and *packaging* of (digital) research information

- How? When? By whom? From where?– > Collection Tools?

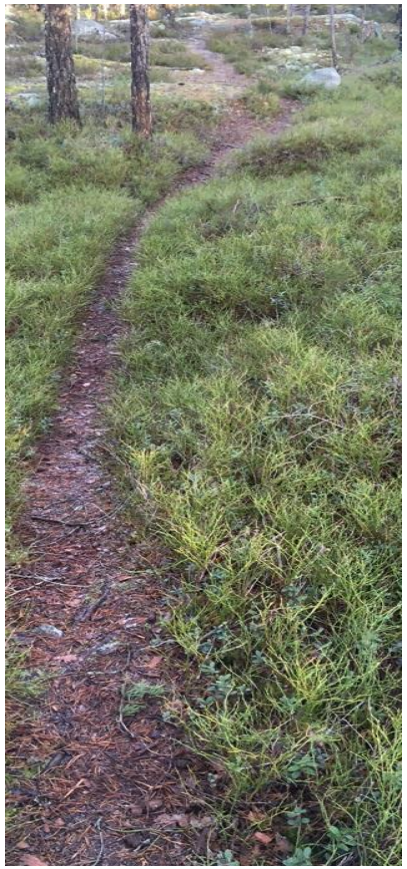
Mass retrieval or more fine-calibrated  capture?

# Collection Events and Processing of Information 4:

Archives for long-term preservation: distinguish between ***creators*** (authors, researchers, producers) of information items Vs. "***collectors***" (processors, retrievers, transformers)

Entrusting act of *collection* / ingathering of research information items (documents) to the creators i.e. the researchers:

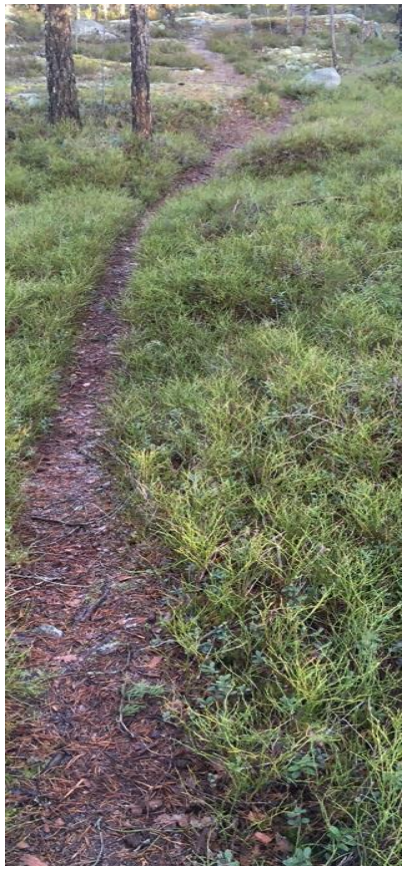
- ignores the stakeholder need of easing the administrative burden on researchers (1)
- likely necessitates increased direct communication from administrative staff with researchers in order to ask for deposit / delivery of documents, completion of web-forms etc. thus, also detrimental to stakeholder need (2)
- potential loss of *control* (automation possibilities, data sources, time) and *quality* of information, thus only passively received by administrative staff from researchers.



## Collection Events and Processing of Information 4:

To *maximize* the degree of potential *automation*, it is important that *information objects* are accessed and *retrieved in a timely manner*, allowing for *metadata enrichment* and *transformation to* a valid ***SIP*** (Submission Information Package), *before ingest* to a digital archive for long-term preservation and further transformation to an AIP, Archival Information Package. (In a recent survey from the Open Preservation Foundation, OPF, pre-ingest was singled out as the area of digital preservation that definitely needs more tooling.)

*Iteration* possible: retrieval of certain information elements sometimes need to be iterated on multiple occasions (e.g. DMPs at different stages, *initial* version & *final* version)



# Packaging of Information

Benefits from keeping associated research information and datasets stored in the same packages:

1) *avoid duplication of effort,* especially when retrieval needs to be iterated; at each collelction event the corresponding package (pre-SIP) will offer an overview of what kind of information has already been collected/retrieved.

2) give an *overview* of available *file formats and metadata* standards for the whole information package.

3) for future dissemination (DIPs) of any or all research information or datasets in a project it will be much easier to search out the desired information in the same place; particularly  if you also keep copies of e.g. PDFs in plain text formats (such as .txt, .csv, .tsv, .md, .html, .json, .xml), thereby enabling more efficient full text search in larger data archives or file libraries, e.g. by means of BASH *grep* commands.

4) enhance metadata and data quality control of separate files, by means of comparison with associated information elements.

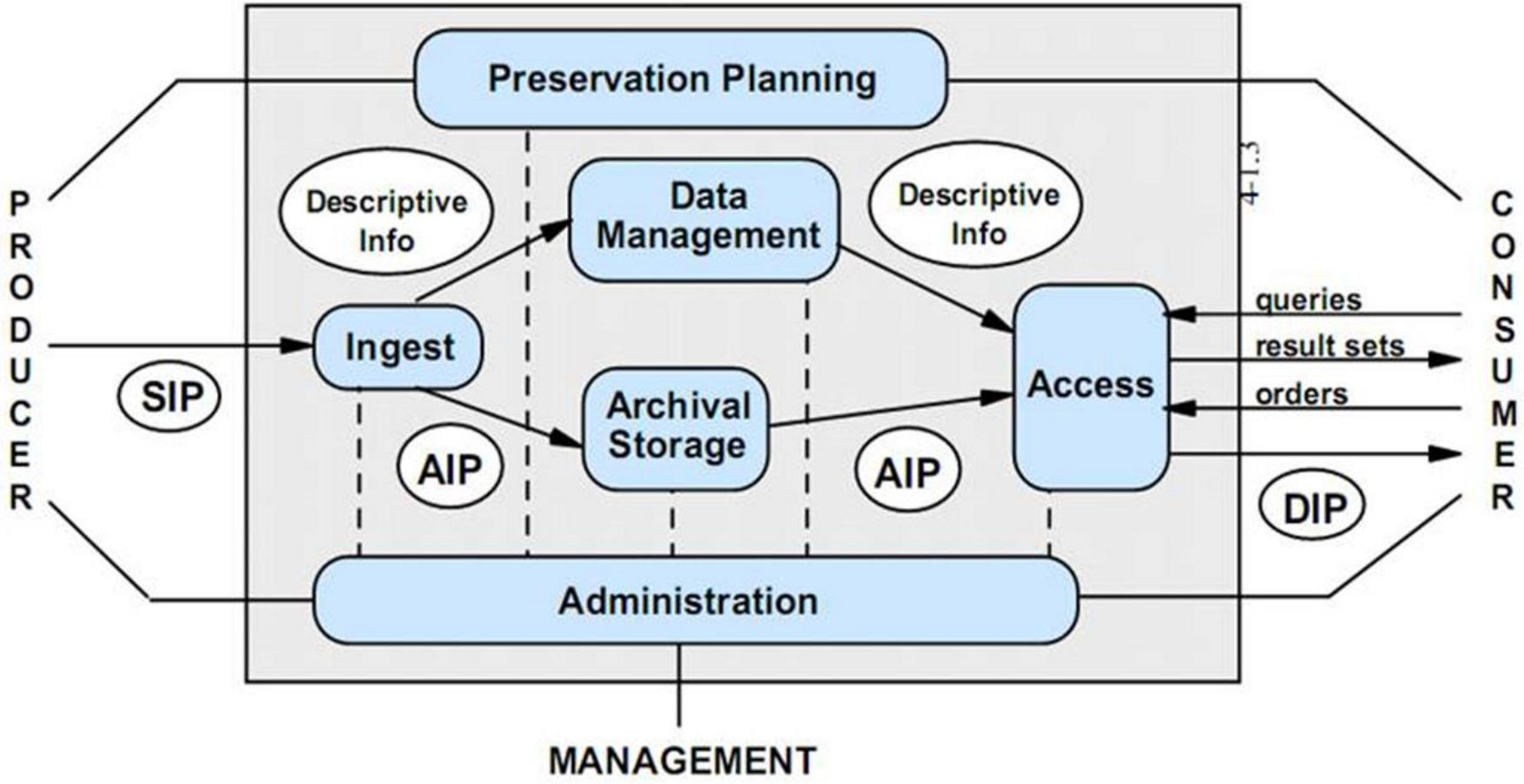
# Prepare for Archiving and Long-Term Preservation:
## Provenance information

Information sources, path (pipelines, workflow), processing (enrichment, tools, templates, transformation, valida
versioning, schemas)

Although Swedish archival regulations say nothing about *how* the information is created, collected or modified, there is a clear recommendation that the organization of records and documents should reflect the *process* through which information was created, collected or modified, and *facilitate disseminating* information that emerged through the *same process together*, possibly in a joint package - a call for *provenance* information.

**Collection events**, iterated or not, representing provenance information could subsequently be recorded, as PREMIS or PROV metadata, in an AIP.

# OAIS model




Consultative Committee for Space Data Systems. (2012). Reference Model for an **Open Archival Information System** (**OAIS**). https://public.ccsds.org/Pubs/650x0m2.pdf

Stockholm University

# Research Data information from DMP to ”Combine Harvester”

A DMP can also contain information about *repositories* where *datasets* created in a research project will be deposited, sometimes already the *identifiers* (DOIs or other Persistent Identifiers, PIDs ) of these datasets, to be extracted from maDMPs (automatically), for possible harvest and transformation to archival format of metadata from repositories and potential necessary file format conversion of data files.

II: Documentation and data quality (3 / 3)

**Q1: How will metadata be created for your dataset? If by *use of a repository* (*recommended*), please specify which, either from the given options, or - if Other - by giving a link(s) / URL(s) [if multiple separated by commas] as Additional Information below. Please, *do not write whole texts here* with line or paragraph breaks, *as this prevents automatic processing and evaluation of the DMP!***

- [x] 1. Dataverse/StockholmUniversityLibrary
- [ ] 2. su.Figshare.com
- [ ] 3. SND
- [ ] 4. Zenodo/StockholmUniversityLibrary
- [ ] 5. GitHub
- [x] 6. README-file
- [ ] 7. Bolin Climate Research DB
- [x] 8. Other: https://...
- [ ] 9. Manually (not recommended)

**Q2: Dataset ID: at this initial planning stage, please find one main identifier (e.g. a DOI, Handle, URL, ... ) for the entire dataset(s) in the project where possible, even if it comprises several data files of different types.**

https://doi.org/10.7910/DVN/MGZBAL



*Prepare* for Archive: **harvest & transform** from **su.figshare.com**, ***Zenodo*** et al. to **Swedish National Archive Common Specification** for information **Packages (*SIP, FGS* - METS)**

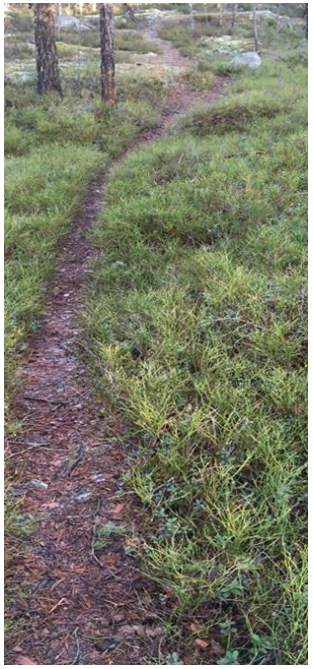
Stockholm University

# The ”Combine Harvester” at Stockholm University

- harvests datafiles and transforms original metadata from several sources / repositories
- is ”semi-automatic” – it takes a human driver!
- *prepares* for long-term preservation in local archive by enriching and transforming original metadata to accord with Swedish National Archives *FGS* – SIP
- is using only open (non-proprietary) file formats for scripts and in processing (json, xml, xslt, xquery)
- is implemented locally using *Oxygen* (licensed software), Git *BASH* for Windows (free), *BaseX* (free)


Stockholm University

# Conclusions



We have seen some interesting attempts at using maDMPs as tools for the automated integration of research information, thereby serving to a certain extent both the stakeholder needs that we identified at the outset:

1) easing the administrative burden on researchers

2) answering the demand for ”scientometric” control and improved tools for curation and review to be used by administrative staff

However, few appear to have yet leveraged these integration models to serve also the *needs of archives* for *long-term preservation* of research information as discrete, self-sufficient packages with extensive *provenance* documentation.

At Stockholm University, we still have a long journey ahead of us to develop efficient workflows covering the whole research cycle from grant applications to long-term preservation of integrated research information packages in a local OAIS-compliant digital archive. We have only taken the first step on that path!

Stockholm University

**Merci pour votre attention!**

**Thanks for your attention!**

Stockholm University

joakim.philipson@su.se