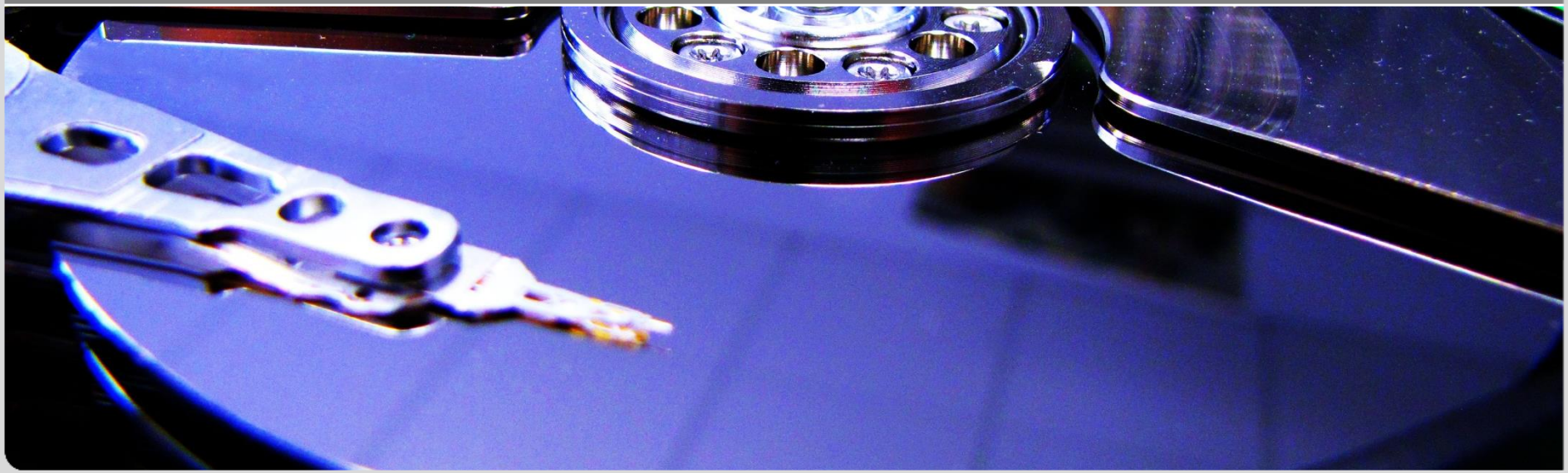
Image Data, RDA and Practical Policies

Rainer Stotzka and many others …

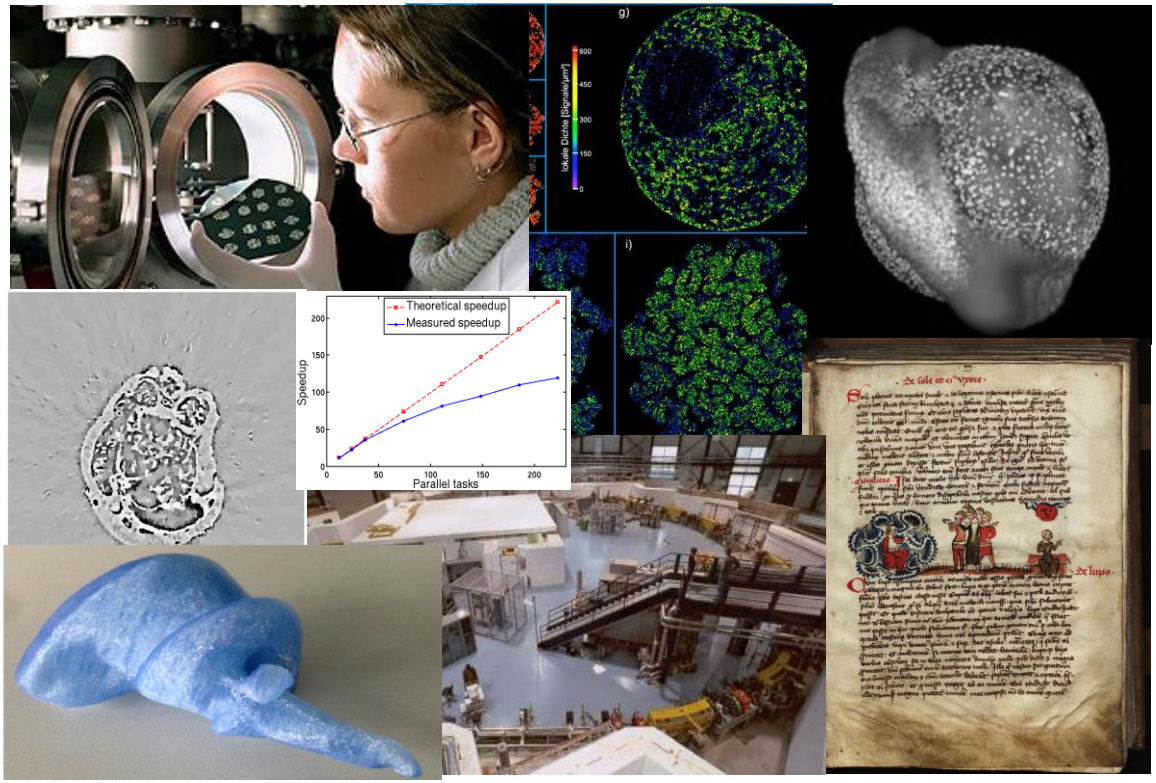
Institute for Data Processing and Electronics



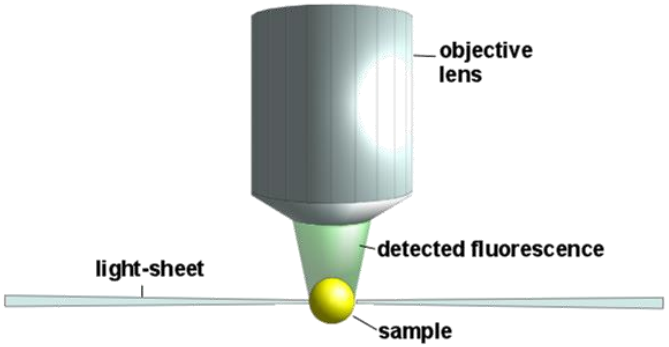
Data Life Cycle Lab “Key Technologies”

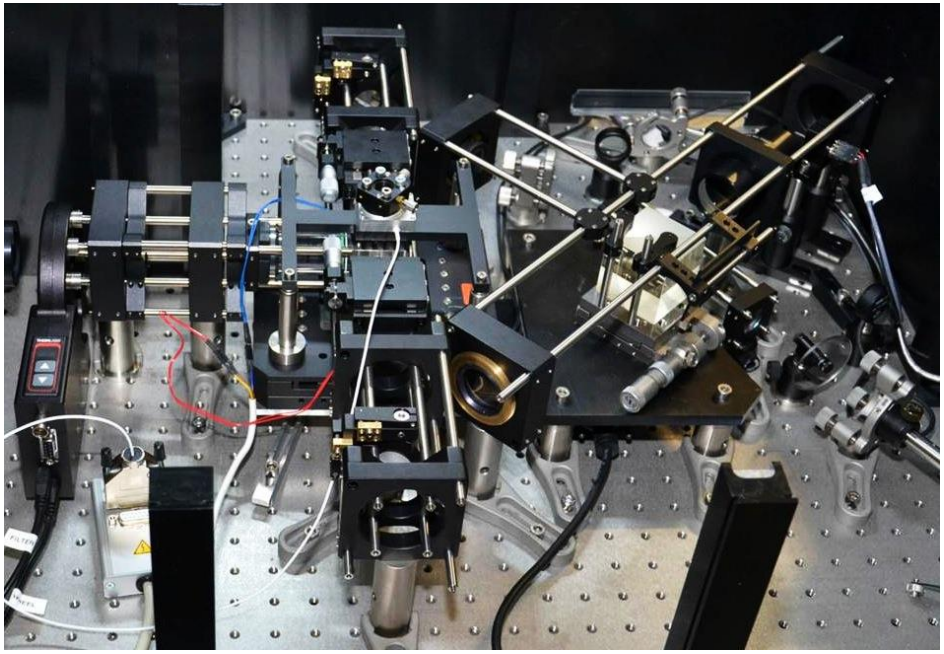
Mostly scientific imaging projects with common needs in

- Data and **metadata** management
- High throughput **analysis** in near real time
- Data archiving and **sharing**
- High performance data **ingest**
- Automatic metadata extraction
- Visualization



Light Sheet Microscopy





Novel microscope to image ***living Zebrafish embryos***

- High 3D resolution
- Extreme short data acquisition time

Growth of an embryo with in 16 h → ***16 TB raw data***

Acknowledgements:
Andrey Kobitskiy, G. Ulrich Nienhaus
Jens C. Otte, Masanari Takamiya, Uwe Strähle
Johannes Stegmaier, Ralf Mikut
Francesca Rindone, Volker Hartmann, Thomas Jejkal
Achim Streit, Christopher Jung, Jos van Wezel

Institute of Applied Physics
Institute of Toxicology and Genetics
Institute for Applied Computer Science
Institute for Data Processing and Electronics
Steinbuch Centre for Computing

Image Data Workflow

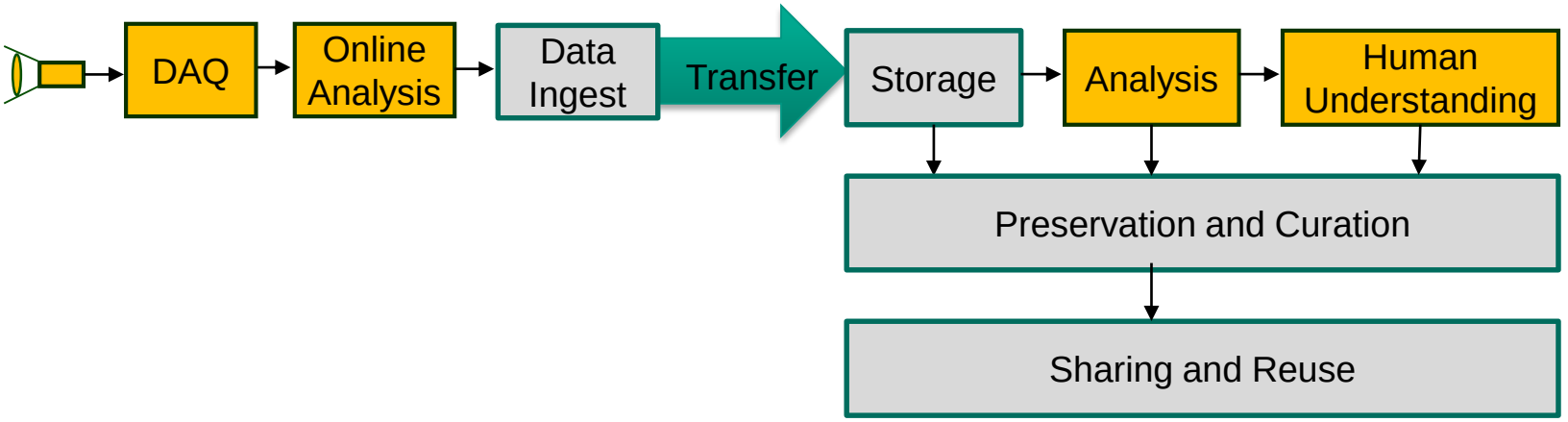
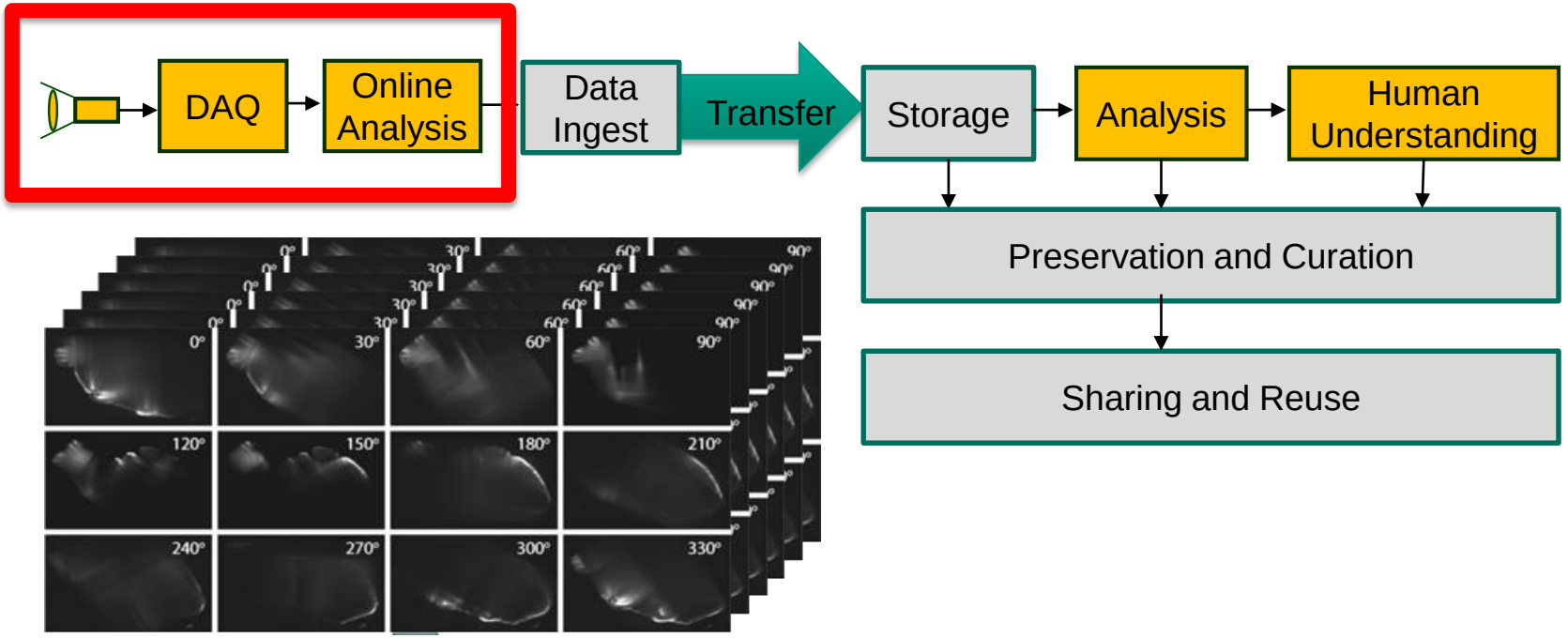
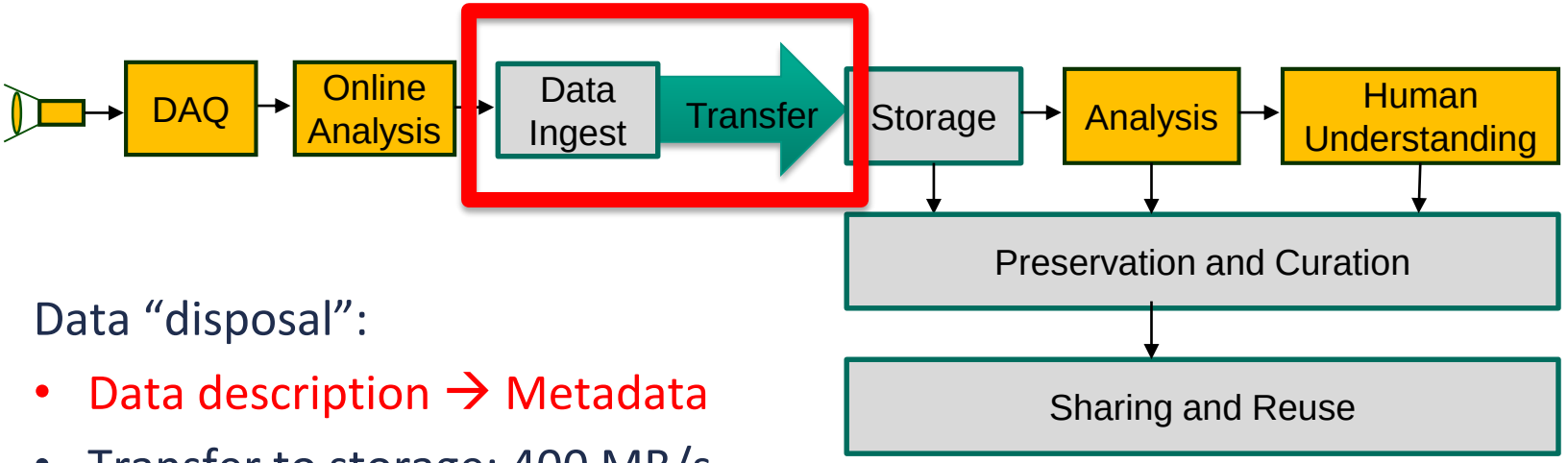


Image Data Production



- Experiment design → Big Data: variety
- Data acquisition and quality control
- 200 MB/s → 16 TB/d

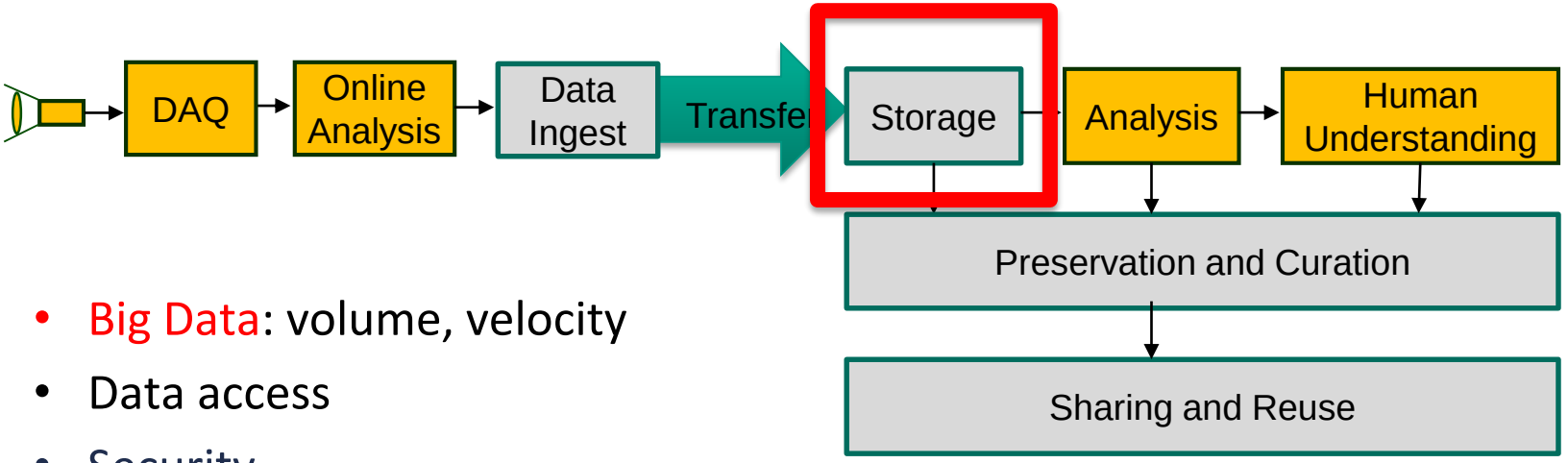
Image Data Transfer



Data “disposal”:

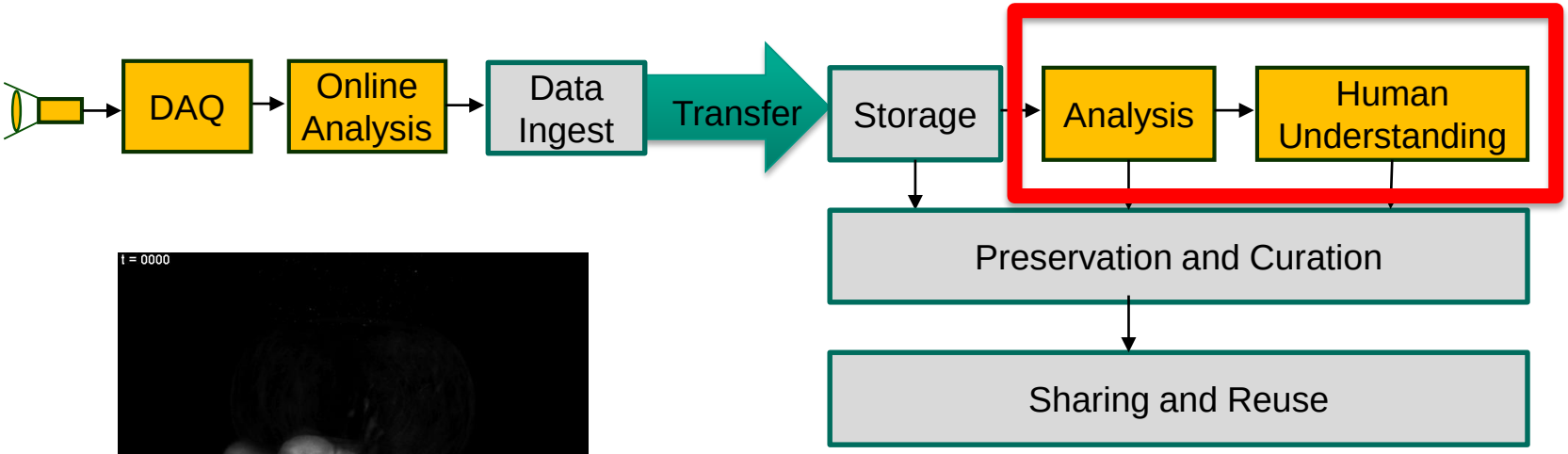
- Data description → Metadata
- Transfer to storage: 400 MB/s

Image Data Storage



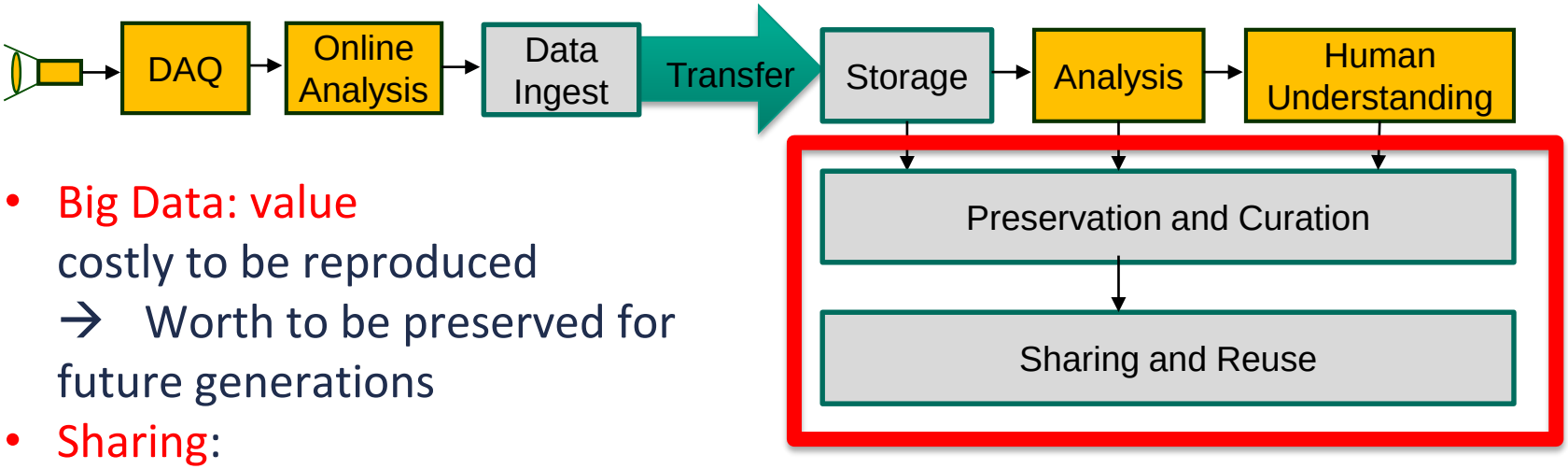
- Big Data: volume, velocity
- Data access
- Security

Image Data Analysis



- Big Data: value

Image Data



- Big Data: value
  costly to be reproduced
  → Worth to be preserved for future generations
- Sharing:
  → Reproducible science
- Re-use:
  → New (interdisciplinary) scientific outcomes
- Big Data: veracity
  → Trust, well maintained

→ Open Data

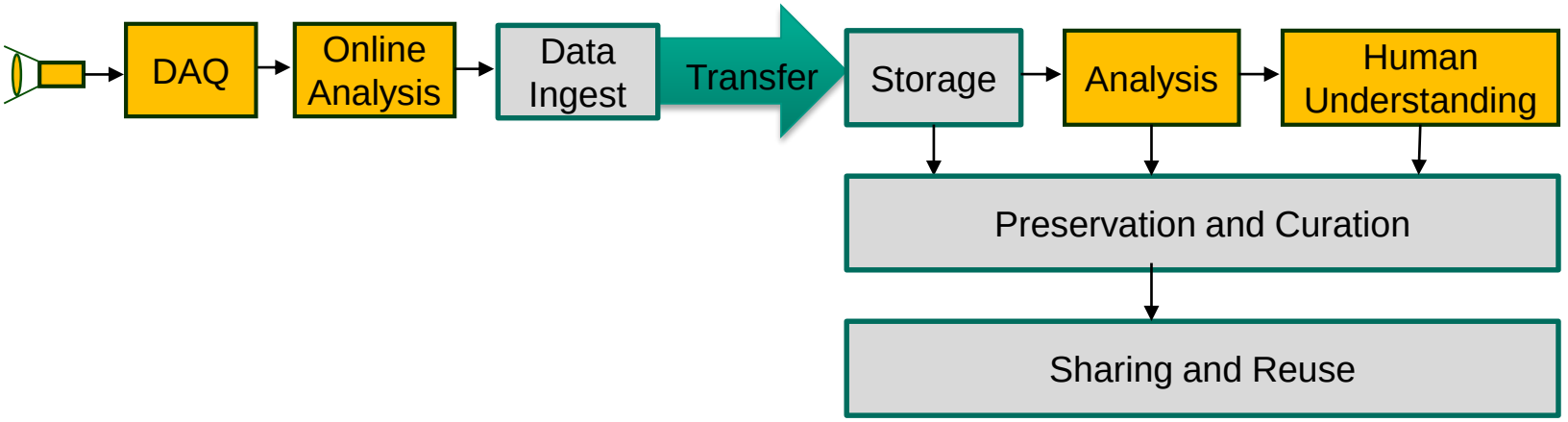
Deutsche Forschungsgemeinschaft DFG

THE FRAMEWORK PROGRAMME FOR RESEARCH AND INNOVATION HORIZON 2020

Bundesministerium für Bildung und Forschung

RDA Research Data Sharing without barriers
RESEARCH DATA ALLIANCE

Image Workflow

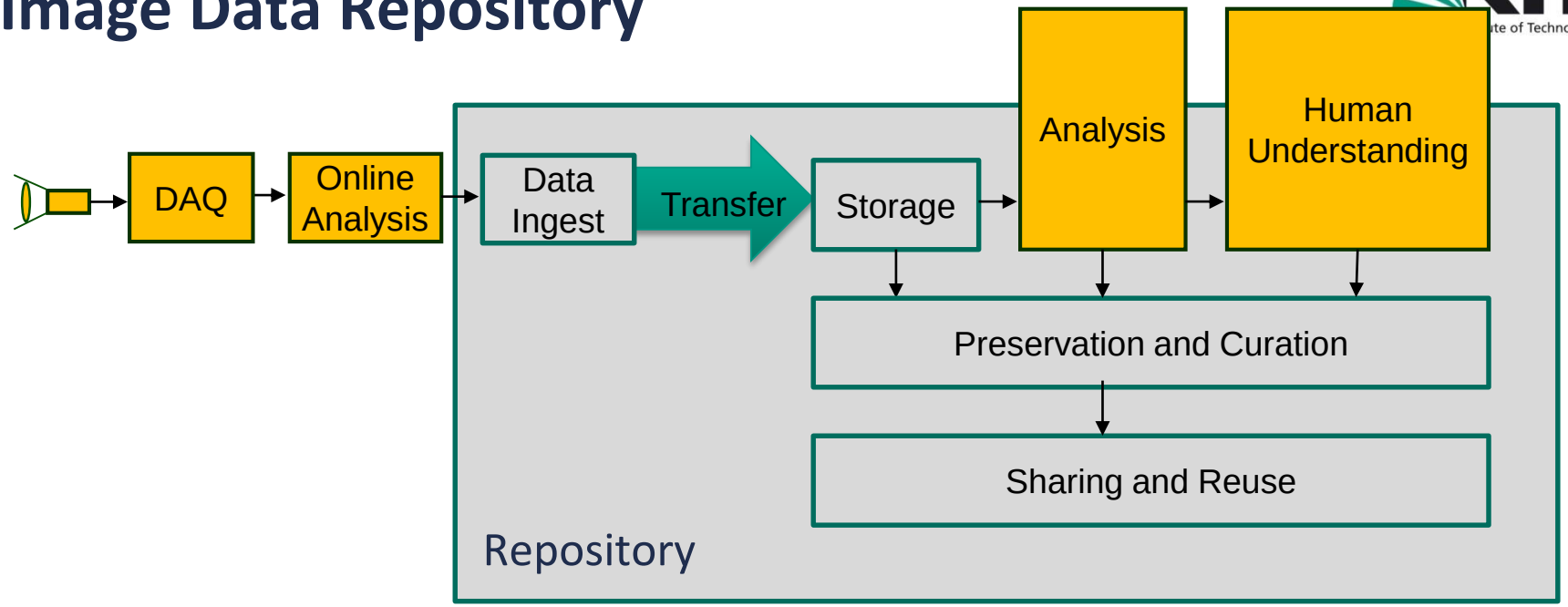


Problems:

HELP: Where is my data?

HOW to manage, preserve, and to curate the data?

Image Data Repository



Repository

Repository:

Managed location/destination/directory/bucket where digital data objects are

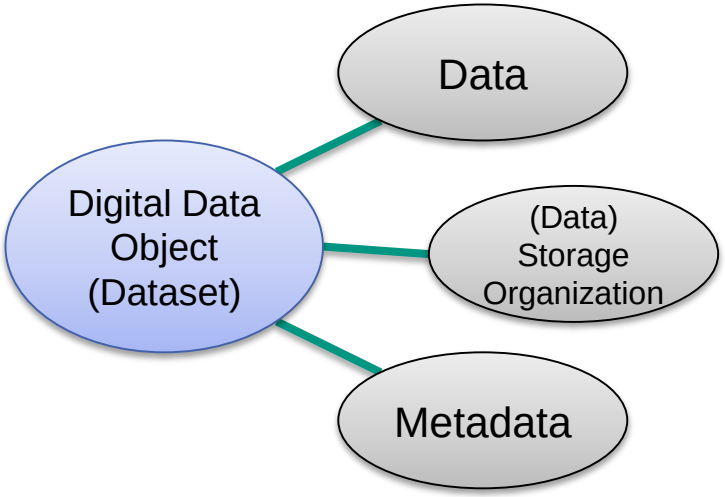
- Registered
- Permanently stored
- Made accessible and retrievable
- Curated

E.g.: libraries & museums

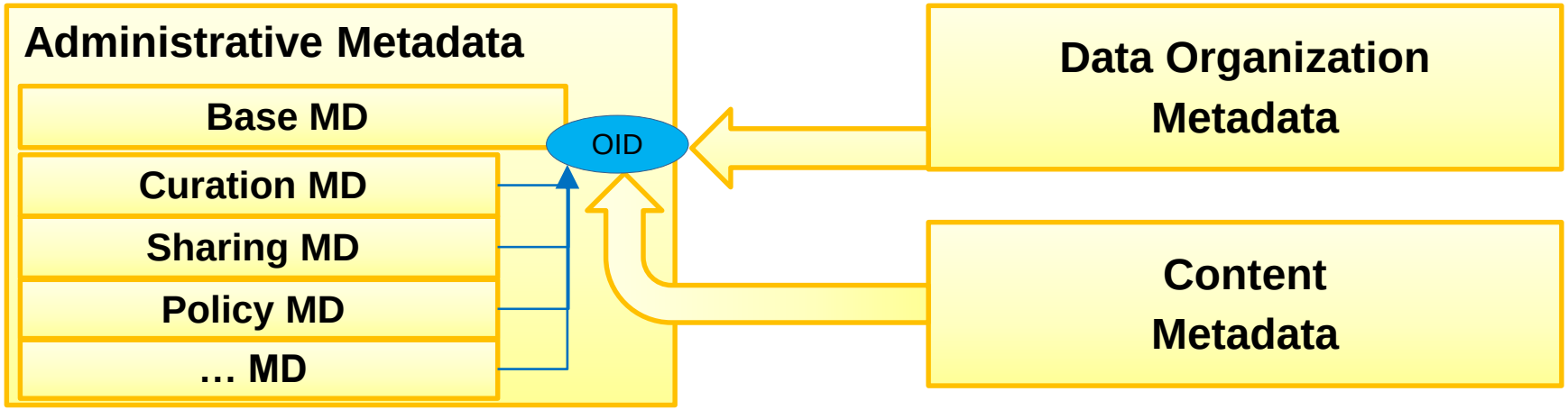
Digital data object (dataset):

Consists of

- Data
- Description for re-use



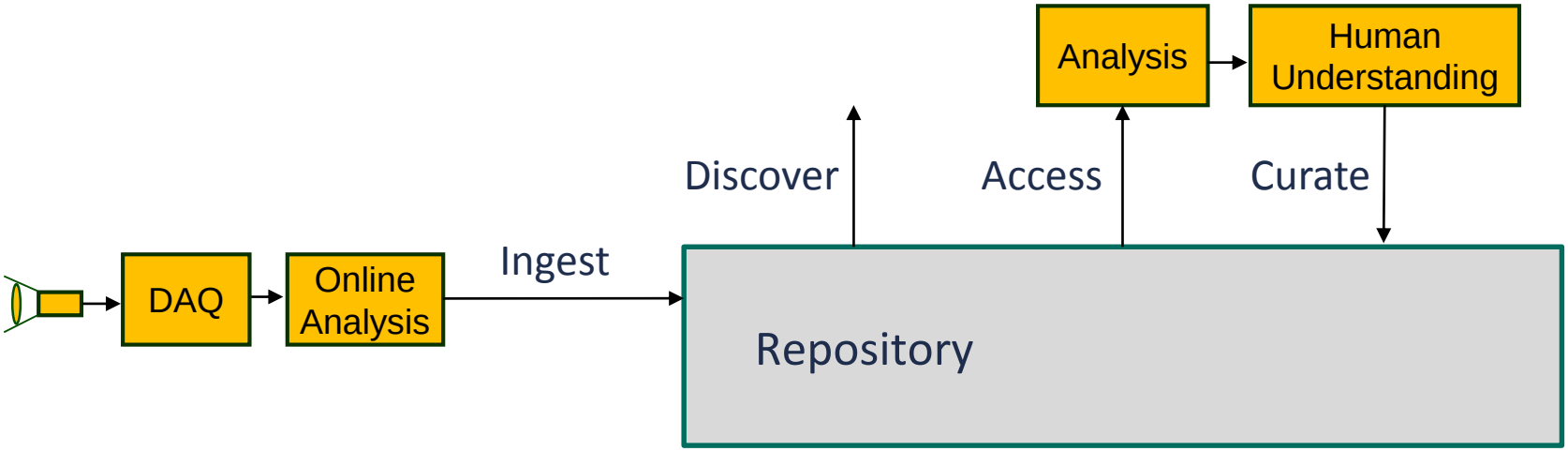
Repository System: KIT Data Manager



Repository System:

- Optimized for large scale and heterogeneous research data
- Internals are based on metadata
- Users & data curators don’t see the complexity:
- Visible:
  - Content metadata
  - Discovery tool: search
  - Representation of data

Repository Design



HELP: Where is my data?

HOW to manage, preserve, and to curate the data?

HOW to design the repository?

- Need for content metadata
- Need for rules for automation → practical policies

RDA Working Group: Practical Policies

Chairs: Reagan Moore, Rainer Stotzka

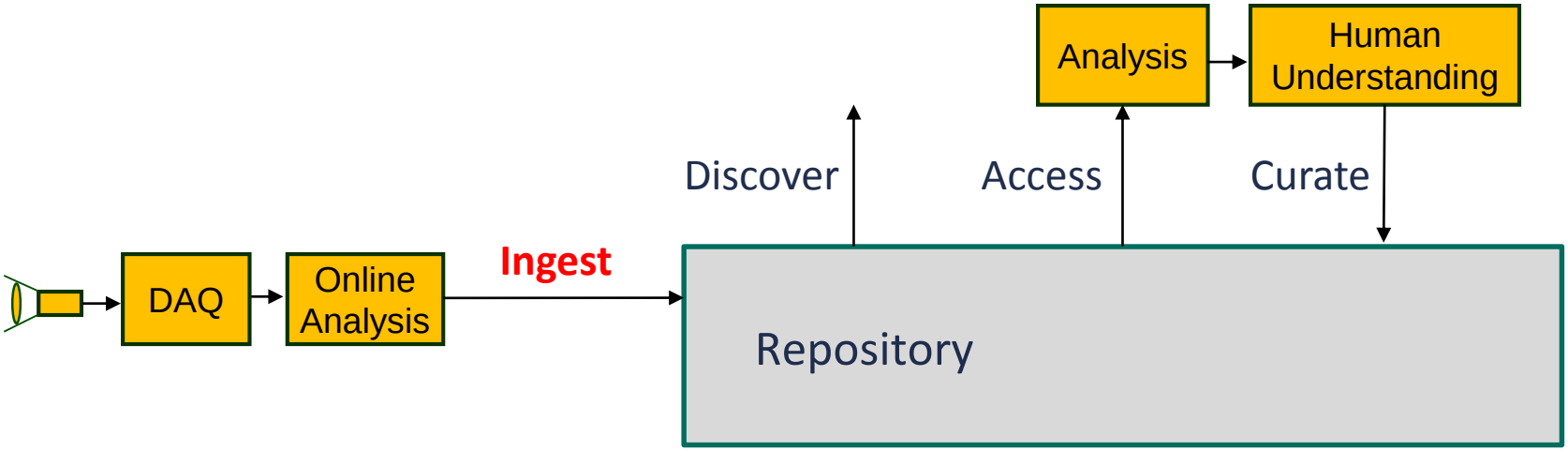
Catalog with 11 important policy areas:

- Contextual metadata extraction
- Data access control
- Data backup
- Data format control
- Data retention
- Disposition
- Integrity (including replication)
- Notification
- Restricted searching
- Storage cost reports
- Use agreements

Example: Contextual Metadata Extraction

Scenario:

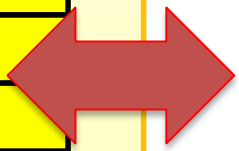
- Extract metadata from an associated document, e.g. DICOM, OME, ...
- Extract metadata from …



Example: Contextual Metadata Extraction

Policy MD

enable Contextual metadata extraction	
On file	File_name
On digital data object	OID
On user	User_ID
Extract metadata	Attribute_name
	Attribute_value
	…
enable Data access control	
enable Data backup	
enable Data format control	
enable Data retention	
enable Disposition	
+++	
+++	

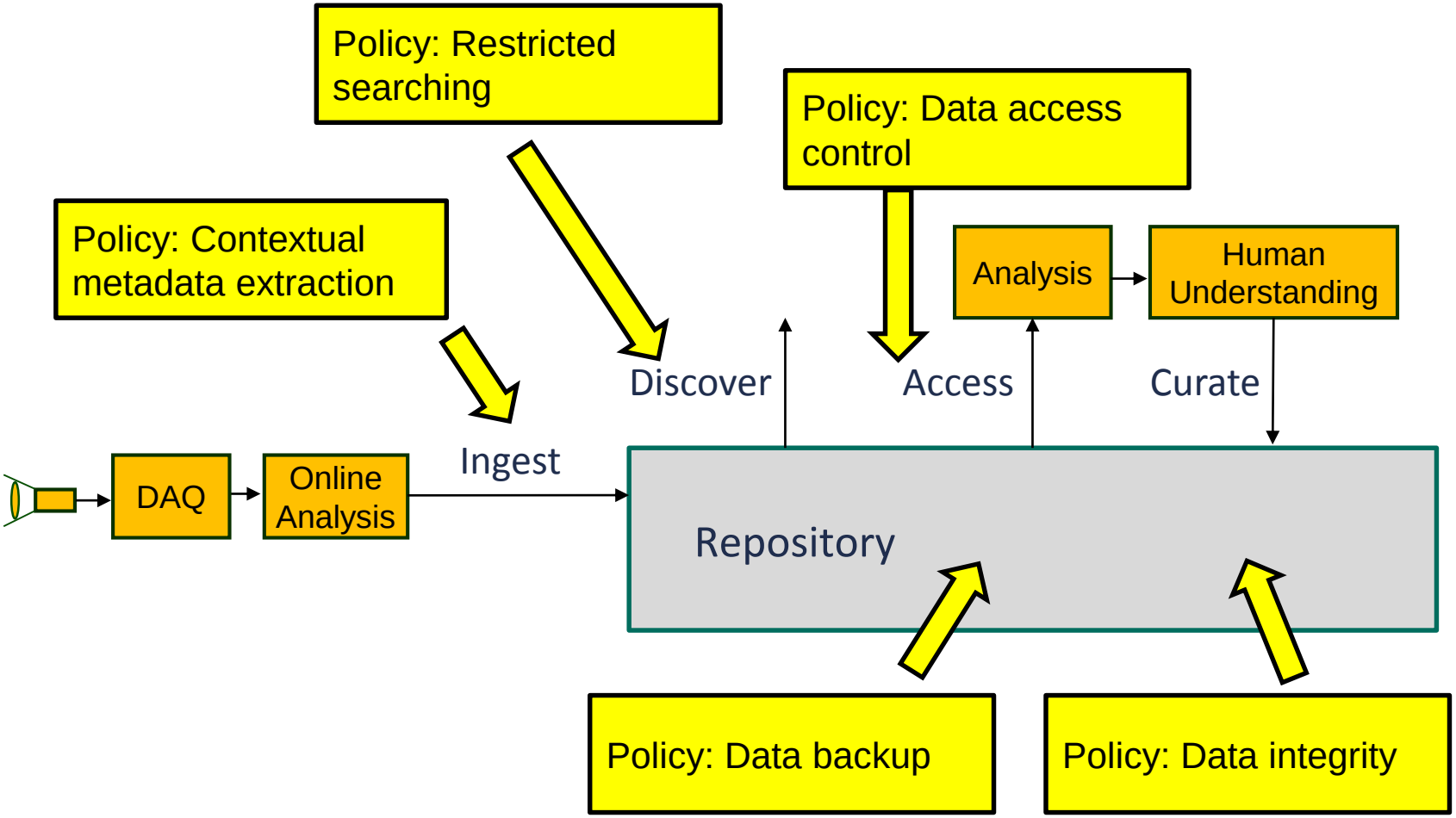


Expert knowledge required

→ Data Life Cycle Lab

Minimize the effort of manually recording metadata !

Policy Examples



Interaction with RDA

Important groups for imaging repositories

- WG Practical Policies
- WG Data Foundation & Terminology
- PID groups
- Metadata groups
- IG Data Fabric
- Repository groups
- Domain related groups

Typical PhD researcher

- Cooperates with domain partners
- Solves their data problems
- Specific research topic

- Monitors the activities of groups

- Transfer to the domain partners

- If funding available → RDA plenaries
- Inhibition threshold: world experts
- Active in RDA groups

- Results:
  - Motivation booster
  - Networking
  - Future “State of the art”

Conclusions

Goal: ***Better and easier management of research data***

→ Sustain the value

→ Enable sharing, reproducibility and re-use

→ Enable “better” research

RDA has real value and impact

- RDA is alive and growing, efficiency needs further improvement
  → Technical Advisory Board, Organizational Advisory Board, Council, Secretariat, and individuals
- Plenaries (group sessions) produce results:
  2 plenaries per year, low registration fee
- Need to invest: contributions → gain results
- RDA is producing internationally connected data experts