Data Life Cycle: Introduction, Definitions and Considerations

EUDAT, Sept. 25, 2014

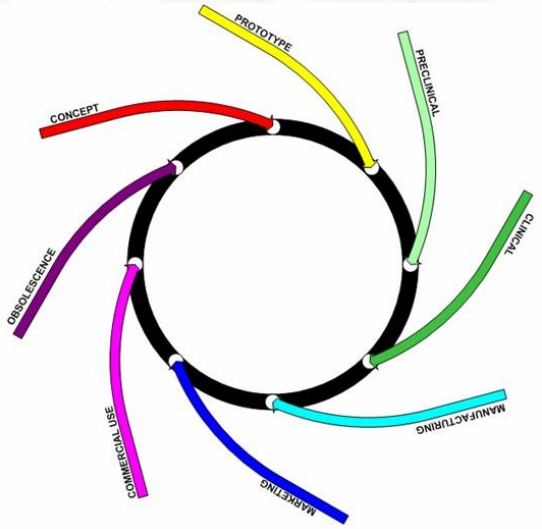
Prof. Peter Fox (pfox@cs.rpi.edu, @taswegian, #twcrpi)
Tetherless World Constellation Chair, Earth and Environmental Science/ Computer Science/ Cognitive Science/ IT and Web Science)
Rensselaer Polytechnic Institute, Troy, NY USA

Definitions

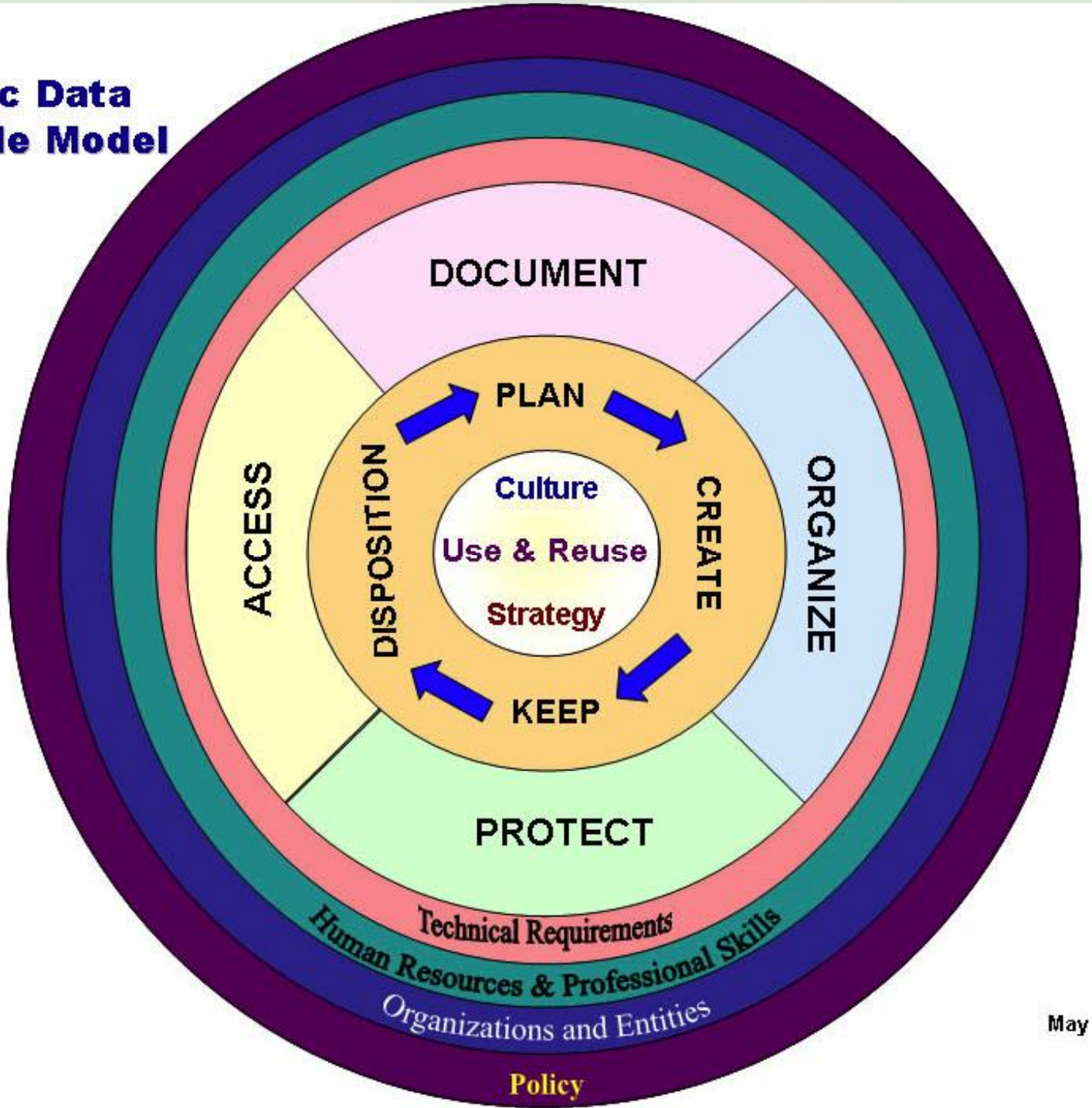
- Data (management) life-cycle broad elements -
  - Acquisition: Process of recording or generating a concrete artefact from the concept (see transduction)
  - Curation: The activity of managing the use of data from its point of creation to ensure it is available for discovery and re-use in the future
  - Preservation: Process of retaining usability of data in some source form for intended and unintended use
- Stewardship: Process of maintaining integrity for acquisition, curation, preservation
- BUT…

Definitions ctd.

- (Data) Management: Process of arranging for discovery, access and use of data, information and all related elements. Also oversees or effects control of processes for acquisition, curation, preservation and stewardship. Involves fiscal and intellectual responsibility.

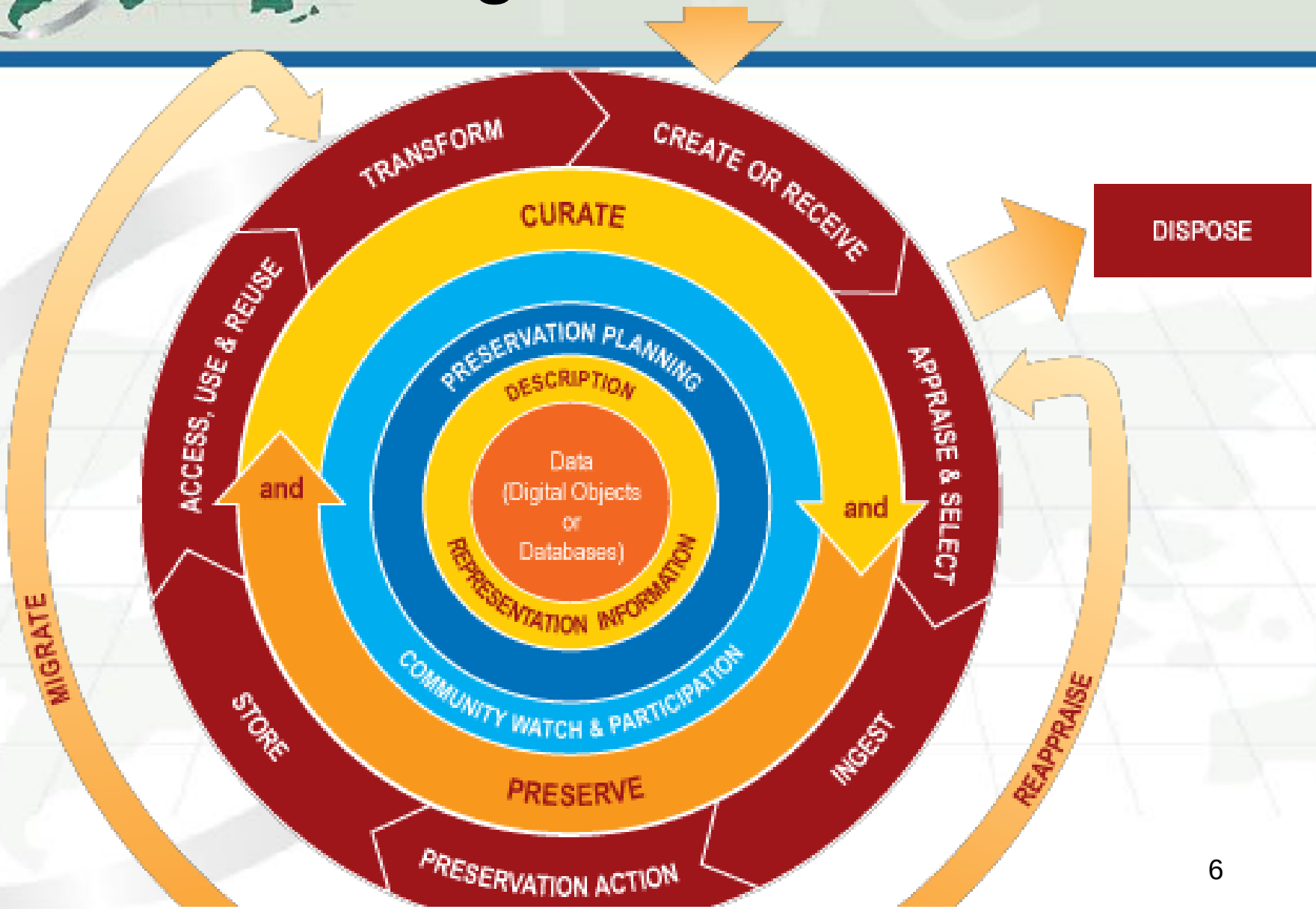


IWGDD Scientific Data Life Cycle Model

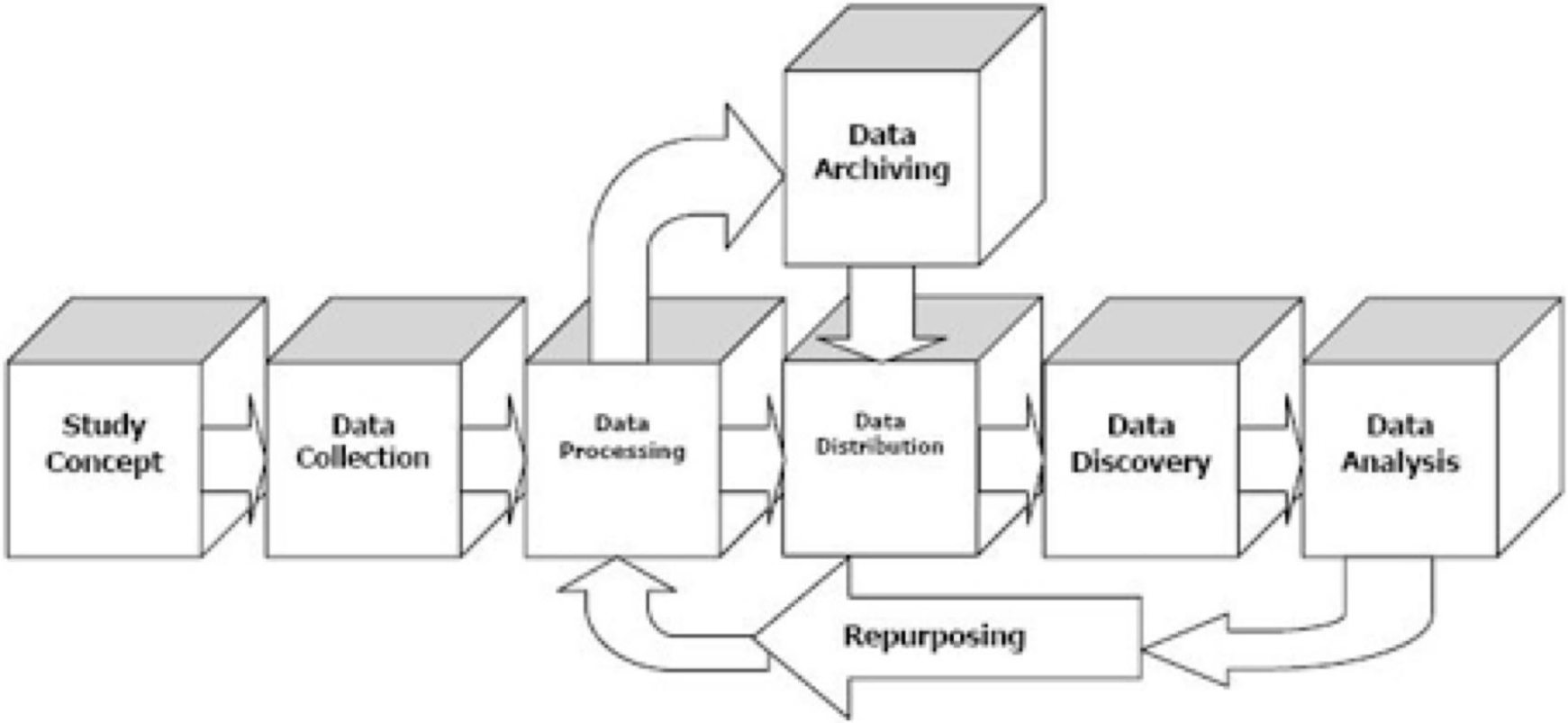


May 2008

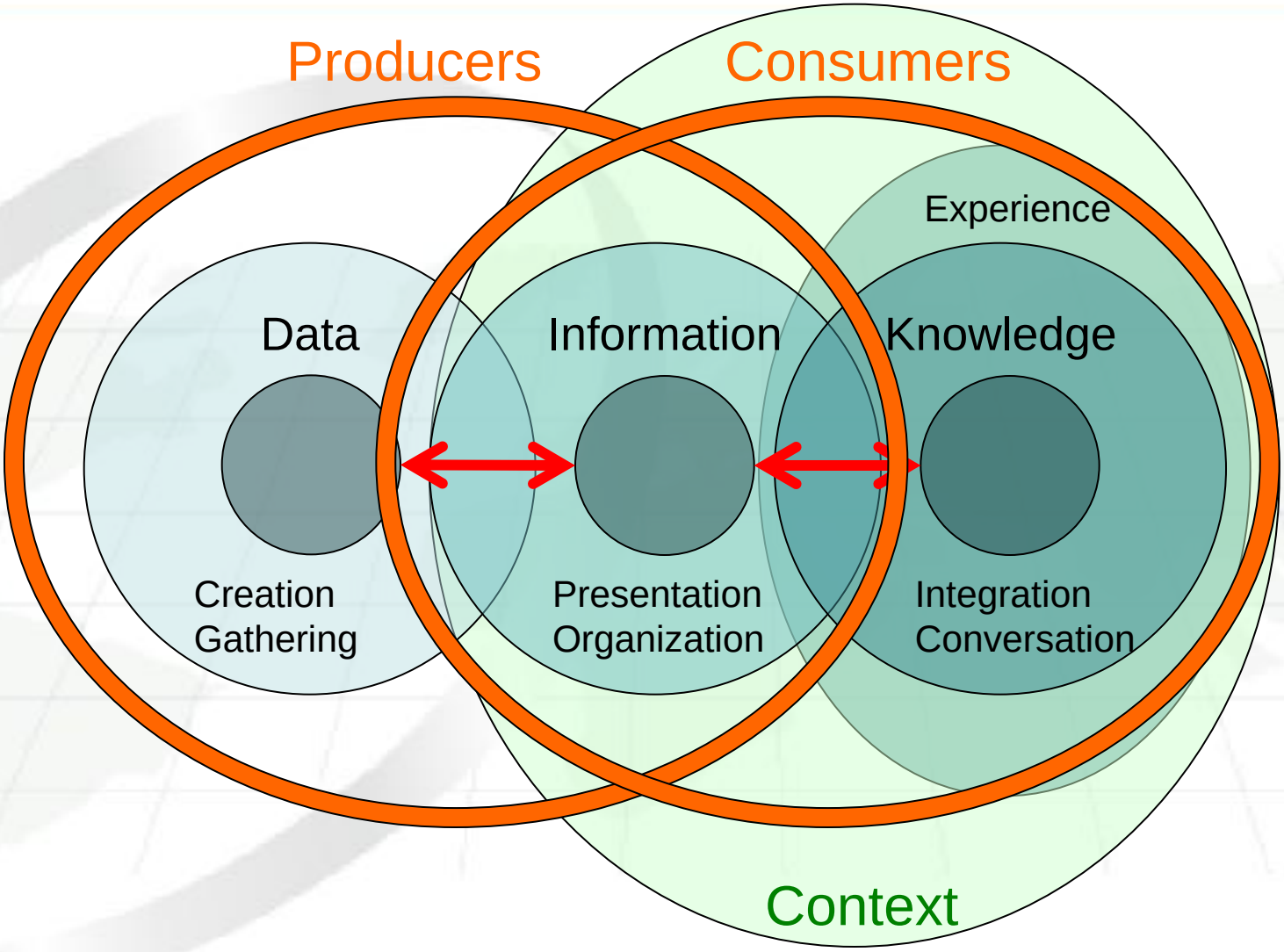
Digital Curation Centre



MIT DDI Alliance Life Cycle

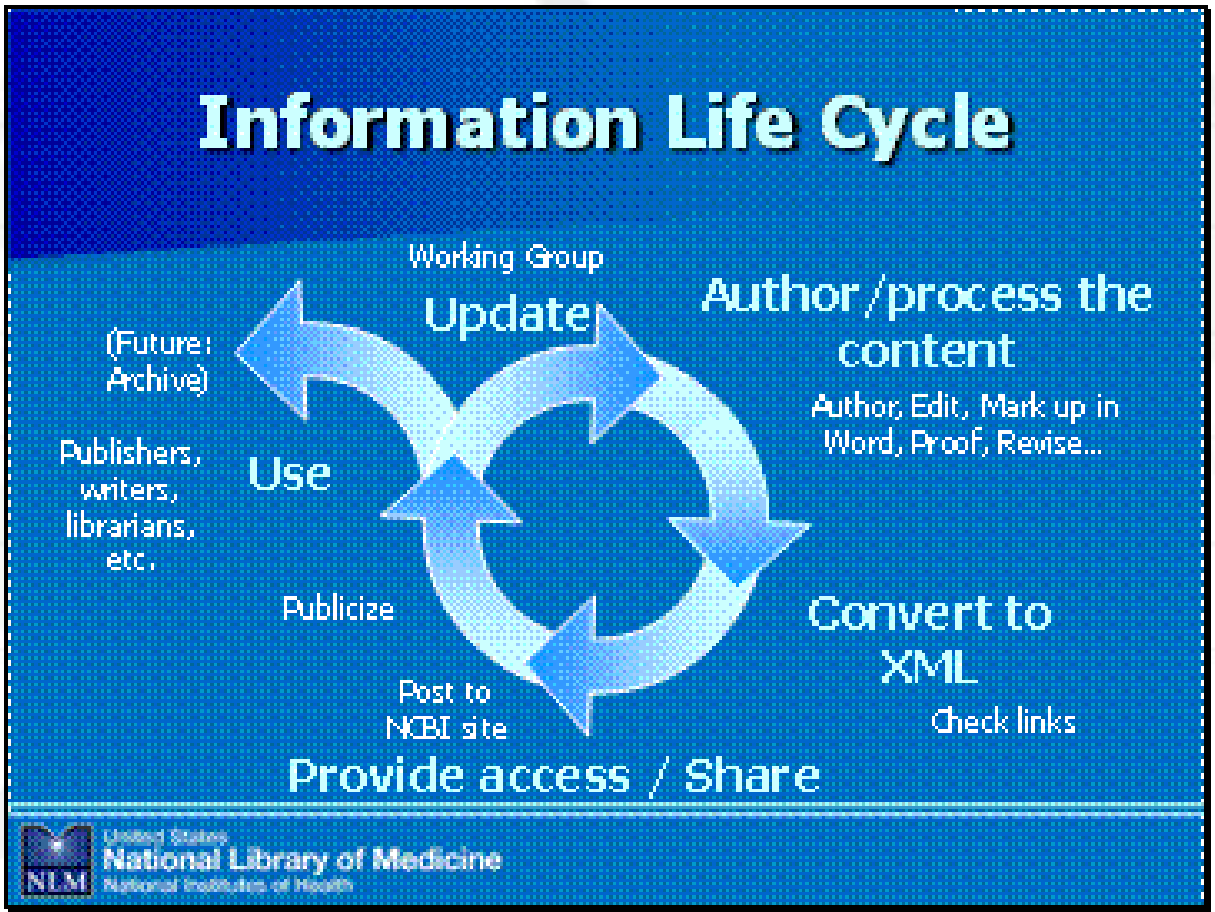


Data-Information-Knowledge Ecosystem



Data Life Cycle embedded in Research Life Cycle

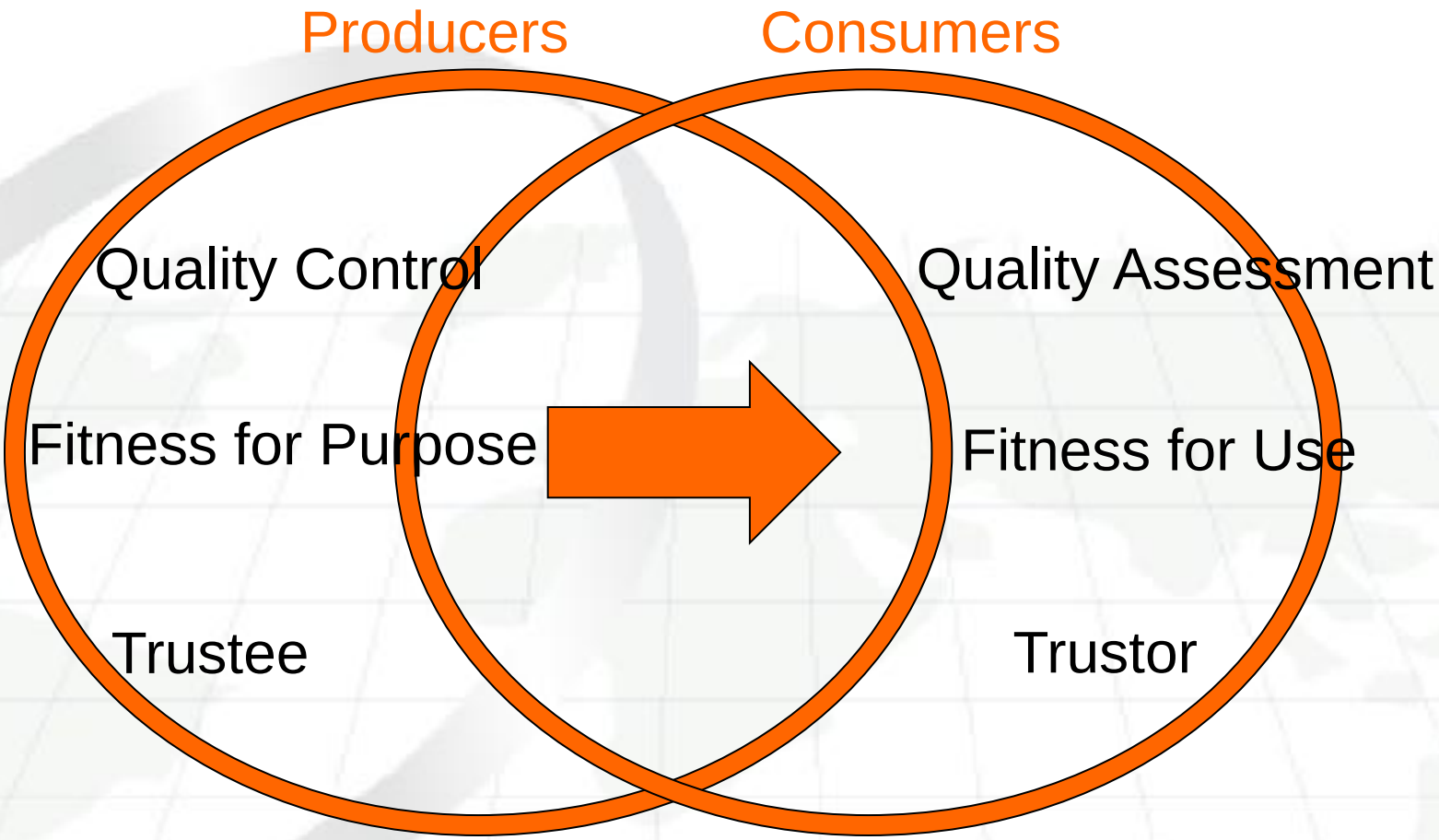
- Information Life Cycle
- Knowledge Life Cycle



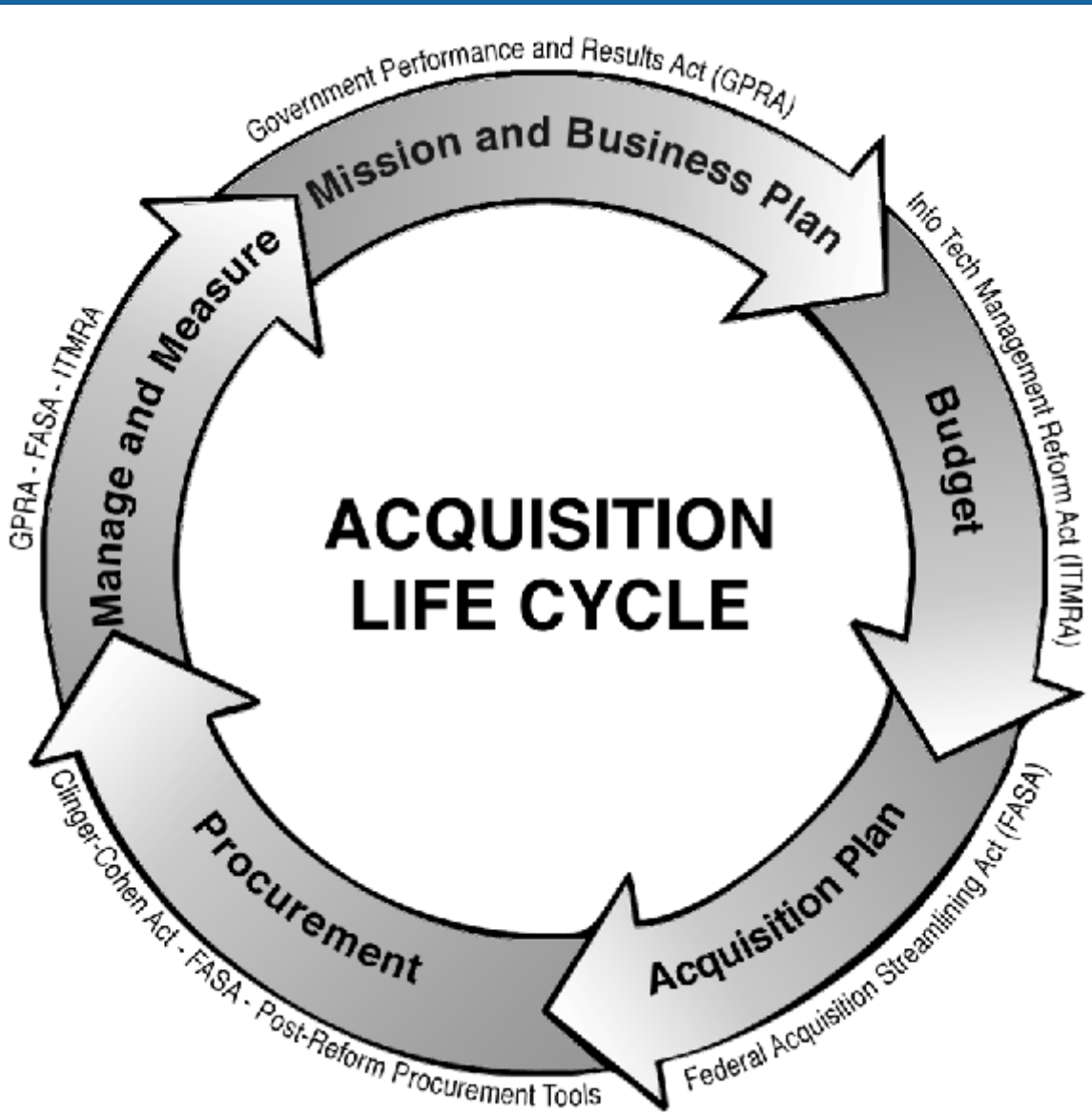
Type of knowledge created

- Tacit (created and stored informally):
  - Human memory
  - Localize, e.g. hard drive of the computer
  - Movement of tacit information into a formalized structure
- Explicit (created and sorted formally):
  - Network shared
  - Network Web site/intranet
  - Informal knowledge-management system
  - Document-management system
  - Formal KM system

Curation…


Producers
Consumers
Quality Control
Fitness for Purpose
Trustee
Quality Assessment
Fitness for Use
Trustor

Acquisition meets science

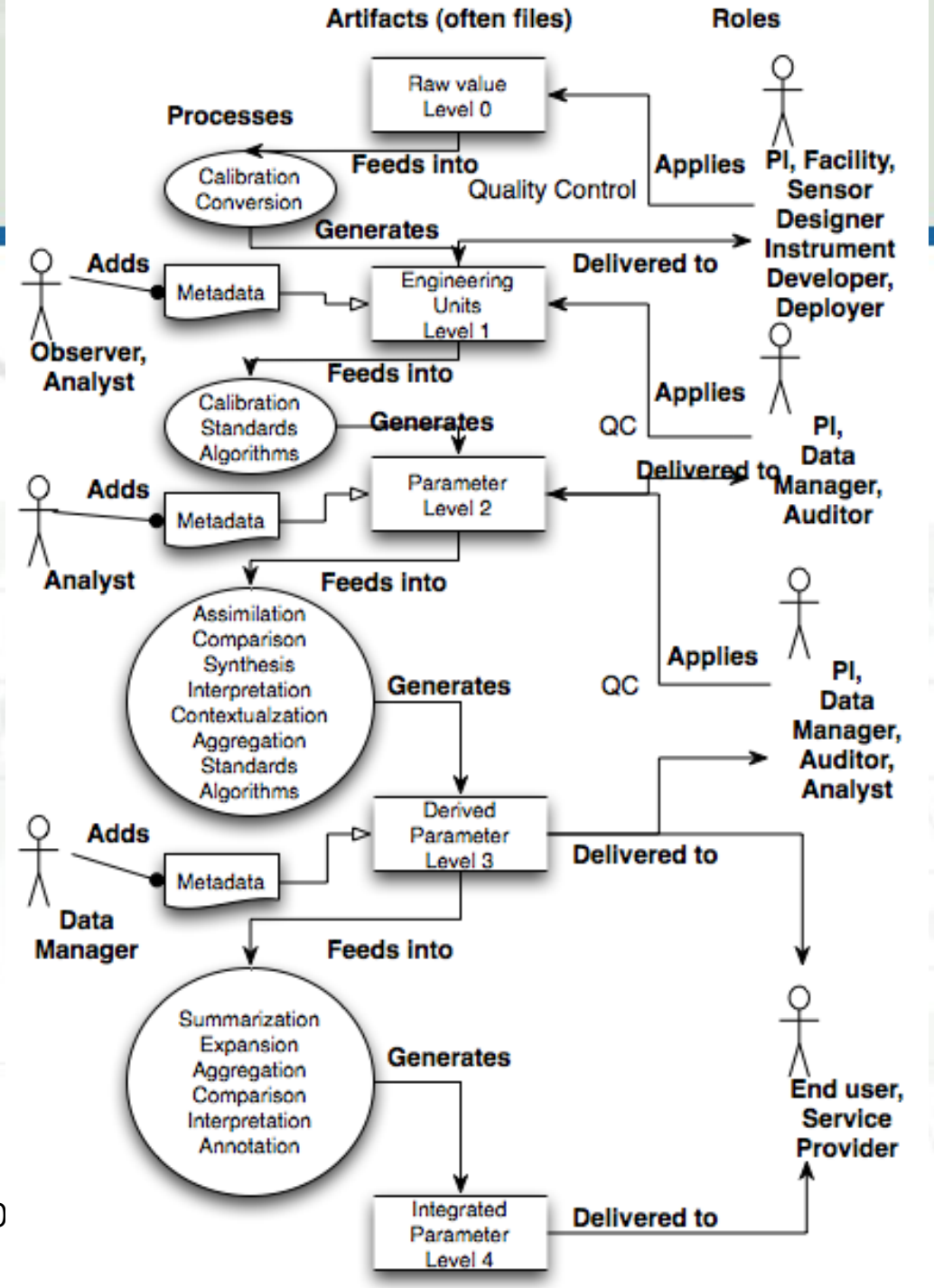


Enough with the unlabelled ARROWS and INTERFACES!

Workflows and Life Cycles

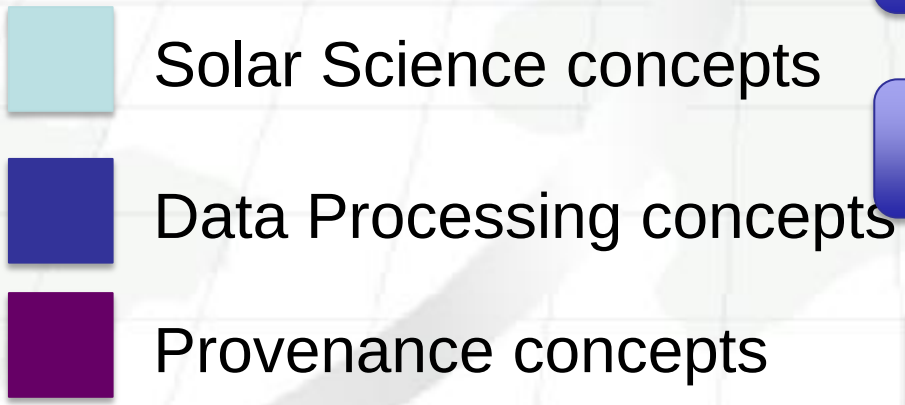
- Yann covered much of this but the question remains:
  - Why is the life cycle not “just” a workflow?
  - Well it is: sort of…..
- Workflows give internal “provenance”
- To capture embedding of data life cycle in research life cycle + external provenance

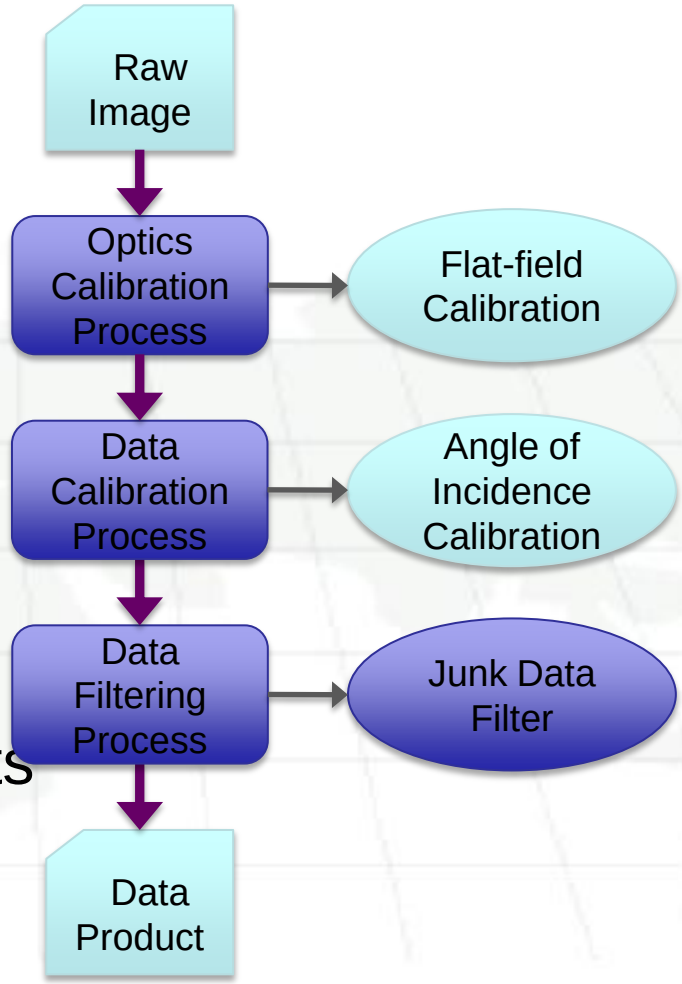
- Provenance in this data pipeline/ life cycle?
- Provenance is metadata in *context*
- What context?
  - Who you are?
  - What you are asking?
  - What you will use the answer for?



Modeling a Provenance Use Case – data workflow

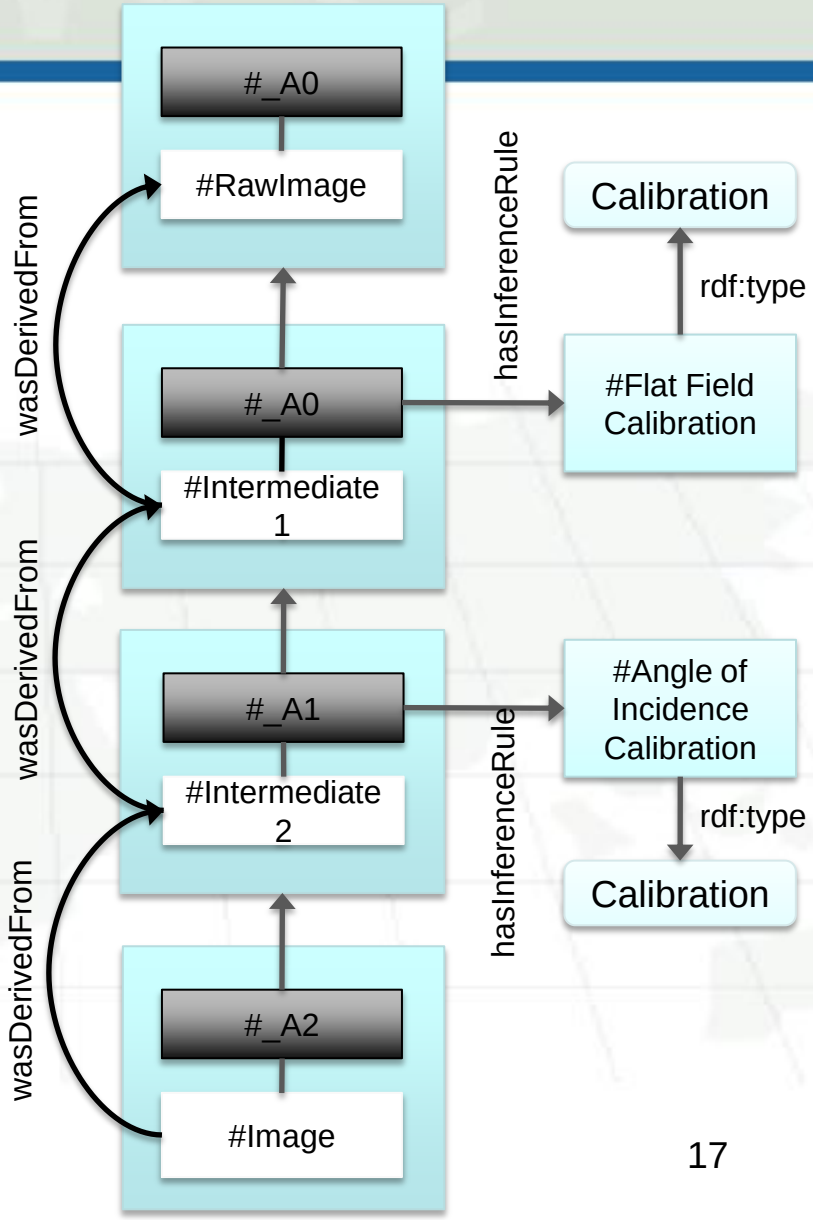
- What *calibrations* have been applied to this *image*?



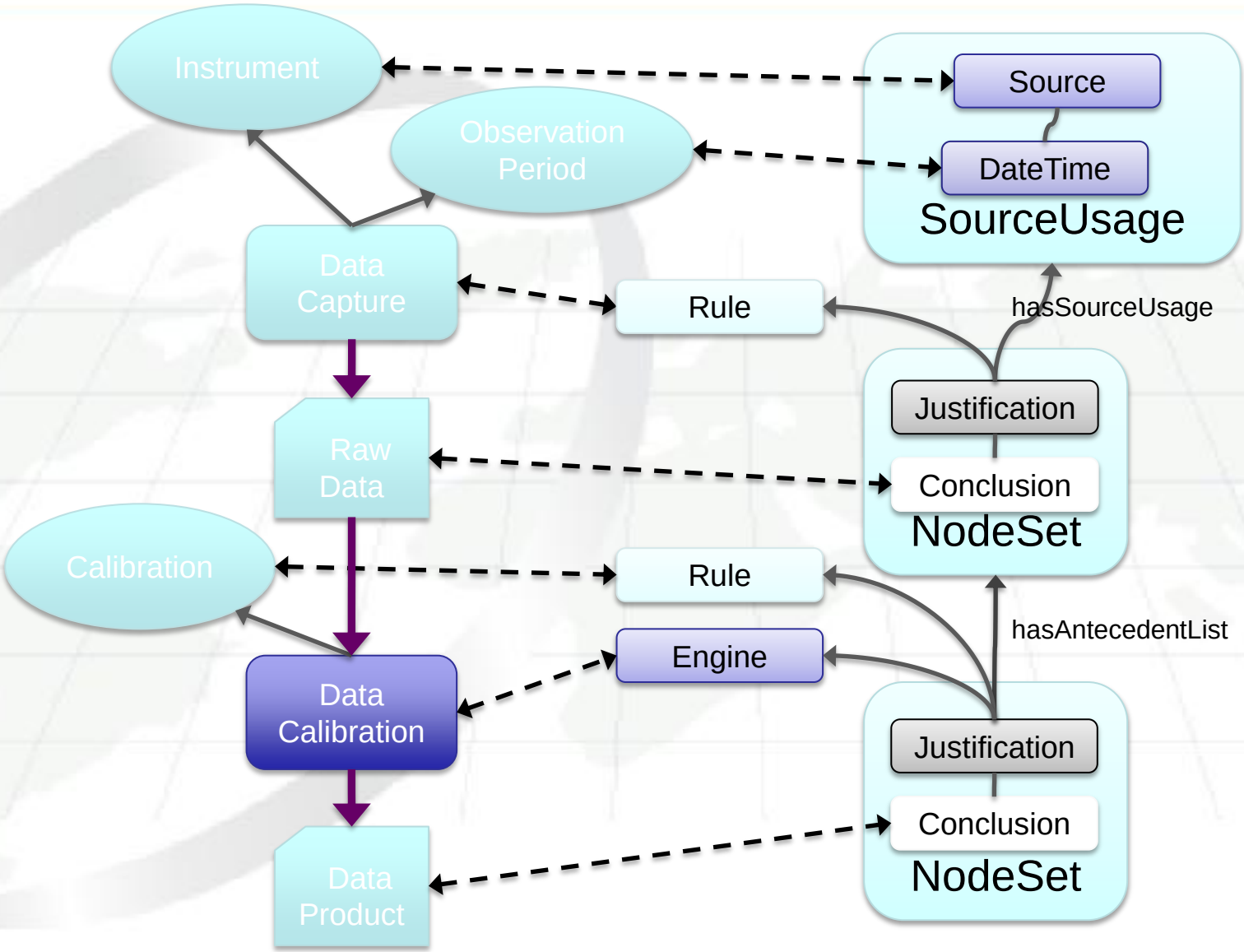


What *calibrations* have been applied to this *image*?

- We construct a query returns any individuals with type Calibration used as the InferenceRule in the justification from any artifact the current artifact was derived from.
- We assume that any calibration applied to an artifact the current artifact was derived from can also be considered as ‘applied’ to the current artifact, and that the wasDerivedFrom property is transitive



Concept Alignment (PML)



Plug for provenance in life cycle

- Provenance concepts describe how domain concepts are related
- Domain and provenance models should be independent, but aligned
- Aligning with a well-supported provenance model can enhance interoperability and tool support
- Aligned knowledge base supports complex multi-domain query and search

Life cycle is a complex issue but no longer intracable

- Must be
  - Managed
  - Modelled
  - Documented
  - Contextualized

✓Information models

✓Provenance

- As part of the use case, but also often outside it (pre-condition, trigger, …)