Provenance concepts and uptake within IVOA

Mireille Louys , CDS & ICube , Strasbourg University
François Bonnarel, CDS, Observatoire de Strasbourg
Mathieu Servillat, LUTH, Observatoire de Paris
and the IVOA Provenance team
within the IVOA DM working group

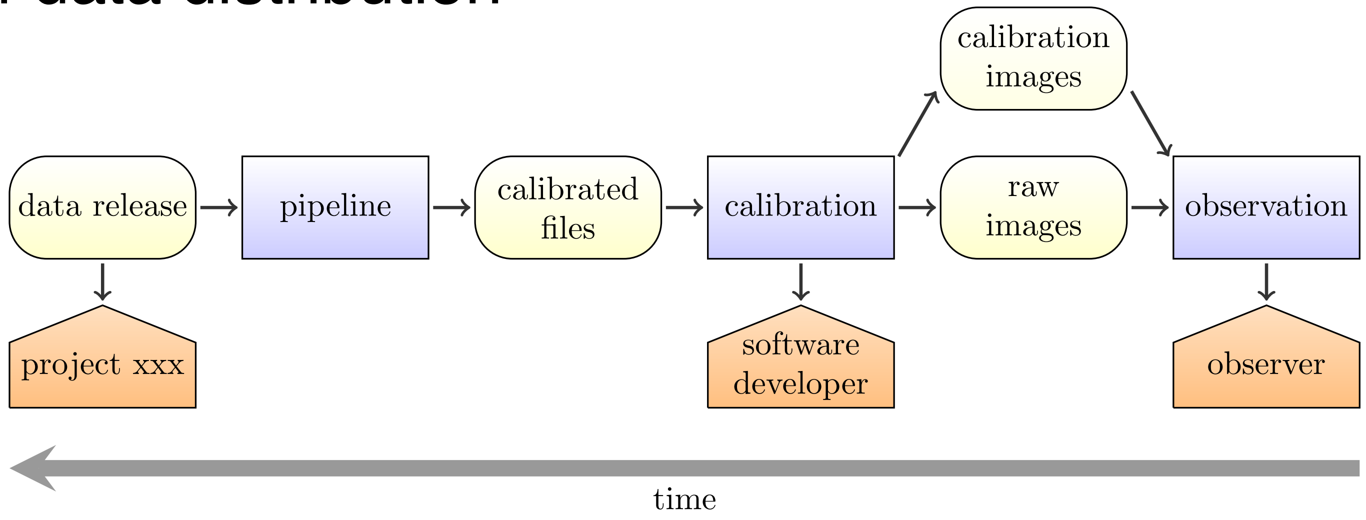
Centre de Données
Astronomiques de Strasbourg

ESCAPE
VO Virtual Observatory

IVOA

Provenance definition

- Provenance is a structured representation of the information necessary to capture, store, and analyse the history of data production. It supports trust and efficiency for the management of data archives in a re-usable and machine readable format.
- A notion that came with the automation of data production pipelines data and the large expansion of data distribution



- Crucial in science, and especially when data are widely distributed and re-used

□ Provenance goals

- Managing Provenance info allows for:

A: Tracking the production history

Find out which steps were taken to produce a dataset and list the methods/tools/software that was involved.

B: Attribution and contact information

Find the people involved in the production of a dataset, that need to be cited or can be asked for more information.

C: Locating error sources

Find the location of possible error sources in the generation of a dataset.

D: Quality assessment

Judge the quality of an observation, production step or dataset.

E: Searching in structured provenance metadata

This would allow one to also do a “forward search”, i.e. locate derived datasets or outputs.

Provenance history (1)

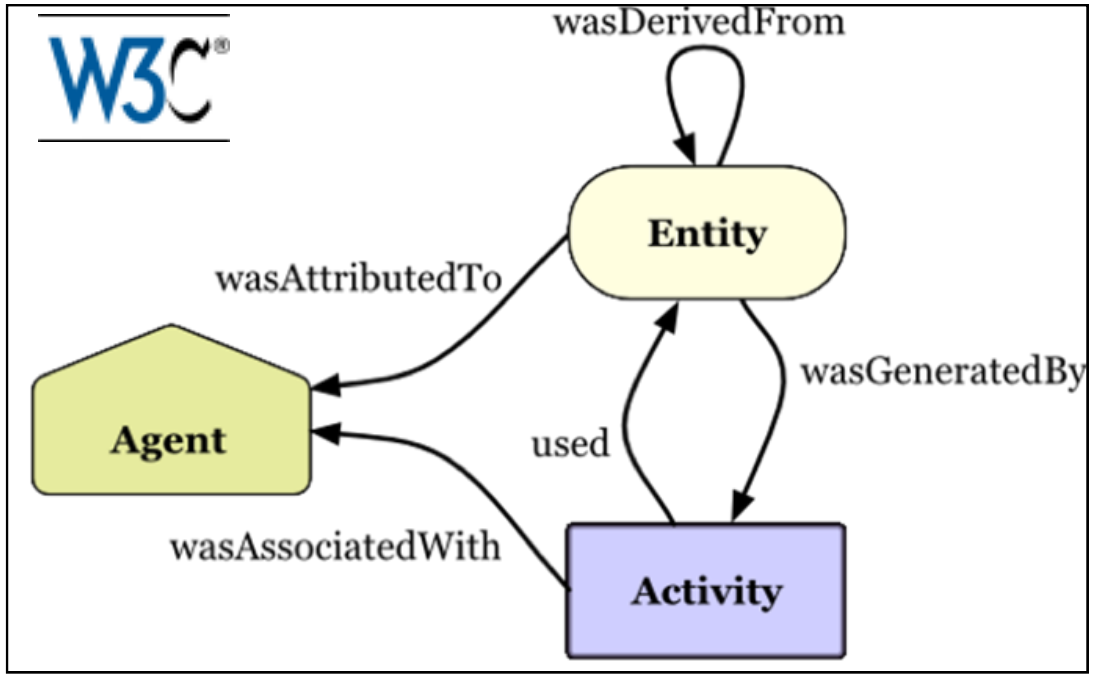
- a concern in various domains and science communities
  - document processing —> design and maintain workflows
  - genetics, molecular biology —> reproduce experiments
  - today for all observing science in general
- considered at beginning of IVOA data modeling group
  - for the Observation Data model 2005, J. Mc Dowell
  - for a reflexion on observing configurations, instruments, across various projects 2010, F. Bonnarel and coll.
  - see the history summary on the wiki page
    - https://wiki.ivoa.net/twiki/bin/view/IVOA/ProvenanceDataModelLegacy

□ Provenance history (2)

- developed and experienced in computer science research
  - 'Provenance Week' events since 1990, **IPAW conference**
  - modeled by the **W3C** consortium : a very general approach **PROV-DM** recommendation (2013)

    https://www.w3.org/TR/prov-overview/
- in parallel, description of Workflows have been developed with distributed computing
  - XML description, adhoc scripting languages, etc.
  - workflow management applications (Common Execution Architecture, Astrogrid)
- http://www.ivoa.net/documents/Notes/AstrogridWorkflow/AstrogridWorkflow-20060227.pdf
  - workflow management systems, e.g Taverna, grid management systems, etc.
- benefits from the development of **Metadata formats** : XML, JSON, …

W3C Provenance Pattern



- In our context

Entity

- data products (files), ancillary data (calibration, instrumental response, etc.), processing parameter files

Activity

- data acquisition, mosaicing, regridding, fusion, calibration, transformation

Agent

- Telescope astronomer, pipeline operator, principal investigator, data engineer, etc.
- W3C relations can make explicit:
  - Processing steps
  - Chain of dependencies
  - Responsibilities

□ Requirements listed for IVOA

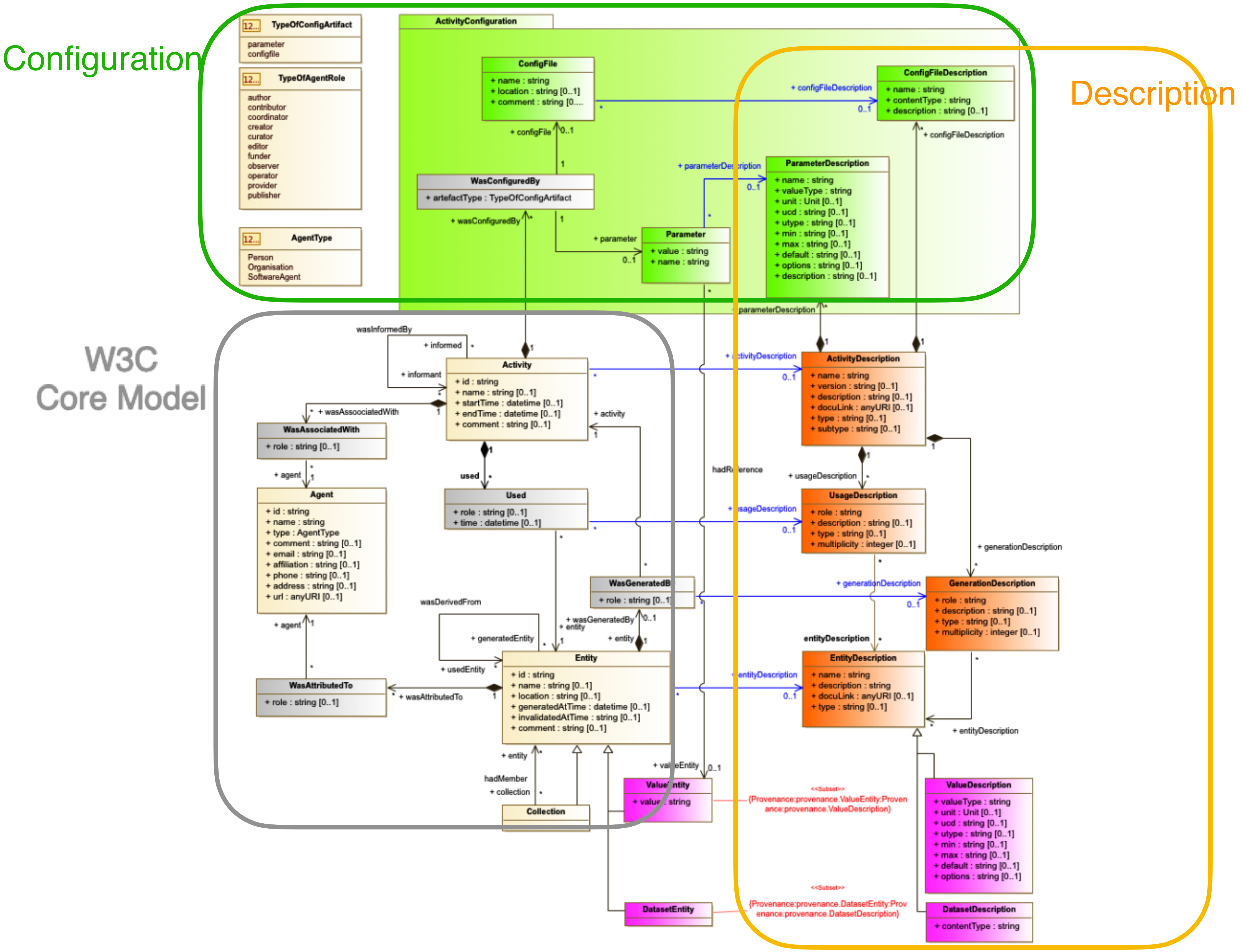
1. Provenance information must be stored in a **standard model**, with **standard serialization formats**.
2. Provenance information must be **machine readable**.
3. Provenance data model classes and attributes should be **linked to IVOA semantics, data models and formats** (DatasetDM, ObsCoreDM, SimDM, VOTable, UCDs, …).
4. Provenance information should be **serializable into the W3C provenance standard formats** (PROV-N, PROV-XML, PROV-JSON) with minimum information loss.
5. Provenance metadata must contain information to find immediate **progenitor(s)** (if existing) for a given entity, i.e. a dataset.
6. An entity must be linked to **the activity that generated it** (if the activity is recorded).
7. Activities must be linked to **input entities** (if applicable).
8. Activities may point to **output entities**.
9. Provenance information should make it possible to derive the **chronological** sequence of activities.
10. Entities, Activities and Agents must be **uniquely identifiable** within a domain
11. Released entities should have a **main contact**.
12. It is recommended that all activities and entities have **contact information** and contain a (short) **description** or link to a description

order of
importance
depends on
projects
-> core profile
+ options
tuning

□ PROV for the astronomical domain

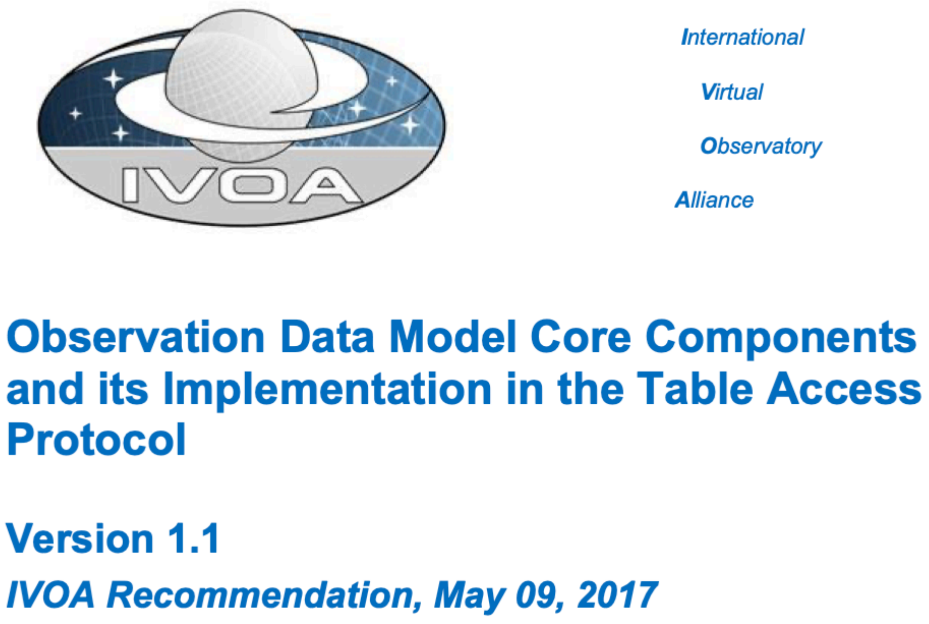
- W3C Provenance DM project is
  - good to understand concepts
  - very general —> not tuned to science data
- PROV-Store : a demo application for W3C ProvDM
  - https://openprovenance.org/store/
  - helps to build up examples
  - to practice various serialization languages: e.g. PROV-XML, PROV-N , PROV-JSON
  - does not scale easily with our datasets production chain
- Need more rules and customization to support project implementations
  - W3C PROV-Constraints REC https://www.w3.org/TR/prov-constraints/
  - Proposal in PROV-ONE (W3C Note) —> workflow and grid management
  - PROV-Template, etc. cf Luc Moreau 's team @Kings College London

IVOA provenance data model



□ IVOA landscape

- IVOA PROV-DM extends and restricts the general W3C PROV-DM
- It adds the data history
- IVOA has stable sets of metadata definitions for
  - data identification
  - data access and format
  - data curation
  - data physical properties (Characterization DM)



International
Virtual
Observatory
Alliance

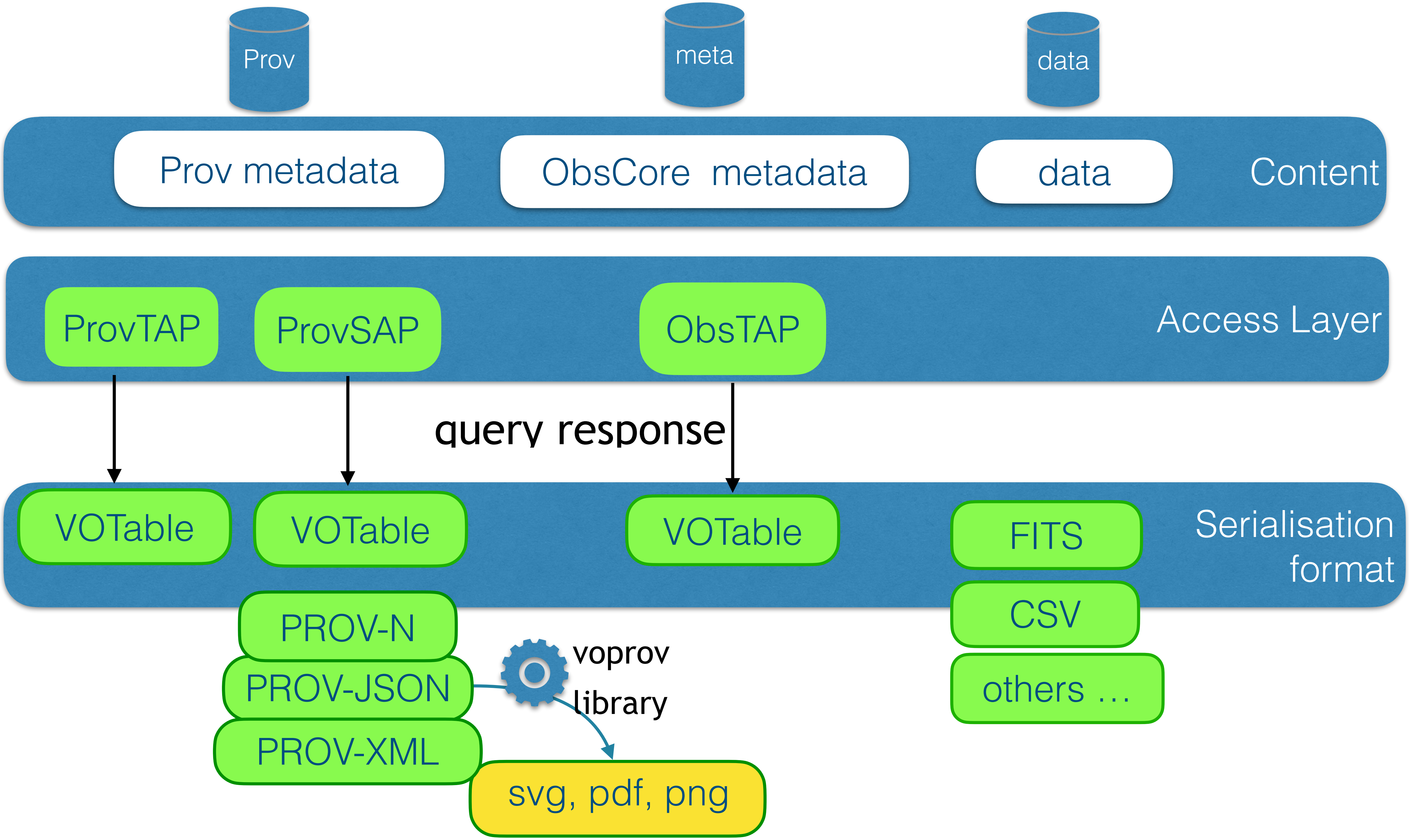
Observation Data Model Core Components and its Implementation in the Table Access Protocol

Version 1.1

IVOA Recommendation, May 09, 2017

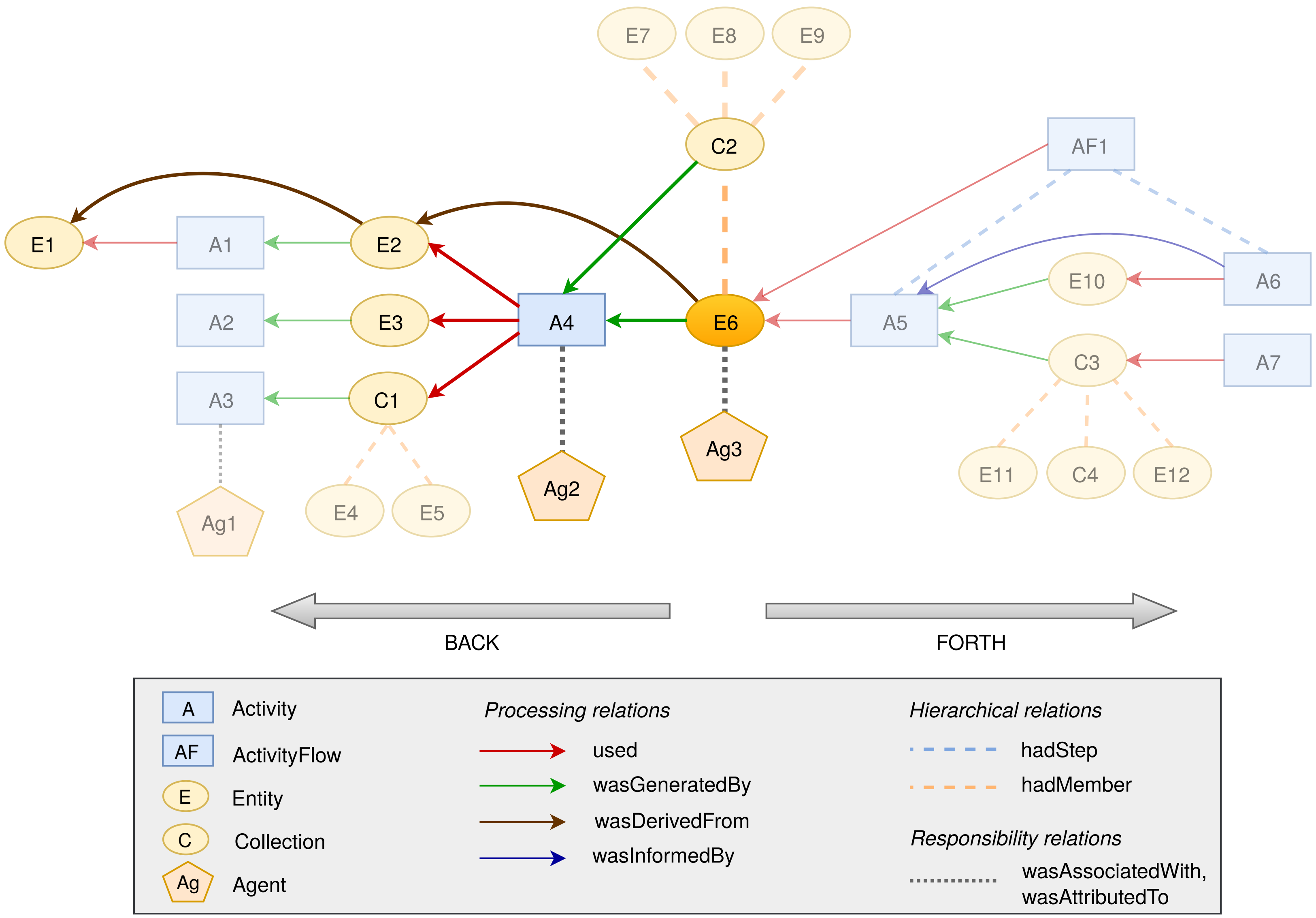
  - Various strategies to consider how to serve data, metadata, and Provenance records

VO building blocks



Implementation choices

- For each project
- Mapping for Entity
- define at which granularity to expose entities
  - datasets, collections of sub-observations, etc…)
- select the entity relations to implement
  - core: used +wasgeneratedBy
  - progenitor view : wasDerivedFrom
  - activity chaining : wasInformedby



credits : K. Riebe, AIP, Postdam

□ Implementation choices (2)

- For each project
- Mapping for Activity
- define granularity for task definition : ActivityDescription
    - existing dictionary of tasks? templates ?
    - interface to existing workflow descriptions language
        - e.g. CWL:Common Workflow Langage, Yaml
    - relations between Used—> UsageDescription, etc. or only Activity—> ActivityDescription
    - format for description : yaml, xml, cwl, scripting language, etc.

□ Conclusion

- The distribution of provenance metadata comes with a best effort strategy.
- On the data provider's side, the cost in implementing these features needs to be balanced with
  - an understandable content exportable outside the project
  - metadata clearly mapped means better queries prepared by the user or by the wrapping API (translator)
  - maintenance benefits to better monitor the archive collections (quality control, reprocessing, …)
- On the client side, an application querying several data centers will have to deal with the various level of completeness chosen by the data centers.
  - define various depth for provenance profiles : Core, Workflow description, all, etc.
  - enhance data search with provenance profile flavor selection

Thanks

Questions ?
Comments ?

□ Serialisation Formats

- Ready :
  - Gammapy Provenance embedded VOTable, PROV-N, PROV-XML
  - CTA Pipe/DIRAC text JSON
  - OPUS job submission and execution (LUTH) VOTable, JSON
  - Image database prototype in Triplestore (CDS) RDF/ttl
  - HiPS Image database (CDS) with PROV-TAP VOTable
  - Applause VOTable
  - RAVE implementation (AIP, Postdam)
    - Simple access (Prov-SAP) prototype Prov-N, PROV-JSON
  - Provenance for Pollux DB & *voprov* library (LUPM) VOTable, Prov-N, PROV-JSON
  - Under study :
  - SVOM pipeline execution tracking JSON FITS embedded