Astronomy ESFRI & Research Infrastructure Cluster
ASTERICS - 653477

Métadonnées de Provenance pour l'Observatoire Virtuel

Mathieu Servillat, Catherine Boisson – LUTH, Meudon

Michèle Sanguillon, Johan Bregeon – LUPM, Montpellier

Kristin Riebe, AIP, POSTDAM

Mireille Louys, François Bonnarel – CDS, ICube, Strasbourg

Préservation et Provenance

- En astronomie la préservation des données est constitutive du domaine
- Pas de vérité terrain mais un processus de validation croisée d'interprétation des observations au fil du temps
- Un enjeu pour l'astronome:

Sélectionner des jeux de données d'intérêts dans un grand ensemble de collections structurées

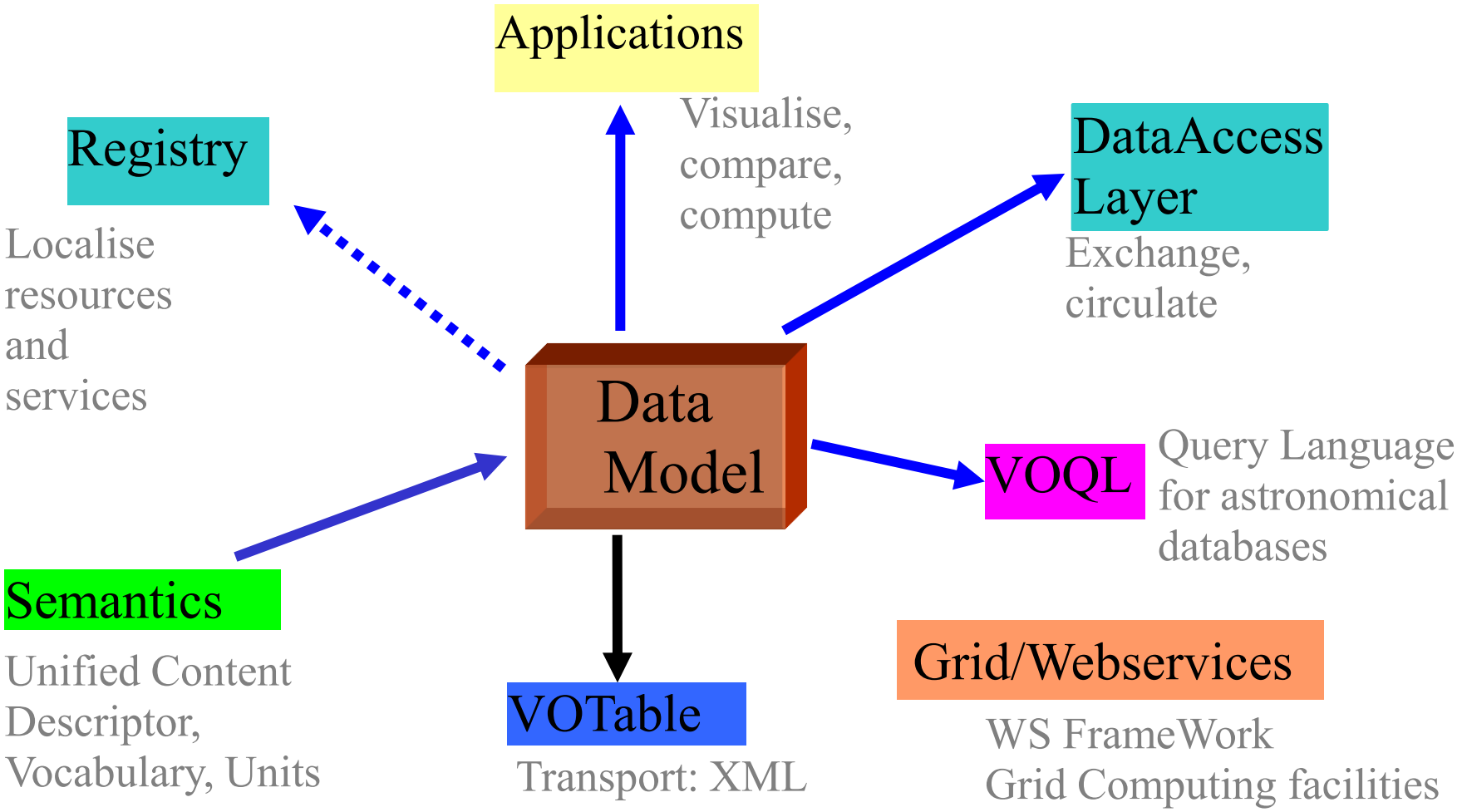
International Virtual Observatory Alliance

ivoa.net

Interopérabilité et fédération de données

- Développe une infrastructure interopérable pour la description, distribution, stockage et circulation des données en astronomie
- Orienté par les besoins scientifiques
- Standardisation process a la W3C
  - Working groups
  - Technical coordination
  - Science priority committee
  - Executive Board representing all national projects

Working Groups / Interactions



Découverte de données

- Quels critères de sélection ?
  - En fonction du sujet d'étude
    - Project, instrument, facility (type et nom de telescope)
    - Position sur le ciel ou classe d'objet
    - Propriétés physiques en position, temps, domaine spectral, flux
    - Types, formats, taille, etc
- Décrire les contenus d'observations
  - Modélisation des métadonnées

Data Models @ivoa.net

Space Time Coordinates http://ivoa.net/Documents/latest/STC.html	Space and Time Coordinates
Characterisation XML Shema http://ivoa.net/Documents/latest/CharacterisationDM.html	Physical axes, coverage, resolution, Precision
Spectral DM http://www.ivoa.net/documents/SpectralDM/index.html	Characterisation + Curation + Data
Observation Core Components DM http://www.ivoa.net/documents/ObsCore/20111028/index.html	Characterisation, Curation, Calibration,
Cube DM	Characterisation, Curation, Calibration for **N-D datasets**
Simulation Data Model	Objects and physical process

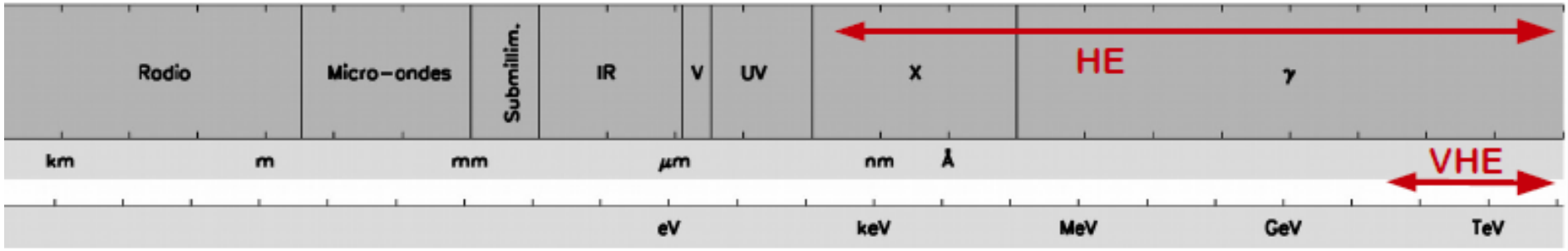
Provenance

- Décrire les étapes de production des données
- Processus et conditions d'observation
- Phases de réduction, sélection et extraction appliquées aux données brutes pour fournir les données interprétables

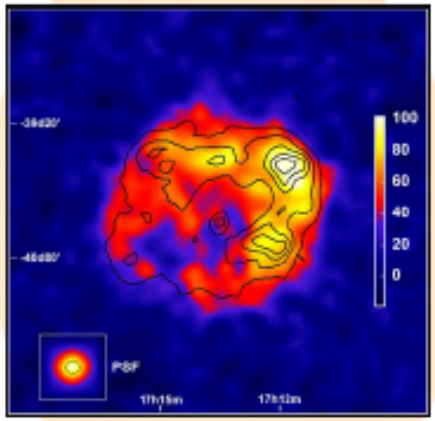
  (catalogues de sources, spectres, courbes de lumières, images, cubes hyperspectraux)
- Aider l'utilisateur à :
  - Définir des critères de sélection pour le choix des données utiles à son but scientifique
  - Choix entre différentes versions/collections de données
  - Recalculer une phase de réduction à partir de données intermédiaires

  →accéder aux progéniteurs des produits réduits

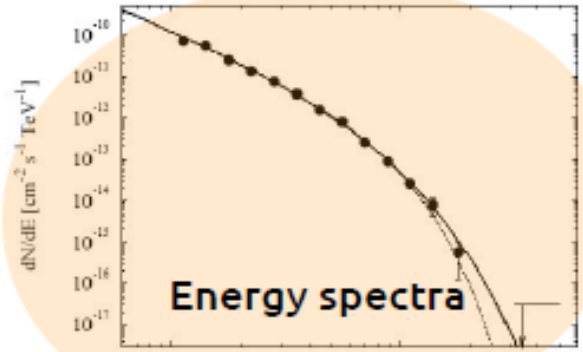
Very high energy data

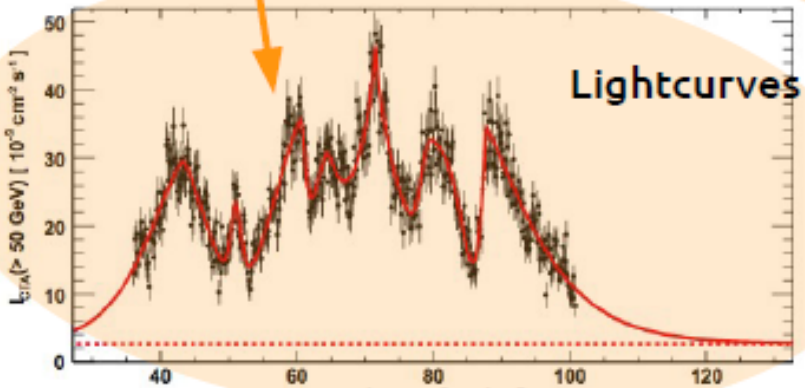


- Several orders of magnitude
- Photon counting
- Low count statistics, high background
- **Event lists**
  (coordinates, time, energy)



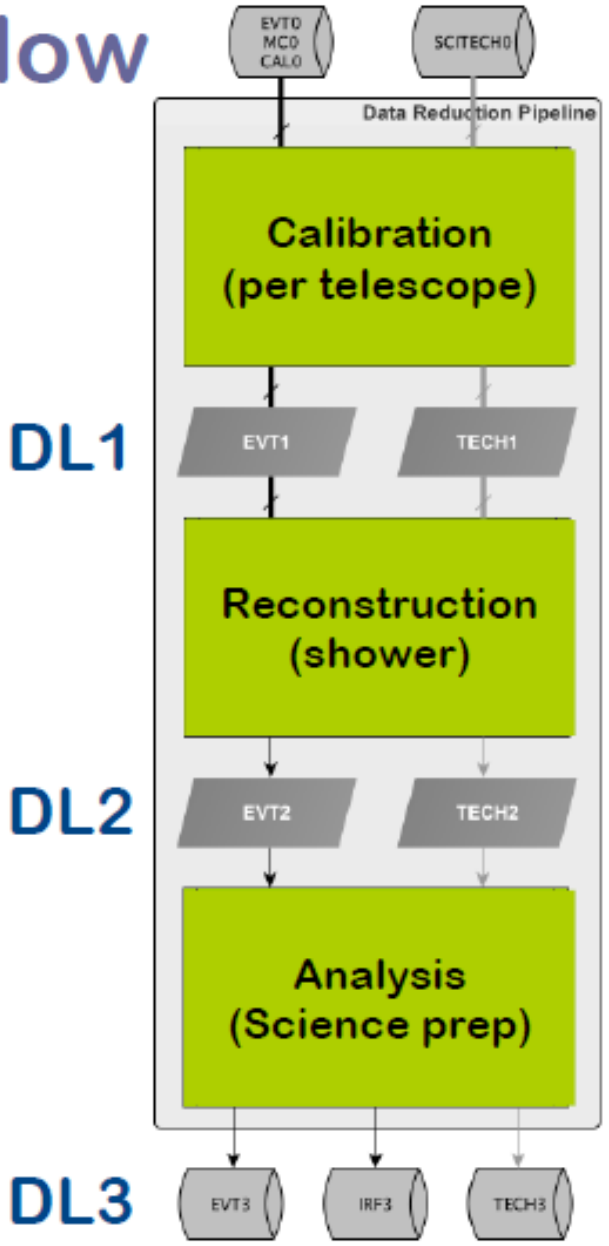
Images





CTA data levels and workflow

Data Level	Short Name	Description
Level 0 (DL0)	DAQ-RAW	Data from the Data Acquisition hardware/software.
Level 1 (DL1)	CALIBRATED	Physical quantities measured in each separate camera: photons, arrival times, etc., and per-telescope parameters derived from those quantities.
Level 2 (DL2)	RECONSTRUCTED	Reconstructed shower parameters (per event, no longer per-telescope) such as energy, direction, particle ID, and related signal discrimination parameters.
Level 3 (DL3)	REDUCED	Sets of selected (e.g. gamma-ray-candidate) events, along with associated instrumental response characterizations and any technical data needed for science analysis.
Level 4 (DL4)	SCIENCE	High Level binned data products like spectra, sky maps, or light curves.
Level 5 (DL5)	OBSERVATORY	Legacy observatory data, such as CTA survey sky maps or the CTA source catalog.



Provenance in the W3C

- W3C Provenance definition

"Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. "

PROV-OVERVIEW (Note), an overview of the PROV family of documents

PROV-PRIMER (Note), a primer for the PROV data model

PROV-O (Recommendation), the PROV ontology, an OWL2 ontology allowing the mapping of the PROV data model to RDF

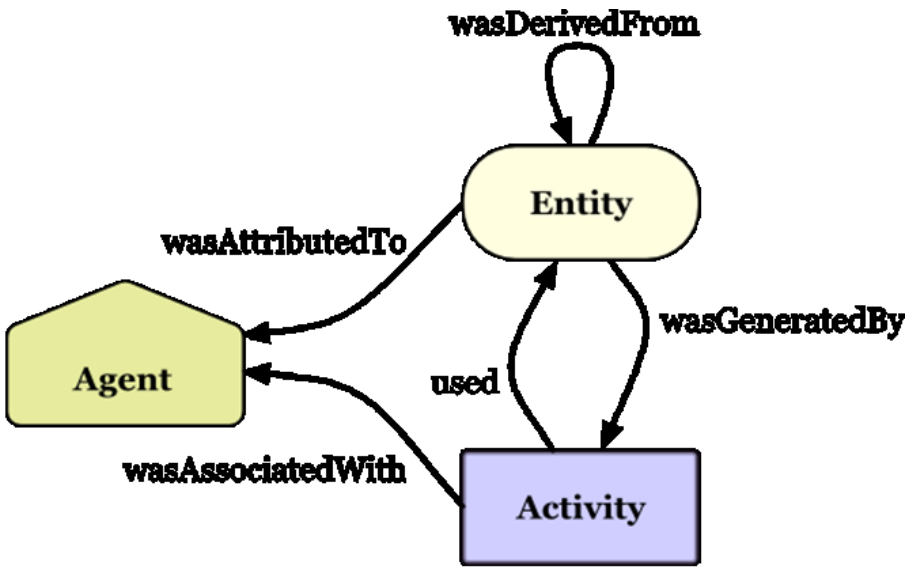
PROV-DM (Recommendation), the PROV data model for provenance

PROV-N (Recommendation), a notation for provenance aimed at human consumption

PROV-XML (Note), an XML schema for the PROV data model

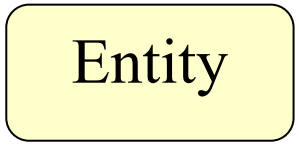
PROV-AQ (Note), mechanisms for accessing and querying provenance

W3C Provenance pattern



- Explicite ainsi les:
  - Etapes du pipeline
  - Dépendances
  - Responsabilités
- s'applique à toute description de chaîne de traitement, pipeline de réduction, analysis workflow, etc.
- Acquisition et Réduction

Dans notre contexte

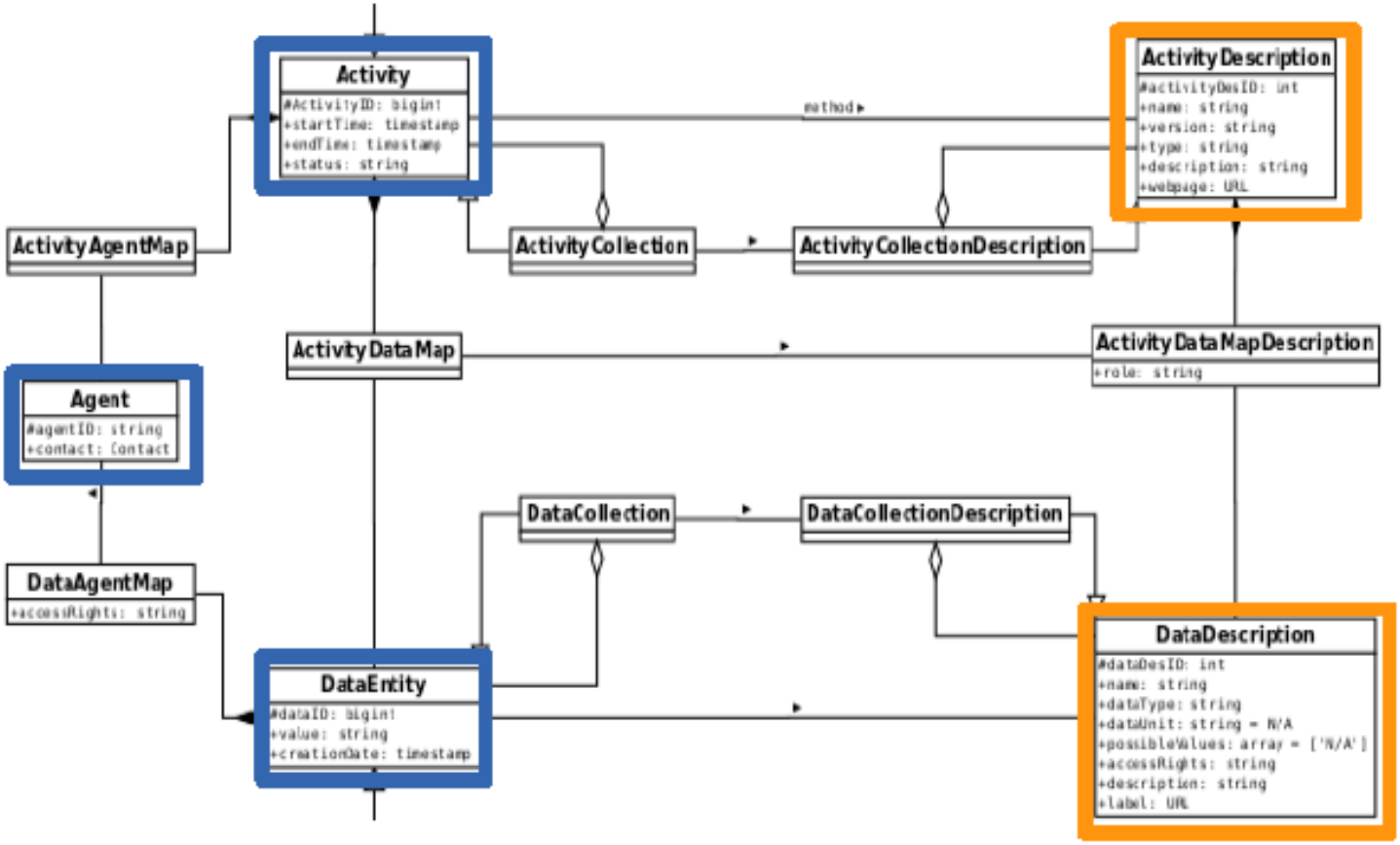


- data products (files), ancillary data (calibration, instrumental response, etc.), processing parameter files

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

- Telescope astronomer, pipeline operator, principal investigator, etc.

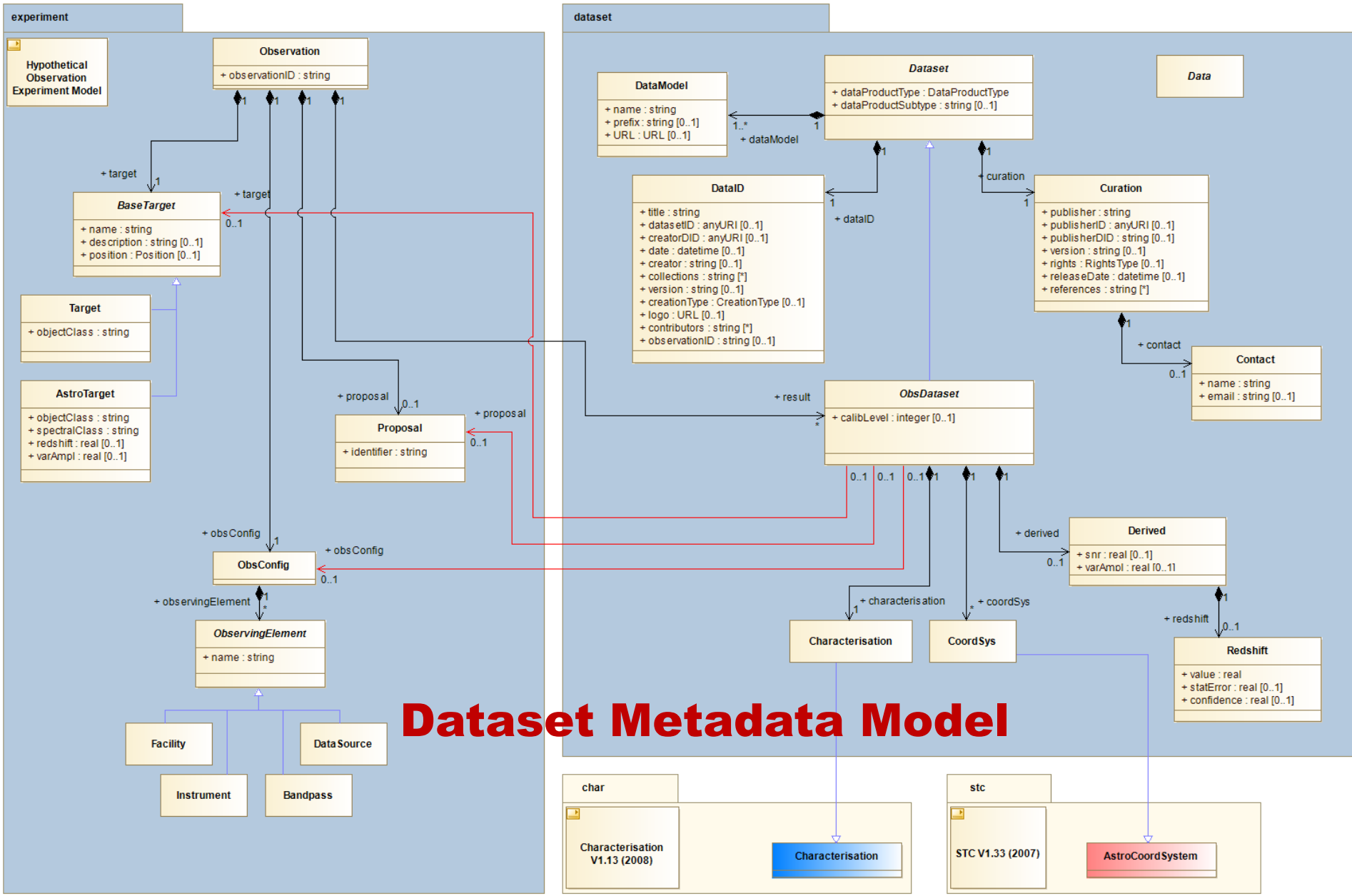
Provenance IVOA data model



Workflow description

Data Level description

Lien avec les modèles existants



Work Package 4

- Etudes en cours:
- Explorer la description du bloc ActivityDescription
  - M.Servillat, C. Boisson, M.Sanguillon, J. Bregeon

    → CTA : 4 niveaux de progéniteurs et leurs transformations
  - High energy physics

    → Ajustements de modèles paramétriques sur des spectres XMM
  - Theoretical spectra

    →Provenance pour la base de données Pollux au LUPM

Quel format de sérialisation ?

- habituellement une liste d'étapes:
  - Fichiers de log, journaux d'exécution
  - Listes de commandes sous forme de commentaires (header FITS)
- W3C offre plusieurs formes de syntaxe + traducteurs
- PROV-N (W3C)
  - Traces le scenario d'exécution en simple texte
  - Définit une grammaire

PROV-N

```
entity(rave:0645m522I0049.fits, [prov:type = 'std:fits']
entity(rave:0645m522I0049.wav.fits, [prov:type = 'std:fits']

agent(aao:Paul_Cass, [prov:type='prov:Person'])
agent(rave:Alessandro_Siviero, [prov:type='prov:Person'])

activity(rave:act_observation, 2008-02-16T13:25:24, -,
      [ prov:type = 'obs:Observation'  ])
activity(rave:act_irafReduction, 2008-03-04T09:46:57, -,
      [ prov:type = 'std:reduction' ])

wasAssociatedWith(rave:act_observation, aao:Paul_Cass, -,
                              [ prov:role = 'obs:Observer' ])
wasAssociatedWith(rave:act_irafReduction, rave:Alessandro_Siviero, -)

wasGeneratedBy(rave:0645m522I0049.fits, rave:act_observation, -)
used(rave:act_irafReduction, rave:0645m522I0049.fits, -)
wasGeneratedBy(rave:0645m522I0049.wav.fts, rave:act_irafReduction, -)
wasDerivedFrom(rave:0645m522I0049.wav.fts, rave:0645m522I0049.fits)
```

@ Kristin Riebe

Conclusion

- Emergence de projets gigantesques (LSST, …)
  - La stratégie « code to data » suppose une description adéquate des étapes de traitements, précise and interopérable.
- Un moment approprié pour relier la Provenance à l'infrastructure OV
- Diverses orientations pour des cas d'utilisation
  - Tracer la qualité des données → science
  - Management des pipelines → reproductibilité

Et dans d'autres disciplines ?

- Thème « Big Data »
- RDA Research Data Alliance

https://rd-alliance.org/groups/research-data-provenance.html