Provenance Metadata
An experience for the astronomical field

Mireille Louys[1,2], François Bonnarel[1],
1, CDS, Observatoire de Strasbourg
2, IPSEO/Images, Laboratoire Cube, Université de Strasbourg

for the Provenance effort of the IVOA Data modeling WG

CENTRE DE DONNÉES ASTRONOMIQUES DE STRASBOURG

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.”

- Base reliability on the fine description of operations, configuration, entities .
- We trust the scientists’ s decision and provide them with what they can interpret in terms of quality and reliability in their context

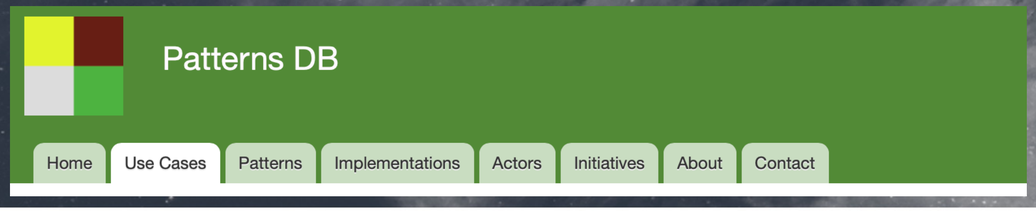
Provenance in RDA

Provenance Interest Group status : withdrawn
https://github.com/RDAProvIG/Group-Status/blob/master/CaseStatement.md

Provenance Patterns Working Group https://patterns.promsns.org
(login RDA nécessaire)
Nicholas Car, David Dubin, Paolo Missier

- Quelles situations récurrentes apparaissent ? Quel design pour les résoudre?

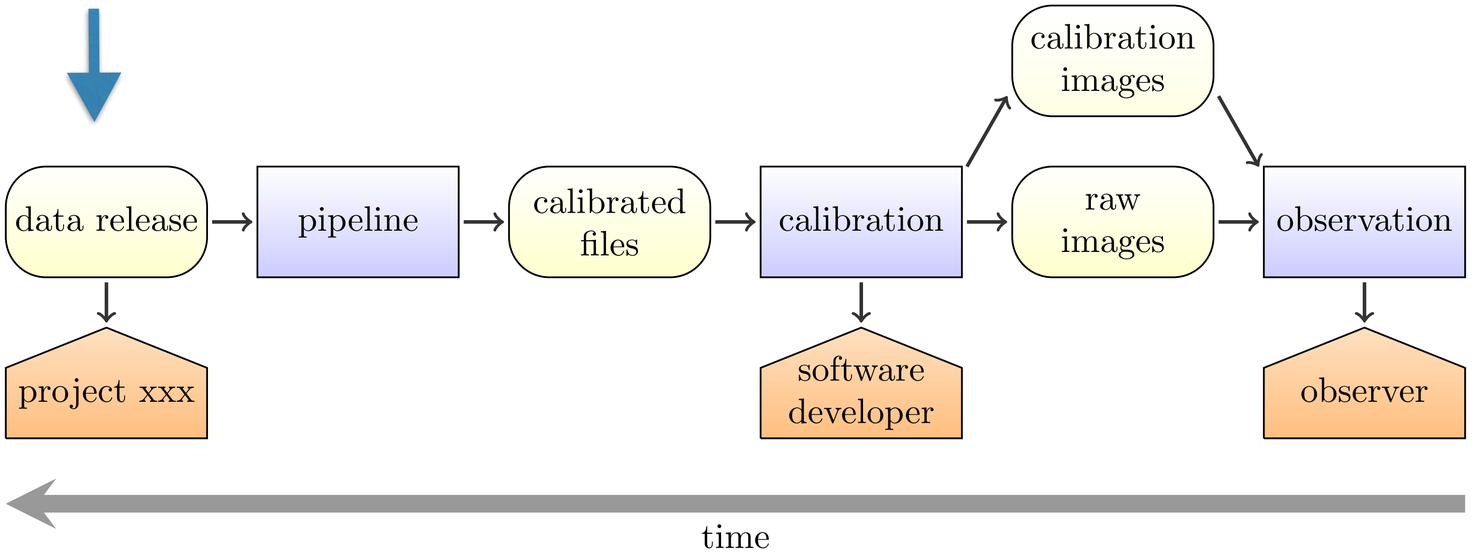
- Visiter le site pour vérifier
- Engagement
- Travaux
- Projets en développements :
  - Australian federal gov. Linke data Dregister,
  - Geoscience programs,
  - Britisch Museum



Use Cases

- Assemble a provenance graph from multiple sub-graphs
- Build a citation suggestion from provenance
- Build a software citation suggestion from provenance
- Checking that data's storage system matches provenance requirements
- Click through the provenance of a registered item and back
- Creation of metadata after the fact and repairing metadata
- Determine trust from the source of a dataset in a data conflation service
- Documenting provenance in an automated federation of spatial information
- Extract provenance metadata in a standardised format ★
- Find all data associated with a person's work
- Find all results derived from a dataset ★
- Find the agent who launched a workflow ★
- Find the basis for a decision
- Find the business rules supporting an outcome/result
- Find the explanation for a result
- Find the lineage of a result ★
- Find the lineage of a workflow
- Find the provenance of an analysis
- Indicate a dataset's funding source
- Link dataset to provenance information
- Locate multiple uses of data elements ★
- Logging a heterogenous-system workflow in a standardised provenance format
- View the usage of a dataset
- Provenance enable a data model
- Record standards-compliant provenance for datasets in a legacy data catalogue
- Scientists can track, list, and examine script executions
- Trace the agent who is responsible for a decision
- Understand how else this data has been used
- Understand who else has used this data ★
- Usage notes
- Use a dcat:Dataset as a subclass of prov:Entity
- Using provenance in the search for relevant spatial data
- Version release date
- View the lineage of a dataset
- View the production history of a dataset

Buts opérationnels en astronomie



- décrire la dépendance des données distribuées
- stocker les différentes étapes de traitements et leurs produits
- décrire les responsabilités des acteurs impliqués
- permettre de rejouer certaines étapes des traitement à partir des états antérieurs des données

Contexte de l’observatoire virtuel

- c’est un projet collectif des acteurs du domaines de l astronomie
- il favorise la mise en place d’une infrastructure de partage de données des fournisseurs vers les utilisateurs
- il définit et assure l’ interopérabilité des services
- il décrit de façon standardisée les contenus et les formats de données
- il définit des protocoles pour les accéder
- il répertorie les services et banques de données dans un annuaire global mis à jour en continu (registry)

- ce cadre permet le développement d’applications communes à destination des astronomes utilisateurs pour les requêtes et traitements scientifiques

Objectifs du modèle Provenance IVOA

- A: Traceability of products
- B: Acknowledgement and contact information
- C: Quality and Reliability assessment
- D: Identification of error location
- E: Search in structured provenance metadata
- F: Enable reproducibility

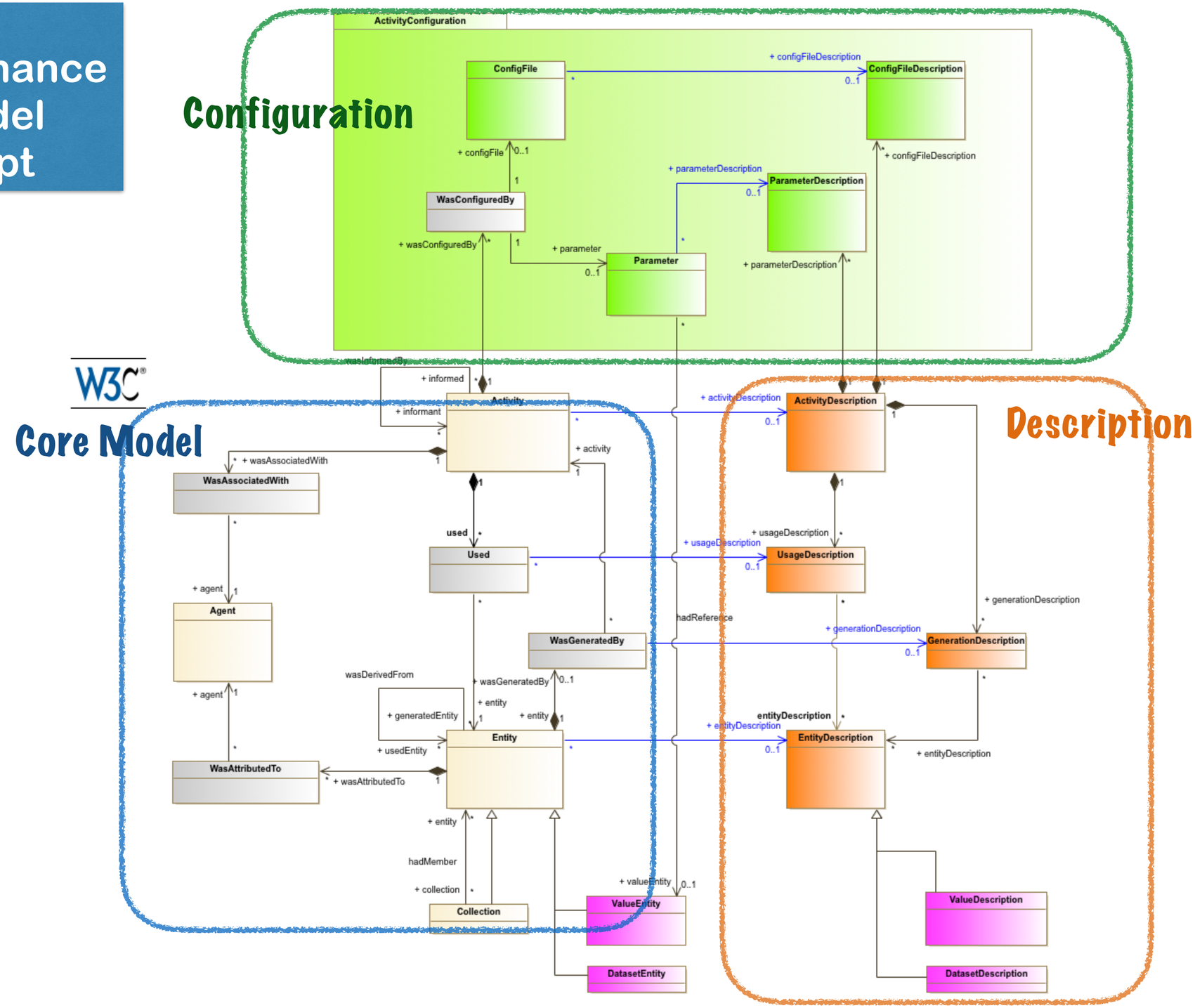
Mise en oeuvre

- à la discrétion du fournisseur de données
- selon le principe « best effort »
- variabilité des implémentations

- pour l‘ interopérabilité:

différents niveaux de compatibilité à définir

IVOA Provenance Data Model 2019 Sept



Provenance IVOA Data Model

- http://www.ivoa.net/documents/ProvenanceDM/
- lien direct

IVOA Provenance Data Model

Version 1.0

IVOA Proposed Recommendation 2019-07-19

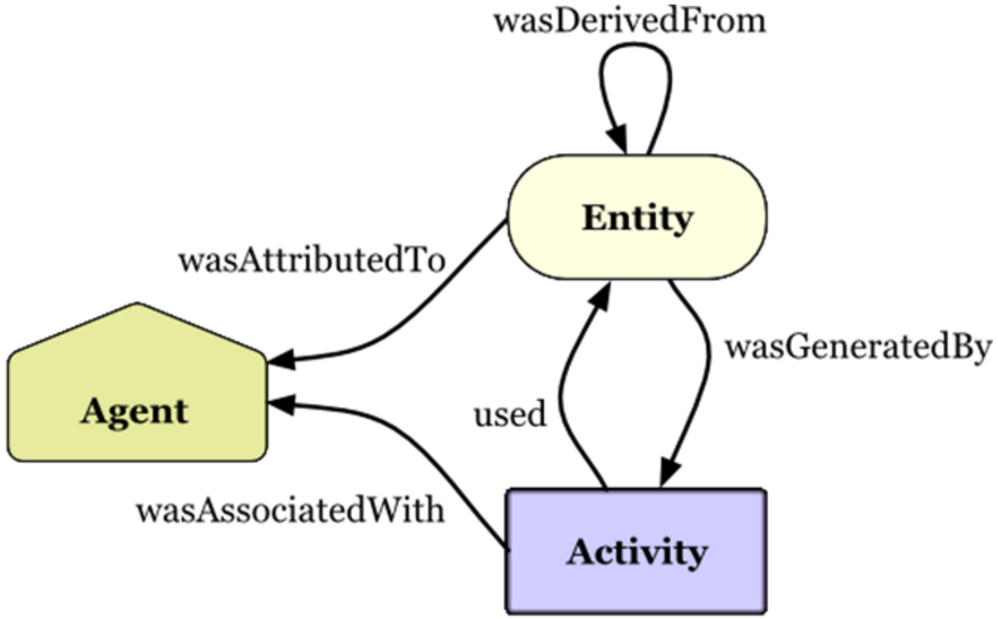
Working group
DM
This version
http://www.ivoa.net/documents/ProvenanceDM/20190719
Latest version
http://www.ivoa.net/documents/ProvenanceDM
Previous versions
WD-ProvenanceDM-1.0-20190614.pdf
PR-ProvenanceDM-1.0-20181015.pdf
WD-ProvenanceDM-1.0-20180530.pdf
WD-ProvenanceDM-1.0-20170921.pdf
WD-ProvenanceDM-1.0-20161121.pdf
ProvDM-0.2-20160428.pdf
ProvDM-0.1-20141008.pdf
Author(s)
Mathieu Servillat, Kristin Riebe, Catherine Boisson, François Bonnarel, Anastasia Galkin, Mireille Louys, Markus Nullmeier, Nicolas Renault-Tinacci, Michèle Sanguillon, Ole Streicher
Editor(s)
Mathieu Servillat

Accès / Distribution

- TAP (Table Access Protocol) est un type de protocole adapté aux données tabulaires très répandu dans l Observatoire virtuel pour l'échange de données
- Interface une base de données relationnelle (PostGres, Oracle, Mariadb, etc.)
- Definit un schéma de métadonnées tables, colonnes en XML : ProvTAP.xml
- S'accède par des applications telles que
- TapHandle (http://saada.u-strasbg.fr/taphandle/), Topcat (http://www.star.bris.ac.uk/~mbt/topcat/) qui permettent de poser des requêtes
- Représentation des réponses :
  - en format ivoa : VOTable ou JSON spécifique
  - en format compatible W3C : PROV-N, PROV-JSON , PROV-RDF, via une bibliothèque de conversion de format

Visualisation des résultats

- Extension des représentations W3C en PROV-N, PROV-RDF, etc.
- Le modèle W3C propose des meta-classes que nous avons dérivées pour enrichir le modèle IVOA et qui peuvent être embarquées dans les formats W3C si besoin



Règles : (en cours d'élaboration)

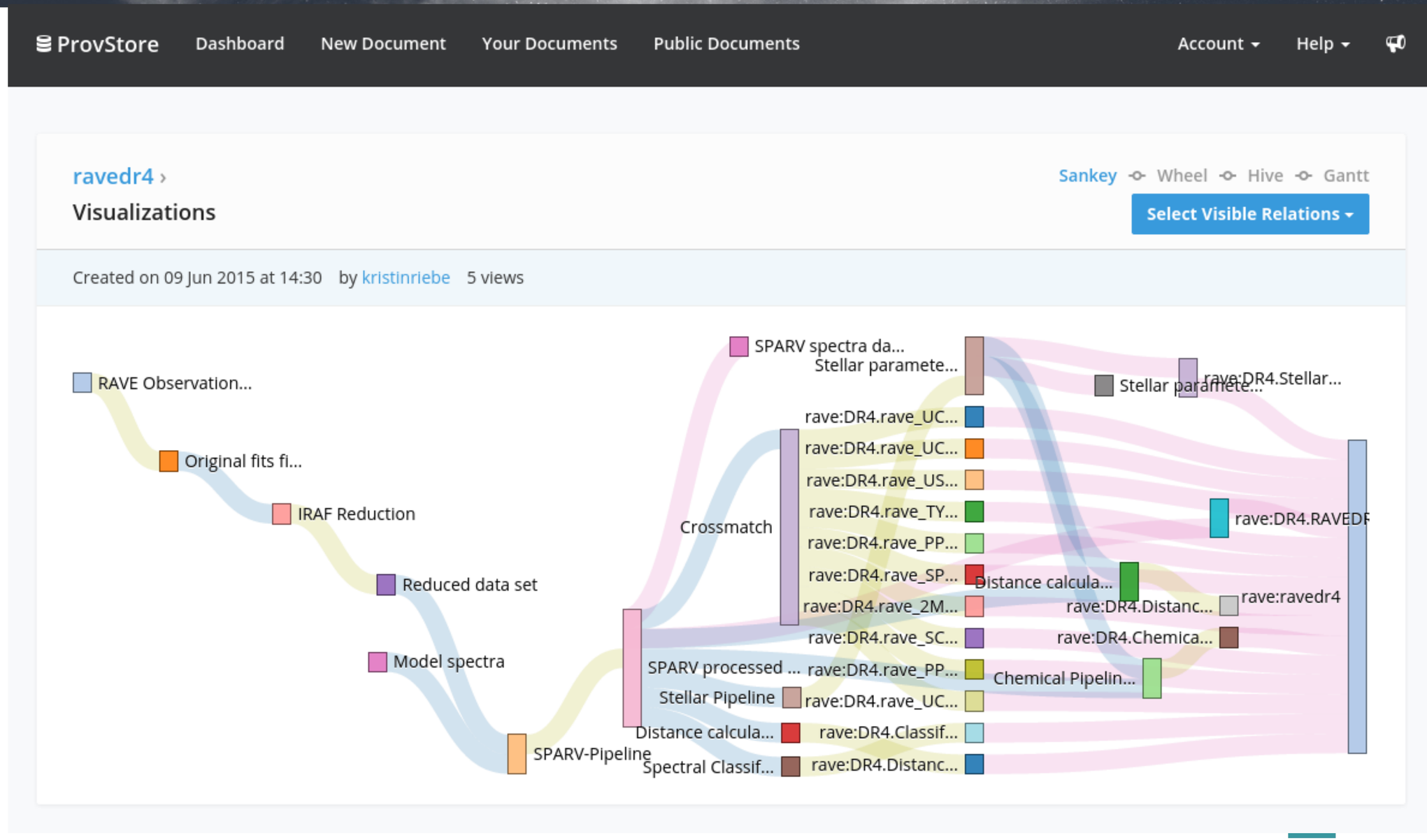
- *wasInfluencedBy* pour les relations gauche droite de description 'isdescribedBy'
- collection des éléments du template *Activitydescription*

Exemple : Job manager application : OPUS / VO-Paris

https://voparis-uws-test.obspm.fr/provsap?ID=6188ab&DEPTH=ALL&AGENT=1&DESCRIPTIONS=1

Visualisation des résultats/ développements

- RAVE provenance prototype (Riebe et al.)
- Outils de visualisations interactifs



Graph data bases

- Triplestore est une architecture intéressante
- Utilisée dans RDA pour Provenance Patterns DB
- CDS prototype basé sur Blazegraph
  - extension de l'ontologie PROV-O de W3C
  - requêtes SPARQL (exemples)

Les +

- ingestion progressive des portions du modèle
- supporte différents profils de compatibilité
- réutilise et étend facilement les représentations W3C comme PROV-O

Provenance Metadata in a Triplestore

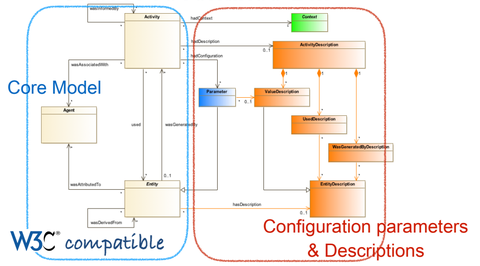
Goal

Evaluate the **triplestore database organisation** for implementing the IVOA Provenance data model.
This model extends the PROV-DM defined by W3C.
In the IVOA framework, Entities typically represent data products, Activities the tasks consuming and producing Entities. Credits or responsibility is given to Agents for each Entity or Activity. Parameters and Descriptions for the methods applied in an Activity together with the roles of Entities in the scenarios are described in specific relations and classes.

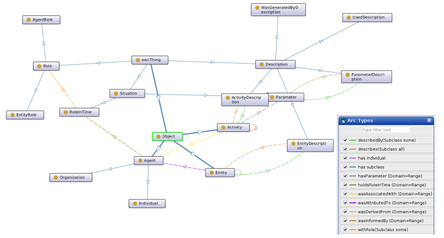
CDS Prototype Image Database

A test database tracing the processing of image data sets, from plates through files with digitization, cut-outs, RGB combination, HiPS conversion has been used for testbed. It implements IVOA Provenance DM in PostGres and supports a TAP/ADQL query interface.

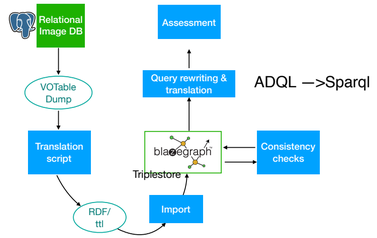
IVOA Provenance DM



PROV-O Ontology Extension



Implementation and Testing



SPARQL query testing

« Give me all agents associated to an entity or to an activity which formerly used this entity named 'E' ».

```
SELECT ?name ?role ?relation WHERE {
{ :*E* ?relation ?x .
                    ?x :refersTo ?name .
                    ?x :holdsRoleInTime/:withRole ?role .
FILTER regex( str(?relation), "wasAttributedTo", '' i"). }
}
UNION
{  ?activity :used/refersTo  :*E* .
   ?activity ?relation ?y .
   ?y :refersTo ?name .
   ?y :holdRoleInTime/:withRole ?role .
   FILTER regex( str(?relation), "wasAssociatedwith", '' i"). }
}
```

Sparql filters relations on their names and so can avoid many joins
Set of comparison queries:
http://wiki.ivoa.net/internal/IVOA/ProvenanceRFC/ProvQuerytest-3store.pdf

Lessons Learned

The **Triplestore RDF/ttl** offers :
- **equivalent support** for queries in SPARQL compared to ADQL
- **flexibility** to code relations and add new properties
- **extensibility** if the model grows with new properties of classes/ relations
- **expressibility** of searching criteria
- **scalability** very stable and efficient with many relations and instances.
- Blazegraph together with a spatial index code scale properly up to 8,5 million objects extracted from the Simbad database.

The **IVOA Provenance data model** can :
- circulate in **multiple serialization formats : IVOA (VOTable) and semantic web (RDF/ttl)**
- propose an interoperable framework to trace provenance info
- adjust to various compliance levels, from simple to rich descriptions
- answer a **large variety** of queries for provenance use-cases
- reuse the W3C provenance concepts with some degree of freedom and adaptability

Mireille Louys, L. Holzmann, F.-X. Pineau, F. Bonnarel
mireille.louys@astro.unistra.fr

ADASS XXVIII - College Park MD, , USA - 2018 - CDS/Observatoire de Strasbourg

Feedback modélisation et mise en place

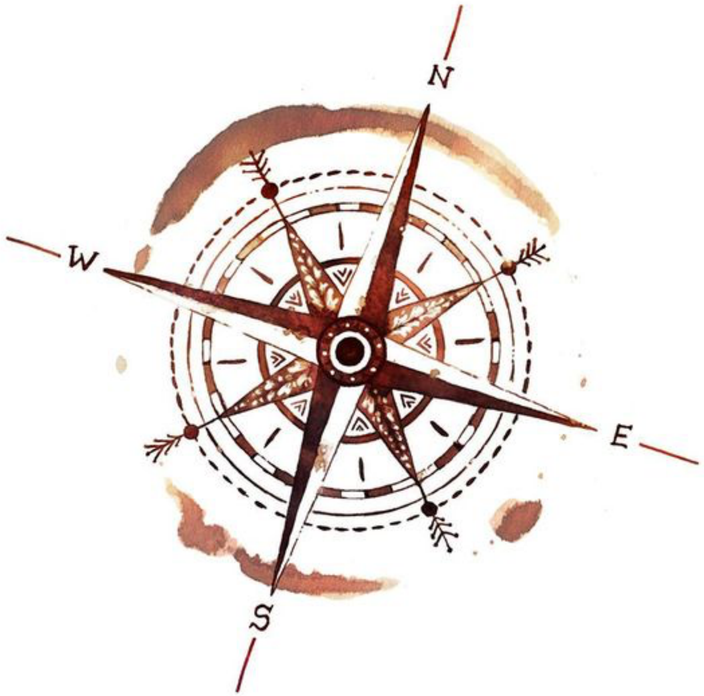
- Choix nécessaires du fournisseur de données
- Isoler les données elles-mêmes et leurs entités de Provenance
- Adapter la granularité
  - aux entités (data, components, objets)
  - aux traitements

- Détailler les descriptions d'activités selon le domaine
  - application en cours pour **ctapipe**, (https://cta-observatory.github.io/ctapipe/) le workflow de traitement des données CTA (Cherenkov Telescope array)
  - enregistre un ensemble de métadonnées de provenance pendant l exécution d'une étape
- Connexion avec des initiatives de partage de codes (cf RDA WG)
  - référencement d'applications
    - Force 11
    - ASCL (astronomie)
    - autres initiatives

Conclusion

- Un modèle complet
- Il peut être décliné selon des profils simplifiés par domaines d'intérêt
- s'applique dans les cas *a priori* (built-in provenance) et *a posteriori* (extraction des données et cahiers d'exécution)
- nécessite des choix d'implémentations du fournisseur pour ajuster la granularité aux entités et aux activités
- Connecter les parties de descriptions de code à représenter dans ActivityDescription vers d'autres initiatives

  —> RDA working groups



Merci de votre attention