

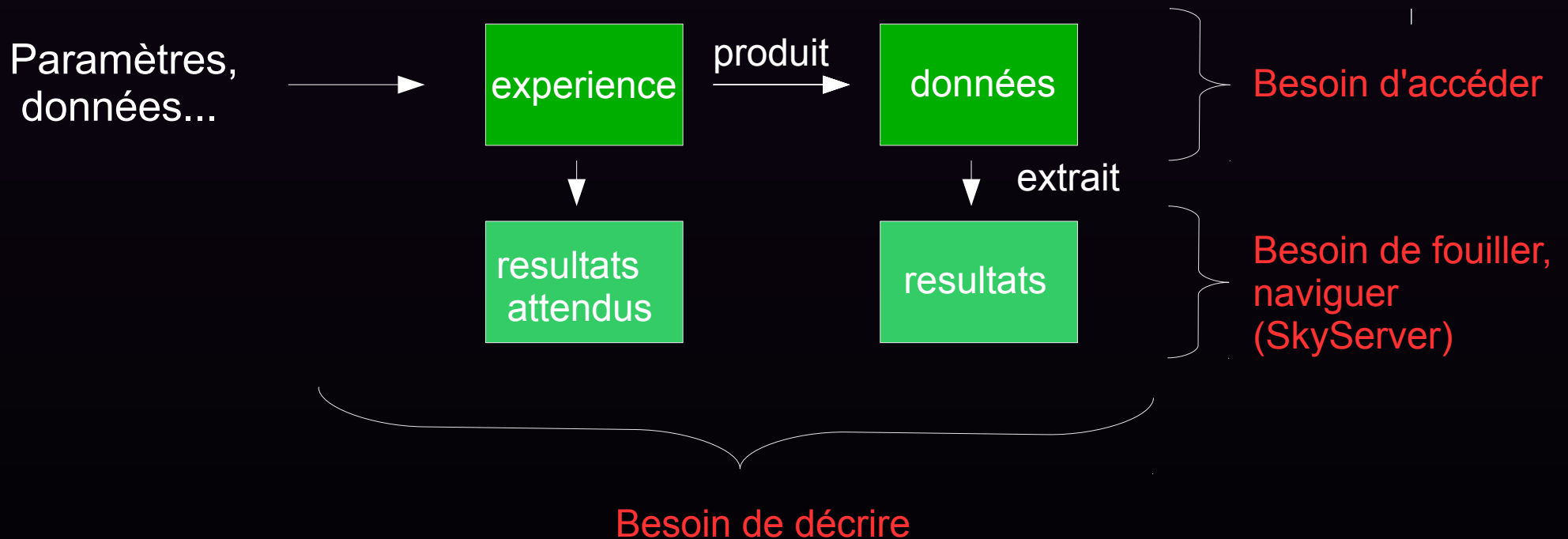
Les très grands ensembles de données des simulations numériques

David Languignon, Franck Le Petit

ASOV meeting - Paris - 17 Janvier 2012



Cas d'utilisation



Actions de standardisations au sein de VO-Theory

- Solutions pour décrire
 - Simulation Data Model (SimDM) **Ok**
- Solutions pour accéder/fouiller
 - Simulation Data Access Layer (SimDAL) **En Cours**
 - Adaptation des solutions DAL pour les simulations
 - Ex : TAP..
 - Couche Sémantique (skos concepts...) **En cours**

Ce qu'on peut faire aujourd'hui

- Décrire de manière **standardisée** (SimDM)
- Accéder de manière **partiellement standardisée**

The image shows a 3D visualization of a point cloud of red dots in a 3D coordinate system (x, y, z). The x and y axes range from 0 to 150, and the z axis ranges from 0 to 150. The points are distributed in a roughly spherical cloud. The interface is labeled '3D' and includes a toolbar with various icons for manipulation and viewing.

The TOPCAT interface shows the 'Table List' with '1: results' selected. The 'Current Table Properties' panel displays the following information:

- Label: results
- Location: WebSampConnector:results
- Name: results
- Rows: 6,386
- Columns: 4
- Sort Order: ↑
- Row Subset: All

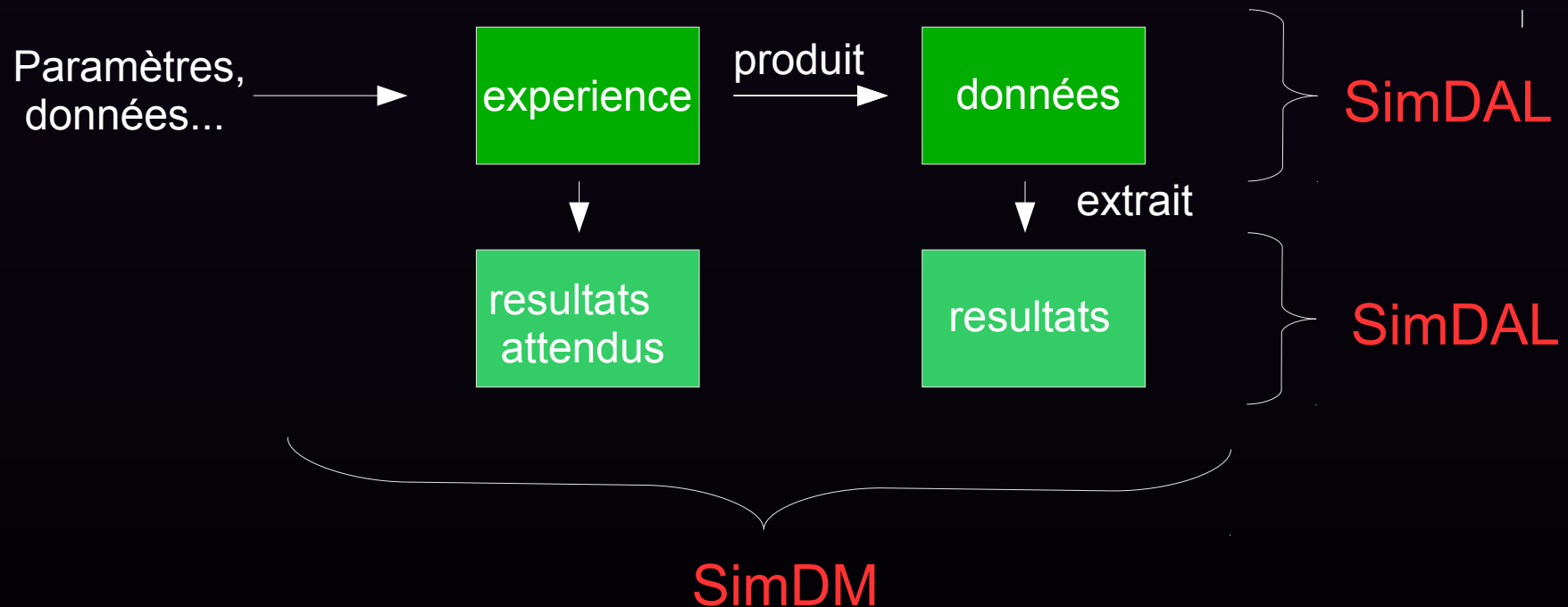
The 'SAMP' section shows a 'Messages' field and 'Clients' with icons for various software packages.

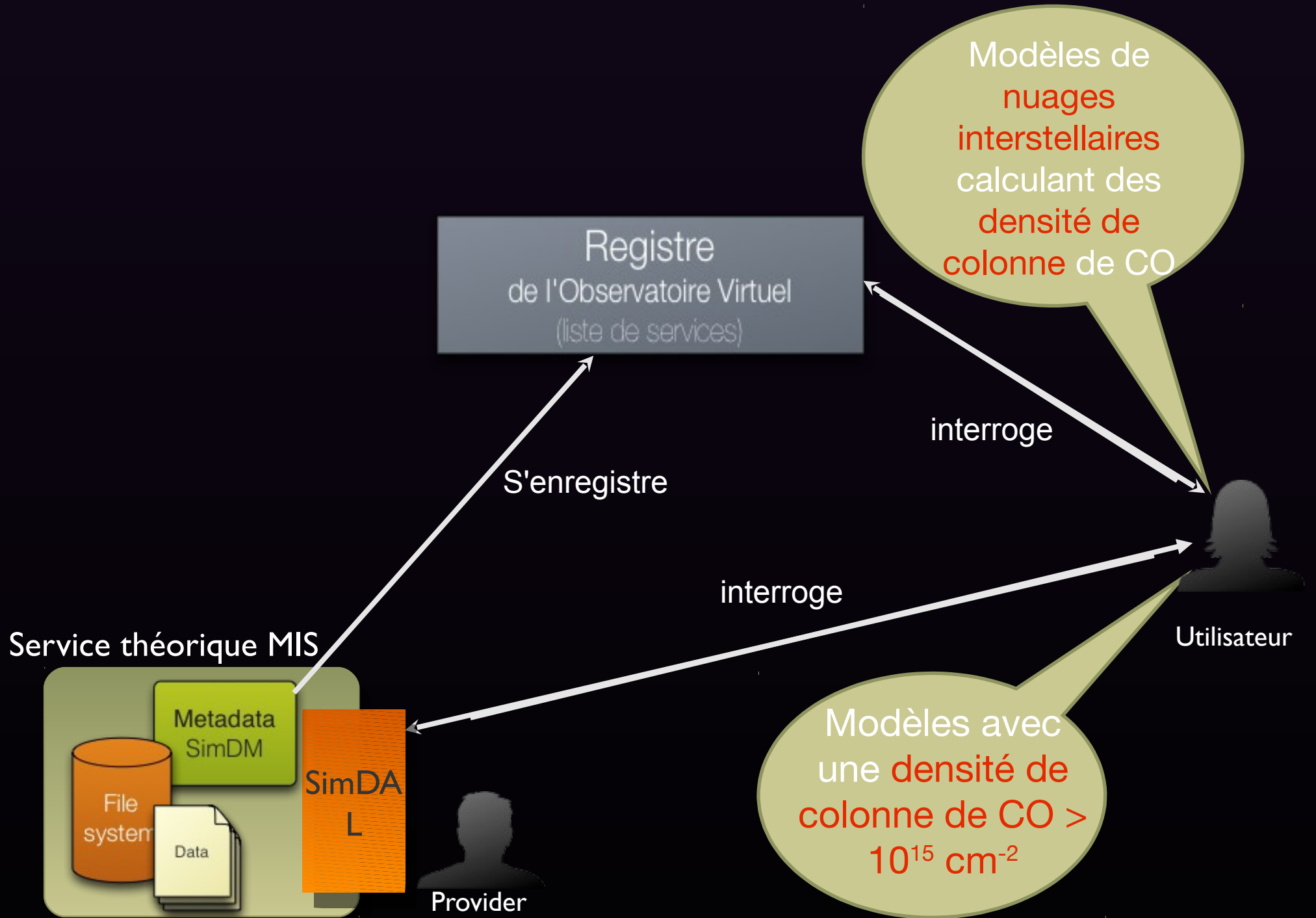
Below the TOPCAT interface, a table of matching objects is displayed. The table has columns for X, Y, Z, mass, and wanted particles (select all). The table contains 7 rows of data:

X	Y	Z	mass	wanted particles (select all)
4.37	160.56	7.16	1.53e+12	<input type="checkbox"/>
2.46	24	14.96	5.99e+12	<input type="checkbox"/>
23.89	15.64	2.13	3.23e+12	<input type="checkbox"/>
23.21	14.7	2.07	1.62e+12	<input type="checkbox"/>
34.92	13.28	6.94	3.35e+12	<input type="checkbox"/>
37.15	14.63	9.78	2.22e+12	<input type="checkbox"/>

Below the table, there is a 'Search matching objects' button and a message: 'There are 6386 matching objects'. Below this message, there are links for 'Extract selected Halos', 'Broadcast VOTable through SAMP', 'Download Text File', 'Download Votable', and 'TopCat: Download | Launch'.

Il manque un standard d'accès



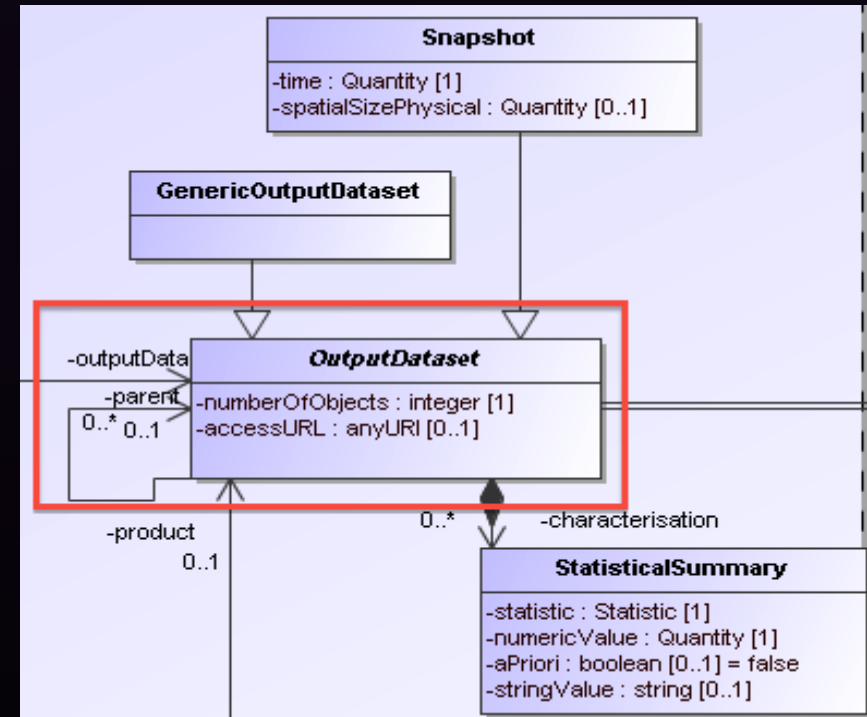
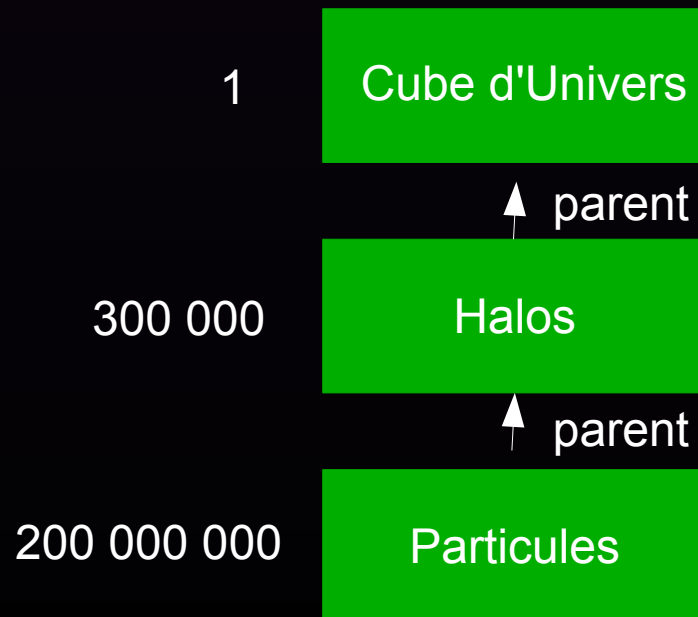


SimDM / SimDAL

Périmètre SimDM, SimDAL ?

Quel grain pour l'arborescence
OutputDataset ?

Ex : Cosmologie



Utilisation des resultats

- Pipe entre plusieurs codes/services
- Preview (sous quelle forme ?)
- Navigation, fouille, analyse (ex : SkyServer)
- Accès direct, récupération (téléchargement ?)

Problème

- Résultats (SimDM Dataset) souvent TRES volumineux
 - Ex : Cosmologie

Comment faire ?

Existant non adapté

- Bases de données relationnelles (transactions)
- Conçu pour des utilisations spécifiques (finance)
- Datawarehouse ? Model étoile relationnel
- Clusters ? Ne fait que repousser les limites du relationnel
- Bases objets ? Orientées colonnes ? Le problème est déplacé

Et Pourtant

SkyServer fonctionne avec SqlServer (cluster),
maintenant avec MonetDB (orienté colonne).

Est-ce encore extensible ?

Loi de Moore dirait oui

Problème

- Certains facteurs ne croissent pas automatiquement avec l'évolution des technologies
 - Manpower nécessaire
 - Disk i/o
 - Net bandwidth

Constats

- Pas besoin de transactionnel
- Il faut distinguer données brutes en sortie d'une expérience et résultats extraits par le scientifique
- Besoin d'un preview (à différentes échelles)
- La science est faite en fouillant les résultats en ligne

Solutions à l'étude

- Étendre au maximum les capacités du relationnel lorsque c'est suffisant
 - bases plates (dénormalisation)
 - cluster, segmentation
- Nouvelles approches
 - Machine learning objets d'intéret et comportement utilisateur sur les ensembles de données
 - Indexation sémantique

Etude de cas

- Deuvo 2010
 - Simulations cosmologie
 - 4 000 000 halos
 - 140 milliards particules
- Grain décrit par SimDM : cube ramses (= stats sur les halos)
- Grain halos stocké dans des tables plates séparées (4 000 000 lignes) (stats sur les particules)
- Grain particules en fichier binaire téléchargeable (cutout possible)

Besoin urgent d'une nouvelle approche

- Deuvo 2012
 - 200 000 000 halos.....

A term

- Est il pertinent de conserver TOUTES les données d'une simulation ?
 - Reproductible à moindre coût dans quelques années (mois) à partir des métadonnées
 - <> données observationnelles
- Approche “run à la demande” si loin que ça ?
 - Devons nous penser les standards dès maintenant ?
 - PDL