

Tâches de service & Big Data

Franck Le Petit



Réforme SO5

Avant la réforme :

- Souvent des services sous-critiques
- Pas de stratégie globale de développements de service

Objectifs de la réforme SO5:

- Remise à plat de la liste des services de référence
 - Centres d'expertise régionaux : assurent les masses critiques / moyens
 - Pôles thématiques nationaux : coordination nationale des services
- > définition de stratégies d'évolution des services

Spécificités des services de référence en SO5 :

- pérennisation des données
- qualité des données (sensibilité, précision ...)
- documentation des données
- interopérabilité des données
- exploitations innovantes / plus-value

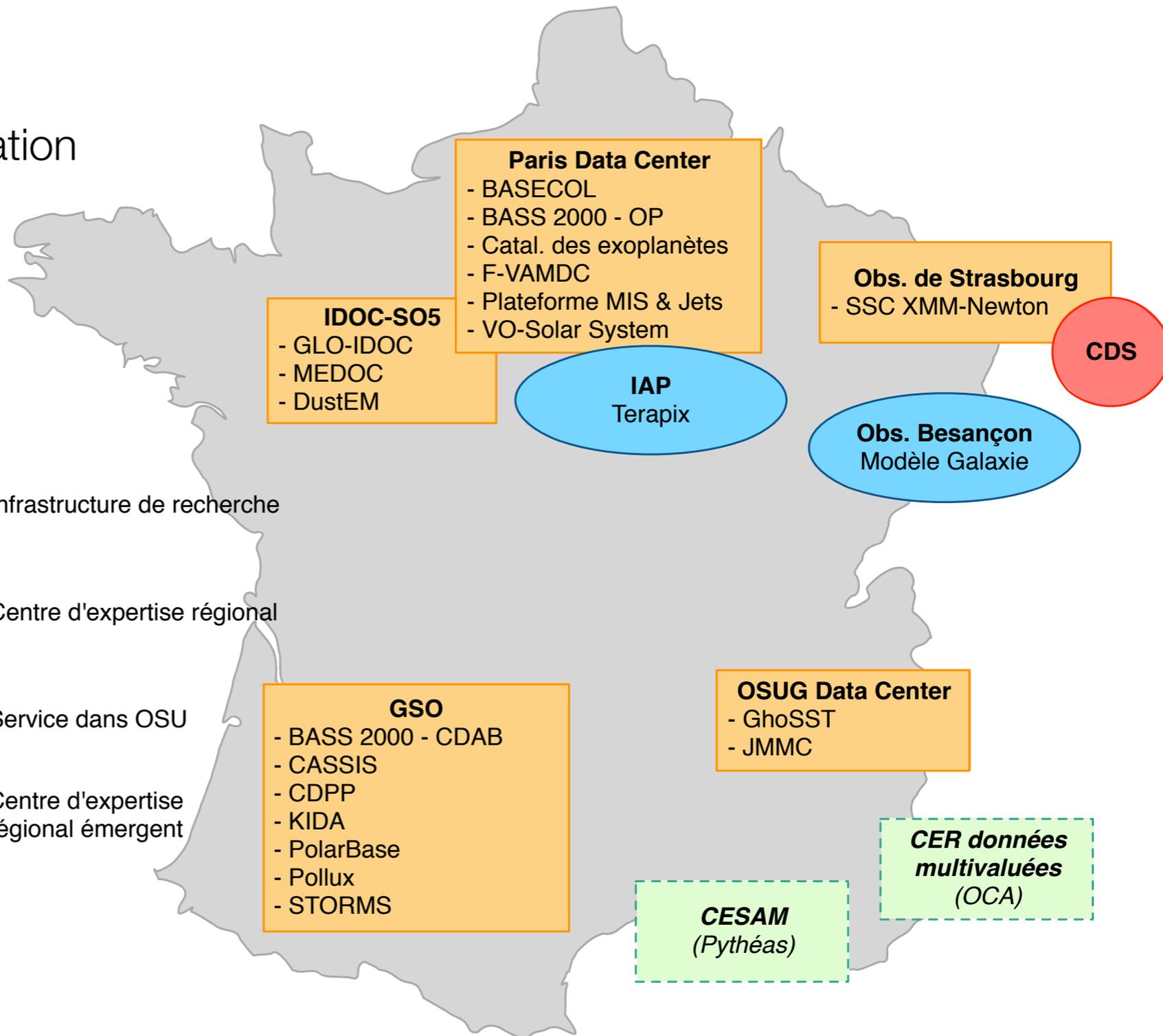
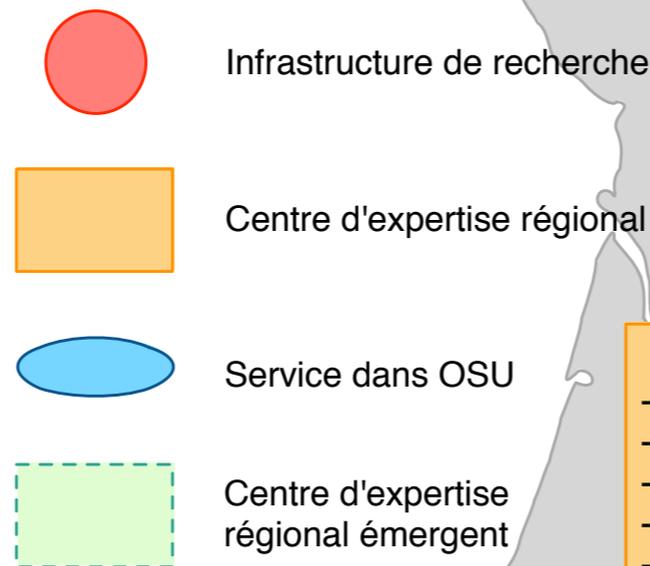
Réforme SO5

2012 : ~ 60 demandes de labellisation

2014 : 21 services de référence

Procédure de labellisation:

- Avis des PN & AS
- Evaluation comité
- Proposition à la CSAA



Pôles thématiques nationaux

- Pôle de diffusion des données de physique atomiques et moléculaires
- Pôle de diffusion de données plasmas planétaires
- Pôle de diffusion des données solaires
- Pôle de traitement des données interférométriques IR/Visible

Réforme SO5

Recommandations pour tous les services :

- Intégrés dans les Observatoires Virtuels
- Comités d'utilisateurs

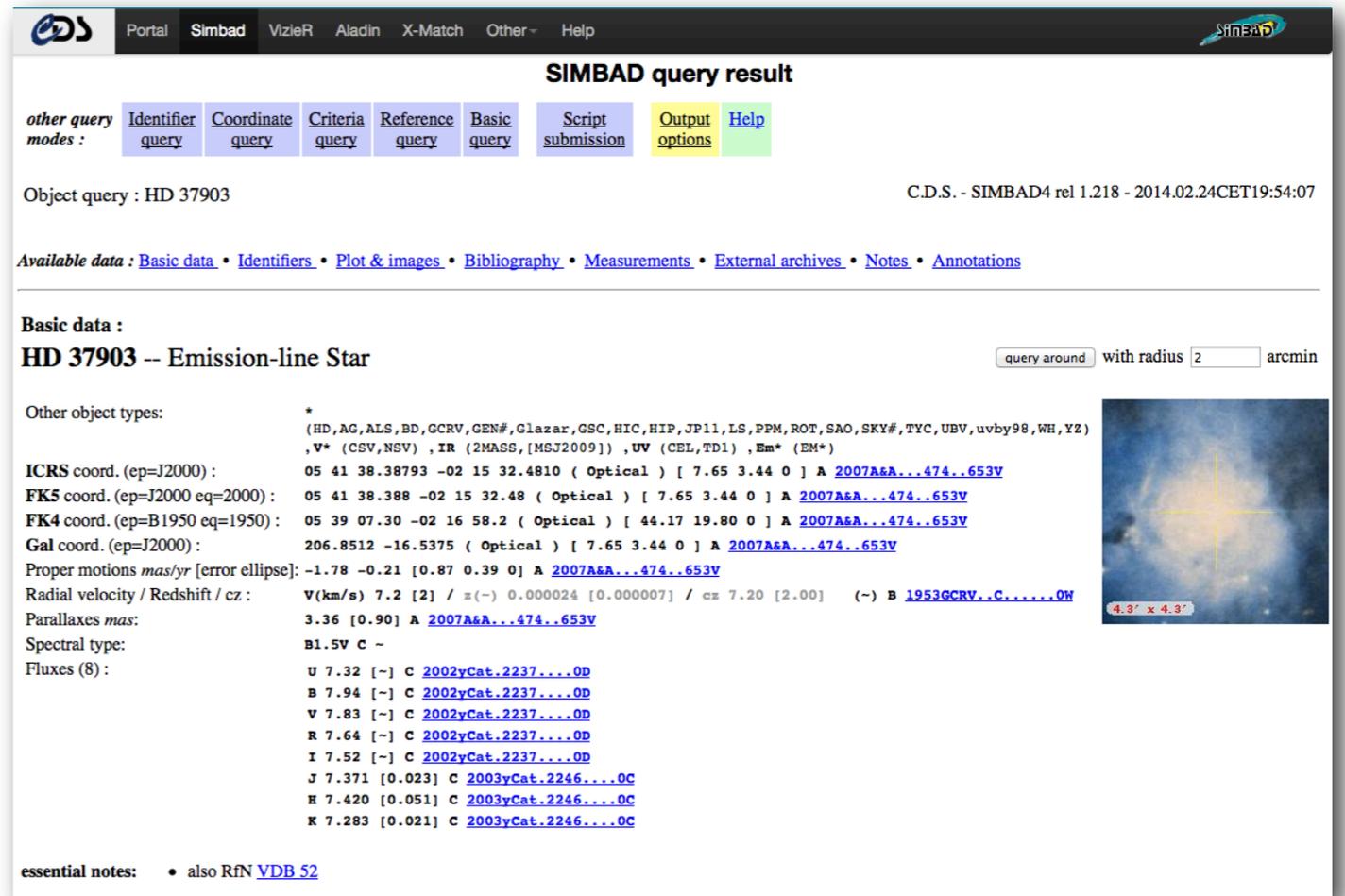
Rôle important des PN

- procédure de labellisation
- stratégie d'évolution des services :
 - comités utilisateurs
 - Pôles thématiques

Services liés au PNPS

- CDS
- JMMC

- Modèle de la Galaxie de Besançon
- POLLUX
- PolarBase
- Plateforme MIS & Jets
- VAMDC



SIMBAD query result

Object query : HD 37903 C.D.S. - SIMBAD4 rel 1.218 - 2014.02.24CET19:54:07

Available data : [Basic data](#) • [Identifiers](#) • [Plot & images](#) • [Bibliography](#) • [Measurements](#) • [External archives](#) • [Notes](#) • [Annotations](#)

Basic data :
HD 37903 -- Emission-line Star query around with radius arcmin

Other object types: *

ICRS coord. (ep=J2000): 05 41 38.38793 -02 15 32.4810 (Optical) [7.65 3.44 0] A [2007A&A...474..653V](#)

FK5 coord. (ep=J2000 eq=2000): 05 41 38.388 -02 15 32.48 (Optical) [7.65 3.44 0] A [2007A&A...474..653V](#)

FK4 coord. (ep=B1950 eq=1950): 05 39 07.30 -02 16 58.2 (Optical) [44.17 19.80 0] A [2007A&A...474..653V](#)

Gal coord. (ep=J2000): 206.8512 -16.5375 (Optical) [7.65 3.44 0] A [2007A&A...474..653V](#)

Proper motions mas/yr [error ellipse]: -1.78 -0.21 [0.87 0.39 0] A [2007A&A...474..653V](#)

Radial velocity / Redshift / cz: V(km/s) 7.2 [2] / z(-) 0.000024 [0.000007] / cz 7.20 [2.00] (-) B [1953GCRV...C.....0W](#)

Parallax mas: 3.36 [0.90] A [2007A&A...474..653V](#)

Spectral type: B1.5V C -

Fluxes (8):

U	7.32	[-]	C	2002yCat.2237....0D
B	7.94	[-]	C	2002yCat.2237....0D
V	7.83	[-]	C	2002yCat.2237....0D
R	7.64	[-]	C	2002yCat.2237....0D
I	7.52	[-]	C	2002yCat.2237....0D
J	7.371	[0.023]	C	2003yCat.2246....0C
H	7.420	[0.051]	C	2003yCat.2246....0C
K	7.283	[0.021]	C	2003yCat.2246....0C

essential notes: • also RfN [VDB 52](#)

Pollux database

<http://www.pollux.graal.univ-montp2.fr>

Spectres synthétiques stellaires

Codes MARCS, ATLAS9, CMFGEN

- Spectres HR : $R > 150\,000$
- SED $R \sim 20000$
- Etoiles O à M

Paramètres couverts :

Teff, log g, Fe / H, ...

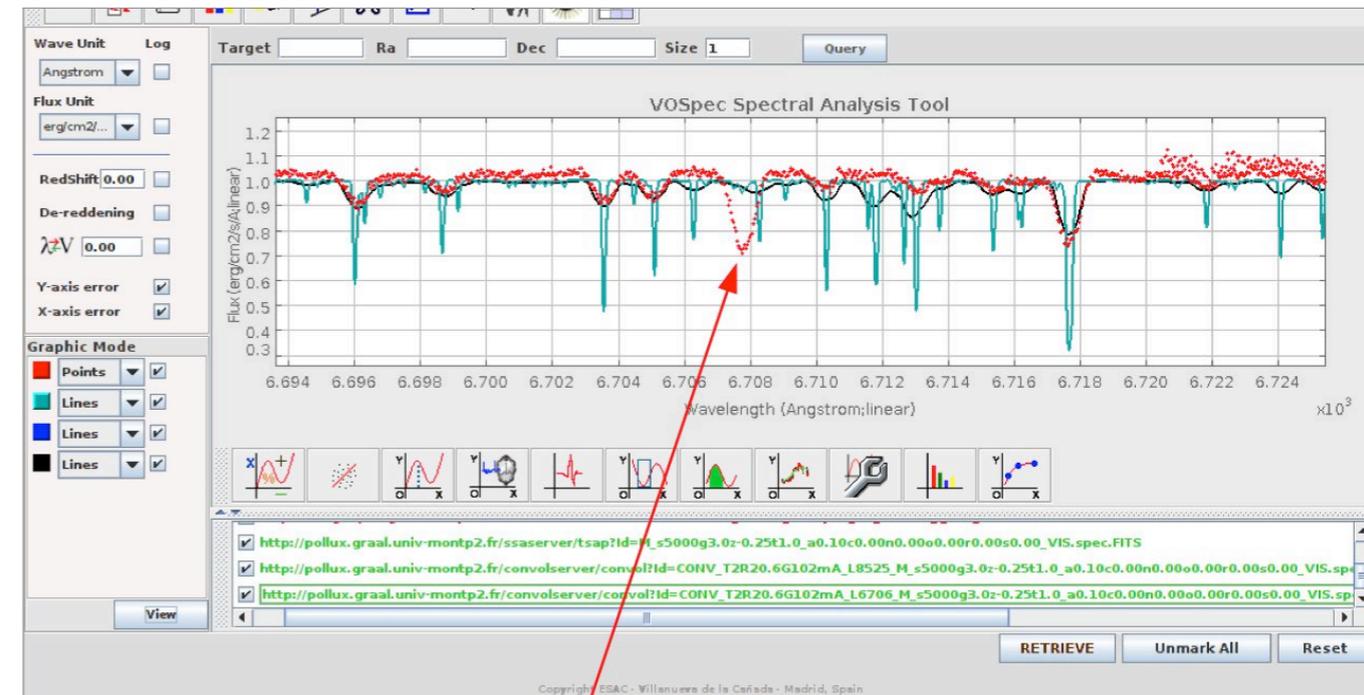
Services :

- Accessible via l'OV
- Service de convolution
VOSPECFLOW
Comparaison aux données NARVAL

The screenshot shows the 'Query Form' interface of the Pollux Database. It features a navigation menu with 'Home', 'User's Guide', and 'Contact'. The main content area is divided into several sections:

- General Parameters Information:** Includes 'Type of data' (SSHR & SED), 'issued from Model Atmosphere' (MARCS & CMFGEN & ATLAS), and 'Type of Model Atmosphere' (parallel & spherical). A search button is present.
- Select Spectra:** A tree view showing folders for 'SSHR', 'SED', 'SSHR & SED', 'MARCS', 'CMFGEN', 'CMFGEN-WR', 'ATLAS', and 'MARCS & CMFGEN & ATLAS'. The latter is expanded to show 'parallel', 'spherical', and 'parallel & spherical' options.
- Spectrum Parameters (at least one recommended):** A table with columns for parameter names, 'lowest', 'low/equal', 'high', and 'highest'. Parameters include effective temperature (K), gravity log10 (cgs), mass (solar mass), luminosity (log10 of solar luminosity), microturbulent velocity (km/s), and metallicity ([Fe/H]).
- Specific Abundances (optional):** A table with columns for parameter names, 'lowest', 'low/equal', 'high', and 'highest'. Parameters include alpha elements [alpha/Fe], Carbon [C/Fe], Oxygen [O/Fe], Nitrogen [N/Fe], r process elements [r elements/Fe], and s process elements [s elements/Fe].
- Cart Status:** A section indicating 'No spectra to be downloaded'.

Logos for LUPM (Laboratoire Montpellier Univers & Particules), um2 (Université Montpellier 2), and CNRS INSU (Observer & comprendre) are visible at the bottom.



Modèle de la Galaxie

<http://model.obs-besancon.fr>

Service fournissant des simulations :

- du contenu stellaire de la Galaxie
- d'images du ciel de l'UV à l'infrarouge
- de catalogues
- de grands relevés
- de microlentilles gravitationnelles



Service basé sur :

- Synthèse de populations stellaires
- Modèle de l'extinction interstellaire
- Simulateur d'évènements de microlentilles gravitationnelles

Model of stellar population synthesis of the Galaxy
Catalogue simulation without kinematics, Johnson-Cousins photometric system

To get help on parameters and values to supply, click on [?](#)

Field of view :

- Distance interval (kpc) [?](#)
[0.000000 , 50.000000]
Distance step [?](#): mode
- specify step value (in parsecs) if linear mode or $\Delta r/r$ if logarithmic mode:
- field:
 - small field** [?](#) (defined by the center of the field and its surface) :
Longitude : Latitude : Solid angle (deg²) :
 - large field** [?](#) (field defined by galactic longitude and latitude):
Coordinate system:
 - If equatorial coordinates, specify equinox:
- Galactic longitude or right ascension (decimal degrees):
minimum maximum step
[,]
- Galactic latitude or declination (decimal degrees):
[,]

Extinction law

[?](#)

Diffuse extinction:
 mag/kpc
(A mean diffuse absorption of 0.7 mag/kpc is recommended for intermediate and high latitude fields. It may be modified.)

Discrete clouds:

cloud n°	absorption (A_V)	distance (pc)	cloud n°	absorption (A_V)	distance (pc)	cloud n°	absorption (A_V)	distance (pc)
1:	<input type="text"/>	<input type="text"/>	2:	<input type="text"/>	<input type="text"/>	3:	<input type="text"/>	<input type="text"/>
4:	<input type="text"/>	<input type="text"/>	5:	<input type="text"/>	<input type="text"/>	6:	<input type="text"/>	<input type="text"/>
7:	<input type="text"/>	<input type="text"/>	8:	<input type="text"/>	<input type="text"/>	9:	<input type="text"/>	<input type="text"/>
10:	<input type="text"/>	<input type="text"/>	11:	<input type="text"/>	<input type="text"/>	12:	<input type="text"/>	<input type="text"/>
13:	<input type="text"/>	<input type="text"/>	14:	<input type="text"/>	<input type="text"/>	15:	<input type="text"/>	<input type="text"/>
16:	<input type="text"/>	<input type="text"/>	17:	<input type="text"/>	<input type="text"/>	18:	<input type="text"/>	<input type="text"/>
19:	<input type="text"/>	<input type="text"/>	20:	<input type="text"/>	<input type="text"/>	21:	<input type="text"/>	<input type="text"/>
22:	<input type="text"/>	<input type="text"/>	23:	<input type="text"/>	<input type="text"/>	24:	<input type="text"/>	<input type="text"/>
25:	<input type="text"/>	<input type="text"/>						

Polarbase

<http://polarbase.irap.omp.eu>

Diffusion des données ESPADOnS, NARVAL

- données des spectropolarimètres INSU

Domaine : 370 - 1000 nm

Polarimétrie

Accès aux données réduites

ASCII / FITS

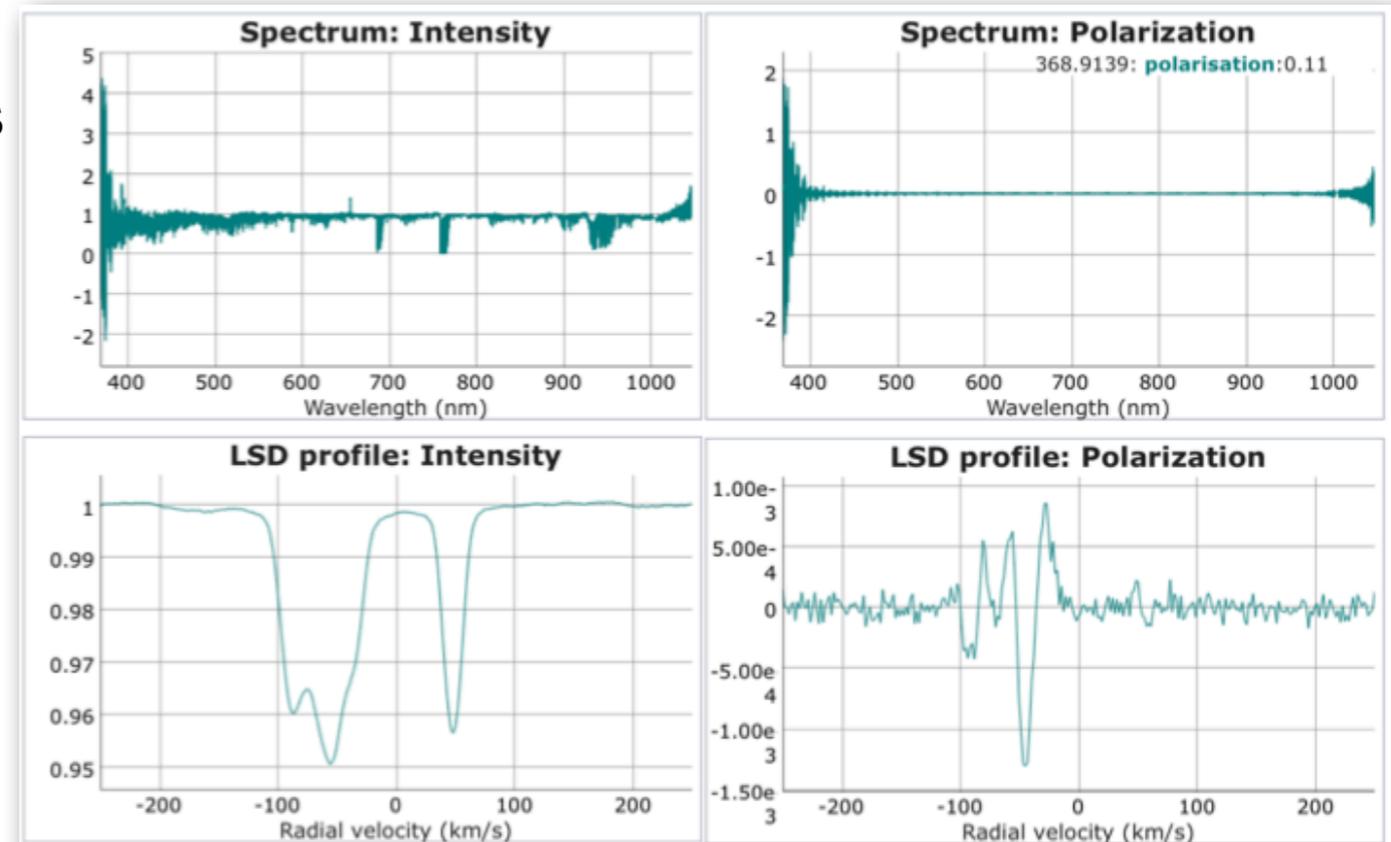
Intégré à l'OV - manque descrip. polarimétrie

- association de données complémentaires

- champ magnétique de surface
- émission chromosphérique
- vitesse radiale, ...

The screenshot displays the Polarbase web interface. On the left, there are search filters for Object Parameters (Name, RA, DEC, V Magnitude, Spectral Type) and Observation Parameters (Instrument: Narval, Espadons; Observation date, time, HJD, GJD, SNR, Obs. mode, Stokes Param, Airmass, Hour angle, LSD results). On the right, the 'Query Results' section shows a table of 10 entries for the target V*V711Tau. The table columns are Sel., Plot, Target, Inst. mode, UT date, Instrument, Stokes, and Bloc. Below the table, there are options to download selected spectra or the table as an ASCII file, and a 'Retrieve Wget file for all results' button.

Sel.	Plot	Target	Inst. mode	UT date	Instrument	Stokes	Bloc
<input type="checkbox"/>		V*V711Tau	pol	2005/12/14	espadons	I	3168
<input type="checkbox"/>		V*V711Tau	pol	2005/12/14	espadons	I	3169
<input type="checkbox"/>		V*V711Tau	pol	2005/12/14	espadons	V	542
<input type="checkbox"/>		V*V711Tau	pol	2005/12/14	espadons	I	3170
<input type="checkbox"/>		V*V711Tau	pol	2005/12/14	espadons	I	3171
<input type="checkbox"/>		V*V711Tau	pol	2005/12/16	espadons	I	3375
<input type="checkbox"/>		V*V711Tau	pol	2005/12/16	espadons	I	3376
<input type="checkbox"/>		V*V711Tau	pol	2005/12/16	espadons	V	574
<input type="checkbox"/>		V*V711Tau	pol	2005/12/16	espadons	I	3377
<input type="checkbox"/>		V*V711Tau	pol	2005/12/16	espadons	I	3378



Plateforme MIS & Jets

<http://ism.obspm.fr>

Services & diffusion de modèles de références et de simulations lourdes pour l'interprétation des observations dans les nuages et les jets

STARFORMAT (P. Hennebelle et al.)

Diffusion de simulations MHD

Dynamique du gaz et formation des étoiles

Accès à plusieurs jeux de simulations

- Coeurs denses

Services fournis :

- Catalogues de propriétés de coeurs denses
- Téléchargement de simulations
- Coupes : cubes, lignes de visée, ...
- Post-traitement RADMC-3D

Evolution du service - nouveaux projets

- Simulations de formation stellaire (B. Commerçon)
- Jets (A. Ciardi, C. Stehlé)
- Disques

ISMServices
Numerical Services for Interstellar and Extragalactic Medium

CODES & DATABASES
access to services

TECHNOLOGIES
standards

PARTNERS
credits

PDR Code
The Meudon PDR code

Starformat
MHD simulations data base

Shock
Paris-Durham Shock model

TDR Code
Turbulence Dissipation Regions

The Interstellar Medium and Jets Platform gathers a set of state-of-the-art numerical services to prepare and interpret observations in the interstellar medium and in astrophysical jets. Our services give access to some of our numerical codes, to databases of pre-computed numerical simulations and to tools to analyze results. These services are developed and maintained by scientists and software engineers of Paris Observatory / VO-Paris Data Center and CEA.

STARFORMAT

To query the models, select first a code version and then choose at least a search criteria:

Formation of molecular clouds in a small box with filaments

Code description

The aim of this run is to study the formation of molecular clouds from the warm atomic neutral medium (reference Hennebelle et al. 143 A&A 488, 2006). Starting the simulation with WNM only, a converging flow is imposed from the left and from the right. The converging flow has a velocity equal to five times the sound speed of the WNM on top of which fluctuations have been superimposed. The magnetic field is initially uniform. The simulation includes atomic cooling and gravity. After few Myrs, dense gas develops and eventually collapses. The run has been performed with the FRODOSS-MHD code (Heylser 2022, A&ASP, A, 388, 2021, Fromang et al., A&ASP, A, 457, 2021). This is a mesh refinement code, implying that it can increase locally the spatial resolution by adding new cells in the computation. It uses the Godunov method and constant transport method to maintain the divergence of the magnetic field equal to zero.

Query on experiment parameters

Select at least one criteria on parameters:

Parameter	Parameter values	Unit
Magnetic Field - X Boundary	1.0	
Magnetic Field - Y Boundary	0.0	
Magnetic Field - Z Boundary	0.0	
Velocity of the incoming flow	13.34792, 17.791	
Lowest AMR level	7.0	
Highest AMR level	10.0	
Initial density within the box	1.0	
Modulation of the incoming flow	0.5, 1.0	

Structure size distribution

Created: Thu Sep 10 18:48:11 CEST 2009
Updated: Thu Sep 10 19:20:36 CEST 2009
Status: published

Snapshots available

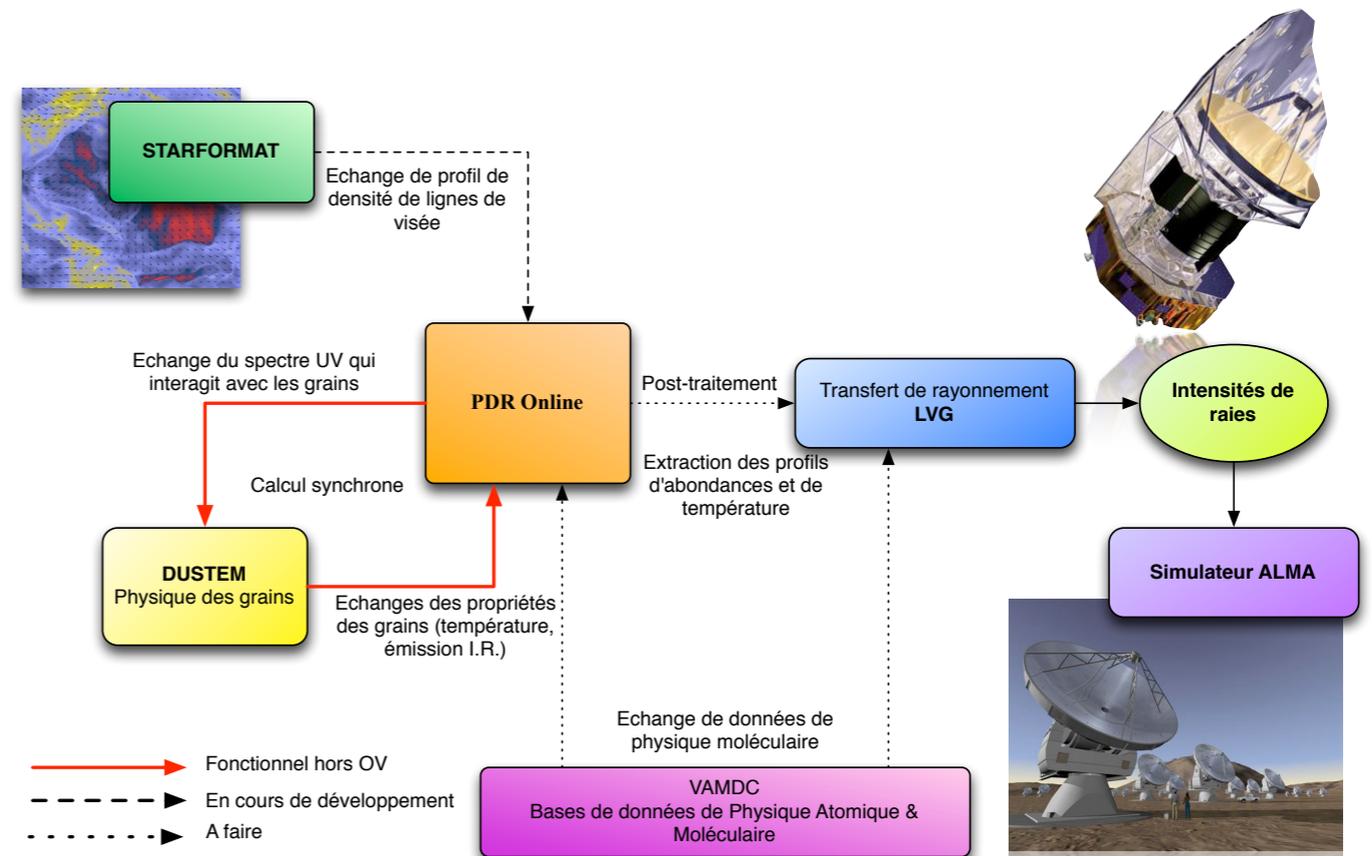
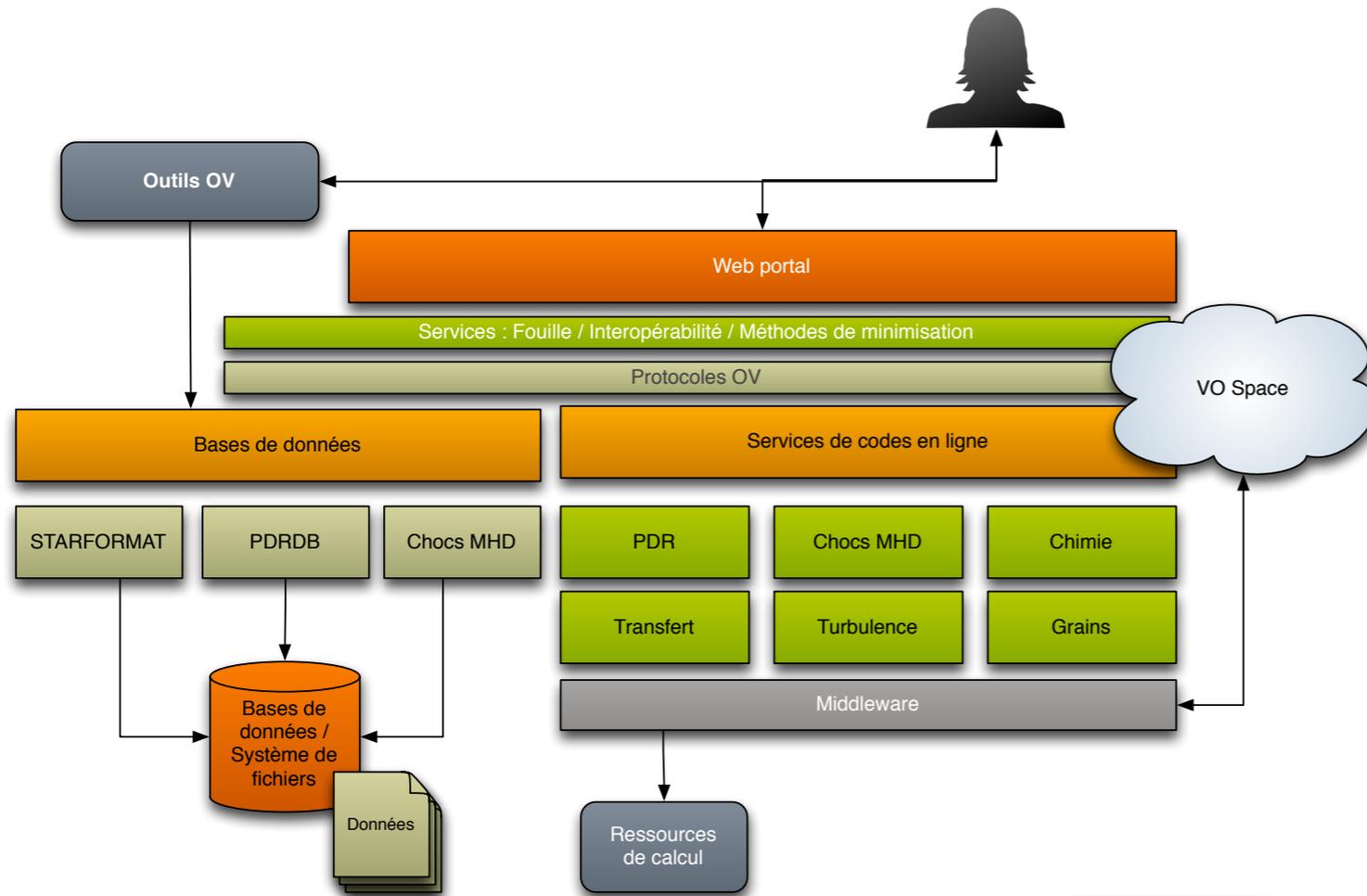
8.47702Myrs 10.44202Myrs

t=10.42 (Myrs)

Structure size (pc)

→ v = 23.7189 (km/s)

Plateforme MIS & Jets



Données de physique atomique & moléculaires

VAMDC : Virtual Atomic and Molecular Data Center

Infrastructure pour faciliter :

- la découverte de jeux de données atomiques & moléculaires
- l'accès aux données

Principales bases de données européennes participant :

VALD, HITRAN, CHIANTI, TIP Base, TOP Base, CDMS, ...

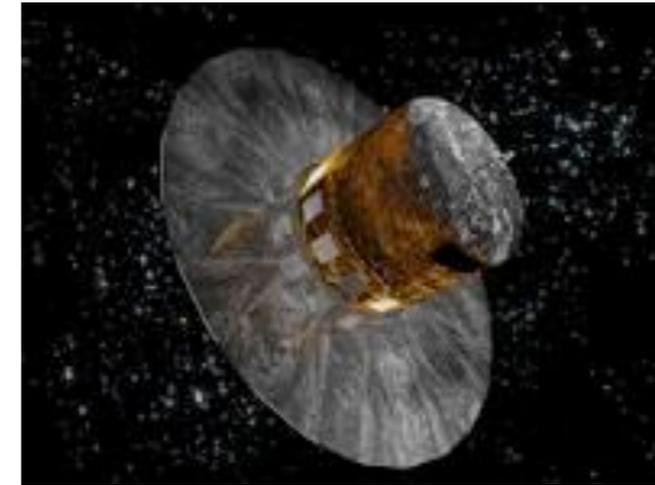
The screenshot displays the VAMDC website interface. At the top, there is a navigation bar with links: Home, VAMDC databases, Query, Saved queries, Info, Known issues, Feedback, Login, and Register. On the left side, there is a 'Query by...' menu with options: Species, Processes, Environment, and Advanced. The main content area features a query form titled 'Atom 1' with fields for Atom symbol, Mass number, Nuclear charge, Ion charge, InChIKey, State energy, and Equivalent to. The State energy field includes a unit dropdown menu set to '1/cm'. To the right of the query form, there are buttons for 'Find data' and 'Save query'. Below these buttons is a 'Legend' section with color-coded text: green for 'available, can answer', yellow for 'available, don't support query', and red for 'unsupported keyword'. A list of databases follows, each with a folder icon and a name: Chianti, GSMA Reims S&MPO, ECaSDa - Ethene Calculated Spectroscopic Database, GhoSST, Stark-b, JPL database: VAMDC-TAP service, HITRAN-UCL resource, VALD sub-set in Moscow (obs), RADAM - Ion Interactions, ALADDIN2, Cologne Database for Molecular Spectroscopy: VAMDC-TAP service, MeCaSDa - Methane Calculated Spectroscopic Database, VALD (atoms), VAMDC species-DB, OACT - LASP Database, TOPbase : VAMDC-TAP interface, DESIRE database (Moscow mirror), BASECOL: VAMDC-TAP interface, UMIST Database for Astrochemistry, IDEADB - Innsbruck Dissociative Electron Attachment Database, and TIPbase : VAMDC-TAP interface.

Big Data

Données de plus en plus volumineuses

Gaïa

- caractérisation d'un milliard d'étoiles
 - type spectral
 - masse, rayon, luminosité, température
 - composition chimique
 - ...



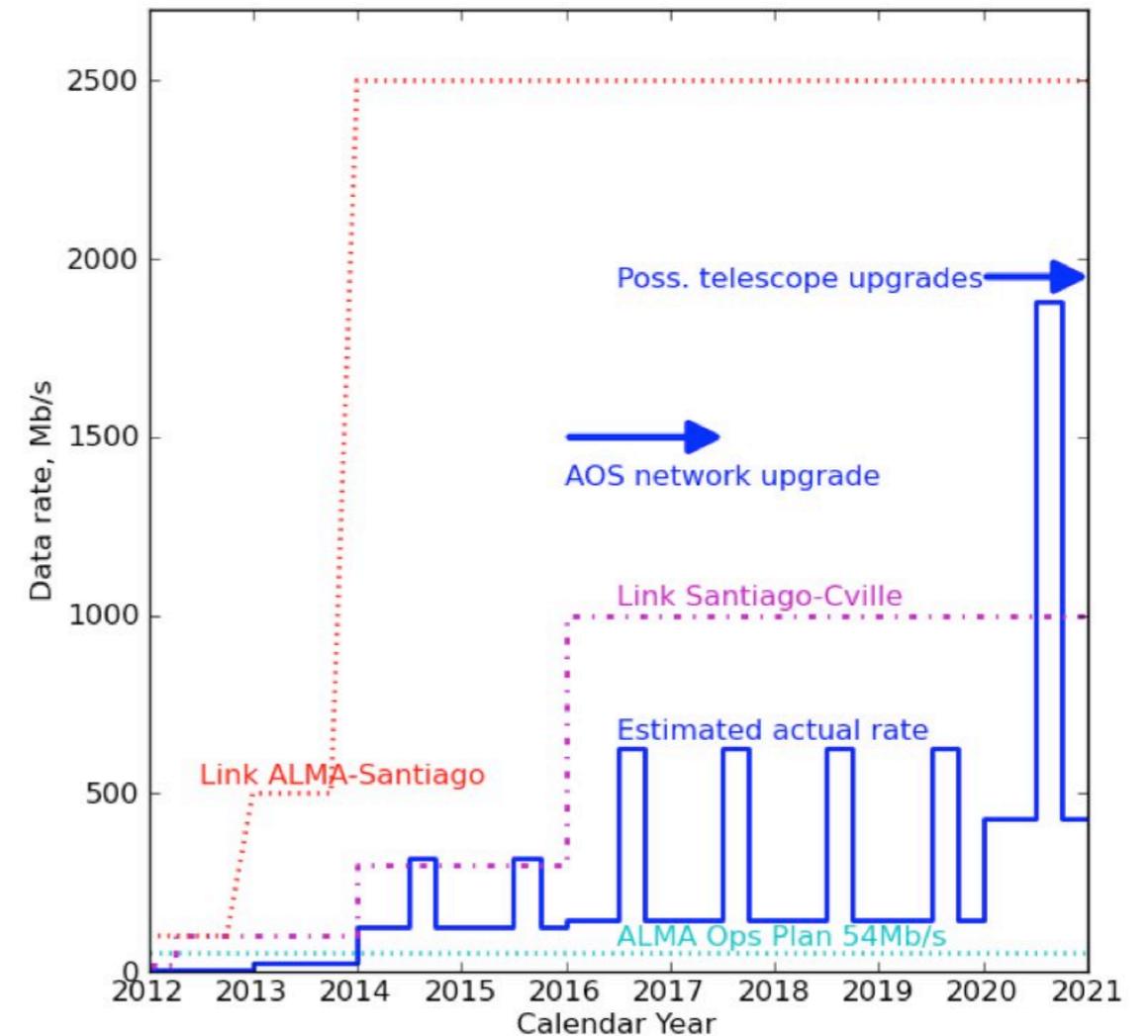
ALMA - Centres d'archivage :

- ~ 1PB actuellement
- ~ 3 PB / yr (2015-2018)
- => problèmes stockages / électriques

Comment stocker ces données ?

Comment les manipuler ?

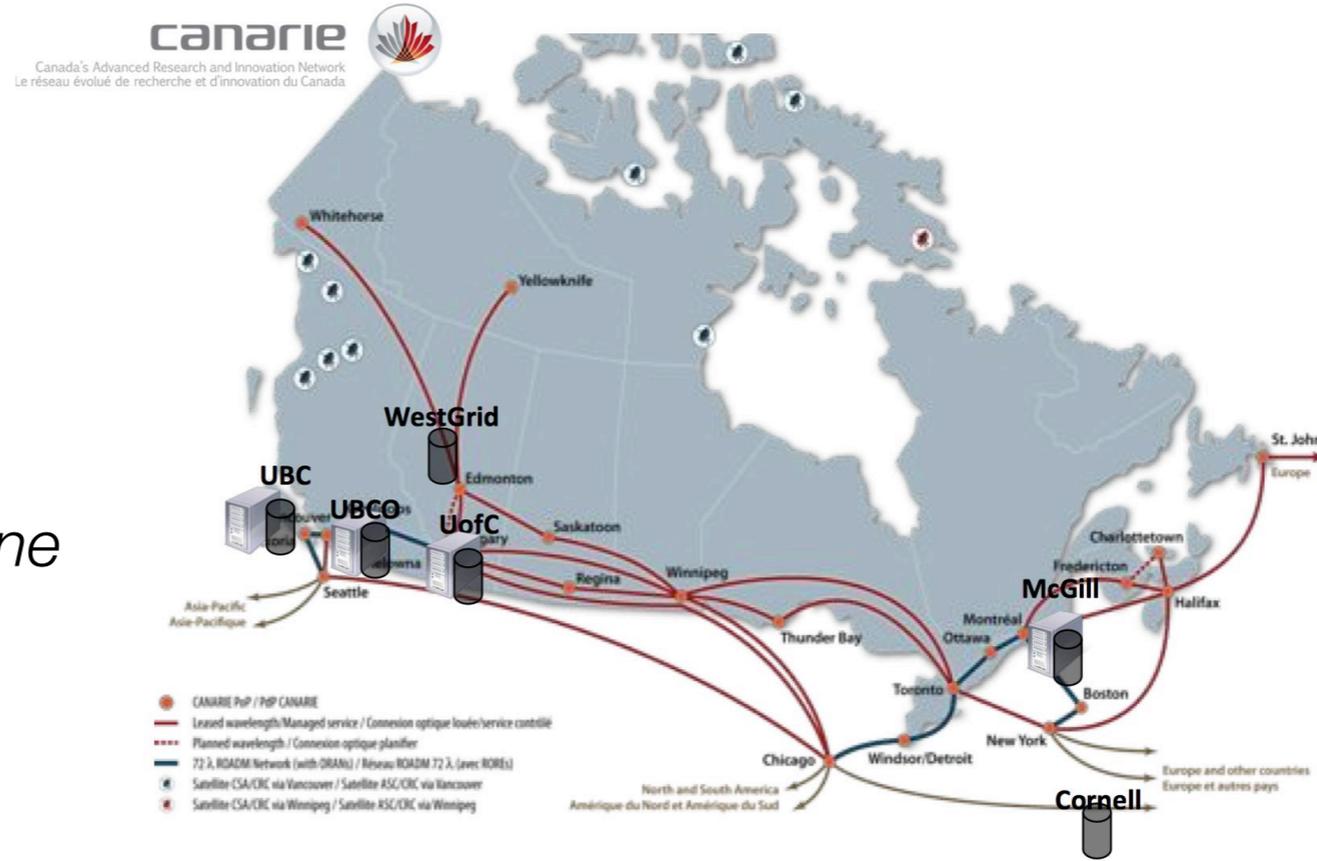
Comment les diffuser ?



Big Data

Exemple de CyberSKA Analyse des données SKA

- système distribué
- réplication des données
- outils collaboratifs
- visualisation des données *online*



Big Data

Problèmes posés par les volumes de données au niveau des services:

- Stockage :

- Besoin d'infrastructures de stockage de plus en plus coûteuses
- Centres de données AA / *Centres de calcul nationaux*
- Centres d'expertise SO5 / *Mesocentres*

- Transfert des données :

Un combat perdu d'avance si approche brute:

- Réseau internet : 10 Gbytes / s -> 100 Gbytes / s (RENATER)
- Goulet d'étranglement pour arriver aux ordinateurs

Solutions :

- Si peu d'utilisateurs intéressés par les grands volumes :
Ex: grands jeux de simulations très spécialisées

- Données fréquemment sollicitées :

=> Développer des méthodes innovantes pour l'extraction et le datamining



Big Data

Solutions pour transférer les volumes de données

Exemple d'ALMA :

Cube de données :

- 1 - 100 Gby
- tout le cube n'est pas pertinent

Système permettant de

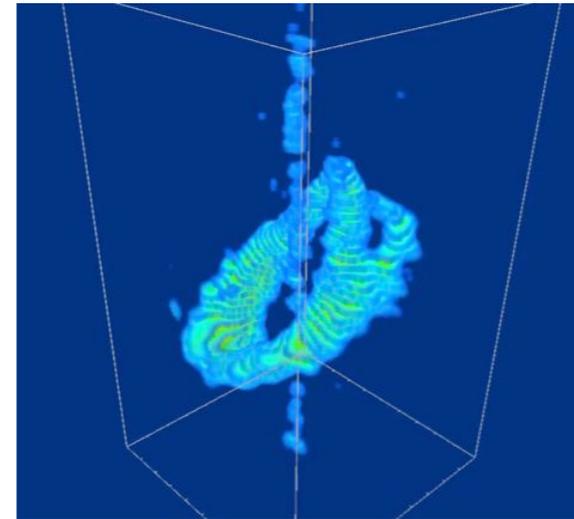
- 1) sélectionner une portion du cube
- 2) récupérer un sous-ensemble

Problème similaire :

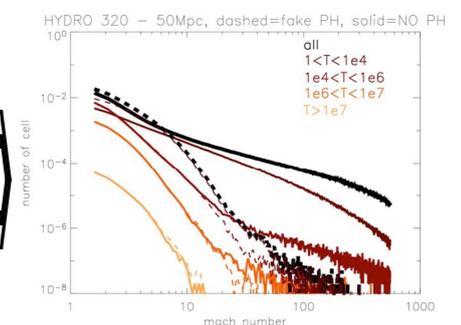
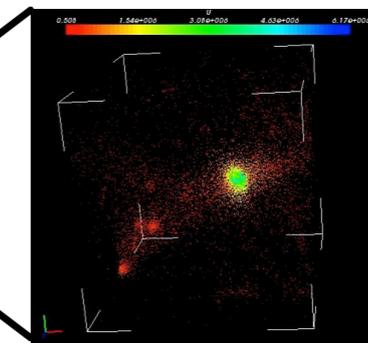
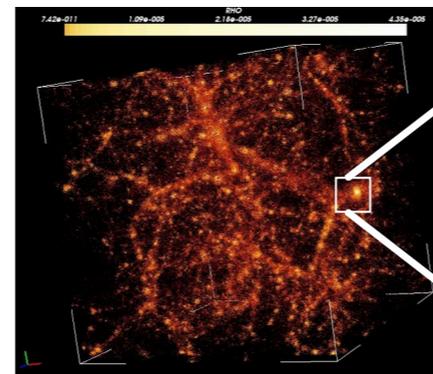
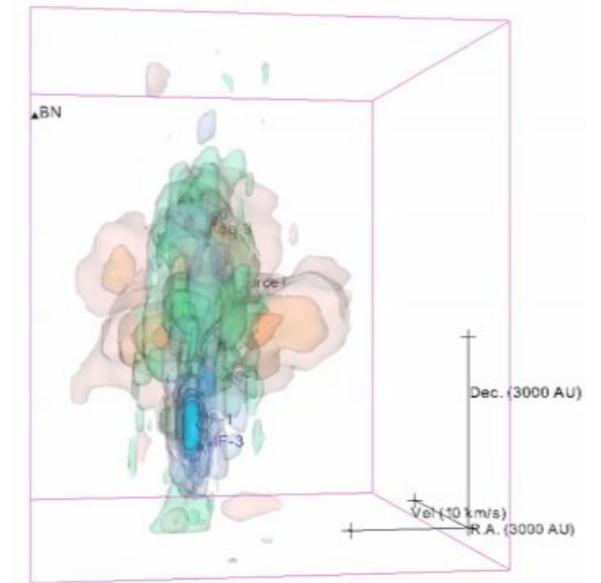
- extraire des informations d'un catalogue
- extraire un cube dans une simulation

Solutions :

- * Envoi progressif des données : imageries / type Google Map / pagination
- * Extraction & transfert : cutout (travaux en cours à l'IVOA)



B. Glendenning



La difficulté n'est pas tant le volume des données ...
... que de trouver ce qu'on veut dedans

Big Data

Exploitation des Big Data implique souvent de nombreuses méta-données

Avant le Big Data

Fouille simples selon des paramètres classiques

Exemple recherche de spectres :

* Spectre HST/STIS de HD 34078

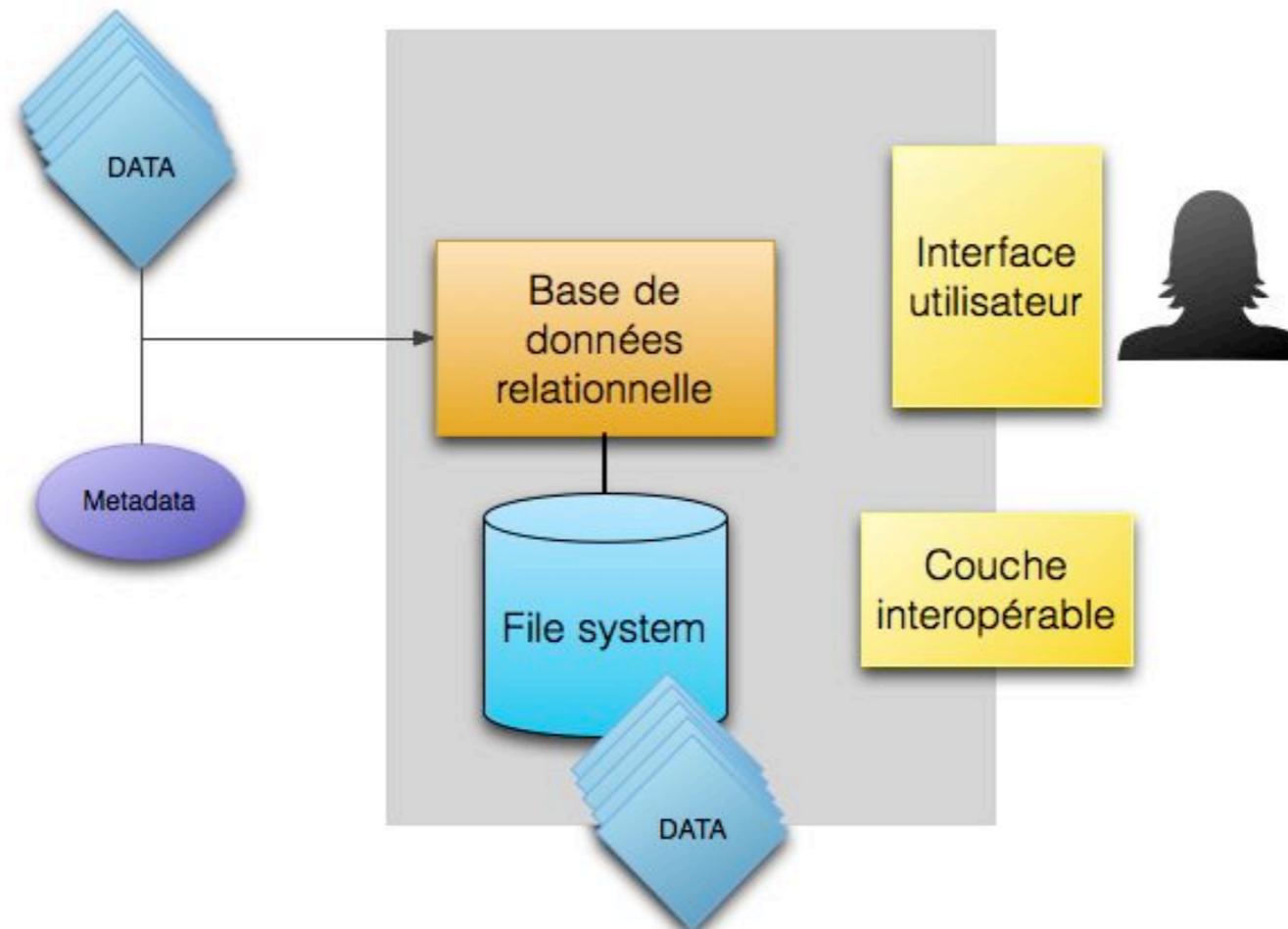
* Modèle stellaire O 9

Système informatique :

Filesystem stockant ~ qq To de données

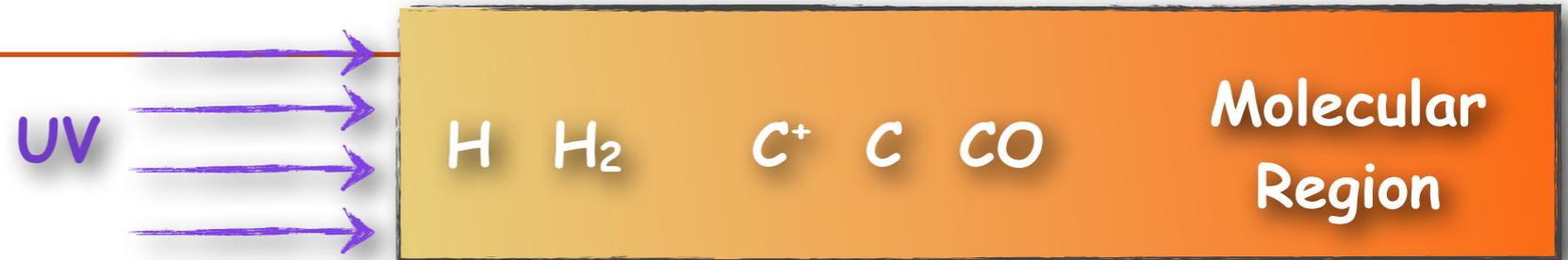
Base de données gérant ~ 10 métadonnées

- position dans le ciel
- instrument
- date
- ...



Big Data

Example : Services PDR



Fouille 1 : on recherche sur les paramètres d'entrée

Ex: $n = 1000 \text{ cm}^{-3}$, flux UV = 10 ISRF

Metadonnées = Paramètres d'entrée
~ 30 métadonnées

Fouille 2 : Problème inverse

quels sont les modèles tels que

$9\text{E-}8 < \text{intensité de CO (1-0)} < 2\text{E-}7 \text{ erg cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1}$

$8\text{E-}5 < \text{intensité de C+ à 158 microns} < 3\text{E-}4 \text{ erg cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1}$

Le cas d'utilisation 2 impose que certaines données soient traitées en tant que métadonnées
=> + 150 000 types de métadonnées

Problèmes techniques :

- les technologies informatiques de data management classiques ne fonctionnent plus !
=> bases de données relationnelles sont peu adaptées au grand nombre de métadonnées

Besoin de technologies informatiques moins / peu connues :

Ex : bases orientées objets / bases orientées document / MongoDB / ...
+ algorithmes de fouille adaptés au problème

Big Data

Big Data peut aussi nécessiter de repenser les interfaces utilisateurs

Formulaire de recherche PDR
150 000 quantités

Formulaire de recherche du NRAO

General Search Parameters :

Telescopes All Jansky VLA Historical VLA VLBA GBT

Project Code
GBT: Project Session
AGBT12A_055 JVA: 12A-256

Observer Name Archive File ID (partial strings allowed) To (2010-06-21 14:20:30)

Position Search :

Target Name Search Type SIMBAD or NED Min. Exposure (secs)

RA or Longitude (04h33m11.1s or 68.29d) DEC or Latitude (05d21'15.5" or 5.352d) Equinox J2000

Search Radius 1.0' (1d00'00" or 0.2d) - OR - Check for automatic VLA field-of-view, freq. dependent.??

Observing Configurations Search :

Telescope All A AB BnA B BC CnB
Config C CD DnC D DA

Sub_array All 1 2 3 4 5

Polarization ALL

Data Type ALL

Observing Bands All 4 P L S C
 X U K Ka Q W

Frequency Range
(In MHz : 1665.401 - 1720.500)

Receiver ID ALL

Backend ID ALL
(GBT only - select GBT in Telescopes list))

NRAO Archive, advanced search form: 23 search parameters (plus some output specification parameters)

Grid of models

Graphical Parameters Space Search

protocol (code) : pdr_1_5_2_rev826

x, y axis : input parameters of the code

x axis : Proton density (initial value)
ex : proton density

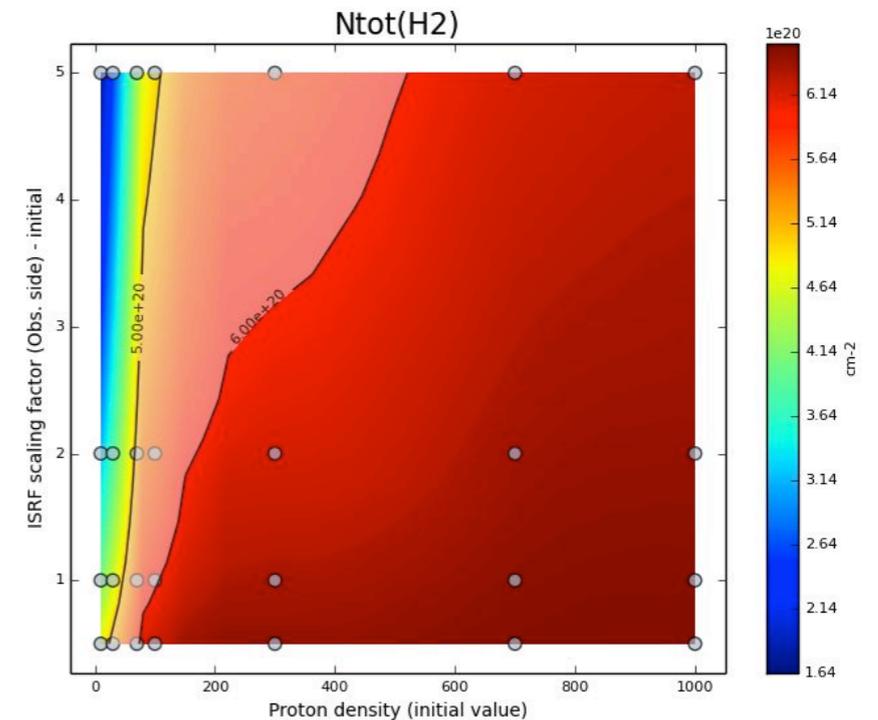
y axis : ISRF scaling factor (Obs. side)
ex : ISRF scaling factor

z axis : output calculated value of the code

z axis : Ntot(C+)
ex : Ntot(C+)

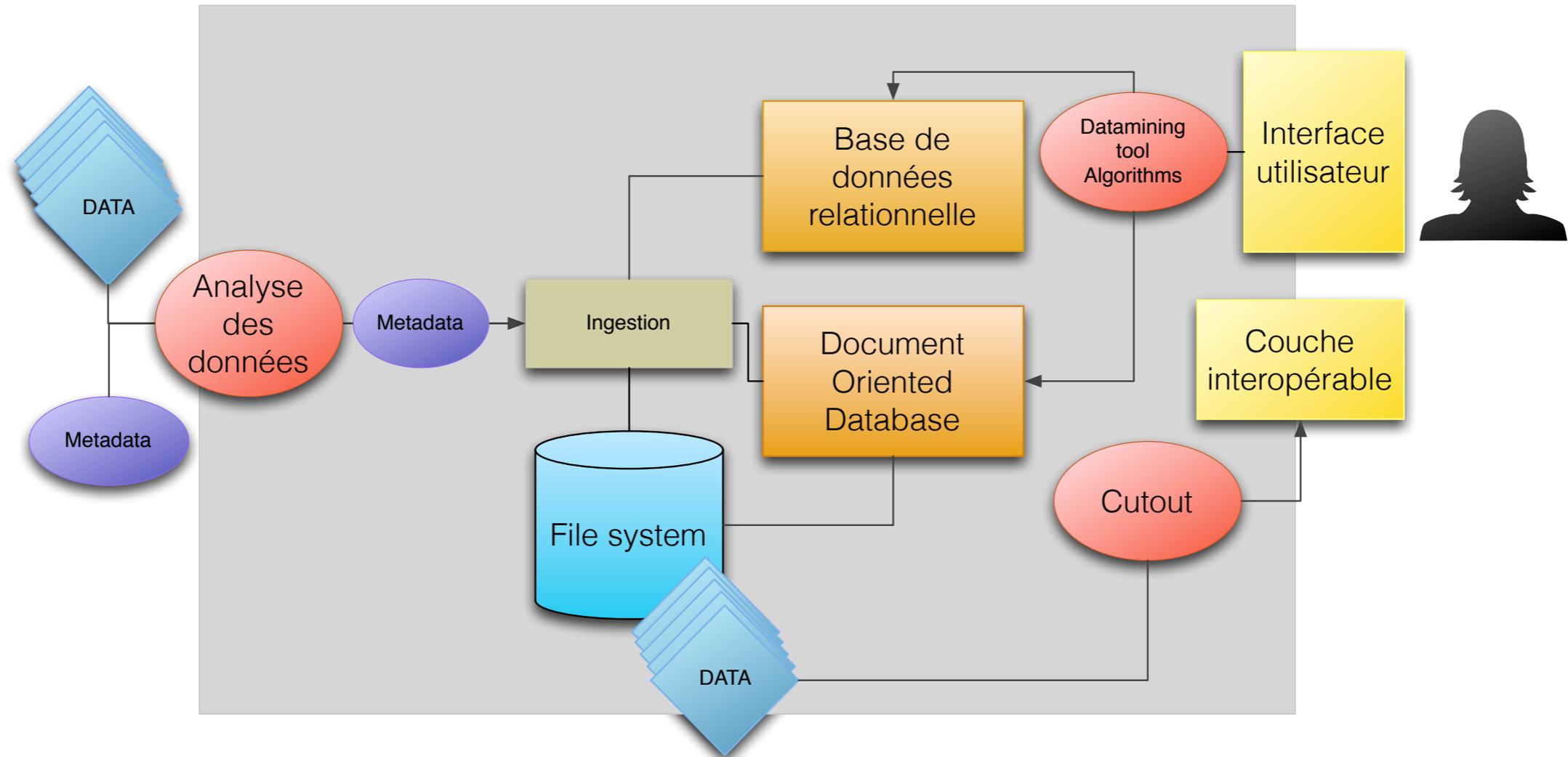
constraints on output calculated values of the code

constraints query: "Ntot(C+)" > 1.6e17 and "Ntot(H2)" < 6e20 and "Ntot(H2)" > 5l
ex : "Ntot(C+)" > 1.6e17 and "Ntot(H2)" < 6e20



«Barre Google»
+
web sémantique

Big Data



Infrastructure informatique plus complexe et spécifique à des cas d'utilisation

- Algorithmes d'analyse côté serveur : interprétation, analyse, visualisation
- Nouvelles technologies informatiques pour gérer les masses de méta-données
- Technologies de système distribué
- Algorithmes de fouille

=> coût humain plus important pour développer un service mais résultat beaucoup plus puissant

Big Data

Plus coûteux à mettre en oeuvre

Pas de solution universelle mais dépendante de

- du type de données
- de la façon dont on souhaite que les utilisateurs interagissent avec le système

Procédure itérative dans le développement

Développement tend à relever d'une R&D informatique à visée opérationnelle

Opportunités pour la recherche

Permet des services très évolués

- accès à des données riches
- fouille / services plus intelligents

Conclusion

Plusieurs services directement liés au PNPS

- CDS
- JMMC
- Pollux
- Polarbase
- Modèle de la Galaxie
- Plateforme MIS & Jets
- VAMDC + services de physique atomique

Définir une stratégie pour le développement des services SO5

- Rôle important des PN
 - Participent à la définition des besoins de la communauté
 - Suggestion sur l'organisation nationale
- => utile à faire apparaître dans la prospective

Big Data

Une opportunité pour la recherche

Plus difficile à mettre en place que les systèmes de diffusion *classiques*

Nécessite de **définir en amont les besoins** pour trouver des solutions

=> PN doivent jouer un rôle pour spécifier les besoins de leur communauté

Développement complexe :

- coordination / partage d'expertise entre les centres d'expertises / services