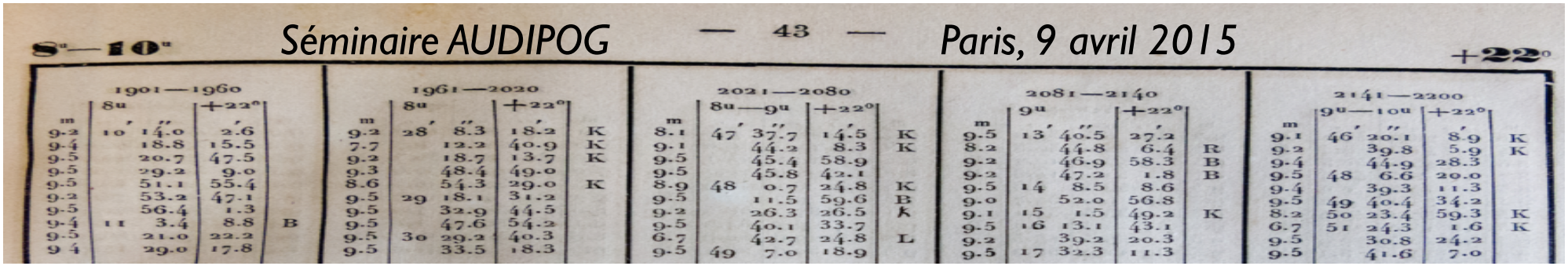


Vers le « Big Data », l'exemple de l'astronomie

André Schaaff

Observatoire astronomique de Strasbourg & Centre de Données astronomiques de Strasbourg

andre.schaaff@astro.unistra.fr

UNIVERSITÉ DE STRASBOURG

Observatoire astronomique de Strasbourg

CENTRE DE DONNÉES ASTRONOMIQUES DE STRASBOURG

cnrs INSU Observer & comprendre

Avant-propos

Une introduction au « Big Data », d'un point de vue global (l'avalanche de données) et d'un point de vue plus orienté fournisseur de données et de services.

L'exemple de l'astronomie avec un focus sur le Centre de Données astronomiques de Strasbourg.

Big Data ?

Big data, littéralement « **grosses données »**, parfois appelées *données massives*, est une expression anglophone utilisée pour désigner des ensembles de données qui deviennent tellement volumineux que leur exploitation avec des outils classiques de gestion de base de données ou de gestion de l'information peut s'avérer ardue.

Quelques expressions courantes:
L'avalanche de données
Surfer la vague des données
Les Masses de données
Etc.

Big Data, l'avalanche de données

Des données hétérogènes générées en quantité exponentielle par des entités de divers domaines (les sciences: astronomie, biologie, etc.; les services publiques en général, l'industrie, le commerce et la finance, etc..).

S'y ajoutent les données déjà existantes, dans un contexte Open Data (les données étant vues comme la mine d'or ou le puits de pétrole des décennies à venir… cf. par exemple les déclarations d'Obama aux USA)

Sans oublier nos données plus personnelles…

D'une production ponctuelle… : achats/réservations en ligne, transports en commun (badges), navigation Web (réseaux sociaux, cheminement sur la toile, …), téléphonie mobile (Voix, SMS, ..), données de localisation (la plupart des outils mobiles, véhicules, …)

… à un « pipeline » de données : avec la probable explosion de l'utilisation des assistants (bracelets et autres) en tous genres qui produisent en permanence des données (stockées par exemple sur un Cloud…de la marque)

Nous sommes tous des sources « inépuisables » de données…

La notion de volume des données

Le **V**olume n'est que l'un des « 3**V** » du Big Data

- **V**olume (en expansion rapide avec un doublement régulier et de plus en plus rapprochée de toutes les données générées par l'humanité)
- **V**ariété (des données structurées, semi-structurées, brutes)
- **V**élocité, un cycle de plus en plus court entre la génération des données et leur exploitation/partage

Référence: **META Group, 2001**

La notion de volume, un critère réducteur d'un point de vue du fournisseur de données et de services

« ensembles de données … volumineux »

- On peut très bien héberger et fournir un accès à un gros volume de données sans être forcément dans la « catégorie » Big Data
- On peut très bien héberger et fournir un accès à un volume de données relativement modeste (quelques Go ou To) et entrer dans la « catégorie » Big Data
  - Nature des données ?, opérations associées ?, …
  - Prendre en compte le contexte
- Il est extrêmement hasardeux de catégoriser, Big Data ou non Big Data

Big Data, la notion (plurielle) d'exploitation des données

« leur exploitation avec des outils classiques … peut s'avérer ardue »

La partie technique

- Le stockage des données à une échelle sans commune mesure avec la dernière décennie qui a déjà connu une croissance extrêmement importante
- Distribution / répartition des données
- Multiplication des « Datacenters »

Big Data, la notion (plurielle) d'exploitation des données (2)

L'accès aux données implique des problématiques variées…

- La représentation des données
- L'indexation des données pour assurer des temps d'accès suffisants pour répondre à divers besoins (recherche multicritères, croisement des données, etc.) et la « scalabilité » des services associés
- La sémantique
- Il ne suffit pas d'utiliser Hadoop !
- Etc.

Big Data, la notion (plurielle) d'exploitation des données (3)

… et des applications multiples

- La science des données: Data Mining, traitement analytique prédictif, apprentissage automatique
- Croisement de multiples sources de données pour profiler une personne (à des fins commerciales (ciblages des offres, etc.) ou de renseignement (sécurité))
- Extraction de données pertinentes dans de nombreuses sources de données, souvent très volumineuses : recherches de similitudes (échantillons de population par exemple), prévision de crises, etc.
- Etc.

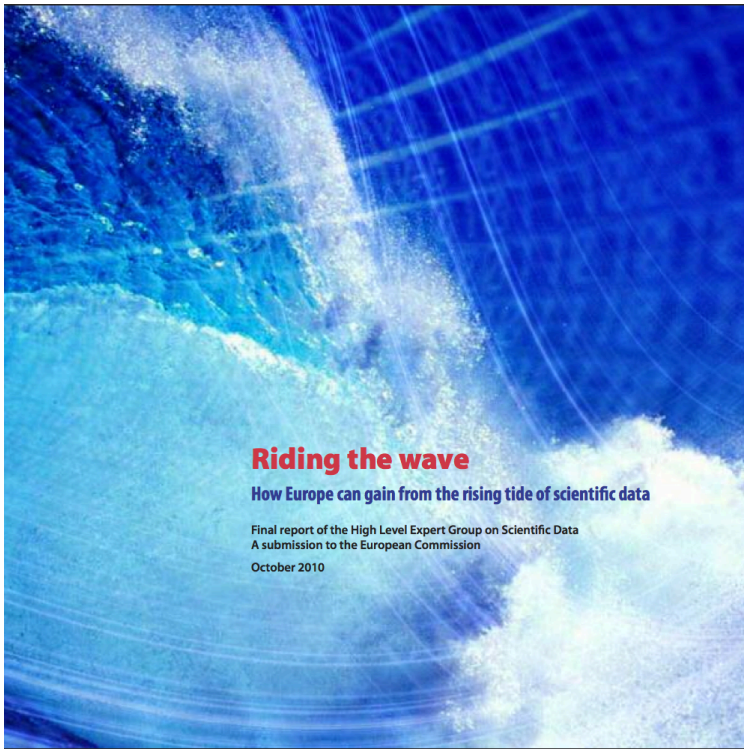
Surfer la vague sans oublier l'essentiel…

- La qualité des données ?
- La provenance des données ?
- La pérennité des données ?
- Les droits concernant les données générées, transformées, remises éventuellement à disposition ?
- Etc.

L'exemple de l'astronomie

Big Data par nature

Le contexte général de la science: la révolution numérique



Réseaux de communication, moyens de calcul massif, acquisition des données, mais aussi **la mise à disposition des données**

Les données doivent devenir une des infrastructures de la recherche

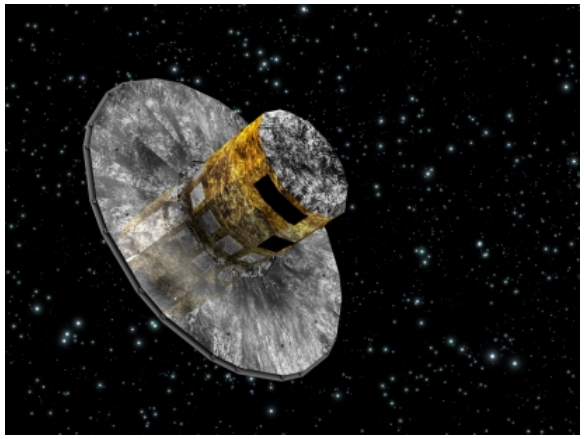
Une révolution dans les méthodes de travail des chercheurs

Découvrir les données, y accéder, les réutiliser, les combiner

L'astronomie est l'exemple d'une discipline qui a effectué cette (r)évolution

De « gros » producteurs de données…

Vue d'artiste de SKA
Crédit: SKA Organisation



Vue d'artiste de Gaia
Crédit: ESA

Vue d'oiseau du Very Large Telescope
Crédit: J.L. Dauvergne & G. Hüdepohl (atacamaphoto.com)/ESO

Hubble, le télescope spatial
Crédit: NASA, 2002

XMM-Newton
Crédit: Image courtesy of ESA

Sous le charme des Nuages de Magellan, ALMA
Crédit: ESO/C. Malin

Les données en astronomie

Des sources multiples

- Observations des télescopes sol et spatiaux
- Très grands relevés du ciel (informations homogènes sur un grand nombre d'objets)
- Bases de données à valeur ajoutée (CDS, NED)
- Données bibliographiques (journaux académiques, base de données ADS maintenue par la NASA)
- Données de modélisation

Les données en astronomie (2)

Des raisons scientifiques majeures de conserver les données et de les réutiliser

- Observations sur le long terme de phénomènes variables
- Un très grand nombre d'objets, des interactions complexes à différentes échelles (des phénomènes à l'oeuvre sur les grains du milieu interstellaire aux échelles cosmiques!)

Le changement de paradigme est effectif, les astronomes utilisant quotidiennement des données qu'ils trouvent dans des services distants

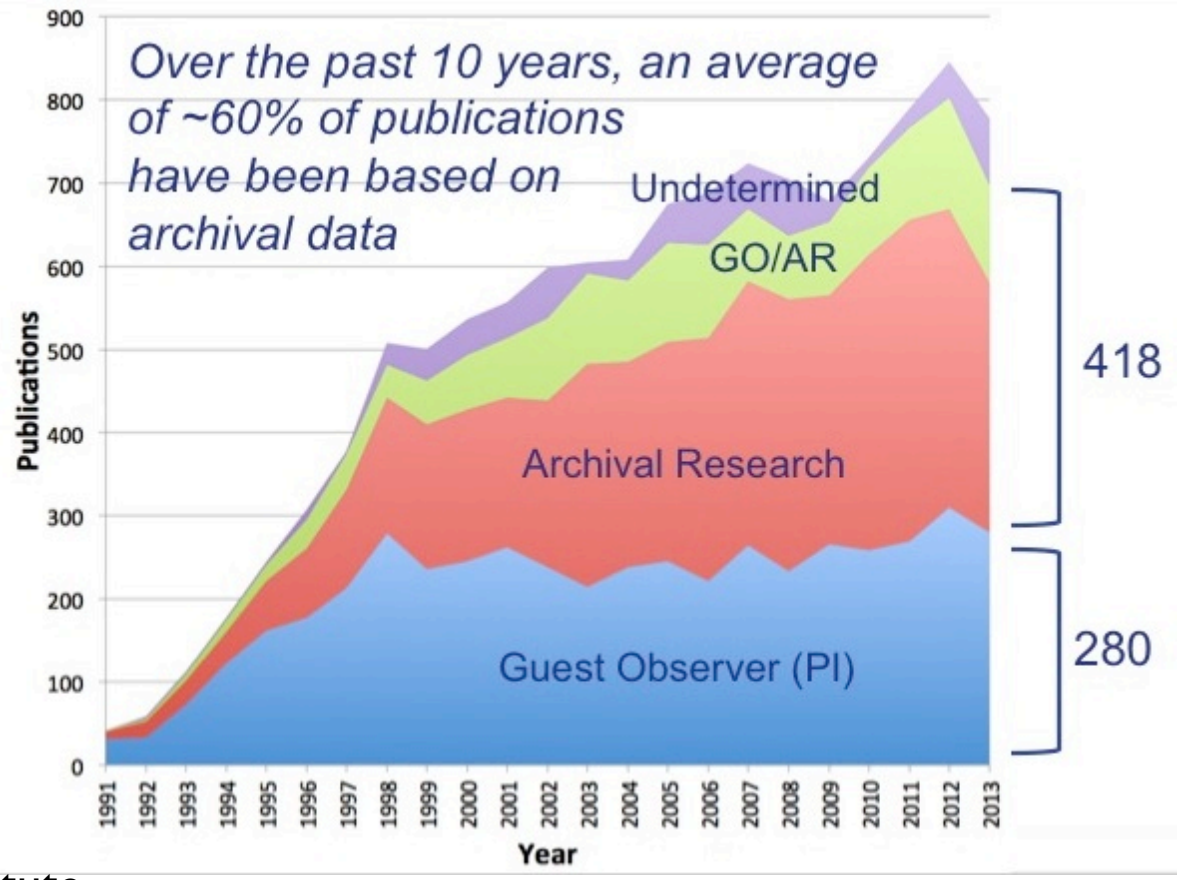
Les données en astronomie (3)

Les données sont réutilisées pour des objectifs différents des objectifs initiaux

- Une augmentation significative du retour scientifique des gros investissements
- La combinaison d'observations par différents instruments permet de comprendre les phénomènes à l'œuvre et engendre une part significative et croissante des publications

Illustration avec un instrument bien connu…

L'exemple du HST (Hubble Space Telescope)



Remerciements:
Robert J. Hanisch
Space Telescope Science Institute

Le Big Data en astronomie, une notion déjà ancienne

Du volume inhérent au grands instruments

- Pipeline de données brutes issues de capteurs CCD
- Stockage en temps réel

Une variété des données

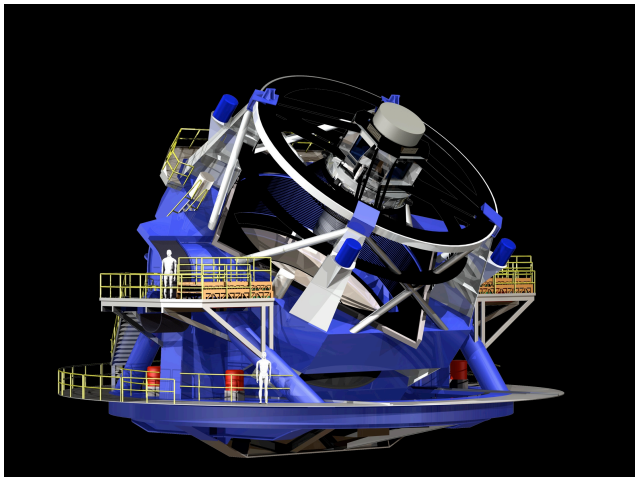
- Mesures physiques, images, spectres, simulations, publications, etc.
- Une approche multi-longueur d'onde pour une vision globale d'un objet astronomique complexe

Centaurus A en X, optique, infrarouge et radio

Des projets à venir encore plus prolifiques

LSST, Large Synoptic Survey Telescope

- Des milliards d'objets observés « sous toutes les coutures »
- Télescope terrestre de 8,4 mètres équipé d'une caméra de 3200 Mégapixels
- 30 Téraoctets de données par nuit



- Un changement d'échelle même pour l'astronomie

Le Centre de Données astronomiques de Strasbourg

Rester l'un des « spots » à surfer

Le Centre de Données astronomiques de Strasbourg existe depuis 1972

Le CDS

- dispose d'une équipe de documentalistes, d'astronomes et d'informaticiens (environ 30 personnes également réparties dans les 3 métiers)
- collecte des données utiles sur les objets astronomiques
- enrichit des données en les évaluant de façon critique et en les combinant
- développe des outils et mène des actions de R&D
- distribue des résultats à la communauté internationale
- conduit des recherches utilisant les données
- participe à des projets dans son domaine de connaissances

Aujourd'hui, environ 1,000,000 requêtes / jour sur ses services…

Un quadra en évolution…

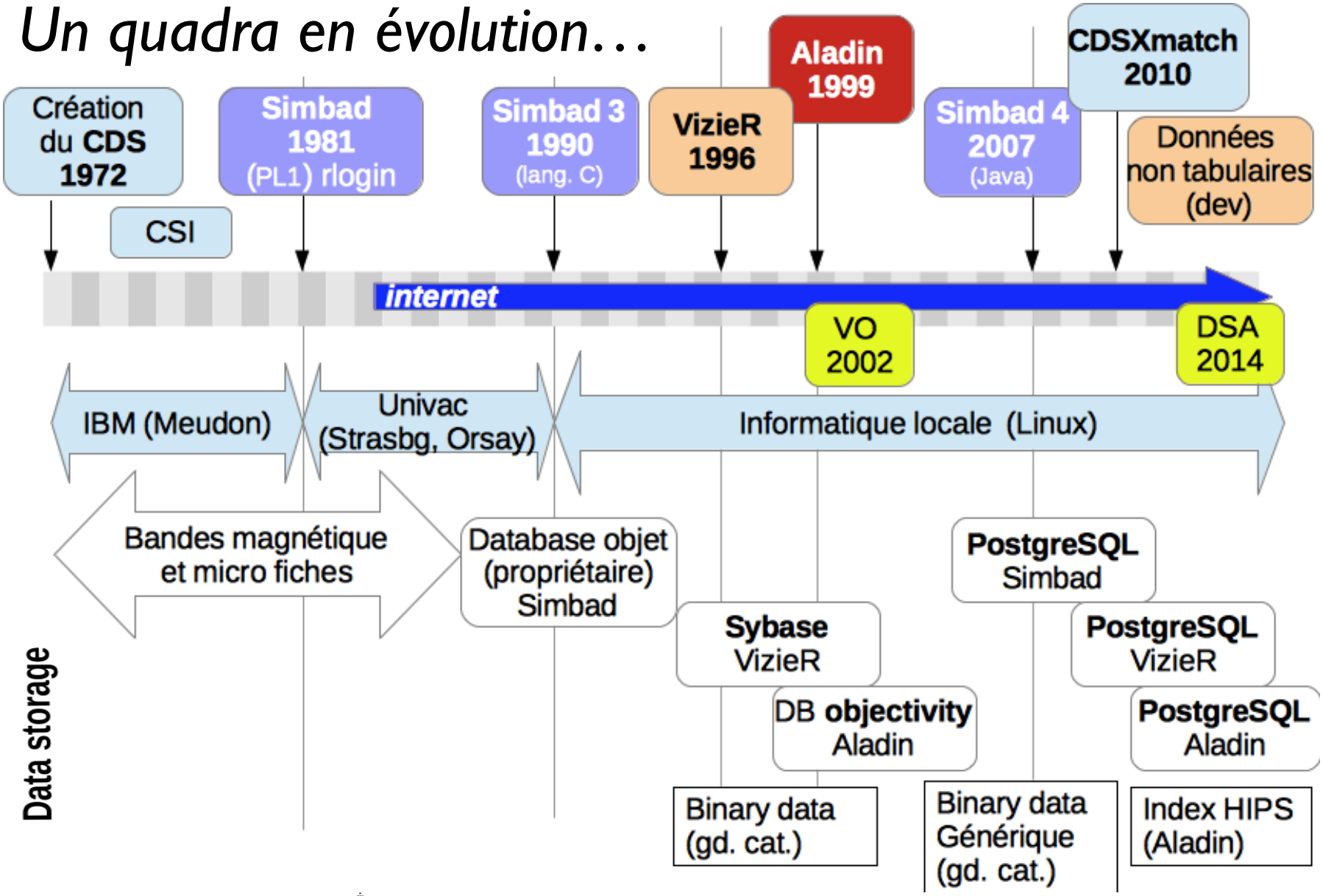
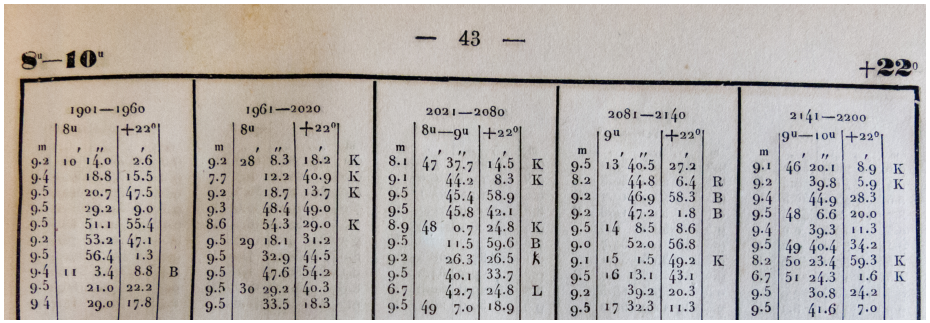


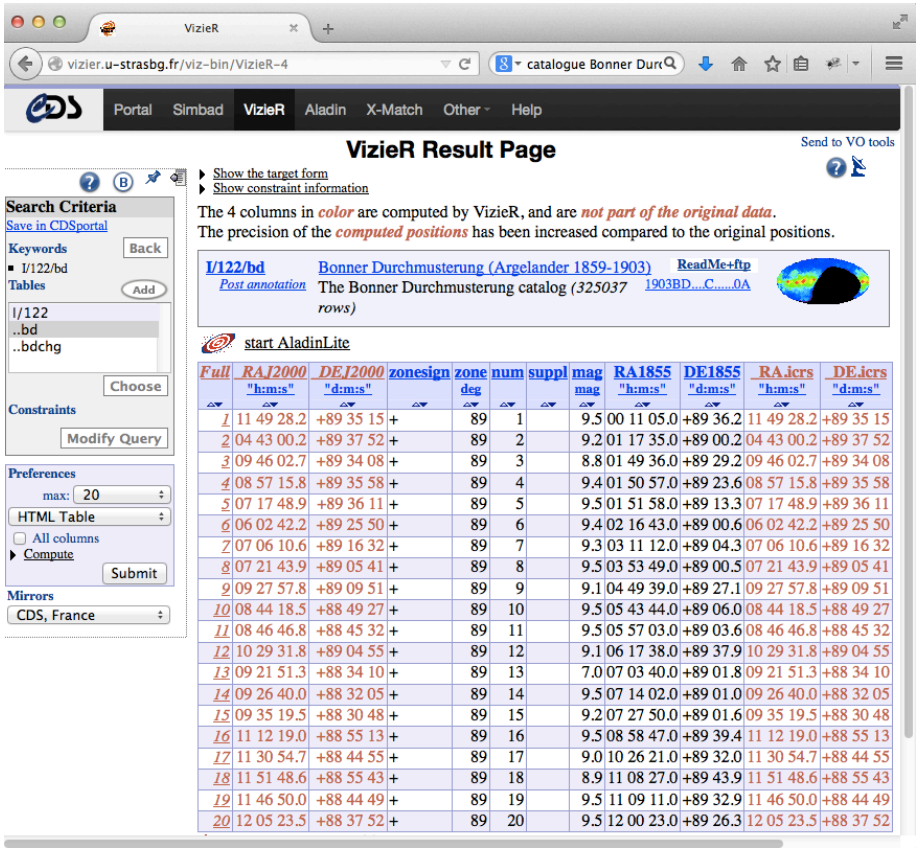
Illustration des migrations / évolutions du CDS



— 43 —

8u—10u +22°

1901—1960				1961—2020				2021—2080				2081—2140				2141—2200			
	8u		+22°		8u		+22°		8u—9u		+22°		9u		+22°		9u—10u		+22°
m	′	″	′	m	′	″	′	m	′	″	′	m	′	″	′	m	′	″	′
9.2	10	14.0	2.6	9.2	28	8.3	18.2	K	8.1	47	37.7	14.5	K	9.5	13	40.5	27.2	9.1	46	20.1	8.9	K
9.4		18.8	15.5	7.7		12.2	40.9	K	9.1		44.2	8.3	K	8.2		44.8	6.4	R	9.2		39.8	5.9	K
9.5		20.7	47.5	9.2		18.7	13.7	K	9.5		45.4	58.9		9.2		46.9	58.3	B	9.4		44.9	28.3	
9.3		29.2	9.0	9.3		48.4	49.0		9.5		45.8	42.1		9.2		47.2	1.8	B	9.5	48	6.6	20.0	
9.5		51.1	55.4	8.6		54.3	29.0	K	8.9	48	0.7	24.8	K	9.5	14	8.5	8.6		9.4		39.3	11.3	
9.2		53.2	47.1	9.5	29	18.1	31.2		9.5		11.5	59.6	B	9.0		52.0	56.8		9.5	49	40.4	34.2	
9.5		56.4	1.3	9.5		32.9	44.5		9.2		26.3	26.5	K	9.1	15	1.5	49.2	K	8.2	50	23.4	59.3	K
9.4	11	3.4	8.8	B 9.5		47.6	54.2		9.5		40.1	33.7		9.5	16	13.1	43.1		6.7	51	24.3	1.6	K
9.5		21.0	22.2	9.5	30	29.2	40.3		6.7		42.7	24.8	L	9.2		39.2	20.3		9.5		30.8	24.2	
9 4		29.0	17.8	9.5		33.5	18.3		9.5	49	7.0	18.9		9.5	17	32.3	11.3		9.5		41.6	7.0	



Full	RAJ2000 "h:m:s"	DEJ2000 "d:m:s"	zonesign	zone deg	num	suppl	mag mag	RA1855 "h:m:s"	DE1855 "d:m:s"	RA.icrs "h:m:s"	DE.icrs "d:m:s"
1	11 49 28.2	+89 35 15	+	89	1		9.5	00 11 05.0	+89 36.2	11 49 28.2	+89 35 15
2	04 43 00.2	+89 37 52	+	89	2		9.2	01 17 35.0	+89 00.2	04 43 00.2	+89 37 52
3	09 46 02.7	+89 34 08	+	89	3		8.8	01 49 36.0	+89 29.2	09 46 02.7	+89 34 08
4	08 57 15.8	+89 35 58	+	89	4		9.4	01 50 57.0	+89 23.6	08 57 15.8	+89 35 58
5	07 17 48.9	+89 36 11	+	89	5		9.5	01 51 58.0	+89 13.3	07 17 48.9	+89 36 11
6	06 02 42.2	+89 25 50	+	89	6		9.4	02 16 43.0	+89 00.6	06 02 42.2	+89 25 50
7	07 06 10.6	+89 16 32	+	89	7		9.3	03 11 12.0	+89 04.3	07 06 10.6	+89 16 32
8	07 21 43.9	+89 05 41	+	89	8		9.5	03 53 49.0	+89 00.5	07 21 43.9	+89 05 41
9	09 27 57.8	+89 09 51	+	89	9		9.1	04 49 39.0	+89 27.1	09 27 57.8	+89 09 51
10	08 44 18.5	+88 49 27	+	89	10		9.5	05 43 44.0	+89 06.0	08 44 18.5	+88 49 27
11	08 46 46.8	+88 45 32	+	89	11		9.5	05 57 03.0	+89 03.6	08 46 46.8	+88 45 32
12	10 29 31.8	+89 04 55	+	89	12		9.1	06 17 38.0	+89 37.9	10 29 31.8	+89 04 55
13	09 21 51.3	+88 34 10	+	89	13		7.0	07 03 40.0	+89 01.8	09 21 51.3	+88 34 10
14	09 26 40.0	+88 32 05	+	89	14		9.5	07 14 02.0	+89 01.0	09 26 40.0	+88 32 05
15	09 35 19.5	+88 30 48	+	89	15		9.2	07 27 50.0	+89 01.6	09 35 19.5	+88 30 48
16	11 12 19.0	+88 55 13	+	89	16		9.5	08 58 47.0	+89 39.4	11 12 19.0	+88 55 13
17	11 30 54.7	+88 44 55	+	89	17		9.0	10 26 21.0	+89 32.0	11 30 54.7	+88 44 55
18	11 51 48.6	+88 55 43	+	89	18		8.9	11 08 27.0	+89 43.9	11 51 48.6	+88 55 43
19	11 46 50.0	+88 44 49	+	89	19		9.5	11 09 11.0	+89 32.9	11 46 50.0	+88 44 49
20	12 05 23.5	+88 37 52	+	89	20		9.5	12 00 23.0	+89 26.3	12 05 23.5	+88 37 52

Du papier (le catalogue Bonner Durchmusterung (F.W.A. Argelander 1799-1875)) au numérique (via l'interface Web de VizieR)

Des services de référence de la communauté

~8,000,000 d'objets

Base de données d'objets astronomiques: identification, bibliographie, données, mesures...

~ 13,000 catalogues, 26,000 tables

Collection de données sous forme de catalogues: issus de journaux, de logs d'observation ou de grand relevés principalement constitués de données tabulaires (mais aussi spectres, images, séries temporelles)

50 To, 175 relevés au format HiPS

Atlas interactif du ciel: découverte, visualisation et manipulation de données, bases de données et archives d'images locales et distantes

La mise en réseau des données et une participation active à l'Observatoire Virtuel

Dès le démarrage du Web (~1993), une mise en ligne et en réseau des bases de données à valeur ajoutée et des publications, puis des archives

Depuis 2002, l'International Virtual Observatory Alliance définit des standards d'interopérabilité et développe des outils pour permettre un accès unifié aux données

Facilite l'accès aux données en éliminant l'étape d'apprentissage des interfaces propres à chacun des services de données

Utilisation de standards génériques lorsque c'est possible (registre des ressources OAI-PMH, vocabulaires SKOS / RDF), pour assurer la communication avec le monde extérieur

L'International Virtual Observatory Alliance

Pourquoi ? Comment ?

L'habitude de travailler dans des collaborations internationales

- Un format de données commun depuis les années 70, FITS (Flexible Image Transport System), toujours maintenu sous l'égide de l'Union Astronomique Internationale
- Les données des observatoires sont « ouvertes », en général après une période « propriétaire » réservé aux personnes qui ont obtenu du temps d'observation sur appel d'offre – les intérêts légitimes des personnes qui ont eu l'idée de l'observation sont préservés, et l'ouverture des données est accepté par la communauté, qui en profite pour ses propres recherches

L'Observatoire Virtuel, cela fonctionne comment ?

Pas de point central, un monde multipolaire

Un modèle ouvert

- Une fine couche assure l'interopérabilité mais chacun organise ses données comme il le souhaite
- Tout le monde peut enregistrer un service de données et développer un outil d'accès aux données

L'OV est « invisible », mais il est utilisé chaque fois que les chercheurs utilisent les outils ou les services

LES DONNÉES SONT UNE DES INFRASTRUCTURES DE L'ASTRONOMIE

La dissémination des données

Des données en libre accès

- Des applications Web ou « standalone » (Aladin) dédiées à chaque service
- Accès via les outils de l'Observatoire Virtuel ou par scripts
- Les formats des résultats des requêtes reconnus par la communauté astronomique: TSV, VOTable, FITS
- Des recherches de ressources (catalogues / tables / images...) facilitées par des métadonnées permettant des recherches multicritères
- Une indexation efficace

Une identification pérenne et reconnue par la communauté

- Un identifiant « bibcode » est attribué à chaque article
- Un identifiant est attribué à chaque catalogue

Le « bibcode » et l'identifiant catalogue sont reconnus et utilisés par l'application de bibliographie ADS (NASA) qui indexe les articles parus en astronomie
L'identifiant catalogue est également présent dans le Registre de l'Observatoire Virtuel

Le rôle de l'OV dans la dissémination des données

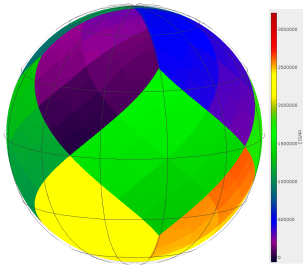
La réutilisation des données est garantie par l'utilisation de standards

- En utilisant des sorties standards (VOTable / XML) comprises par les applications de l'OV comprenant les données + les métadonnées
- En utilisant une nomenclature standard pour désigner le contenu des colonnes des tables
- En implémentant les services OV pour accéder aux données:
  - Recherche spatiale
  - Interrogation des tables par un langage commun basé sur SQL et incluant des fonctionnalités nécessaires pour l'astronomie

Une visibilité des données du CDS à travers l'Observatoire Virtuel
Le « registre OV », un « dictionnaire de ressources » basé sur le standard OAI-PMH, est interrogé par les outils de l'OV

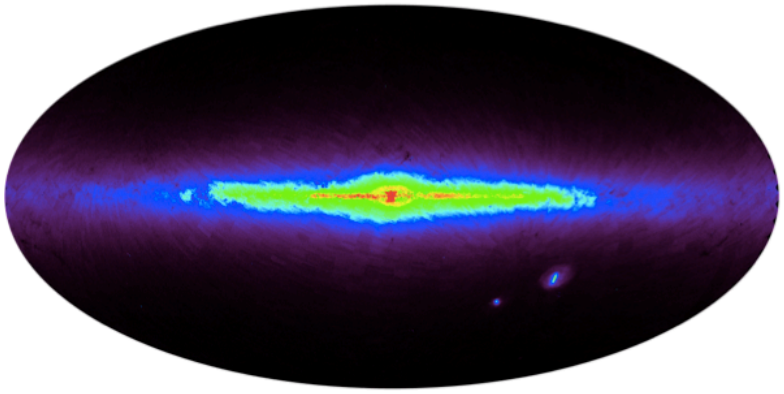
Une R&D soutenue

Exemple de « corrélation croisée » de deux catalogues astronomiques 2MASS et USNO-B1

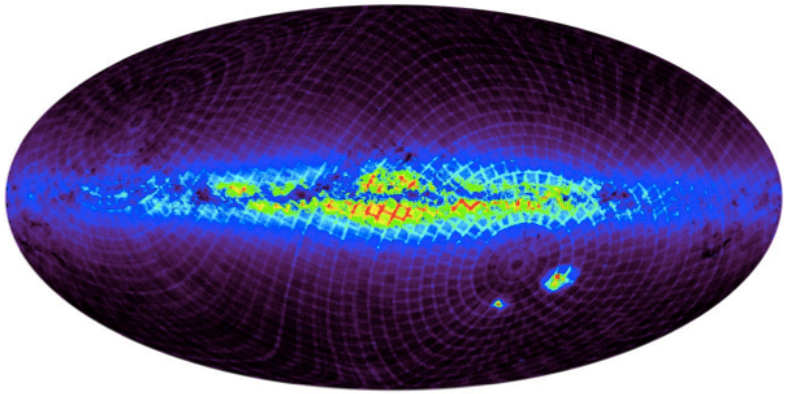


- Pixellisation du ciel
- Une trentaine de minutes de traitement sur du matériel standard
- 583 300 000 associations trouvées

F.X. Pineau



2MASS (~470 000 000 sources)



USNO-B1 (~1 046 000 000 sources)

Et des efforts de labellisation

Le CDS a obtenu le « **D**ata **S**eal of **A**pproval » en août 2014

Un label international pour la préservation des données scientifiques sur le long terme

Une réévaluation tous les 2 ans portant sur 16 critères pouvant se résumer aux règles suivantes:

- Les données sont accessibles via Internet
- Les licences / droits d'utilisations des données sont clairement exprimés
- Les données sont réutilisables
- Les données sont fiables
- Les données sont identifiées de manière unique selon une nomenclature permettant un référencement externe

Organisation de l'infrastructure selon un modèle reconnu internationalement et adaptée au cas spécifique, « OAIS like »

En complément
RDA, une initiative mondiale (depuis 2013)
pour renforcer l'échange des données de la recherche

La 6ème réunion plénière aura lieu à Paris en septembre 2015

Research Data Alliance

Research Data Sharing without barriers

Les chercheurs et les innovateurs partagent ouvertement les données par delà les frontières des technologies, des disciplines et des pays pour résoudre les grands défis..



… construire des passerelles (techniques, sociologiques) pour permettre le partage des données au niveau mondial

Les chercheurs, les praticiens des données, les « technologistes » sont invités à travailler ensemble

Conclusion

L'astronomie

- Une discipline concernée de facto par le Big Data
- Une forte implication scientifique et technique
- Des travaux initiés (Observatoire Virtuel par exemple) depuis de nombreuses années pour préparer la révolution numérique
- Un travail constant en amont (R&D) pour améliorer la disponibilité et l'utilisation des données et pas seulement par une augmentation des capacités de stockage !
- La définition des standards et des métadonnées sont des briques essentielles pour faciliter le passage à l'échelle

Questions ?