L’archivage des données en astronomie

Francoise Genova

Observatoire Astronomique de Strasbourg

Gilles Duvert

IPAG/OSUG et JMMC

Le contexte: les données en astronomie

Les infrastructures de recherche en astronomie


Gaia


ESO/VLT chili


ESO/ELT chili


SKA chili australie


Herschel


CTA chili europe


HESS namibie


Planck


IRAM hautes-alpes

+ LES DONNEES

Les données en astronomie

- La réutilisation des données et l'interopérabilité sont au cœur de la démarche scientifique de la discipline
  - Astronomie multi-longueur d'onde et multi-messagers (ex : ondes gravitationnelles LIGO/VIRGO, particules hautes energie CTA,HESS, etc)

    La combinaison d'observations par différents instruments permet de comprendre les phénomènes à l'oeuvre – une part significative et croissante des publications
  - Variabilité des objets
  - Comparaison modèles/observations
  - Etc …
- Optimisation du retour scientifique des investissements dans les télescopes sol et spatiaux

Les données en astronomie

- Les données
  - Observations des télescopes sol et spatiaux – archives des observatoires
  - Très grands relevés du ciel (informations homogènes sur un grand nombre d'objets)
  - Bases de données à valeur ajoutée (CDS Centre de Données Stellaires, NED : NASA Extragalactic Database.)
  - Données bibliographiques (journaux académiques, base de données ADS maintenue par la NASA)
  - Données de modélisation
- « Big data » et données « de longue traîne » - résultats attachés aux publications – données utiles, validées et documentées
- Les astronomes peuvent trouver les données, y accéder, les réutiliser et les faire interopérer – des outils de travail dans la recherche de tous les jours

Des standards partagés au niveau international

- Il faut des standards (et qu'ils soient utilisés par les producteurs de données!) pour découvrir les données, y accéder, les réutiliser et les rendre interopérables
- Point de départ: le format FITS (1977)
  - Intègre données et métadonnées: les données sont **réutilisables**
  - Permet de partager les observations des télescopes
  - Permet de développer des outils communs
  - Maintenu par l'Union Astronomique Internationale. Voir fits.gsfc.nasa.gov
- Depuis 2002: l'Observatoire Virtuel astronomique
  - Trouver les données, y accéder, interopérabilité des outils et des données
  - Standards maintenus par l'International Virtual Observatory Alliance (IVOA) (voir ivoa.net, france-ov.org)
  - Un système ouvert, sans point central: tout le monde peut déclarer un service dans le registre des ressources de l'IVOA et construire un outil d'accès aux données sur la base des standards
  - Le VO est sous-jacent, invisible pour les utilisateurs

Les acteurs du cycle de vie des données

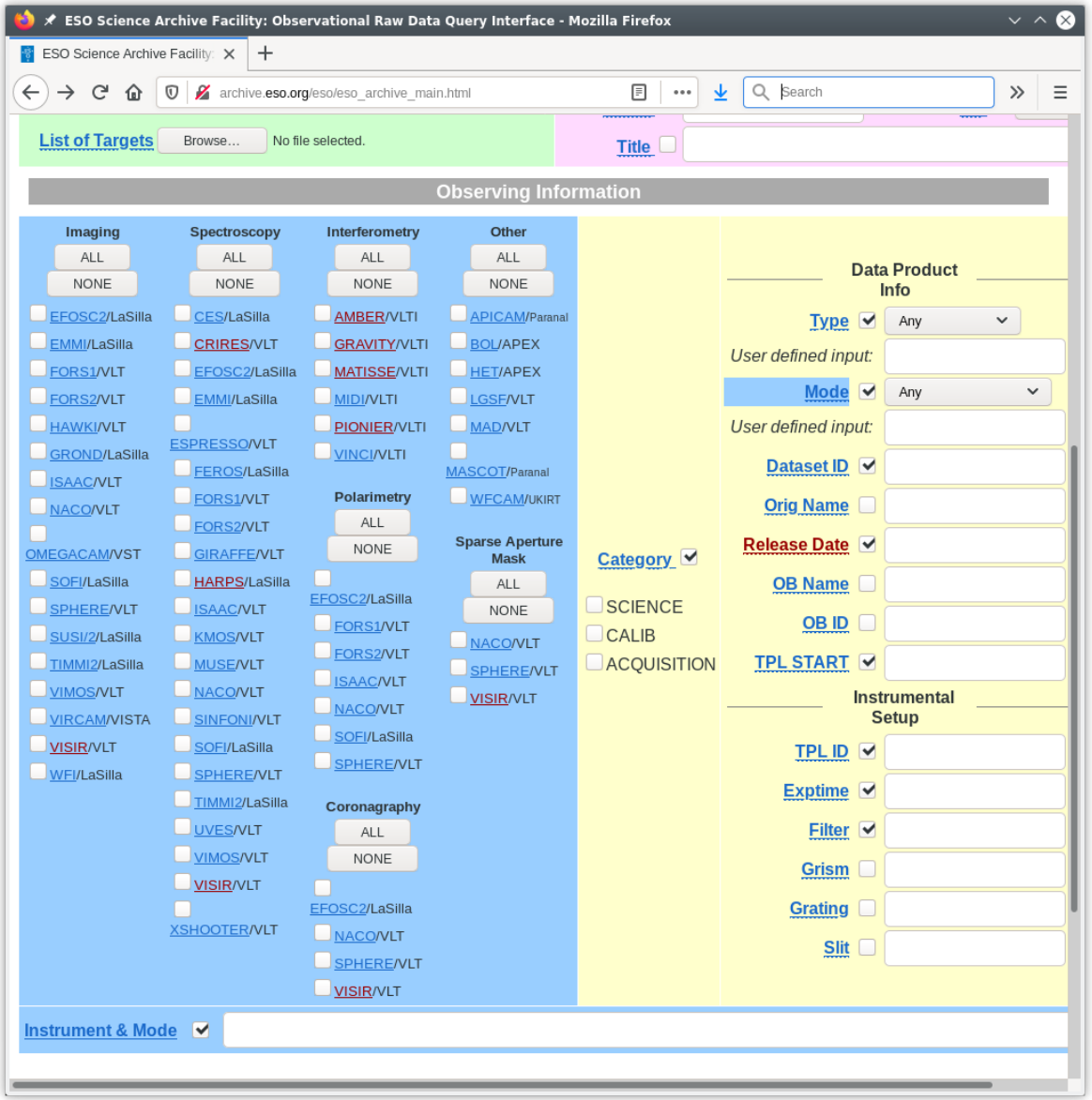
Les grands observatoires sol et spatiaux

- **Ils jouent un rôle majeur dans le cycle de vie des données**
- **Les observatoires ou les agences qui en sont responsables**
  - Acquièrent les données
  - Les rendent disponibles pour la communauté, souvent après une période propriétaire pour les données obtenues par des équipes scientifiques sur appel d'offre pour du temps d'observation
  - **Gardent la responsabilité des données sur le long terme**
    - Pour les missions spatiales de la NASA, un transfert de responsabilité est fait de la mission à un centre thématique (optique/UV, infrarouge, hautes énergies)
    - L'ESA (spatial) et l'ESO (sol) sont en charge de l'archivage de leur données

Les grands observatoires sol et spatiaux

Exemple de l'ESO :

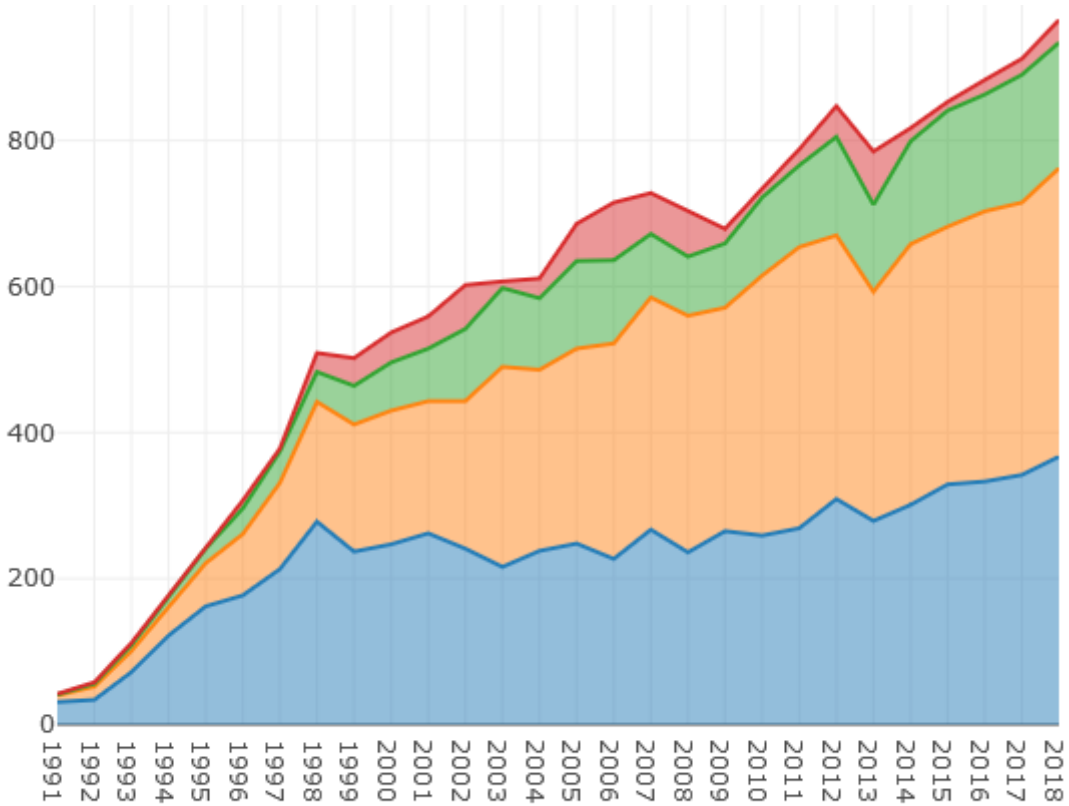
~30 instruments sur 35 ans, plusieurs modes (spectroscopie, polarimetrie, imagerie, interféro-métrie, …



Les grands observatoires sol et spatiaux

- **Le partage des données sur le long terme est un élément important de l’impact des télescopes**
- les retrouver facilement à partir de critères de position, date, résolution spectrale ou angulaire, qualité, nom d'objets, qualité, etc…
- et de manière intéropérable…
- c'est l'observatoire virtuel.

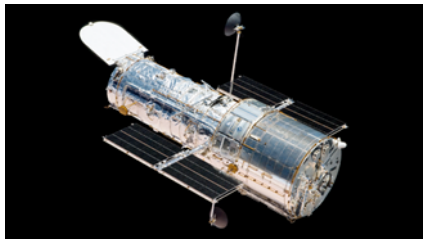
Publications avec les données du Télescope Spatial Hubble (HST)



Les deux

Archive

« Guest Observers »



Plus d'articles publiés à partir de l'archive que directement à partir du temps d'observation sur appel d'offre

https://archive.stsci.edu/hst/bibliography/pubstat.html

Ces chiffres sont globalement valables aussi pour les données dans l'archive l'ESO

Le Centre de Données astronomiques de Strasbourg (CDS)

- Depuis 1972
  - Responsabilité conjointe du CNRS/INSU et de l'Université de Strasbourg
- Dans une structure de recherche de l'Université, l'Observatoire Astronomique de Strasbourg (UMR 7550)
- Equipe intégrée de chercheurs/documentalistes/informaticiens (~35 personnes, 1/3-1/3-1/3)
- « Big and smaller data »: données résultats de recherche, validées par des publications + données d'observation « de référence »

Le rôle du CDS

- Reste fidèle à sa mission initiale
  - Collecter les données utiles sous forme électronique
  - Les améliorer en les évaluant et en les comparant
  - Distribuer les résultats à la communauté scientifique internationale
  - Mener des recherches utilisant les données
- Pour la communauté
  - Collecte de données attachées à des publications, en liaison avec les éditeurs scientifiques des journaux
  - Met à disposition des services largement utilisés par la communauté internationale – ~1 million de requêtes/jour en moyenne

Le CDS et l’archivage des données

- La préservation des données est une conséquence collatérale de la collecte de données et de la fourniture de services sur le long terme
- Préservation = Conservation sur le long terme de données sous une forme qui permet leur réutilisation

- La question de documenter la fonction d’archivage au CDS s’est posée quand le CDS a envisagé de candidater pour une Certification comme dépôt de données de confiance dans le cadre du Data Seal of Approval, devenu depuis Core Trust Seal 

en fusionnant avec le cadre de certification du World Data System

Critères Core Trust Seal liés à la préservation

- Parmi les critères relevant de l'infrastructure Opérationnelle

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

- Parmi les critères relevant de la gestion des objets numériques

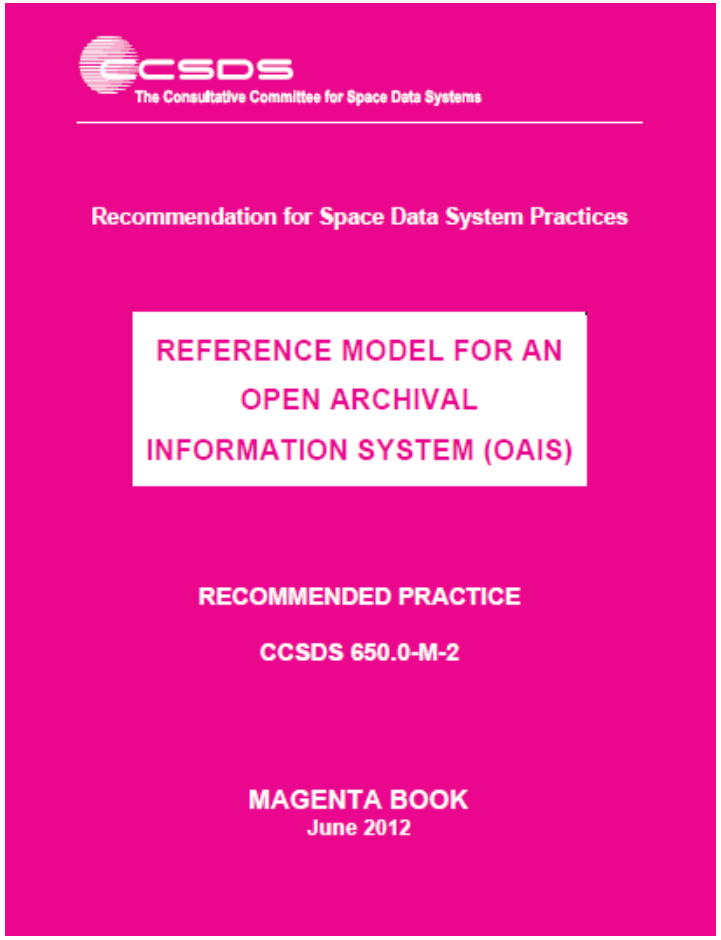
R9. The repository applies documented processes and procedures in managing archival storage of the data.

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

R12. Archiving takes place according to defined workflows from ingest to dissemination.

- Le questionnaire rempli par le CDS: https://www.coretrustseal.org/wp-content/uploads/2019/02/Strasbourg-Astronomical-Data-Centre.pdf

Description des process du CDS



- Basé sur le modèle OAIS – Open Archive Information System

https://public.ccsds.org/Pubs/650x0m2.pdf

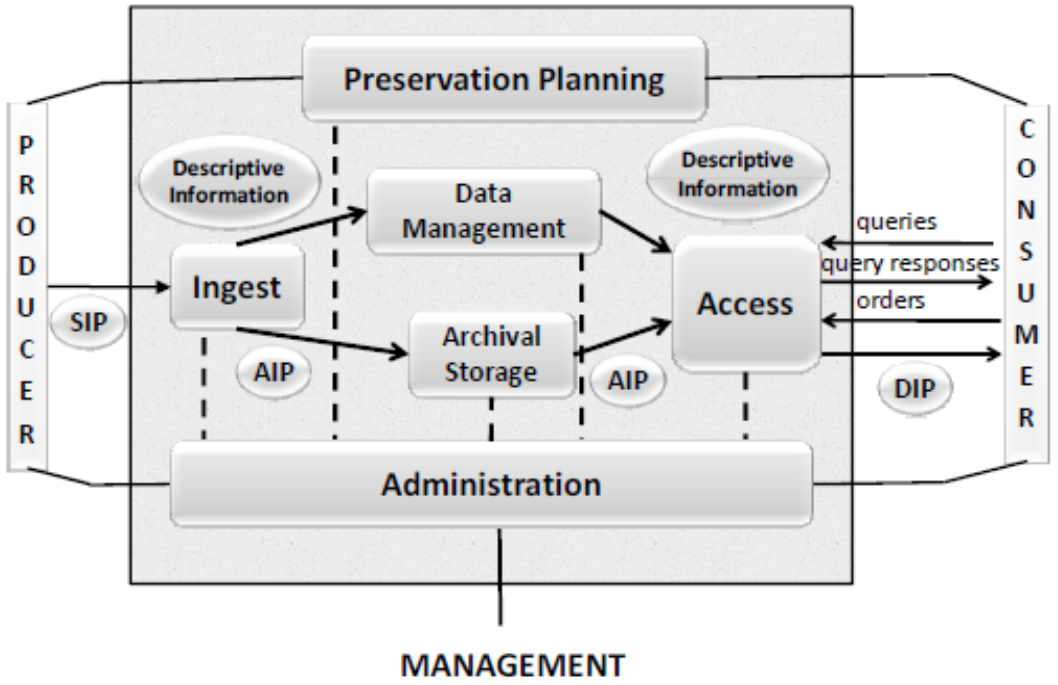
- Site en français

https://www.cines.fr/archivage/un-concept-des-problematiques/le-modele-de-reference-loais

L’environnement d’une archive OAIS

Les acteurs selon le modèle OAIS

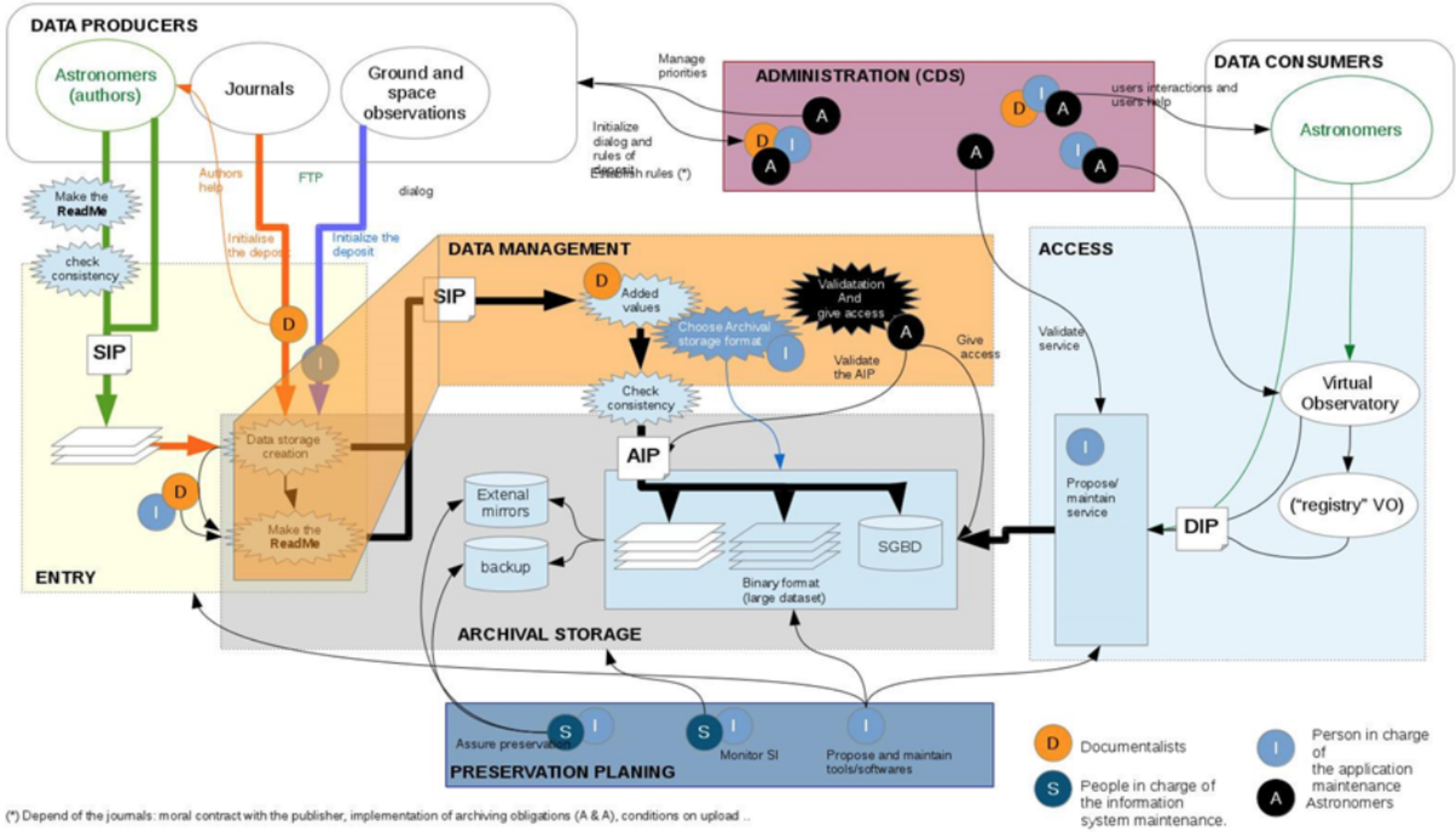
Les entités fonctionnelles de l'OAIS



SIP: Submission Information Package
AIP: Archival Information Package
DIP: Dissemination Information Package

Le pipeline de données du CDS dans le modèle OAIS

Les éléments du pipeline et les intervenants



Parmi les éléments indispensables

- Des formats permettant la réutilisation sur le long terme
- Archiver les données et les metadonnées
- Migrations en temps et heure, sans perte d'utilisabilité
- Une très longue histoire, résumée dans le document Core Trust Seal du CDS
  - 1972-1972: plusieurs serveurs génériques, succesivement à Meudon, Orsay, Strasbourg, Orsay
  - 1972- : plusieurs générations de stations gérées par l'Observatoire de Strasbourg

Conclusion

- L’archivage est un élément du cycle de vie des données, qui doit être préparé en amont
- Il doit prendre en compte les pratiques disciplinaires pour la préservation, qui doit permettre la réutilisation
- Core Trust Seal est un outil utile pour vérifier les procédures…
  - La certification des entrepôts de données est une recommandation du Plan National pour la Science Ouverte et une priorité de RDA-France