Les données scientifiques au CNES

Cycle de vie, processus de gestion

Danièle BOUCON

Réunion PIN du 4 janvier 2013

SOMMAIRE

- Contexte des données scientifiques au CNES
- Production d'une base de données et cycle de vie
- Processus d'archivage
- Quelques exemples de modélisation de cycle de vie
- La vision ESA/LTDP
- Conclusion

Les données scientifiques au CNES

Patrimoine important de données issues de missions spatiales

- Le CNES a plus de 50 ans d'existence
- Fort investissement scientifique (le CNES a pour missions « de développer et d'orienter les recherches scientifiques et techniques poursuivies dans le domaine des recherches spatiales », loi 61-1382 du 19/12/1961) :
  - participation au développement d'instruments scientifiques en partenariat avec les laboratoires,
  - missions dans un cadre multilatéral sur des objectifs scientifiques (et technologiques),
  - pôles thématiques pour le traitement et l'archivage de données spatiales …
- A conduit ou participé à plus de 100 missions spatiales
- Certaines données ont plus de 30 ans
- 4 grands domaines scientifiques « observation de la terre », « sciences de l'univers », « sciences de la vie », « sciences de la matière »
- Des caractéristiques propres aux données scientifiques

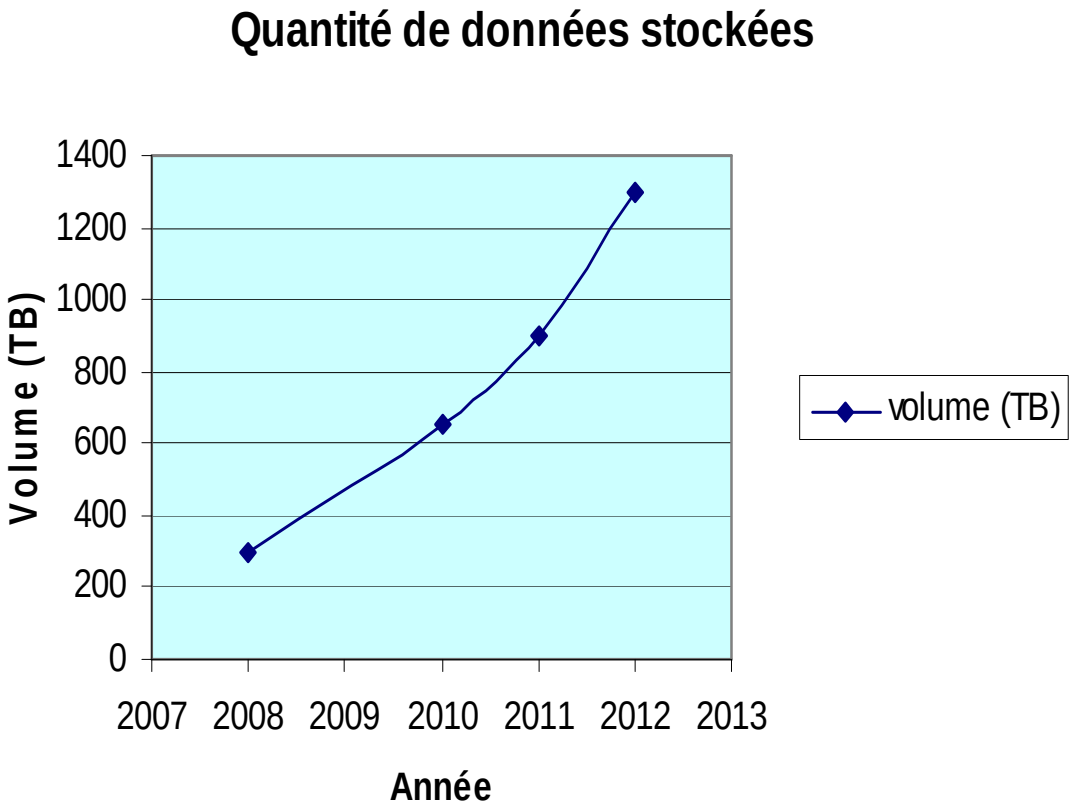
Quelques caractéristiques des données scientifiques spatiales

- Données uniques et difficilement reproductibles (événements exceptionnels, séries temporelles …)
- Collecte et gestion = entreprises lourdes nécessitant des moyens importants
- Masse élevée d'informations (sur la planète) -> des volumes de données importants (et en forte croissance)
- Données hétérogènes -> forte diversité de formats
- Diversité des producteurs : thématiques, localisation
- Cycle de vie étendu dans le temps :
  - Durée de vie d'un projet spatial (initialisation jusqu'à phase de retrait) : variable (environ 10 à 20 ans),
  - Développement d'un projet spatial entre 5 et 12 ans,
  - Durée d'exploitation d'un instrument : besoin initial de 1 (mission Picard) à 5 ans, mais peut aller bien au-delà (presque 18 ans pour mission Wind),
  - Durée de conservation des données : a priori illimitée

Evolution de la quantité de données stockées

Accroissement de la volumétrie au CNES (STAF)

- 2008
  - > 300 TB de données stockées
    - » 7 500 000 fichiers
    - » 1,5 TB/ semaine
- 2010
  - > 650 TB de données stockées
    - » 11 000 000 fichiers
    - » 3 TB /semaine
- 2011 : > 900 TB de données sto
  - > 5 TB /semaine
- 2012 : ~ 1,3 PB de données stockées,
  - > 7 TB /semaine



Cycle de vie des données spatiales

Les données spatiales, quelques exemples

- SPOT (début 2010, 5 satellites, depuis 1986 pour SPOT 1)
  - ✦300 TB archivés au Cnes
- Pléiades (2 satellites, depuis déc 2011 pour PLEIADES 1A)
  - ✦> 5000 TB sur la durée de la mission
- SWOT (Haute résolution Hydrologie)
  - ✦> 6000 TB à échanger entre NASA et CNES
- GAIA (le Data Processing Center du Cnes)
  - ✦1000 TB de données
  - ✦1 Milliard d'étoiles cataloguées
  - ✦80 Milliards d'objets gérés

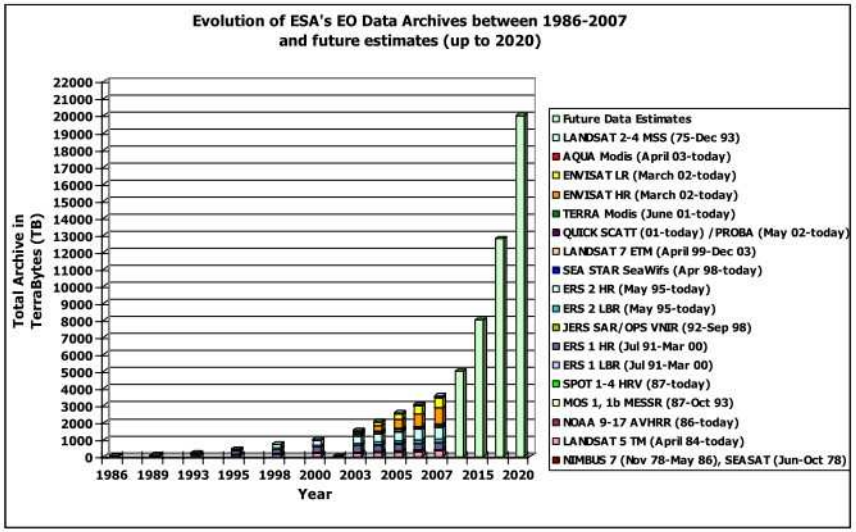


Figure 5 – ESA expected volume of archives

Techniques de réduction de données massivement collaboratives

Travaux en cours

Axes stratégiques

Maîtriser la gestion du cycle de vie de l'information et des données,
Maitriser l'augmentation des volumétries des données et les coûts de transport/stockage/archivage/traitement/accès.

➔ Réflexion pour améliorer la maitrise du patrimoine informationnel (données et documents)

- Données spatiales
- Données technologiques
- Données techniques
- Données administratives

➔ Compléter politiques, directives et procédures pour la gestion des bases de données

- Cycle de vie
- Processus de gestion et pérennisation
  - Intégrer dans les projets, dès les phases amont
  - En concertation avec les travaux européens

➔ Recenser les bases, évaluer les données à pérenniser, établir un plan d'actions

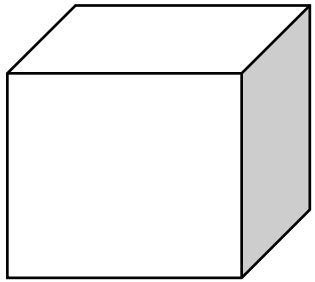
Définitions

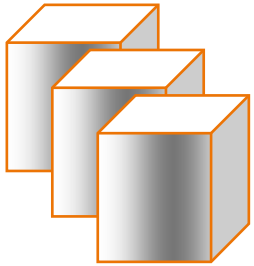
- Base de données
  - recueil d'oeuvres, de données ou d'autres éléments indépendants, disposés de manière systématique ou méthodique, et individuellement accessibles par des moyens électroniques ou par tout autre moyen.
  - Une base de données peut être constituées de différents jeux de données. Elle intègre également l'ensemble des informations nécessaires à la compréhension et à l'utilisation de ces jeux de données.

- Jeu de données
  - Un jeu de données désigne un groupe d'objets « données » ayant des caractéristiques communes motivant leur regroupement. Dans la plupart des cas, un jeu de données correspond à un ensemble homogène d'objets « données » (par exemple, les données de niveau 1 d'un instrument). Dans ce cas, les fichiers du jeu ont la même structure et contiennent les mêmes paramètres. Mais un jeu de données peut aussi correspondre à un ensemble d'objets « données » d'origines diverses (par exemple, l'ensemble des observations disponibles sur une zone géographique, quelle que soit la provenance des observations).

- Niveaux de données
  - Résultats de différentes transformations opérées sur les mesures effectuées par des instruments en orbite afin d'obtenir des paramètres physiques (ou combinaisons). En général :
    - Données brutes, N0, N1 (données étalonnées en grandeurs physiques) -> compétence organismes spatiaux
    - Données N2 et supérieurs -> compétence laboratoire de recherche scientifiques

Production de la base des données





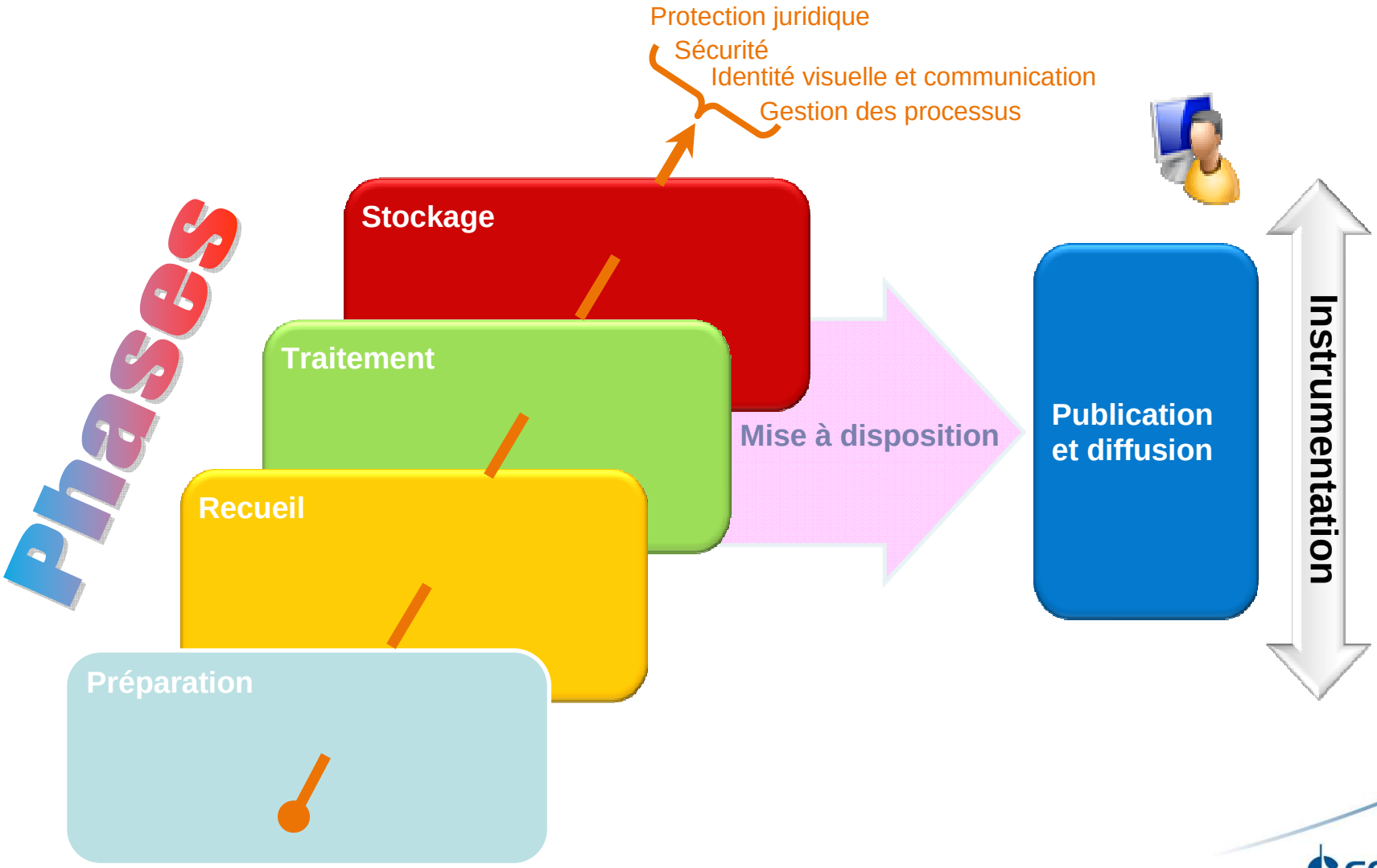
Données scientifiques

Jeux de données brutes produites par les instruments embarqués ou de niveau supérieur

Données complémentaires

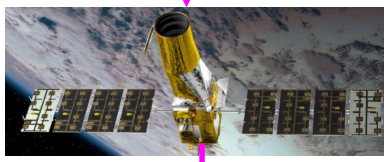
- **Données auxiliaires scientifiques et/ou techniques**
- **Paramètres de calibration**
- **Quick look**
- **Métadonnées**
- **Logiciels**
- **Documentation**

Phases du cycle de vie et activités transverses



Production de la base des données

préparation



diffusion

recueil

traitement

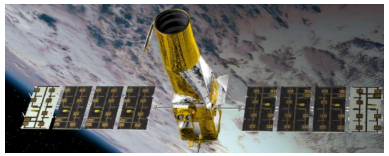
Données
N0

stockage

archives
courantes

Production de la base des données

préparation



diffusion

préparation

recueil

traitement

recueil

traitement

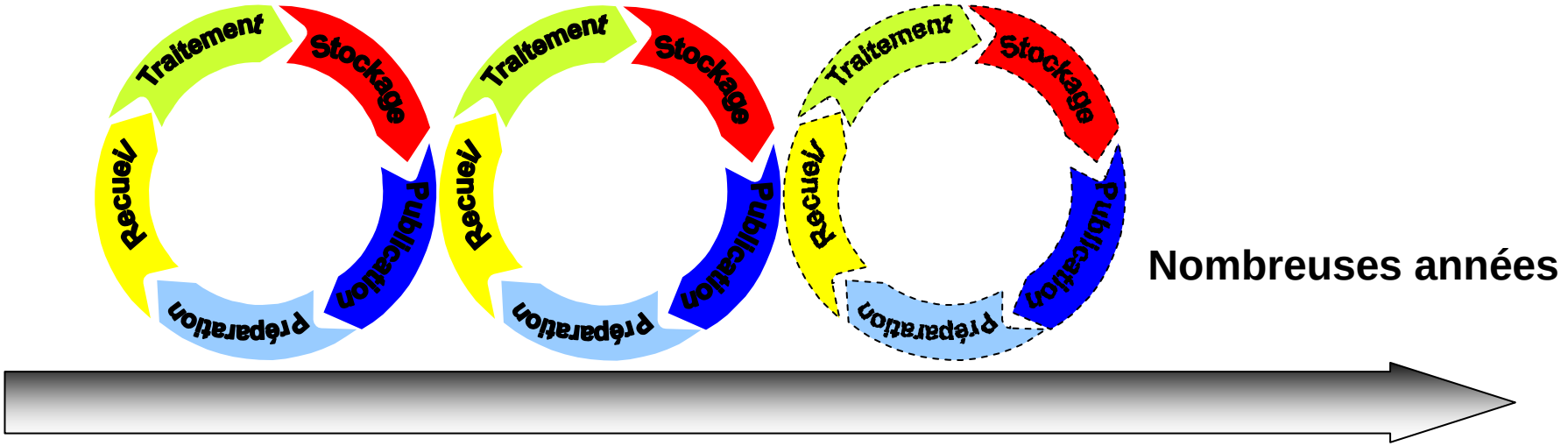
Données N0

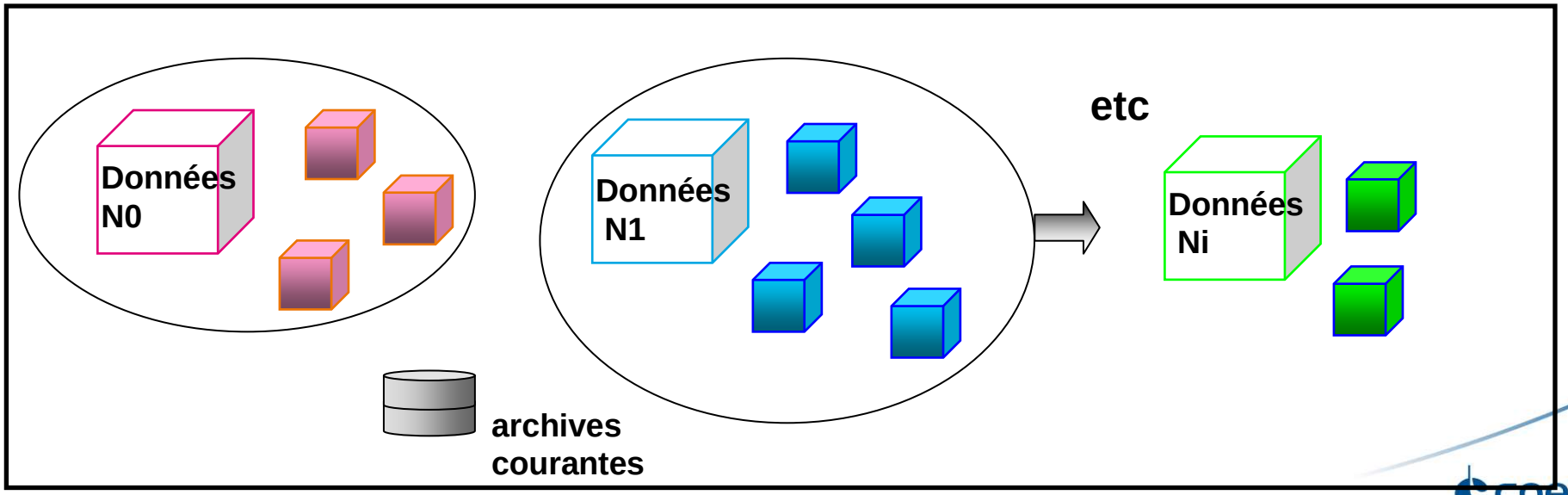
Données N1

stockage

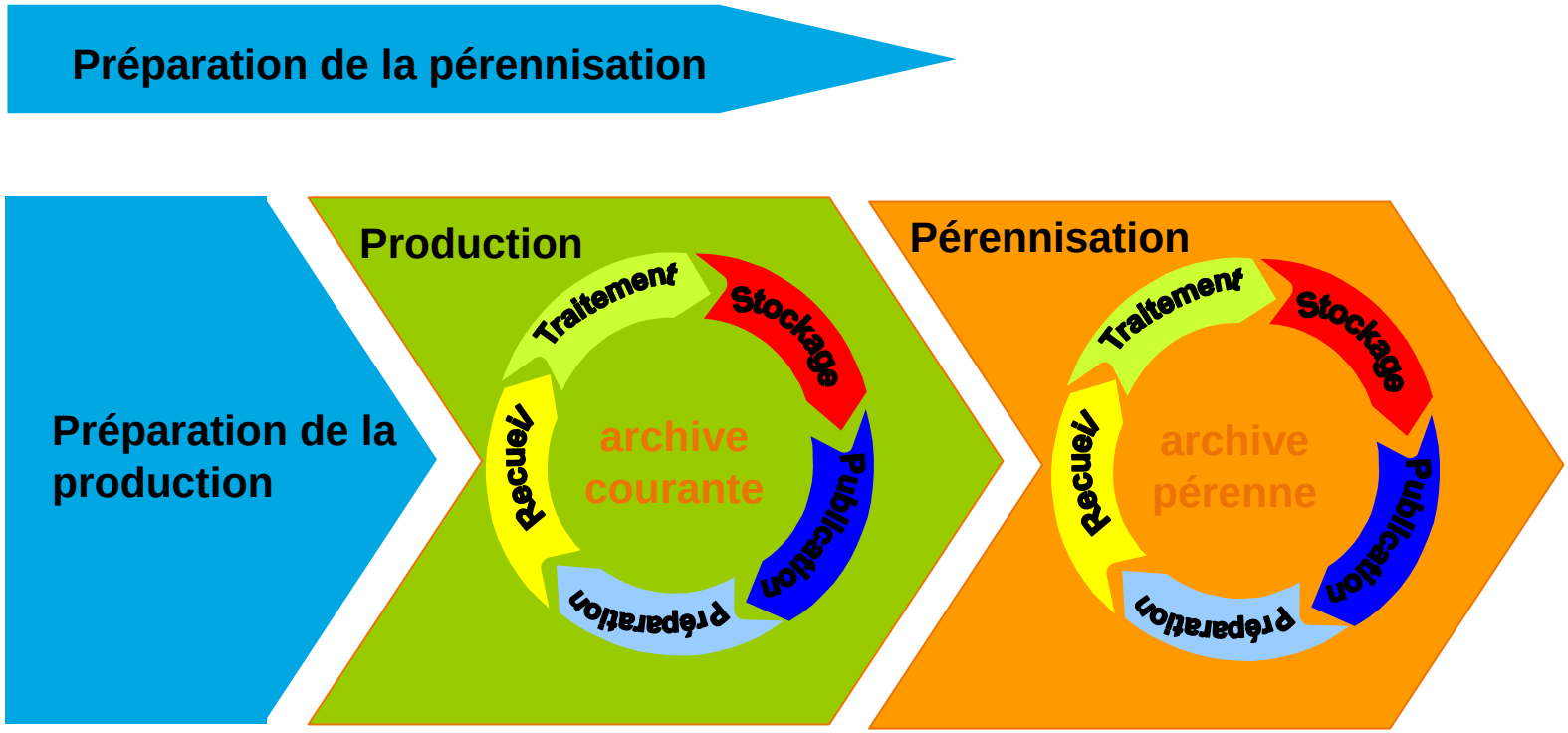
archives courantes

Production de la base des données



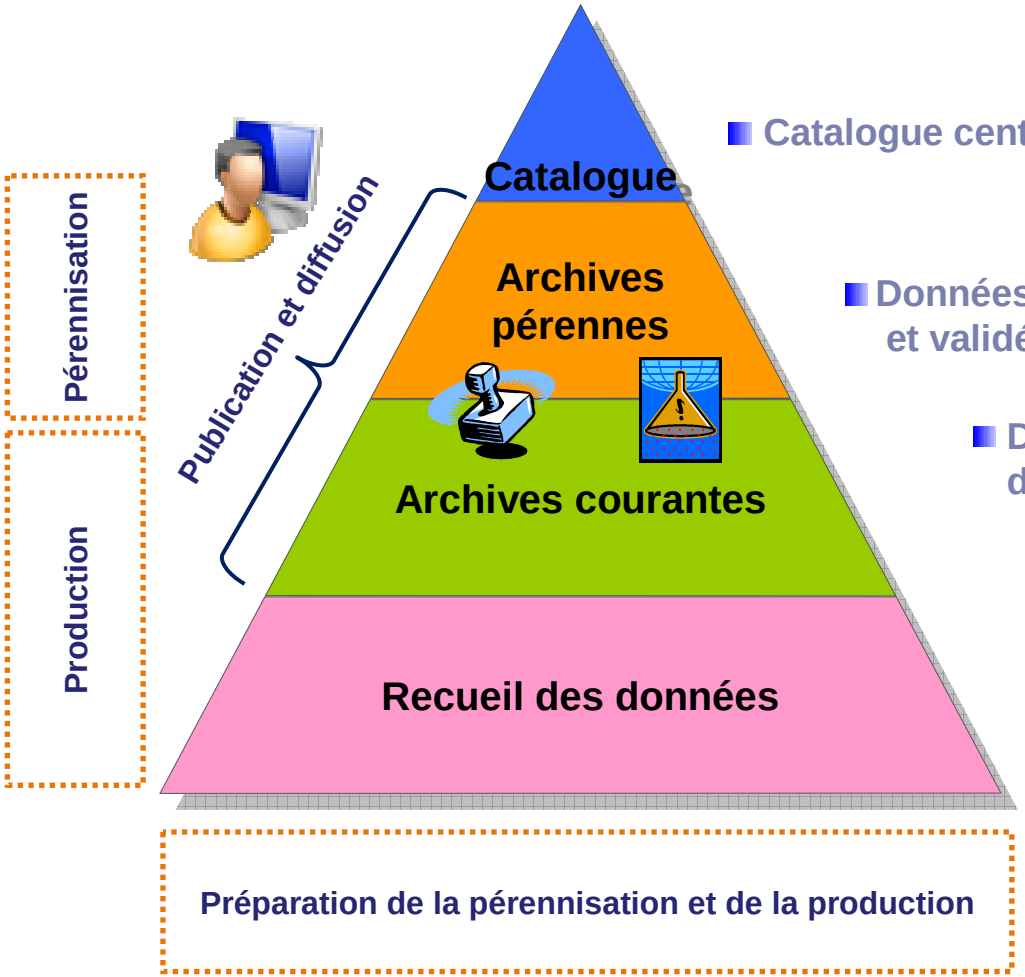


Processus de gestion des données



3 grandes étapes : préparation, production, pérennisation

Espaces de gestion des données au cours de leur cycle de vie

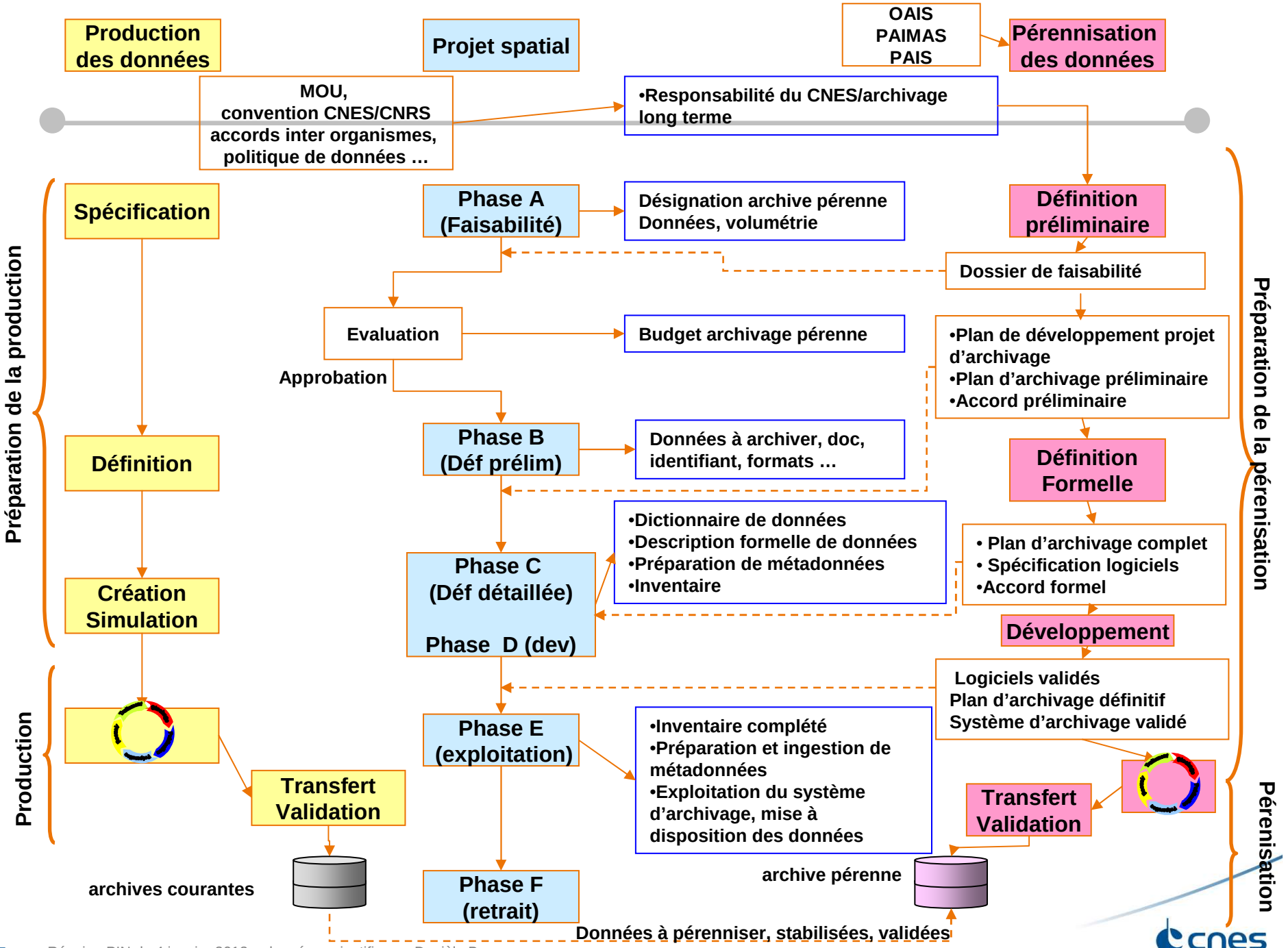


- Catalogue centralisé
- Données stabilisées patrimoniales et validées
- Données utilisées dans le cadre d'un projet spatial
- Données collectées ou saisies (peuvent provenir de calcul)

Processus de gestion des données spatiales

➔ Intégrer la pérennisation tôt dans le projet spatial (avant la production des données)

- Le cycle de vie des données est conçu pour fournir un cadre de gestion des données (indépendant des applications/systèmes dans lesquels résident les données)
- Ce cycle de vie est non linéaire
- Les données sont gérées dans le cadre d'un projet spatial, ou par une autre structure au retrait de celui-ci (par exemple SERAD –Service de Référencement et d'Archivage de Données pour les données orphelines)
- La pérennisation des données doit s'appuyer sur un plan d'archivage, elle respecte les normes OAIS et PAIMAS, et PAIS (à venir).
- La préparation de la pérennisation doit démarrer en parallèle de la préparation de la production, et doit se phaser avec les grandes étapes d'un projet spatial

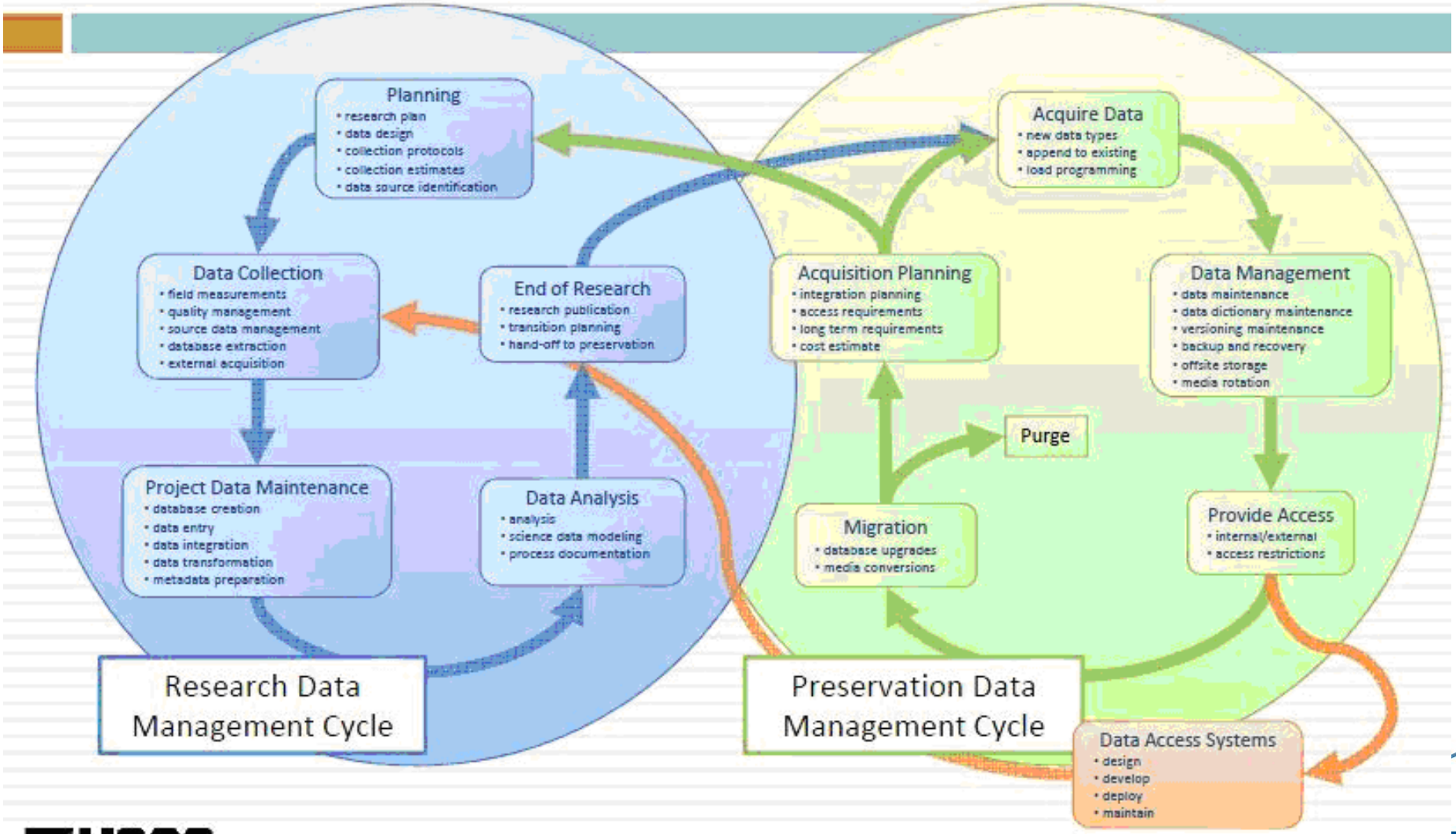

Production des données
Projet spatial
OAIS
PAIMAS
PAIS
Pérennisation des données
MOU,
convention CNES/CNRS
accords inter organismes,
politique de données …
•Responsabilité du CNES/archivage long terme
Spécification
Phase A (Faisabilité)
Désignation archive pérenne
Données, volumétrie
Définition préliminaire
Dossier de faisabilité
Evaluation
Budget archivage pérenne
Approbation
•Plan de développement projet d'archivage
•Plan d'archivage préliminaire
•Accord préliminaire
Définition
Phase B (Déf prélim)
Données à archiver, doc, identifiant, formats …
Définition Formelle
•Dictionnaire de données
•Description formelle de données
•Préparation de métadonnées
•Inventaire
Phase C (Déf détaillée)
Phase D (dev)
• Plan d'archivage complet
• Spécification logiciels
•Accord formel
Développement
Création Simulation
Logiciels validés
Plan d'archivage définitif
Système d'archivage validé
Phase E (exploitation)
•Inventaire complété
•Préparation et ingestion de métadonnées
•Exploitation du système d'archivage, mise à disposition des données
Transfert Validation
Transfert Validation
archives courantes
archive pérenne
Phase F (retrait)
Données à pérenniser, stabilisées, validées
Préparation de la production
Production
Préparation de la pérenisation
Pérenisation

Cycle de vie des données, exemples

- Nombreux travaux dans la littérature concernant la modélisation du cycle de vie des données
- Différents points de vue (scientifique/utilisateur, gestionnaire archive courante, gestionnaire archive pérenne)
- Représentations complexes

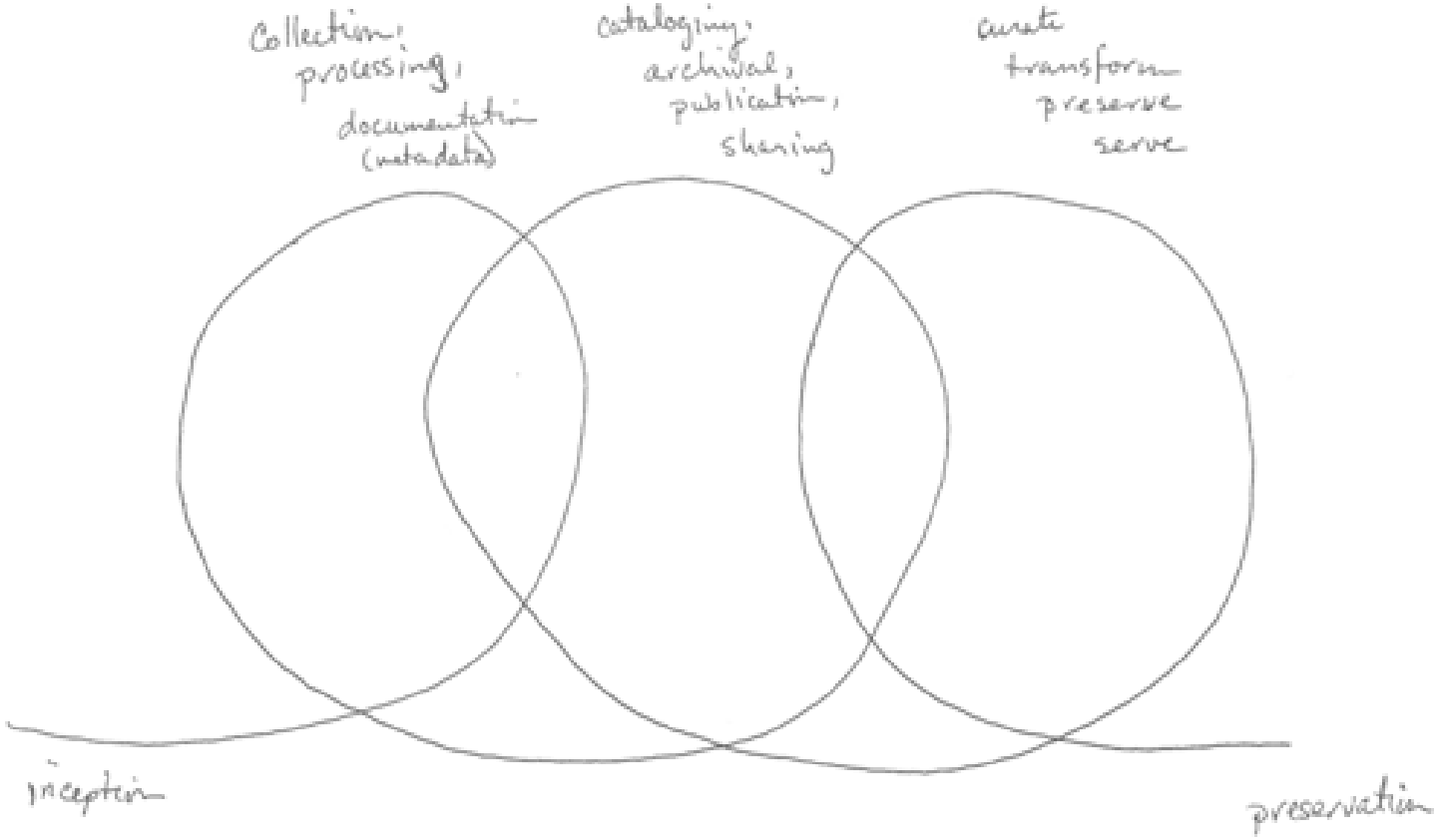
USGS Data Management Plan Framework (DMPf) – Smith, Tessler, and McHale, 2010 [Climate Effects Network (CEN) and Alaska Science Center (ASC)]

Research and Preservation

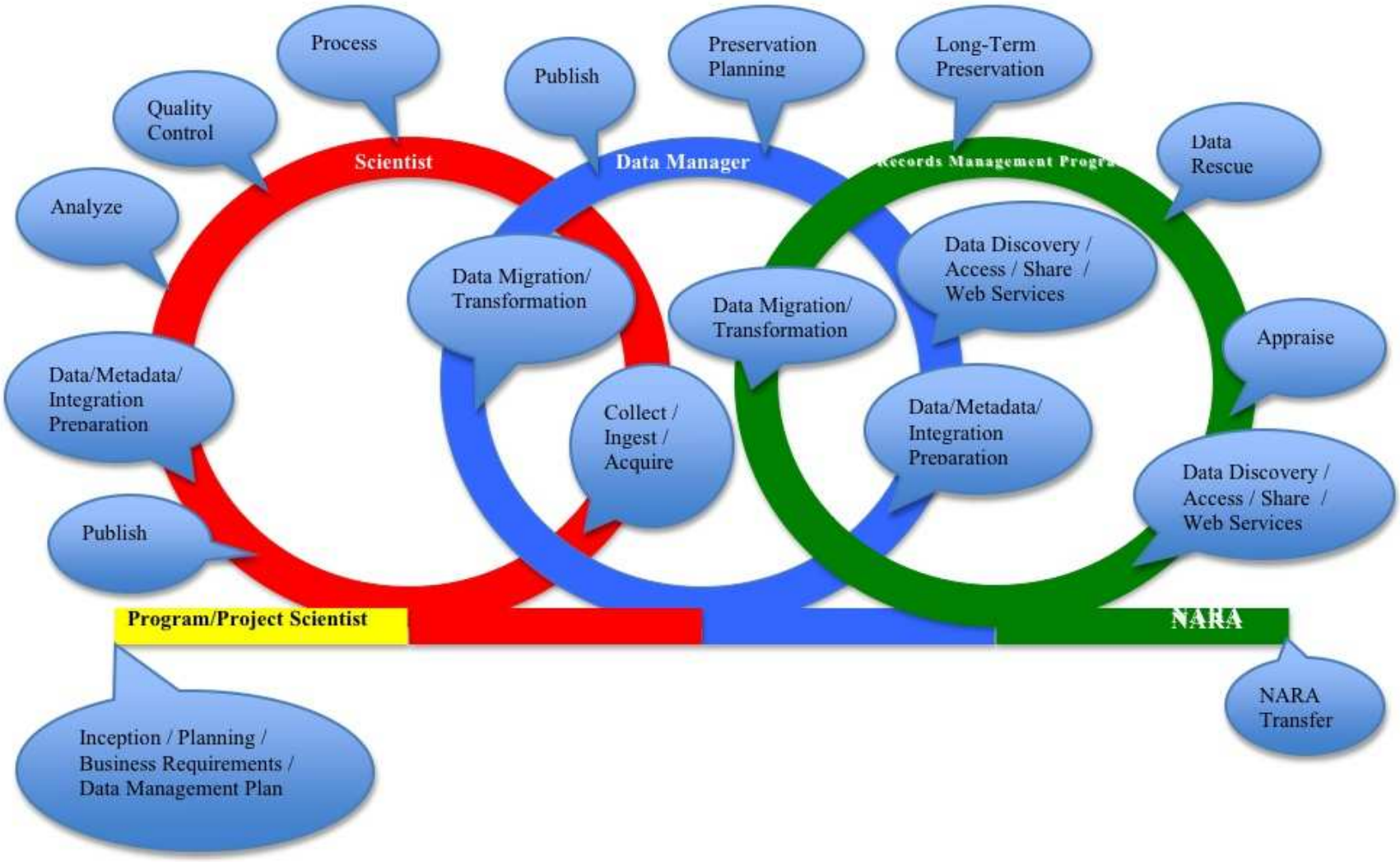


USGS

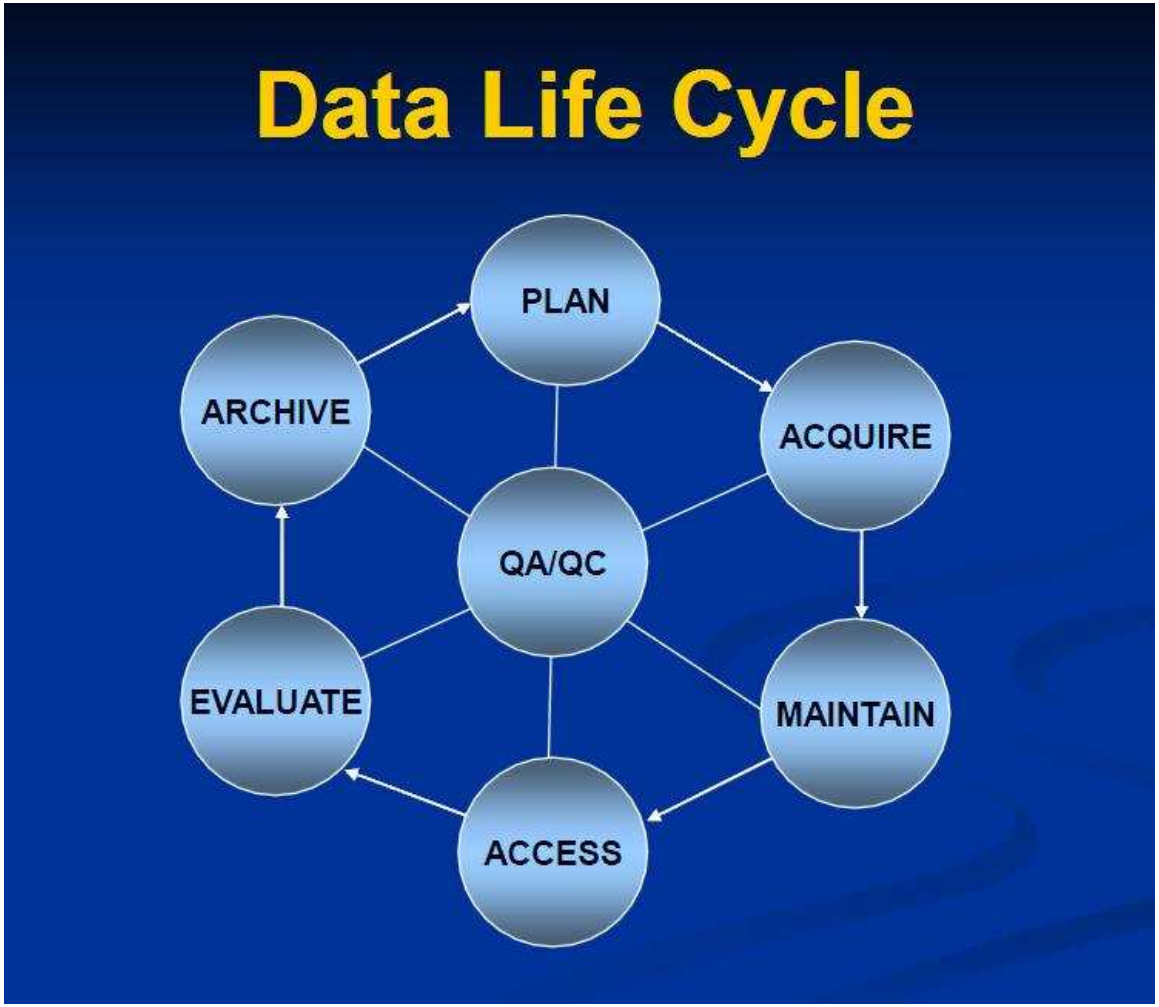
THE ELLYN MONTGOMERY, USGS, DATA LIFECYCLE DIAGRAM



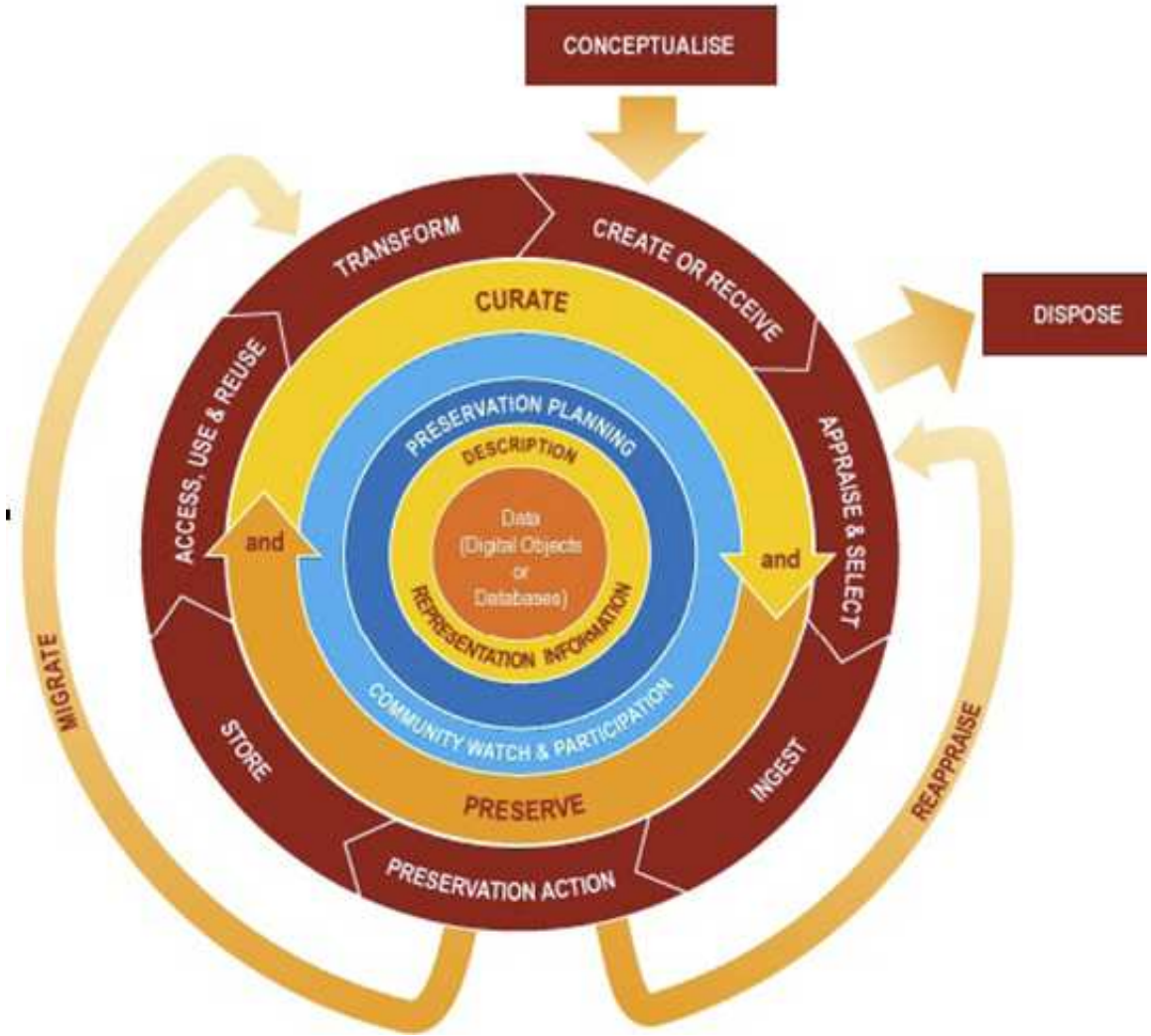
USGS Data Lifecycle Model– John Faundeen & Ellyn Montgomery “Spins”



BLM Data Management Handbook



Digital Curation Centre Lifecycle Model



Conclusion

- Axes stratégiques
  - Maîtriser la gestion du cycle de vie de l'information et des données
    -> directives, procédures
  - Maitriser l'augmentation des volumétries des données et les coûts de transport/stockage/archivage/traitement/accès.
    - » Concentrer données et traitements
    - » Limiter les déplacements de données
    - » Favoriser l'accès à l'information « déspatialisée »
- Enjeux
  - Pérennisation et valorisation des données spatiales
  - Interopérabilité des centres