Comment décrire un corpus à des fins d'archivage?

- Un corpus c'est quoi ?
- Archiver
  - Archiver c'est quoi ?
  - Où archiver ?
  - Comment archiver ?
- Décrire
  - Pourquoi décrire ?
  - Quoi décrire ?
  - Comment décrire ?
- Exemple de la plate-forme Cocoon
  - C'est quoi ?
  - Comment on décrit ?
  - Qui fait quoi ?
  - L'évaluation ?

Un corpus c'est quoi ?

- Un ensemble indissociable de données
  - Des données primaires (textes, images, enregistrements de parole)
  - Des annotations de ces données (transcriptions, traductions, analyses linguistiques, etc.)
  - De la documentation décrivant le contexte de production de ces ressources (lieux, dates, intervenants, etc.)

Archiver c'est quoi ?

- L'archivage couvre toutes les opérations sur les données de leur production ou de leur collecte jusqu'à leur éventuelle destruction
  - Première période : on a besoin du Corpus dans le cadre de sa recherche. On cherche donc des éléments de sécurité et des outils/services facilitant son travail (redondance, gestion de versions, contrôle d'intégrité, logiciels de traitement, etc.)
  - Deuxième période : on conserve le corpus comme un élément d'une démarche scientifique (reproductibilité des observations, cumul des connaissances). On cherche alors des outils/services facilitant l'accès, le partage, le référencement.
  - Dernière période : on peut se poser la question de détruire le corpus car il n'y a plus d'usage dessus ou de le conserver définitivement parce qu'il existe d'autres usages que l'usage initial. On cherche alors à faciliter sa réutilisation par d'autres communautés (documentation, mise en contexte).

Où archiver ?

- Pour les 2 premières périodes
  - C'est le producteur qui reste responsable (l'institution recherche) de la conservation de ses données.
- Pour la dernière période de conservation définitive
  - Il s'agit d'une fonction régalienne confiée au réseau des archives :
    - Le ministère de la recherche a confié au CINES la conservation définitive des thèses au format numérique. Mission élargie à la conservation intermédiaire d'archives publiques sous condition de l'obtention d'un agrément du ministère de la culture.
    - La TGIR Huma-Num a négociée pour les SHS une convention avec le CINES pour l'archivage intermédiaire des données de la communauté.
    - Le CINES reçoit des données en provenance de plate-formes mutualisantes (Cocoon, Archeo-vision, EFEO, etc.), effectue des contrôles techniques, prends en charge leur pérennisation (migration de formats et de supports) sur la période intermédiaire avant de les transférer aux Archives nationales

Comment archiver ?

- La solution : ne pas rester tout seul
- Suivant le moment du projet
  - S'adresser aux plate-formes de stockage, de partage, de diffusion
  - S'adresser
    - aux plate-formes connectées au CINES
    - directement aux institutions de conservation (archives, bibliothèques, musées)

Pourquoi décrire ?

- Parce que les responsables des données (conservateurs, gestionnaires des plate-formes) en ont besoin pour gérer dans le temps la conservation des données et leur communication en garantissant
  - Leur authenticité (de qui proviennent les données, quand sont-elles entrées dans le système)
  - Leur intégrité (pas d'altération des données, lutte contre la falsification)
  - Leur lisibilité (les données restent exploitables et compréhensibles)
  - La traçabilité des opérations touchant la donnée (Par exemple lors des changements de formats ou lors des communications)
- Parce que les utilisateurs en ont besoin pour découvrir son existence et pour interpréter correctement l'information trouvée

Quoi décrire ?

- Pour qu'un corpus soit compréhensible et réutilisable il faut décrire :
  - Son contenu intellectuel (qui, quoi, où, quand, etc.)
  - Ses informations de gestion (communicabilité, licence d'utilisation, etc.)
  - Sa forme (format, codage, dictionnaire de données, etc.)
  - Son contexte de production (organismes ayant participé, objectifs de la recherche, etc.)

Comment décrire ?

- Il existe différents codages permettant d'exprimer cette description
  - EAD, MODS, Dubin-Core, TEI-header, OLAC, CIDOC-CRM, etc.
- Il existe différents outils/plate-forme pour aider à la saisie de ces informations
  - ATOM, Nakala, Cocoon, Ortolang, etc.
- Il existe aussi des personnes (documentalistes, bibliothécaires, archivistes) qui peuvent faire ou aider à faire cette description.
  - À la BnF, aux Archives nationales, dans les centres de documentation et les bibliothèques.

Exemple de la plate-forme Cocoon

- Définition / réduction du périmètre (contraintes)
  - La ressource primaire est forcément un enregistrement de parole.
  - Les données doivent être numériques (nativement ou numérisées)
  - Les déposants doivent faire partie de la communauté SHS en France
  - Les déposants doivent avoir obtenu l'autorisation de disposer de ces ressources de la sorte

Comment faire la description dans Cocoon ?

- L'interface web de dépôt permet au déposant de décrire les documents déposés avec les champs OLAC. Essentiellement la description du contenu intellectuel, du contexte de production et des droits.
- La description technique (format, codage, taille, etc.) est effectué par la plateforme.
- La description se fait aux différents niveaux (collections et ressources)
- Des catalogueurs de la BnF peuvent préciser et compléter cette description (indexation, liage à des référentiels) ou entrer en contact avec le déposant pour complément d'information

Les modèles de description dans Cocoon

- La description des document (enregistrement, texte, etc.) est faite sur le modèle OLAC (Dublin-core qualifié + vocabulaires contrôlés)
- La description des données de référence (personnes, organisations, lieux, sujets, langues) se fait dans les référentiels
  - Pour les auteurs (référentiel externe VIAF [schema.org])
  - Pour les locuteurs (référetiel interne [foaf])
  - Pour les sujets (référentiel externe RAMEAU [skos])
  - Pour les langues (référentiel externe Lexvo)

Qui fait quoi dans Cocoon ?

- Exemples :
- Le déposant : identification des participants (nom, prénom)
- Le catalogueur : recherche dans les référentiels auteurs (VIAF) et liage.

- Le déposant : rédaction d'un résumé, d'une table des matière ou d'une simple description du contenu.
- Le catalogueur : indexation à partir de cette description ou par lecture/écoute de la ressource primaire quand c'est du français ou de la traduction quand elle est disponible.

Cocoon une plate-forme technique pour quoi faire?

- Les services offerts par la plate-forme adressent les besoins des 3 périodes
  - Temps 1 : Stockage sécurisé, accès contrôlé, aide à la description.
  - Temps 2 : Aide au partage (interfaces d'accès, de consultation et d'interrogation), aide à la citation (identifiants pérennes), protocole de moissonnage (OAI-PMH) et référencement auprès de services (OLAC, Isidore, CLARIN, etc.), facilitation de la réutilisation (exposition en Web de données, rdf, sparql-endpoint, etc.).
  - Temps 3 : Fabrication des paquets d'archivage pour le le CINES et suivi des échanges.

Quelle est la part d'évaluation effectuée dans Cocoon ?

- Outre les contraintes de périmètre, les principes qui sous-tendent l'évaluation par la plate-forme technique tournent autour de la lisibilité technique de l'information
  - Le nommage et le classement des données
  - Le formatage et de codage de l'information
  - La complétude, la précision et la cohérence de l'information
- Il n'y a aucune évaluation scientifique
  - Celle-ci peut passer par la politique éditoriale du responsable de la collection (par exemple la collection « Corpus de la parole » du Ministère de la culture) ou ou par de la labellisation (une labellisation CORLI?).