Huma-Num
la TGIR des humanités numériques

La pérennisation des données numériques

Nicolas Larrousse, Michel Jacobson, Adrien Desseigne, Joel Marchand
Huma-Num
Les rencontres d'Huma-Num – Lyon - 11-14-juin-2018

cnrs Aix*Marseille université CAMPUS CONDORCET Paris–Aubervilliers

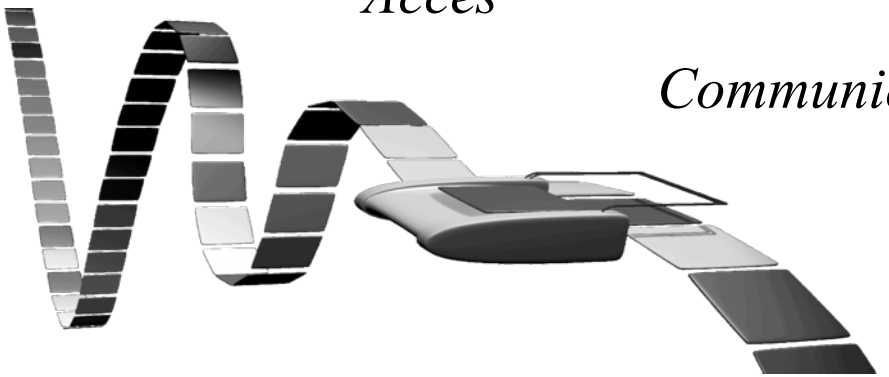
Des Humanités Classiques aux Humanités Numériques

Accès

Natif

Communication

Simulation



Duplication

...

Des données coûteuses à produire

Encodages de textes

Numérisations manuscrits

Modélisations 3D

Transcriptions d'enregistrements

...

Une typologie foisonnante

Textes

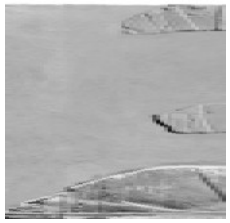
Transcriptions

Enrichissements

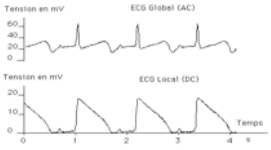
Sons

Vidéos

Images



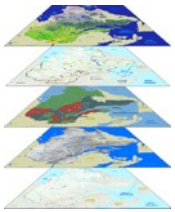
Enregistrements physiologiques



Géomatique

Modèle 3D

Cartes

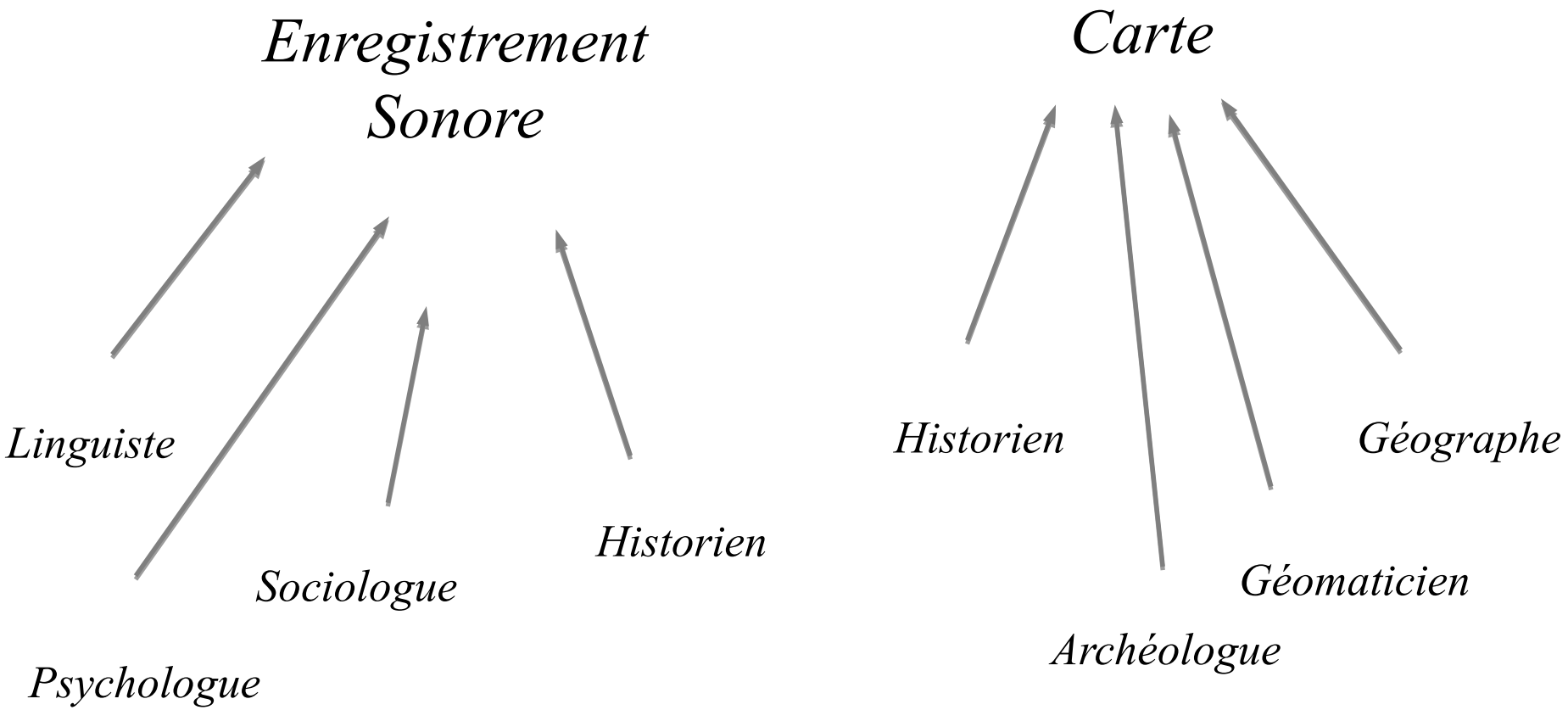


Données SIG

SMS

...

Des usages multiples pour une même donnée

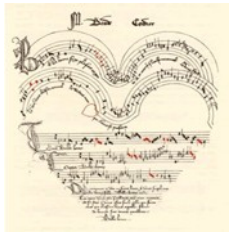


Un intérêt patrimonial

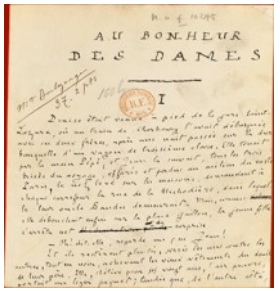
Données de l'archéologie

Enregistrements de langues disparues

Musique



Numérisation de manuscrits plus consultables



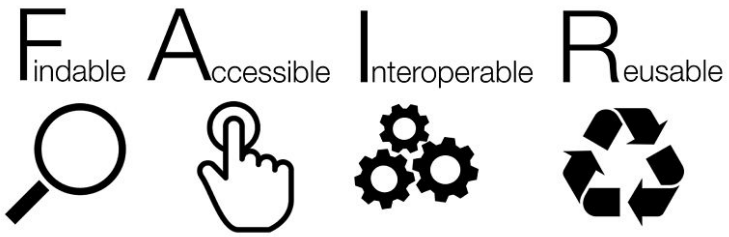
Peu d'intérêt des tutelles

Définition des données de la recherche ?

Travail non valorisé pour la carrière d'un chercheur

Coûts en général non intégrés dans les projets

Un timide avancée avec le
« Data Management Plan »
du programme H2020



...

SERVICES POUR LES DONNÉES NUMÉRIQUES

Huma-Num propose des services pour les données numériques produites en SHS.
A chaque étape du cycle de vie des données correspond un service dédié.

CINES
Centre Informatique National de l'Enseignement Supérieur

STOCKER
Entreposer . Organiser

ARCHIVER
Préservation à long terme

TRAITER
Outils . Logiciels

SIGNALER
Enrichissement sémantique
Accès unifié

isidore

DONNÉES
DE LA RECHERCHE

DIFFUSER
Machines virtuelles
Diffusion web

EXPOSER
Documenter . Partager

nakala
nakalona

Partenariat avec le CC-IN2P3
et le CINES

La pérennisation des données numériques

Comment bien perdre ses données ?

- Vous ne savez pas comment relire une cartouche Iomega JAZ
- Votre disquette 3½ pouces est endommagée
- Vous avez effacé par mégarde le dossier où était votre fichier
- Le codec utilisé pour votre image n’est plus reconnu par aucun de vos outils
- Vous ne savez plus comment se nomme le fichier ni sur quel disque vous l’avez mis
- Vous avez 40 exemplaires d’un même fichier mais vous ne savez pas lequel est le bon
- La lecture de votre fichier par la nouvelle version du logiciel vous donne un autre résultat qu’avant
- Vous ne pouvez plus installer votre unique outil de visualisation sur la nouvelle version de votre système d’exploitation
- Vous revenez 10 ans plus tard sur un de vos fichiers, mais vous ne comprenez plus sa structure ni son contenu que vous n’avez jamais documenté
- ...

La pérennisation des données numériques

Les causes de la perte

- Obsolescence
- Vieillissement
- Mauvaises manipulation
- Manque d’organisation
- Manque de documentation
- ...

La pérennisation des données numériques

1. Organiser ses données
   - Définir un plan de nommage
     - Simple, court, documenté, durable
     - Éviter quelques chausse-trapes (accents, espaces, casse,…)
   - Exemple : le plan de nommage de l'IRHT pour la cotation d'un manuscrit numérisé:
     - 2 chiffres du département
     - 3 chiffres de la commune
     - 2 chiffres indiquant le type de bibliothèque
     - 2 chiffres, séquentiel, allant de 01 à 99
   - Exemple pour Chartre = 280856201_MS1038_t1

La pérennisation des données numériques

1. Organiser ses données
   - Définir un plan de classement
     - Pas trop profond
     - Classificatoire : tout fichier à une place et une seule possible
   - Exemple du plan de classement de l’IRHT
     - Pays/Ville/Etablissement/TIFF/Cote_manuscript/

La pérennisation des données numériques

1. Organiser ses données
   - Choisir ses formats
     - Ouvert, libre, normalisé, largement utilisé, existence d’outils de contrôle de conformité
   - Cf. Guide méthodologique du CINES et la liste des formats acceptés (facile.cines.fr)

La pérennisation des données numériques

2. Stocker ses données
   - Avoir plusieurs copies des données
     - Sur des lieux distants (lutte contre des sinistres type incendie, inondation…)
     - Sur des types de support et ou d'accès distincts (lutte contre les risques de marché [éditeur défaillant ou arrêtant son support] ou d'attaques)
   - Définir une politique de sauvegarde
     - Combien de copies sont gardées ? Combien de temps ? À quelle périodicité ? (lutte contre les fausses manipulations)
   - La chasse aux doublons
     - Les différentes copies ne doivent pas être isolées mais répondre à une organisation, une synchronisation et idéalement offrir une vue unifiée sur les données

La pérennisation des données numériques

3. Décrire ses données
   - Les données doivent être associées à des descripteurs
     - Métadonnées de provenance, de contexte, d'identification, d'intégrité, de droits, de codage, etc.
   - Il existe des standards pour exprimer ses métadonnées
     - Dublin-Core
     - EAD
     - CIDOC-CRM
     - EDM
     - …
   - Il vaut mieux expliciter dans un simple format texte non standardisé que de garder ces informations dans sa tête.

La pérennisation des données numériques

4. La pérennisation à long-terme
   - Objectif : maintenir l'authenticité, l'intégrité la traçabilité et la lisibilité de l'information
     - Migrations des supports et des formats (lutte contre le vieillissement et l'obsolescence)
     - Apporter la preuve que l'information n'a pas été altérée. Le cas échéant, restauration de données dégradées à partir de copies intègres
     - Journalisation de toutes les opérations sur les archives
     - Apporter la preuve de la provenance (qui, quand, quoi) des informations

La pérennisation des données numériques

Le dispositif en collaboration avec le CINES

- Couvre la partie 4
- Demande à ce que les parties 1 à 3 soient déjà franchies
- S’ajoute une partie administrative (ouverture de compte), une explicitation des règles de gestion (statut des archives, durée de conservation, communicabilité), une partie protocolaire (organisation des échanges), éventuellement des travaux préalables pour la prise en charge de nouveaux formats
- Représente pour le producteur l’avant chambre des Archives nationales
- Il s’agit d’un transfert de la responsabilité de la conservation

La pérennisation des données numériques

Pérenniser des données numériques

- Ne peut pas se faire sans le producteur. C’est lui qui en connaît le contenu, le contexte de production/collecte, l’usage envisagé.
- Ne peut se faire tout seul. Il faut d’autres compétences que scientifique ou du domaine. Compétences techniques numériques, compétences archivistiques
- Le coût humain est important
- Les délais sont généralement assez longs : on compte plutôt en années

La pérennisation des données numériques

Métaphores possibles

- L’archivage comme l’alpinisme. On ne part pas seul, on s’entraîne avant, ça va être dur, ça va être long, ça va être coûteux. On y va par étapes progressivement : promenade en pleine, en montagne à vaches, petite course. Bien maîtriser chaque étape avant de passer à la prochaine.
- Photo de classe ou de famille dont plus personne ne sait qui est dessus. Ce qu’on transmet aux futurs chercheurs doit être décrit maintenant avant que les souvenirs ne s’effacent.
- Pyramide des besoins : la base correspond à la production de données, le sommet à ce qui est préservé pour les futures générations.
- INA : 10 % du budget pour l’infrastructure, 90 % pour la description.

Le rôle de l'infrastructure Huma-Num

Huma-Num

CINES Centre Informatique National de l'Enseignement Supérieur

Un projet pilote basé sur les données orales

Un besoin « urgent »

S'appuyer sur des ressources existantes et les adapter

- *Regrouper les objets en collections*
- *Prise en compte de nouveaux formats*
- *Metadonnées spécifiques*
- *Notion de version*

Fonctionner sur le long terme dans un environnement instable …

Le rôle de l'infrastructure Huma-Num

Huma-Num

Identification des besoins

Repérage en collaboration avec les communautés

Identification des formats de données et de metadonnées

Accompagnement des projets d'archivage

Sélection et organisation des données à préserver

Appui technique

Faire le lien avec les autres organismes

Lien avec les institutions d'archives et les autres partenaires

Sensibilisation des tutelles et des agences de financement

Huma-Num

Quelques effets secondaires attendus du service de préservation

Une prise en compte de l'importance des données dès le début du projet

Une incitation à produire de « meilleures » données

Préparer l'interopérabilité des données

Quelques perspectives pour la préservation

Un évidente fragilité des données

Les migrations permanentes des données ne les préservent pas

Des pertes par simple ignorance

Une prise de conscience nécessaire

Informer les producteurs dès le début du projet

Promouvoir de bonnes pratiques