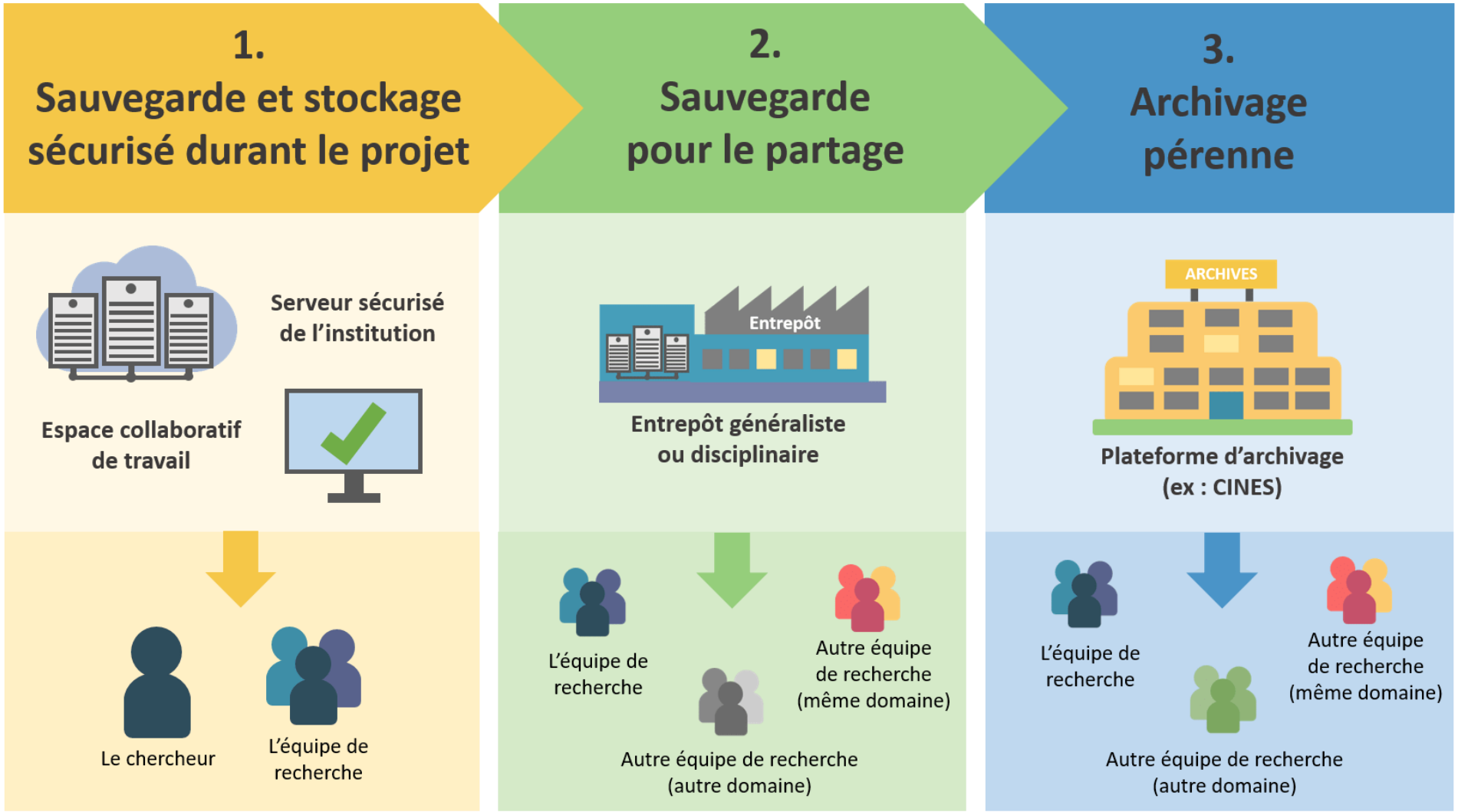
Daniele Franco
Humathèque Condorcet (Aubervilliers)
Centre Gilles Gaston Granger (Aix-en-Provence)

Le circuit des données dans la constitution d'un fonds de documents numérisés

- Les opérations préalables
- Le circuit de numérisation
- Le stockage
- L’exploitation

- Les données de la recherche peuvent être définies comme "*des* ***enregistrements factuels*** *(chiffres, textes, images, sons) utilisés comme* ***sources principales pour la recherche scientifique*** *et généralement reconnus par la communauté scientifique comme nécessaires* ***pour valider les résultats de la recherche****. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche* » (OCDE, 2007).



Source : https://doranum.fr/stockage-archivage/les-trois-niveaux-de-sauvegarde-des-donnees-de-la-recherche/

Entrepôts de données

- Une base de données destinée à accueillir, conserver, rendre visibles et accessibles des données de recherche.

- Démarche de partage et d'ouverture des données selon les **principes FAIR** pour que les données soient "Facile à trouver, Accessible, Interopérable et Réutilisable" (en anglais : Findable, Accessible, Interoperable, Reusable).

- Le dépôt d'un jeu de données s'accompagne de la saisie ou de la collecte d'informations (les métadonnées) sur les données déposées facilitant la compréhension et l'interprétation de ces données.

Ex : Les métadonnées standard comme celles du format Dublin Core permettent de décrire l'auteur, le titre, l'année de création… d'un jeu de données. Un entrepôt propose généralement un ensemble de métadonnées spécifiques aux sujet, thème, discipline (par exemple données biologiques, astronomiques, environnementales, etc.) des données qu'il accueille.

Idref

IdRef (Identifiants et Référentiels) est une application Web permettant à des utilisateurs et à des applications clientes d’interroger et consulter les autorités des catalogues Calames, Sudoc et theses.fr.

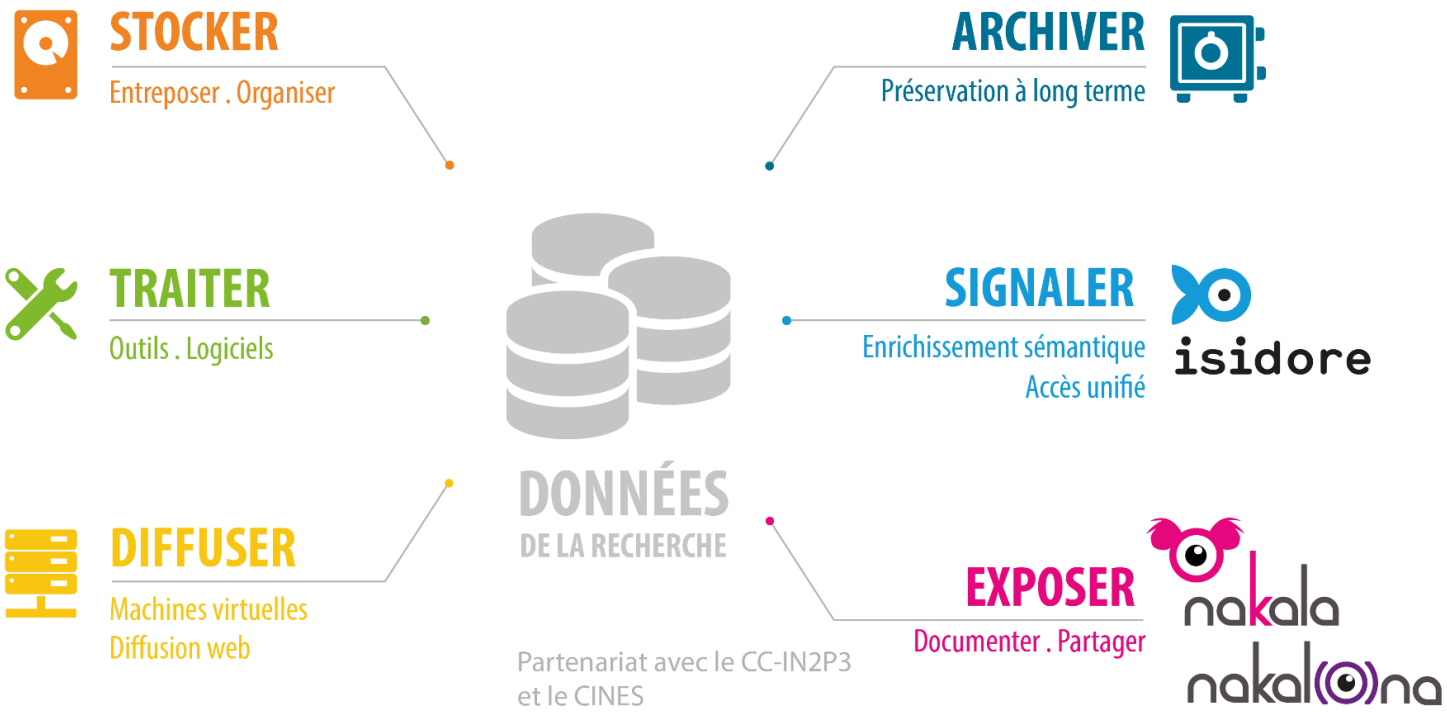
Geonames

Base de données géographiques qui couvre tous les pays avec plus d’onze millions de noms de lieu en libre accès.

Humanum

- La **TGIR Huma-Num** propose un ensemble de **services et outils** pour les données numériques produites dans les projets de recherche en SHS.
- Ces services et outils sont construits sur un ensemble de **technologies d'infrastructure** (serveurs) et de **systèmes informatiques** pour mutualiser, diffuser et stabiliser l'accès aux données et documents.
- Loin d'être un simple guichet, l'hébergement et la diffusion des données se font, au sein de la TGIR, en responsabilisant les équipes sur le rôle qu'elles ont à jouer dans la pérennisation des données et des outils de traitement qui leurs sont associés. **La mission première est d'assurer la préservation du patrimoine scientifique des laboratoires**, et plus particulièrement des données et documents acquis ou réalisés dans le cadre d'opération de recherche. **Cette mission sous-tend également une stratégie économique visant à diminuer les coûts récurrents,** par la mise en commun d'une infrastructure en co-gérant des outils, instruments et systèmes de gestion des données.

SERVICES POUR LES DONNÉES NUMÉRIQUES



Source : Services et outils https://www.huma-num.fr/services-et-outils

Nakala

- NAKALA est un entrepôt de données de recherche destiné à accueillir, conserver et rendre visible et accessible les données de recherche en SHS

- Ce service HumaNum est proposé en partenariat avec le Centre de Calcul de l'IN2P3

- Nakala permet à une équipe de projet en SHS de déposer des données numériques documentées dans un entrepôt sécurisé afin de les partager

Les services proposés par NAKALA

Accès et citation
par un identifiant pérenne
(handle)

Stockage sécurisé

Accès permanent

Services d'accès aux données

Exposition dans un
Triple Store RDF

Accès via un
entrepôt OAI-PMH

Services interopérables
de présentation des métadonnées

Ces services sont proposés en partenariat avec le CC-IN2P3.

Source : Exposer ses données avec Nakala https://www.huma-num.fr/services-et-outils/exposer

Les acteurs identifiés dans la chaîne de numérisation

Le chargé de projet

•Coordonne le projet,

•Interlocuteur entre le GED et le chercheur,

•Identifie les besoins du chercheur,

•Analyse la faisabilité du projet (sélection du corpus en lien avec la politique de numérisation, gère la question juridique, précise les besoins techniques et le niveau de qualités des descriptions des métadonnées) en lien avec le chargé des opérations de la numérisation et le responsable de la bibliothèque numérique,

•Cadre le projet et suit l'avancée du projet,

•Transmet la commande au chargé des opérations de numérisation.

Le chargé des opérations de numérisation

•Coordonne les opérations de la chaîne de numérisation,
•S’assure de la qualité des métadonnées qui lui sont transmises et de la cohérence avec les modèle de données de la bibliothèque numérique,
•Donne les consignes pour la préparation matérielle du corpus en fonction de la commande
•Réalise le cahier de spécification en lien avec le chargé de projet,
•Importe les métadonnées dans Numa Hop,
•Réalise le mapping si besoin dans Numa Hop,
•Exporte les métadonnées dans le serveur de stockage,
•Interlocuteur avec le prestataire (bordereau d’enlèvement et de réception, bon de commande, rejet des scan,etc…),
•Valide le pré-contrôle des images et valide les métadonnées,
•Importe les scans dans le serveur de stockage.

Le chargé de la qualité des métadonnées

- Fournit les métadonnées existantes,
- Enrichit les métadonnées en fonction de la commande.

Le chargé du récolement et du constat d’état

- Réalise le récolement sur un tableur,
- Réalise le constat d’état dans Numa Hop ou sur le tableur de récolement.

Le chargé de la numérisation

- Scanne et produit les métadonnées techniques,
- Reprend les rejets,
- Interlocuteur avec le chargé des opérations.

Le chargé du contrôle qualité des scans

- Contrôle les scans dans Numa Hop,
- Contrôle la réception physique des documents et leur état matériel.

Le chargé de la qualité des ocrs

- Contrôle la qualité des ocrs,
- Corrige les ocrs,
- Transmet les fichiers corrigés au chargé des opérations.

Le chargé de la bibliothèque numérique

•Définit les prérequis techniques pour diffuser dans la bibliothèque numérique (modèles de données et les formats de fichiers à diffuser,
•Importe les métadonnées et les scans dans la bibliothèque numérique via Huma Num Box.
•En fonction du projet, une seule personne peut être amener à jouer tous les rôles.

Les numérisations Collex-Persée

- Le GT numérisation Collex réalise une étude de besoins auprès des différents types de structures documentaires (BU, BIU, BM, Archives, Laboratoires, Grands Établissements etc.). Cette étude vise à repérer des outils de gestion de numérisation, de production et de diffusion de corpus numérisés, en vue de les mutualiser ou de les faire évoluer pour faciliter les projets de numérisation.
- Cette étude permet de nourrir la réflexion pour favoriser l'émergence au niveau national d'une chaîne de numérisation enrichie solide, constituée de briques techniques interopérables et/ou d'une ou plusieurs plateformes ouvertes. Cette chaîne pourrait être utilisée par les structures documentaires lors de leurs projets de numérisation, pour une ou plusieurs étapes du projet.

Trois axes de développement sont identifiés :

- Le soutien à la production et à la diffusion de corpus liés aux programmes de recherche (notamment via les appels à projets collaboratifs en 2018, 2019 et à venir en 2021), ainsi que la promotion d'une expertise des outils, en vue de leur recommandation (par exemple, le soutien à NumaHop).
- Le portage de programmes disciplinaires d'envergure nationale : en privilégiant une approche massive, par la mise à disposition des collections, publications, archives, thèses et en travaillant sur les outils et la priorité de convergence EAD et TEI.
- La promotion de la bibliothèque comme lieu de recherche et d'expertise, notamment par la mise en œuvre d'un appel à projets «résidence» en 2020.

Collection 1

Projet de numérisation d’un grand fonds de photographies issues des terrains effectués par les membres du laboratoire aux années 1960 et 1970

Collaboration active avec le laboratoire pour le classement et la description

Points de vigilance ou de l’importance du PGD

Métadonnées pas exploitables car incomplètes.

Pour cela une personne a été embauchée durant l’été afin d’enrichir les champs sujets et localisation géographique

Le projet échoue à cause d’un problème lié à la cession des droits à l’image.

Importance de se doter d’un plan de gestion de données établi avec tous les acteurs impliqués dès le début du projet.

Plan de gestion des données (DMP)

Le plan de gestion de données est rédigé en **amont du projet**, pour anticiper les différentes étapes. Il s'agit d'un document qui est amené à **évoluer**, afin de prendre en compte les avancées du projet pour préciser ou développer certains aspects. Il permet de définir les différentes modalités de gestion pour aider à répondre aux principes FAIR (*Findable, Accessible, Interoperable, Reusable*).

Plan de gestion des données (DMP)

- Tutoriel en ligne : https://doranum.fr/tutoriel-sur-loutil-de-redaction-dmp-opidor (DORANUM est la plateforme de formation en ligne sur les Données de la recherche, gérée par l'INIST et le réseau des URFIST)
- Manuel papier : https://www.pasteur.fr/fr/file/34473/download
- Vidéos : https://oaamu.hypotheses.org/2208

Collection 2

Après avoir pris contact avec un laboratoire qui souhaitait numériser un fonds d’archives de la recherche, nous n’avons pas donné suite au projet car il manquait les autorisations nécessaires.

Une autre raison bloquante était l’état de conservation du fonds, qui ne permettait une visibilité globale sur les documents dans la phase préalable du projet.

Collection Mai 68

- Fonds d’archives, numérisation financée par Collex
- Les métadonnées sont extraites de la plateforme Calames
- Plus de 30k images, elles sont associées à des métadonnées au niveau de dossier, pour différencier les images une chaine de caractères est crée
- La navigation dans la collection est extrêmement compliqué, car le logiciel de gestion d’une bibliothèque numérique n’est pas conçu pour afficher l’arborescence des archives

Calames

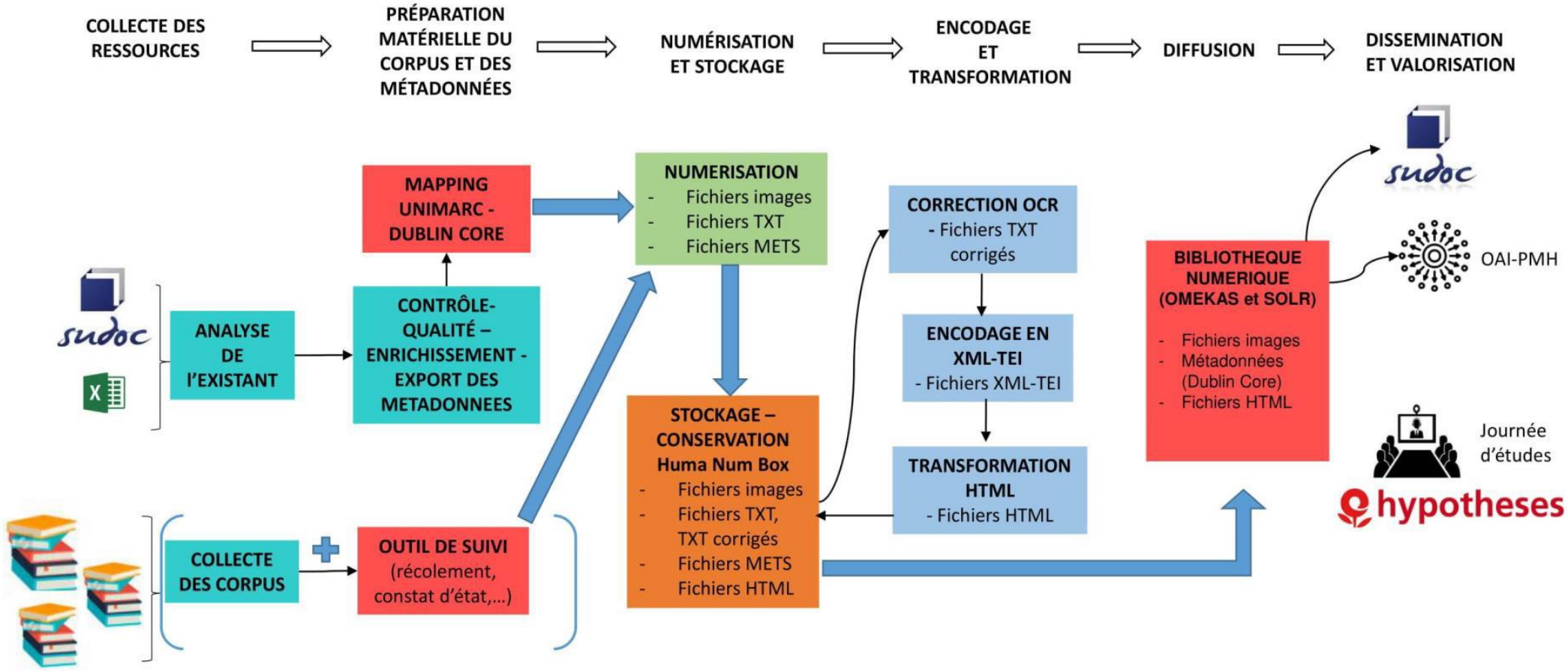
- Calames est le catalogue des archives et des manuscrits des bibliothèques universitaires françaises, mais aussi de grands établissements nationaux (Institut de France, Muséum d'histoire naturelle,...) et de plusieurs établissements de recherche (Bibliothèque Littéraire Jacques Doucet...).

- C'est un catalogue vivant, qui s'enrichit des nouvelles acquisitions et du travail de description effectués par ces établissements ; de nombreux instruments de recherche sont en cours de constitution ou partiellement publiés.

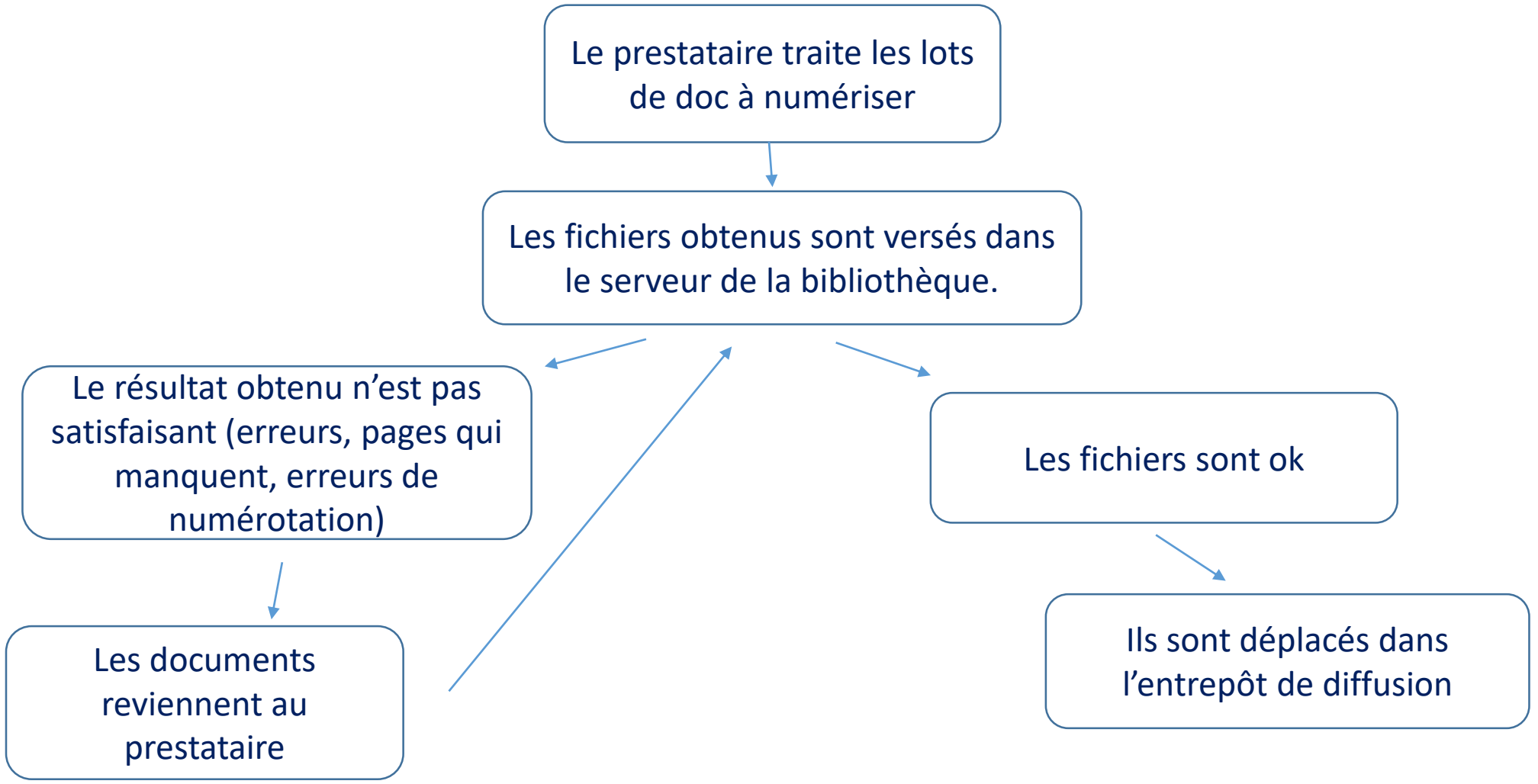
- Dans Calames les collections sont représentées avec leur arborescence mais celle-ci s’arrête au niveau de dossier. Pour une bibliothèque numérique le niveau de métadonnée requis est à la pièce

Collection Palibr

- Collection de brochures issues de différentes bibliothèques
- Numérisation Collex
- Enrichissement en parallèle des métadonnées par les collègues du laboratoire partenaire qui connaissent le fonds


COLLECTE DES RESSOURCES
PRÉPARATION MATÉRIELLE DU CORPUS ET DES MÉTADONNÉES
NUMÉRISATION ET STOCKAGE
ENCODAGE ET TRANSFORMATION
DIFFUSION
DISSEMINATION ET VALORISATION
sudoc
ANALYSE DE l'EXISTANT
MAPPING UNIMARC - DUBLIN CORE
CONTRÔLE-QUALITÉ – ENRICHISSEMENT - EXPORT DES METADONNEES
NUMERISATION
- Fichiers images
- Fichiers TXT
- Fichiers METS
COLLECTE DES CORPUS
OUTIL DE SUIVI (récolement, constat d'état,…)
STOCKAGE – CONSERVATION Huma Num Box
- Fichiers images
- Fichiers TXT, TXT corrigés
- Fichiers METS
- Fichiers HTML
CORRECTION OCR
- Fichiers TXT corrigés
ENCODAGE EN XML-TEI
- Fichiers XML-TEI
TRANSFORMATION HTML
- Fichiers HTML
BIBLIOTHEQUE NUMERIQUE (OMEKAS et SOLR)
- Fichiers images
- Métadonnées (Dublin Core)
- Fichiers HTML
sudoc
OAI-PMH
Journée d'études
hypotheses

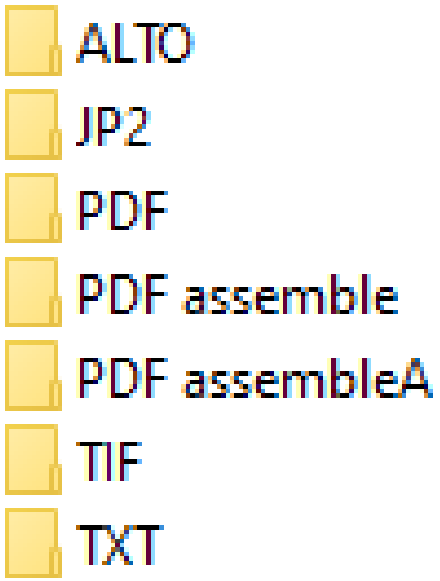
Circuit de numérisation



Numérisation

Nous avons fait recours à un prestataire externe pour la numérisation du fonds en question

Le prestataire enrichissait les md des images selon les indications données et produisait pour chaque images le jeu de données suivant :



Les différents formats d’image livrés

- ALTO (*Analysed Layout and Text Object*) est un standard XML permettant de rendre compte de la mise en page physique et de la structure logique d'un texte transcrit par reconnaissance optique des caractères (OCR). Il est très adapté à la conservation à long terme des données issues de la conversion ; il permet une réutilisation ultérieure du mode texte.

- JP2000 pour la diffusion sur la bibliothèque numérique. La compression utilisée dans JP2 permet de stocker des images avec ou sans perte de qualité, ce qui n’est pas possible dans les fichiers JPEG standard dans lesquels chaque compression de l’image diminue sa qualité. La taille du fichier final est réduite par rapport à un niveau de conversion similaire.

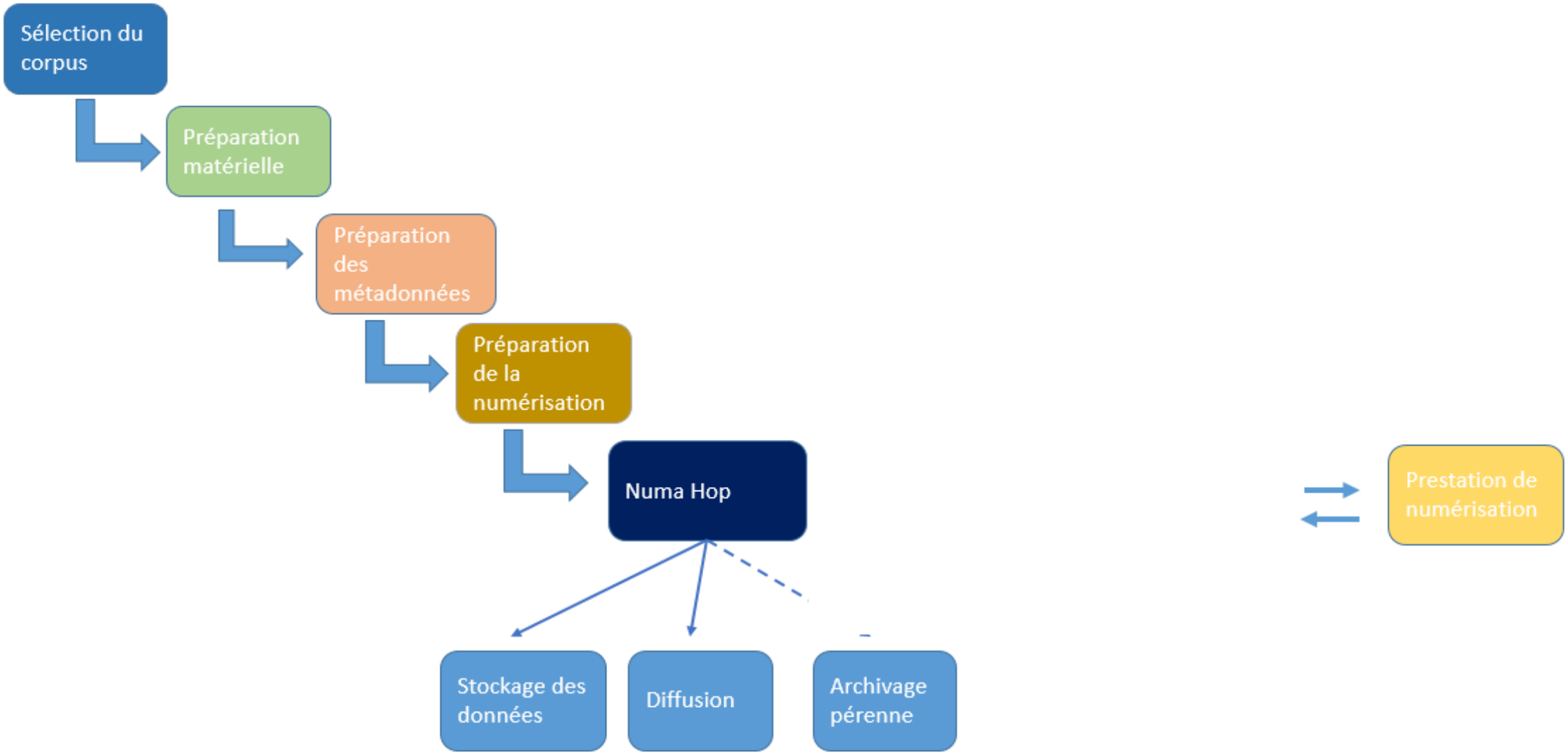
Numahop

NumaHOP permet de gérer une chaîne de numérisation de documents de l'import des notices et du constat d'état des documents physiques à la diffusion et à l'archivage grâce à un interfaçage largement automatisé entre les différentes étapes de la numérisation impliquant les acteurs concernés (prestataires de numérisation, bibliothèques, diffuseurs, CINES).
Le bénéfice de cette réalisation est triple :

- privilégier l'usage de formats normalisés ;
- favoriser la standardisation des méthodes de travail ;
- permettre la mutualisation et l'échange des savoir-faire entre les établissements qui utilisent cette plate-forme.

NumaHOP est composé de plusieurs modules fonctionnels permettant :

- de convertir des notices au format UNIMARC ou EAD dans des formats interopérables : Dublin Core, Dublin Core qualifié ;
- de réaliser des constats d'état pour les lots de documents à numériser envoyés vers les prestataires de numérisation ;
- de recevoir les lots numérisés par le prestataire (images et métadonnées) et de les contrôler ;
- d'utiliser des fonctions de workflow, de contrôle et de structuration des projets ;
- de valider les unités documentaires numérisées (images + métadonnées) et de les exporter vers les diffuseurs et les archiveurs ;
- de produire des fichiers OCR, METS, images dérivées...


Sélection du corpus
Préparation matérielle
Préparation des métadonnées
Préparation de la numérisation
Numa Hop
Stockage des données
Diffusion
Archivage pérenne
Prestation de numérisation

La sauvegarde des données se fait actuellement dans Huma Num Box. Huma Num Box est un système de stockage distribué proposé par la TGIR Huma-Num. Dans ce cadre, l'architecture de stockage s'appuie sur :

- un stockage de haute sécurité et redondé ;
- des solutions de sauvegarde sur bandes magnétiques de l'IN2P3 (Institut national de physique nucléaire et de physique des particules) ;
- une copie mensuelle sur le site de Paris d'Huma-Num.

Exploitation des données

- Partage
- Diffusion
- Valorisation

Omeka S

- Omeka S est une nouvelle version du logiciel Omeka qui vient compléter plutôt que remplacer Omeka Classic. Bien que les principes et objectifs soient les mêmes, Omeka S a été complètement réécrit. Cette version est plus particulièrement destinée aux institutions qui gèrent plusieurs sites et qui souhaitent publier des données liées (*linked data*). La conception du logiciel a été initiée en 2012, sept versions alpha (2015-2016) et quatre versions bêta (2016-2017) ont été nécessaires avant d'aboutir à la version stable 1.0.0 en novembre 2017.
- Plutôt que de définir une requête qu'Omeka S exécute pour déterminer quels *items* se trouvent dans un site, les *items* peuvent désormais être attachés et détachés un par un, de sorte qu'il est possible d'avoir un contrôle fin et direct sur les *items* qui apparaissent dans tel ou tel site.

Omeka S

Principales caractéristiques d'Omeka S :

- Gestion d'un nombre illimité de sites
- Rôles : superadmin, administrateur des sites (CMS), administrateur des données, auteur, ...
- Focus sur l'échange de données
- Conçu avec les principes du web des données (format natif JSON-LD, URI, description des ressources à l'aide de vocabulaires RDF)
- Interoperabilité avec d'autres systèmes (connecteurs Fedora et DSpace, API REST)

Les expositions dans Omeka S

Omeka S permet de créer des pages d’expositions virtuelles grâce notamment au resolveur des liens IIIF .

IIIF se base sur le manifest, un ensemble de données interopérables qui permet aux visionneuses d’afficher des images sauvegardées dans des entrepôts différents.

Module IIIF

- IIIF – *International Image Interoperability Framework*™ – désigne à la fois une **communauté** et un **cadre d'interopérabilité** pour diffuser, présenter et annoter des images et documents audio/vidéo sur le Web. Il s'est imposé en quelques années comme un standard et une brique technologique essentielle pour décloisonner les collections numérisées des institutions patrimoniales à l'échelle mondiale.

- Dans cet environnement distribué et interopérable, chaque entrepôt devient un point d'accès distant potentiel auquel des applications tierces peuvent se "brancher" et réutiliser les ressources à d'autres fins, sans avoir à les dupliquer et sans en perdre le contexte.
- En rendant accessibles de vastes corpus d'images et de ressources audiovisuelles, IIIF agit comme un facilitateur pour de nombreux projets et applications utilisant les collections des institutions culturelles et susceptibles de toucher des publics variés : portails thématiques, expositions virtuelles, plateformes de *crowdsourcing*, événements de type hackathon, constitution de corpus de recherche, reconstitutions virtuelles, etc.

Outils de transcription

Transkribus et eScriptorium partagent des workflow similaires. Dans les deux cas, on utilise des réseaux neuronaux pour l'apprentissage profond. Il faut entrainer les logiciels avec des images et des transcriptions correspondantes pour qu'il apprenne à lire une écriture manuscrite particulière.
Ils permettent :
-l'import de documents
-la segmentation des pages de texte
-la création ou fournitures de données d'entrainement (créer ou importer des transcriptions alignées grâce à la segmentation à l'image numérisée) = ce que l'on appelle les « vérités de terrain », ou « ground truth » qui sont à la base de l'apprentissage
-la création à partir des données d'entrainement de modèles de reconnaissance d'écritures manuscrites
-la transcription automatique et d'exportation des données créées
Les outils diffèrent par leur ouverture, leur ergonomie, leurs fonctionnalités et la puissance de calcul des serveurs qui soutiennent les outils.

- La transcription automatique sert à
- Massifier les transcriptions, soit une opération chronophage mais souvent essentielle dans de nombreux projets d'Humanités Numériques
- Condition de possibilité pour d'autres traitements informatiques des textes manuscrits (textométrie/lexicométrie, balisage TEI, fouille de texte etc.)
- Améliorer l'accès aux informations contenues dans les textes manuscrits, une promesse de recherche plein texte, alors même si le taux de reconnaissance n'est pas parfait. Soit une promesse de révolution dans l'indexation et l'accessibilité des textes manuscrits et des informations qu'ils contiennent
- Elle permet possiblement d'enrichir de données textuelles les numérisations exposées dans les bibliothèques numériques

Archivage pérenne

Le CINES propose une solution performante pour la conservation à long terme du patrimoine numérique des établissements

- Conforme aux normes du domaine
- Respectant le protocole standard d'échange
- Agrée par le Service Interministériel des Archives de France

Sont concernées :

- Les données scientifiques : issues d'observations ou de calculs
- Les données patrimoniales : pédagogiques, publications, etc.
- Les données administratives

L'expertise du CINES repose sur :

- Une équipe multidisciplinaire de spécialistes à l'écoute des besoins
- Des collaborations avec les principaux acteurs : BNF, Archives de France, Archives nationales…
- L'expérience d'un centre informatique dans l'exploitation des données numériques

Le CINES

Le C.I.N.E.S. (Centre Informatique National de l'Enseignement Supérieur) est un établissement public national, basé à Montpellier (France) et placé sous la tutelle du Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI). Le C.I.N.E.S. a trois missions stratégiques nationales : le calcul numérique intensif, l'archivage pérenne de données électroniques, l'hébergement de plates-formes informatiques d'envergure nationale.

Le CINES est également impliqué le projet européen EUDAT visant à mettre en place une infrastructure européenne d'échange et de conservation de données.

Merci pour votre attention