Huma-Num IR*

Une infrastructure pour les SHS

Journées AUDACES - 1 juin 2023 (Clermont-Ferrand)

Aurelia Vasile (CNRS) : MSH Clermont-Ferrand

Huma-Num IR*

MSH CLERMONT-FERRAND

cnrs

Plan de la présentation

- Historique
- Les missions d’Huma-Num
- Les consortiums
- L’offre de services
- Adresses et contacts

Historique

2013 : la naissance de la TGIR (Très Grande Infrastructure de Recherche) Huma-Num par la fusion de deux infrastructures :

- **l'infrastructure de recherche Corpus** (IR) → numérisation massive des documents
- le **très grand équipement Adonis** (TGE) → accès unifié aux données en SHS grâce à l'intéropérabilité et à la structuration selon des schémas identifiables et acceptés par la communauté.

2021 : nouvel acronyme labélisé par la Ministère = IR « étoile »

Les missions d'Huma-Num

1. La **principale mission de l'IR*** est de construire une infrastructure numérique de niveau international → nœud français des ERIC (*European Research Infrastructure Consortium*) pour les SHS : DARIAH et CLARIN.
    - ➢ à partir d'un pilotage scientifique qui décide les orientations et les priorités
    - ➢ avec les acteurs des communautés scientifiques:
        - ✓ **Réseau des correspondants dans les MSH** (aiguilleurs vers les services d'Huma-Num)
        - ✓ **Consortiums**
2. Proposer des **services et outils** qui répondent aux principes de la Science Ouverte → favoriser l'ouverture et la qualité des données en SHS

Consortiums

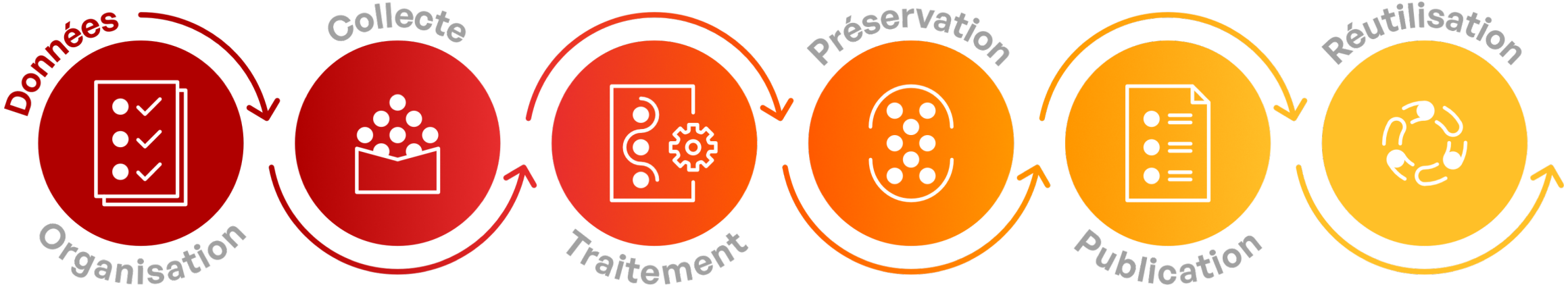
- réunissent plusieurs personnels d'unités et équipes de recherche françaises (chercheurs, ingénieurs, archivistes, documentalistes, etc.) autour de thématiques ou d'objets communs pour lesquels ils définissent des procédures et standards numériques (méthodes, outils, expériences).
- prônent et forment aux bonnes pratiques numériques ses membres, mais aussi les communautés gravitant autour du consortium
- labellisés pour 4 ans : évalués par le CS et financés par Huma-Num

9 Consortiums

- ARIANE (Analyses, Recherches, Intelligence Artificielle et Nouvelles Editions numériques)
- SoL (Sound of Life)
- PTM (Projets Time Machine)
- MASA (Mémoires des archéologues et des sites archéologiques)
- DISTAM (DIgital STudies Africa Asia and the Middle east)
- CANEVAS (Consortium pour l'annotation, l'analyse et l'archive de la vidéo appliquées aux activités scientifiques)
- Consortium Musica 2
- Consortium 3D
- CORLI 2 (Corpus, Langues et Interactions 2)

Des services pour les données en Sciences Humaines et Sociales

https://documentation.huma-num.fr/



Des services pour organiser le travail collaboratif autour de vos données.

– ShareDocs
– GitLab
– Kanboard
– Mattermost

Des services de stockage sécurisé pour la collecte et la création de vos données.

– ShareDocs
– Huma-Num Box

Des services et outils spécifiques pour le traitement et l'analyse de vos données.

– Calcul statistique et environnements R
– Logiciels d'enquête et d'analyse de données
– Reconnaissance de caractères
– Puissance de calcul (+ CC-IN2P3)

Huma-num vous accompagne pour le dépôt et la documentation de vos données dans Nakala, entrepôt pour les données en SHS.

– Nakala
– Huma-Num Box
– Préservation à long terme (+ CINES)

Vos données peuvent être publiées depuis Nakala sur le web et signalées dans Isidore, moteur de recherche pour les SHS.

– Hébergement Web
– Machines Virtuelles
– Nakala
– Isidore

Vos données entreposées dans Nakala et signalées dans Isidore sont réutilisables.

– Portail web
– API
– Triplestore
– OAI-PMH

Huma-Num IR*

Des services pour organiser le travail collaboratif autour de vos données.

– ShareDocs
– GitLab
– Kanboard
– Mattermost

- **ShareDoc**: outil de gestion de fichiers basé sur l'application FileRun
- Ce service permet de stocker, organiser et partager les données de manière sécurisée. (un système de cryptage simple - par mot de passe)
- Il est adapté pour le travail quotidien et la mise à jour régulière des fichiers.
- Accès : soit par navigateur, soit par un client de synchronisation WebDAV
- Volumétrie : 100go / utilisateur et 500go / projet, max: 1to

- **GitLab**: instance sur les serveurs d'Huma-Num de l'outil de gestion de version des fichiers et du code informatiques GitLab et qui met en place les principes *git*
- Utilisé pour déposer du code, réaliser de l'intégration continue et générer des sites web

- **Kanboard** : instance sur les serveurs d'Huma-Num d'un outil de gestion de projets basé sur *Kanban project management software* : permets d'organiser visuellement les tâches et le flux d'activité.

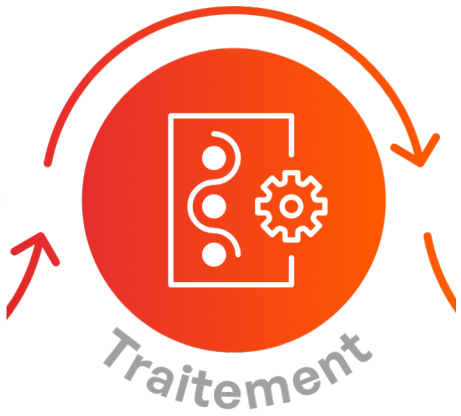
- **Mattermost** : une plateforme de messagerie instantanée

Des services de stockage sécurisé pour la collecte et la création de vos données.

– ShareDocs
– Huma-Num Box

Outils pour stocker les données brutes, hétérogènes et volumineuse

- **Huma-Num Box**
- pour la préservation de volumes importants de données (plusieurs téraoctets)
- service de stockage basé sur un système distribué qui consiste à stocker les données sur des clusters de nœuds dispersés géographiquement.
- Données froides ou tièdes = peu d'accès en écriture et en lecture
- Accès: via un client utilisant le protocole SFTP
- Limitation technique : pas plus de 10 000 fichiers dans un dossier et pas plus de 10 millions de fichiers par projet/structure.



Des services et outils spécifiques pour le traitement et l'analyse de vos données.

– Calcul statistique et environnements R
– Logiciels d'enquête et d'analyse de données
– Reconnaissance de caractères
– Puissance de calcul (+ CC-IN2P3)

Répond aux besoins de transformation et d'analyses de données typiques pour les SHS

1. Outils de traitement intégré dans l'espace ShareDocs → dans le dossier hnTools_watchFolder :
   - **OCR** (*Optical Character Recognition*) : générer du texte brut à partir des documents numérisés (imprimés et manuscrits): 3 services: abbyyCloud, abbyyServer, Tesseract
     - Limites de abbyy : 900 fichiers par personnes / an. (pas de cota pour tesseract)
   - **SpeechToText** : permet de retranscrire automatiquement la parole (fichiers audio et video) en texte brut (bibliothèque Whisper de Python)
   - **Audio / Video** : outils de conversion et trascodage des fichiers audio et vidéo vers certains formats (mp3, mp4, wav, webM)
   - **PDF**: outils de conversion et de conversion à partir des fichiers pdf: conversion vers *ebook, printer, prepress, screen* et transcription vers du texte.

2. Outils de calcul scientifique : en partenariat avec le Centre de calcul de l'IN2P3 (CC-IN2P3)
   - Niveau 1 : la mise à disposition d'un serveur pour du calcul interactif mutualisé entre plusieurs utilisateurs. (integrated development environment IDE R et Jupyther)
   - Niveau 2 : l'accès à la ferme de calcul du CC-IN2P3 par soumission de jobs pour des traitements nécessitants une puissance plus importante.

- Entrepôt de données de recherche pour les Sciences Humaines et Sociales.
- Permet d'enregistrer des données, de les décrire en vue de les exposer et les rendre réutilisables.
- Attribue automatiquement un DOI enregistré auprès de DataCite
- Utilise la norme Dublin Core
- Propose un bac à sable

Huma-num vous accompagne pour le dépôt et la documentation de vos données dans Nakala, entrepôt pour les données en SHS.

– Nakala
– Huma-Num Box
– Préservation à long terme (+ CINES)

Deux niveaux de préservation :

Un **niveau par défaut** qui est mis en pratique dès lors qu'une donnée est enregistrée dans NAKALA. La donnée est décrite, contextualisée et stockée de manière sécurisée.

Un **niveau avancé** qui s'inscrit dans un partenariat avec le **CINES**. Dans ce circuit de dépôt avancé, la préservation à long terme est assurée par le CINES.

Huma-Num IR*

Vos données peuvent être publiées depuis Nakala sur le web et signalées dans Isidore, moteur de recherche pour les SHS.

– Hébergement Web
– Machines Virtuelles
– Nakala
– Isidore

- **Nakala Press** : un module de publication permettant de créer un site web autour de données publiques déposées dans Nakala

- **Hébergement web mutualisé** → toute application Web utilisant les technologies classiques PHP, MySQL, PostgreSQL, Java

- **Hébergement de machines virtuelles** → applications Web et de traitements complexes. Ce service donne de l'autonomie logicielle aux projets.

- **Isidore:** moteur de recherche, agrégateur de données de plusieurs fournisseurs (Hal, Hypothèse, Thèse.fr, OpenEdition, Nakala, Gallica, etc)
- Offre de nombreuses fonctionnalités pour organiser la veille scientifique.
- Rechercher en texte intégral dans plusieurs millions de documents
- Enrichir l'expérience utilisateur, en proposant des contenus similaires, par l'attribution automatique des mots-clés, etc.
- Modèle pour GoTriple

Les outils et les services → orientés pour permettre et faciliter la réutilisation des données hébergées chez Huma-Num.
Les outils informatiques → favorisent la réutilisation des données en s'appuyant sur des protocoles d'échanges.

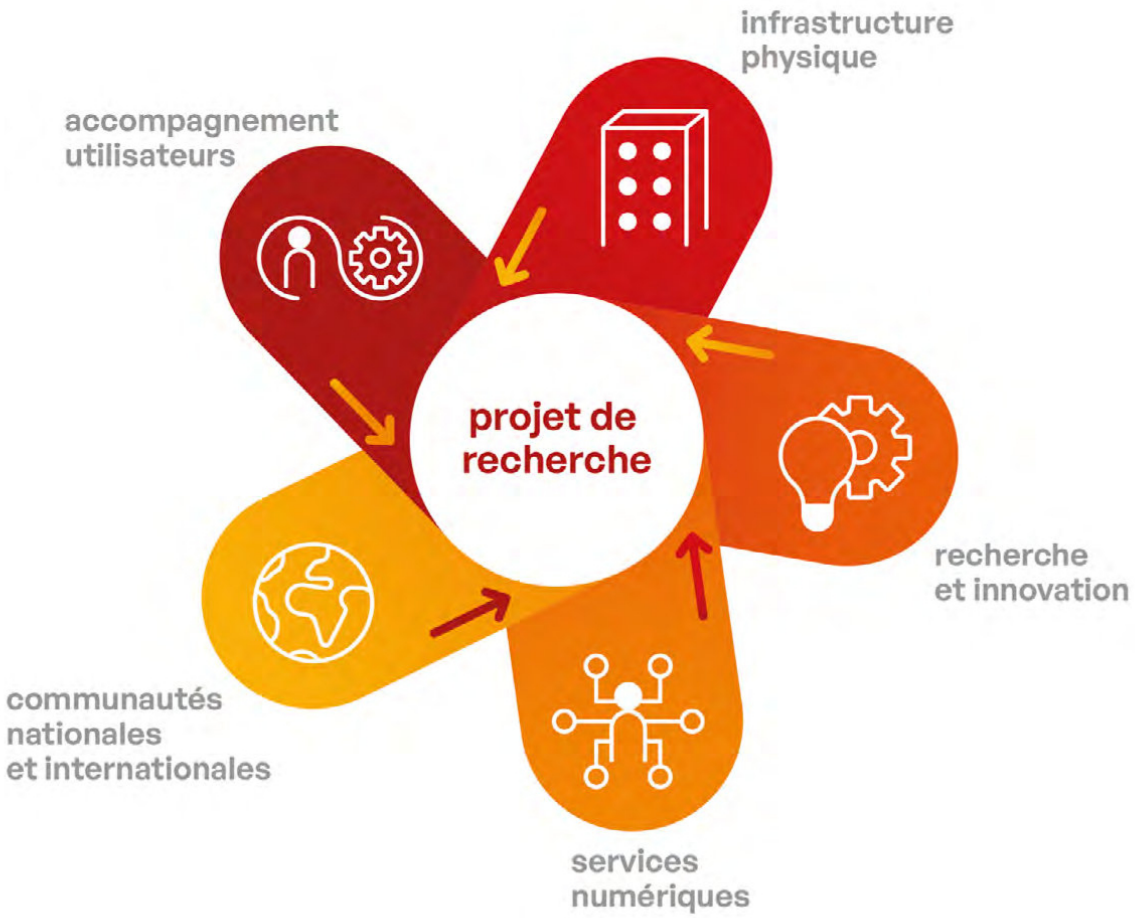
- **API : Nakala**
  - Récupérer et de réutiliser des métadonnées
  - Déposer des données et des métadonnées
  - Faire des mises à jour
  - Assurer la gestion des utilisateurs
- **API d'Isidore**: GET pour des recherches ciblées
- **Serveur IIIF** → préconisation du IIIF (*International Image Interoperability Framework*
  - diffuser, présenter et annoter des images de manière standardisée
- **Protocole Oai-PMH** point d'accès offert par Nakala: diffuser des ensembles (collections)
- **Triplestore d'Isidore et de Nakala** exposition des données en format RDF et langage de requête sparql

Vos données entreposées dans Nakala et signalées dans Isidore sont réutilisables.

– Portail web
– API
– Triplestore
– OAI-PMH

Huma-Num IR*

Des services pour les données en Sciences Humaines et Sociales



Huma-Num IR*

Adresses, contact et liens

Sur le web

- Le site web : https://www.huma-num.fr/
- Le carnet de recherche : https://humanum.hypotheses.org/
- Le site de documentation : https://documentation.huma-num.fr/
- Site de Nakala : https://nakala.fr/
- Site d'Isidore : https://isidore.science/
- Api Nakala: https://api.nakala.fr/doc
- Api Isidore: https://isidore.science/api
- Présentation détaillée par Huma-Num: https://anf2021-humanum.sciencesconf.org/resource/page/id/1

Listes et réseaux sociaux

- Liste de diffusion : humanum-diffusion@listes.huma-num.fr
- Twitter : @Huma_Num

Via les acteurs-relais

- Contacter la MSH la plus proche du porteur de projet
- Annuaire public des correspondant.e.s HN (https://www.huma-num.fr/carte-des-relais-huma-num-dans-les-msh/)

Contact direct

- Par la biais de humanid : https://humanid.huma-num.fr/
- Pour demander l'ouverture d'un service : cogrid@huma-num.fr
- En cas de problème d'utilisation d'un service : assistance@huma-num.fr
- Question spécifique à Nakala : nakala@huma-num.fr

Huma-Num IR*