Huma-Num
la TGIR des humanités numériques

ANF-2021

Gérer ses données en SHS avec les services et outils proposés par la TGIR Huma-Num

Module 1 « Préparer ses données et ses métadonnées pour NAKALA »

16-septembre-2021

cnrs Aix*Marseille université CAMPUS CONDORCET Paris–Aubervilliers

Huma-Num
la TGIR des humanités numériques

Objectifs du module 1

Positionner Nakala dans le cycle de vie des données

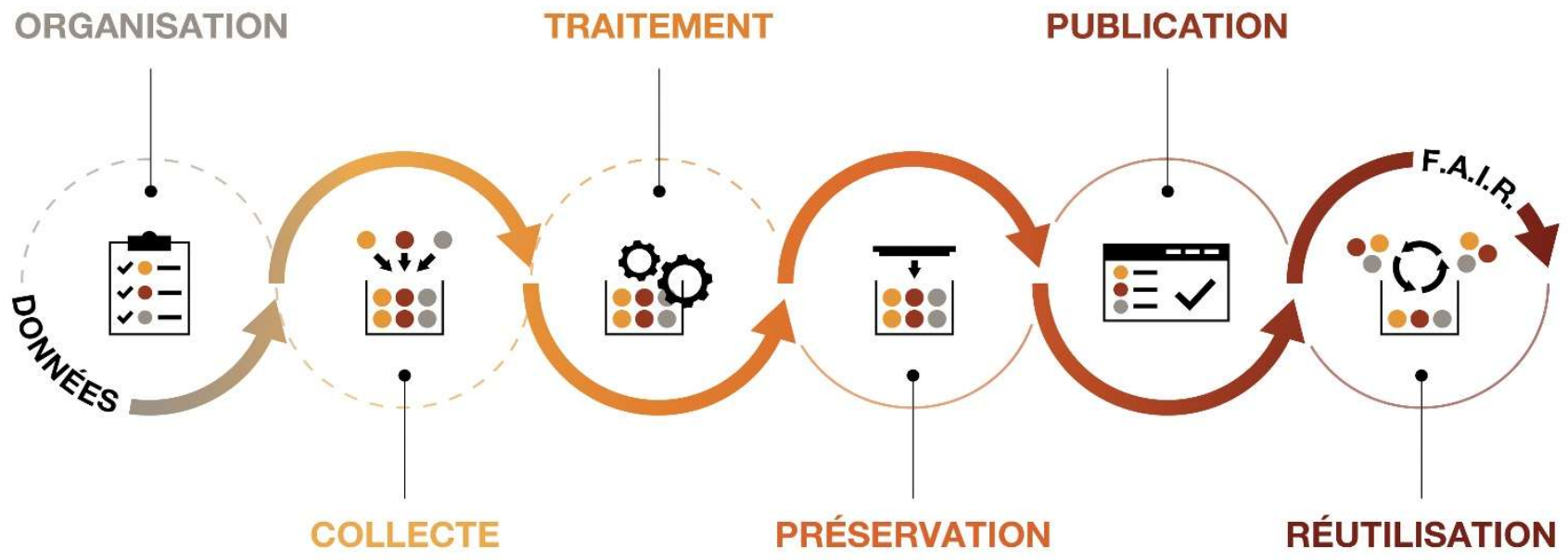
Définir le périmètre de Nakala

Identifier les étapes de préparation des données et métadonnées avant leur versement dans Nakala

Passer en revue des outils d’aide pour les phases de préparation

Huma-Num
la TGIR des humanités numériques

Le cycle de vie des données en SHS



Des données brutes aux données F.A.I.R.

Facile à trouver • Accessible • Interopérable • Réutilisable

Huma-Num

Huma-Num
la TGIR des humanités numériques

Des services pour les données en SHS

ORGANISATION

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

- ShareDocs
- GitLab
- Kanboard
- Mattermost

TRAITEMENT

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

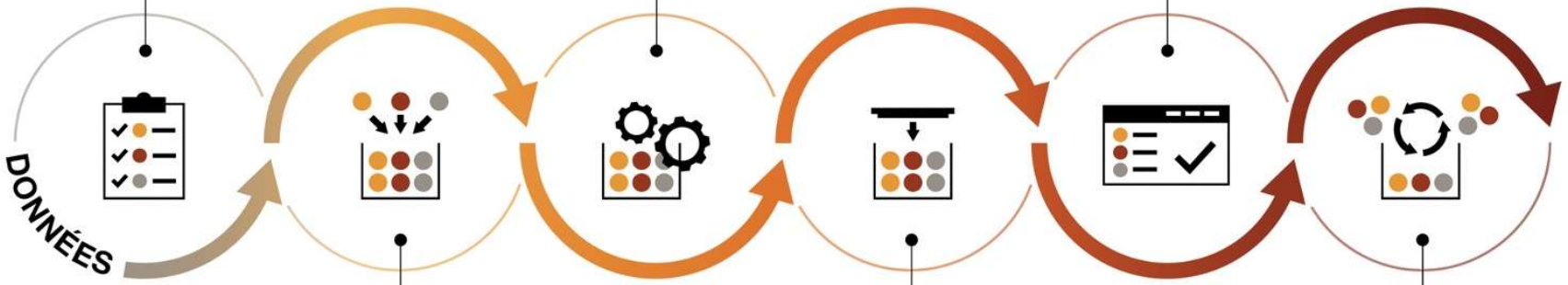
- Calcul statistique et environnements R
- Logiciels d'enquête et d'analyse de données
- Reconnaissance de caractères
- Puissance de calcul (+ CC-IN2P3)

...

PUBLICATION

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



COLLECTE

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

- ShareDocs
- Huma-Num Box

PRÉSERVATION

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

RÉUTILISATION

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

- Portail web
- API
- Triplestore

Huma-Num

Huma-Num
la TGIR des humanités numériques

Positionnement

* Positionnement de Nakala dans le cycle de vie des données

À cheval sur les préoccupations de préservation et de publication

Préservation	Publication
• Dépôt, stockage des données et des métadonnées • Attribution d'identifiants pérennes pour citer et accéder aux données • Gestion des versions	• Accès (OAI-PMH, SPARQL, API REST) • Gestion des droits d'accès : embargo • Visionneuses pour les formats (image, CSV, audio, vidéo, PDF, Markdown, Zip, Code (XML, HTML, JSON…)) • Recherche (dans les métadonnées et dans les données textuelles)

Huma-Num
la TGIR des humanités numériques

Les étapes amont de Nakala

1) L’organisation (gestion de projet)

2) La collecte des données

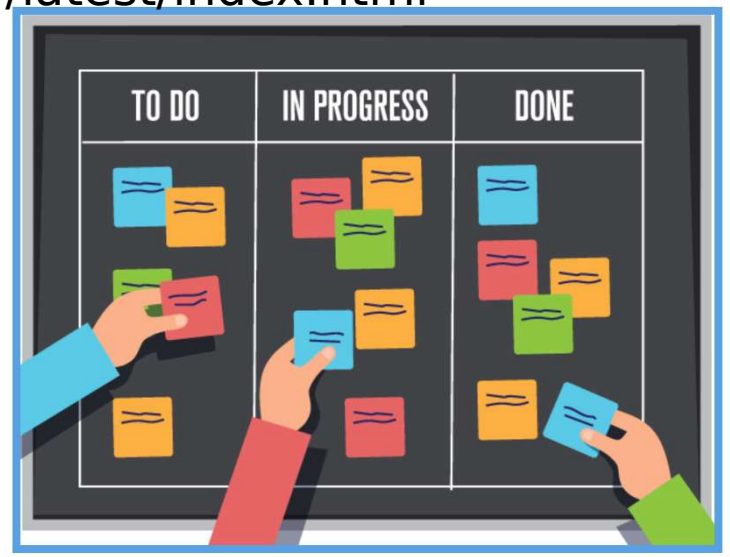
3) Le traitement des données

Huma-Num

1) Organisation
Exemple de l’outil Kanboard

Tutoriel complet :
https://docs.kanboard.org/fr/latest/index.html

La méthode KANBAN



- Méthodologie de gestion de projet basée sur un flux continu flexible (*opposé à la méthode Scrum basé des sprints figés)
- La visualisation est centrale dans cette méthodologie: système de cartes (1 tâche = 1 carte)
- Division du flux de travail dès le début du projet de recherche (chronologique)
- A démarré au début du projet de recherche et à utiliser tout au long de celui-ci

L'outil Kanboard: modalités techniques

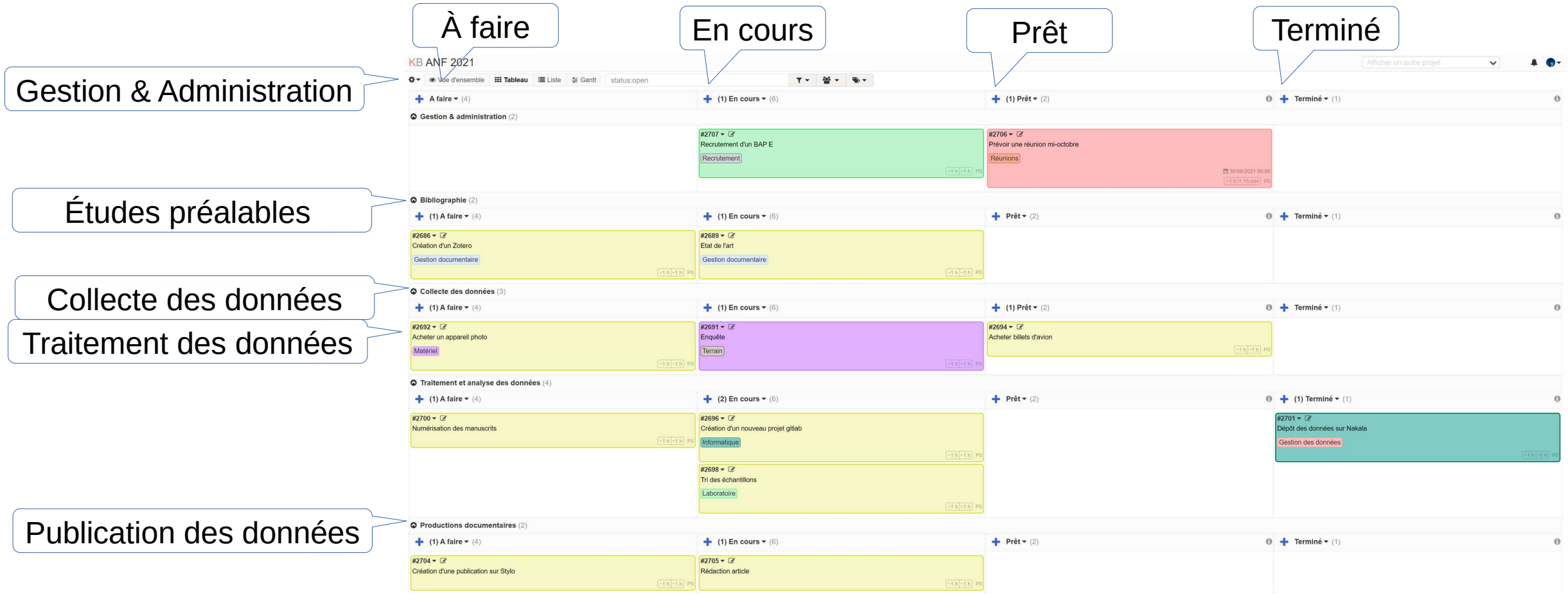
—> Pour tout type de projet

—> Utilisation individuelle ou collaborative (multi-utilisateurs)

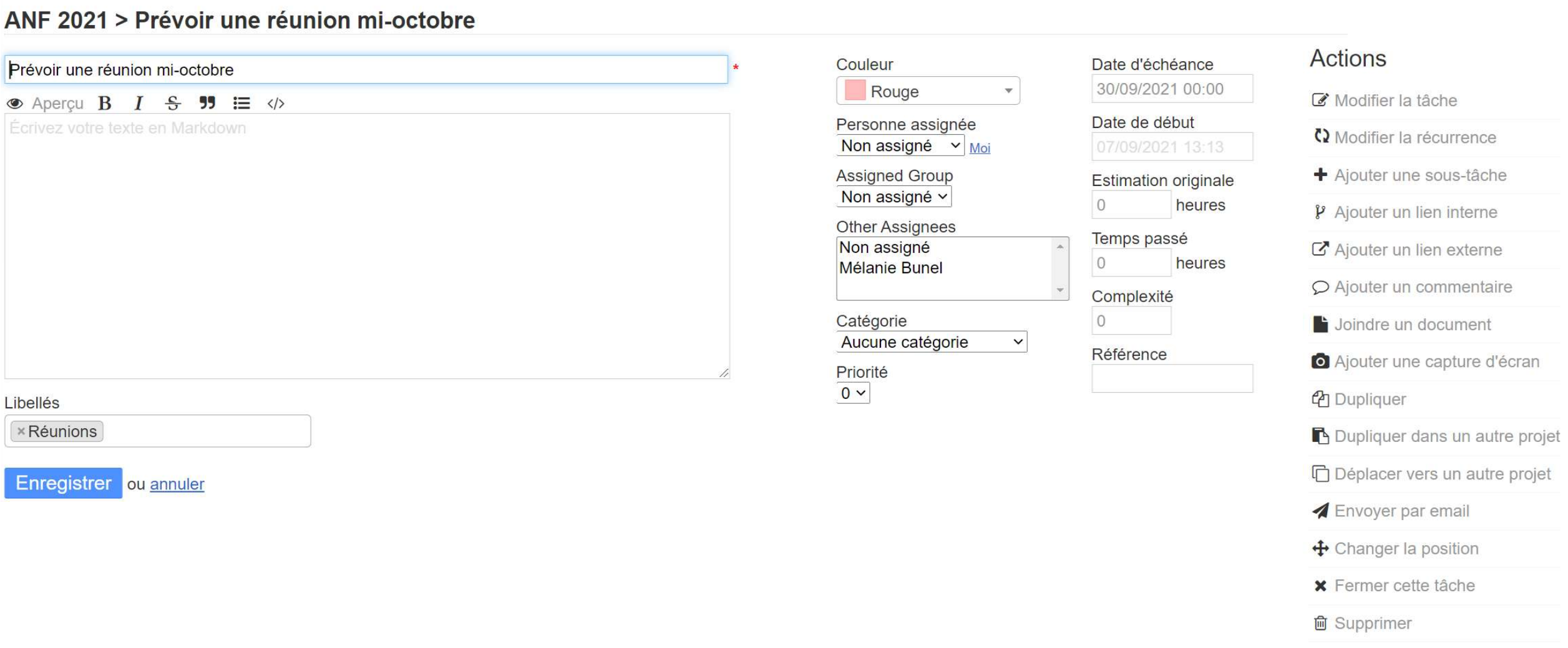
—> Connexion via HumanID sur: https://kanboard.huma-num.fr/

- Possibilité de paramétrage pour une utilisation personnalisée
- Possibilité d'assigner des utilisateurs à des tâches
- Possibilité d'ajout de milestones
- Structuration grâce à des 'swimlanes' (diviser un projet en sous-projet)
- Utilisation des couleurs et d'étiquettes/catégories
- Plusieurs vues dont la plus visuelle est le tableau

L'outil Kanboard: Visualisation “Tableau” du projet ANF

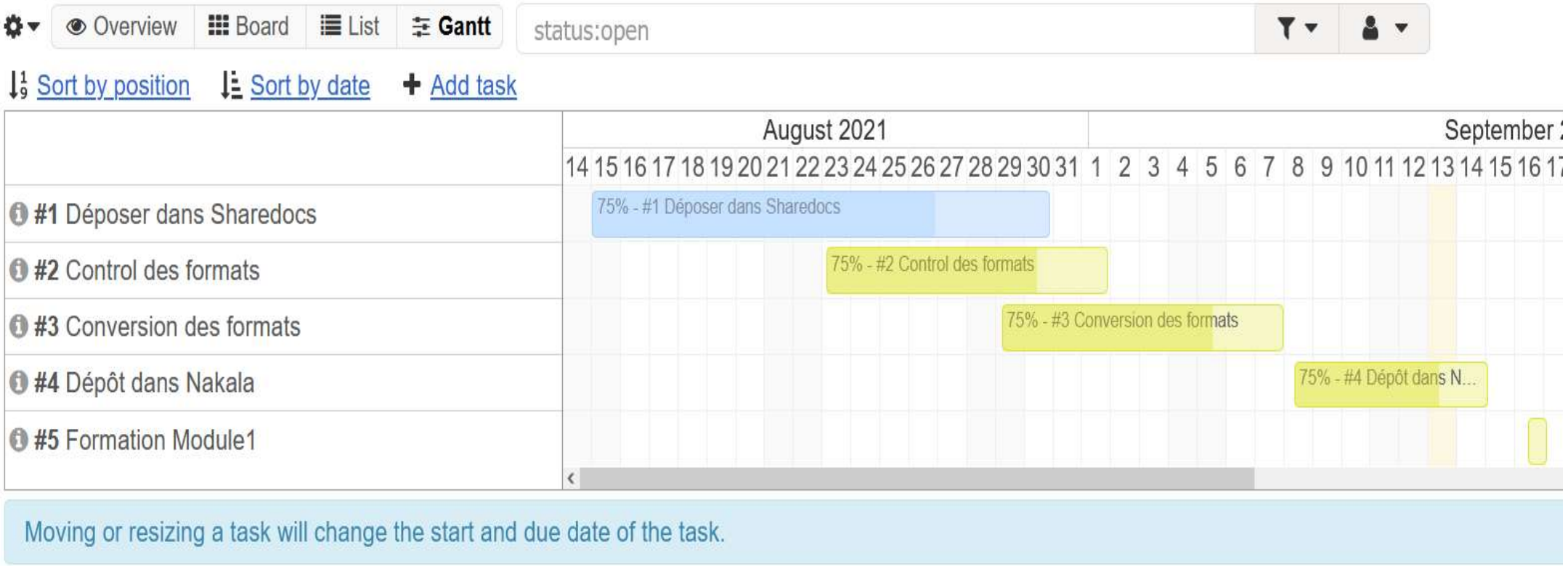


L'outil Kanboard: Création d’une carte/tâche



L'outil Kanboard: Visualisation en diagramme de Gantt

utilise les dates ajoutée lors de la création des tâches



Huma-Num
la TGIR des humanités numériques

1) Organisation

Des outils de communication (listes de diffusion : Sympa, messagerie instantanée : Mattermost)

Huma-Num

Huma-Num
la TGIR des humanités numériques

2) Collecte

Huma-Num

Huma-Num
la TGIR des humanités numériques

La collecte des données avec les services d’Huma-Num, illustration avec le service ShareDocs

Huma-Num

Quelles sont les données dont il est question

- Il s’agit de données de la recherche dans le périmètre des Sciences Humaines et Sociales
- Elles contiennent de la connaissance et forment des ensembles uniques
- Elles sont diverses (formats, typologie) et peuvent être massives

Huma-Num

Dans quel contexte ces données sont produites ou collectées

- Gestion (des données), Fairisation (des données), Science ouverte
- Les agences de financement ont mis en place des règles pour la gestion (DMP) et pour l’ouverture des données (aussi ouvert que possible, aussi fermé que nécessaire)
- Le Plan Science Ouverte 2021-2024

Huma-Num

Dans quel contexte ces données sont produites ou collectées

« L'ANR participe à l'alignement européen et international en faveur de la structuration et de l'ouverture des données de la recherche. Le principe « aussi ouvert que possible, aussi fermé que nécessaire » est au cœur de sa démarche. L'Agence attire l'attention des coordinateurs sur l'importance de considérer la question de la gestion et du partage des données dès le montage du projet. Elle demande l'élaboration d'un Plan de Gestion des Données (PGD) pour les projets financés à partir de 2019, document qui synthétise la description et l'évolution des jeux de données, et prépare le partage, la réutilisation et la pérennisation des données. »

Huma-Num

Comment la collecte se fait

- Des données vont pouvoir être collectées du début du projet et tout au long
- Il pourra s’agir de création de données ou collecte de données existantes
- Les outils d’acquisition et de production sont variés

Huma-Num

Quels sont les principaux besoins pour réaliser la collecte

- La mise en sécurité des données
- La facilité d’accès à ce dispositif de stockage sauvegardé
- La gestion des accès à ces données pour les membres du projet
- Le fait de pouvoir partager en équipe le travail sur les données
- Le fait de pouvoir travailler sur ces données, les mettre à jour, les organiser, les faire évoluer.

Huma-Num

ShareDocs, un outil au service de la collecte de données

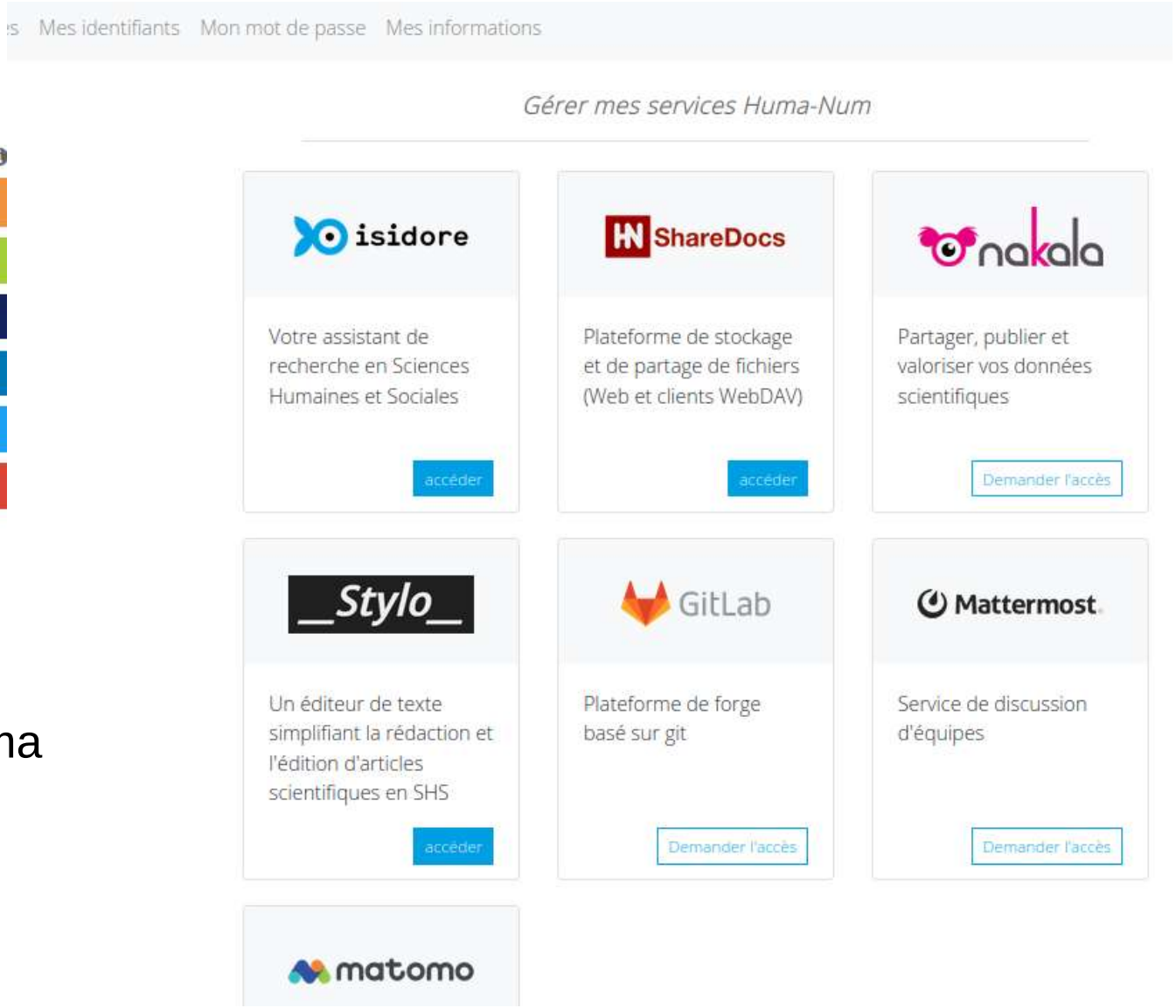
- ShareDocs est un gestionnaire de fichiers en ligne
- Son accès se fait par un navigateur web
- Il permet de stocker des fichiers
- D’organiser les fichiers en dossiers
- De manipuler les fichiers
- De gérer l’accès aux dossiers et fichiers
- De partager de manière restreinte les fichiers/dossiers

Info Nakala

Nakala est destiné au stockage de données stabilisées (décrites et complètes), ShareDocs est fait pour le processus de travail sur les données

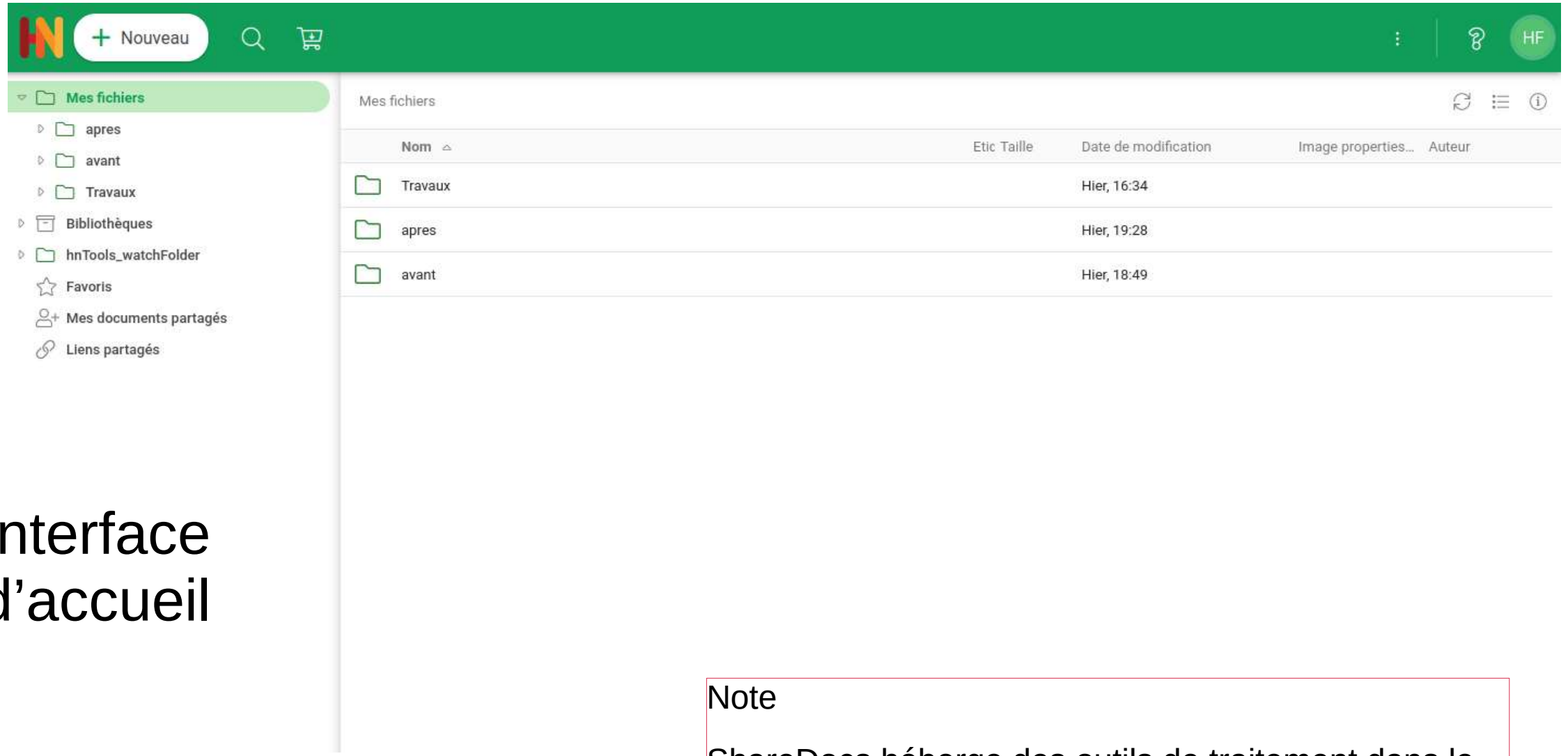
Huma-Num

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



Accès à ShareDocs
par HumanID

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



Interface d’accueil

Note

ShareDocs héberge des outils de traitement dans le dossier hnTools_watchFolder

Quels ordre dans les priorités au cours du processus de collecte

- Avant de débuter la collecte : réaliser l'étude des formats d'acquisition identifier les outils pour transformation en vue de la diffusion
- Départ de la collecte : mettre en sécurité les données
- Ordonner et organiser les données
- Réunir et consigner toutes les informations sur les données en vue de leur description
- Définir et mettre en place le nommage harmonisé des fichiers
- Planifier et organiser le dépôt dans un entrepôt de Données

Et aussi : les traitements sur les fichiers, manipulations, transformations etc.

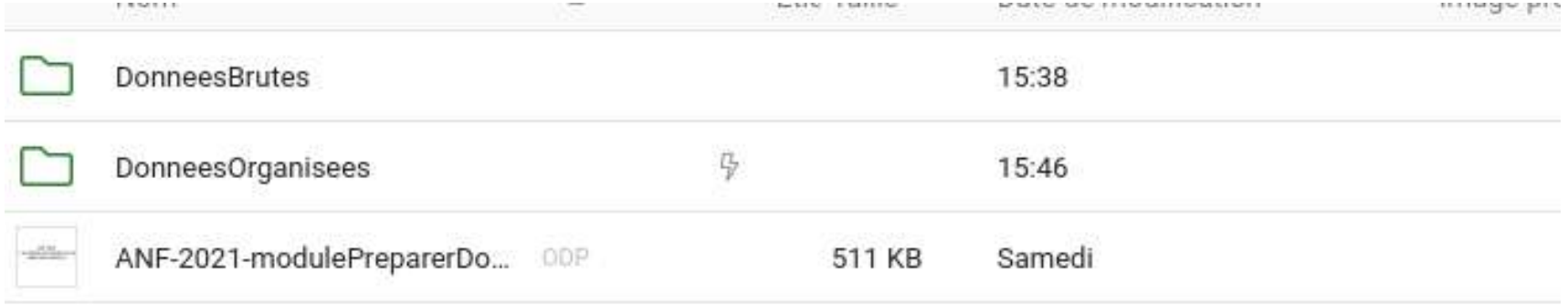
Huma-Num

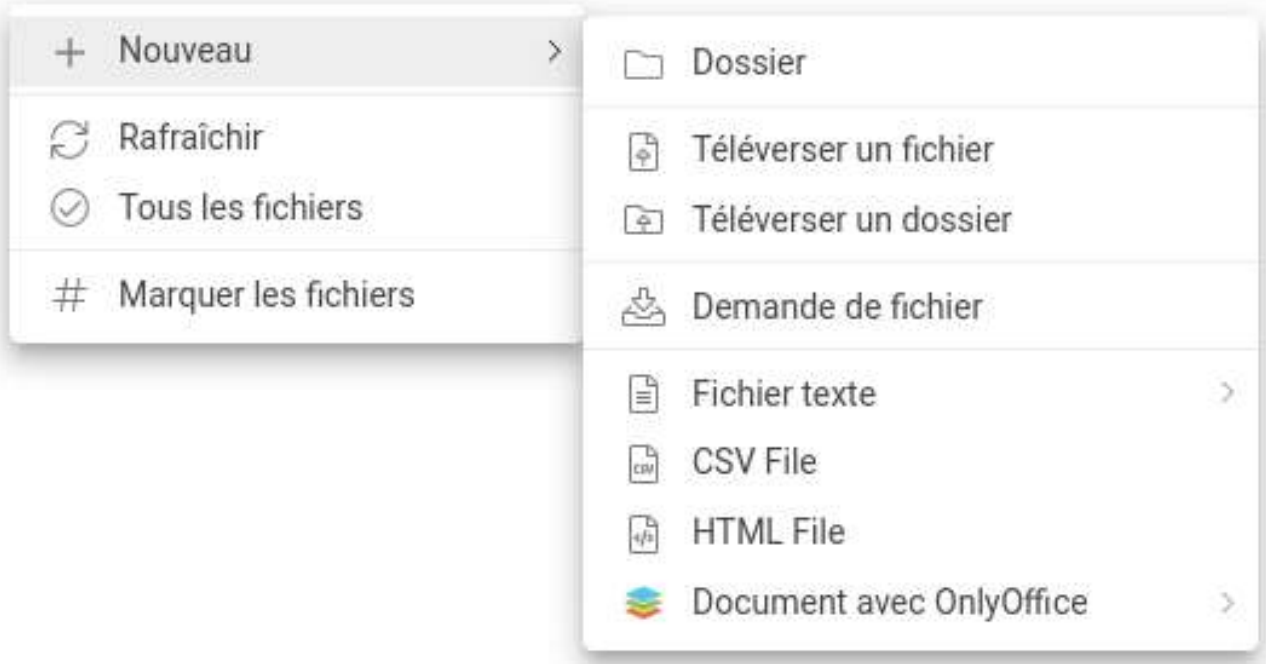
Quels peuvent être les formats des données collectées

- Pas de restriction sur le(s) format(s) des fichiers collectés ou crées dans les outils de collecte mis à disposition par Huma-Num
- Nakala accepte tous les formats de fichiers pour le dépôt et permet la visualisation de certains formats. Distinguer le fait que les données sont entreposées (sauvegardes et accessibles), et le fait que leur visualisation possible ou non dans la landing page (dans le viewer)
- Par contre la question des formats de données entre en jeu dans les principes de Fairisaion, dans les enjeux de réutilisation et de préservation. C’est un paramètre primordial à prendre en compte au moment de la collecte
- Il est possible de prévoir déposer plusieurs fichiers d’une même donnée (de formats ou des qualités différents par exemple)

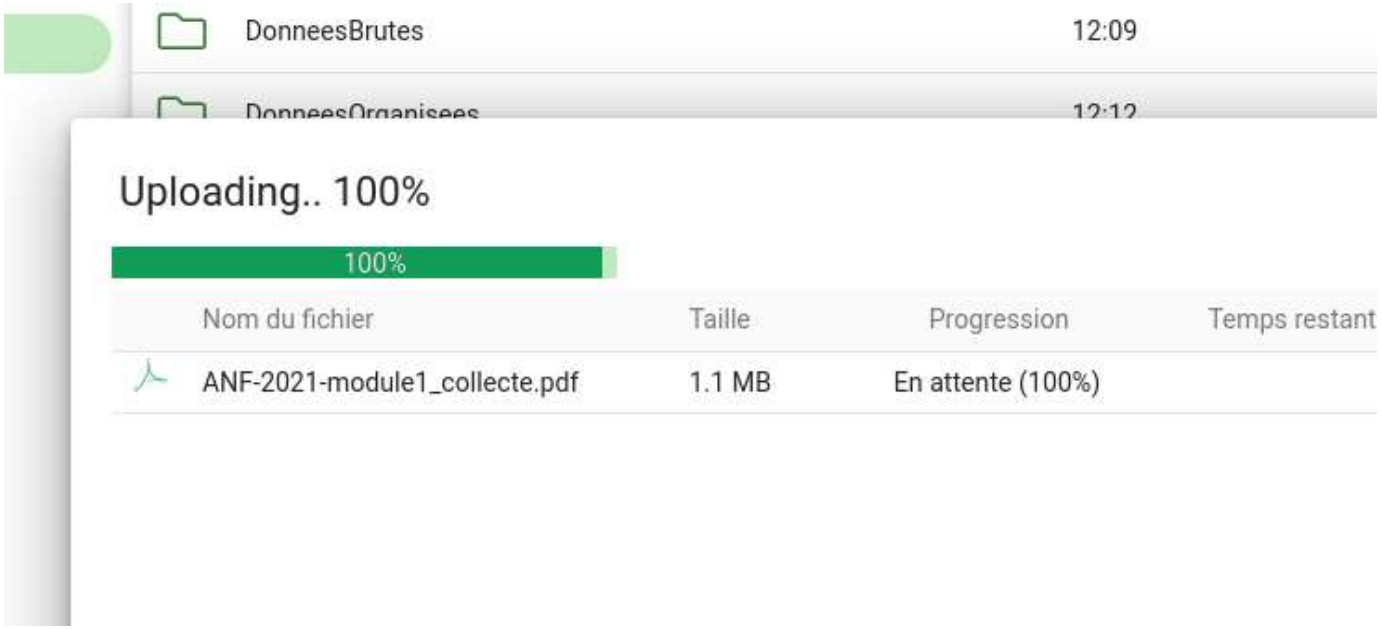
Huma-Num

Chargement de fichiers



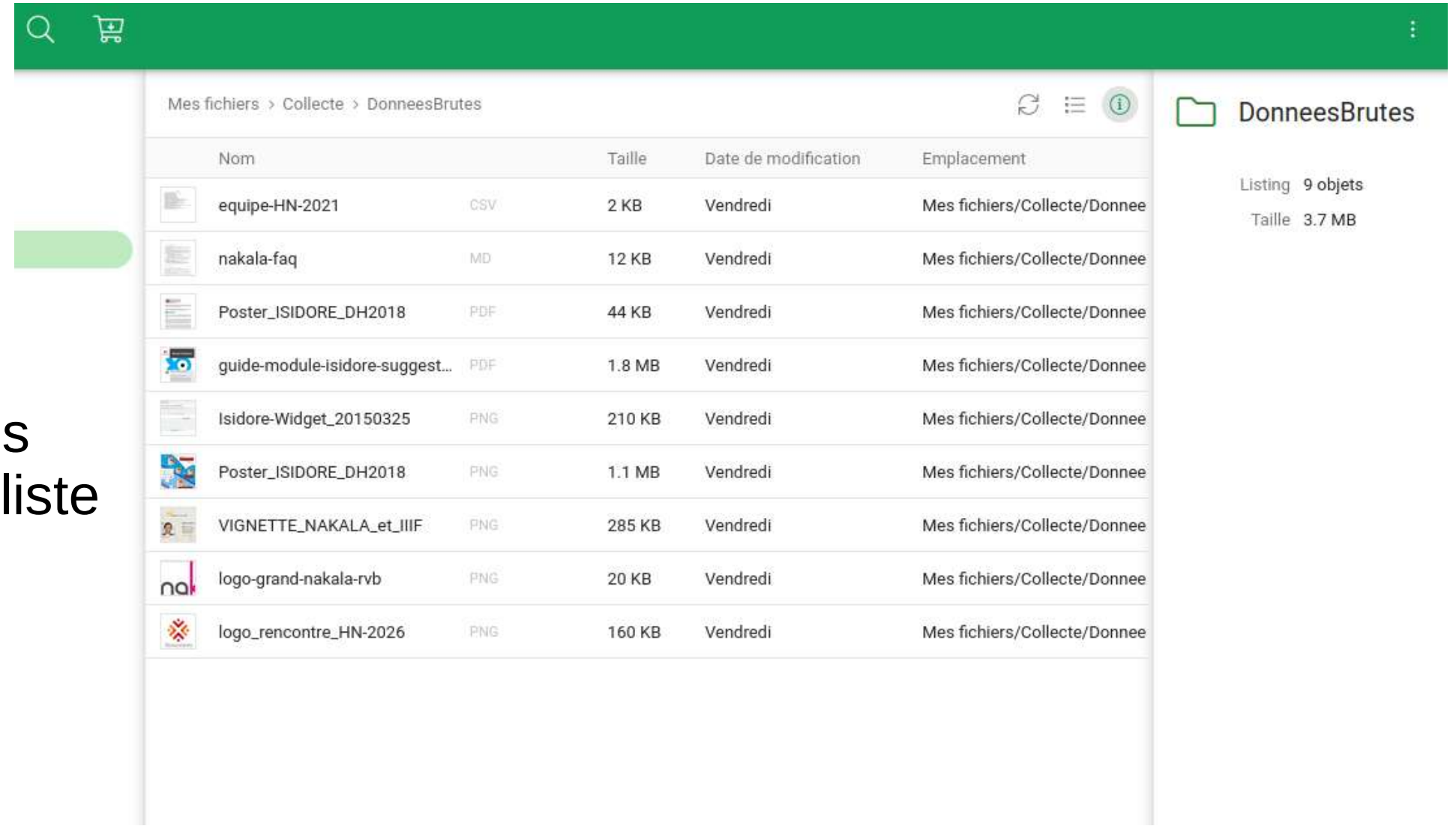


Chargement de fichiers

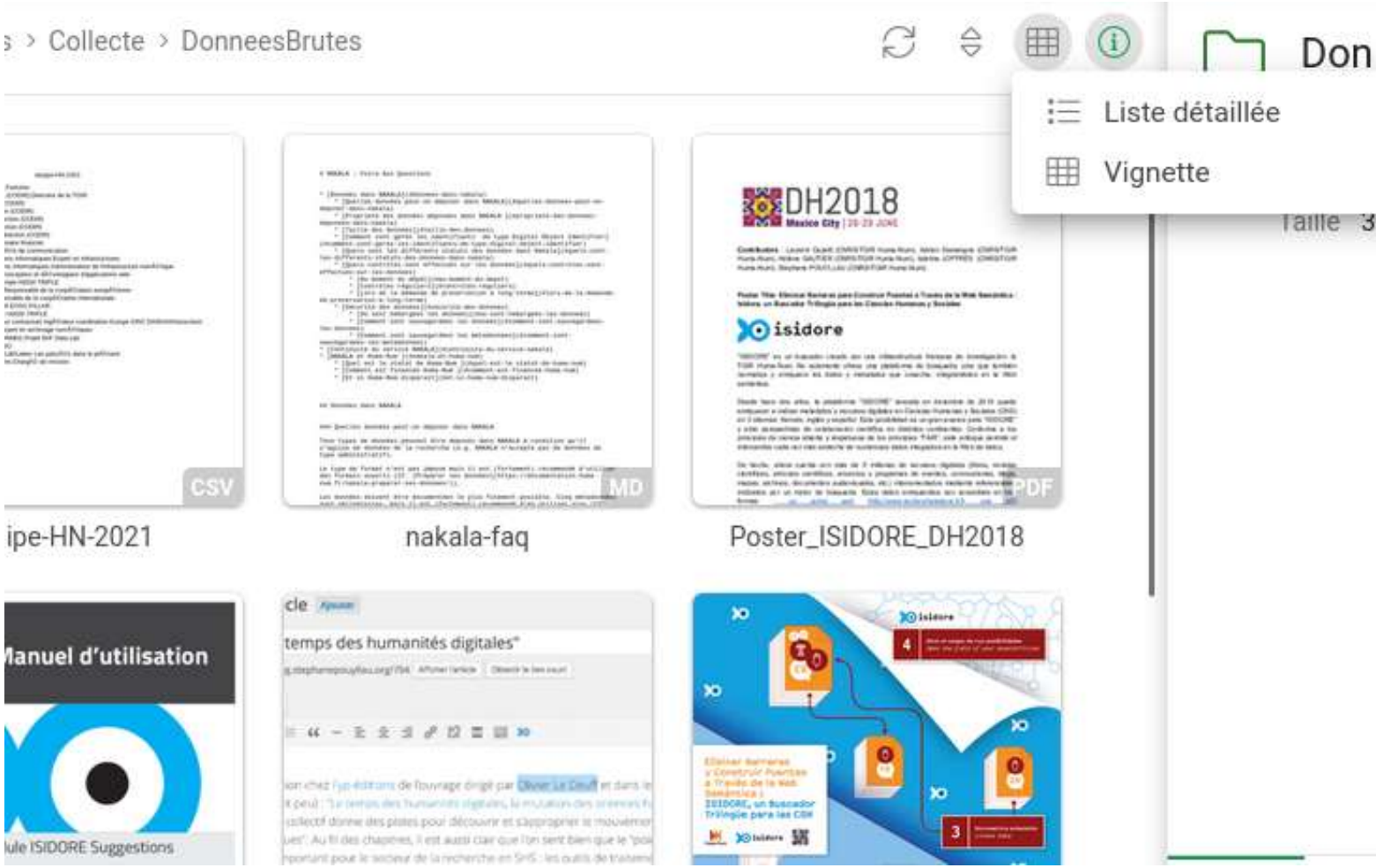


Affichage des données en liste

Colonnes



Affichage des données vignettes



Quelle organisation possible pour les données

- Il est possible de prévoir 2 temps dans l'organisation
- Le temps du travail de recherche (collaboratif généralement) : les fichiers seront alors organisés d'après les besoins du processus de recherche
- Les regroupements pourront être faits en vue : de réaliser des traitements (passer de fichiers bruts à des fichiers traités, de répartir le travail entre collaborateurs etc).

Huma-Num

Quelle organisation possible pour les données

- Le temps du futur dépôt dans un entrepôt
- Penser une répartition des dossiers compréhensible et logique pour des personnes extérieures au projet
- Penser l'organisation des données pour faciliter leur réutilisation, réaliser des regroupements qu'on va nommer : jeu de données
- Les jeux de données ont pu être identifiés dans le DMP ou permettent la mise à jour du DMP

Info Nakala

Nakala permet des regroupements de données en collections

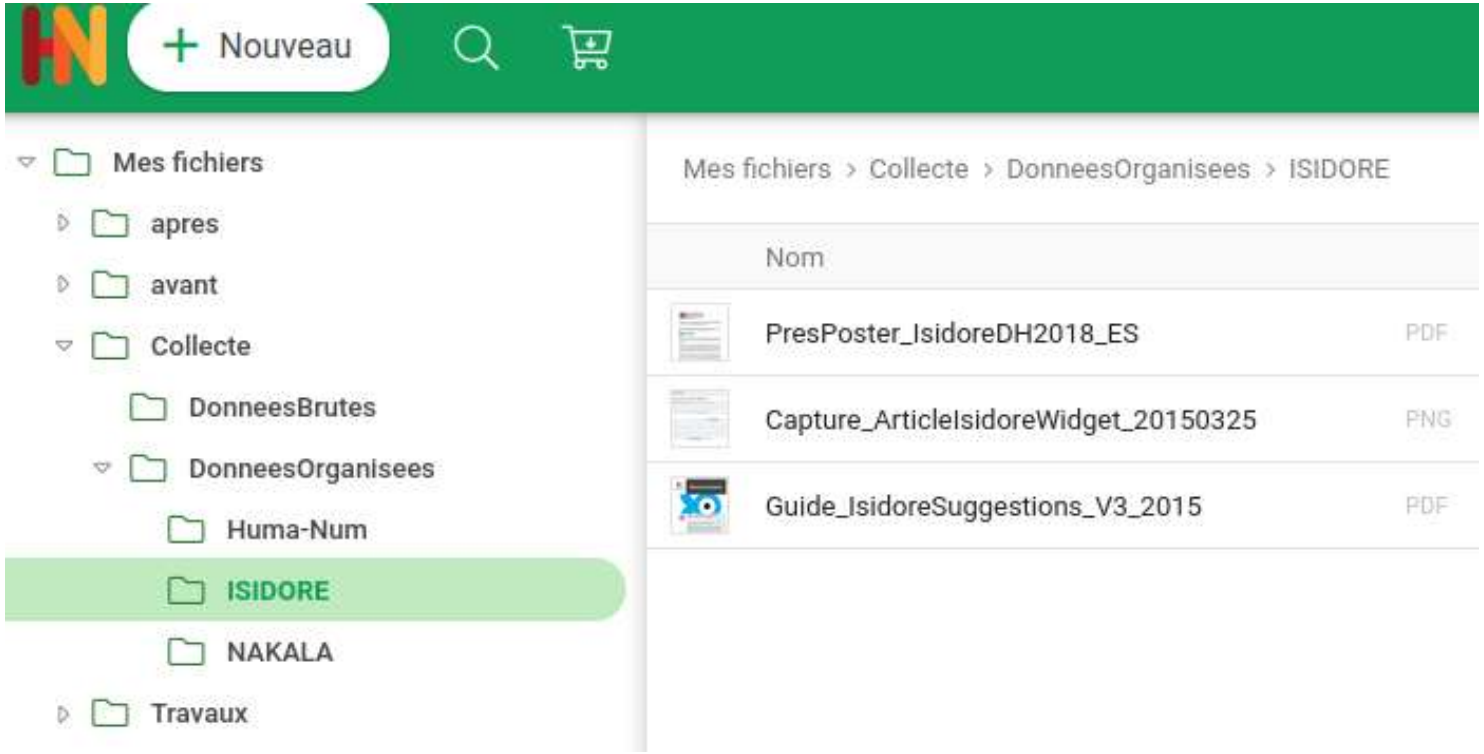
Nakala stocke les collections sans organisation hiérarchique

Des liens intellectuels entre les collections peuvent se faire par les mots clés collection

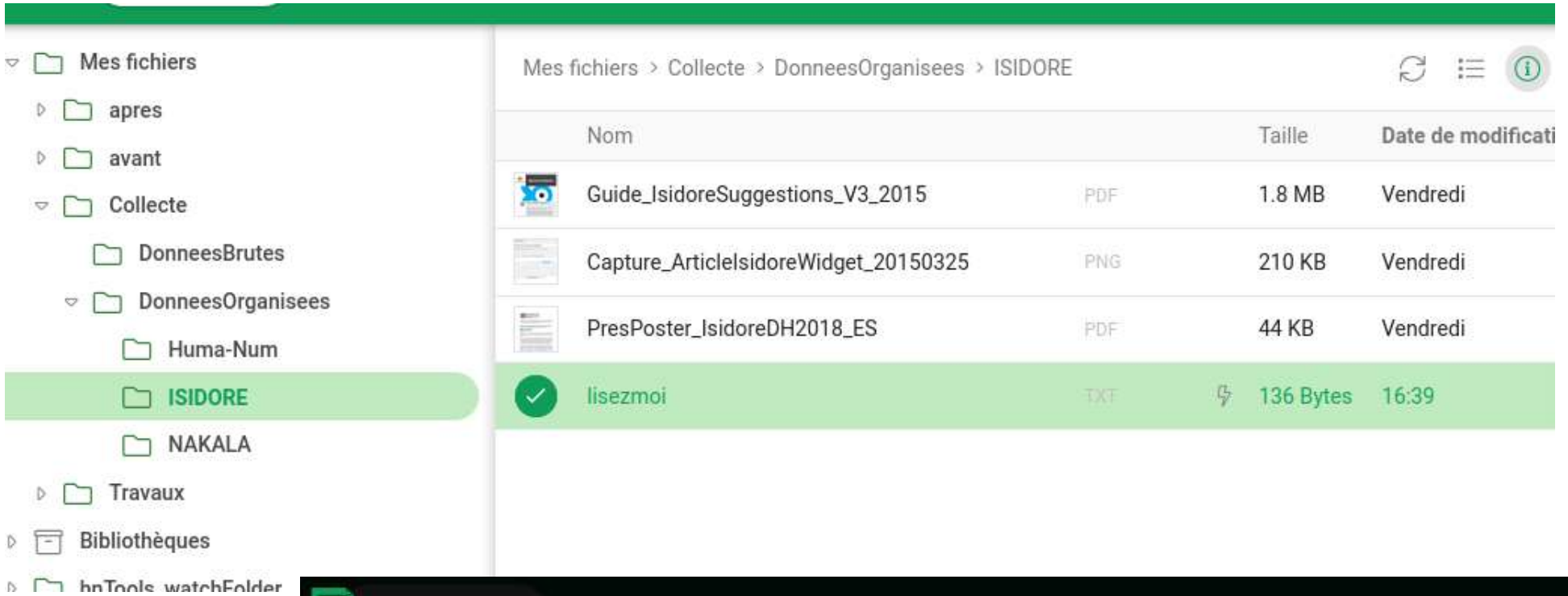
Dans ce contexte il est utile de penser l'organisation en dossiers sans lien de hiérarchie entre eux

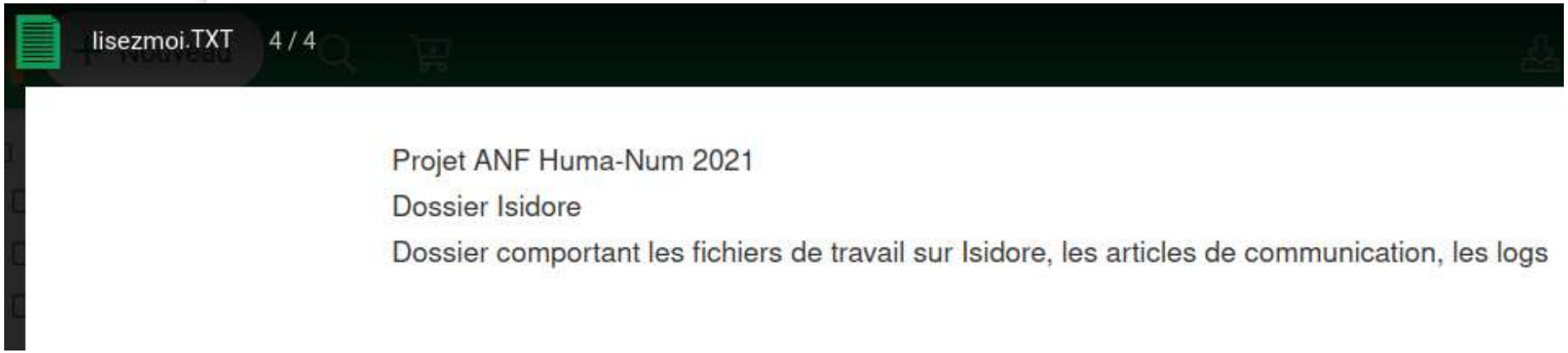
Plan de classement

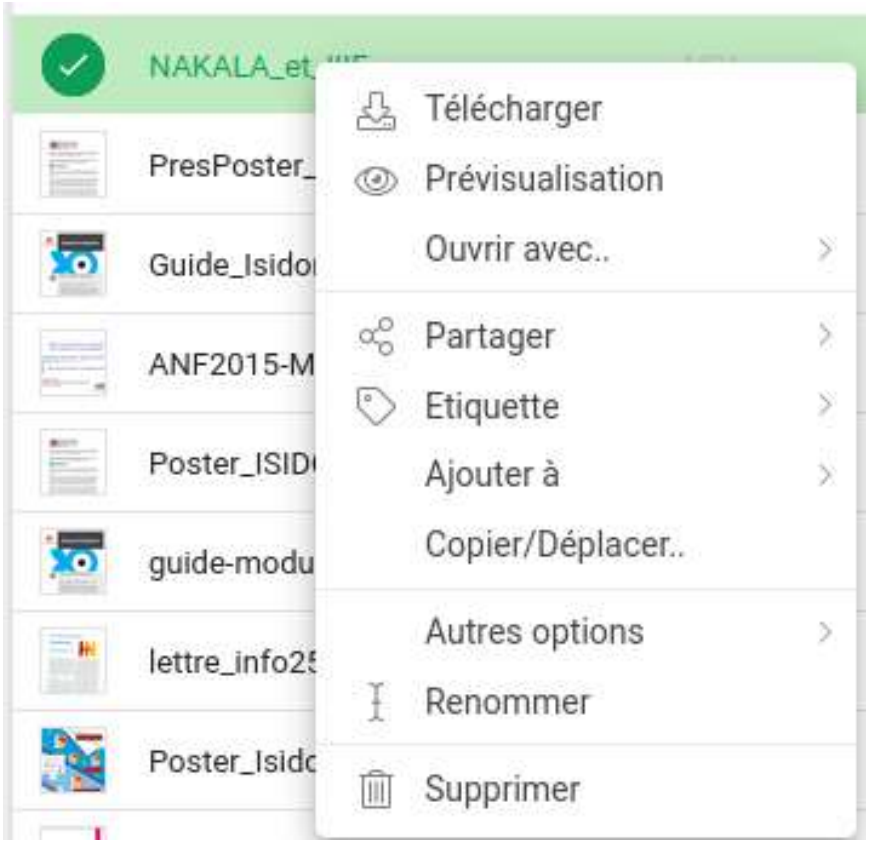
- Arborescence des dossiers
- Affichage des dossiers
- Possibilité de mettre des fichiers lisezmoi pour documenter le classement



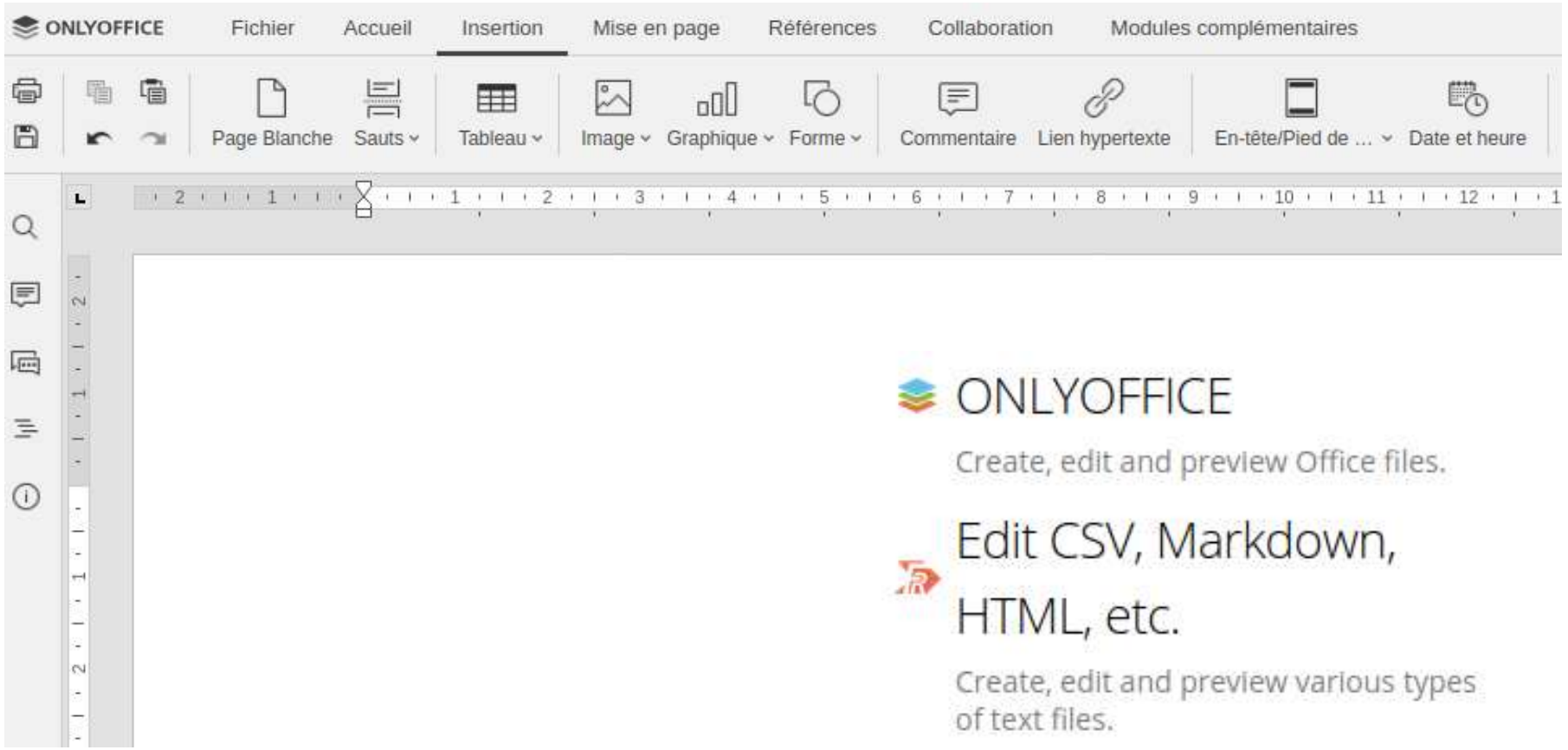
Aperçu du fichier lisez-moi







Aperçu des fonctionnalités de manipulation de fichiers et des dossiers de ShareDocs



ShareDocs dispose d’outils d’édition pour les fichiers Microsoft Office et de type texte (html, md, csv...)

Définir des conventions de nommage des fichiers

- Élaborer des règles communes de nommage et les mettre à jour tout au long du projet
- Avec l'organisation des dossiers, le nommage des fichiers sont 2 opérations importantes pour
  - faciliter l'accès aux données
  - optimiser le partage et le tri des documents
  - Faciliter la compréhension de l'ensemble des données du projet
  - Éviter la perte et la duplication erronée de fichiers

Ressources

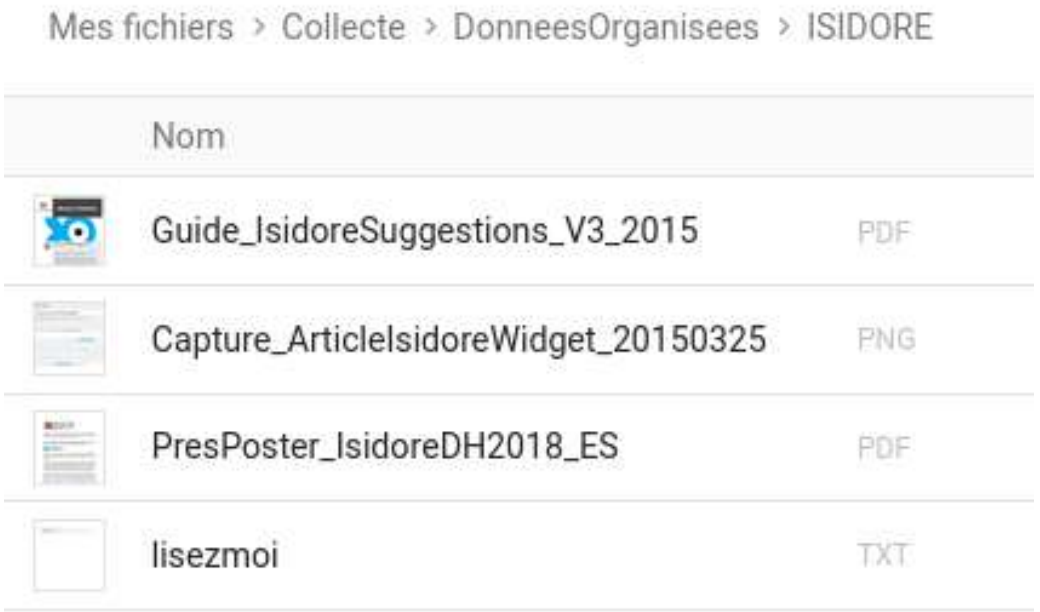
https://doranum.fr/stockage-archivage/comment-nommer-fichiers/

http://qualite-en-recherche.cnrs.fr/IMG/pdf/guide_tracabilite_activites_recherche_gestion_connaissances.pdf

Huma-Num

Convention de nommage (exemple)

- Nature du fichier (pres,liste,logo…)
- Sujet (mots avec majuscules non séparés)
- Date (année, ou année mois jour)
- Informations séparées par des _



Quelles seront les réutilisations possibles des données

- Identifier les données qui contiennent des informations à caractère personnel et/ou sensibles
- Qui entrent dans le périmètre du RGPD (règlement général sur la protection des données)
- Réaliser une déclaration de traitement au DPO/DPD (Délégué à la Protection des données)
- Le cas échéant s'informer sur la pseudonymisation ou anonymisation possible

Les services de stockage de la TGIR sécurisent les données dans leur intégrité « numérique ».

Les fichiers contenant des données dites sensibles devraient, idéalement, être chiffrés avant leur stockage dans ShareDocs

Ressources

https://www.inshs.cnrs.fr/sites/institut_inshs/files/pdf/Guide_rgpd_2021.pdf

https://www.cnil.fr/fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen-data

Huma-Num

Préparer la description des données

- Accumuler les informations sur les données au moment de la collecte : à la fin du projet ces informations seront difficiles à trouver
- Répertorier le maximum d'informations même de type sensibles si cela aide à la compréhension des données.
- Les données pourront être anonymysées ou pseudonymisées plus tard
- Avoir en tête le dépôt : décrire la donnée pour qu'elle soit compréhensible hors contexte.
- Relever toutes les informations possible sur le contexte de production de la donnée.
- Point d'attention sur les dates (de création de la donnée, de création d'une version numérique etc.)

Info Entrepôt

La description des données est une caractéristique commune à tous les entrepôts de données

Les formats de description et champs obligatoires varient d'un entrepôt à l'autre

Huma-Num

Consigner la description des données en vue de leur dépôt

Utiliser un fichier séparé explicite pour :

- faciliter le dépôt (rassemble en 1 endroit les informations)
- faciliter l'enrichissement des métadonnées tout au long du projet

Info Nakala

5 champs minimum

Type, Titre, Auteur, Date, Licence

+ tous les champs du DublinCore

Info Nakala

1 donnée dans Nakala correspond à l'association de :

-> métadonnées (= notice descriptive)

ET

-> 1 ou plusieurs fichier(s)

Identifiants des données :

→ Chaque donnée reçoit un DOI

→ Chaque fichier reçoit une url

Huma-Num

Quelles seront les réutilisations possibles des données

- Connaître le statut juridique des données si elles sont collectées
- Définir leur statut si les données sont produites
- Qui en est propriétaire
- Quels droits de réutilisation il leur sera donné
- Choisir et attribuer une licence

Info Nakala

Les métadonnées sont toujours visibles dès lors qu'une donnée est publiée

Les fichiers peuvent être mis sous embargo (avec durée ou sans limite de temps)

Les données déposées dans Nakala doivent être éligibles au dépôt

L'information de licence de réutilisation est une information obligatoire pour les données déposées

Ressource

https://doranum.fr/aspects-juridiques-ethiques/fiche-synthetique/

Huma-Num

Exemple de fichier consignant les informations sur les données du projet :
Fichier tabulé (de type tableur)

Les métadonnées vues dans ShareDocs sont internes au logiciel, elles ne sont pas exploitables (exportables etc.)

A	B	C	D	E	F	G	H	I	J	K	L	M
file	nakala:title	nakala:creator	nakala:created	nakala:license	nakala:type	dcterms:description	dcterms:language	dcterms:subject	dcterms:creator	dcterms:created	dcterms:publisher	nakala:collection
Guide_IsidoreSuggestions_V3_2015.pdf	Manuel d'utilisation Module ISIDORE Suggestions	Desseigne,Adrien,0000-0002-8272-6125	2015-12-25	Creative Commons Attribution 4.0 International	Cours	ISIDORE Suggestions pour WordPress est un module pour le système de gestion de contenu WordPress. Il permet de proposer des suggestions de lectures, issues des ressources moissonnées et enrichies par ISIDORE, en utilisant soit les mots du texte, soit les mots-clefs d'un billet. Ce guide vous présente en détail les différents paramétrages nécessaires à son fonctionnement.	fr	ISIDORE;Wordpress;IST			TGIR Huma-Num	TGIR Huma-Num;ISIDORE;Widgets ISIDORE
ISIDORE Suggestions_low.mp4	ISIDORE widget - tutoriel vidéo	Pouyllau, Stephane,0000-0002-9619-1002	2015-05-11	Creative Commons Attribution 4.0 International	Cours	Tutoriel vidéo pour ISIDORE Suggestion : installation et paramétrage	fr	ISIDORE;Wordpress;IST			TGIR Huma-Num	TGIR Huma-Num;ISIDORE;Widgets ISIDORE
Capture_ArticleIsidoreWidget_20150325.png	ISIDORE widget - 2.0	Anonyme	Iconnue	Creative Commons Attribution 4.0 International	Image	Capture d'écran illustrant l'utilisation du widget ISIDORE Suggestions.	fr	ISIDORE;Wordpress;IST			TGIR Huma-Num	TGIR Huma-Num;ISIDORE;Widgets ISIDORE
NAKALA_et_IIIF.mp4;VIGNETTE_NAKALA_et_IIIF.png	Nakala et IIIF	Desseigne,Adrien,0000-0002-8272-6125;Capelli,Laurent,0000-0002-1873-3857	2021-03-24	Creative Commons Attribution 4.0 International	Présentation	Présentation du service Nakala et de son API IIIF Image à l'occasion du Rendez-vous IIIF360 organisé le 24 mars 2021 par Biblissima et la TGIR Huma-Num.	fr	Biblissima,IIIF,Campus Condorcet,OpenSeadragon,IIIF,NAKALA			TGIR Huma-Num	TGIR Huma-Num;NAKALA
logo-grand-nakala-rvb.png	Logo NAKALA - jpg	Larrousse,Nicolas,0000-0002-4968-797X	2015	Creative Commons Attribution 4.0 International	Image	Logo de NAKALA au format jpg		NAKALA			TGIR Huma-Num	TGIR Huma-Num;NAKALA
nakala-faq.md	NAKALA : Foire Aux Questions	Anonyme	2021-04-21	Creative Commons Attribution 4.0 International	texte	Foire Aux Questions sur le service NAKALA, un entrepôt de données pour les SHS mis en place par la TGIR Huma-Num	fr	NAKALA	TGIR Huma-Num		TGIR Huma-Num	TGIR Huma-Num;NAKALA
				Creative								

Le fichier de description des données

Les colonnes

- Champs Nakala :
(File),
nakala:title,
nakala:creator,
nakala:created,
nakala:license,
nakala:type,

- Champs supplémentaires (exemples) :
dcterms:description,
dcterms:language,
dcterms:subject,
dcterms:creator,
dcterms:created,
dcterms:publisher,

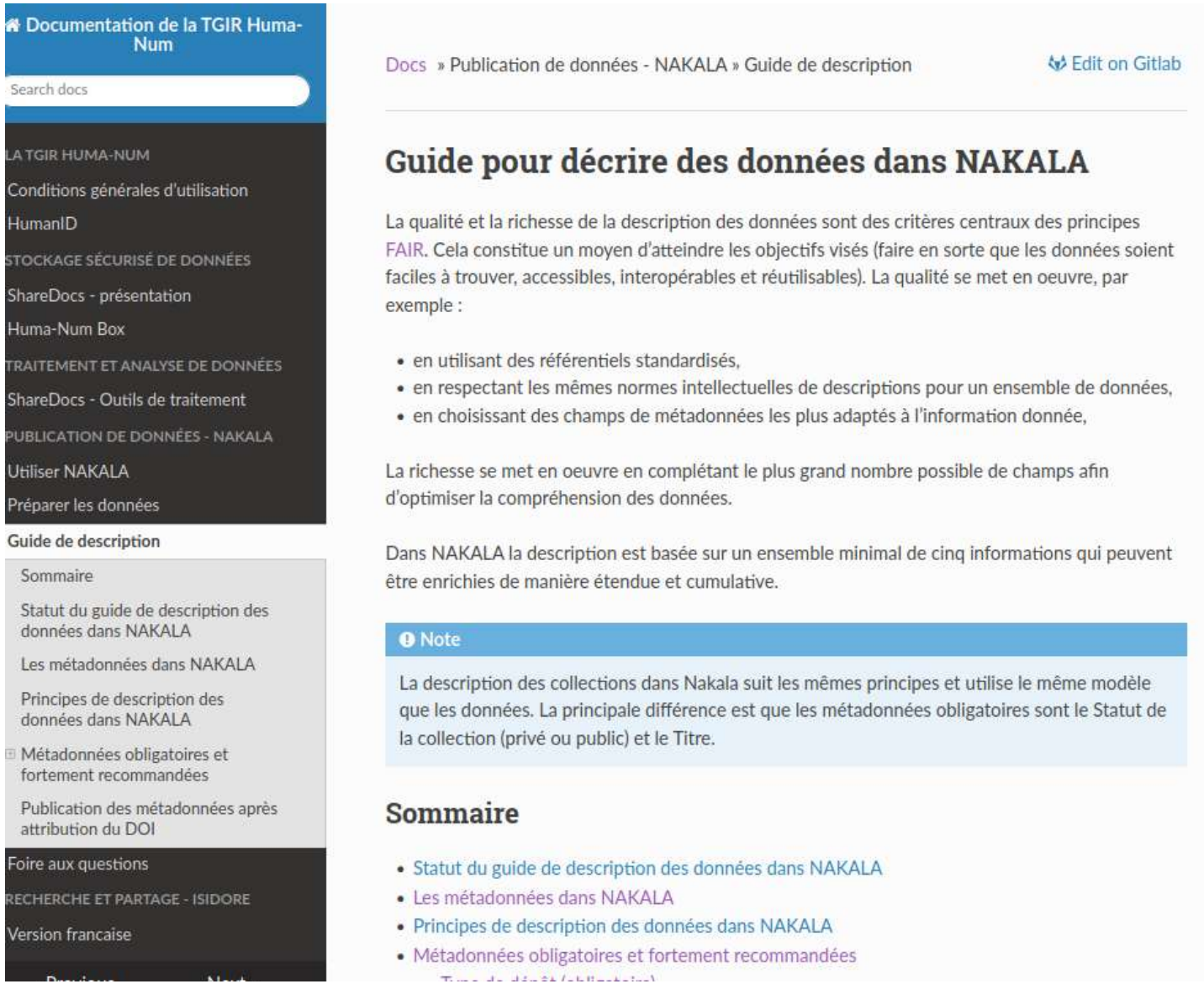
https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

- Indication collection : nakala:collection,

Guide pour décrire des données dans NAKALA :

https://documentation.huma-num.fr/nakala-guide-de-description/

Version actuelle :
Ensemble de conseils et bonnes pratiques pour les champs de métadonnées obligatoires et complémentaires de premier niveau.



Le choix des formats de métadonnées

- Un autre standard que Dublin Core ?

Info Nakala

Les champs « Nakala » issus du dublin core sont obligatoires

Il est possible de déposer avec la donnée, un fichier de métadonnée différent

Ressources :

- https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
- https://doranum.fr/metadonnees-standards-formats/standard-metadonnees/
- https://www.dcc.ac.uk/guidance/standards/metadata

Huma-Num

Préparer la description des collections

- Les collections sont des objets à décrire dans Nakala
- Il est utile de préparer leur description pour une meilleure harmonie lors du dépôt

Info Nakala

4 champs

Titre, créateur, Description

(+ type = collection)

ShareDocs, un outil au service de la collecte de données

- Distinguer les services rendus par ShareDocs de ceux d'un entrepôt de données
- Comme Nakala, ShareDocs assure une sauvegarde sécurisée des données
- ShareDocs permet de charger rapidement un ensemble de fichiers (glisser/déposer)
- Il n'est pas adapté à la description de données

Des besoins spécifiques

- Les données de santé qui relèvent d'un hébergeur de données de santé (HDS)
- Nécessite une infrastructure certifiée HDS
- https://esante.gouv.fr/labels-certifications/hds/liste-des-herbergeurs-certifies

Huma-Num

L’assurance d’un stockage sécurisé des données dans ShareDocs

- Infrastructure de haute fiabilité :
- Serveurs hébergés par le Centre de Calcul de l’IN2P3 (CC-IN2P3), Unité d’Appui et de Recherche du CNRS. https://cc.in2p3.fr/nos-services/hebergement
- Stockage répliqué multi sites
- Utilisation du réseau académique Renater
- Exploitation répartie sur un ensemble d’ingénieur(e)s qualifiés (BAP E)

Gestion des sauvegardes dans ShareDocs :

- Mode incrémental (logiciel TSM (Tivoli) d'IBM)
- de tous les fichiers modifiés par rapport à la veille
- avec conservation de 6 versions du même fichier
- avec rétention durant 1 an d'un fichier totalement supprimé à la source
- sur deux robotiques de bandes LTO dans les deux bâtiments distincts du CC-IN2P3

Huma-Num

Autres spécificités de ShareDocs

- Sharedocs est une application web, dont l’accès est protégé par un Firewall régulièrement mis à jour
- Sharedocs est accessible via le réseau public Internet
- L’utilisateur se connecte à l’application par un login/pass (sécurisé SSL)
- L’édition de fichiers par le plugin OnlyOffice s’opère à l’extérieur des serveurs d’Huma-Num.

Les comptes ShareDocs

- Tout accès à ShareDocs passe par un compte HumanID https://humanid.huma-num.fr
- Validation de la demande et ouverture du compte par le Comité de la Grille d'Huma-num
- Volumétrie : sur la base des prévisions de stockage pour 2021-2023 la volumétrie maximale actuelle sur ShareDocs par compte est limitée à 90G par défaut.
  Pour les projets, la volumétrie plus importante est à adapter en lien avec l'équipe Huma-Num
- (Il est recommandé pour limiter l'usage du stockage de gérer au plus proche les fichiers, notamment en compressant les fichiers volumineux sur les formats qui peuvent l'être (comme les tif, mjpeg, ply, json), de ne pas utiliser cette espace pour des backups de serveurs, d'applications, des fichiers de logs, des fichiers temporaires)

Huma-Num

Utiliser ShareDocs en équipe

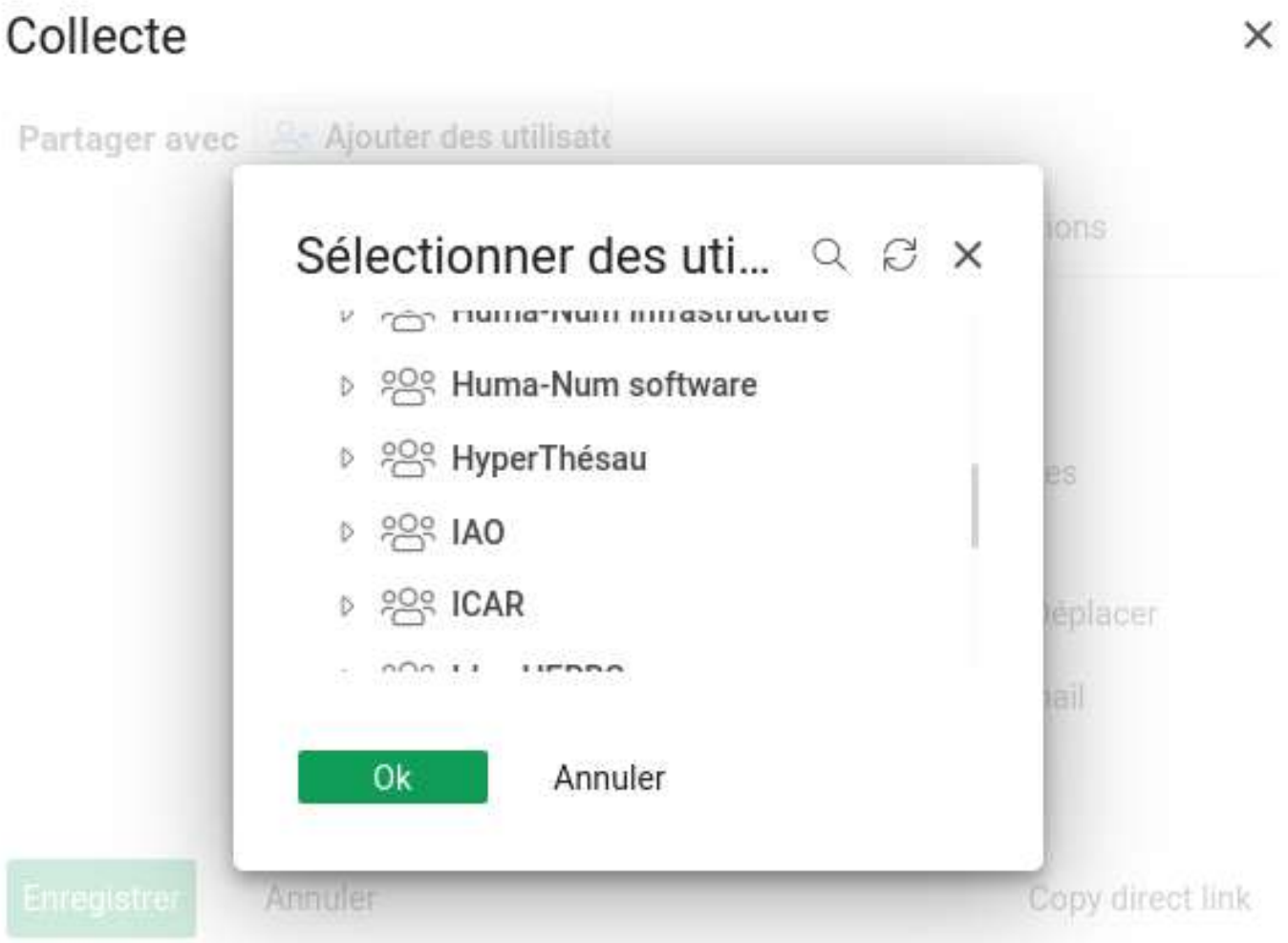
Création d'un compte HumanID et demande d'accès à Sharedocs pour chaque membre en indiquant le projet.

a) Vous partagez votre dossier en équipe : depuis votre compte vous créez un répertoire du projet et vous le mettez directement en partage avec les personnes ce qui vous permet de gérer directement les droits

B) Vous voulez séparer les données du projet de votre compte : Huma-Num déclare un partage au niveau du serveur, et vous nous indiquez la liste des personnes qui ont le droit d'y accéder

C) Dernière possibilité plus lourde en gestion : Huma-Num crée un compte spécifique pour le projet, qui va servir à définir des partages et des droits spécifiques sur chacun.
C'est plus lourd en gestion, mais cela permet de subdiviser les droits.
Cela permet de gérer très finement les différents droits selon des groupes (coordinateur, équipe scientifique, externe, workspace...) et de classer les documents suivant les droits de chaque groupe.

Huma-Num



Le partage en équipe :

sélection des utilisateurs à autoriser sur le dossier

Utiliser ShareDocs

- https://documentation.huma-num.fr/sharedocs-stockage/
- https://filerun.com/

Synchroniser une répertoire ShareDocs avec son ordinateur :

ShareDocs peut également s'utiliser en synchronisant un répertoire de son ordinateur avec un répertoire ShareDocs stocké chez Huma-Num :

avec un logiciel de synchronisation de fichiers comme NextCloud ou encore avec tout autre logiciel client WebDAV permettant la synchronisation de fichiers

https://documentation.huma-num.fr/sharedocs-stockage/#synchronisation-de-stockage-par-webdav

Info Nakala

Il n'y a pas de passerelle directe entre ShareDocs et Nakala

Utiliser ShareDocs pour citer des fichiers

Vigilance : les liens ne sont pas pérennes, ne l'utiliser que ponctuellement pour un public limité

Huma-Num

Focus Gitlab pour la collecte et le travail collaboratif

Merci à nos collègues du HN LAB pour cette explication

GITLAB – qu’est ce que c’est ?

GIT, une “forge logicielle” + Lab, une plateforme collaborative

- Un protocole d’échange et de synchronisation de fichiers
- Un système de suivi de versions
- Un registre (principe de la forge)

→ produire du code, du texte, des données selon un protocole rigoureux

- Fonctionnalités communautaires
- Gestion de groupes, de membres
- Documentation (wiki, readme)
- Gestion de projets (tickets, jalons, tableau de bord)

→ travailler en collectif, ouvrir et partager ses travaux

Fonctionnalités: *issues*, *board*, *milestones*

Principe du ticket (issue) pour la gestion de tâches :

- description, caractérisation, organisation et attribution des tâches
- discussion et suivi de la production

→ gestion de projet sans modalités d’usage prédéfinis. L’équipe peut établir sa propre méthodologie.

Fonctionnalités: *wiki*, *snippets*, *readme*

Des espaces d’écriture destinés :

- à la documentation du projet
- la circulation de fragments de code ou d’éléments d’information

→ Favorise les approches collaboratives

Fonctionnalités: *fork*, *branch*, *pull request*, *merge*

Actions spécifiques au protocole Git (inscriptions au registre du répertoire)

- Gestion collaborative des fichiers et des modifications

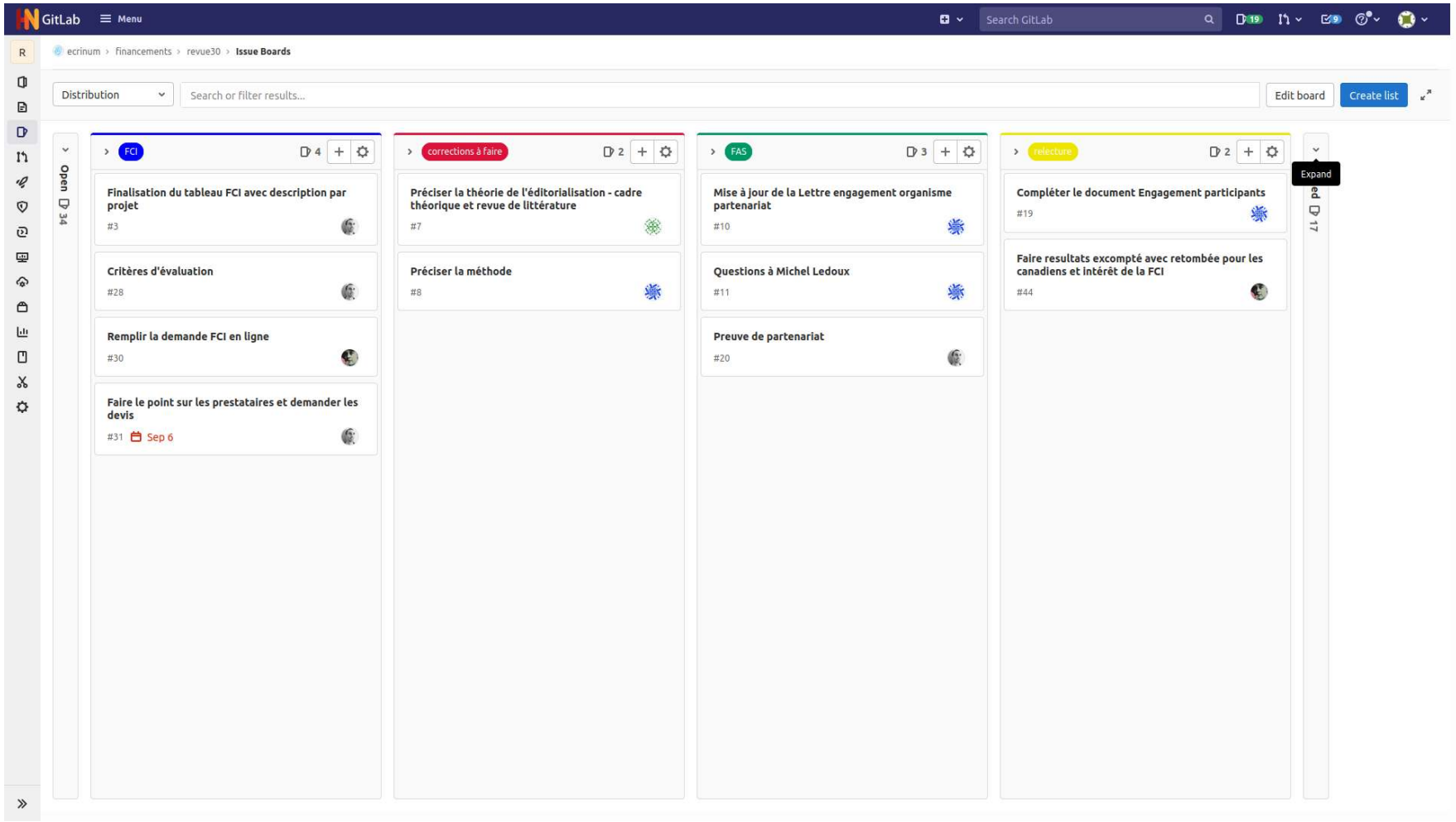
→ Permet l’appropriation du travail par d’autres équipes, la soumission et l’évaluation par les pairs d’une contribution, la distribution des tâches d’évaluation, de correction et de validation.

Fonctionnalités: *history*, *blame*, *graph*

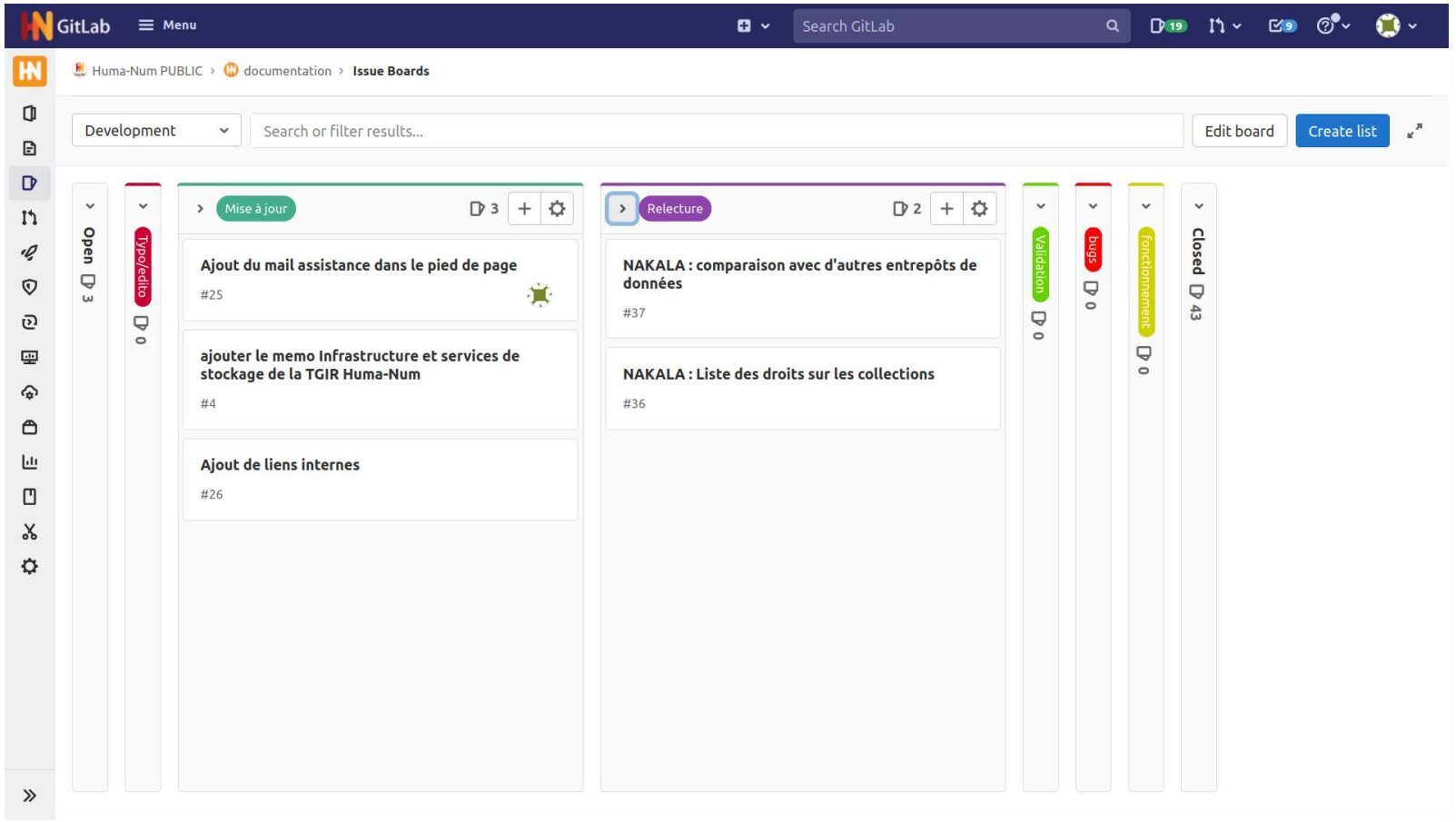
Représentations diverses des actions : listes, log, comparateur ligne à ligne, visualisation en graphe des branches et des commits

→ consolident et synthétisent les activités unitaires et individuelles en une activité collective.

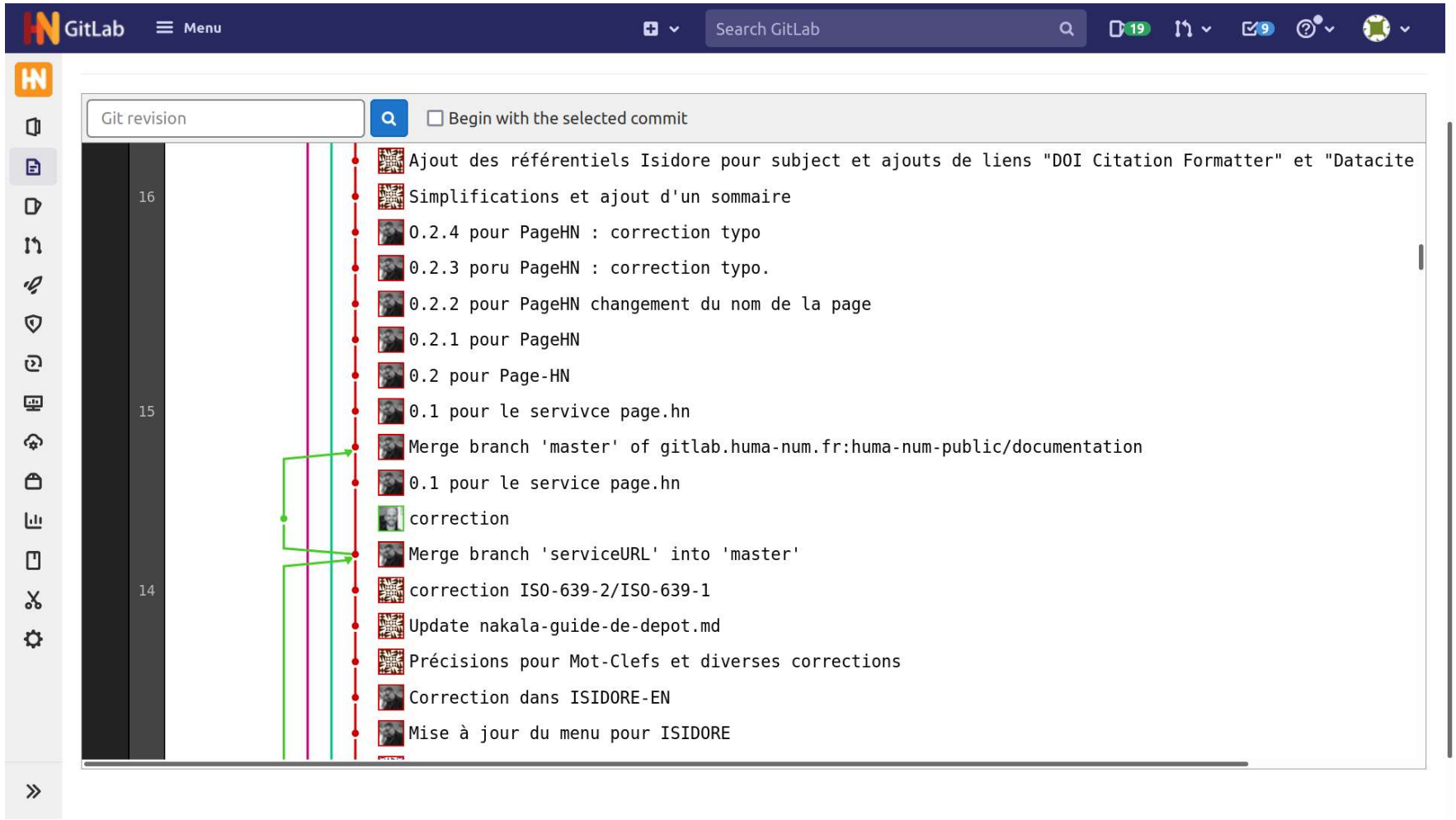
- Boards ou tableau de bord: permet d’organiser à sa guise les tickets/tâches, selon un double système de mot-clés et d’avancement.



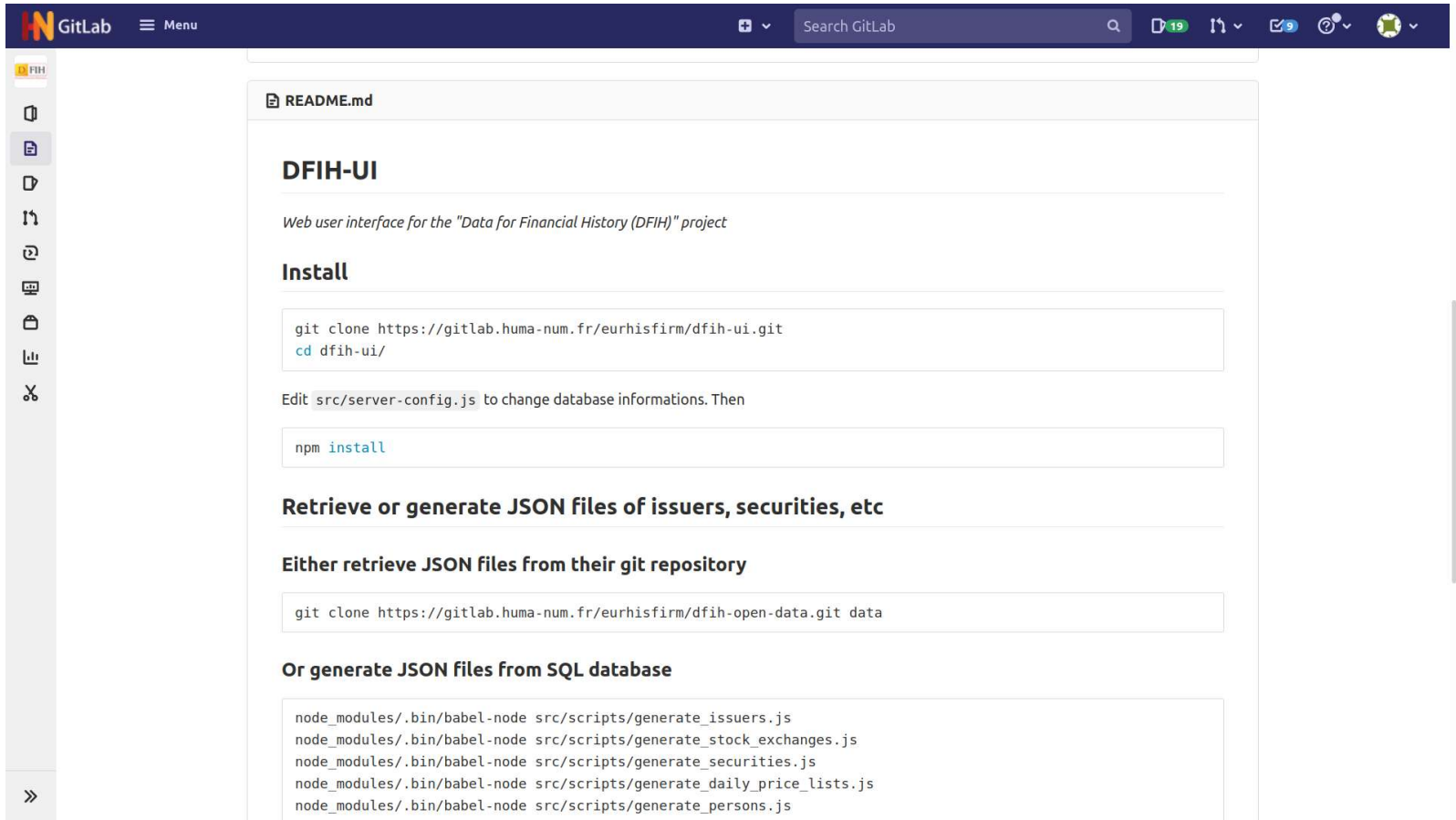
- Boards#2


GitLab
Menu
Search GitLab
Huma-Num PUBLIC › documentation › Issue Boards
Development
Search or filter results...
Edit board
Create list
Open 3
Typo/edito 0
Mise à jour 3
Ajout du mail assistance dans le pied de page
#25
ajouter le memo Infrastructure et services de stockage de la TGIR Huma-Num
#4
Ajout de liens internes
#26
Relecture 2
NAKALA : comparaison avec d'autres entrepôts de données
#37
NAKALA : Liste des droits sur les collections
#36
Validation 0
bugs 0
fonctionnement 0
Closed 43

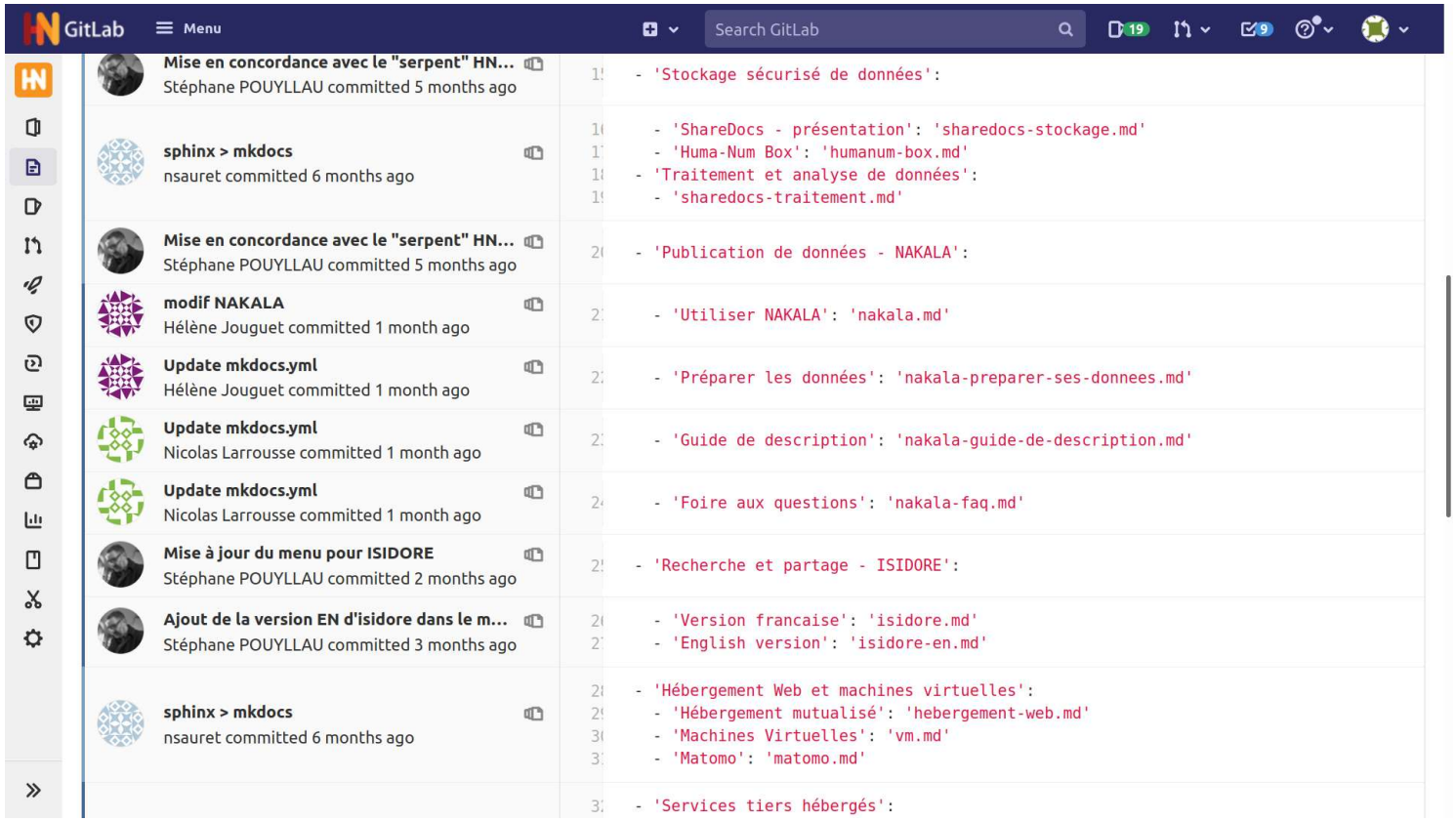
- Graph / Merge / fork: visualisation en graphe d’une branche et de sa fusion dans la branche principale



- Espace d’écriture et de publication, le readme vient documenter un répertoire ou un sous-dossier.



- Blame: présente ligne par lignes les contributions à un document.



- Accès au service Gitlab : http://humanid.huma-num.fr
- https://documentation.huma-num.fr/gitlab/
- Il s'agit d'une implémentation du logiciel gitlab hébergée sur l’infrastructure Huma-Num et maintenue par l’équipe

Huma-Num
la TGIR des humanités numériques

3) Le traitement

Points d’attention sur les formats de fichiers.

- Quels enjeux dans le choix des formats et codages ?
- Comment choisir les formats pour le projet ?
- Comment vérifier que les données sont bien formatées ?
- Comment convertir les données pour atteindre les formats choisis ?

Huma-Num
la TGIR des humanités numériques

Quels sont les risques ?

Dépendance

ex un format fermé lisible que par un logiciel propriétaire d'un éditeur dont on ne connaît pas la feuille de route

Non maîtrise de la qualité des données

ex impossibilité de valider par un outil autre que celui de l'éditeur

Perte de l'information

ex
- format propriétaire dont l'éditeur ne maintient plus les outils d'édition/lecture

Comment évaluer les risques ?

Critères basés sur les spécifications (Cf. guide du CINES)

- Existantes (et non pas embarquée dans l'outil d'édition)
- Publiques / Standardisées / Normalisées
- Avec ou sans entrave légale (brevet, copyright…)
- Qualité des spécifications
  - Complexité, taille des spécifications
  - Spécifications qui en cachent d’autres (dépendances, formats conteneurs...)
  - Des zones d’ombre non spécifiées
- Spécifications auditables. Par exemple le format CSV spécifié dans le RFC 4180
- « Due to lack of a single specification, there are considerable differences among implementations. Implementors should "be conservative in what you do, be liberal in what you accept from others" (RFC 793 [8]) when processing CSV files »

Critères liés au marché (existence d’outils, d’alternative, d’utilisateurs)

Huma-Num
la TGIR des humanités numériques

Preservation Risk matrix

- NARA
- https://github.com/usnationalarchives/digital-preservation/tree/master/Digital_Preservation_Risk_Matrix

	A	B	C	D	E	F	G	H	I	J
1	NARA Guidance: Preferred	NARA Guidance: Acceptable	Numeric Risk Rating	Risk Level	NARA Format ID	Format Name	File Extension(s)	Category/Plan(s)	Is the format proprietary?	Does the format have a published open specification?
11			-26,00	High Risk	NF00182	Executable file	exe	Software and Code	-1	-1
12			-25,00	High Risk	NF00702	BlackBerry Binary Executable	cod	Software and Code	-1	-1
17			-25,00	High Risk	NF00473	Nikon RAW (NRW)	nrw	Digital Still Image	-1	-1
18			-25,00	High Risk	NF00474	Olympus RAW	orf	Digital Still Image	-1	-1
19			-25,00	High Risk	NF00356	PCPaint Image	pic; clp	Digital Still Image	-1	-1
20			-25,00	High Risk	NF00475	Pentax RAW	pef; ptx	Digital Still Image	-1	-1
21			-25,00	High Risk	NF00476	Sigma RAW	x3f	Digital Still Image	-1	-1
22			-25,00	High Risk	NF00477	Sony RAW	srf	Digital Still Image	-1	-1
634	X		45,00	Low Risk	NF00650	Broadcast Wave (BWF) unspecified version	wav	Digital Audio	-1	2
635	X		45,00	Low Risk	NF00561	eXtensible Markup Language 1.1	xml	Web Records; Softwar	-1	2
636			45,00	Low Risk	NF00646	Portable Document Format/Archiving (PDF/A-4)	pdf	Presentation and Publ	2	2
637	X		45,00	Low Risk	NF00406	SIARD 1.0	siard	Databases	-1	2
647			47,00	Low Risk	NF00618	SIARD 2.0	siard	Databases	-1	2
648	X	X	48,00	Low Risk	NF00602	Portable Document Format/Archiving (PDF/A-2a) accessible	pdf	Presentation and Publ	2	2
649	X	X	48,00	Low Risk	NF00634	Portable Document Format/Archiving (PDF/A-2b) basic	pdf	Presentation and Publ	2	2
650			48,00	Low Risk	NF00642	Portable Document Format/Archiving (PDF/A-2u) unicode	pdf	Presentation and Publ	2	2
651			49,00	Low Risk	NF00619	SIARD 2.1.1	siard	Databases	-1	2

Huma-Num
la TGIR des humanités numériques

Liste des formats validables

- CINES
- https://facile.cines.fr/

Liste des formats validables

⚠ Attention : le validateur de formats permet de valider certains formats qui ne sont pas pris en charge par la plateforme d'archivage du CINES.

Format	Nom	PRONOM PIUD	Type MIME	Commentaire	Archivable dans PAC
AAC AAC	Advanced Audio Codings	[fmt/199]		Format Mpeg-4 contenant uniquement un flux audio au format AAC.	✔
AIFF PCM	Audio Interchange File Format	[fmt/414]	[audio/x-aif, audio/x-aiff]	Format audio contenant uniquement un flux PCM.	✔
APNG	Animated Portable Network Graphics	[fmt/935]	[image/vnd.mozilla.apng, image/apng]	L'APNG est une extension du format PNG permettant de réaliser des animations graphiques.	✖
DAE UTF-8 1.4.1	Collada		[application/xml]	Format permettant de stocker des données géométriques sous forme de scènes (plusieurs objets combinés dans le même référentiel), et d'y ajouter des informations supplémentaires pour décrire la scène et les objets (matériaux, environement lumineux, animations, ...) ou pour ajouter des notions sémantiques (relations entre les objets, découpage d'un objet en plusieurs éléments fonctionnels, etc...).	✖
FLAC FLAC 1.2.1	Free Lossless Audio Codec	[fmt/279]	[audio/ogg, audio/x-flac]	Format audio compressé sans perte.	✔
GIF 87a	Graphics Interchange Format	[fmt/3]	[image/gif]	Format image pouvant contenir également des animations.	✔
GIF 89a	Graphics Interchange Format	[fmt/4]	[image/gif]	Format image pouvant contenir également des animations.	✔
GeoTIFF	Geographic Tagged Image File Format	[fmt/155]	[image/tiff]	Format dérivé du TIFF contenant des informations de géoréférencement et de géolocalisation.	✔
HDF5 1.0	Hierarchical Data Format	[fmt/286]		Format de données à caractère scientifique.	✖
HDF5 2.0	Hierarchical Data Format	[fmt/287]		Format de données à caractère scientifique.	✖

Huma-Num

Huma-Num
la TGIR des humanités numériques

Les formats : un enjeux pour la fairisation des données

Pour qu'un fichier de données soit lisible et réutilisable

1) Son format doit être identifié explicitement et précisément
2) Sa forme doit être valide (conforme aux spécifications des formats et codages utilisés)
3) Le format doit être maintenable ou alors il convient d'envisager une conversion vers un autre format

Huma-Num

Huma-Num
la TGIR des humanités numériques

1) Comment identifier un format ?

- Comment « deviner » le format d'un fichier ?
  - Il existe des indices : l'extension dans le nom du fichier ; l'outil de lecture ; les fonctions « enregistrer sous... »
  - Il existe des outils d'aide à l'identification (DROID, FIDO, file,...)
- Où l'indiquer ?
  - Dans le DMP ; dans les métadonnées ; de manière normalisée en utilisant des identifiants issus de référentiels (type-mime, PRONOM)
- Pourquoi ?
  - Guider l'interprétation des données

Huma-Num
la TGIR des humanités numériques

Exemple d'informations sur un dans PRONOM

The National Archives

Simple search | File format | PRONOM Unique Identifier | Software | Vendor | Lifecycles | Migration Pathways

Details for: Acrobat PDF 1.5 - Portable Document Format 1.5

Save as... XML | CSV Print

Go to: Summary | Documentation › | Signatures › | Compression › | Character encoding › | Rights › | Reference files › Properties ›

Summary

Name	Acrobat PDF 1.5 - Portable Document Format
Version	1.5
Other names	PDF (1.5)
Identifiers	MIME: application/pdf Apple Uniform Type Identifier: com.adobe.pdf PUID: fmt/19
Family	
Classification	Page Description
Disclosure	Full
Description	Portable Document Format is a platform-independent format for representing formatted documents, developed by Adobe Systems Incorporated. It is the native format of Adobe's Acrobat family of software products, version 1.5 corresponding to the release of Acrobat 6.0. PDF is based on, and shares the same imaging model as, the PostScript page description language. A PDF file comprises a Header section, a Body section containing the objects which make up the document, a Cross Reference Table, and a Trailer section. PDF files can contain a wide variety of content, including text, images, video and audio.
Orientation	Binary
Byte order	Big-endian (Motorola)
Related file formats	Has lower priority than Acrobat PDF/A - Portable Document Format (1a) Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange 1:1999 Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange 1:2001 Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange 1a:2003 Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange 2:2003 Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange 3:2003 Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange 1a:2001 Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange 3:2002 Has lower priority than Acrobat PDF/A - Portable Document Format (1b) Has lower priority than Acrobat PDF/A - Portable Document Format (2a) Has lower priority than Acrobat PDF/A - Portable Document Format (2b) Has lower priority than Acrobat PDF/A - Portable Document Format (2u) Has lower priority than Acrobat PDF/A - Portable Document Format (3a)

Huma-Num

Huma-Num
la TGIR des humanités numériques

1. Identification

Exemple de découverte des formats avec DROID (TNA)

Demo

Resource	Extension	Format	Version	PUID	Mime type
C:\ANF2021-HN\data-avant\ANF2015-Francart.zip	zip	ZIP Format		x-fmt/263	application/zip
ANF2015-Francart					
01 - OAI-PMH-v1.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
02 - Web de donnees et interoperabilite-v1.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
03 - Web de Donn+€es - une introduction - v10 - HumaNum.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
04 - URIs permanentes et exemples.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
05 - rdf-in-a-nutshell-v2.9-fr.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
06 - SPARQL 1.0 - v5.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
07 - SPARQL - Exercices Nakala Isidore - v1.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
08 - SKOS-tf-v9-fr.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
C:\ANF2021-HN\data-avant\ANF2015-MOREL-C.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
C:\ANF2021-HN\data-avant\equipe-HN-2021.csv	csv	Comma Separated Values		x-fmt/18	text/csv
C:\ANF2021-HN\data-avant\guide-module-isidore-suggestions.pdf	pdf	Acrobat PDF 1.6 - Portable Document Format	1.6	fmt/20	application/pdf
C:\ANF2021-HN\data-avant\HUMA-NUM_PPTX_prez.pptx	pptx	Microsoft Powerpoint for Windows	2007 onwards	fmt/215	application/vnd.openxmlformats-officedocument.presentatio.
C:\ANF2021-HN\data-avant\Human-numGBverticale.ai	ai ⚠	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
C:\ANF2021-HN\data-avant\Human-numGBverticale.jpg	jpg	Raw JPEG Stream		fmt/41	image/jpeg
C:\ANF2021-HN\data-avant\Human-numGBverticale.png	png	Portable Network Graphics	1.0	fmt/11	image/png
C:\ANF2021-HN\data-avant\ISIDORE Suggestions_low.wmv	wmv ⚠	Windows Media Audio		fmt/132	audio/x-ms-wma
C:\ANF2021-HN\data-avant\ISIDORE_widget_2-0.png	png	Portable Network Graphics	1.2	fmt/13	image/png
C:\ANF2021-HN\data-avant\lettre_info25-iINSHS_TribuneHumaNum.pdf	pdf	Acrobat PDF 1.4 - Portable Document Format	1.4	fmt/18	application/pdf
C:\ANF2021-HN\data-avant\logo-grand-nakala-rvb.png	png	Portable Network Graphics	1.2	fmt/13	image/png
C:\ANF2021-HN\data-avant\logo_rencontre_HN-2026.png	png	Portable Network Graphics	1.2	fmt/13	image/png
C:\ANF2021-HN\data-avant\nakala-faq.md	md	Markdown		fmt/1149	text/markdown
C:\ANF2021-HN\data-avant\NAKALA_et_IIIF.mp4	mp4	MPEG-4 Media File		fmt/199	application/mp4, video/mp4
C:\ANF2021-HN\data-avant\Poster_ISIDORE_DH2018.pdf	pdf	Acrobat PDF 1.3 - Portable Document Format	1.3	fmt/17	application/pdf
C:\ANF2021-HN\data-avant\Poster_ISIDORE_DH2018.png	png	Portable Network Graphics	1.0	fmt/11	image/png
C:\ANF2021-HN\data-avant\VIGNETTE_NAKALA_et_IIIF.png	png	Portable Network Graphics	1.2	fmt/13	image/png

Resource	Extension	Format	Version	PUID	Mime type
C:\ANF2021-HN\data-avant\ANF2015-Francart.zip	zip	ZIP Format		x-fmt/263	application/zip
ANF2015-Francart					
01 - OAI-PMH-v1.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
02 - Web de donnees et interoperabilite-v1.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
03 - Web de Donn+ es - une introduction - v10 - HumaNum.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
04 - URIs permanentes et exemples.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
05 - rdf-in-a-nutshell-v2.9-fr.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
06 - SPARQL 1.0 - v5.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
07 - SPARQL - Exercices Nakala Isidore - v1.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
08 - SKOS-tf-v9-fr.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
C:\ANF2021-HN\data-avant\ANF2015-MOREL-C.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
C:\ANF2021-HN\data-avant\equipe-HN-2021.csv	csv	Comma Separated Values		x-fmt/18	text/csv
C:\ANF2021-HN\data-avant\guide-module-isidore-suggestions.pdf	pdf	Acrobat PDF 1.6 - Portable Document Format	1.6	fmt/20	application/pdf
C:\ANF2021-HN\data-avant\HUMA-NUM_PPTX_prez.pptx	pptx	Microsoft Powerpoint for Windows	2007 onwards	fmt/215	application/vnd.openxmlformats-officedocument.presentatio.
C:\ANF2021-HN\data-avant\Human-numGBverticale.ai	ai ⚠	Acrobat PDF 1.5 - Portable Document Format	1.5	fmt/19	application/pdf
C:\ANF2021-HN\data-avant\Human-numGBverticale.jpg	jpg	Raw JPEG Stream		fmt/41	image/jpeg
C:\ANF2021-HN\data-avant\Human-numGBverticale.png	png	Portable Network Graphics	1.0	fmt/11	image/png
C:\ANF2021-HN\data-avant\ISIDORE Suggestions_low.wmv	wmv ⚠	Windows Media Audio		fmt/131	audio/x-ms-wma
C:\ANF2021-HN\data-avant\ISIDORE_widget_2-0.png	png	Portable Network Graphics	1.2	fmt/13	image/png
C:\ANF2021-HN\data-avant\lettre_info25-iINSHS_TribuneHumaNum.pdf	pdf	Acrobat PDF 1.4 - Portable Document Format	1.4	fmt/18	application/pdf
C:\ANF2021-HN\data-avant\logo-grand-nakala-rvb.png	png	Portable Network Graphics	1.2	fmt/13	image/png
C:\ANF2021-HN\data-avant\logo_rencontre_HN-2026.png	png	Portable Network Graphics	1.2	fmt/13	image/png
C:\ANF2021-HN\data-avant\nakala-faq.md	md	Markdown		fmt/1149	text/markdown
C:\ANF2021-HN\data-avant\NAKALA_et_IIIF.mp4	mp4	MPEG-4 Media File		fmt/199	application/mp4, video/mp4
C:\ANF2021-HN\data-avant\Poster_ISIDORE_DH2018.pdf	pdf	Acrobat PDF 1.3 - Portable Document Format	1.3	fmt/17	application/pdf
C:\ANF2021-HN\data-avant\Poster_ISIDORE_DH2018.png	png	Portable Network Graphics	1.0	fmt/11	image/png
C:\ANF2021-HN\data-avant\VIGNETTE_NAKALA_et_IIIF.png	png	Portable Network Graphics	1.2	fmt/13	image/png

Huma-Num
la TGIR des humanités numériques

2. Comment tester la validité d'un fichier ?

- La lecture avec son éditeur ne suffit pas
  - Peu fiable : les éditeurs sont souvent très tolérants aux erreurs afin de ne pas bloquer la lecture
  - Problème de juge/partie
- Il existe des outils dont c'est la finalité
  - Spécialisés : jpylyzer (Open Planets Fondation) pour le JPEG2000
  - Généralistes : Jhove (JSTOR and the Harvard Library)
  - Intégrateurs : facile (CINES), FITS (Harvard Library)

Qui valide ?

- En dernier ressort, les organismes en charge de la conservation des l'information effectuent des tests de validation pour guider leur choix de prise en charge ou de niveau de service

CINES (Centre informatique national de l'enseignement supérieur)

- Un guide méthodologique
- Une liste de formats acceptés
- Un outil de test : Facile

Matrice des risques de la NARA (National Archives and Records Administration)

- Une méthodologie
- Des critères d’évaluation
- Une liste de formats avec évaluation

Huma-Num
la TGIR des humanités numériques

2. Validation

Validation avec facile

- Formats non pris en charge
  - jpeg-raw, md et mwv, pptx
- 1 fichier pdf non valide
  - Message : « Nombre d'objet ou de flux d'objet invalide »
- 1 format obsolete
  - PDF v. 1.3

Détails	Fichier	Format identifié	Bien formé	Valide	Archivable dans PAC	Commentaire
	nakala-faq.md		✗	✗	✗	
	Poster_ISIDORE_DH2018.pdf	PDF 1.3	✓	✓	✗	
	Human-numGBverticale.png	PNG 1.0	✓	✓	✓	
	equipe-HN-2021.csv	TXT UTF-8	✓	✓	✓	
	VIGNETTE_NAKALA_et_IIIF.png	PNG 1.2	✓	✓	✓	
	Poster_ISIDORE_DH2018.png	PNG 1.0	✓	✓	✓	
	logo-grand-nakala-rvb.png	PNG 1.2	✓	✓	✓	
	logo_rencontre_HN-2026.png	PNG 1.2	✓	✓	✓	
	Human-numGBverticale.jpg	JPEG RAW	✓	✓	✗	
	guide-module-isidore-suggestions.pdf	PDF 1.6	✓	✓	✓	
	ISIDORE_widget_2-0.png	PNG 1.2	✓	✓	✓	
	Human-numGBverticale.ai	PDF 1.5	✓	✓	✓	
	ANF2015-MOREL-C.pdf	PDF 1.5	✗	✗	✗	Corriger automatiquement votre fichier Consulter nos tutoriels pour corriger votre fichier
	lettre_info25-iINSHS_TribuneHumaNum.pdf	PDF 1.4	✓	✓	✓	
	ANF2015-Francart.zip		✗	✗	✗	
	ISIDORE Suggestions_low.wmv		✗	✗	✗	
	HUMA-NUM_PPTX_prez.pptx		✗	✗	✗	
	NAKALA_et_IIIF.mp4	MPEG-4 AVC/AAC LC	✓	✓	✓	

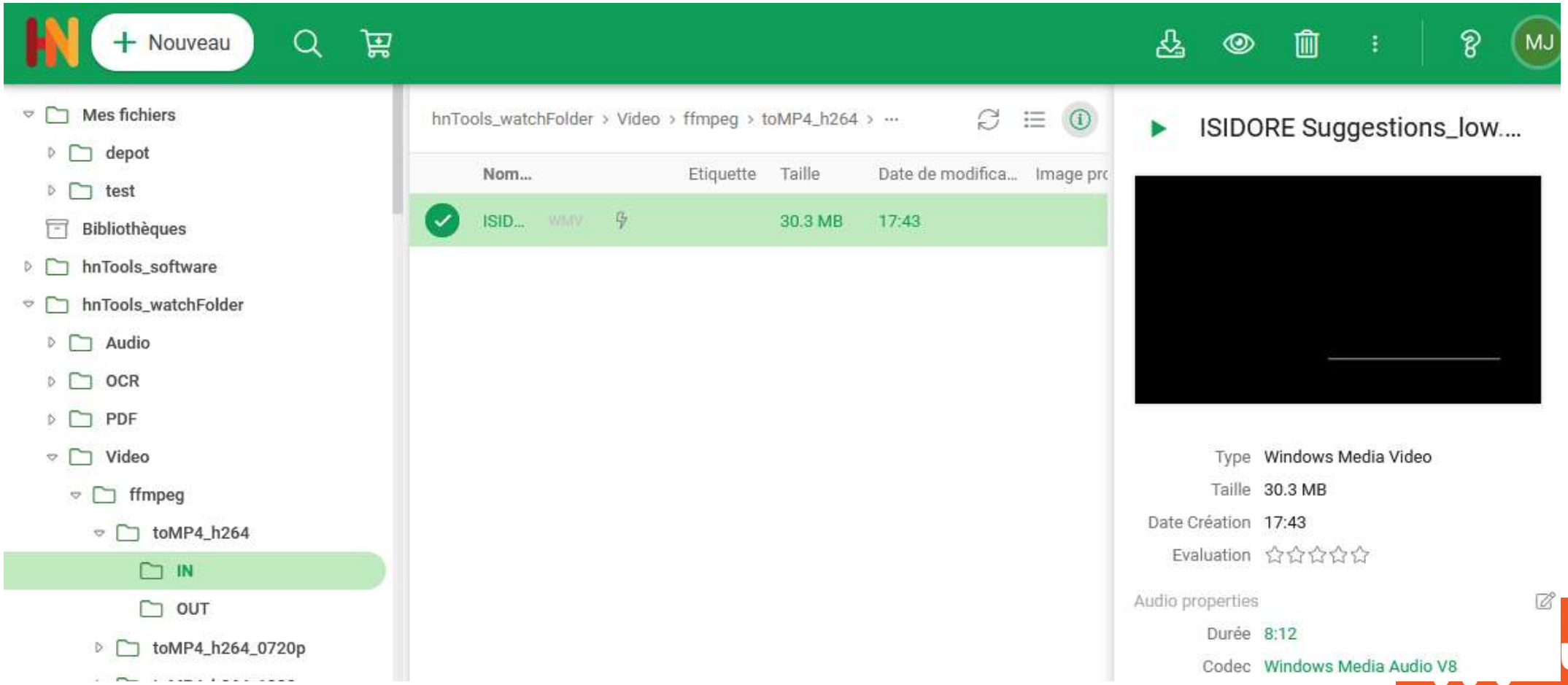
Huma-Num
la TGIR des humanités numériques

Si besoin de conversion de format

- Pour lutter contre l'obsolescence, pour maîtriser les risques, pour faciliter des usages
  - Trouver un format équivalent pour exprimer toute l'information et permettre de conserver les usages peut s'avérer difficile. Cela se fait parfois avec une perte d'information qu'il faut évaluer
  - Trouver des outils de conversion (il en existe plein) et les qualifier (résultat fidèle et valide)
    - Par ex. ffmpeg pour l'audio-visuel (utilisable à travers Sharedocs)
  - Documenter dans les métadonnées le statut de chaque format et/ou leur mode de production
    - Original vs. Diffusion
    - Issue de la conversion de...

Huma-Num
la TGIR des humanités numériques

Conversion de format pour normalisation WMV → MP4/H264



Huma-Num
la TGIR des humanités numériques

Conversion de format pour normalisation WMV → MP4/H264

[HUMA-NUM] Video transcoding is finished Boîte de réception ×

sysadmin@huma-num.fr
À moi

anglais › français Traduire le message

HN-Tools : job information

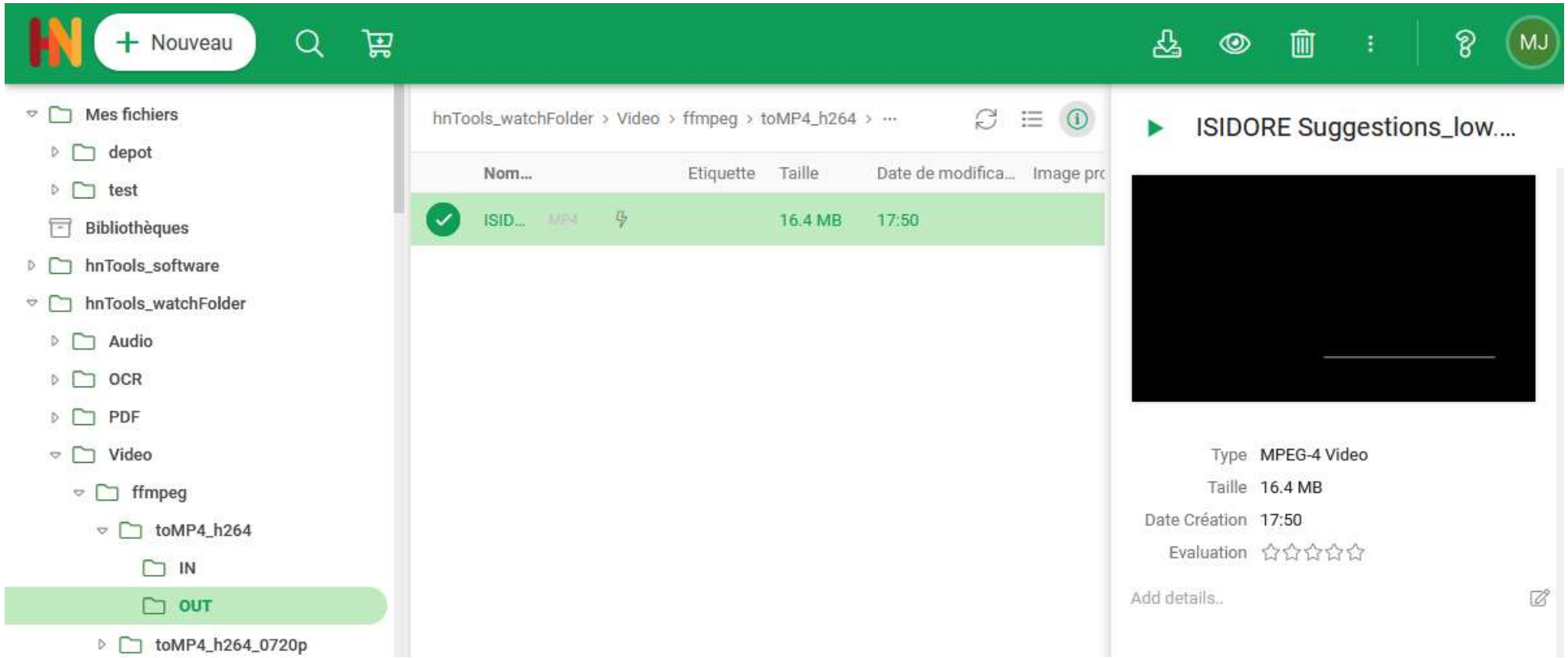
```
--------------------------------------------------------------------------------
--- INPUT
--------------------------------------------------------------------------------
File name        : ISIDORE Suggestions_low.wmv
File path        : mjacobson/Video/ffmpeg/toMP4_h264/IN
File size        : 30.26mo
File extension   : wmv
File type        : video/x-ms-asf
File hash        : 7f6af8e56647ea7079e21af04dc65dc337572ef254613fbf8005d14e
File date        : 1631202208
Submit by        : Jacobson Michel (michel.jacobson@gmail.com)

--------------------------------------------------------------------------------
--- OUTPUT
--------------------------------------------------------------------------------
Output file      : mjacobson/Video/ffmpeg/toMP4_h264/OUT/ISIDORE Suggestions_low.mp4
Tool             : Video
Engine           : ffmpeg
Preset 1         : toMP4_h264
Execution time   : 2mn 24.191s
```

Huma-Num

Huma-Num
la TGIR des humanités numériques

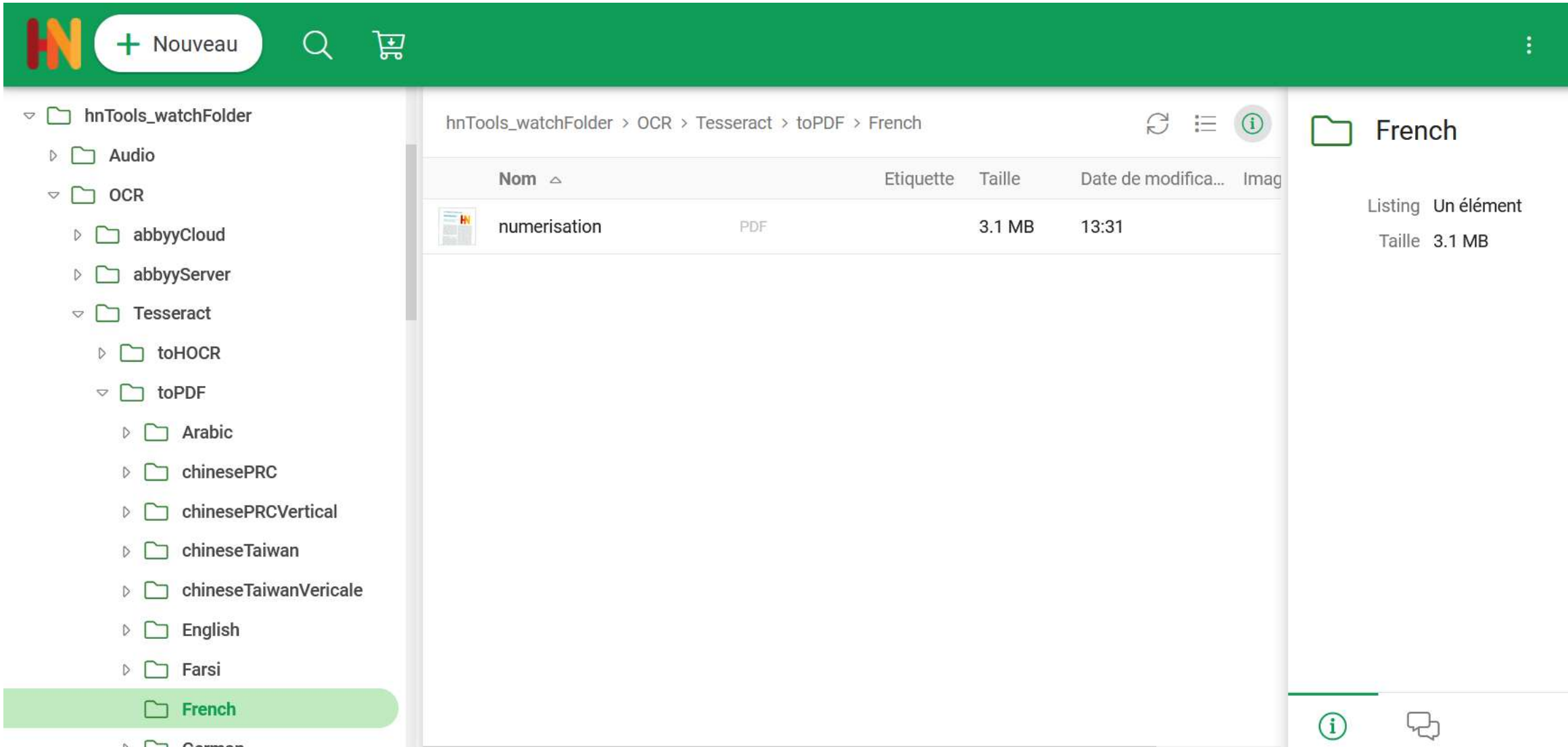
Conversion de format pour normalisation WMV → MP4/H264



Huma-Num

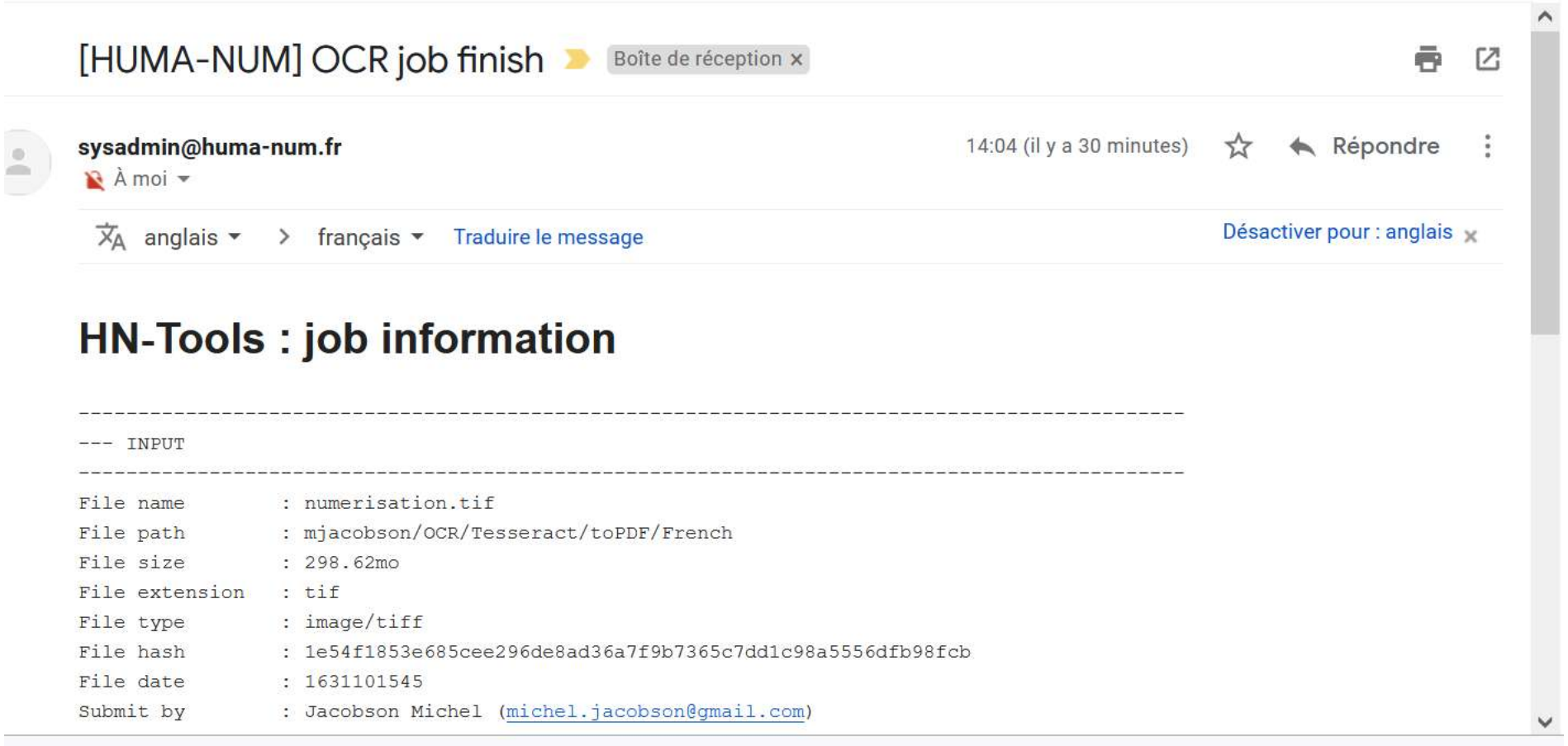
Huma-Num
la TGIR des humanités numériques

OCRisation (conversion pour enrichir)



Huma-Num
la TGIR des humanités numériques

OCRisation (conversion pour enrichir)



[HUMA-NUM] OCR job finish Boîte de réception ×

sysadmin@huma-num.fr
À moi
14:04 (il y a 30 minutes) Répondre

anglais > français Traduire le message Désactiver pour : anglais ×

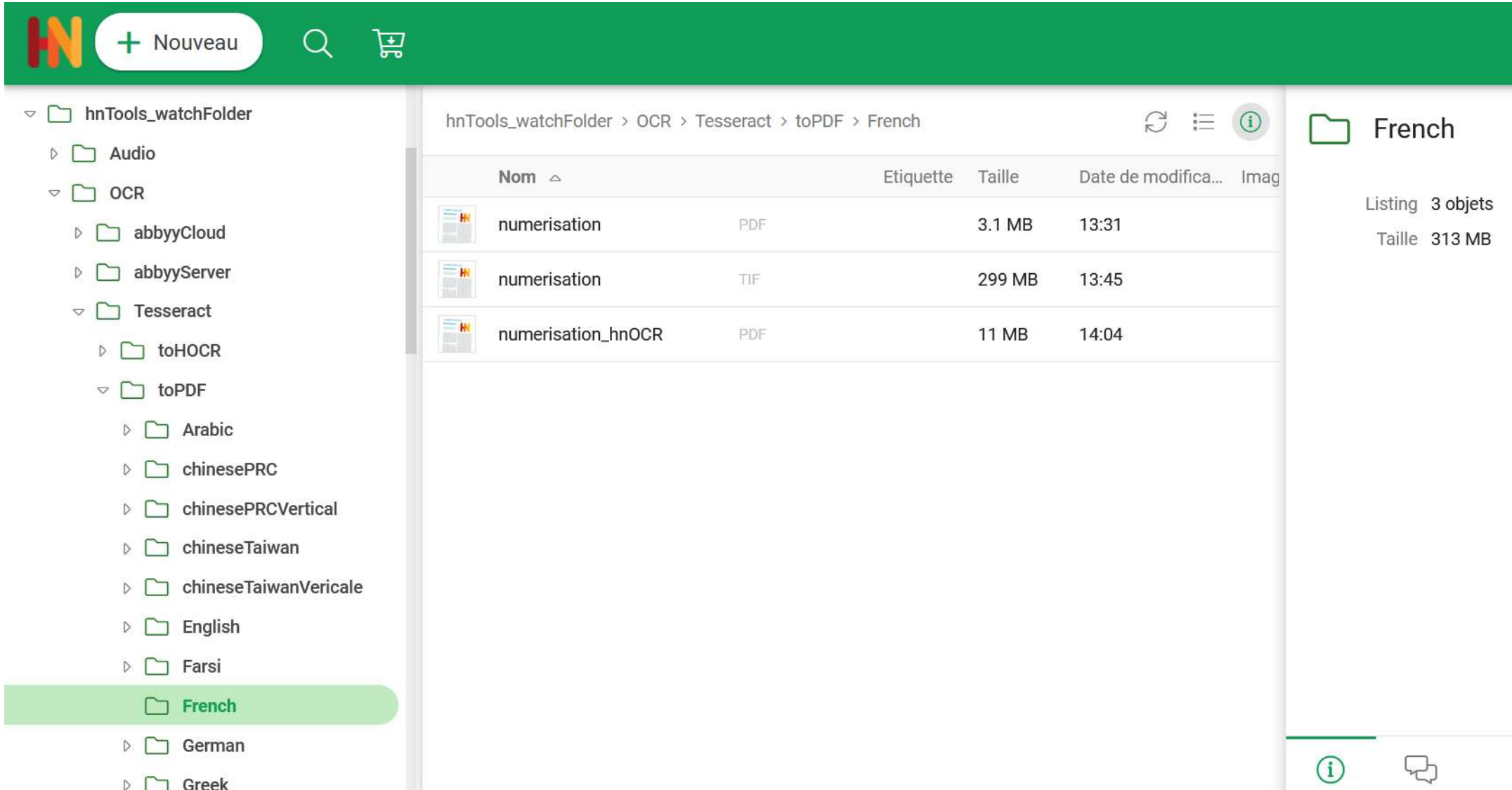
HN-Tools : job information

```
--------------------------------------------------------------------------------
--- INPUT
--------------------------------------------------------------------------------
File name        : numerisation.tif
File path        : mjacobson/OCR/Tesseract/toPDF/French
File size        : 298.62mo
File extension   : tif
File type        : image/tiff
File hash        : 1e54f1853e685cee296de8ad36a7f9b7365c7dd1c98a5556dfb98fcb
File date        : 1631101545
Submit by        : Jacobson Michel (michel.jacobson@gmail.com)
```

Huma-Num

Huma-Num
la TGIR des humanités numériques

OCRisation (conversion pour enrichir)



Huma-Num
la TGIR des humanités numériques

Retour sur identification/validation

Resource	Extension	Format	Version	Mime type	PUID
C:\ANF2021-HN\data-apres					
ANF2015-MOREL-C_corrige.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	application/pdf	fmt/19
HUMA-NUM_PPTX_prez.pdf	pdf	Acrobat PDF 1.6 - Portable Document Format	1.6	application/pdf	fmt/20
Human-numGBverticale.jpg	jpg	JPEG File Interchange Format	1.01	image/jpeg	fmt/43
ISIDORE Suggestions_low.mp4	mp4	MPEG-4 Media File		application/mp4, vide...	fmt/199
lettre_info25-iINSHS_TribuneHumaNumOCR.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	application/pdf	fmt/19
nakala-faq.txt	txt	Plain Text File		text/plain	x-fmt/111

Détails	Fichier	Format identifié	Bien formé	Valide	Archivable dans PAC
	nakala-faq.txt	TXT UTF-8	✔	✔	✔
	Human-numGBverticale.jpg	JPEG 1.01	✔	✔	✔
	ANF2015-MOREL-C_corrige.pdf	PDF 1.5	✔	✔	✔
	HUMA-NUM_PPTX_prez.pdf	PDF 1.4	✔	✔	✔
	lettre_info25-iINSHS_TribuneHumaNumOCR.pdf	PDF 1.5	✔	✔	✔
	ISIDORE Suggestions_low.mp4	MPEG-4 AVC/AAC LC	✔	✔	✔

Huma-Num

Huma-Num
la TGIR des humanités numériques

Synthèse

- Le choix des formats et des codages est très important pour la FAIRisation des données
  - Ils conditionnent la lisibilité et la réutilisabilité des données
  - Il sont discriminants pour les systèmes de conservation à long terme
- Ces choix doivent être consigné dans le DMP
- Des points de contrôle doivent être mis en place tout au long du cycle de vie

Huma-Num

Huma-Num
la TGIR des humanités numériques

Merci pour votre écoute

-

Place aux questions /réponses

Huma-Num