Choisir un entrepôt de confiance : les principes FAIR en pratique avec Nakala

Intervenant : Xiaoou Wang, référent Huma-Num en Humanités numériques à la MSHS Sud-Est

Table des matières

Bonjour à toutes et à tous, et merci de venir aussi nombreux aujourd'hui.

Je me présente rapidement : je m'appelle Xiaoou Wang et je suis référent Humanités numériques à la MSHS Sud-Est.
Je me charge notamment de tout ce qui relève de l'application des outils numériques et de l'informatique aux problématiques en sciences humaines et sociales.

Aujourd'hui, le thème du webinaire est :

> **« Choisir un entrepôt de confiance : les principes FAIR en pratique avec Nakala »**

Vous avez probablement déjà entendu parler des principes FAIR.
Aujourd'hui, l'idée n'est pas vraiment de refaire un grand discours sur le fait que les principes FAIR sont géniaux ou indispensables à la science ouverte.
Je voudrais plutôt qu'on regarde leurs implications très concrètes dans vos pratiques de recherche.

Qu'est-ce qu'un entrepôt de confiance ?

Dans un premier temps, on va essayer de comprendre ce qu'est un entrepôt de confiance.

Le Collège des données de la recherche a établi un ensemble de critères d'exclusion.
Tous ces critères ont vraiment pour objectif de garantir la qualité des dépôts.

Par exemple :

- la nécessité de modérer les dépôts ;
- la présence d'identifiants pérennes ;
- la garantie de pérennité de l'infrastructure ;
- ou encore certaines questions liées aux licences et à la liberté académique.

Toute cette liste sert finalement à trier les entrepôts présents sur Internet, parce qu'aujourd'hui ils existent en très grand nombre et avec des niveaux de qualité très variables.

Et évidemment, ces entrepôts doivent aussi répondre aux principes FAIR :

- **Findable**
- **Accessible**
- **Interoperable**
- **Reusable**

On va d'ailleurs s'attarder un peu plus longtemps sur la notion d'interopérabilité, parce que c'est souvent le critère le plus abstrait, mais aussi l'un des plus importants.

Où trouver des entrepôts de confiance ?

Il existe ce qu'on appelle des entrepôts thématiques.

Sur le portail Recherche Data Gouv, vous pouvez trouver des entrepôts correspondant à différentes disciplines ou champs scientifiques.

Cette liste a été triée et validée par le Collège des données de la recherche.

Vous pouvez faire des recherches par discipline, par domaine, et consulter les différents critères d'évaluation.

Et dans le cas où il n'existe pas d'entrepôt spécifique à votre discipline, il existe aussi des entrepôts généralistes, comme Recherche Data Gouv.

Si vous cliquez sur les descriptions détaillées, vous trouverez toutes les informations concernant :

- les institutions porteuses ;
- les modérateurs ;
- les politiques de conservation ;
- et l'ensemble des critères utilisés pour évaluer les entrepôts.

Pourquoi déposer ses données dans un entrepôt de confiance ?

Premièrement, pour garantir la pérennité des données.

Et deuxièmement — et c'est souvent le point le plus concret — pour garantir leur visibilité maximale.

Parce qu'effectivement, ces entrepôts servent ensuite à faire identifier et moissonner automatiquement les données par différents catalogues scientifiques.

Exemple concret : Nakala

On va maintenant passer à un exemple concret avec Nakala, qui est un entrepôt de confiance en sciences humaines et sociales, et que vous pouvez retrouver sur la liste de Recherche Data Gouv.

Pour accéder à Nakala, il faut d'abord obtenir ce qu'on appelle un HumanID.

La procédure est assez simple.

Il existe plusieurs moyens de l'obtenir, notamment via HAL.

Une fois connecté, vous arrivez sur le portail des services Huma-Num, où vous pouvez demander l'accès à Nakala.

L'identifiant HumanID est créé instantanément, mais l'accès effectif à Nakala peut prendre quelques jours, parfois jusqu'à deux semaines.

Findable — Trouvable

On va maintenant passer au premier principe FAIR :

> **Findable — trouvable**

L'idée ici est que les données doivent être facilement retrouvables.

Le point le plus important est l'attribution d'un DOI, donc d'un identifiant pérenne unique.

Je vais prendre ici l'exemple d'un dataset issu de ma thèse.

Quand on ouvre le dataset, on voit immédiatement qu'il possède un DOI.

Pour déposer un dataset, les métadonnées obligatoires restent relativement limitées :

- le titre ;
- la date ;
- le créateur ;
- le type de données ;
- la licence ;
- la description.

Mais dans la pratique, plus les métadonnées sont riches, plus les données seront visibles et faciles à retrouver.

Par exemple ici :

- j'ai ajouté des mots-clés ;
- des descriptions en français et en anglais ;

- des informations sur les financements ;
- et surtout des relations vers d'autres publications.

L'idée est vraiment de créer un réseau entre :

- articles ;
- données ;
- publications ;
- corpus ;
- projets.

Mauvaise pratique : dépôt multiple

Je vais aussi montrer une mauvaise pratique que j'ai moi-même faite à la fin de ma thèse.

À l'époque, j'ai essayé de déposer mes datasets un peu partout.

Le problème, c'est qu'à chaque dépôt, un nouveau DOI est créé.
Et une fois créé, ce DOI continue d'exister.

Résultat :

- les citations se fragmentent ;
- plusieurs versions du dataset circulent ;
- et les utilisateurs ne savent plus à quelle version se référer.

Donc si vous avez déjà fait ce type d'erreur, il est important d'indiquer clairement quelle est la version canonique du dataset.

Accessible — Accessible

On va maintenant passer au principe :

> **Accessible — accessible**

Ici, j'aimerais vraiment faire une nuance importante entre :

- accessible ;
- et ouvert à tous.

Quand les gens entendent « dépôt de données », ils pensent souvent que les données seront automatiquement ouvertes publiquement.

Mais ce n'est pas forcément le cas.

On peut choisir :

- quelles parties rendre accessibles ;
- et quelles parties garder restreintes.

Dans l'exemple que je vais montrer :

- les métadonnées sont publiques ;
- le DOI est visible ;
- mais les fichiers eux-mêmes ne sont pas accessibles aux utilisateurs externes.

L'idée est donc aussi de partager l'existence et la description des données sans forcément exposer directement leur contenu.

Interoperable — Interopérable

On arrive maintenant au principe :

> **Interoperable — interopérable**

C'est probablement le principe le plus technique.

Formats ouverts

L'idée est de privilégier des formats ouverts :

- CSV ;
- JSON ;
- XML ;
- etc.

Cela permet de garantir une meilleure pérennité et facilite les échanges entre logiciels.

Vocabulaires contrôlés

Ensuite, il y a la question des vocabulaires contrôlés et des thésaurus.

L'objectif ici est d'utiliser des terminologies communes afin d'assurer une sémantique partagée.

Par exemple, sur FAIRsharing.org, on peut rechercher des thésaurus spécialisés selon les disciplines.

On trouve parfois des terminologies extrêmement spécifiques.

Métadonnées normalisées

Il est aussi important d'utiliser des schémas de métadonnées normalisés.

Il existe plusieurs centaines de schémas de métadonnées, mais en SHS, des standards comme :

- Dublin Core ;
- Dublin Core Qualifié ;
- ou DataCite

sont particulièrement utilisés.

Et on va voir pourquoi ces standards sont importants :
ils facilitent énormément l'interopérabilité et le moissonnage automatique.

Moissonnage automatique

Dans Nakala, ce qui se passe en coulisses, c'est que les métadonnées internes sont automatiquement transformées vers différents standards internationaux.

C'est ce qu'on appelle le mappage des métadonnées.

Concrètement :

vous déposez vos données une seule fois sur Nakala, puis elles peuvent être automatiquement moissonnées par différents catalogues scientifiques.

Par exemple :

- DataCite ;
- Google Dataset Search ;
- BASE ;
- et bientôt Recherche Data Gouv.

Donc il suffit finalement de déposer les données à un seul endroit pour qu'elles soient ensuite visibles dans plusieurs moteurs de recherche scientifiques.

Et ici, j'insiste vraiment sur l'importance des métadonnées riches.

Plus les métadonnées sont détaillées et standardisées, plus le moissonnage automatique fonctionnera correctement.

Dans le cas où les métadonnées sont incomplètes, certaines plateformes risquent même de ne pas référencer les données.

Schema.org et Google Dataset Search

Autre point intéressant :

certains catalogues réexposent ensuite les métadonnées dans d'autres formats.

Par exemple, DataCite réexpose les métadonnées en Schema.org, qui est le schéma privilégié par Google.

C'est ce qui explique pourquoi certains datasets deviennent visibles automatiquement dans Google Dataset Search.

Valorisation avec NakalaPress

Je voudrais aussi attirer votre attention sur un autre avantage de Nakala : NakalaPress.

L'idée est de valoriser les données via un site web statique généré automatiquement à partir du dataset.

Créer un site web prend normalement du temps, mais ici on peut générer quelque chose de relativement propre en quelques minutes seulement.

Dans mon cas, le site que je vais montrer a été réalisé en une dizaine de minutes.

Reusable — Réutilisable

On arrive maintenant au dernier principe :

> **Reusable — réutilisable**

Ici, la question principale concerne les licences.

Je ne vais pas faire aujourd'hui une introduction complète aux différents types de licences, mais il faut savoir que chaque entrepôt propose des politiques différentes.

Par exemple :

- Nakala propose plusieurs centaines de licences ;
- alors que d'autres plateformes, comme OpenNeuro, limitent beaucoup plus les choix et privilégient aujourd'hui CC0.

Donc lorsqu'on choisit un entrepôt, il faut aussi regarder :

- quelles licences sont proposées ;
- et quel niveau d'accompagnement juridique existe autour des données.

Conclusion

Nous approchons maintenant de la fin de cette présentation.

J'espère que ce webinaire vous aura permis de mieux comprendre :

- les principes FAIR ;
- ce qu'est un entrepôt de confiance ;
- et surtout les implications très concrètes de ces notions dans vos pratiques de recherche.

Si vous avez des questions, n’hésitez pas à me contacter.

Pour toutes les questions liées au cycle de vie des données, vous pouvez également contacter le guichet Recherche Data de l’université.

Prochain webinaire : anonymisation des données

Et j’en profite aussi pour faire une petite annonce concernant le prochain webinaire, qui portera sur l’anonymisation des données en sciences humaines et sociales.

Je présenterai notamment un outil que j’ai développé pour faciliter l’identification automatique des entités nommées :

- noms de personnes ;
- lieux ;
- dates ;
- numéros de téléphone ;
- etc.

L’outil permet ensuite :

- d’anonymiser automatiquement ces éléments ;
- mais aussi de générer un rapport documentant tout le processus d’anonymisation.

Donc si le sujet vous intéresse, n’hésitez pas à venir au prochain webinaire.

Merci beaucoup pour votre attention, et merci encore pour votre participation.