IIe Colloque international de l'association HUMANISTICA
Rennes, 10-12 mai 2021

"FAIR"iser des données : état des lieux, barrières et choix.

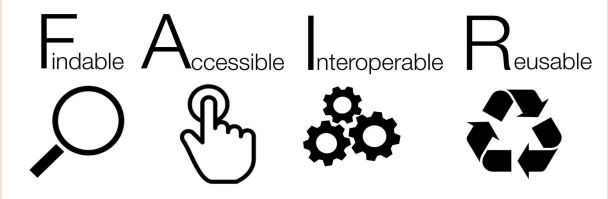
Une réflexion à partir des données des corpus d'auteurs.

Ioana Galleron (Université Sorbonne Nouvelle)
ioana.galleron@sorbonne-nouvelle.fr

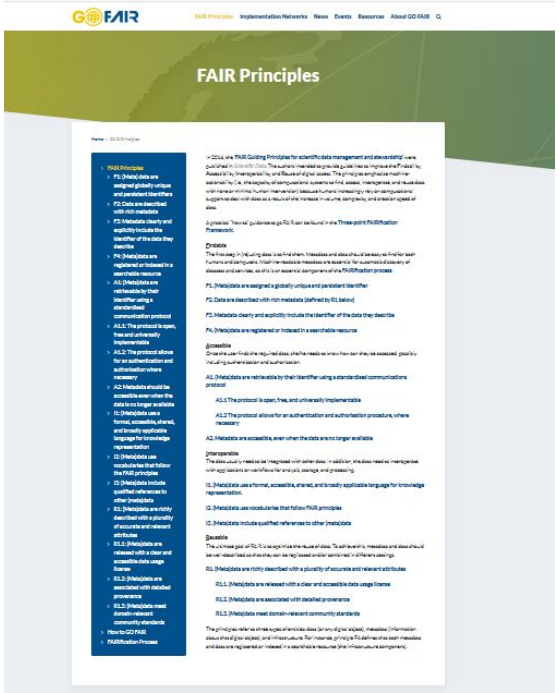
Fatiha Idmhand (Université de Poitiers)
fatihaidmhand@yahoo.es

Introduction

Pour quoi FAIR?



FAIRiser = ouvrir (vraiment) la science



Ouvrir la science = permettre la **réutilisation** des données

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

The FAIR Guiding Principles for scientific data management and stewardship (2016)

"There is an urgent need to improve the infrastructure supporting the **reuse of scholarly data**. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, **the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles**, and includes the rationale behind them, and some exemplar implementations in the community."

www.nature.com/sdata/

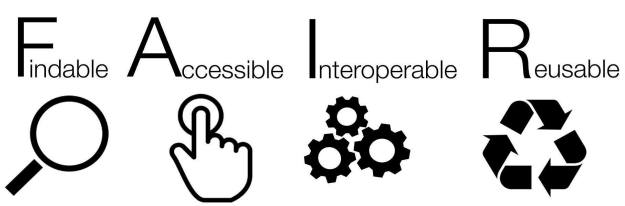
Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

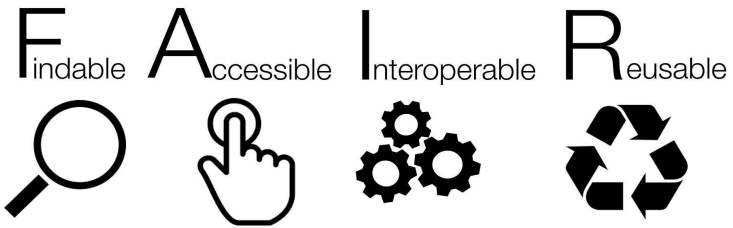
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards



Ouvrir la science = lever les barrières

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

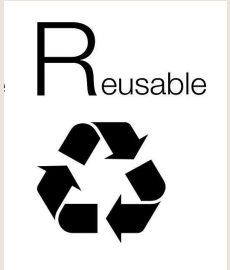


"Les principes FAIR définissent les caractéristiques que les ressources, outils, vocabulaires et infrastructures de données contemporaines devraient présenter pour faciliter la découverte et la réutilisation par des tiers. En définissant de manière minimale chaque principe directeur, **la barrière à l'entrée pour les producteurs, éditeurs et gestionnaires de données qui souhaitent rendre leurs fonds de données FAIR est volontairement maintenue aussi basse que possible.**" (Wilkinson *et. al.* 2016)

"Ces principes directeurs FAIR de haut niveau **précèdent les choix de mise en œuvre et ne suggèrent aucune technologie spécifique ou solution de mise en œuvre, normes, etc. De plus, les principes FAIR ne sont pas eux-mêmes une norme ou une spécification.** Ils agissent comme un guide pour les éditeurs de données et les responsables de la gestion des données pour pour les aider à évaluer si leurs choix particuliers de mise en œuvre rendent leurs artefacts numériques de recherche trouvables, accessibles et faciles à utiliser. Nous prévoyons que ces principes de haut niveau permettront un large éventail de comportements intégratifs et exploratoires, basés sur un large éventail de choix et d'implémentations technologiques." (Wilkinson *et. al.* 2016)

Qu'impliquent les principes FAIR? Réflexions à partir du cas des corpus d'auteurs du Consortium CAHIER

4. "Réutilisabilité"



Réutiliser: quoi? pourquoi? comment?

consortium cahier

Huma-Num

Le consortium CAHIER

Corpus d'Auteurs pour les Humanités

Informatisation, Édition, Recherche

De l'interopérabilité à la réutilisabilité des éditions électroniques

https://journals.openedition.org/revuehn/350

Freins:

- commettre des ressources à travailler "pour les autres"
- envisager les usages des autres
- activité de recherche non reconnue…

3 modalités:

- l'enquête préliminaire: CAHIER comme espace de rencontre et de discussion en amont de l'adhésion des projets
- la conceptualisation du travail scientifique (ex. "scholarly primitives" Unsworth)
- un mélange des deux

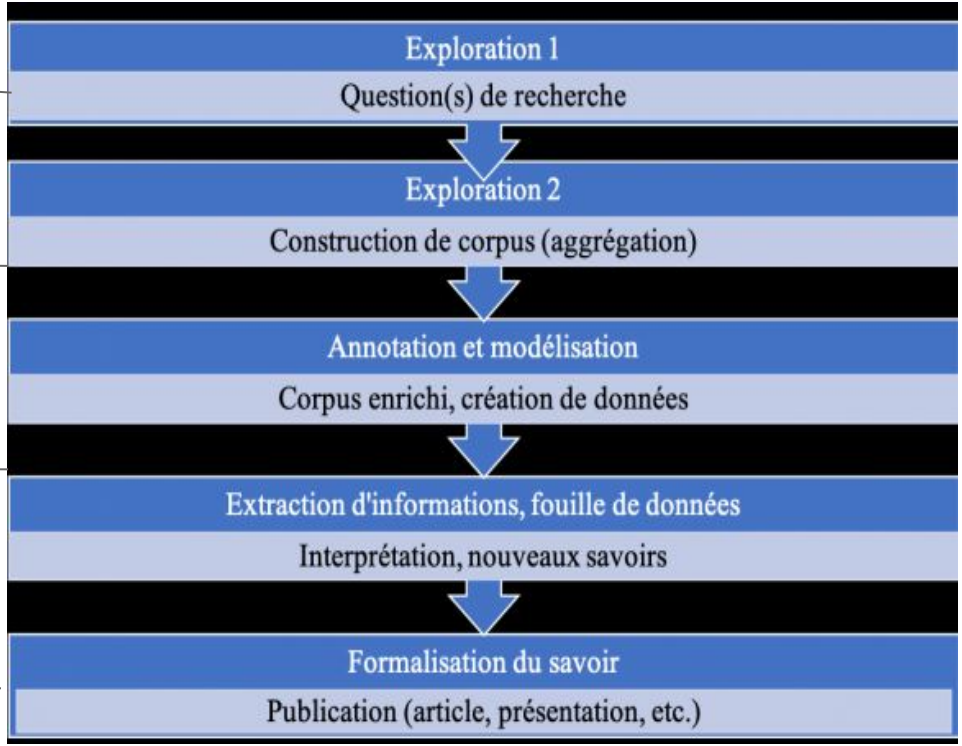
Penser la “reutilisabilité” en lien avec chaque étape du projet d’édition

Créer des espaces de publicité, afin de montrer l’intérêt du texte édité

Métadonnées “riches”, moissonnage par un aussi grand nombre de moteurs de recherche, formats “machine readable”, possibilité de télécharger les textes annotés

Transcription respectant les graphies d’origine, annotation des entités nommées, mise en lien

Assurer la citabilité



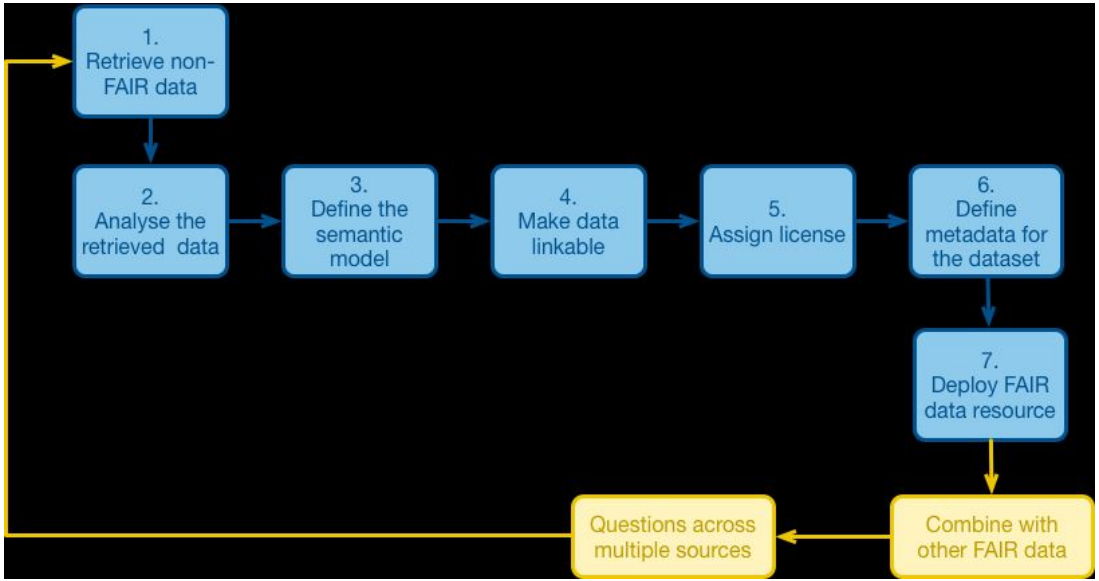
De l'interopérabilité à la réutilisabilité des éditions électroniques
https://journals.openedition.org/revuehn/350

ATTENTION
Les *FAIRdata principles* concernent les données ET les métadonnées

"Les principes des données FAIR s'appliquent aux métadonnées, aux données et à l'infrastructure de soutien (par exemple, les moteurs de recherche). **La plupart des exigences en matière de facilité de recherche et d'accessibilité peuvent être satisfaites au niveau des métadonnées.** L'interopérabilité et la réutilisation exigent davantage d'efforts au niveau des données.

Le schéma ci-joint décrit le processus de FAIRisation adopté par GO FAIR, en se concentrant sur les données, mais en indiquant également **le travail nécessaire pour les métadonnées**"

https://www.go-fair.org/fair-principles/fairification-process/



Les indicateurs FAIR

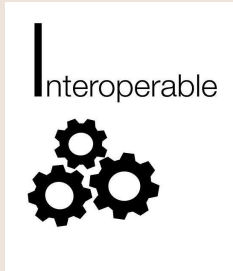
Voir le guide FAIR de CAHIER V2 et les recommandations de la Research Data Alliance
RDA : https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines#comment-form

Reusable

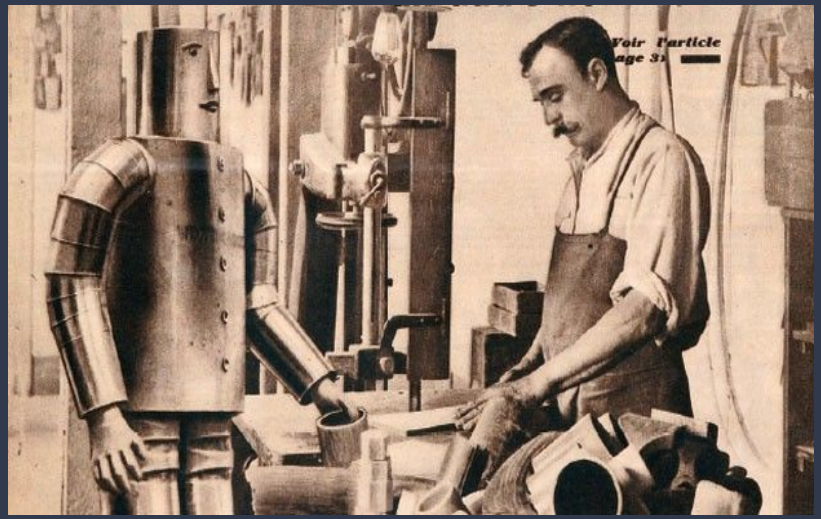
Reusable	Une pluralité d'attributs précis et pertinents sont fournis pour permettre la réutilisation	Essentiel
Reusable	Les métadonnées comprennent des informations sur la licence en vertu de laquelle les données peuvent être réutilisée	Essentiel
Reusable	Les métadonnées font référence à une licence de réutilisation standard	Important
Reusable	Les métadonnées font référence à une licence de réutilisation compréhensible par la machine	Important
Reusable	Les métadonnées comprennent des informations sur la provenance selon des normes communautaires spécifiques	Important
Reusable	Les métadonnées comprennent des informations sur la provenance selon une langue intercommunautaire	Utile
Reusable	Les métadonnées sont conformes à la norme d'une communauté	Essentiel
Reusable	Les données sont conformes à la norme d'une communauté	Essentiel
Reusable	Les métadonnées sont exprimées conformément à la norme d'une communauté compréhensible par la machine	Essentiel
Reusable	Les données sont exprimées conformément à la norme d'une communauté compréhensible par les machines	Important

Qu’impliquent les principes FAIR? Réflexions à partir du cas des corpus d’auteurs du Consortium CAHIER

3. “Interopérabilité”



Qu'est-ce qu'une donnée "interopérable"?



L'homme et la machine

Définition informatique :

capacité, pour deux ou plusieurs systèmes informatiques, à fonctionner ensemble

Définition humaine :

capacité, pour deux ou plusieurs communautés à s'entendre pour décrire leurs données de façon à pouvoir les faire fonctionner ensemble dans des systèmes informatiques différents

Technologies partagées / malentendus multiples

Freins:

- prendre conscience des divergences de pratique
- manque d'expérience dans l'identification de bonnes pratiques, la rédaction de vocabulaires contrôlés
- conflit rigueur/ complexité
- activité de recherche non-reconnue…

- pratiques différentes dans la définition d'un auteur, d'un créateur, d'un contributeur, d'un éditeur
- quelle typologie des titres?
- description des contenus (ex. genres textuels): la jungle des mots clés

etc.

CAHIER comme espace de discussion au sujet des:

- typologies des genres littéraires
- pratiques d'annotation
- headers minimaux/ partagés

e. a.

Les indicateurs FAIR

Voir le guide FAIR de CAHIER V2 et leRDA :
https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines#comment-form

Interoperable

Interoperable	Les métadonnées utilisent une représentation des connaissances exprimée dans un format standardisé	Important
Interoperable	Les données utilisent une représentation des connaissances exprimée dans un format standardisé	Important
Interoperable	Les métadonnées utilisent une représentation des connaissances compréhensible par la machine	Important
Interoperable	Les données utilisent une représentation des connaissances compréhensible par la machine	Important
Interoperable	Les métadonnées utilisent des vocabulaires conformes aux principes FAIR	Important
Interoperable	Les données utilisent des vocabulaires conformes aux principes FAIR	Utile
Interoperable	Les métadonnées comprennent des références à d'autres métadonnées	Important
Interoperable	Les données comprennent des références à d'autres données	Utile
Interoperable	Les métadonnées comprennent des références à d'autres données	Utile
Interoperable	Les données comprennent des références qualifiées à d'autres données	Utile
Interoperable	Les métadonnées comprennent des références qualifiées à d'autres métadonnées	Important
Interoperable	Les métadonnées comprennent des références qualifiées à d'autres données	Utile

Qu’impliquent les principes FAIR? Réflexions à partir du cas des corpus d’auteurs du Consortium CAHIER

2. “Accessibilité”

L'accessibilité est un .... rêve

https://vidensportal.deic.dk/sites/default/files/uploads/A%20FAIRy%20tale%20book%20digital_ny.pdf

Freins:

- FAIRE un site web n'est pas rendre ses données FAIR : l'imaginaire du "site web"
- manque d'imagination au sujet des usages/ requêtes potentielles
- restrictions imposées par le droit d'auteur
- réflexes communautaires
- activité de recherche non reconnue…

Les indicateurs FAIR

Voir le guide FAIR de CAHIER V2 et les recommandations de la Research Data Alliance
RDA : https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines#comment-form

Accessible

Accessible	A1	RDA-A1-01M	Metadata contains information to enable the user to get access to the (	Les métadonnées contiennent des informations permettant à l'utilisateur d'accéder aux données	Important
Accessible	A1	RDA-A1-02M	Metadata can be accessed manually (i.e. with human intervention)	Les métadonnées peuvent être consultées manuellement (c'est-à-dire avec une intervention humaine)	Essentiel
Accessible	A1	RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)	Les données peuvent être consultées manuellement (c'est-à-dire avec une intervention humaine)	Essentiel
Accessible	A1	RDA-A1-03M	Metadata identifier resolves to a metadata record	L'identificateur de métadonnées se transforme en un enregistrement de métadonnées	Essentiel
Accessible	A1	RDA-A1-03D	Data identifier resolves to a digital object	L'identificateur de données est un objet numérique	Essentiel
Accessible	A1	RDA-A1-04M	Metadata is accessed through standardised protocol	Les métadonnées sont accessibles par le biais d'un protocole standardisé	Essentiel
Accessible	A1	RDA-A1-04D	Data is accessible through standardised protocol	Les données sont accessibles via un protocole standardisé	Essentiel
Accessible	A1	RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)	Les données peuvent être consultées automatiquement (c'est-à-dire par un programme informatique)	Important
Accessible	A1.1	RDA-A1.1-01M	Metadata is accessible through a free access protocol	Les métadonnées sont accessibles via un protocole d'accès libre	Essentiel
Accessible	A1.1	RDA-A1.1-01D	Data is accessible through a free access protocol	Les données sont accessibles via un protocole d'accès libre	Important
Accessible	A1.2	RDA-A1.2-02D	Data is accessible through an access protocol that supports authentica	Les données sont accessibles via un protocole d'accès qui prend en charge l'authentification et l'autorisation	Utile
Accessible	A2	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer avail	Les métadonnées resteront disponibles lorsque les données ne sont plus disponibles	Essentiel

Qu'impliquent les principes FAIR? Réflexions à partir du cas des corpus d'auteurs du Consortium CAHIER

1. F/Facile à (re)trouver

Les problèmes de la “trouvabilité”

Quand?

⇒ réponses de CAHIER (les états du projet)

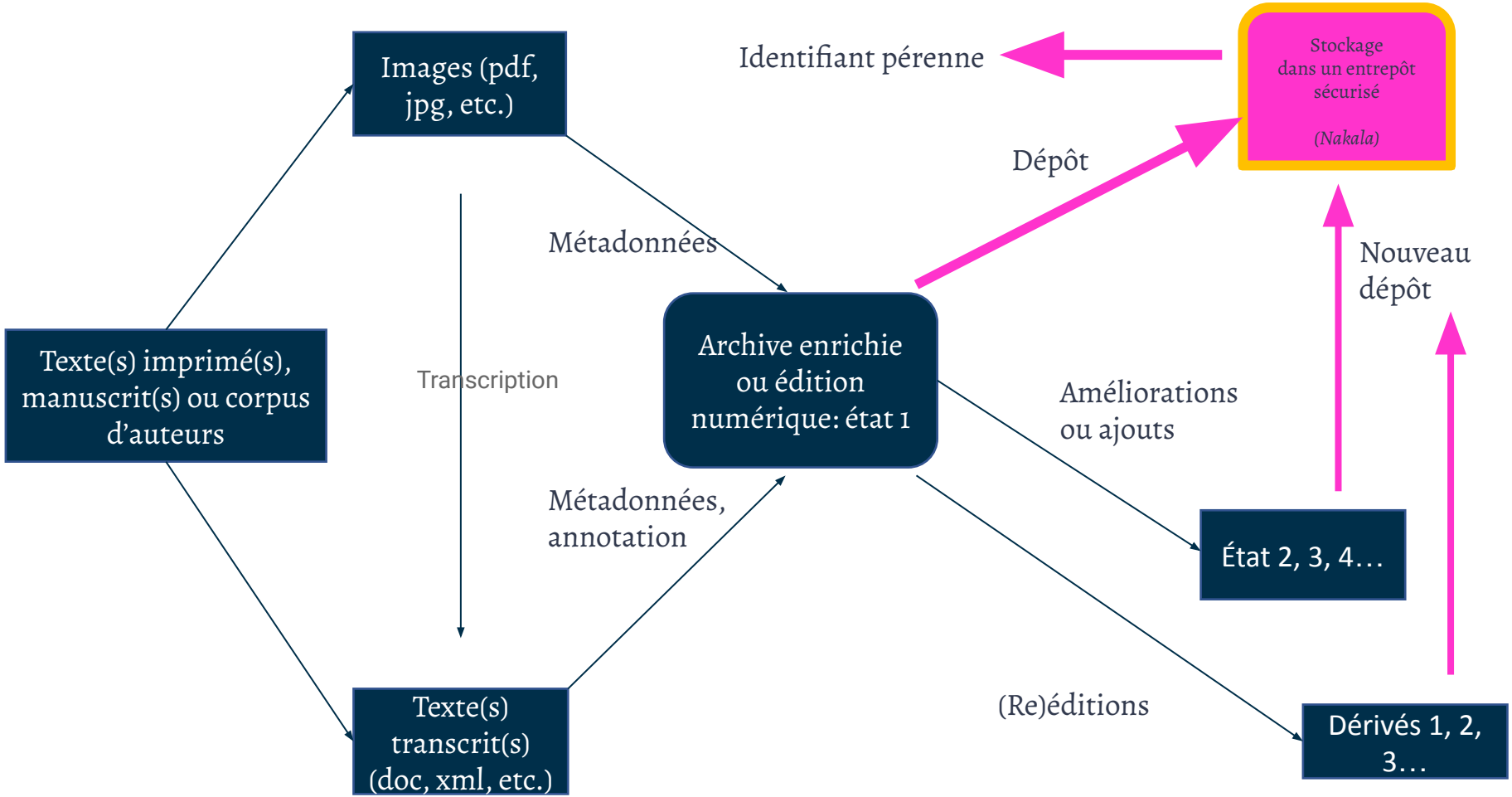
Comment?

⇒ grâce à l’infrastructure HUMA-NUM (mynakala)

FREINS

⇒ Activité de recherche non reconnue…

Vie d'un projet numérique sur corpus d'auteur(s)



Réutiliser différents “états” du projet

Exemple : définition d’un “Etat 1” du projet d’édition

Élément	État qualitatif	État quantitatif
Images de la source		Au moins 30% des images déjà acquises
Texte	Protocole de transcription défini	Au moins 30000 mots transcrits selon ce protocole
Métadonnées	Jeu de métadonnées défini (incluant les recommandations CAHIER)	Les 30% d’images acquises et/ ou les 30000 mots transcrits équipés de métadonnées selon le schéma défini
Annotations	Schéma d’annotation stabilisé dans une V1	Les 30000 mots transcrits annotés selon le schéma d’annotation stabilisé

Les indicateurs FAIR

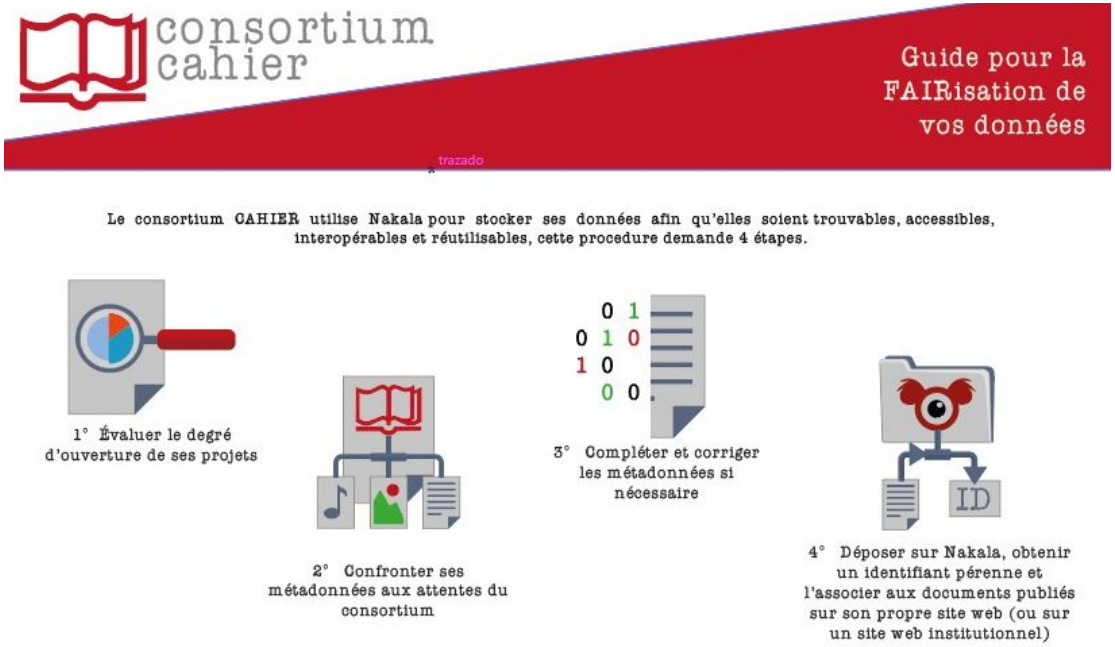
Voir le guide FAIR de CAHIER V2 et les recommandations de la Research Data Alliance
RDA : https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines#comment-form

FAIR data maturity model Indicators

2020-04-08/version 0.05

Principle		Degré de priorité
Findable	Les métadonnées sont identifiées par un identifiant permanent	Essentiel
Findable	Les données sont identifiées par un identifiant permanent	Essentiel
Findable	Les métadonnées sont identifiées par un identifiant unique au niveau mondial	Essentiel
Findable	Les données sont identifiées par un identifiant unique au niveau mondial	Essentiel
Findable	Des métadonnées riches sont fournies pour permettre la découverte	Essentiel
Findable	Les métadonnées comprennent l'identifiant des données	Essentiel
Findable	Les métadonnées sont proposées de telle sorte qu'elles peuvent être diffusées, récoltées et indexées	Essentiel

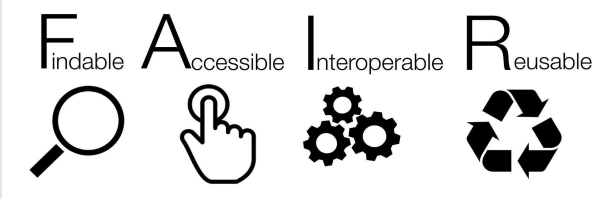
Développer les services adéquats



mynakala

Conclusion

Pour quoi FAIR?



Bilan provisoire

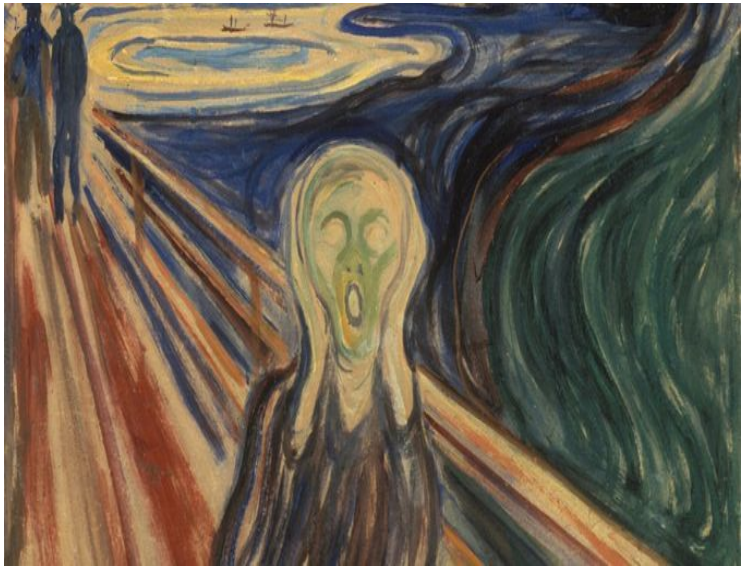
Potentiel CAHIER…

- 60 projets
- approx. 327.000 fichiers
- cca. 500.000 images

consortium cahier

… données FAIR

- ~20 projets
- approx. 13200 fichiers
- cca. 100.000 images



Bilan provisoire

Comment FAIR quand, fondamentalement, cette activité de recherche n'est pas reconnue ?

- ⇒ section CNU?

- ⇒ lobbying?

- ⇒ label de doctorat?

Merci pour votre attention

Plus d'informations sur
CAHIER : https://cahier.hypotheses.org
HUMA-NUM : https://www.huma-num.fr/