Vive les métadonnées ! Les bases de Data Documentation Initiative (DDI)

Alina Danciu, Christophe Dzikowski, Simon Hodson, Hilde Orten, Nicolas Sauger

Vive les métadonnées !
Les bases de Data Documentation Initiative (DDI)

Lundi 13 Juin 2022

Série de webinaires CODATA-Alliance DDI

Sommaire

1) Introduction : Data Documentation Initiative (DDI), un standard FAIR
2) Valoriser les données avec DDI
3) DDI-Codebook
4) DDI-Lifecycle
5) DDI-CDI
6) Conclusion

Introduction

Les principes FAIR

- **F**acilement trouvable
- **A**ccessible
- **I**nteropérable
- **R**éutilisable

“Les principes FAIR sont un ensemble de principes directeurs pour gérer les données de la recherche visant à les rendre faciles à trouver, accessibles, interopérables et réutilisables par [les chercheurs et leurs machines].”
https://www.ouvrirlascience.fr/fair-principles/

Les principes FAIR quasi omniprésents…

- Recommandation de l'UNESCO sur une science ouverte: la gestion des données de recherche devrait suivre les principes FAIR (et CARE).
- Union Européenne Horizon Europe: les données de recherche devraient être FAIR.
- "FAIR est le langage commun du EOSC"
- Quasi impossible de trouver un projet qui ne s'appelle FAIR qqch… (FAIRsFAIR, FAIR Impact, FAIR Pillar, WorldFAIR, FAIRsFAIRsFAIR…)
- Plus qu'une tendance?
- Oui, parce que ce sont de bons principes pour la gestion et la réutilisation des données avec les technologies informatiques.
- Rapport PWC: coût d'opportunité minimum de 10.2 Bn euros du fait de ne pas avoir les données FAIR: https://data.europa.eu/doi/10.2777/02999

Les principes FAIR

- **F**acilement trouvable
  - F1 Les données et les **métadonnées** sont identifiées par un identifiant global unique et pérenne.
  - F2 Les **métadonnées** décrivant les données sont riches.
  - F3 Les données et les **métadonnées** sont enregistrées et indexées dans un dispositif permettant de les rechercher.
  - F4 Les **métadonnées** spécifient l'identifiant de la donnée.
- **A**ccessible
  - A1 Les données et les **métadonnées** sont accessibles par leur identifiant via un protocole de communication standardisé.
    - A1.1 Le protocole utilisé est ouvert, libre et peut être implémenté de manière universelle.
    - A1.2 Le protocole utilisé permet l'authentification et l'autorisation si besoin.
  - A2 Les **métadonnées** sont accessibles même quand les données ne le sont plus.
- **I**nteropérable
  - I1 Les données et les **métadonnées** utilisent un langage formel, accessible, partagé et largement applicable pour la représentation des connaissances.
  - I2 Les données et les **métadonnées** utilisent des vocabulaires qui respectent les principes FAIR.
  - I3 Les données et les **métadonnées** incluent des liens vers d'autres (méta)données.
- **R**éutilisable
  - …

https://datapartage.inrae.fr/Produire-des-donnees-FAIR

DATA DOCUMENTATION INITIATIVE

Valoriser les données avec DDI

Qu’est-ce que DDI ?

- DDI = Data Documentation Initiative
- Un standard international de métadonnées
  - Utilisé principalement en sciences sociales et comportementales, économie, santé
  - Un standard ouvert créé pour le partage et la réutilisation des données
- Une structure pour décrire les données et les informations en lien avec celles-ci
- Décrit des données d’enquête et des données provenant d’autres méthodes de collectes basés sur de l’observation
  - est en train d’évoluer pour couvrir de nouveaux types de données, ainsi que des données provenant d’autres domaines que les SHS ou la santé

Pour en savoir plus

- Site web DDI
  - http://www.ddialliance.org/
  - Un site de référence, en anglais
    - Apprendre (ressources auto-formation)
    - Produits
    - Événements
    - Publications
    - Collaborations
    - À propos de

Pourquoi utiliser DDI?

*DDI encourage la **description complète des données pour leur exploration et analyse et soutient le partage des données**. DDI étant une norme structurée, elle facilite l'interopérabilité des systèmes et peut être utilisée pour piloter ceux-ci. Une autre caractéristique de DDI est l'accent que la norme met sur la **réutilisation des métadonnées ; « saisir une fois, utiliser souvent »** signifie que vous pouvez ré-utiliser les métadonnées tout au long de leur cycle de vie pour éviter une duplication coûteuse des efforts.*

Source: http://www.ddialliance.org/training/why-use-ddi

DATA DOCUMENTATION INITIATIVE

Atouts de DDI

- Interopérabilité
- Contenu riche
  - Granularité fine
- Accroît la « visibilité » des données
  - Précision dans la recherche
- Communauté internationale

Challenges de DDI

- Complexité
- Niveau d'adhésion des chercheurs

Utilisateurs de DDI

Organisations

- Banque mondiale
- Université de Harvard
- Sciences Po
- Data Liberation Initiative (Statistics Canada)
- US Census Bureau/MPC
- ICPSR
- INSEE France (et d'autres instituts statistiques à travers le monde)
- Norwegian Agency for Shared Services in Education and Research (Sikt)
- UK Data Archive
- Zentralarchiv für Empirische Sozialforschung (GESIS)
- RODA (Romanian Social Data Archive)
- …

Projets

- CESSDA Data Portal
- Australian Social Science Data Archive
- DAMES Project (UK)
- DataFirst (at University of Cape Town)
- Israel Social Science Data Center
- ODESI (Canada)
- Statistics New Zealand
- ResearchDataGouv.fr
- …

DATA DOCUMENTATION INITIATIVE

Utilisateurs de DDI

Source: https://ddialliance.org/community/join

Public cible DDI

- Ingénieurs données
- Administrateurs d’entrepôts/banques de données
- Financeurs
- Producteurs de données
- Chercheurs
- Développeurs

Commencer à utiliser DDI

- Déroutant au début
  - Le processus est décomposé en étapes
- Ressources utiles
  - DDI Alliance
  - http://www.ddialliance.org/training/getting-started
  - Collègues
- Listes d'utilisateurs DDI

Interprétation de DDI

- Écrite en format XML
- Besoin d'un outil pour l'interpréter
- Les plus connus :
  - Nesstar
  - Colectica
  - Dataverse

Pour en savoir plus : https://ddialliance.org/resources/tools

Interprétation de DDI (2)

- Écrite en XML
  - Le schéma XML est une manière de baliser le texte en fonction de son sens et non de son apparence
  - Définition
    - Des balises disponibles
    - L'ordre dans lequel les balises apparaissent
    - Si les balises sont obligatoires ou optionnelles
    - Si les balises se répètent ou non

Exemples de balises DDI

<titl>**Canadian Community Health Survey, 2012: Annual Component** </titl>
<labl>**Questionnaire (.pdf)**</labl>

<dataDscr><notes>**The variables in this study are identical to earlier waves.** </notes></dataDscr>

<titl>**Canadian Gallup Poll, May 2000**</titl>
<dataChck>**Quality checks were performed by Carleton University Data Centre.** </dataChck>

<titl>**Survey of Household Spending, 2001 [Canada]**</titl>
<varQnty>**255**</varQnty>

<titl>**Canadian Gallup Poll, May 1949, #186**</titl>
<copyright>**Copyright Gallup Canada Inc., 1950**</copyright>

En résumé

- DDI est un standard de métadonnées puissant à condition que
  - *l'information correcte soit rentrée dans les champs corrects*
- Besoin d'outils pour éditer les métadonnées et les publier

Produits DDI

- Le standard DDI s'est développé à travers le temps
  - Continue à se développer en fonction des besoins des utilisateurs
- Trois produits principaux existent actuellement
  - DDI Codebook
  - DDI Lifecycle
  - DDI CDI (à venir)
- Chacun a été développé pour un but différent

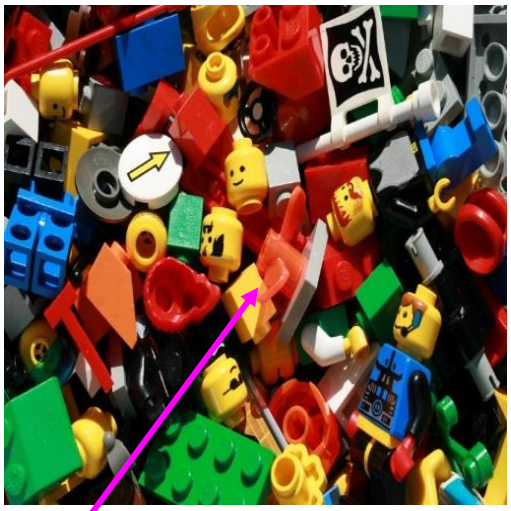
DDI Codebook

Besoins en termes de métadonnées

- Quelles sont les informations qui sont indispensables pour l'utilisation d'un fichier de données ?
- Quelles sont les informations supplémentaires que vous aimeriez avoir, au cas où elles seraient disponibles ?

DDI Codebook

Métadonnées non structurées



Data manager perdu

DDI Codebook structures

Métadonnées structurées avec le standard DDI Codebook

DDI Codebook

- Une structure qui facilite la production de *codebook* qui peuvent être lus par l'homme et la machine
- Utilisé pour documenter des enquêtes déjà produites

- Construit pour produire un codebook physique
  - c'est-à-dire, décrire un jeu de données, une seule étude ou une seule édition ou vague d'une étude répétée
- La version la plus récente est 2.5

DDI Codebook (2)

- Assez “simple”
- Sections
  - *Description du document*
  - *Description de l’étude*
  - *Description des fichiers de données*
  - *Description des variables*
  - *Autres documents liés à l’étude*

Colectica for Excel

- Importer des fichiers de données à partir de SPSS, Stata ou SAS pour les documenter dans Excel
- Exporter les métadonnées au format DDI

Nesstar Publisher

- Créer et éditer des métadonnées au format DDI
- Extraire les métadonnées des logiciels statistiques
- Valider les métadonnées et les variables
- Documentation au niveau de la variable
- Création de codebook

Dataverse

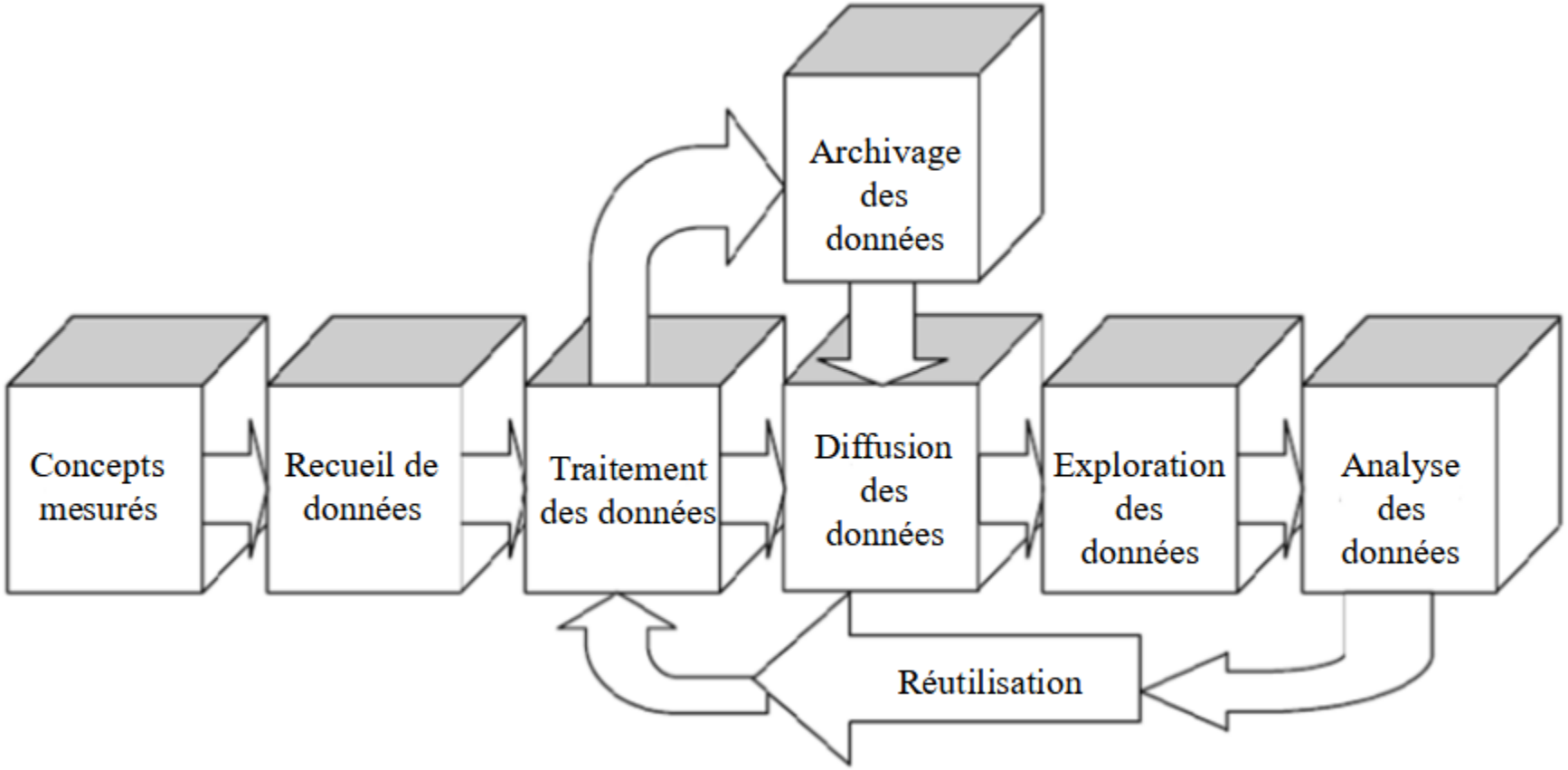
- **Édition de métadonnées**
  - Créer et éditer des métadonnées au format DDI (description de l'enquête)
  - Utilisation de vocabulaires contrôlés
- **Entrepôt de données et métadonnées**
  - Moteur de recherche puissant
  - Plug-ins existent pour la documentation et l'exploration des variables (ex : Data Explorer)

DDI Lifecycle

Les spécifications DDI-L

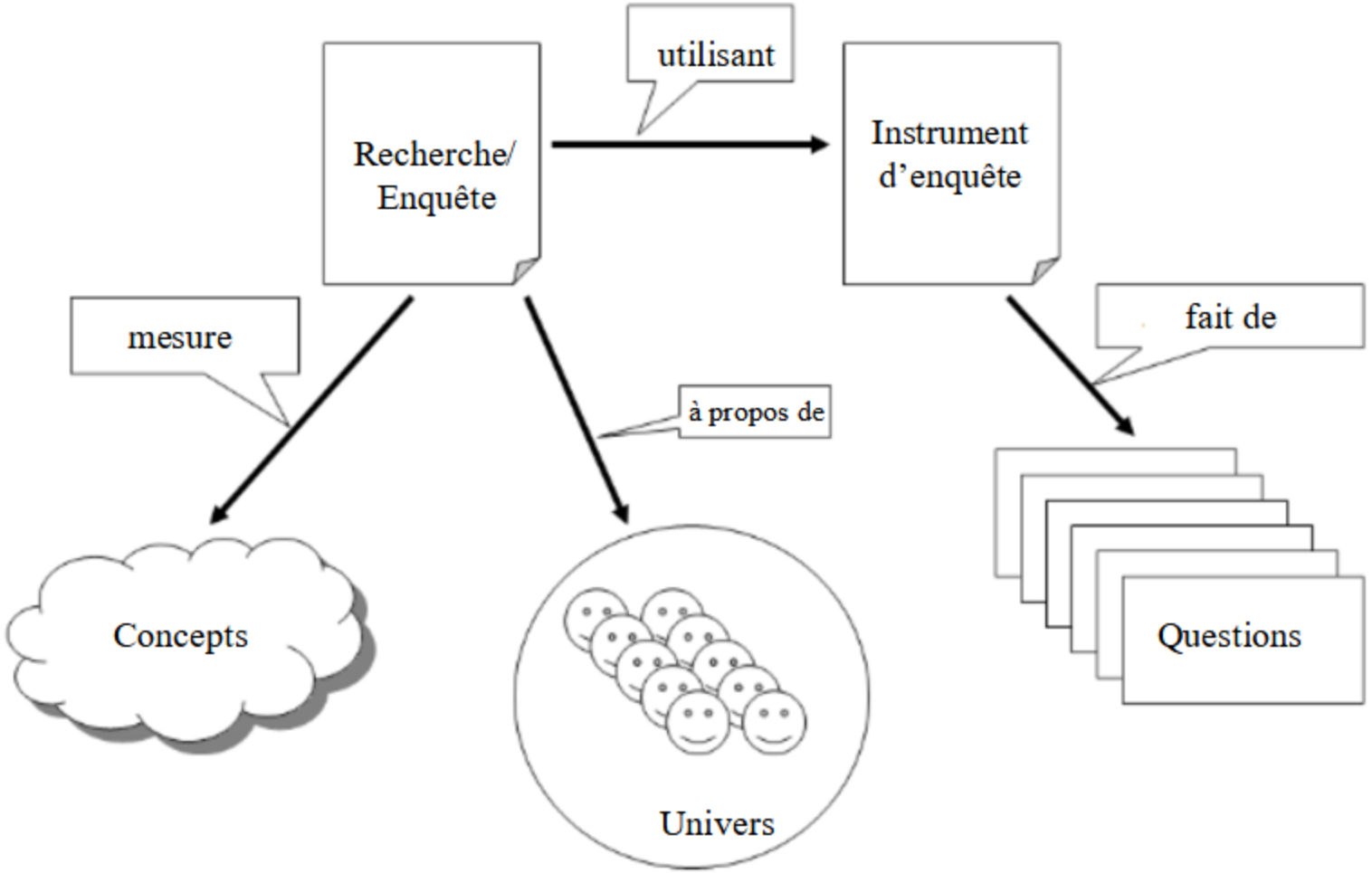
- Conçu :
  - *Pour répondre à un large éventail d'exigences typiques de la gestion et de l'utilisation des métadonnées*
  - *Pour prendre en charge tous les types de réutilisation et pour fonctionner avec des approches par registre et par référentiel*

DDI-L : le cycle de vie des données

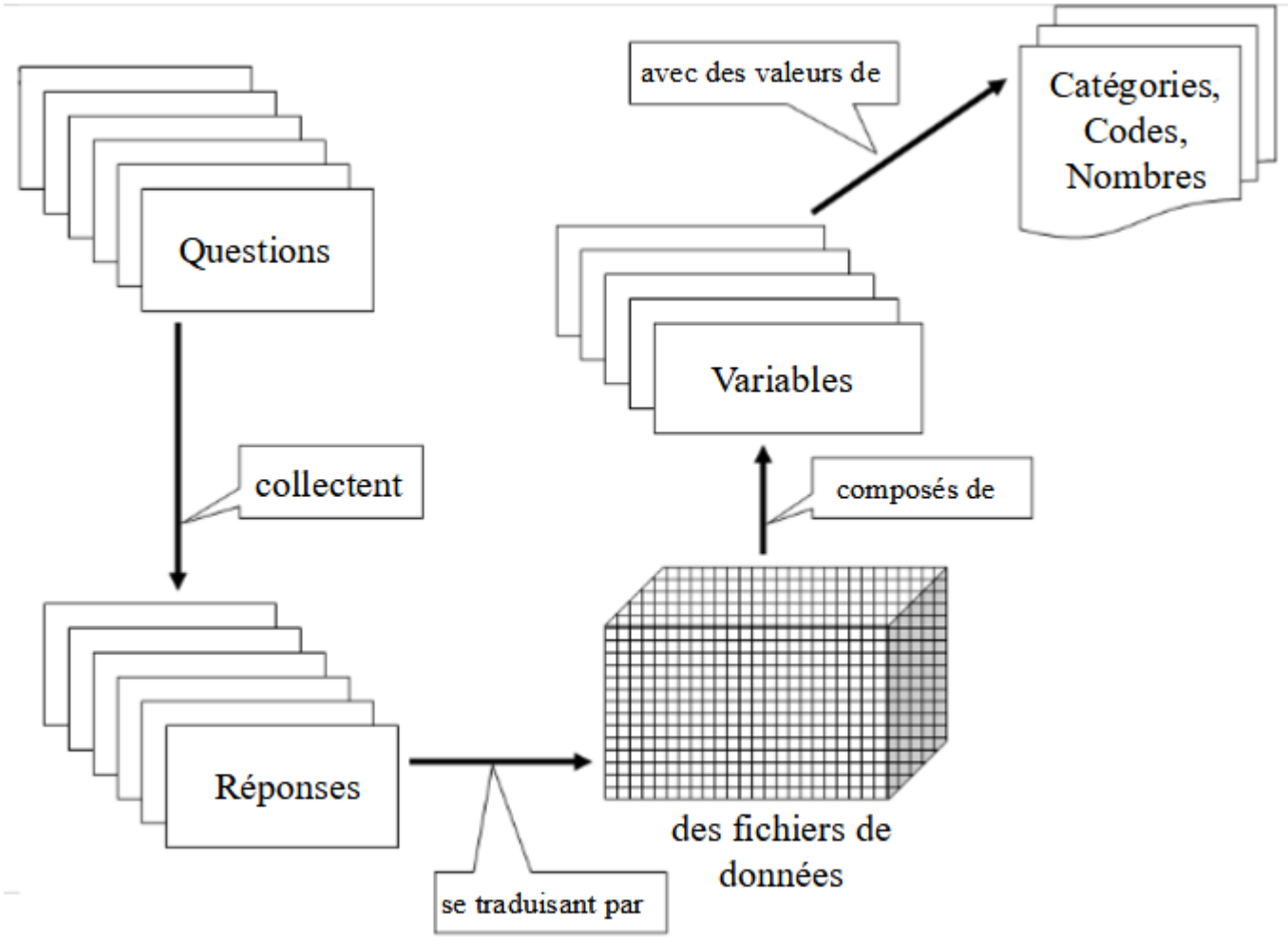


DATA DOCUMENTATION INITIATIVE

DDI-L en 60 secondes



DDI-L en 60 secondes

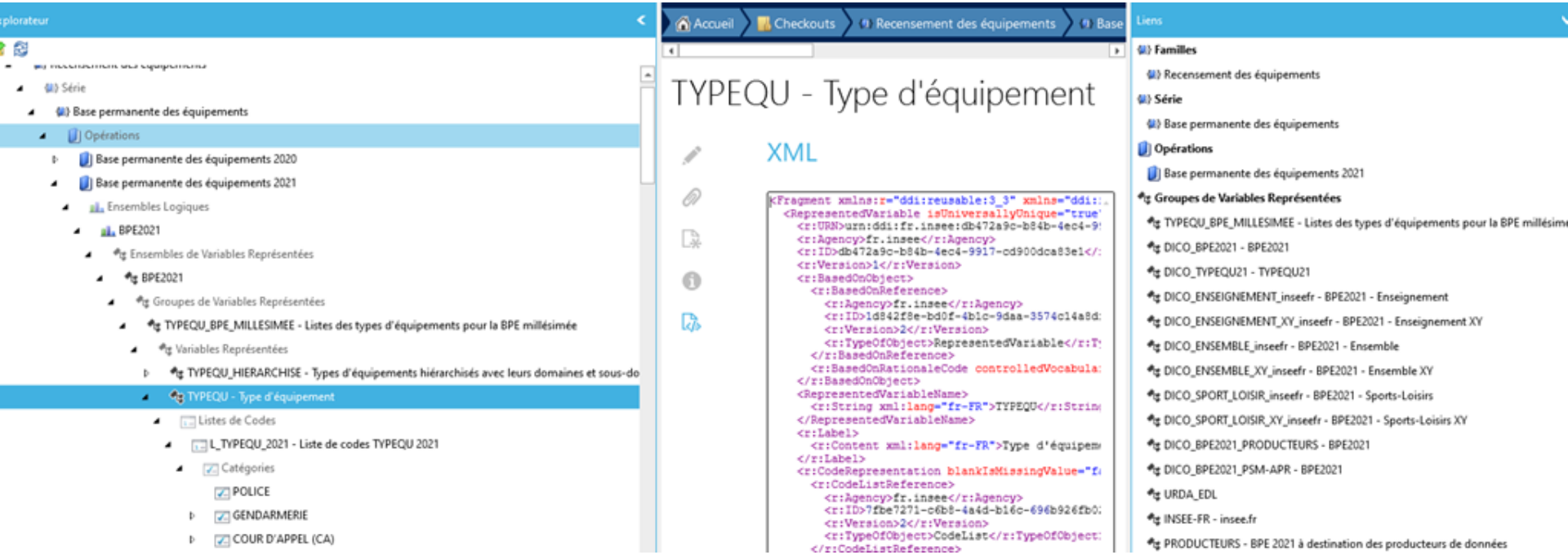


DATA DOCUMENTATION INITIATIVE

DDI-L : spécifications et métadonnées

DDI-L : réutilisation des métadonnées

DDI-L en action : Base permanente des équipements



DATA DOCUMENTATION INITIATIVE

Autres standards DDI

XKOS pour décrire les classifications
https://rdf-vocabulary.ddialliance.org/xkos.html

DISCO pour faciliter la découverte des jeux de données
https://rdf-vocabulary.ddialliance.org/discovery.html

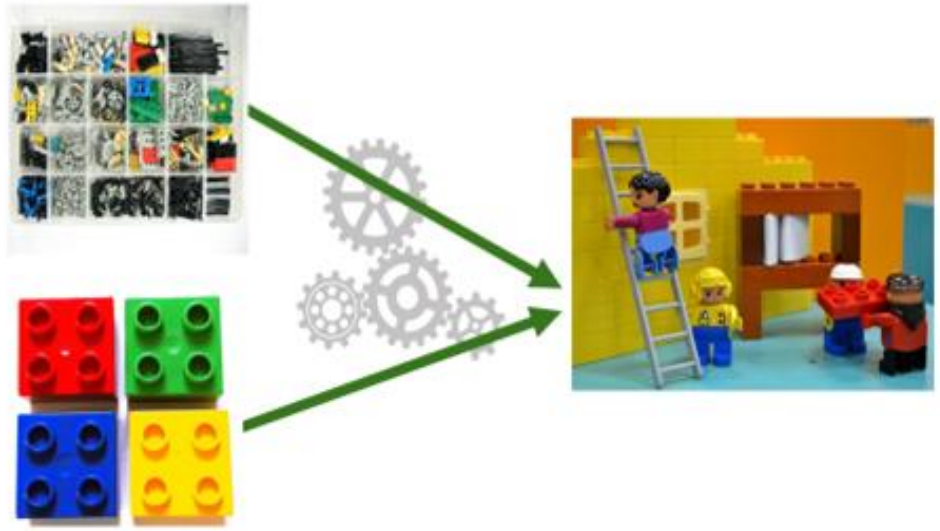
SDTL pour représenter les commandes de transformation de données structurées
https://ddialliance.org/products/sdtl/1.0

DATA DOCUMENTATION INITIATIVE

DDI - Cross-Domain Integration (DDI-CDI)

DDI-CDI

- Décrit différentes structures de données
- Conçu pour fonctionner avec d'autres standards et décrire des données de domaines différents
- Capture la provenance des données



DDI-CDI

- Complémentaire à DDI-C et DDI-L
- Permet de décrire différents types de structuration de données
- Restitue la provenance de données hétérogènes
- Destiné à être utilisé dans un large éventail de domaines
  - Différents types de données et modèles
  - Terminologie plus abstraite/générales

DDI-CDI

- Nouveau type de spécification
- À utiliser avec de nombreux autres standards
- À utiliser comme un “compagnon” de DDI-C, DDI-L et autres standards (DCAT, PROV, etc.)
- Ne les remplace pas
- Ajoute un cadre pour décrire des types de données différents
- Élargit la capacité de décrire les processus et la provenance

DATA DOCUMENTATION INITIATIVE

DDI-CDI et FAIR

- Beaucoup d'activités dans le monde des données pour rendre les données scientifiques plus **F**acilement trouvable et **A**ccessible
- Jusqu'à présent, pas autant d'intérêt pour les points **I**nteropérable et **R**éutilisable
- DDI-CDI se concentre sur ces aspects FAIR également
- Il s'avère aussi très utile pour l'exploration des données
- L'interopérabilité et la réutilisation des données dépendent des métadonnées
- Historiquement, ces aspects de la gestion des données sont coûteux et n'ont pas été pleinement encouragés par les financeurs – FAIR change cette situation
- L'accent mis aujourd'hui sur les données FAIR *exige* que nous fassions plus !

DDI-CDI , un nouveau type de produit

DDI Codebook et DDI Lifecycle sont des spécifications de métadonnées pour les sciences sociales, comportementales et économiques

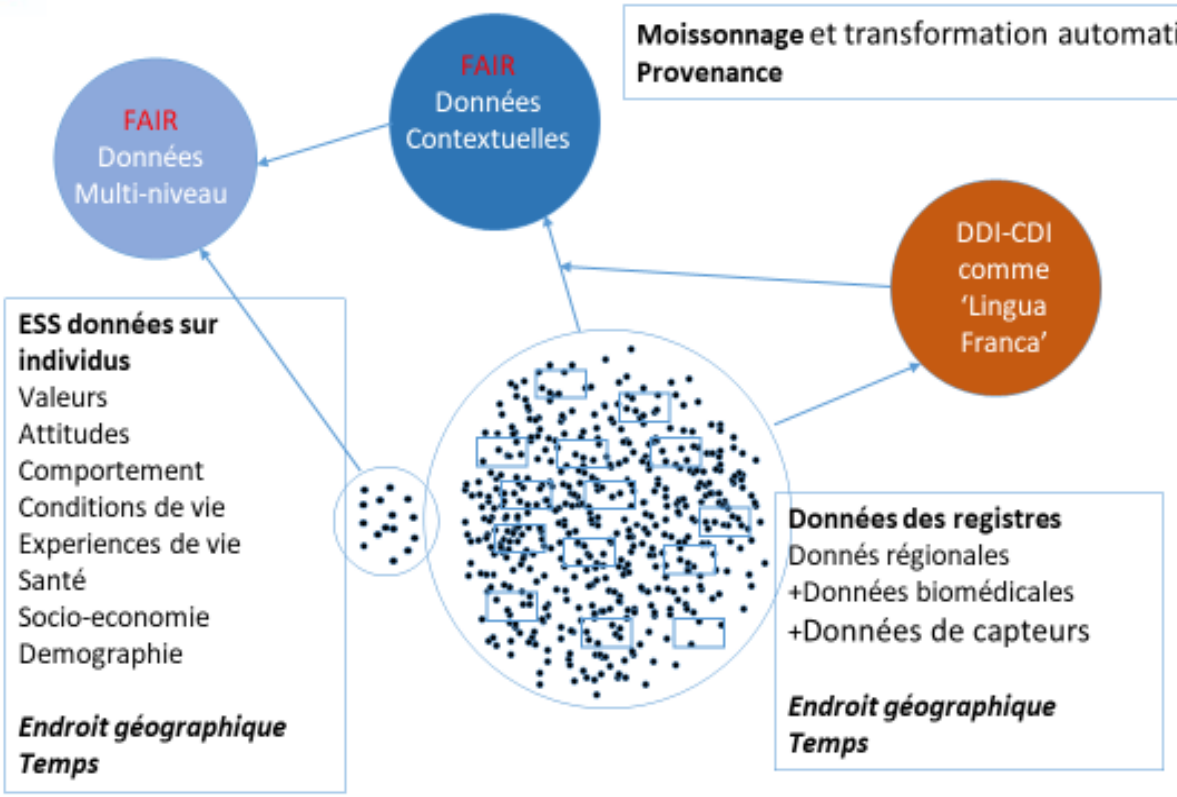
- Ils sont suffisamment génériques pour être utilisés dans des domaines similaires (statistiques officielles, santé publique…)
- Ils utilisent des termes et modèles spécifiques au domaine des SHS
- DDI-CDI est différent : il est destiné à être utilisé dans un plus grand nombre de domaines
- Différents types de données/modèles
- Terminologie plus abstraite/générale

DDI-CDI est un nouveau type de spécification, destinée à être utilisée avec de nombreuses autres normes, dans le cadre des SHS et en dehors

DATA DOCUMENTATION INITIATIVE

DDI-CDI : Étude de cas

Enquête sociale européenne
Application multi-niveaux -Idées pour des améliorations futures



Conclusion

Conclusion

- Ne pas négliger l'importance des métadonnées

  Coûts et bénéfices*
- Prévoir le plus tôt possible la gestion des métadonnées dans le cycle de vie des données
- Respecter les standards comme DDI et utiliser des vocabulaires contrôlés
- Adapter sa stratégie au type de projet (données collectées, données hétérogènes…) et aux ressources (humaines, financières)

Yes, we can do it!

*Cf. *Cost of not having FAIR research metadata* *doi: 10.2777/02999*

Stratégies

Le scénario A

- Création de métadonnées standard, importation de métadonnées à partir de fichiers de données, prise en charge des systèmes d'exploration de données (portails, catalogues)

Le scénario B

- Centralisation de la gestion des métadonnées : *Single Source of Truth* (entrepôt de données)
- Usage optimal des outils et des processus de documentation existants dans le 1er niveau pour la création, l'édition des métadonnées
- Gestion des versions possibles

Le scénario C

- La collecte de métadonnées tout au long du cycle de vie des données dès la conception du projet (enquêtes, flux de données, données agrégées...)
- La gestion et la documentation des données selon les producteurs et à travers le temps

Outils DDI : sommaire

- Liste outils recommandés par l'Alliance DDI
- Références complémentaires :
  - European DDI Conference 2021. Training FAIR, Track 3: DDI Tools and Services
- Outils les plus utilisés :
  - Colectica (pour Excel, Designer, Portal…)
  - Dataverse (édition + entrepôt)
  - Nesstar (édition)

Questions ?

Remerciements

Contenu traduit et adapté à partir de diapositives rédigées par le groupe de travail *DDI Training Group*. Basé sur du contenu développé lors du workshop DDI *Train-the-Trainers* qui a eu lieu à Dagstuhl en 2018.

Traduction : Alina Danciu, Christophe Dzikowski
Révision de la traduction : François Loretan, Benjamin Peuch

La partie DDI-C et DDI-CDI est une adaptation en français de Jane Fry, Arofan Gregory, Hilde Orten. (2020, December 4) The DDI, FAIR Convergence Symposium 2020
La partie sur DDI-L et la diapo Stratégies ont été reprises à partir de Alina Danciu, Alexandre Mairot. (2019, March 14). Data Documentation Initiative (DDI), un standard de documentation des données. Webinaires Tuto Mate. https://doi.org/10.5281/zenodo.6590698

DATA DOCUMENTATION INITIATIVE

Remerciements

DDI Train-the-Trainers Dagstuhl workshop 2018 participants

Alina Danciu
Guillaume Duffes
Adrian Dușa
Lauren Eickhorst
Dan Gillman
Arofan Gregory
Taras Günther
Lea Sztuk Haahr
Sanda Ionescu
Jon Johnson
Chifundo Kanjala
Kaia Kulla
Amber Leahey
Alexandre Mairot
Johan Fihn Marberg
Hayley Mills
Olof Olofsson
Hilde Orten
Anja Perry
Dan Smith
Wendy Thomas
Joachim Wackerow
Knut Wenzig

DATA DOCUMENTATION INITIATIVE

Références

- DDI Alliance. *Data Documentation Initiative*. http://www.ddi-alliance.org/
- Fry, J., Cooper, A., Mowers, S., & Carrington, C. (2019). "Best Practices Document: based on DDI 2.x, version 3.1". https://bit.ly/3mhLmmH
- Jacobs, J. (2006). "Evolution of Data Documentation". Workshop "A Gentle Introduction to DDI: What's in it for Me?" presented at IASSIST 2006.
- Orten, H., Beuster, B., & Jääskelainen, T. (2019). «What can DDI do for you? AN introduction to the DDI. Presented at EDDI 2019. DOI: 10.5281/zenodo.3597192
- Perry, A. & Fry, J. "Introduction to DDI: Basic Concepts and How to Develop Skills for Training Researchers" IASSIST 2019.
- Schloss Dagstuhl, October 2014. "DDI Basics". https://bit.ly/2ZkdoTu
- Vardigan, M. & Wackerow, J. (2013). DDI – A metadata standard for the community. Paper presented at the North American Data Documentation Initiative Conference (NADDI) 2013. https://bit.ly/2J7RDTQ

Picture credits

	**Steine unsortiert**	woodleywonderworks, https://www.flickr.com/photos/wwworks/2472232245 (CC-BY)
	Steine sortiert	Windell Oskay, https://www.flickr.com/photos/17425845@N00/2156888497 (CC-BY)
	Bagger (3516880947_0f44a89c1c_z.jpg)	Stephen Edmonds: https://www.flickr.com/photos/popcorncx/3516880947/ (CC-BY-SA)
	Bagger (lego-717196_960_720.png)	Desktopnexus.com: https://abstract.desktopnexus.com/wallpaper/2425074/
	Bagger (3514881626_be3e87cc58_o.jpg)	Stephen Edmonds: https://www.flickr.com/photos/popcorncx/3514881626/ (CC-BY-SA)
	Bagger (4485538519_d4ef5e284b_o.jpg)	Stephen Edmonds: https://www.flickr.com/photos/popcorncx/4485538519/in/album-72157617802609935/ (CC-BY-SA)
	Health (health-2640352_640.jpg)	Pixabay.com: https://pixabay.com/photos/health-nurse-rescue-hospital-2640352/ Simplified Pixabay License