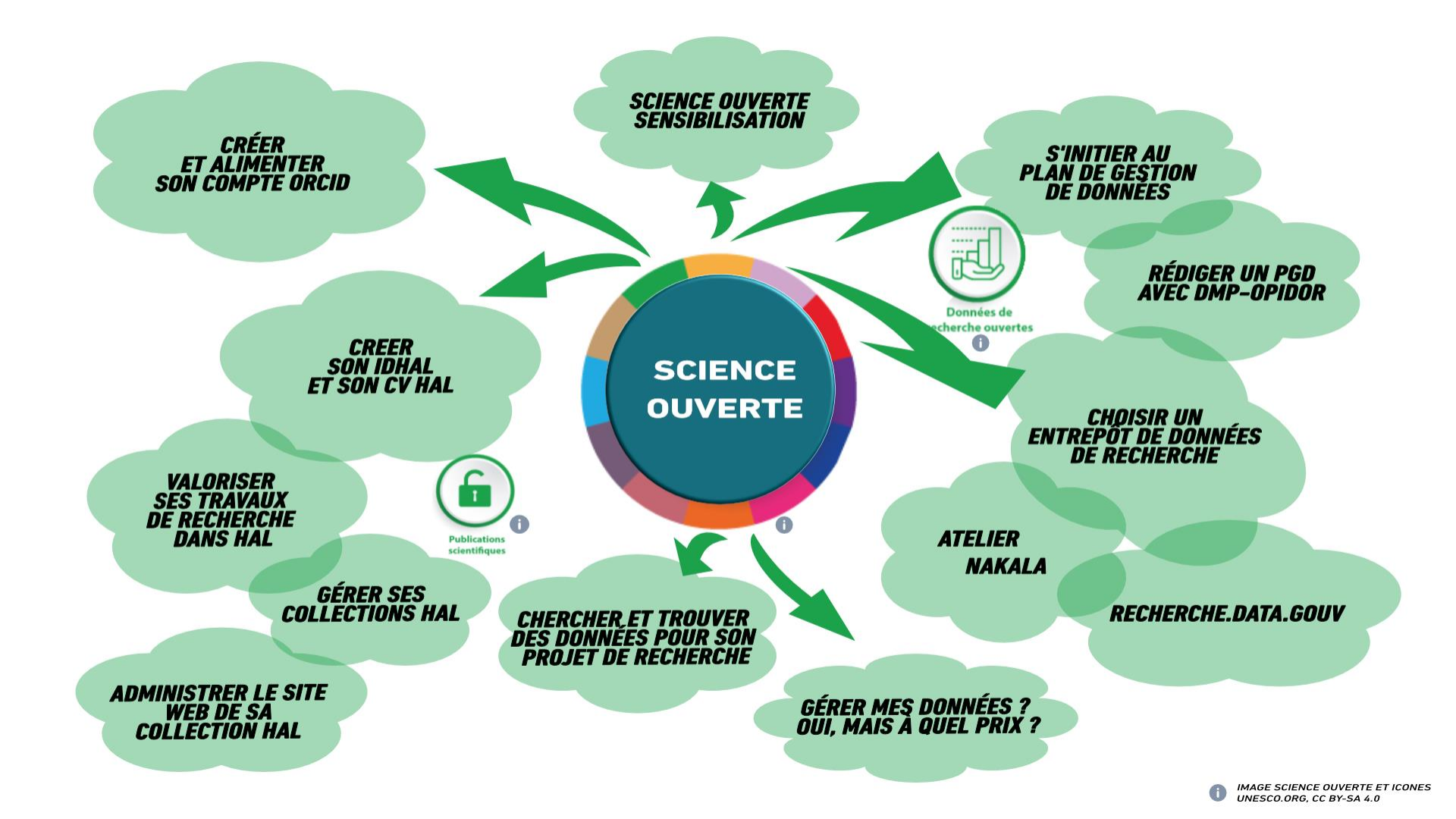
CAMPUS CONDORCET | HUMATHÈQUE

Chercher et trouver des données pour son projet de recherche

Emmanuel Collier

Science ouverte : offre de formations



Programme

Programme

- Pourquoi chercher des données et lesquelles ?
- Données de recherche et données produites par l’administration
- Chercher des données : aspects méthodologiques
- Trouver des données : atelier pratique

Pourquoi chercher des données et lesquelles?

Pourquoi chercher des données ?

- Commencer de nouvelles recherches
- Compléter d’autres travaux
- Vérifier des hypothèses de travail
- Vérifier les sources et résultats avancés dans des publications
- Comparer avec ses propres données
- Gagner du temps et des moyens : collecte, traitement, analyse, description des données, dépôt, *etc.*

Données de la recherche et données produites par l’administration

Données de la recherche

« Définition » de l’OCDE, 2007

« Les données de la recherche sont définies comme des **enregistrements factuels** (chiffres, textes, images et sons), qui sont utilisés comme **sources principales pour la recherche** scientifique et sont généralement reconnus par la communauté scientifique comme **nécessaires pour valider les résultats de recherche**. »

https://doi.org/10.1787/9789264034020-en-fr

Données de la recherche : aperçu typologique

- Notes de terrain
- Photos, enregistrements audio ou vidéo…
- Enquêtes, questionnaires, transcriptions d'entretiens, notes d'entretiens, réponses à des tests...
- Données textuelles structurées (tableaux...)
- Bases de données
- Références bibliographiques
- [...]

Données produites par l’administration

Les « données ouvertes » (*open data*)

« Données en accès libre et gratuit et facilement réutilisables par toutes et tous. Ces données sont produites par l’administration (ministères, collectivités locales, *etc.*) mais aussi par des acteurs privés ou encore des citoyens. »

Lexique de l’open data | *Guides de data.gouv.fr*, https://guides.data.gouv.fr/autres-ressources-utiles/lexique-de-lopen-data , 22 décembre 2023

Données produites par l'administration

Les « données publiques » : aperçu typologique

- Données géographiques (adresses, références cadastrales)
- Financières (budgets, commande publique, subventions, etc.),
- Environnementales (émissions, vente de produits, *etc.*), *etc*.

Lexique de l'open data | Guides de data.gouv.fr, https://guides.data.gouv.fr/autres-ressources-utiles/lexique-de-lopen-data , 22 décembre 2023

Jeu de données

- **Agrégation d’enregistrements de données** organisés pour former un **ensemble cohérent**. Ils sont formatés de telle sorte qu’ils soient communicables, interprétables et adaptés à un traitement informatisé.

- Un jeu de donnée est un **ensemble de ressources ou d’informations** (fichiers de données, fichiers d’explications, API etc.) **et de métadonnées** (description, producteur, date de publication, mots-clefs, couverture géographique temporelle etc.) sur un thème donné.

Rappel du cadre législatif

Loi pour une République numérique (7 octobre 2016)

- Ouverture des « données d'intérêt général »
  - Obligation pour les administrations publiques de publier sur Internet leurs bases de données sous réserve d'anonymisation et de protection du secret statistique, industriel et commercial
  - Facilitation de l'ouverture et de la réutilisation des données pour les chercheurs

Rappel du cadre législatif

Règlement général sur la protection des données (2016, applicable en 2018)

- Donnée personnelle : « Toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres »

- Renforcer et harmoniser en Europe la protection des données à caractère personnel

Chercher des données : aspects méthodologiques

Rappels des principes de la recherche documentaire

- **Cadrer le sujet : 3QOCP**

  Qui ? Quand ? Quoi ? Où ? Comment ? Pourquoi ?

- **Créer un corpus de mots clés pour interroger les outils de recherche**

Définir les types et structures des données

- Données qualitatives VS Données quantitatives
- Structure : formats des fichiers, taille, documentation détaillée, métadonnées

Définir les producteurs potentiels de données

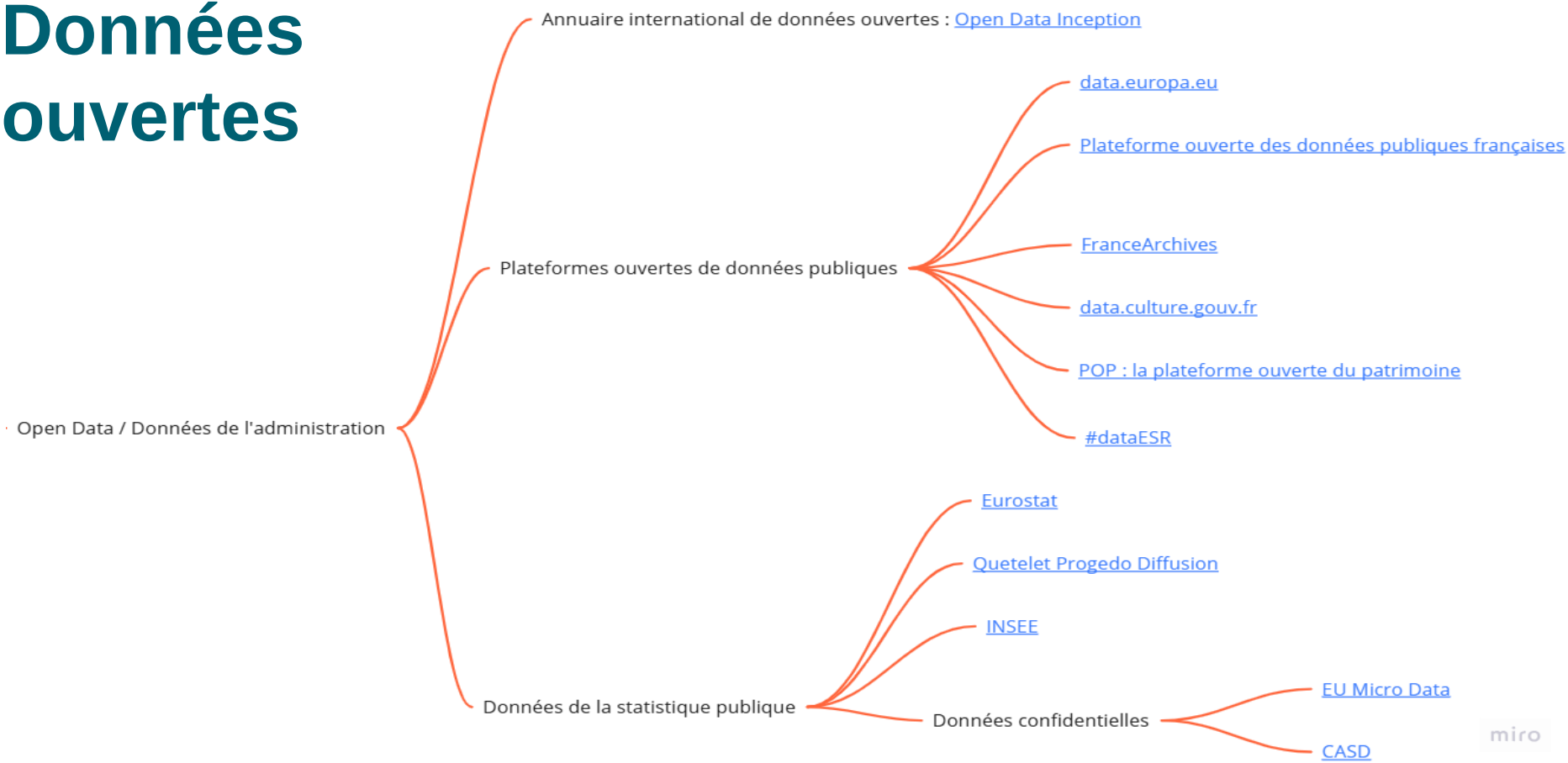
⇒ Équipe de recherche, institution publique produisant des statistiques, collectivité territoriale, services d'un ministère, *etc.*

Sélectionner des outils de recherche de données

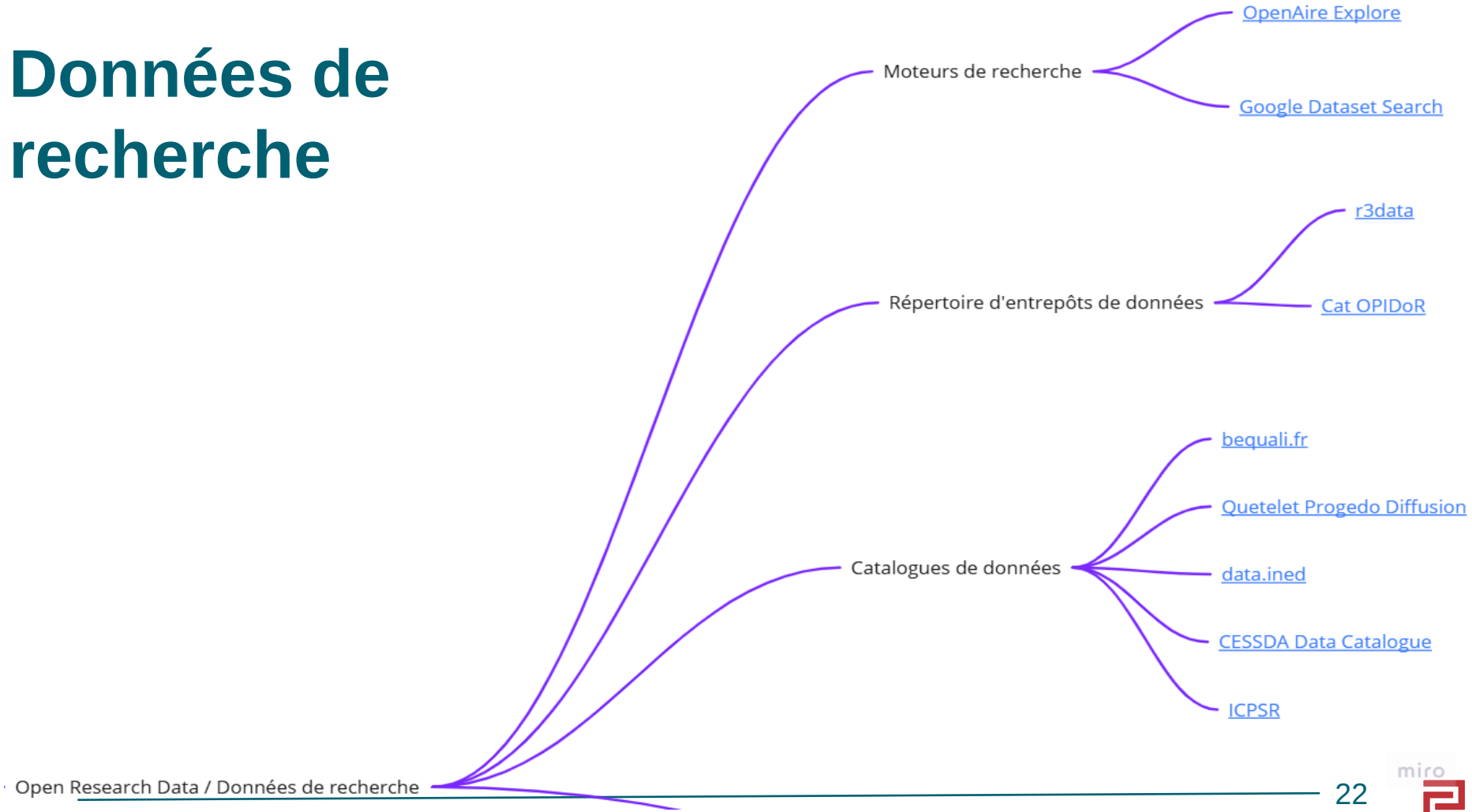
Les outils de recherche de données : panorama typologique

- **Carte mentale :** https://miro.com/app/board/uXjVMTYYk5g=/
  - Open Data / Données ouvertes
  - Données de recherche

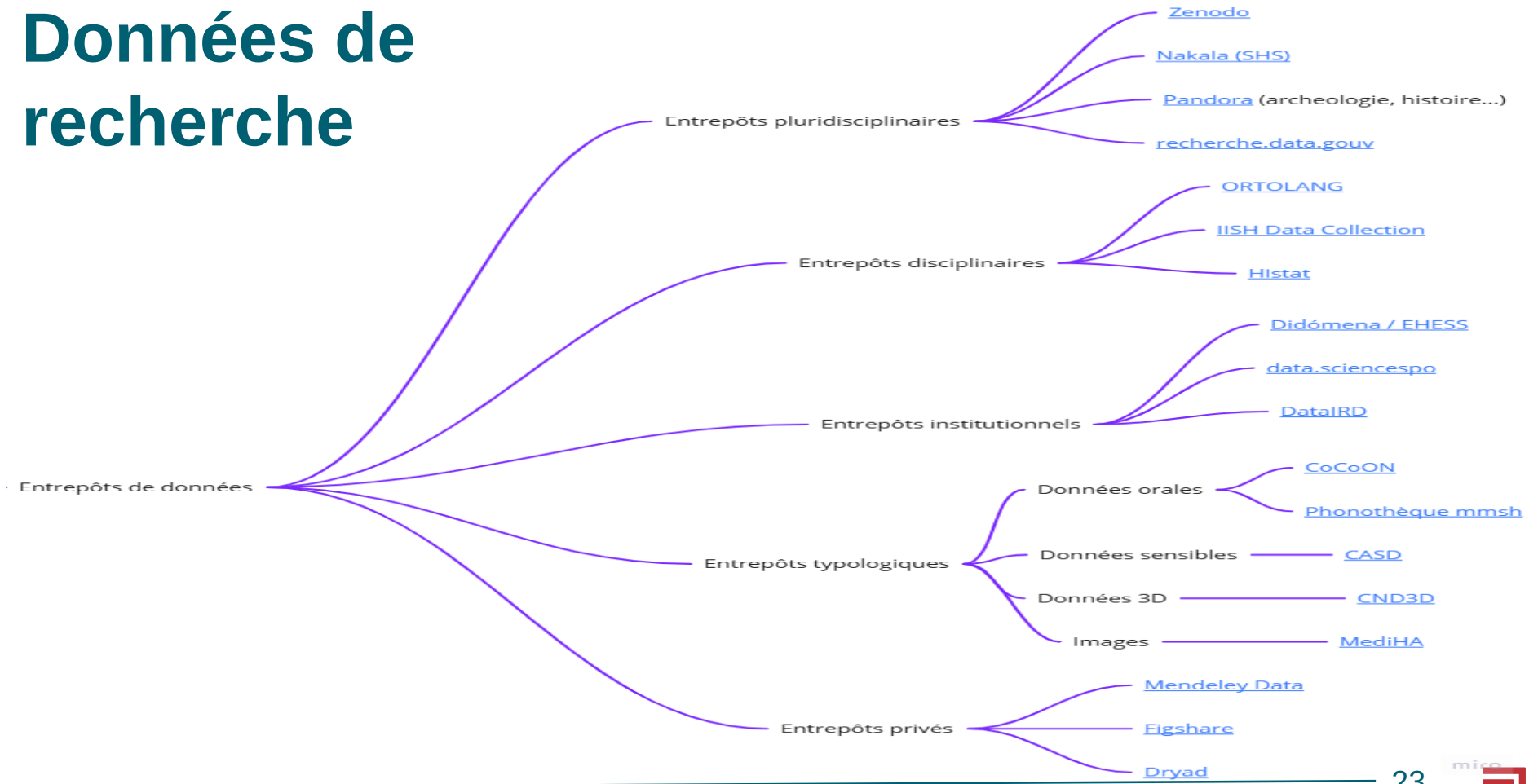
Données ouvertes



Données de recherche



Données de recherche



Conclusion

Points abordés

- Pourquoi chercher des données et lesquelles
  - Données de recherche et données produites par l’administration

- Chercher des données : aspects méthodologiques
  - Rappel : recherche documentaire
  - Types, structures et producteurs potentiels

Points abordés

- Panorama des outils de recherche de données
  - Carte mentale : typologie des outils de recherche
  - Grille d’analyse des outils de recherche

CAMPUS CONDORCET
HUMATHÈQUE

formations.humatheque@campus-condorcet.fr

Annexes

Bibliographie

BENJELLOUN Omar, CHEN Shiyu et NOY Natasha, « Google Dataset Search by the Numbers ». https://arxiv.org/abs/2006.06894

BERNHAUER David, NEČASKÝ Martin, ŠKODA Petr, KLÍMEK Jakub et SKOPAL Tomáš, « Open dataset discovery using context-enhanced similarity search », *Knowledge and Information Systems*, décembre 2022, vol. 64, nº 12, p. 3265-3291. https://link.springer.com/article/10.1007/s10115-022-01751-z

CHAPMAN Adriane, SIMPERL Elena, KOESTEN Laura, KONSTANTINIDIS George, IBÁÑEZ Luis-Daniel, KACPRZAK Emilia et GROTH Paul, « Dataset search: a survey », *The VLDB Journal*, janvier 2020, vol. 29, nº 1, p. 251-272. https://link.springer.com/article/10.1007/s10115-022-01751-z

GREGORY Kathleen, *Findable and reusable?: Data discovery practices in research*,maastricht university, s.l., 2021. https://cris.maastrichtuniversity.nl/en/publications/findable-and-reusable-data-discovery-practices-in-research

GREGORY Kathleen, GROTH Paul, SCHARNHORST Andrea et WYATT Sally, « Lost or Found? Discovering Data Needed for Research », *Harvard Data Science Review*, 30 avril 2020. https://hdsr.mitpress.mit.edu/pub/gw3r97ht/release/6

Bibliographie

GREGORY Kathleen, KHALSA Siri Jodha, MICHENER William K., PSOMOPOULOS Fotis E., WAARD Anita DE et WU Mingfang, « Eleven quick tips for finding research data », *PLOS Computational Biology*, 12 avril 2018, vol. 14, nº 4, p. e1006038. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006038

KERN Dagmar et MATHIAK Brigitte, « Are There Any Differences in Data Set Retrieval *Compared to Well-Known Literature Retrieval? » dans Sarantos Kapidakis, Cezary Mazurek* et Marcin Werla (eds.), *Research and Advanced Technology for Digital Libraries*, Cham, Springer International Publishing (coll. « Lecture Notes in Computer Science »), 2015, p. 197-208. https://link.springer.com/chapter/10.1007/978-3-319-24592-8_15

Fiche d'analyse des outils de recherche de données

Présentation de l'outil

Nom	
URL	
Typologie	
Créateur(s)	
Y-at-il des données spécifiques dans cet outil ? Par exemple : images, sons, vidéos, statistiques, enquêtes, *etc.*	

Modalités de recherche des données

Existe-il des **modalités de recherche spécifiques** dans cet outil ? Par exemple : utilisation de filtres, module de recherche avancée, navigation dans les jeux de données, *etc.*	
Les modalités de recherche vous semble-t-elle facile à appréhender et à utiliser ?	
Les résultats de recherche vous apparaissent-ils pertinents par rapport à la requête formulée ?	

Informations sur les jeux de données	
Les éléments de description des jeux de données sont-il présents ? Par exemple version, date, producteur, format, *etc.*	
Les informations présentés vous semblent-elles claires ?	
Autres informations notables ?	

Modalités d'accès aux données

Les **modalités d'accès aux données** sont-elles bien spécifiées, par exemple : téléchargement libre, embargo, demande d'accès, procédures d'accès, *etc.*	
Les informations présentés vous semblent-elles claires ?	
Autres informations notables ?	

Modalités de réutilisation des données

Les **modalités de réutilisation des données** sont-elles bien présentes et explicitées ? Présence de licences par exemple, restrictions d'usage, spécifications précises, *etc.*		
Les informations présentées vous semblent-elles claires ?		
Autres informations notables ?		