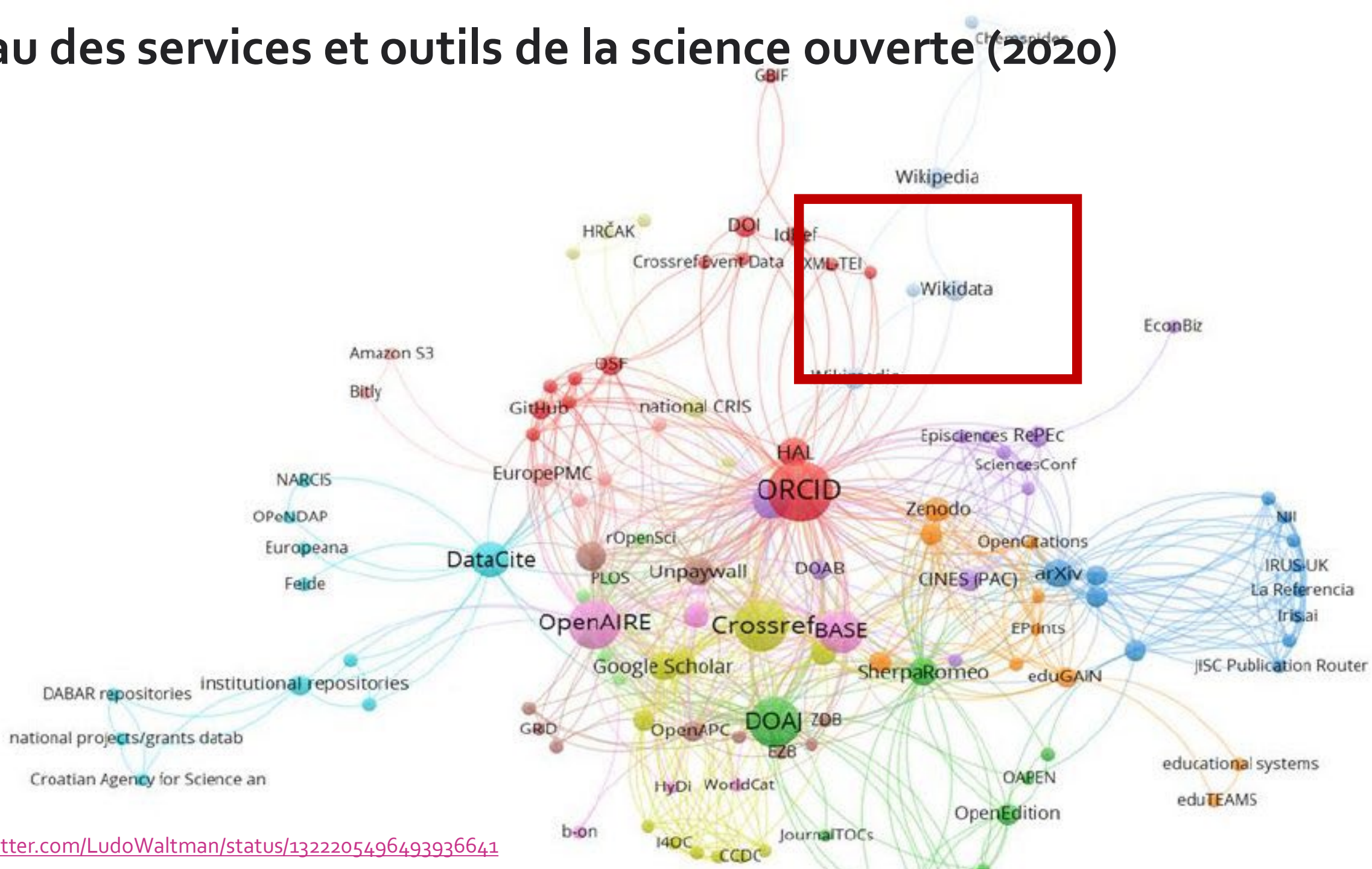




Wikidata et la curation des identifiants pérennes des chercheurs

Pascal Martinolli
Bibliothécaire
Lettres et sciences humaines
2022 CC-BY

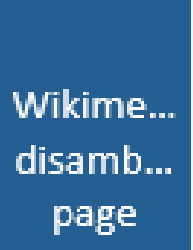
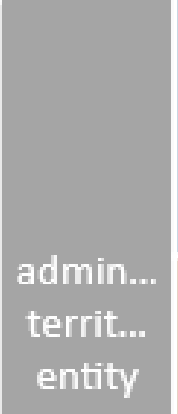
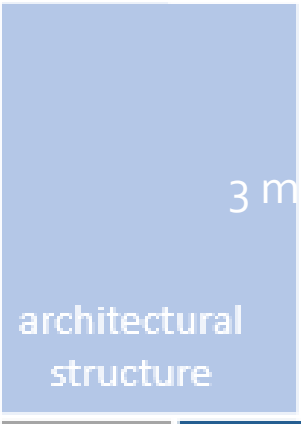
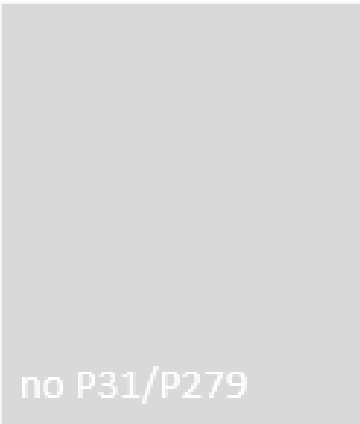
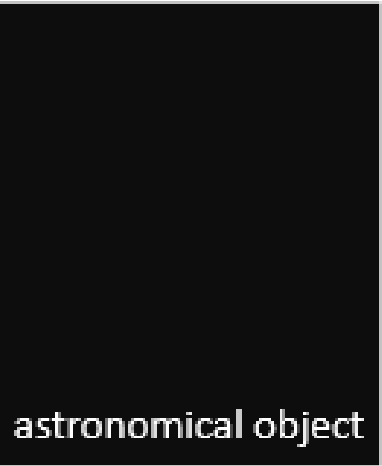
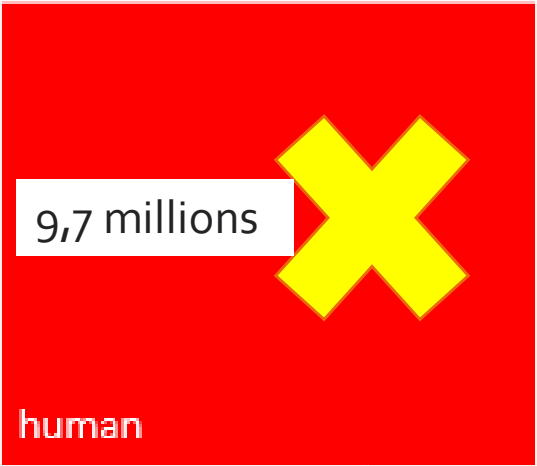
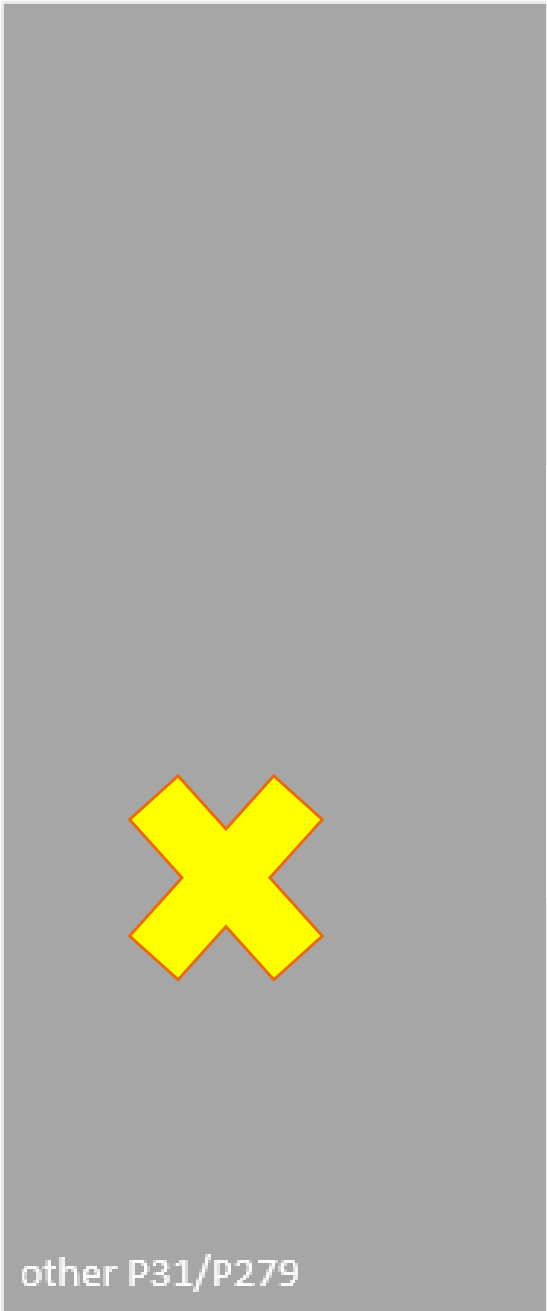
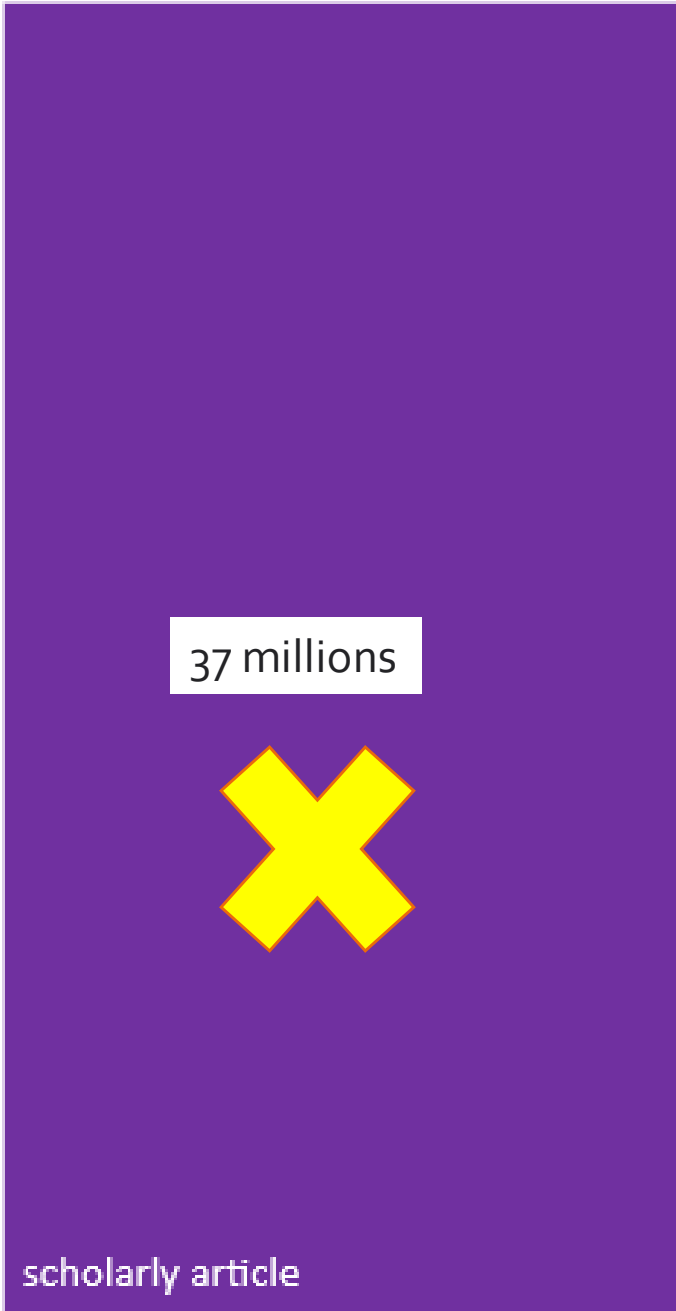
Réseau des services et outils de la science ouverte (2020)



QUOI

Wikidata ou... WikiMetadata ?

- Référentiel centralisé, ouvert, collaboratif
 - d'éléments uniques
- créé à l'origine pour coordonner les métadonnées des projets *Wikipédia, Wikimedia Commons, WikiQuotes, WikiBooks, ...*
- maintenant créant des liens entre des milliers de bases de données



Critères de notoriété :

- Beaucoup plus bas que Wikipédia
- Un article révisé par les pairs
OU
une monographie
(consensus du *Bistrot*)

Marie-Claire Daveluy (Q3291676)

Canadian librarian and writer


[In more languages](#)
Configure

Language	Label	Description
English	Marie-Claire Daveluy	Canadian librarian and writer
French	Marie-Claire Daveluy	bibliothécaire, historienne et écrivaine québécoise
Italian	Marie-Claire Daveluy	No description defined
German	No label defined	kanadische Schriftstellerin, Bibliothekarin und Historikerin

[All entered languages](#)

Statements

instance of human [1 reference](#)

image 

<https://www.wikidata.org/wiki/Q3291676>

Wikidata : un référentiel fédérateur d'identifiants



« devenue progressivement le point de convergence mondial des identifiants ouverts (...) »

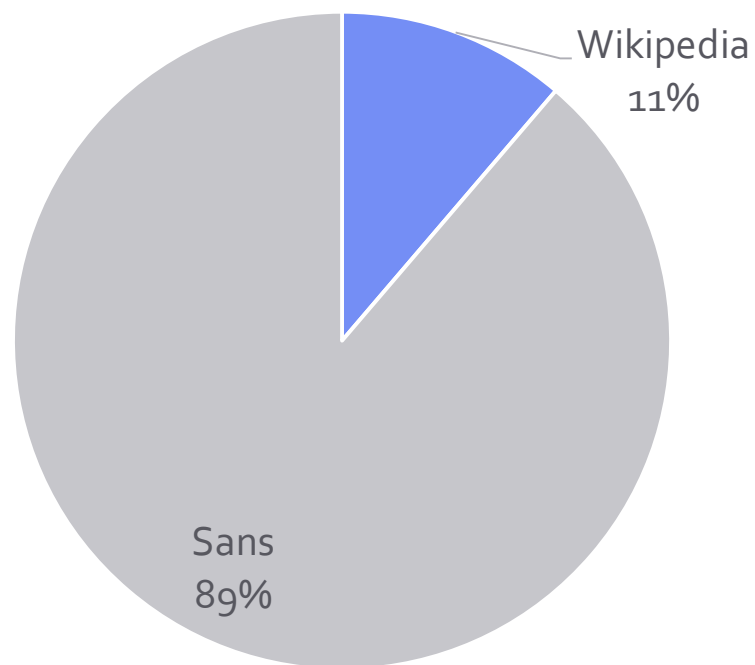
Le référentiel ouvert le plus riche (...) **dont la gouvernance et le modèle économique ne garantissent pas encore complètement la pérennité** (...) »

Faire que tous les chercheurs français, vivants ou ayant vécu aient (...) **un Q Wikidata**, aligné avec leur(s) identifiant ORCID, IdHAL (...) »

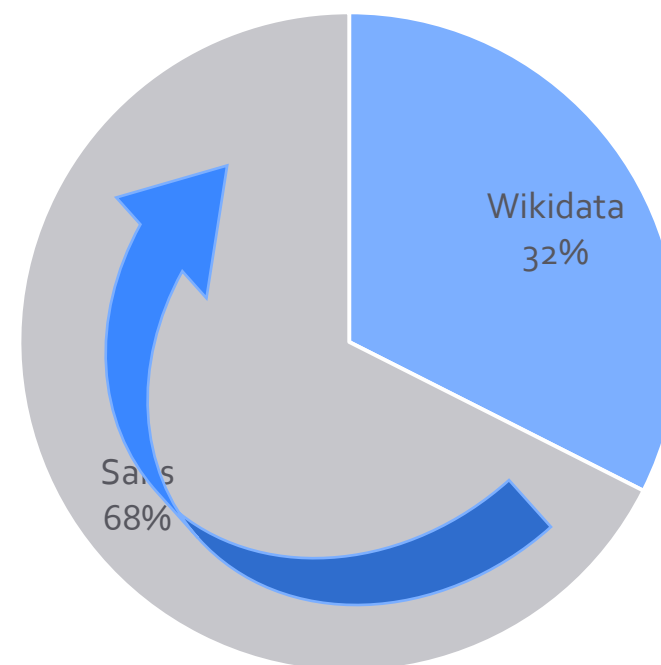
Départements liés à la Bibliothèque des lettres et sciences humaines (n=453, oct. 2019)

Enquête e-profil BLSH
2019

Nombre de profils Wikipédia
EN ou FR



Nombre d'identifiants
Wikidata



Mais aussi :

Les publications des chercheurs

Ambitieux : demande de maîtriser les données de citation, les sujets, les auteurs, les éditions, etc.

les versements en lot,
les indexations automatiques
identifiants DOI, ISBN, etc.

Les laboratoires, les départements, les facultés, les presses universitaires, les revues, etc.

Facile :

Leurs histoires (créé, disparu, renommé, fusionné, etc.)

Leurs domaines

Leurs coordonnées

Leurs affiliations, etc.

identifiants ROR, ISSN, etc.

Les partenaires : fonds de recherche, les prix, les associations, etc.

Facile

Les domaines de recherche : **Expert** : ontologie, hiérarchie, traduction, sous-discipline, etc.



POURQUOI

POURQUOI

- Fédération et alignement **d'identifiants pérennes** disparates (identifiants pérennes de chercheurs)
- **Explorer** les données
- **Compléter ou enrichir** son propre jeu de données
- **Faciliter l'accès** à des grands jeux de données
- Alignement sémantique et ontologique **multilingue**

Métadonnées
≈
Passerelles

Pourquoi II (pour bibliothécaire universitaire)

- Tisser des liens avec les professeurs, discuter avec eux sur eux
- Histoire des départements, des labos, des prix, des fonds, etc.
- Contribuer à la science ouverte et la communication scientifique
 - Chercheurs
 - Institutions
 - etc.
- Se pratiquer au SPARQL / RDF

POURQUOI

1) Identifiants pérennes de chercheurs

Données alignées

- ORCID
- Google Scholar ID
- site web perso
- Twitter ID
- Freebase ID
- VIAF ID
- LOC ID
- ResearcherID
-

Library of Congress authority ID	ORCID	nr96012906	▼ 0 references
Mathematics Genealogy Project ID	ORCID	73102	▼ 0 references
ORCID iD	ORCID	0000-0002-9322-3515	▼ 0 references
ISNI	ORCID	0000 0000 7397 8908	▼ 0 references

Pourquoi aligner ces identifiants pérennes ?

- Aider à la désambiguation
- Aider à la construction de plateformes ouvertes
- Enrichir les bases de données liées
- Produire des rapports statistiques simples
- Améliorer les résultats de recherche ? *Search Engine Optimization*

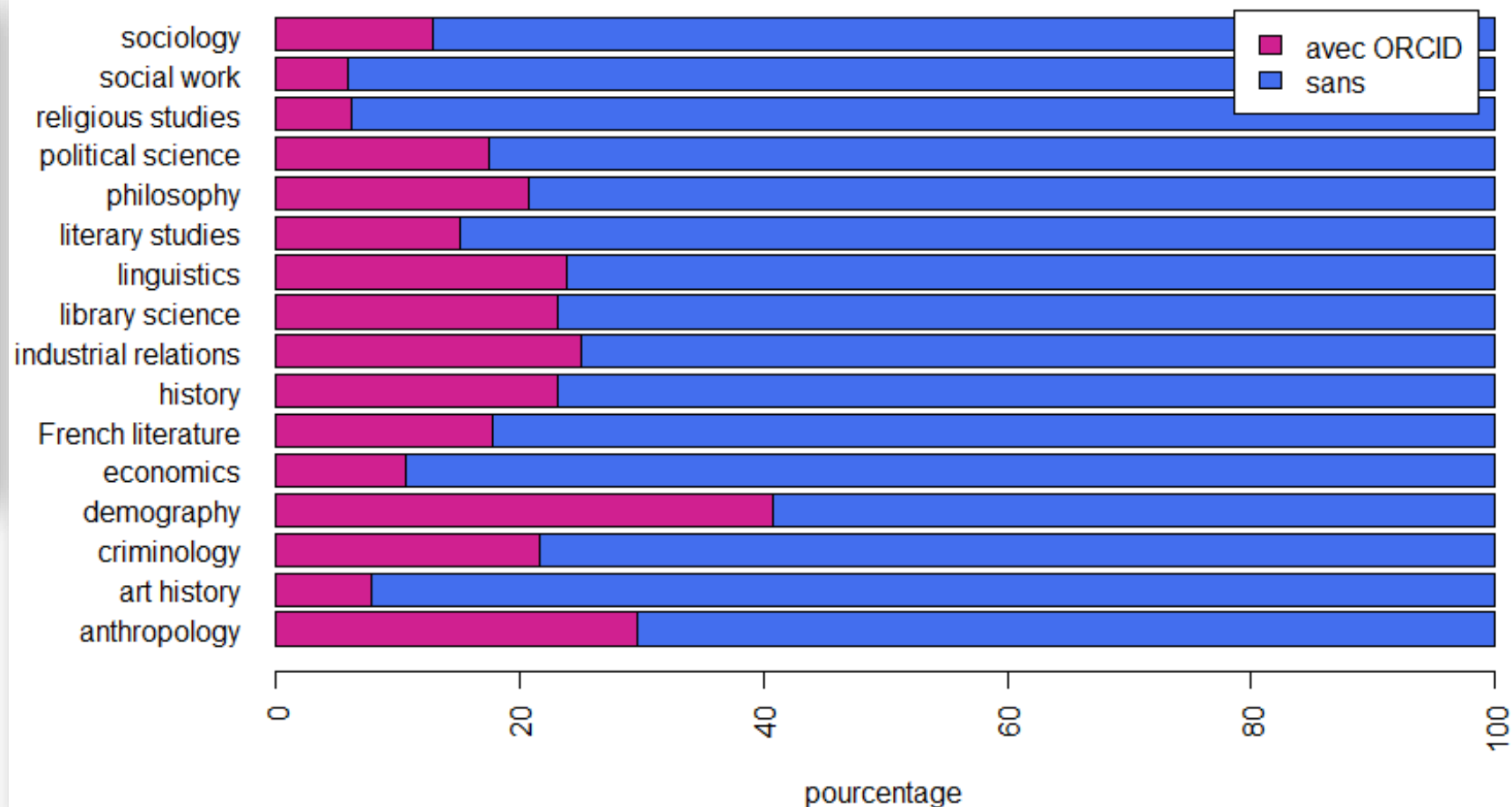
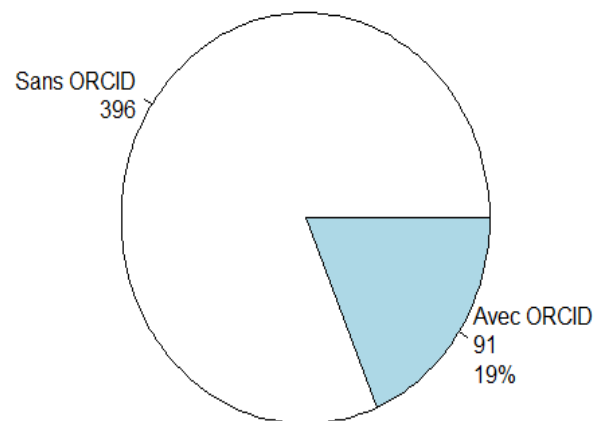
Pas d'amélioration significative sur le PageRank des chercheurs UdeM

https://www.wikidata.org/wiki/User:Pmartinolli/Curation_chercheurs_UdeM/Impact_on_PageRank

Exemples de rapports statistiques simples

ORCID/ BLSH UdeM

(qui a un ORCID ?)



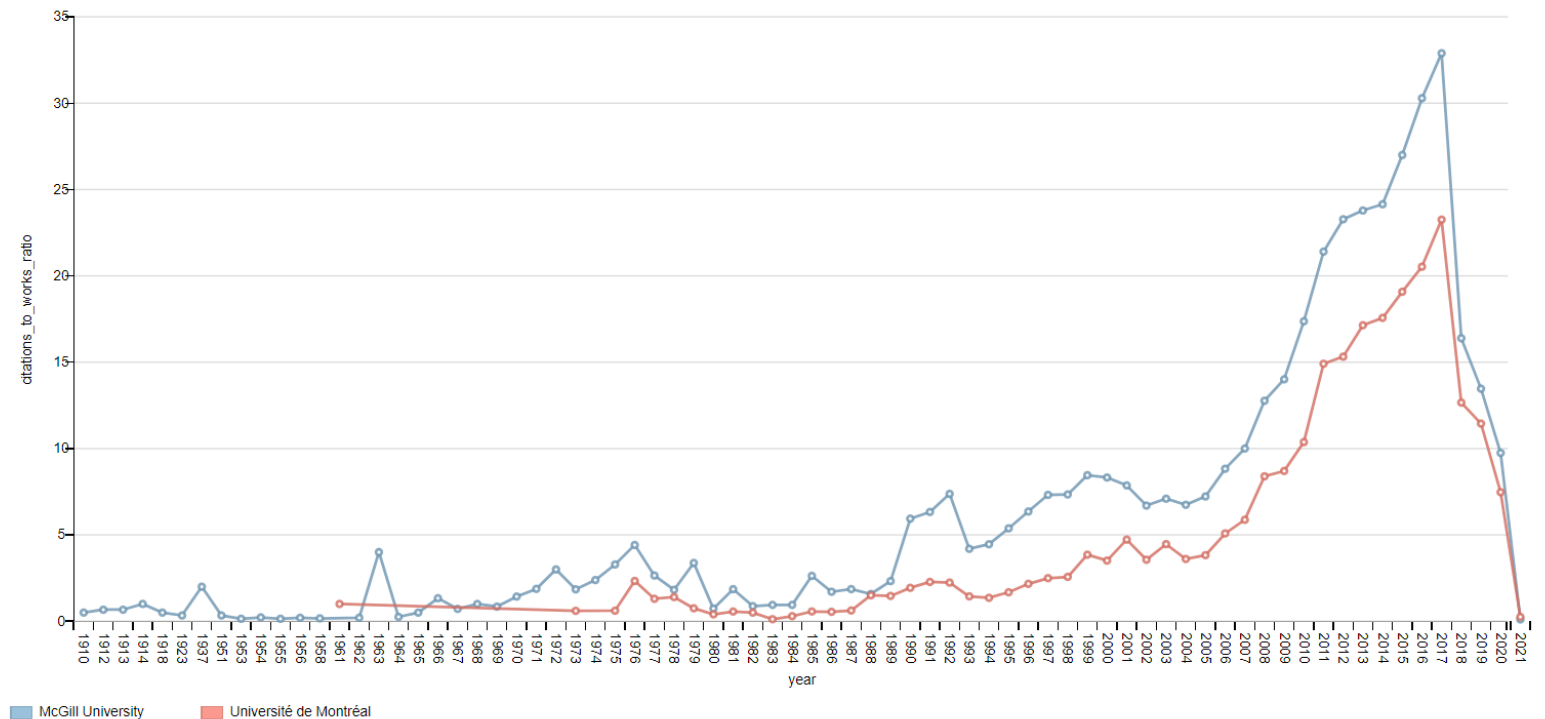
Produire des rapports simples avec Scholia

SCHOLIA : scientométrie

- researchers
- organizations
- journals
- publishers
- papers
- topics

Citations to works ratio

The ratio between the number of citations received and the works authored per organization per year.



author

Yoshua Bengio (Q3572699)

Yoshua Bengio (born March 5, 1964 in Paris, France) is a Canadian computer scientist, most noted for his work on artificial neural networks and deep learning. He is a professor at the Department of Computer Science and Operations Research at the Université de Montréal and scientific director of the Mila Institute for Learning Algorithms (MILA). Bengio received the 2018 ACM A.M. Turing Award, together with Geoffrey Hinton and Yann LeCun, for deep learning. ... (from the English Wikipedia)

Related: Alimuddin Zumla · John Kuriyan · Philarète Chasles · Jean-Yves Veillard · Semion Rostovtsev · Max Heinze · Dmitry Konstantinovich Bobylev · Alekseyevich Kinsky · Arnold Munnich · Émile Saisset

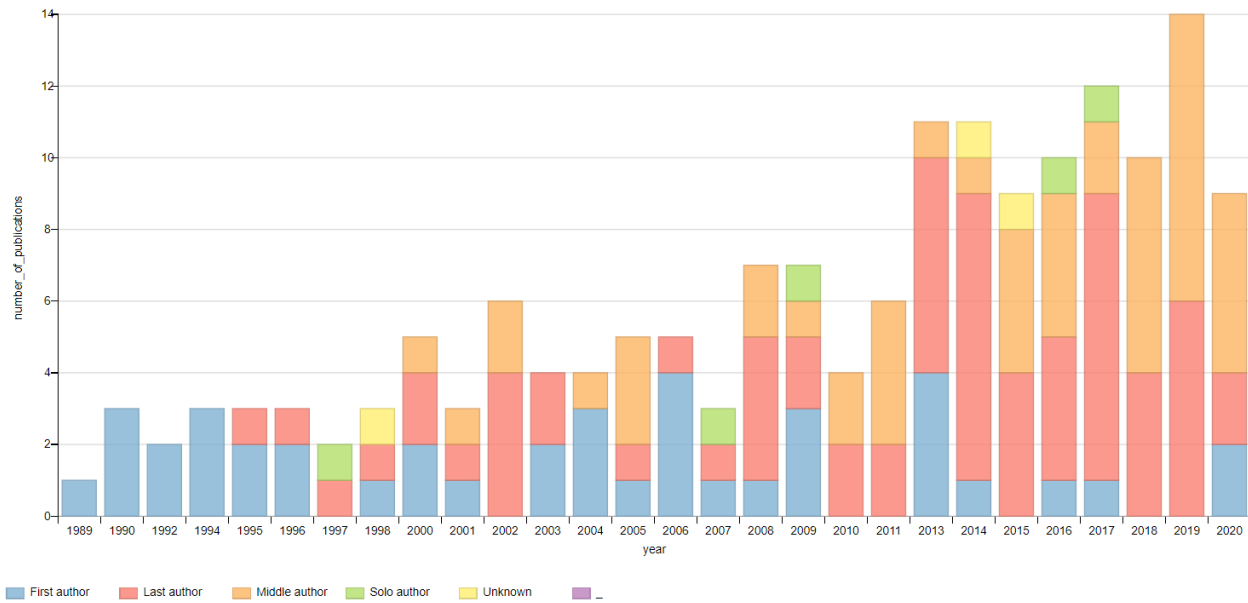
<https://orcid.org/0000-0002-9322-3515>

List of publications

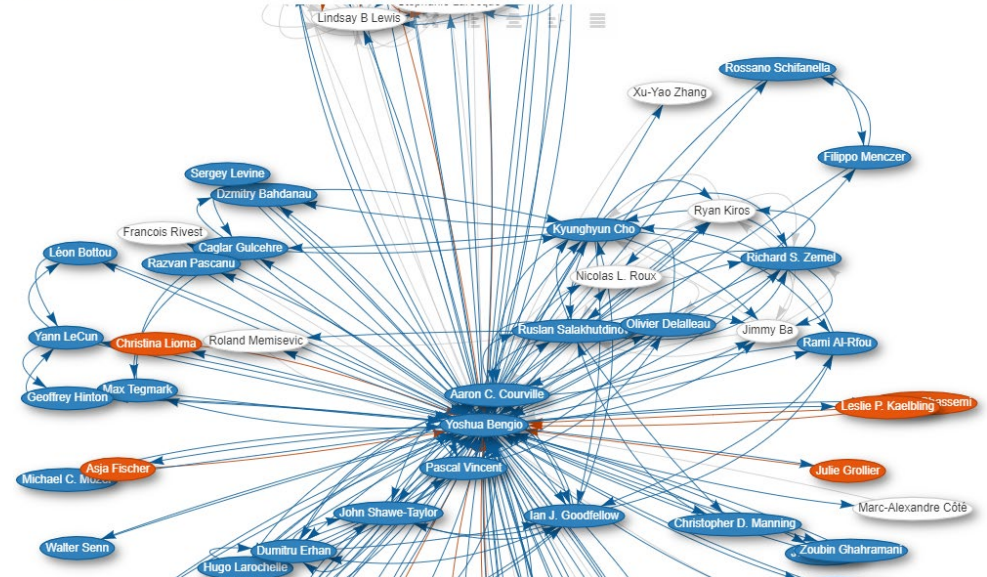
Show 10 entries

Search:

Number of publications per year

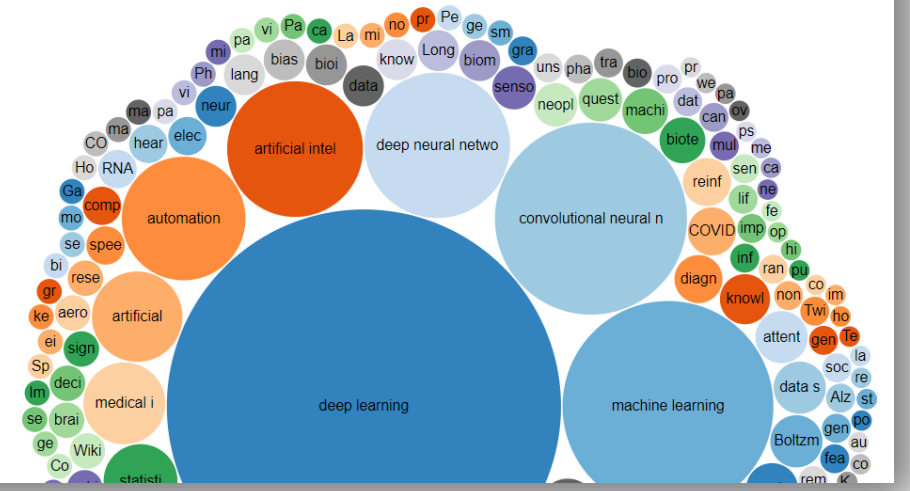


Co-author graph

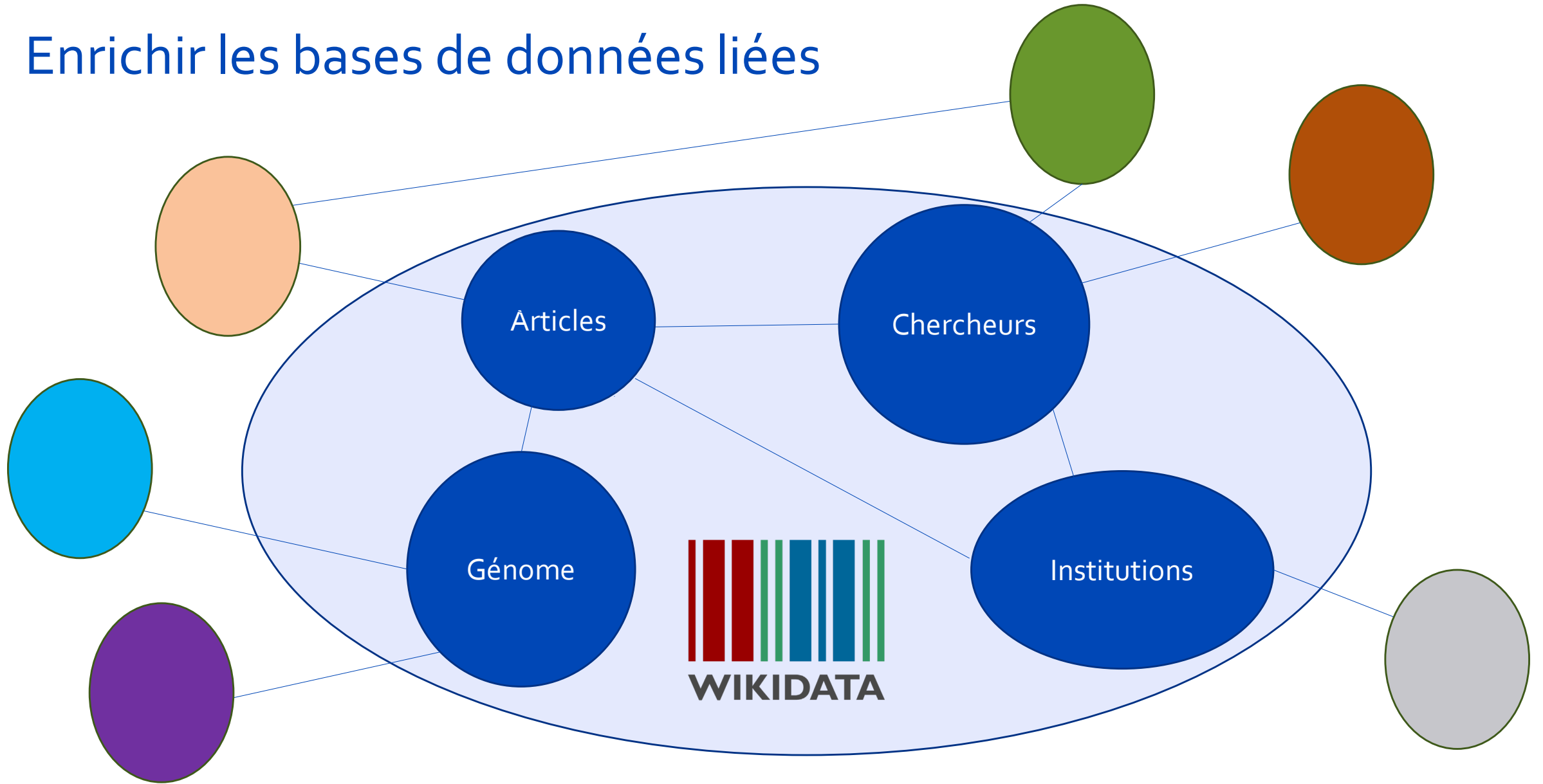


Topic scores

Topics based on a weighting between fields of work, topics of authored works and topics of citing works.



Enrichir les bases de données liées



Exemples d'utilisation

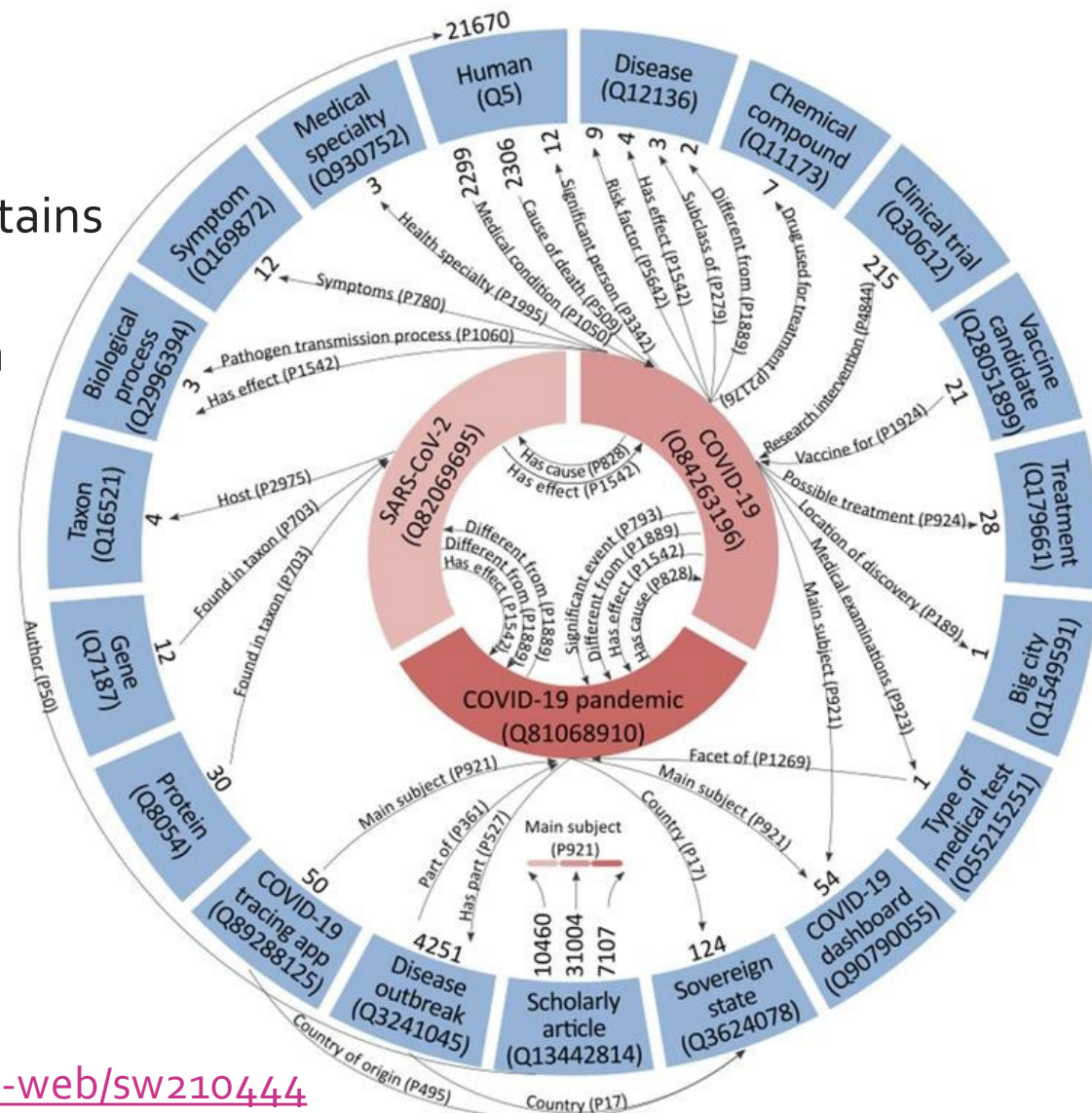
Faire de la bibliométrie avancée

(inaccessible si on n'a pas des accès privilégiés à certains outils)

- Repérer les collaborations entre chercheurs sur un thème en particulier
- Repérer les chercheurs qui ont tendance à publier souvent chez des éditeurs « à la limite »

Rassembler/visualiser la science sur un thème :

Zika, Covid, etc.



POURQUOI

2) Aligner le catalogue OCLC

Zone 024 du catalogue OCLC

Liste d'autorité
partagée
par universités québécoises
BAC et BAnQ
(tous les chercheurs UdeM)

Mais pas exploité par le logiciel
d'OCLC «pour le moment»

024	7		Q72334 \$2 wikidata
035			(CaOONL)411428
042			nlc
046			\$f 1931-02-18 \$g 2019-08-05 \$2 edit
053		4	PS3563.O8749 \$5 CaQMU
100	1		Morrison, Toni
368			Lauréats du Prix Nobel \$2 rvmgd
370			Lorain (Ohio) \$b New York (N.Y.) \$2 lacnaf
372			Littérature \$2 rvm
374			Romanciers \$a Essayistes \$a Critiques \$a Dramaturges \$a Librett

Communications

Wikidata: From “an” Identifier to “the” Identifier

Theo van Veen

ABSTRACT

« pour certaines institutions, Wikidata peut servir de **mécanisme de contrôle d'autorité.**»

« **2000** bases de données bibliographiques »

« avantages de Wikidata comme **identifiant universel** »

« Quand Wikidata > VIAF et ISNI ? »

<https://doi.org/10.6017/ital.v38i2.10886>



The Journal of Academic Librarianship

Volume 47, Issue 2, March 2021, 102326



Much more than a mere technology: A systematic review of Wikidata in libraries

Karim Tharani

Show more

+ Add to Mendeley Share Cite

<https://doi.org/10.1016/j.acalib.2021.102326>

[Get rights and content](#)

A systematic literature review on Wikidata

Marçal Mora-Cantallops, Salvador Sánchez-Alonso and
Elena García-Barriocanal

«...libraries generally embrace Wikidata as an open and reusable knowledge base of **structured data capable of linking their local metadata with a network of global metadata.**

In terms of application in libraries, Wikidata is **primarily being used as a central platform to link authority data to improve local and global metadata quality and processes.**»



Wikidata pour les données de recherche

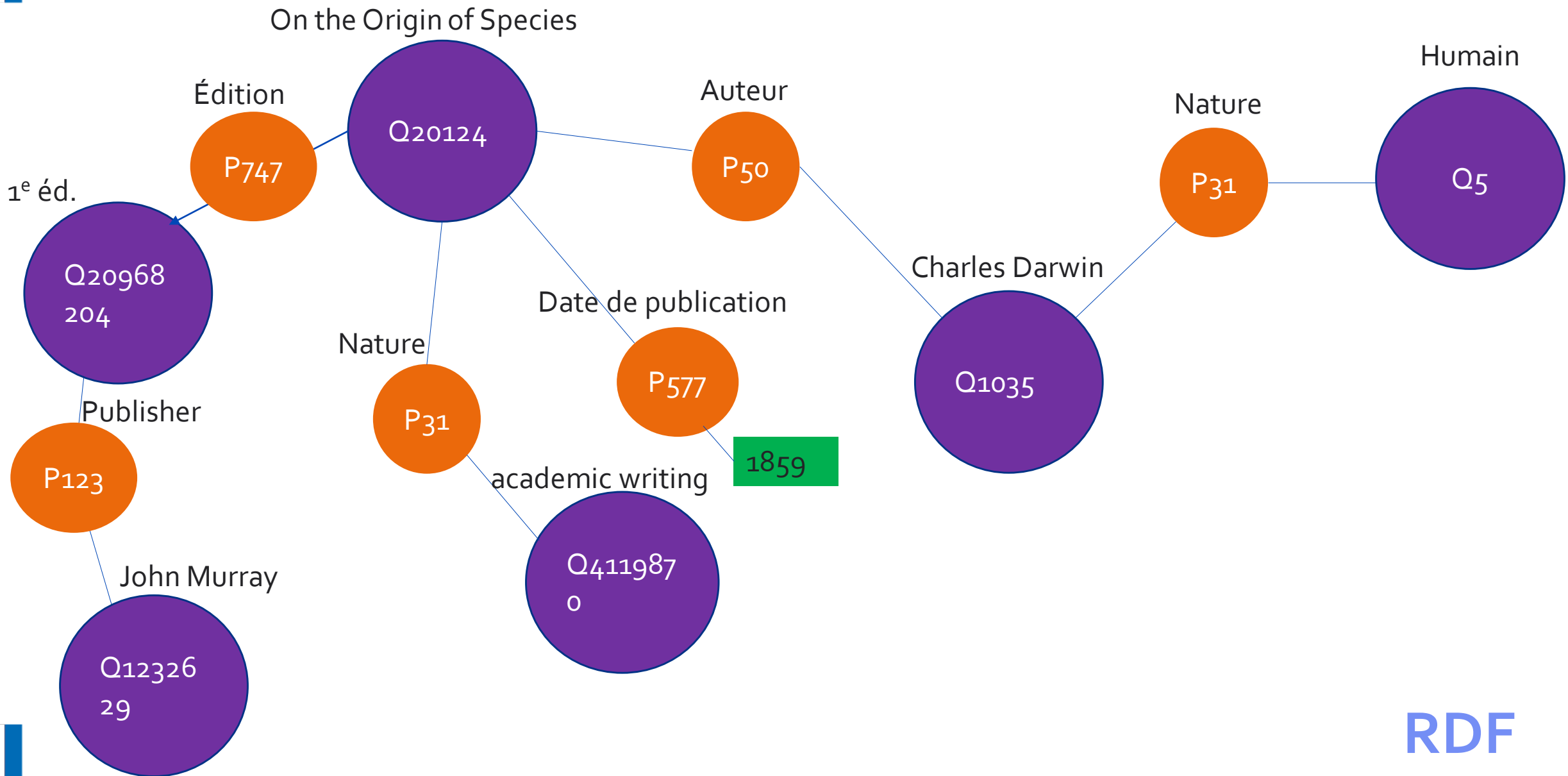
- Alignement : Rendre les données de recherche plus accessibles
- Versement : Diffuser une partie de ses données de recherche

https://www.wikidata.org/wiki/Wikidata:WikiProject_Weather_observations



POURQUOI

3) Explorer les données



RDF

Explorer les données avec le langage SPARQL et la structure RDF des données

- Quels sont tous les chercheurs de l'UdeM qui sont nés dans un lieu qui est situé dans le Québec ?
- Les diplômés de l'UdeM sont-ils décédés plus loin de leur lieu de naissance que les diplômés de McGill U ?

Explorer les données avec le langage SPARQL et la structure RDF des données

- **Recherche de réviseurs par les pairs:**
Quels sont les chercheurs cités dans les travaux de tel chercheur mais qui n'ont pas été des co-auteurs ?
- **Aide à la revue de littérature**
Lister toutes les publications sur un thème et trier/pondérer/marker par :
 - Revues appartenant à des organismes prédateurs ou « à la limite »
 - Chercheurs ayant travaillé avec tel autre chercheur après telle date

Limites de l'exploration et quelques solutions

- Capacité à **coder une requête SPARQL** plus ou moins complexe
 - Demander à un codeur
 - Requêtes préfabriquées de Scholia
- Existence de **déclarations pertinentes** (propriétés)
 - Demander la création de propriétés
- **Qualité et quantité** des données dans Wikidata
 - Ajouter les données soit même

Bibliothécaires universitaires : expertise SPARQL

- Depuis les années 1960-1970 : opérateurs booléens
- Nouvelle expertise : requêtes SPARQL et données RDF
 - Simple au début
 - Mais devient rapidement très complexe

```
SELECT ?day ?count ?PID {
  values (?property) {"921"} ("50") ("2093") ("577") ("2860") ("1476")} #("108")
  bind (concat("|", ?property, "=") as ?p)
  bind (concat(".+\\", ?p, "(\\d+).+") as ?r)
  service wikibase:mwapi {
    bd:serviceParam wikibase:api "Generator" ; wikibase:endpoint "www.wikidata.org"
    mwapi:gapfrom "Property_uses" ; mwapi:gapto "Property_uses" ; mwapi:gapname "Property_uses"
    mwapi:prop "revisions" ; mwapi:rvprop "content|timestamp" ; mwapi:rqlimit "100"
    ?t1 wikibase:apiOutput "revisions/rev[1]/@timestamp" . ?r1 wikibase:apiOutput
  }
  bind (if(contains(?r1, ?p), xsd:integer(replace(?r1, ?r, "$1")), -1) as ?count)
  filter (?count != -1)
  bind (xsd:dateTime(?t1) as ?day)
  bind (concat("P" ?property) as ?PID)
```

COMMENT

1) Éditer

Chaque P ou Q possède sa propre page

- +/- détaillée
- +/- de références
- +/- de base de données liées

/m/0jwvf5b

▼ 1 reference

stated in	Freebase Data Dumps
publication date	28 October 2013

Identifiers

Freebase ID	/m/0jwvf5b
-------------	------------

Page de discussion

Item Discussion

Historique d'évolution

View history

Alerte de suivi



Parfois créés par des robots ou des traitements en lot

- Voir **Historique de l'élément** pour voir comment il a été créé

<https://www.wikidata.org/w/index.php?title=Q54234666&action=history>

- (cur | prev) 15:14, 26 May 2018 QuickStatementsBot (talk | contribs) . . (1,369 bytes) (+1,369) .
| *(Created a new Item: #quickstatements; invoked by*
| *SourceMD:CreateFromWikipediaDOIs) (restore)*

Projets d'importations massives

- Irréguliers (peu de flux automatisé systématique)
- Initiatives individuelles
- Traçabilité et étendue du corpus : connue

WikiProject Source Metadata

https://www.wikidata.org/wiki/Wikidata:WikiProject_Source_Metadata

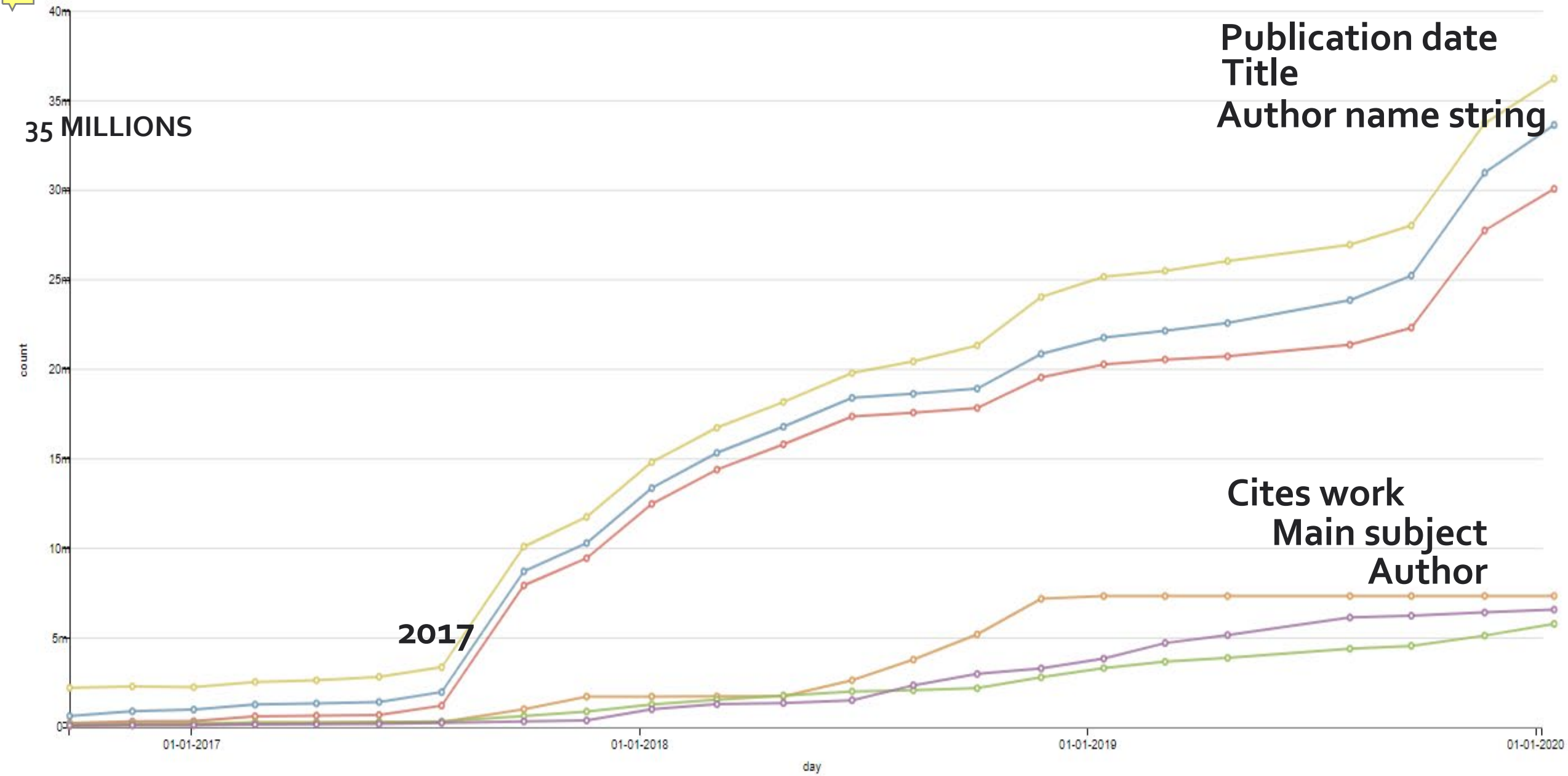


35 MILLIONS

Publication date
Title
Author name string

Cites work
Main subject
Author

2017



P1476 P2093 P2880 P50 P577 P921

Évolution des propriétés

Comment éditer Wikidata

- Avoir un compte Wikimedia
- À la main
- Par des logiciels (en lot)
 - *QuickStatements* + «LibreOffice CALC, le roi du CSV»
 - *Author Disambiguation* (author name string -> author)
 - *ORCIDator* (complète les données Wikidata à partir de ORCID)
 - *Mix'n'match* (données liées)
 - *OpenRefine*

COMMENT

2) Réutiliser



Scholia
Wikigenome
...



Google
(via Knowledge Graph)



Rawgraph.io



COMMENT

3) Robustesse

FAIR

Facile à repérer : Wikimedia, FAIRsharing, identifiers.org

Accessible : SPARQL, API, éditer-lire

Robustesse 

FARIR ?

Interopérable : PIDs, URIs, Json, XML, RDF

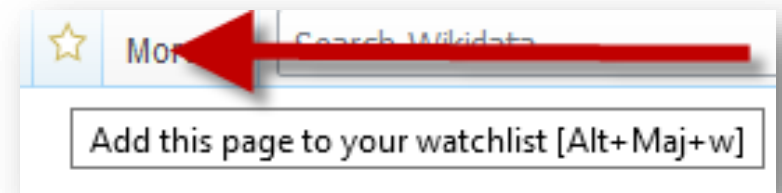
Réutilisable : CC0 (métadonnées : pas toujours évident)

Risque de diffamation ou vie privée faible

En raison des **déclarations** suivantes (et d'autres non listées), il est recommandé aux chercheurs ayant un élément Q à leur nom de faire une veille active sur leur profil.

- Place of birth (P19)
- Date of birth (P569)
- Located at street address (P6375)
- Phone number (P1329)
- E-mail address (P968)
- Sexual orientation (P91)
- Political ideology (P1142)
- Religion (P140)
- Medical condition (P1050)
- Cause of death (P509)

Pour cela, il faut avoir un compte Wikimedia et ensuite cliquer sur l'étoile d'activation de suivi :



Oversight : 12 par an en 2018 environ (dans le monde)

Logiciel - Serveurs

- Dump download (1 fois /semaine)
- WikiBase : Logiciel libre-> algorithme connu
<https://fr.wikipedia.org/wiki/Wikibase>
- Montée en charge/ Croissance
 - Nouveaux éléments : léger
 - Liens entre les éléments : moyen
 - Requêtes avancées : lourd

Données et ontologies

Tissé serré :

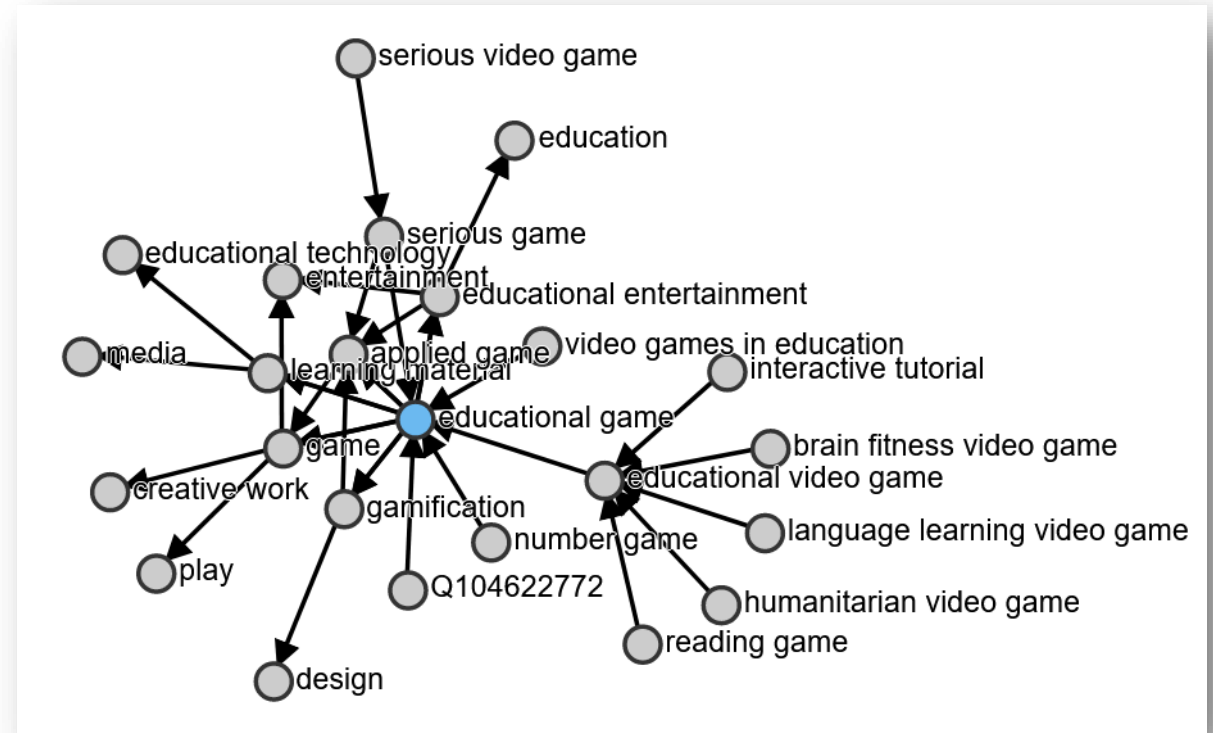
- Contraintes des propriétés
- Cascade des données liées

- Vaste et inégal
- Très incomplet
- Biais

Données et ontologies

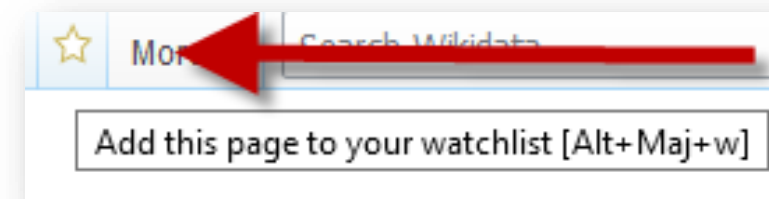
- Ontologies des sujets : embryonnaire et instable dans des sous-disciplines

OpenAlex : topic lié à Wikidata



Données et ontologies

- Moins de vandalisme que Wikipédia (moins connu?)
 - *Wikidata vandalism dashboard*
 - Rapidement traité (facile à cibler + répercussion)
- Patrouilles Wikidata
- «Auto-patrouille» : suivi de ses éléments
 - Ex: Miguel Tremblay (8700 stations météo)



Gouvernance Wikimedia

- Financement
- Wikimedia Canada : bibliothécaires
- Wikimedia Enterprise : changement de modèle ?

Taux d'erreur

- Surtout des **Nature de l'élément**

- Trop génériques
- Erronées

- Exemple : un **chercheur** qui a écrit un **livre** est indexé comme **écrivain** par un robot qui a moissonné un catalogue.



- «problematic classification and taxonomic statements, related to an inadequate use of instantiation and sub-classing in certain Wikidata hierarchies.» [Brasileiro et al. \(2016\)](#)


Robustesse sociale

- Importation en masse -> acceptation de la communauté
- Oligarchie biaisée par l'aspect technique ou l'ancienneté ?
- Renouvellement générationnel
 - Diversité
 - Valeurs et engagement

COMMENT

4) Et vous ?

- Contribuer à un (groupe d') élément qui vous tient à cœur
 - Vos hobbies  : facilitateur de pratique
 - Annuaire partagé d'institutions pour la recherche
 - Le soir, devant la tv  : peu demandant en terme d'attention exclusive

- Bibliothécaire universitaire :
prendre soin des éléments de vos chercheurs 
 - https://www.wikidata.org/wiki/User:Pmartinolli/Curation_chercheurs_UdeM
 - -> prochaine réunion départementale

Pour aller plus loin

- Projets WikiCite
 - WikiProject Source Metadata
https://www.wikidata.org/wiki/Wikidata:WikiProject_Source_MetaData
- Tutoriel et requêtes d'exemple pour la curation des chercheurs UdeM
https://www.wikidata.org/wiki/User:Pmartinolli/Curation_chercheurs_UdeM



Conclusion

Un outil

**pour créer des liens
améliorer les métadonnées**

universel

FAIR

transparent

incomplet

à suivre sur le temps long

Merci
Questions?