COLLECTIONS NUMÉRIQUES ET LEURS DONNÉES

ACCÈS ET PROTECTION À L'ÈRE DES DONNÉES LIÉES (*LINKED DATA*)

Lyne Da Sylva, EBSI

Colloque sur le web sémantique au Québec. « Web sémantique : culture de la donnée et développement socio-économique »

Montréal – 6 juin 2019

PROBLÉMATIQUE

- Mise en ligne de documents vs protection de la confidentialité
- Modes d'accès : métadonnées (« traditionnelles ») et, de plus en plus, maillage via données liées
- Objectif : envisager diverses options pour maximiser l'accès et limiter les effets négatifs

PLAN DE L'EXPOSÉ

- Web sémantique, données liées et milieux archivistiques
- Ouverture des données et protection de la confidentialité

WEB SÉMANTIQUE, DONNÉES LIÉES…

EXEMPLE DE BESOINS « SÉMANTIQUES » EN CONTEXTE ARCHIVISTIQUE : RECHERCHE ÉLABORÉE

Je cherche des images du patrimoine architectural du centre-ville

définition de « patrimoine architectural »
délimitation du territoire – ajout de plan
identification d' « images »

évaluation de « notable »
BD d'acteurs économiques

Quelles politiques culturelles ont eu des effets notables sur le développement économique local?

… ET MILIEUX ARCHIVISTIQUES

PARTICULARITÉS DES COLLECTIONS D'ARCHIVES

- Documents uniques
  - Limites des outils de description génériques (voir Guitard, 2018)
  - Couverture informationnelle limitée
  - Potentiel élevé de complémentarité avec jeux de données externes
- Cycle de vie
  - Archives définitives vs « documents d'activités »
- Présence des documents dans la collection liée au mandat du service
  - BAC vs BAnQ vs société historique d'une municipalité
- D'intérêt particulier ici : services d'archives définitives voués à l'ouverture de leurs données

LIER DES JEUX DE DONNÉES POUR LES ARCHIVES

assurer ma visibilité

lier à sources externes diverses

Documents d'archives

Personnes

Événements

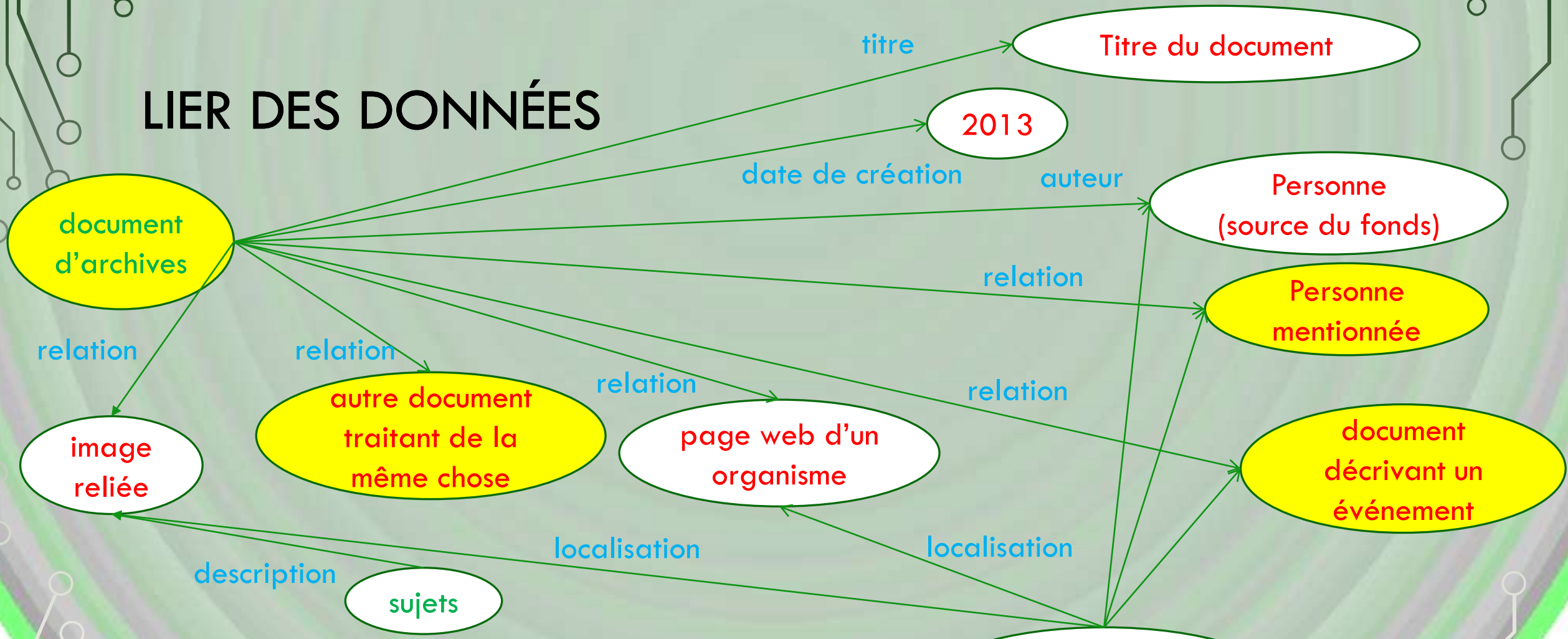
partager avec mon réseau

Organismes

Lieux

exploiter les données en les combinant

LIER DES DONNÉES



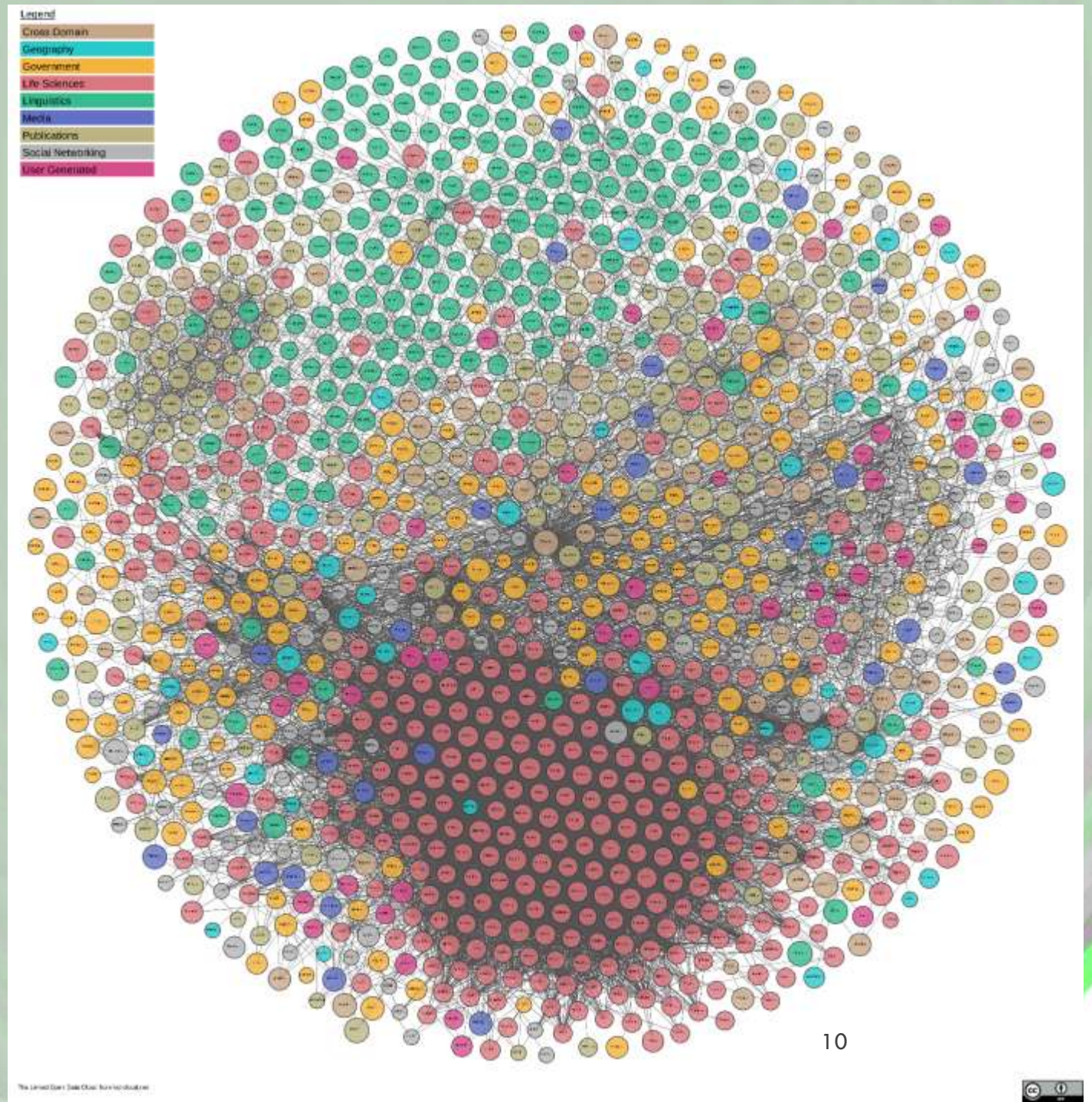
- Provenant de jeux de données distincts
- Différents statuts de confidentialité

LOD : ÉTAT EN MARS 2019

(https://lod-cloud.net/versions/2019-03-29/lod-cloud.png)

Un seul jeu de données contenant « archiv » :

- http://data.archiveshub.ac.uk/
- pas mis à jour depuis 2013

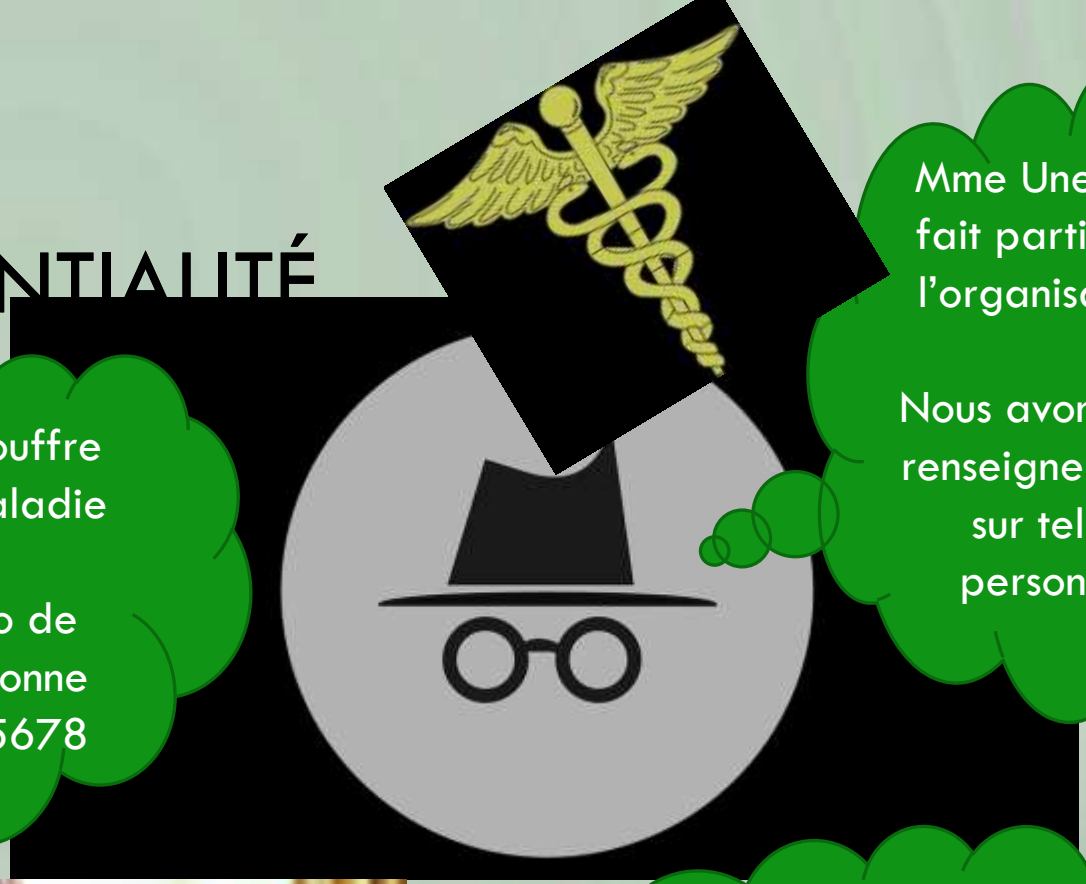


OUVERTURE DES DONNÉES ET PROTECTION DE LA CONFIDENTIALITÉ

PROBLÈMES DE CONFIDENTIALITÉ

M. Untel souffre de telle maladie

Le numéro de cette personne est 12345678



Mme Unetelle fait partie de l'organisation

Nous avons des renseignements sur telle personne

83% des Américains pourraient être identifiés par la combinaison de 3 informations : zip code, date de naissance et genre

M. Untel et Mme Unetelle collaborent…

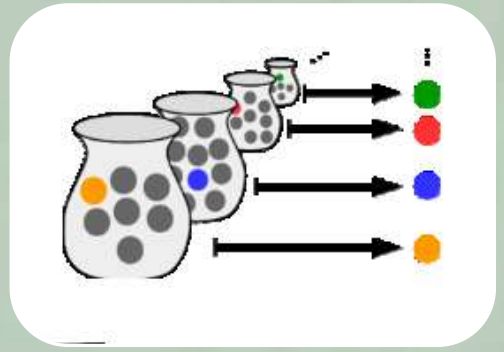
M. Untel a consulté telle personne

Samarati, P., & Sweeney, L. (1998). *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression* (pp. 101-132). technical report, SRI International.

SOLUTIONS POSSIBLES

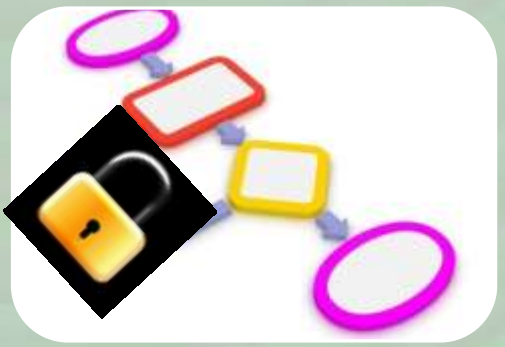
Documents ou données non accessibles

Anonymisation



Choix des métadonnées et des liens

Censure sur les réponses retournées



Contraintes sur les règles de raisonnement utilisées

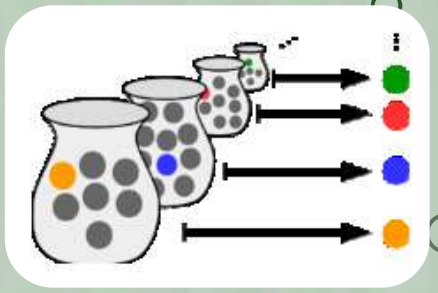
DOCUMENTS OU DONNÉES NON ACCESSIBLE

- Globalement (documents ou données cachés)
- Pour certains utilisateurs (restrictions d'accès selon le profil)

ANONYMISATION

- Manuelle (pour documents sélectionnés)
  - Ajout de fausses données
  - « k-anonymité » (chaque combinaison de quasi-identifiants associée à au moins k-1 entrées)
  - L-diversité (au moins L valeurs sensibles différentes pour chaque combinaison de quasi-identifiants)
- Semi-automatique
  - Techniques d'encodage
  - Troncation de valeurs
- Automatique
  - Repérage d'entités nommés

CHOIX DES MÉTADONNÉES ET DES LIENS



- Exclure des métadonnées publiées

CENSURE SUR LES RÉPONSES RETOURNÉES

- Algorithmes
  - Utilisation de « censeurs » qui s'assurent que les réponses à des requêtes ne contreviennent pas aux politiques de confidentialité (ex. Cuenca Grau et al., 2015)

Cuenca Grau, Bernardo; Kharlamov, Evgeny; Kostylev, Egor V.; Zheleznyakov, Dmitriy. Controlled Query Evaluation for Datalog and OWL 2 Profile Ontologies. *IJCAI 2015*, pp. 2883-2889.

CONTRAINTES SUR RÈGLES DE RAISONNEMENT

- Repose sur les logiques de description et d'inférences utilisées
- Exemples
  - Reformulation des requêtes (pour exclure des éléments « secrets »)
  - Filtrage des axiomes (effet de censure moins prononcée) (ex. Knechtel et Stuckenschmidt, 2010)

Knechtel, Martin et Stuckenschmidt, Heiner. Query-based access control for ontologies. In : *International Conference on Web Reasoning and Rule Systems*. Springer, Berlin, Heidelberg, 2010. p. 73-87.

EN GUISE DE CONCLUSION

- Mouvement de plus en plus important dans les milieux documentaires et culturels
- Sans doute plus d'avantages que de dangers
  - si on reste maître des solutions appliquées à l'interne
  - mais il importe de comprendre les conséquences
- Moins risqué que l'analyse de données massives interreliées
  - mais conséquences potentielles peu étudiées

MERCI!
QUESTIONS?