HTW Chur
Schweizerisches Institut für Informationswissenschaft

DONNÉES DE LA RECHERCHE

MODULE 3-1:
L’utilisation de Linked Data

Elena Mastrandrea
Schweizerisches Institut für Informationswissenschaft (SII)

All cartoons courtesy of Jørgen Stamp, digitalbevaring.dk
CC BY 2.5

DATA SHARING: UTILISATION ET POTENTIEL

Qu’est-ce que l’utilisation de données de la recherche publiées?

Valeur culturelle (par exemple en cas d’une collecte de données unique dans l’histoire et impossible à répéter, ou collecte de données sur des groupes difficilement accessibles et des phénomènes rares).

Rendre possible de nouvelles recherche grâce aux données générées.

De nouveaux potentiels de recherche sont créés pour des analyses secondaires avec de nouvelles questions et méthodologies de recherche, pour des comparaisons temporelles et/ou d’échantillons, ou pour interconnexion des données avec d’autres sources.

UTILISATON DES DONNÉES: CONDITIONS

Je dois comprendre les données.

- De quoi traitent les données?
- D'où viennent-elles?
- Quand ont-elles été collectées?

La qualité des données doit être bonne.

- Ai-je toutes les données dont j'ai besoin?
- Y a-t-il des erreurs ostensibles dans mes données?
- Dois-je nettoyer mes données?
- Dois-je corriger certaines erreurs?
- Dois-je compléter certaines données manquantes?

UTILISATION DES DONNÉES: CONDITIONS

Une réutilisation ultérieure des données peut être limitée quand

la qualité des données de suffit pas.

certaines conditions juridiques ne sont pas remplies.

il est impossible d'interpréter les données.

les données sont perdues ou endommagées.

"Producing data of high quality is essential to the advancement of science" (ICPSR)

„Metadata are often the only form of communication between the secondary analyst and the data producer" (ICPSR)d

Source: http://wiki.bildungsserver.de/bilder/upload/checkliste_datenmanagement.pdf

PROBLÉMATIQUE DE L'UTILISATION DES DONNÉES

Flux de données croissants

L'exploration et l'analyse de grandes quantités de données est toujours plus complexe.

Augmentation de la complexité des données

Comme puis-je créer avoir une visualisation de l'ensemble des données et en tirer des conclusions?

Traitement de données abstraites

telles que des cours d'actions, statistiques, relations, etc. (par opposition à des données comme des mesures de température, de temps, etc.)

VISUALISER LES DONNÉES

La visualisation nous aide

- à découvrir des relations et des particularités
  discovery
- à analyser les données de manière exploratoire
  exploration
- à trouver des explications pour des échantillons, des lois scientifiques ou des caractéristiques
  explanation
- à prendre des décisions
  decision making
- à représenter visuellement des données abstraites
  communication

La visualisation permet par exemple d'identifier des valeurs extrêmes ou de découvrir des corrélations.

COMPARER ET RELIER LES DONNÉES

Use Case:

- Nous souhaitons comparer les données de divers dépôts.

Situation de base:

- Deux tableaux dans des formats différents (ex. fichier SPSS et fichier CSV)

Problématique:

- Différents standards de métadonnées, vocabulaires et modèles de données

INTEROPÉRABILITÉ

“ ... the ability of two or more systems or components to exchange information and to use the information that has been exchanged.” (http://en.wikipedia.org/wiki/Interoperability)

- Interopérabilité structurelle (modèle de données commun)
- Interopérabilité syntaxique (codification syntaxique commune)
- Interopérabilité sémantique (vocabulaire commun)

Source: http://www.kim-forum.org/Subsites/kim/DE/Materialien/Glossar/glossar_node.html

INTERPRÉTATION AUTOMATIQUE

Exemple:

- *Tableau 1 – auteur: Sir Tim Berners-Lee*
- *Tableau 2 – nom: Timothy John Berners-Lee*

Comment puis-je (ou mon ordinateur) savoir qu'il s'agit de la même personne?

Ou comment les contenus peuvent-ils être interprétés par des machines?

Sir Tim Berners-Lee

Berners-Lee in 2014.

Born	Timothy John Berners-Lee 8 June 1955 (age 60)[1] London, England
Institutions	World Wide Web Consortium University of Southampton Plessey MIT
Alma mater	University of Oxford (BA)
Notable awards	OM (2007) KBE (2004) FRS (2001)[2] FREng[*when?*] FRSA[*when?*] DFBCS (1995) See full list of honours
Spouse	Nancy Carlson (m. 1990) (divorced)[*when?*] Rosemary Leith (m. 2014)
Website	www.w3.org/People/Berners-Lee

Source: https://en.wikipedia.org/wiki/Tim_Berners-Lee

INTÉGRATION DES DONNÉES

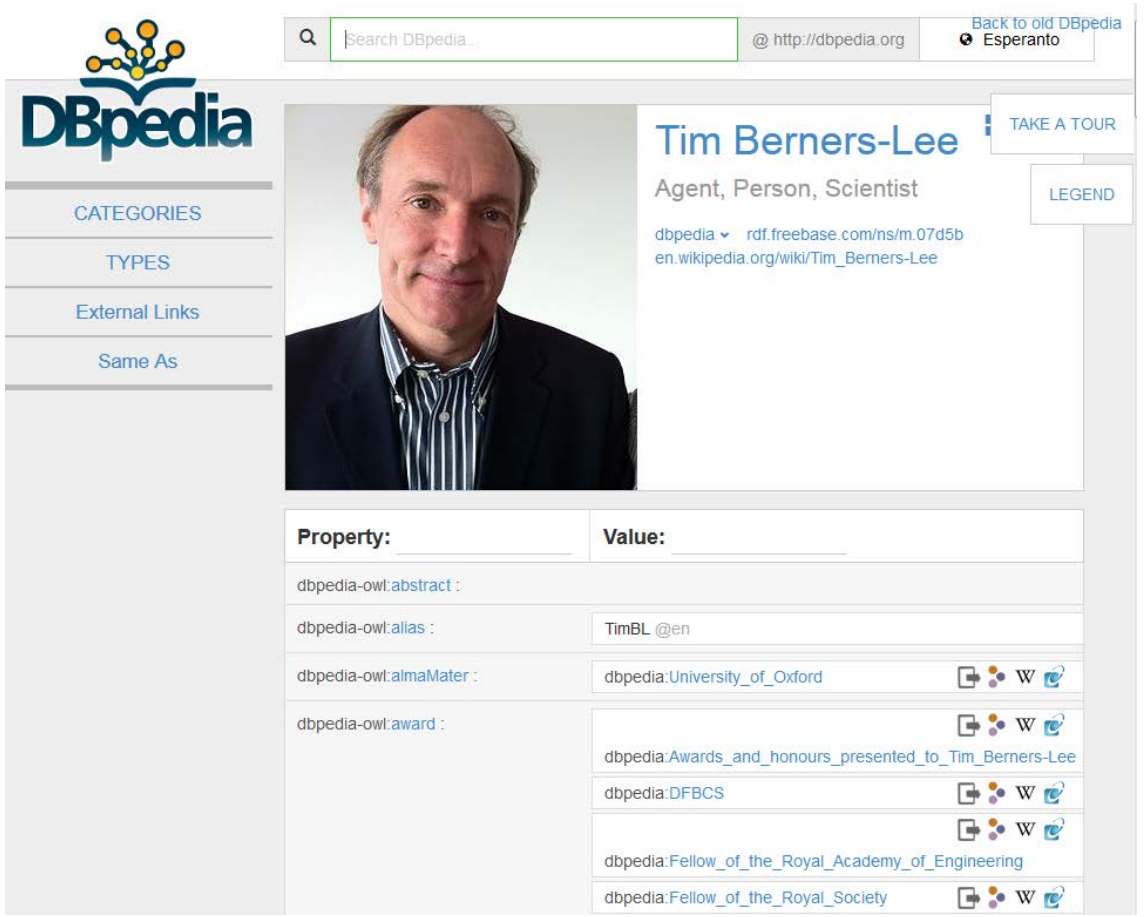
Solution possible:

- Linked Data (avec RDF en tant que modèle de données)

Nous relions des entités avec leurs descriptions et leurs relations à diverses sources.

Exemple DBpedia:

http://uk.dbpedia.org/page/Tim_Berners-Lee



LINKED DATA

- Modèle de données: RDF (Resource Description Framework)
- Sémantique: ontologies connues, telles que Dublin Core, DCAT, etc.
- Média: le web (HTTP)
- Langage d'interrogation: SPARQL
- Base de données: tripelstore RDF
- Interrogation des données: SPARQL endpoint
- Liens (URIs): en tant qu'identifiants
- Interprétation automatique

UN TRIPLET RDF

- Au lieu de décrire un document, je décris une seule information.

sujet	prédicat	objet
(entité/ressource)	(attribut)	(valeur de l'attribut)

UN TRIPLET RDF

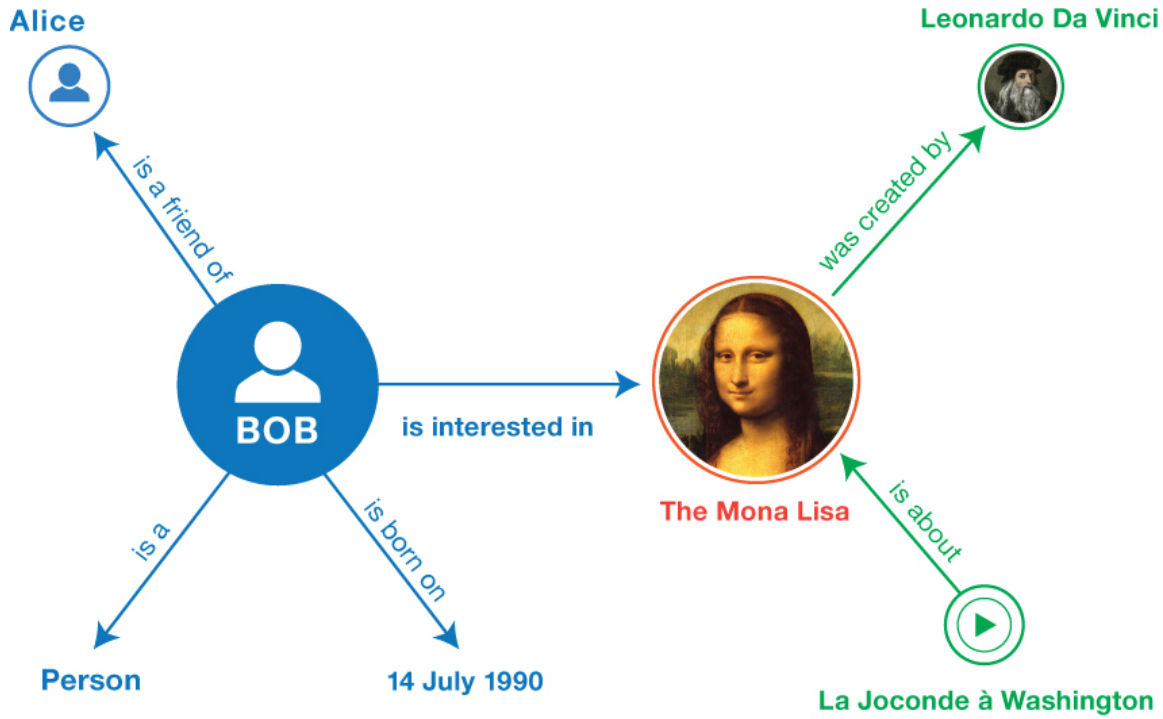
<The Mona Lisa> <was created by> <LeonardoDa Vinci>.

<Bob> <is interested in> <The Mona Lisa>.

<Bob> <is a> <person>.

<Bob> <is born on> <the 4th of July 1990>.

<Bob> <is a friend of> <Alice>.



Source: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/

GRAPHE RDF: RÉSEAU DE TRIPLETS RDF

<the Mona Lisa>

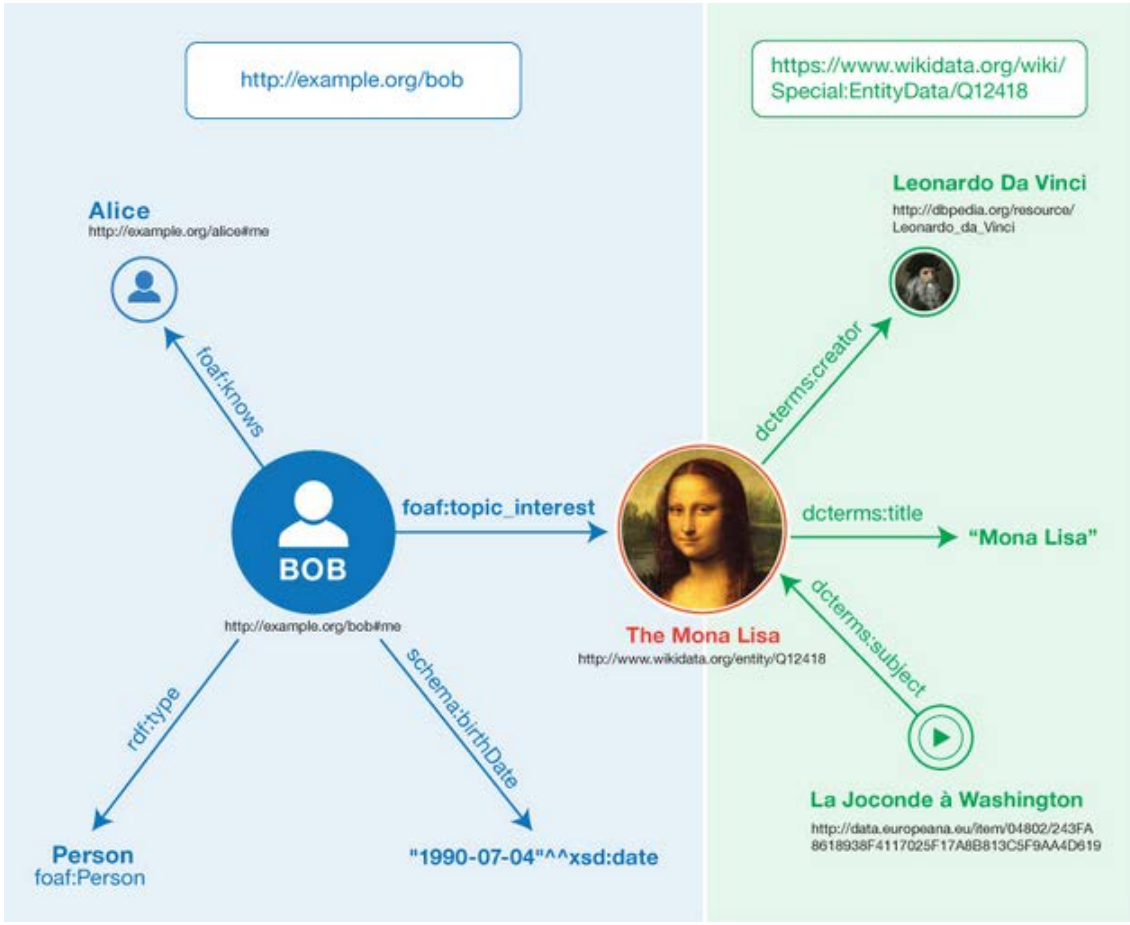
<http://www.europeana.eu/portal/record/03919/FCD38BDE7A03579F24BEDA5D157943B75BB36F11.html>

<was created by>

<http://purl.org/dc/terms/creator> (dcterms:creator)

<LeonardoDa Vinci>.

<http://dbpedia.org/resource/Leonardo_da_Vinci>



VOCABULAIRE RDF

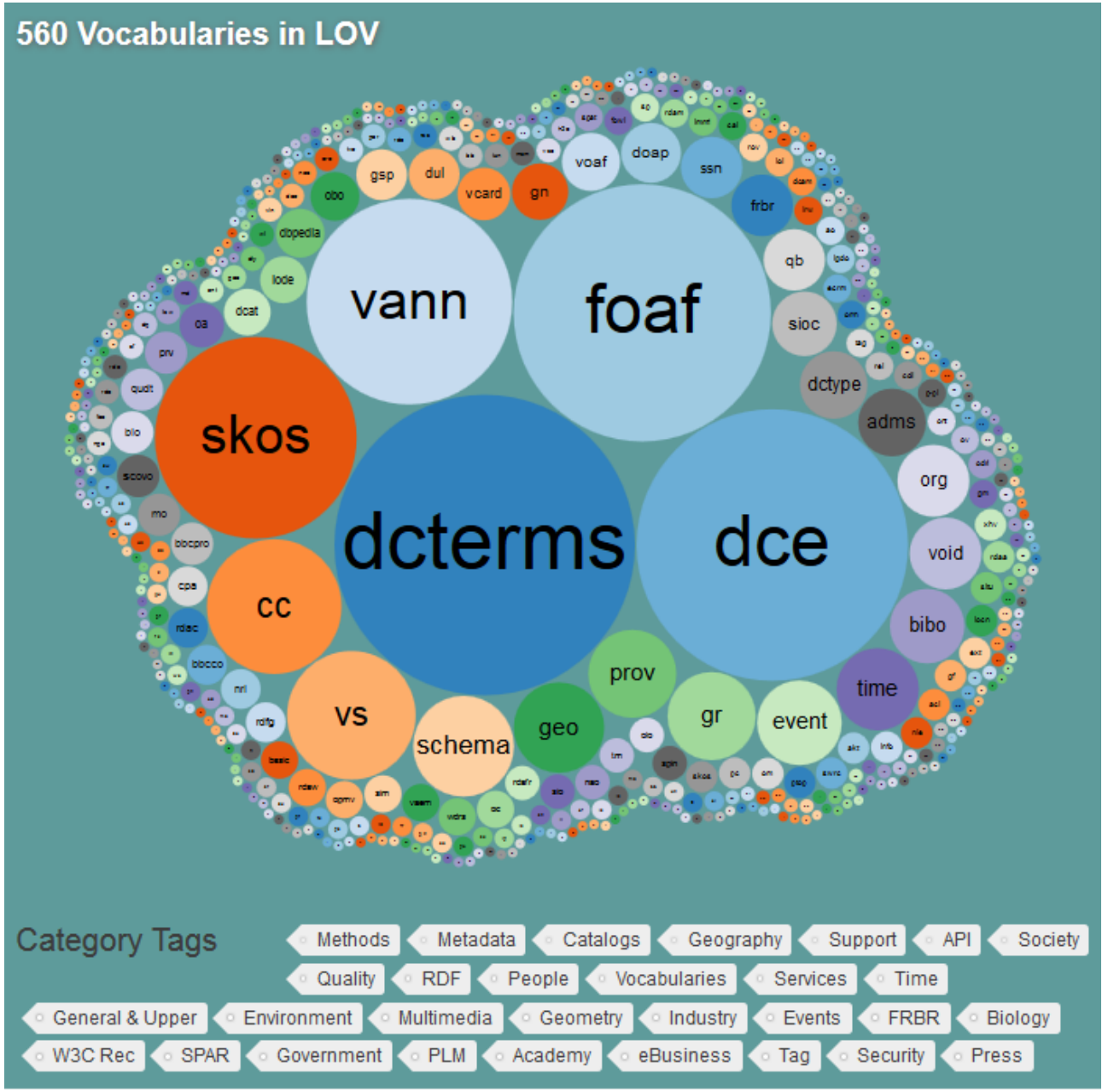
Il faut des règles pour représenter des triplets:

- Description des données au moyen d'ontologies
- Interopérabilité sémantique (vocabulaires communs)

Où chercher des vocabulaires?

Linked Open Vocabularies (LOV)

http://lov.okfn.org/dataset/lov/



‘REUSE OF VOCABULARY’

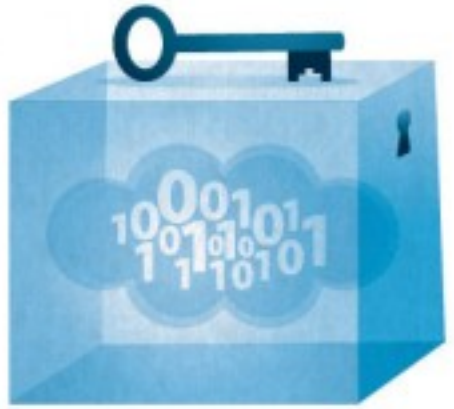
Exemple:

Archives Pina Bausch

Vokabular	Beschreibung	URL
Functional requirements for bibliographic records (FRBR)	Vokabular zur Organisation von Daten über Werke in Bibliotheken und Archiven	http://purl.org/vocab/frbr/core#
Simple knowledge organization system (SKOS)	Vokabular zur Organisation von Wissen, das typischerweise in Thesauren, Klassifikationsschemata oder Taxonomien enthalten ist	http://www.w3.org/2004/02/skos/core#
Friend of a friend (FOAF)	Vokabular zur Verlinkung von Personen und Informationen über sie	http://xmlns.com/foaf/0.1/
Gemeinsame Normdatei der Deutschen Nationalbibliothek (GND)	Beschreibung von Personen und Institutionen	http://d-nb.info/standards/elementset/gnd#
vCard	Standard zur Beschreibung von Visitenkarten	http://www.w3.org/2006/vcard/ns#
Dublin core (DC, DCT)	Beschreibung von Dokumenten	http://purl.org/dc/elements/1.1/, http://purl.org/dc/terms/
Media	Beschreibung von Multimedia-Dokumenten, baut auf DCT auf	http://purl.org/media#
Places	Beschreibung von Orten und ihren Relationen zueinander	http://purl.org/ontology/places#
Geonames (GN)	Vokabular mit offiziellen geographischen Namen	http://www.geonames.org

LANGAGE DE REQUÊTE SPARQL

SPARQL
Protocol
And
RDF
Query
Language

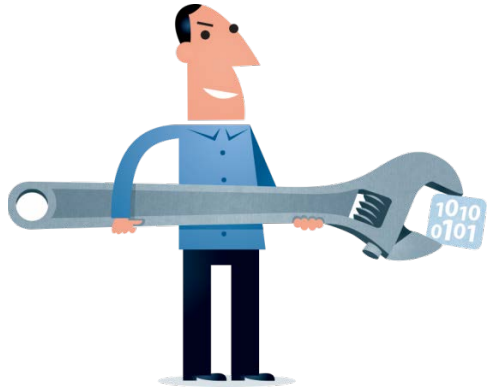


- Un standard du W3C
- Permet de réaliser des requêtes complexes
- Offre diverses fonctions, par ex. pour des traitements à la volée (on the fly)
- Supporte les Federated Queries et l'interconnexion à la volée

SPARQL QUERY

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX ex: <http://example.org/property/>

SELECT DISTINCT * WHERE {
  ?s ?p ?o
}
LIMIT 10
```



Les requêtes se font via des interfaces dites *SPARQL endpoints* (ex. de DBpedia: http://dbpedia.org/sparql).

RÉSUMÉ

Il est relativement simple de générer des triplets RDF et de les réutiliser

- Chercher des vocabulaires RDF déjà existants et en réutiliser si possible les termes
- Flexible et extensible – Réutilisable aussi en-dehors de ce cas d'application
- Les triplets RDF peuvent être reliés à diverses données externes.
- De bons outils sont disponibles – des APIs en RDF pour divers langage de programmation
- Outil d'intégration des données et triplestores RDF échelonnables

SPARQL est un langage de requête puissant pour RDF

- Semblable à SQL, certaines connaissances spécifiques étant néanmoins nécessaires
- Les ontologies et les schéma doivent être connus.
- Les Federated Queries et requêtes de SPARQL endpoints externes sont possibles.

RESSOURCES PÉDAGOGIQUES & OUTILS

- RDF 1.1 Primer:
  https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/
- SPARQL in 11 minutes:
  https://www.youtube.com/watch?v=FvGndkpa4K0
- Semantic University:
  http://www.cambridgesemantics.com/semantic-university/
- LOD Laundromat (entry point to a collection of existing data):
  http://lodlaundromat.org/

- Report on Available Linked Data Training Resources:
  https://www.loc.gov/aba/pcc/sct/documents/PCCSCTFinalReportonAvailableLinkedDataTrainingResources.docx
- Survey of Available RDF/Linked Data Creation Tools:
  https://wiki.duraspace.org/pages/viewpage.action?pageId=69014248