Web de données
et web sémantique

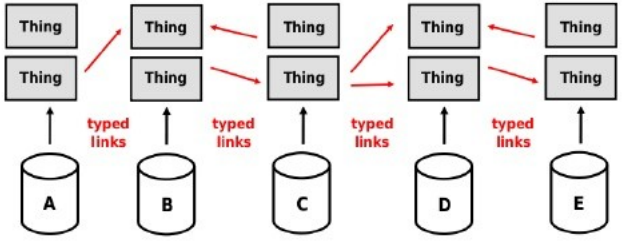
Bref historique du web

- Web 1.0
  - Web statique
  - 1 page = 1 document
  - Traitement des ressources limité à leur mise en forme
  - Navigation entre les pages avec les liens hypertexte

- Web 2.0
  - Exploitation de volumes importants d'informations (bases de données, moteurs de recherche)
  - Personnalisation de l’accès à l’information
  - Naissance du Web 2.0 = Web contributif, web social
    - Les utilisateurs font partie du processus documentaire
    - Ajout de connaissances et de commentaires aux contenu

Web de données

- Le **web de données** (ou **linked data**) est souvent considéré comme le web 3.0 :
  - "Rendre le contenu *(données)* indépendant du contenant *(sites web)*"
  - L'objectif est de considérer le web comme un ensemble d'entrepôts de données qui pourront être récupérées, réutilisées, enrichies



Tim Berners-Lee, principal inventeur du World Wide Web, a défini quatre piliers pour soutenir l'initiative « Web de données »...

Les 4 piliers du web de données

Pilier 1 : nommer les ressources avec des URI (Uniform Resource Identifier)

Toute "chose" qui possède une identité est une **ressource**.

- des "choses" numériques (ex : une page, un service, une image, une vidéo...)
- des "choses" physiques (ex : une personne, un bâtiment, un livre...)
- des concepts abstraits (ex : un nombre, la liberté, l'amour...)

Pour identifier des ressources on utilise la **syntaxe URI** (Uniform Resoure Identifier). Exemples :

- http://viaf.org/viaf/9847974/
- http://isni.org/isni/0000000121200982
- ark:/12148/cb11907966z
- http://dbpedia.org/resource/Victor_Hugo

Les 4 piliers du web de données

Pilier 2 : utiliser des URI HTTP qui existent sur le web

Une ressource n'est jamais manipulée directement, mais toujours à travers des **représentations** (pour la créer, la consulter, la modifier).

Une ressource peut avoir plusieurs représentations :

- Une personne pourra être représentée via sa photo, sa biographie ou son CV
- Un texte aura plusieurs traductions, plusieurs éditions

Les URI HTTP utilisées pour nommer les ressources sont donc **déréférençables**, c'est à dire qu'elles mènent à une représentation de la ressource.

- Exemple : http://viaf.org/viaf/9847974/ est l'URI déréférençable qui mène à une représentation de la ressource *Victor Hugo*, à savoir une notice d'autorité dans le référentiel VIAF.

Les 4 piliers du web de données

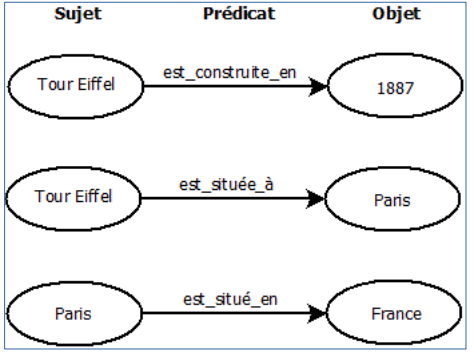
Pilier 3 : fournir à travers l’adresse URI des renseignements exploitables, lisibles par les humains et par les machines

On rajoute donc une couche pour « donner du sens » aux données : c’est le **web sémantique**

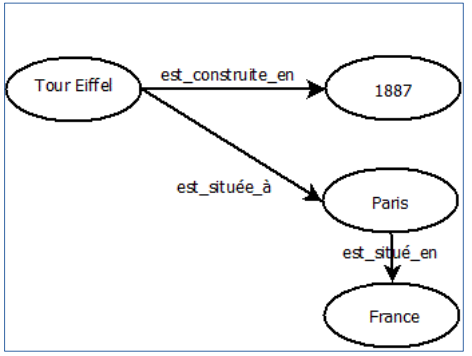
L’un des objectifs du web sémantique est de créer un internet « intelligent », avec des moteurs de recherche capables de gérer par exemple les homonymies (si je recherche « ballon », le moteur pourra différencier les résultats parlant de « football », de « montgolfière », de « ballon d’eau chaude », de « ballon de rouge » etc...)

Le web sémantique

- Basé sur des graphes de connaissance...
- ... eux-même basés sur des **triplets** "Sujet-Prédicat-Objet".
  - **Sujet** : ressource à décrire
  - **Prédicat** : type de propriété applicable à la ressource
  - **Objet** : valeur de la propriété



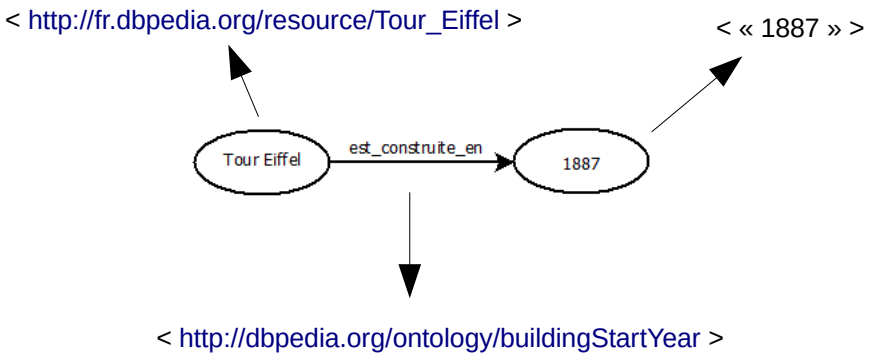
Ces trois triplets peuvent être représentés sous la forme d'un **graphe de connaissance**

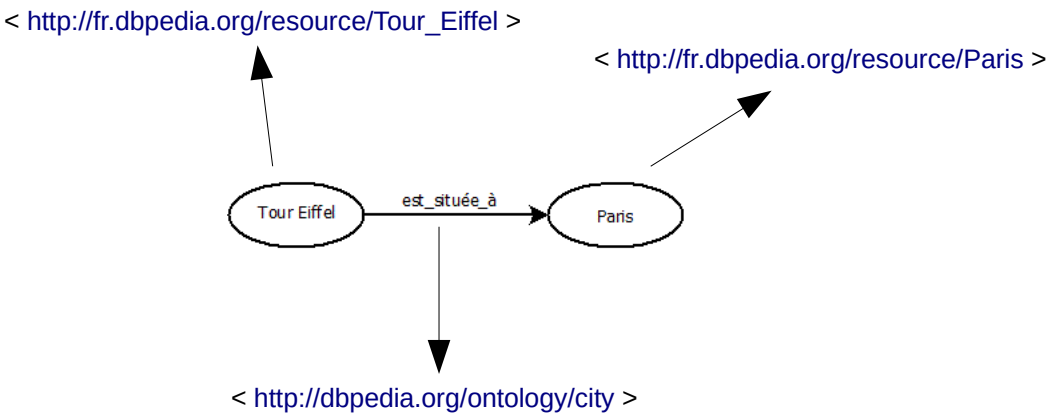


*On parle de **modèle RDF** (Resource Description Framework)*

Le web sémantique

- Les **sujets** sont des URI : http://fr.dbpedia.org/resource/Tour_Eiffel
- Les **prédicats** sont :
  - des URI : http://dbpedia.org/ontology/city
  - décrits par des **ontologies** (une ontologie est une modélisation d'un domaine de connaissance)
- Les **objets** sont
  - des URI : http://fr.dbpedia.org/resource/Paris
  - ou des littéraux





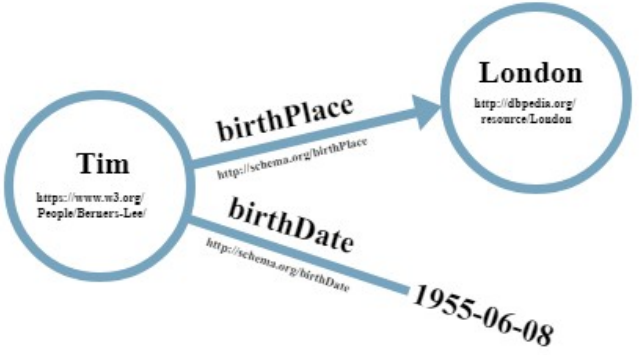
Retour aux 4 piliers du web de données

- **Pilier 4 : mailler l'adresse URI initiale en lui associant des adresses URI externes**

  En d'autres termes, faire pointer les arêtes de son graphe vers d'autres :

  - pour améliorer la découverte d'autres informations sur le Web
  - pour fournir du contexte
  - ...et ainsi créer un « graphe de graphes »

Le web de données envisage donc le web comme un unique et gigantesque graphe de connaissances, où les ressources sont identifiées de façon unique et liées entre elles grâce au modèle RDF.

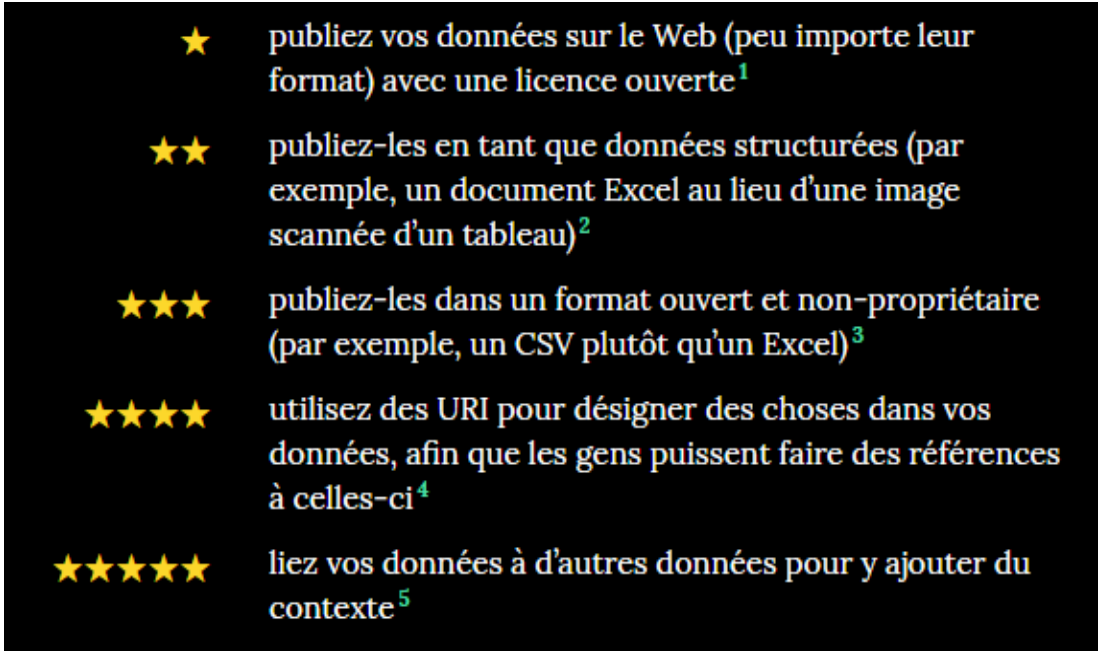


Open data

Open data : données ouvertes, dont l'accès est public et libre de droit, tout comme leur exploitation.

Tim Berners Lee propose une échelle de qualité basée sur 5 étoiles, pour évaluer jusqu'à quel point des données sont réutilisables.

L'idéal pour s'intégrer au Linked Open Data est donc d'atteindre les 5 étoiles...

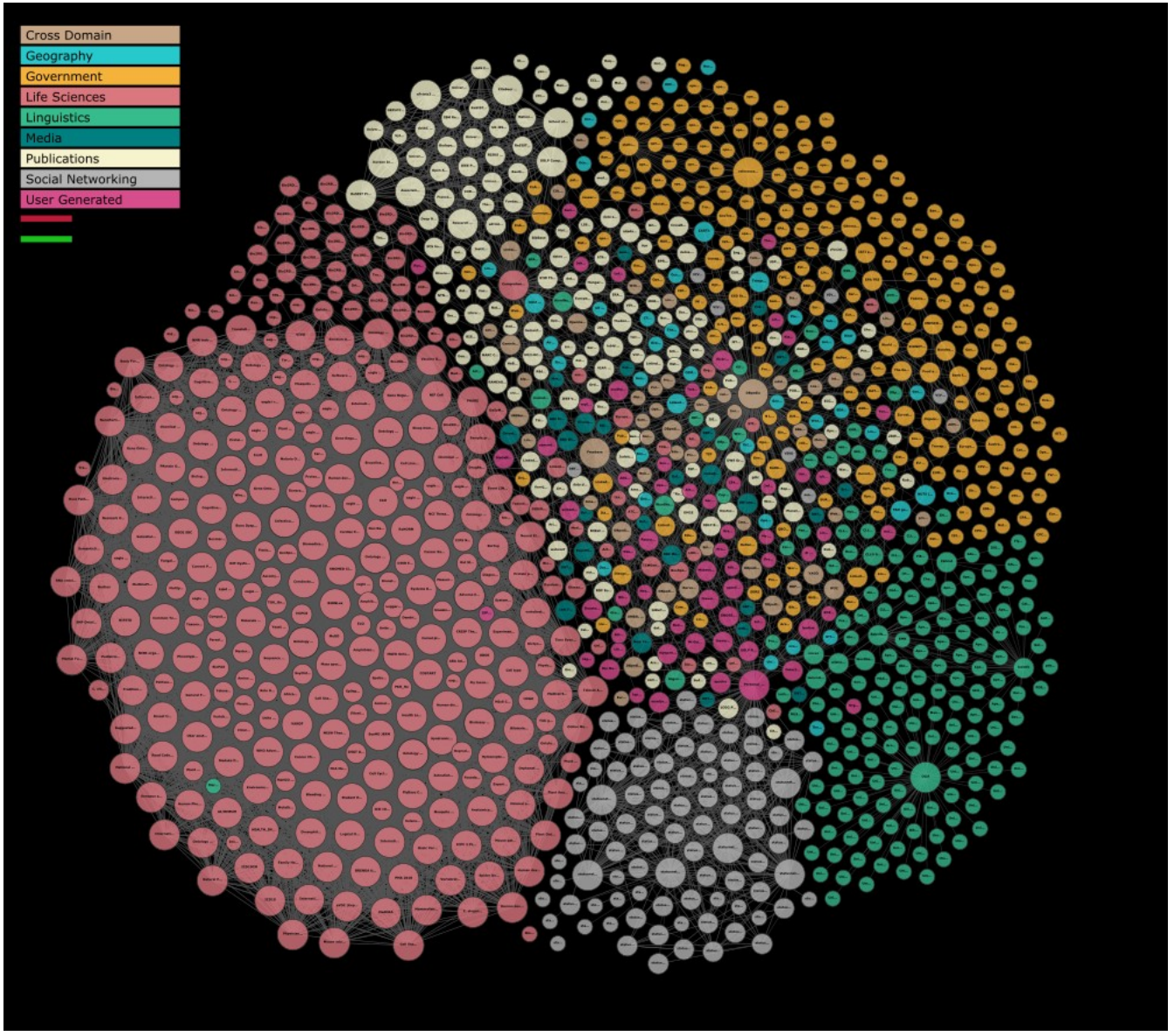


Linked data + open data

Si les données mises en oeuvres dans le web de données sont libres de toute licence, on parle alors de **Linked Open Data**.

Exemple : DBpedia, extraction des données de Wikipedia au format RDF

- 3 milliards de triplets
- téléchargeables librement
- Interrogeable via un **Sparql endpoint**
- 4,87 millions de liens vers des datasets externes



The Linked Open Data Cloud

https://lod-cloud.net/

Les référentiels

Définition : « Ensemble d'informations servant de références, parce qu'elles font autorité, ou parce qu'elles représentent un point de vue privilégié ou offrent une description stable d'une réalité. Un dictionnaire, une nomenclature, un système de coordonnées sont des référentiels. […] Plus généralement on appelle souvent référentiel un thésaurus vérifié et contrôlé permettant d'enrichir des données au sein d'un système d'information. » (glossaire DIGIT_HUM)

L'alignement avec des référentiels pour

- ajouter du contexte
- ouvrir ses données
- enrichir les données d'un projet (en les complétant avec d'autres ressources)

Nombreux référentiels existants : Data.bnf.fr, idRef, ISNI (enseignement supérieur et recherche), VIAF (personnes), Rameau (indexation matière de la BNF), Pactols (archéologie), etc...