Alignement et import de données

Simon Chabot (GI03)

Logilab

2 sites

Paris & Toulouse

3 spécialisations

Web Sémantique et gestion de données
Simulation numérique
Outils & Systèmes

CubicWeb

Plateforme pour le Web Sémantique et la gestion de connaissances

web sémantique	OWL, RDF
multi-source	SQL, LDAP, Mercurial
interrogeable	RQL

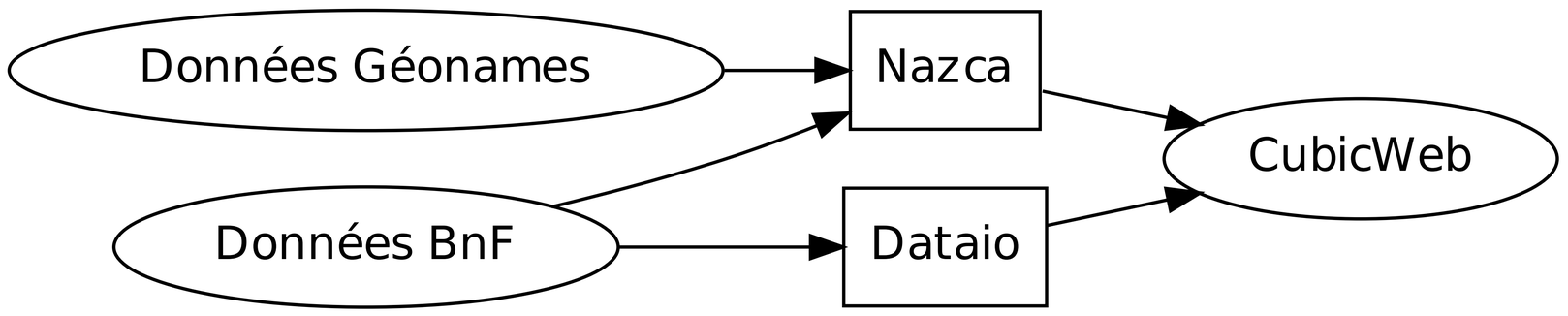
Problématique du stage (par un exemple)

BnF Bibliothèque nationale de France ⟶ 250 mille entités (“villes” uniquement)

Géonames Base de données géographique ⟶ 7 millions d’entités

Deux questions fondamentales :

- Comment importer des données massivement ? ⟶ *dataio*
- Comment enrichir ces données ? ⟶ *Nazca*



Nazca – L’enrichissement de données

Données à aligner	
id	label
uri_0	Paris
uri_1	Compiègne
uri_2	Londres

Données cibles	
id	label
uri_a	Paris
uri_b	Londres
uri_c	New-York

Nazca

Résulats			
id source	label	id cible	label
uri_2	Londres	uri_b	Londres
uri_0	Paris	uri_a	Paris

normalisation	stopwords lemmatization
recherche de voisins	arbre-kd minhashing
calcul de distances	Levensthein Temporelle Géographique

Taux d’alignement BnF / Géonames : **83.4%**

dataio – Import massif

Travailler le plus possible en SQL

- Importer des données massivement grâce à des requêtes SQL
- Établir des relations entre les données importées
- Insérer les métadonnées nécessaires à *CubicWeb*

Temps d’import de Géonames ≈ **80 minutes**

Outils utilisés

python™ CubicWeb mercurial PostgreSQL

Améliorations futures

Nazca

parallélisation normalisation / distance entre voisins

recherche de voisins ajouter d’autres méthodes

dataio

parallélisation création des méta-données CubicWeb

format d’entrées support différents formats d’entrée