De la description des documents à l'exploitation des données : data.bnf.fr

Journée Open Data Aix-en Provence

19 avril 2012

Agnès Simon (Bibliothèque nationale de France)

Ce qui se passe:

- moins de passage par les pages d'accueil

accès direct aux pages du site

- recherches par mots-clés (recherche d'un titre par exemple)

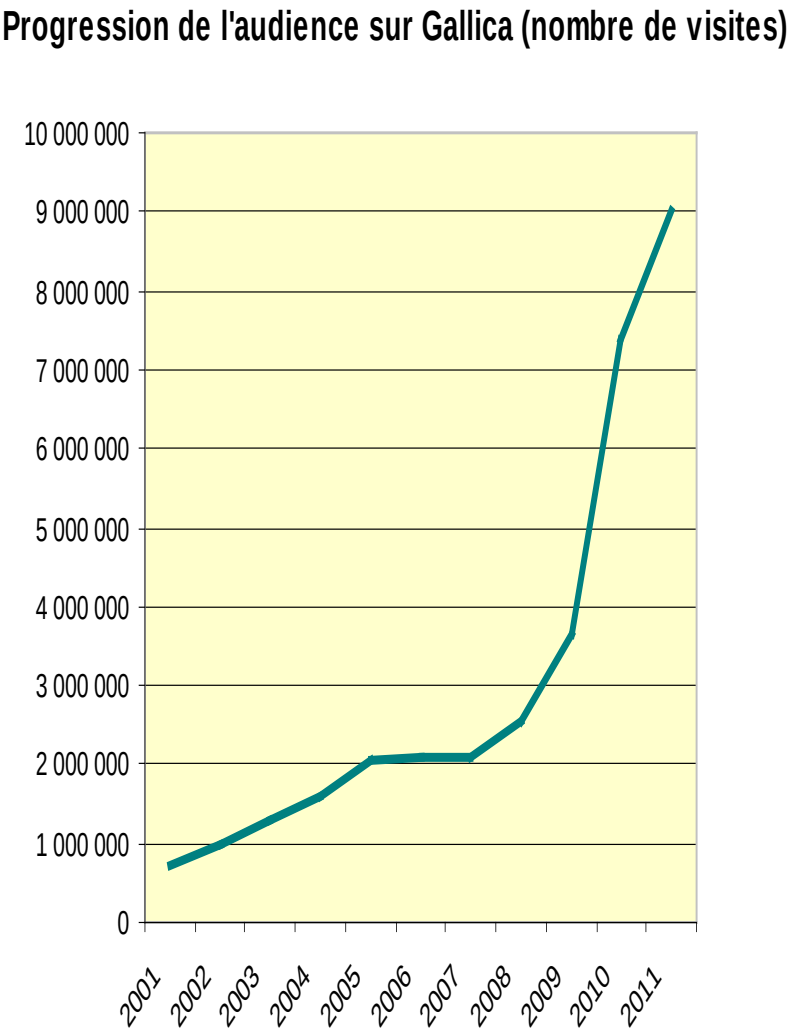
- suivi de liens

Source gallica.bnf.fr / Bibliothèque nationale de France

Environnement favorable

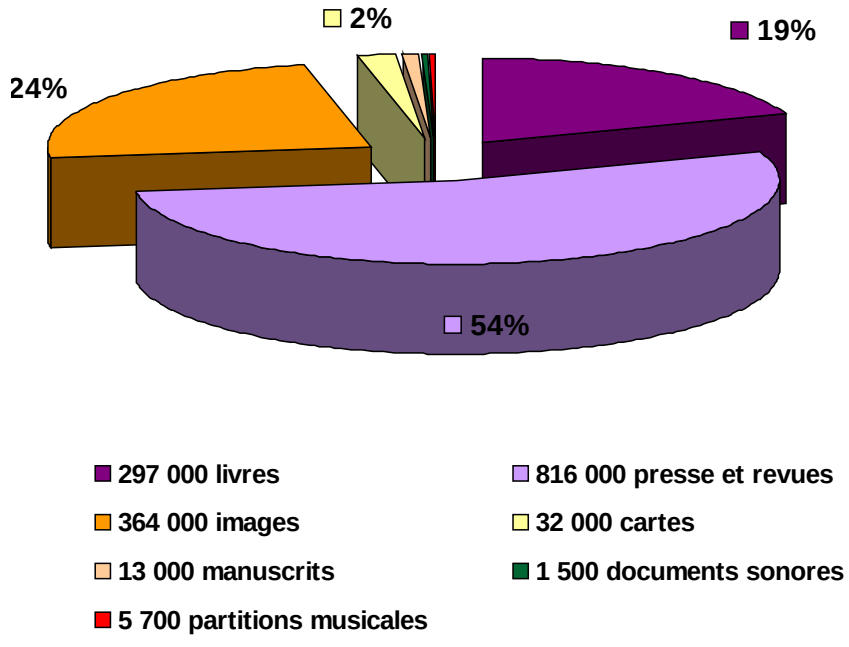
- Audience en croissance rapide
- Public demandeur de documents
- Interlocuteurs demandeurs de données

28 000 visites/jour



Les données publiques culturelles disponibles à la BnF

Les documents Gallica : 1,5 millions d'objets



Les métadonnées

- 15 millions de notices bibliographiques
- 5,1 millions de notices d'autorité sur les personnes, les œuvres, les sujets
- Des ressources spécialisées : archives et manuscrits

gallica
BIBLIOTHÈQUE NUMÉRIQUE

Des ressources difficiles à trouver

- Structures de silos :
  - Séparés et indépendant
  - Non interopérables

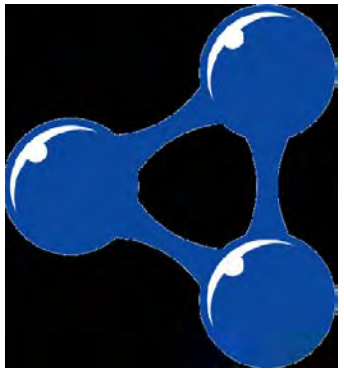
- Bases non indexables par les machines
- « Web profond » ou « Web caché »

data.bnf.fr

- Objectifs du projet :
  - Mettre les ressources de la BnF dans **l'écosystème du Web**
  - Fédérer les informations pour une **navigation plus intuitive** en développant **l'interopérabilité** entre les données
  - Favoriser la **réutilisation** de données
  - Imaginer de nouveaux **services** pour les bibliothèques et d'autres communautés d'utilisateurs

Organiser l'information

- Plus de 200 000 pages et plus de 2 millions de documents reliés.
- Regroupe et organise des données sur les **œuvres,** les **auteurs** et les **thèmes**
- Regrouper les contenus, liens et services
  - Des pages pour les humains
  - Des données pour les machines



Data.bnf.fr

- page *Christine de Pisan*
- page *Les misérables*
- page *Web sémantique*

Comment ça marche ?

Les documents numérisés

gallica BIBLIOTHÈQUE NUMÉRIQUE

BnF Archives et Manuscrits

BnF Catalogue général

BnF catalogue général

data.bnf.fr

Regroupement - Alignement

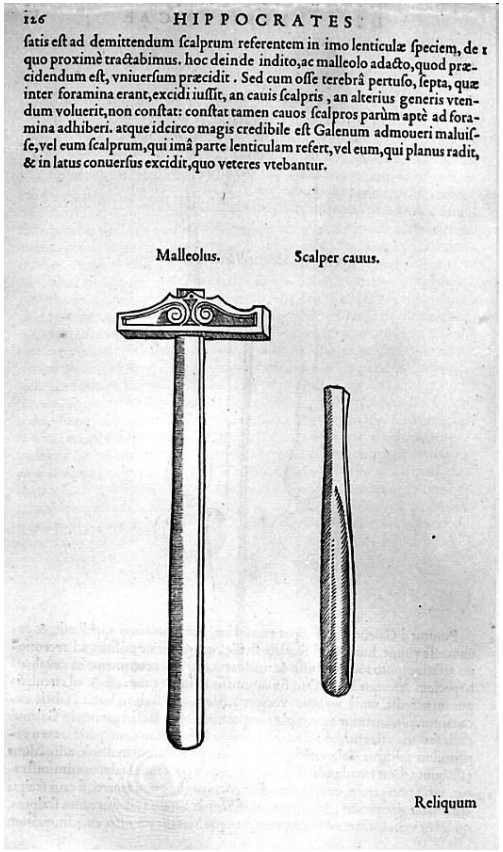
Pages Web pour les lecteurs

Données pour les machines

Organiser l'information autour de concepts

En se basant sur des **outils existants** :

- Identifiants ARK
- Des notice d'autorité
- Des standards de description
- Des techniques d'alignement



Outils du Web sémantique

Objectif :

- Aligner et regrouper des données
- Relier des données sur le Web

Principes :

- Des URI pour les ressources
- Des URI http
- Des informations (URI actionnables)
- Un cadre de description (RDF)

Données brutes en RDF

Plus de 6 millions de « triples RDF » disponibles au téléchargement.

```
- <skos:editorialNote>
    Correspondance / Abélard et Héloïse ; texte traduit [du latin] et présenté par Paul Zumthor, 1979
  </skos:editorialNote>
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <skos:prefLabel xml:lang="fi">Pierre Abélard (1079-1142)</skos:prefLabel>
  <skos:altLabel>Pierres Abaelart (1079-1142)</skos:altLabel>
  <skos:altLabel>Peter Abelard (1079-1142)</skos:altLabel>
  <skos:altLabel xml:lang="la">Petrus Abelardus (1079-1142)</skos:altLabel>
  <skos:altLabel>Peter Abaelard (1079-1142)</skos:altLabel>
  <skos:altLabel xml:lang="la">Petrus Abaelardus (1079-1142)</skos:altLabel>
  <skos:altLabel xml:lang="it">Pietro Abelardo (1079-1142)</skos:altLabel>
  <skos:altLabel>Pierre Abailard (1079-1142)</skos:altLabel>
  <skos:altLabel>Peter Abailard (1079-1142)</skos:altLabel>
  <skos:altLabel>Abélard (1079-1142)</skos:altLabel>
  <rdfs:seeAlso rdf:resource="http://catalogue.bnf.fr/ark:/12148/cb11885557w"/>
</rdf:Description>
```

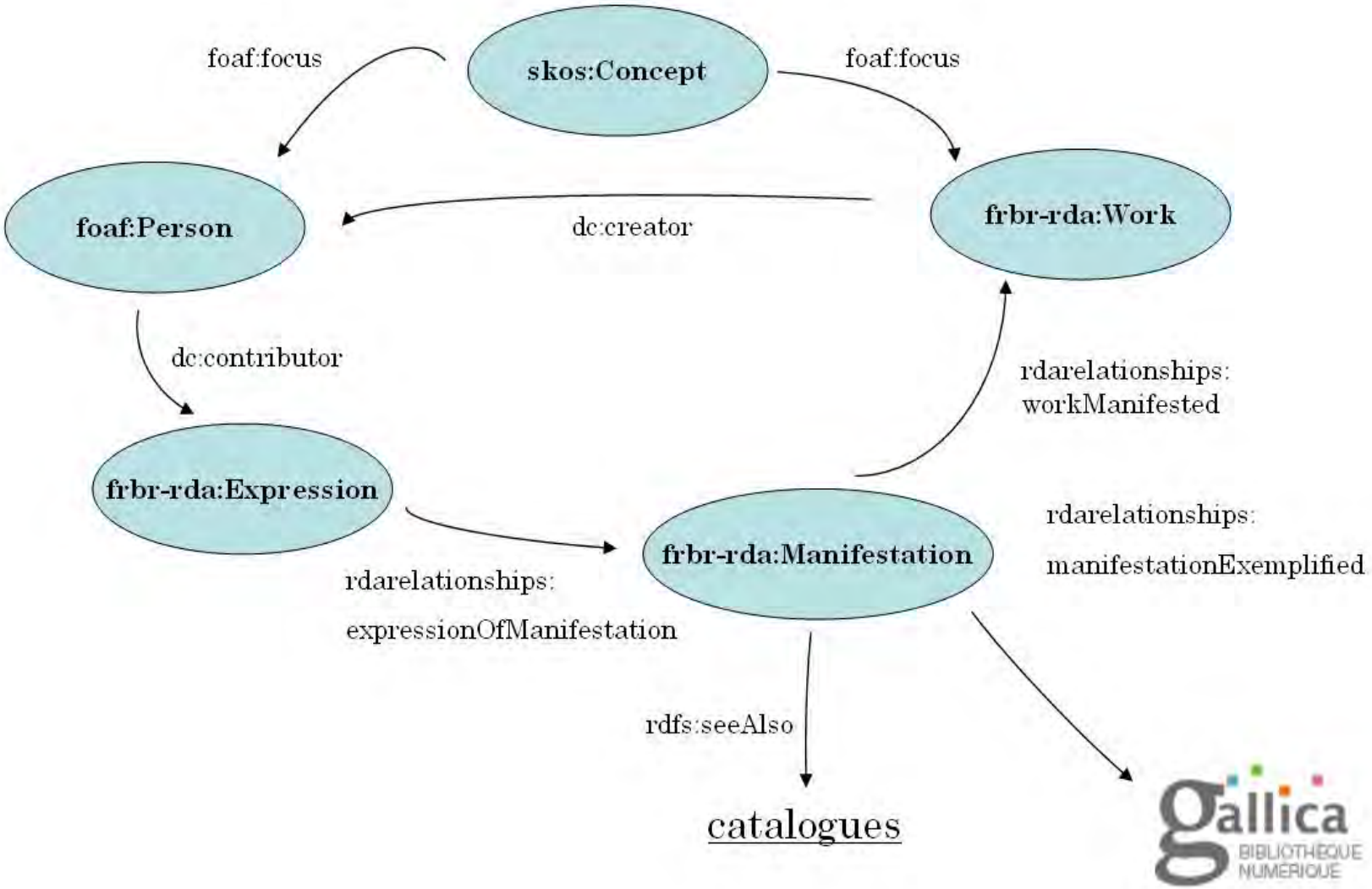
Les données brutes en RDF

Pas de « notices », seulement des données

- SKOS pour les concepts
- FOAF pour les auteurs
- RDA pour les œuvres

Source gallica.bnf.fr / Bibliothèque nationale de France

Modèle RDF simplifié



Triple ouverture

- **Accès technique** aux données

  Téléchargement RDF à la volée et de la base entière (dump)

- **Ouverture juridique**

Open data avec licence ouverte de l'Etat (licence d'attribution)

Présence sur ***data.gouv.fr***

- **Logiciel libre**

LICENCE OUVERTE
OPEN LICENCE

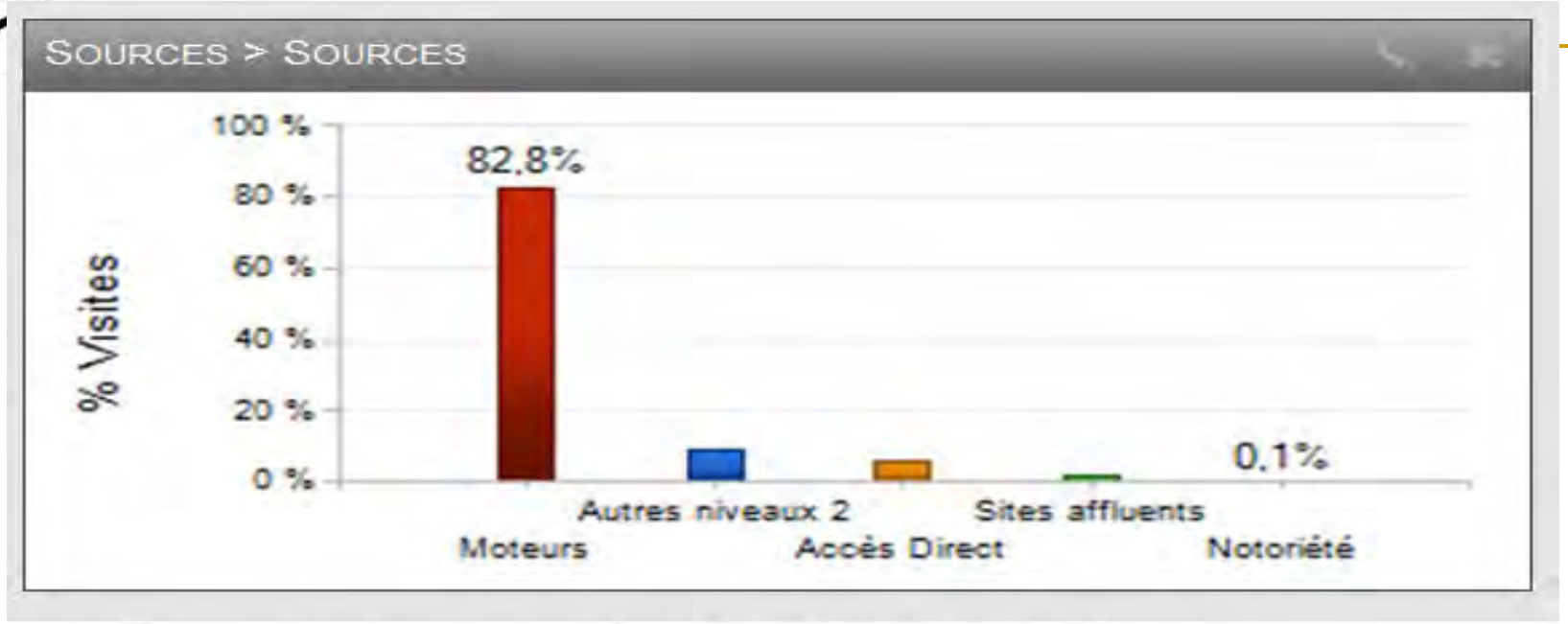
CubicWeb

Corpus :

- **Un socle classique au lancement (Juillet 2011) : les plus *attendus***

- **Ajout de corpus complets et numérisés: les plus *demandés***
  - Auteurs antiques
  - Juristes anciens
  - Musiciens essentiellement français

- **A venir : intégration des auteurs et œuvres *les plus rares* et pour lesquels la BnF apporte une forte valeur ajoutée sur le Web: « la longue traîne »**

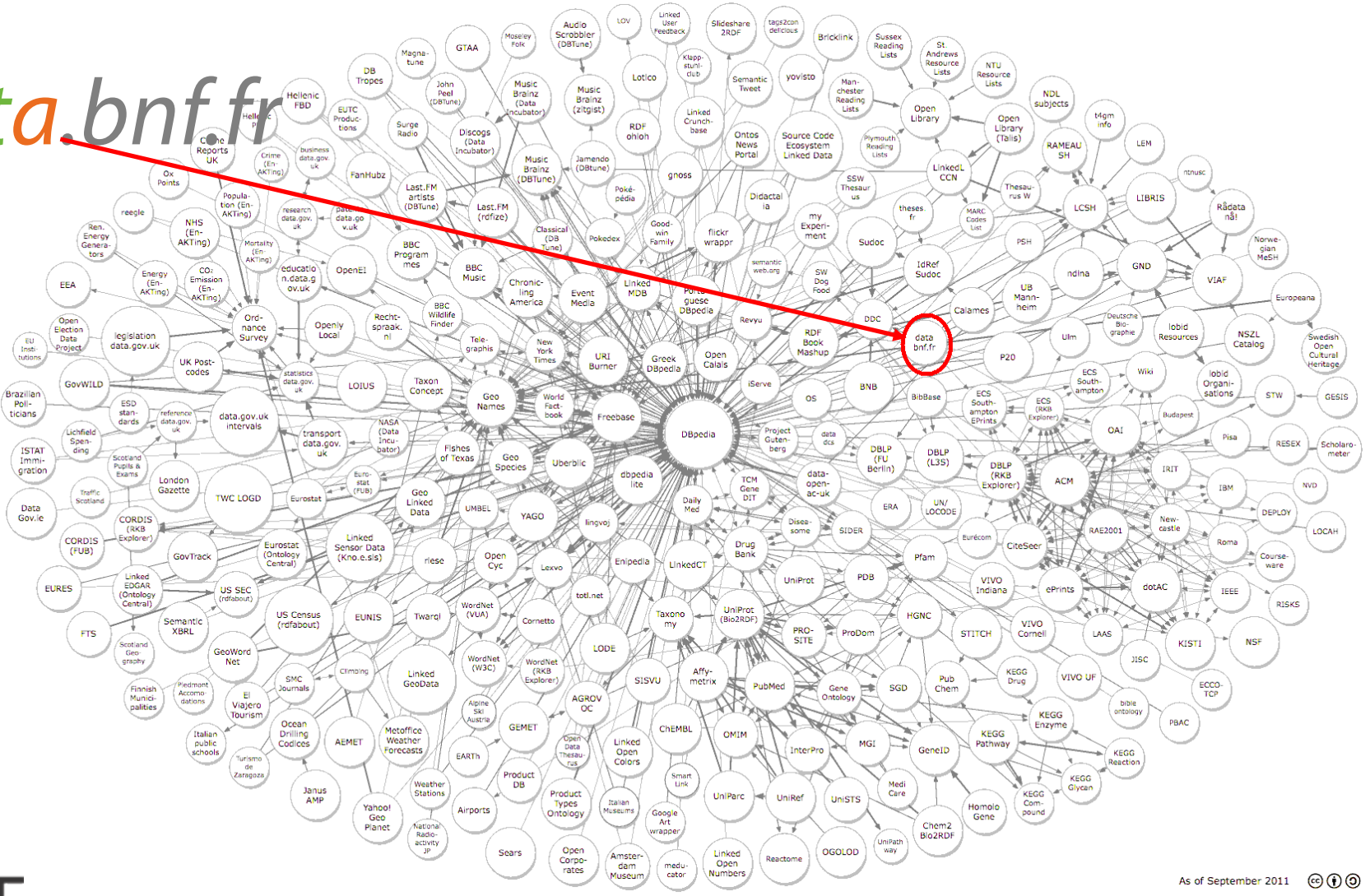


Sources des visites de data.bnf.fr en mars 2012

80 % viennent depuis les moteurs de recherche

Autres provenances : Wikipedia et ***data.gouv.fr***

dans le « Linked Open data »



Linking Open Data cloud diagram, http://lod-cloud.net/

Merci de votre attention

Contacts :

agnes.simon@bnf.fr

romain.wenz@bnf.fr

data.bnf.fr