Ressources documentaires et web des donnés: the missing link

Jérôme Euzenat

Inria & LIG
Montbonnot, France
Jerome.Euzenat@inria.fr
http://exmo.inrialpes.fr

March 28, 2012

Outline

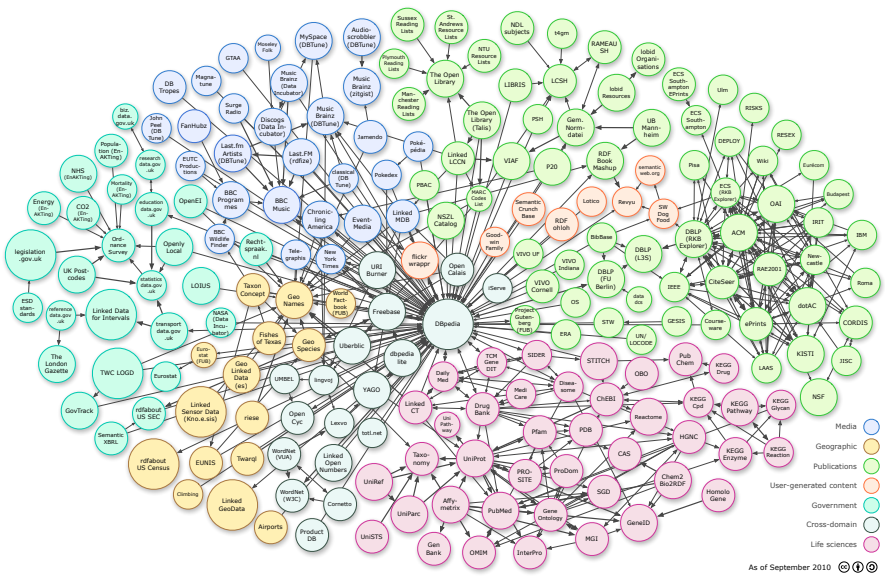
Introduction au web des données

Web des données et documentation

Quelques contributions INRIA

Conclusions

Linked (open) data cloud



Publier des données sur le web

Quatre principes

1. Utiliser des URIs pour identifier les ressources
2. Utiliser des URIs déréférençables
3. Lorsqu'une URI est déréférencée, une description de la ressource doit être retournée
4. Les données publiées doivent être connectées à d'autres jeux de données

Si possible, utiliser les technologies du web sémantique: URI, HTTP, RDF, OWL, SPARQL

Pourquoi?

- Si vous êtes une autorité disposant de données (gov), cela permet de les publier et de les connecter aux données d'autres autorités;
- Si vous êtes quelqu'un qui sait comment lier les données, vous disposez d'entités de référence pour le faire;
- Si vous êtes un développeur d'applications, vous pouvez utiliser ces données disponibles sur le web.

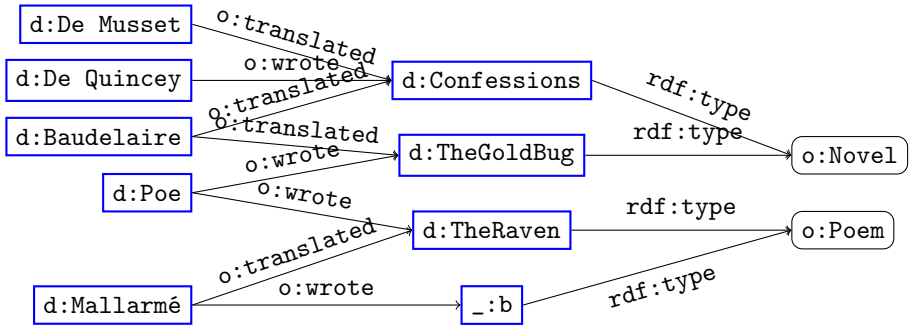
et

- Les données peuvent être constamment à jour;
- Les données peuvent être liées librement;
- On peut naviguer dans les données.

See also
`http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/`

Un peu de données liées

Baudelaire a traduit la nouvelle "The gold bug" d'Edgar Poe; Mallarmé a traduit le poème "The raven" du même Poe et écrit un poème.



Que peut-on trouver à la BNF?

Des URI

`http://data.bnf.fr/11890582/charles_baudelaire/`
`http://data.bnf.fr/11914131/stephane_mallarme/`
`http://data.bnf.fr/11917290/alfred_de_musset/`
`http://data.bnf.fr/11920094/edgar_allan_poe/`
`http://data.bnf.fr/11898910/thomas_de_quincey/`
`http://data.bnf.fr/12038661/edgar_allan_poe_le_corbeau/`
`http://data.bnf.fr/12144801/edgar_allan_poe_le_scarabe_d_or/`

(pourquoi De Quincey est classé à D, De Cervantes à C et De Musset à M?)

D'autres URIs

`http://dbpedia.org/resource/Charles_Baudelaire`
`http://dbpedia.org/resource/Stéphane_Mallarmé`
`http://dbpedia.org/resource/Alfred_de_Musset`
`http://dbpedia.org/resource/Edgar_Allan_Poe`
`http://dbpedia.org/resource/Thomas_De_Quincey`
`http://dbpedia.org/resource/The_Raven`
`http://dbpedia.org/resource/The_Gold-Bug`

Des URI dé-ré-fé-ren-ça-bles

```
curl -H "Accept:application/rdf+xml"
http://data.bnf.fr/11890582/charles_baudelaire/

curl -H "Accept:application/rdf+n3"
http://data.bnf.fr/11898910/thomas_de_quincey/

curl -H "Accept:text/rdf+n3"
http://dbpedia.org/resource/Thomas_De_Quincey

curl -H "Accept:application/turtle"
http://dbpedia.org/resource/The_Raven
```

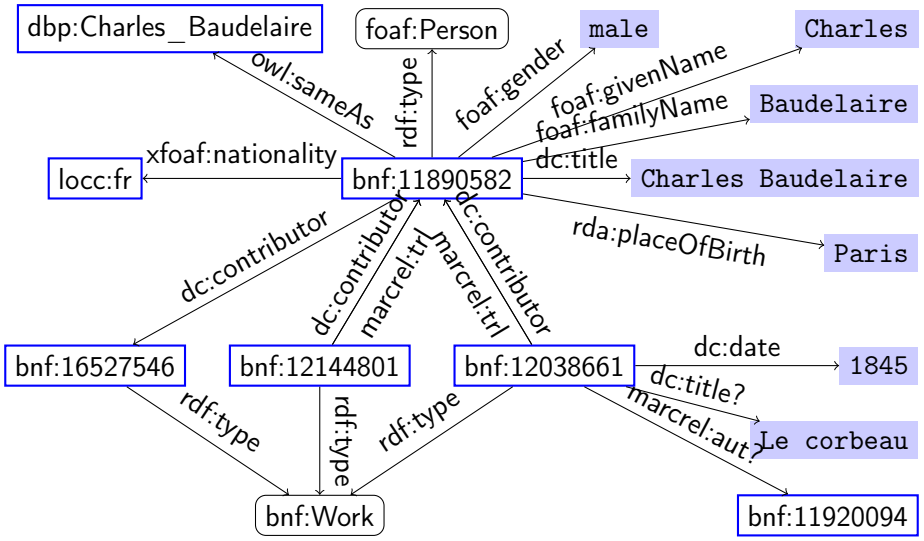

Baudelaire à la BNF

```
<rdf:Description rdf:about="http://data.bnf.fr/11890582#foaf:Person">
  <rdagroup2elements:dateOfDeath>31-08-1867</rdagroup2elements:dateOfDeath>
  <dc:date>1821-1867</dc:date>
  <foaf:gender>male</foaf:gender>
  <dc:contributor rdf:resource="http://data.bnf.fr/13911358#frbr:Work"/>
  <dc:contributor rdf:resource="http://data.bnf.fr/15708067#frbr:Work"/>
  <dc:contributor rdf:resource="http://data.bnf.fr/16527546#frbr:Work"/>
  <dc:title xml:lang="fr">Charles Baudelaire</dc:title>
  <rdagroup2elements:fieldOfActivityOfThePerson>Litt<C3><A9>ratures</rdagro
  <rdagroup2elements:fieldOfActivityOfThePerson rdf:resource="http://dewey
  <owl:sameAs rdf:resource="http://dbpedia.org/resource/Charles_Baudelaire"
  <rdagroup2elements:placeOfBirth>Paris</rdagroup2elements:placeOfBirth>
  <rdagroup2elements:dateOfBirth>09-04-1821</rdagroup2elements:dateOfBirth>
  <foaf:page rdf:resource="http://data.bnf.fr/11890582/charles_baudelaire/"
  <rdagroup2elements:languageOfThePerson rdf:resource="http://id.loc.gov/vo
  <foaf:familyName>Baudelaire</foaf:familyName>
  <foaf:givenName>Charles</foaf:givenName>
  <rdagroup2elements:biographicalInformation>Po<C3><A8>te. - Critique
d'art. - Traducteur d'Edgar Allan Poe. - Signa ses premi<C3><A8>res
```

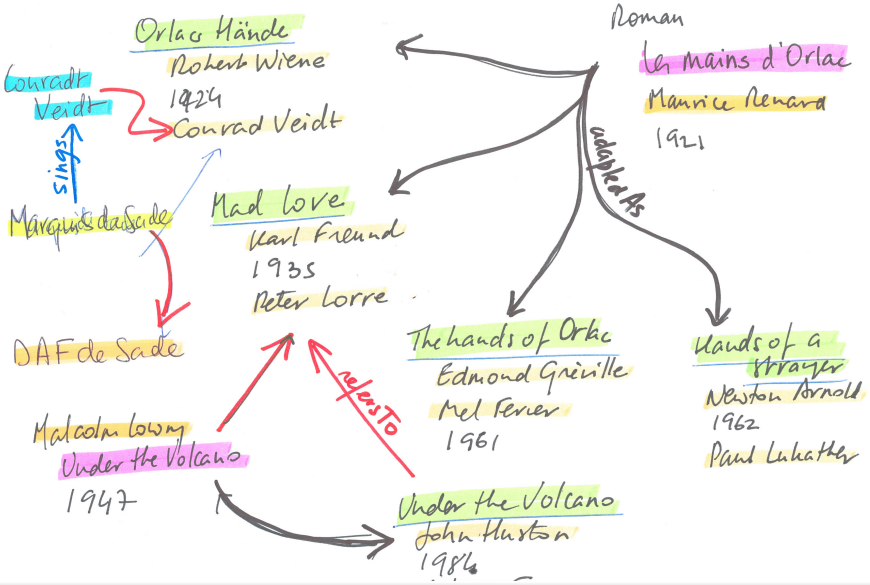


```
  <foaf:depiction rdf:resource="http://gallica.bnf.fr/ark:/12148/btv1b69033
  <foaf:depiction rdf:resource="http://upload.wikimedia.org/wikipedia/commo
```


Graphe



Plus de graphe



Triplets

sujet	prédicat	objet
bnf:11890582,	rdf:type,	foaf:Person
bnf:11890582,	foaf:gender,	"Male"
bnf:11890582,	dc:title,	"Charles Baudelaire"
bnf:11890582,	owl:sameAs,	dbp:Charles_Baudelaire
bnf:11890582,	xfoaf:nationality,	locc:fr
bnf:11890582,	rda:placeOfBirth,	"Paris"
bnf:12038661,	rdf:type,	bnf:Work
bnf:12038661,	dc:date,	"1845"
bnf:12038661,	marcrel:trl,	bnf:11890582
bnf:12038661,	dc:contributor,	bnf:11890582
bnf:12038661,	marcrel:aut,	bnf:11920094
bnf:12038661,	dc:title,	"Le corbeau"
	...	

Connection avec d'autres resources

http://data.bnf.fr/11895774/miguel_de_cervantes_saavedra/

owl:sameAs

http://datos.bne.es/page/resource/XX1718747

http://valdemoro.dia.fi.upm.es/bne/api/persons?tieneNombreDePersona=Cervantes+Saavedra%2C+Miguel+de

Requêtes

- Les jeux de données sont souvent disponibles à travers un accès SPARQL;
- SPARQL est un langage de requête pour RDF;
- Il fonctionne à l'aide de petits graphes contenant des variables (graph patterns);
- Que l'on peut projeter dans le graphe des données.

SPARQL

```
SELECT ?x, ?y
FROM http://data.bnf.fr
WHERE {
    ?z rdf:type o:Novel .
    ?z marcrel:aut ?x .
    ?z marcrel:trl ?y .
    ?z dc:date ?a .
    FILTER { 1800 < ?a && ?a < 1900 }
}
```

retournera:

⟨bnf:Poe, bnf:Baudelaire⟩
⟨bnf:DeQuincey, bnf:Baudelaire⟩
⟨bnf:DeQuincey, bnf:DeMusset⟩
...

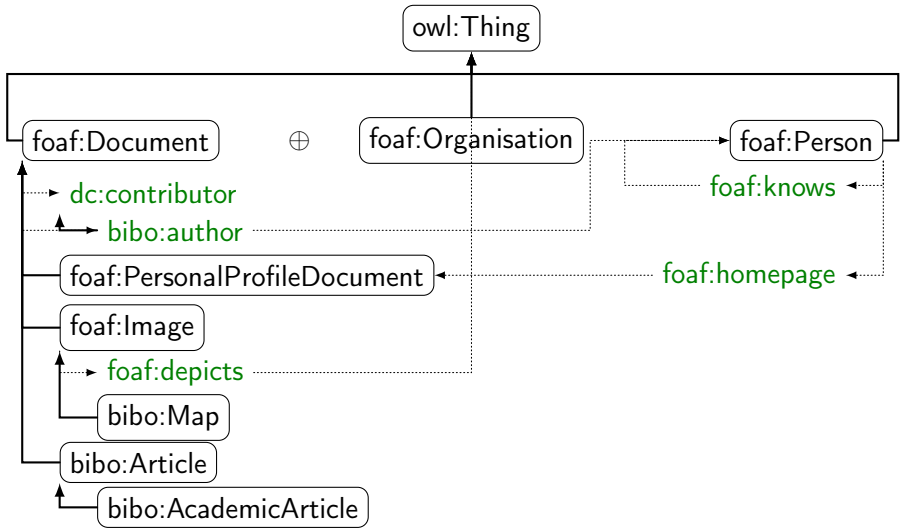
Ontologies? Quiz!

- ☐ RDF est une ontologie;
- ☐ OWL est une ontologie;
- ☐ SKOS est une ontologie;
- ☐ La classification de Dewey est une ontologie;
- ☐ MARC est une ontologie;
- ☐ La liste des codes ISO de pays est une ontologie;
- ☐ Rameau est une ontologie;
- ☐ Le catalogue de la BNF est une ontologie;
- ☐ Le fonds de la BNF est une ontologie.

Ontologies?

- Pour moi, une ontologie c'est une théorie logique. Ni plus, ni moins.
- MARC est une ontologie; Rameau est une ontologie; SKOS est une ontologie;
- Votre catalogue n'est pas une ontologie (quoique);
- En ce qui concerne RDF, une ontologie est quelque chose qui définit, d'une manière ou d'une autre, le vocabulaire

Bibliographic ontology



Contraintes d'ontologies

cites = citedBy^{-1}

```
<owl:Class rdf:about="#Biography">
  <rdfs:equivalentClass>
    <owl:Class>
      <owl:and rdf:parseType="Collection">
        <owl:Class rdf:resource="#Book"">
        <owl:Restriction>
          <owl:onProperty rdf:resource="#topic" />
          <owl:allValuesFrom rdf:resource="foaf:Person" />
        </owl:Restriction>
      </owl:and>
    </owl:Class>
  </rdfs:equivalentClass>
</owl:Class>
```

Les ontologies peuvent être déréférençables

- xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
- xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
- xmlns:owl="http://www.w3.org/2002/07/owl#"
- xmlns:skos="http://www.w3.org/2004/02/skos/core#"
- xmlns:foaf="http://xmlns.com/foaf/0.1/"
- xmlns:xfoaf="http://www.foafrealm.org/xfoaf/0.1/"
- xmlns:dc="http://purl.org/dc/terms/"
- xmlns:rdagroup2elements="http://RDVocab.info/ElementsGr2/"
- xmlns:marcrel="http://id.loc.gov/vocabulary/relators/"

curl -H "Accept:application/rdf+xml" http://id.loc.gov/vocabulary/relators/

SKOS

- Permet d'exprimer des thesauri (Knowledge organisation schemes);
- Deux types d'entités: SKOSConceptScheme, SKOSConcept;
- Trois relations principales: broader, narrower, seeAlso;
- Quelques propriétés: prefLabel, altLabel, definition, note, example
- Vocabulaire secondaire: exactMatch, broadMatch, narrowMatch, closeMatch

Autres sujets

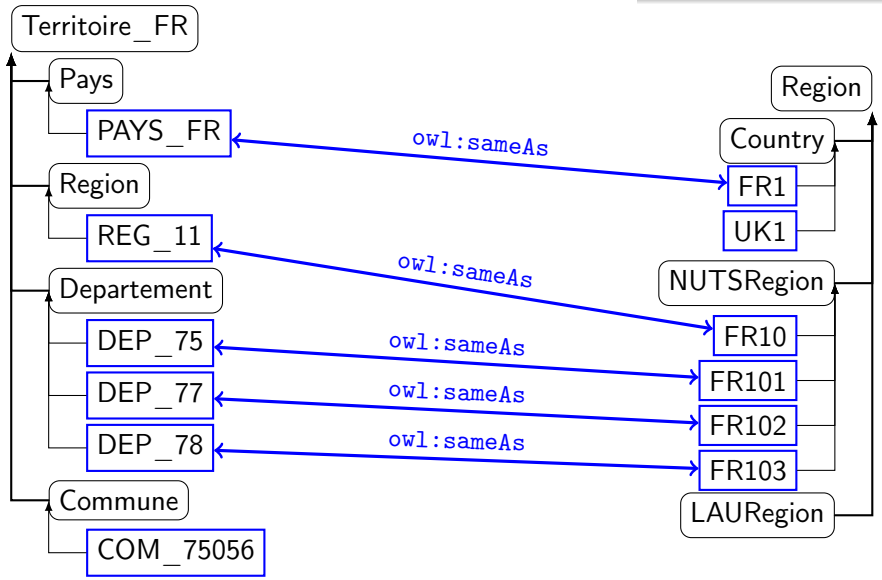
- Schema.org;
- RDFa;
- Répertoires d'ontologies (LOV, Watson, Bioportal);
- Répertoires de sources de données (CKAN, LOD cloud);
- Unique scholarly identifier.

Liage des données

Le but est de produire les liens owl:sameAs entre sources de données.

- Trouver ces liens n'est pas simple:
- Trop fastidieux manuellement;
- Peut parfois se faire en transformant les URIs;
- Peut tirer parti des propriétés (clés);
- Peut utiliser des alignements entre ontologies pour guider la recherche;
- Peut utiliser des techniques statistiques et d'apprentissage;
- C'est un sujet chaud en recherche.

Exemple: Lier INSEE et NUTS



La BNF, d'accord, mais encore?

Le web des données est très important dans les activités documentaires:

- musées (→ CIDOC-CRM)
- bibliothèques nationales (BNF → Rameau; D → Brinksmann)
- centres de docs institutionels (OPOCE → Eurovoc; FAO → Agrovoc)
- instituts géographiques et statistiques (INSEE, IGN, Ordnance survey, Eurostat → NUTS)

En général, un fonds unique qui place l'institution en position d'autorité
D'où les data.gov, data.gov.uk

Producteur ou consomateur

La documentation est en position de médiateur:

- producteur de données concernant son fonds;
- consomateur de données pour satisfaire les requêtes des usagers.

Mais, deux nouveaux modes d'intermédiation peuvent apparaître:

- La production de nouveaux liens;
- La production de requêtes et catégories spécifiques à certains types d'activités.

Le lien fait la force, pas le fonds!

L'intérêt du web des données est dans les liens:

- On peut se référer à des entités décrites ailleurs;
- On peut toujours ajouter de nouveaux liens, par exemples de nouvelles catégories pour un ouvrage (rdf:type);
- On peut joindre des informations provenant de différentes sources (sameAs).

MACS

Multilingual access to subjects

- Groupement de bibliothèques nationales (BNF, Deutsche Bibliothek, British library, Swiss National Library);
- Liage des sujets utilisés pour annoter les œuvres (LCSH, SWD/RSWK, RAMEAU);
- Utilise les principes du web des données;
- Agrège/aligne les thesauri des différentes institutions.
- Objectif: l'utiliser dans The European Library (`http://www.theeuropeanlibrary.org/`)

STITCH

Expériences en Hollande avec l'interconnexion de divers thesauri:

- Liage de différents thesauri utilisés pour annoter les œuvres (Brinkman, Iconclass, GTT, RAMEAU);
- Exporte les données en SKOS (`http://www.cs.vu.nl/STITCH/rameau/`);
- Koninklijke Bibliotheek, Vrije Universiteit, Max Planck Institute for Psycholinguistics, NWO;
- `http://www.cs.vu.nl/STITCH/`.

Europeana

Projet pilote LOD

- Groupement de bibliothèques nationales et de musée;
- Portail offrant un accès aux catalogues des membres;
- Utilise les principes du web des données;
- Utilise uniquement Dublin core
- `http://www.europeana.eu/`

`http://data.europeana.eu/data/proxy/provider/92037/25F9104787668C4B5148BE8E5AB8DBEF5BE5FE03`

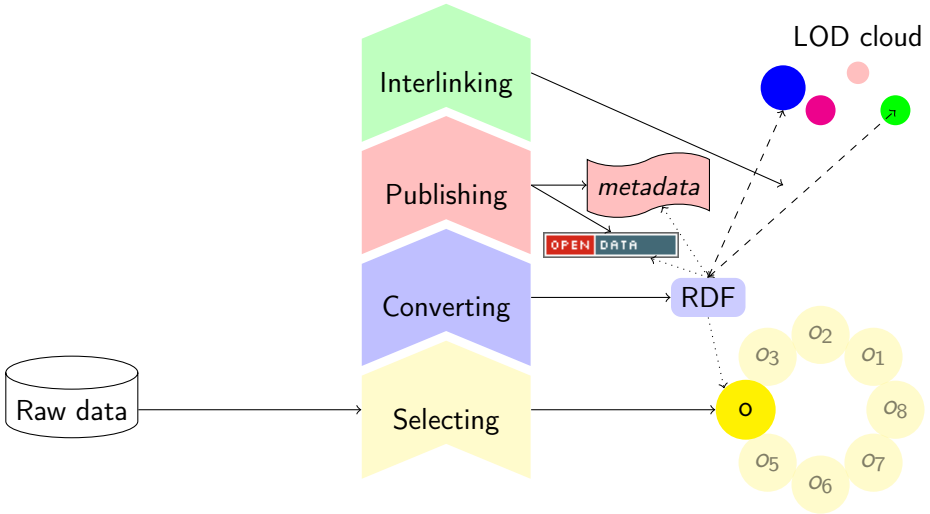
DBPedia.fr

- DBPedia n'indexe que les pages ayant une version anglaise;
- Il lui manque donc des pages et des liens sur des phénomènes typiquement français;
- Objectif: produire la version française de DBPedia;
- Ministère de la culture/INRIA Sophia Antipolis;
- `http://dbpedia.fr`.

Datalift

- Projet ANR 2010-2013
- Recherche: INRIA, LIRMM, Eurecom
- Entreprises: ATOS, Mondeca
- Producteurs de données: IGN, INSEE
- But: produire l'ascenseur pour les données convertissant des sources de données locales en sources de données liées
- `http://datalift.org`

La plate-forme Datalift



TAE/OPOCE

- Eurovoc: thesaurus multilingue de 6700 termes utilisé à l'OPOCE (Luxembourg);
- Maintenu en SKOS sous ITM (Mondeca);
- Termes en 25 langages européens;
- Objectif: exporter sur le web des données (SKOS);
- lier semi-automatiquement à d'autre thesauri (Gemet, ETT, EurLex);
- Mondeca/INRIA Grenoble Rhône-Alpes.

Data.gouv.fr

- Supposé faire le pendant de data.gov et data.gov.uk;
- Mais, on est dans l'open data à ne pas confondre avec linked data;
- Est chargé d'aider à l'ouverture des données des administrations françaises;
- Les données sont produites sous n'importe quels formats (excel, pdf, autre);
- Les données du recensement de l'INSEE y tiennent une place prépondérante
- Placé directement sous la tutelle du premier ministre.
- `http://data.gouv.fr`

Data publica

- Société privée issue de l'INRIA;
- Objectif: produire une place de marché pour les données;
- Exploite le fonds de data.gouv.fr mais aussi d'autres (Eurostat);
- Offre des outils de manipulation des données;
- Très active dans le domaine du "data journalisme";
- `http://www.data-publica.com/`

Questions?

`http://exmo.inrialpes.fr`

`Jerome . Euzenat @ inria . fr`

Conclusion

- Le web des données rompt les silos dans lesquels les données sont enfermées;
- Une part importante de sa valeur ajoutée est dans les liens;
- Ils peuvent être difficiles à trouver.