Utiliser les données ouvertes (*open data*).

Un exemple avec les débats en séance publique à l'Assemblée nationale.

Clément Plancq. LaTTiCe, CNRS, ENS & Univ. Paris 3

Collecter et produire des données pour la recherche en SHS
Fréjus, 15-18 novembre 2016

Définition

- L'open data ou donnée ouverte est une **donnée numérique** dont l'accès et l'usage sont laissés libres aux usagers. [...] Elle est diffusée de **manière structurée** selon une méthode et une **licence ouverte** garantissant son **libre accès** et sa **réutilisation par tous, sans restriction technique, juridique ou financière**.

  https://fr.wikipedia.org/wiki/Open_data

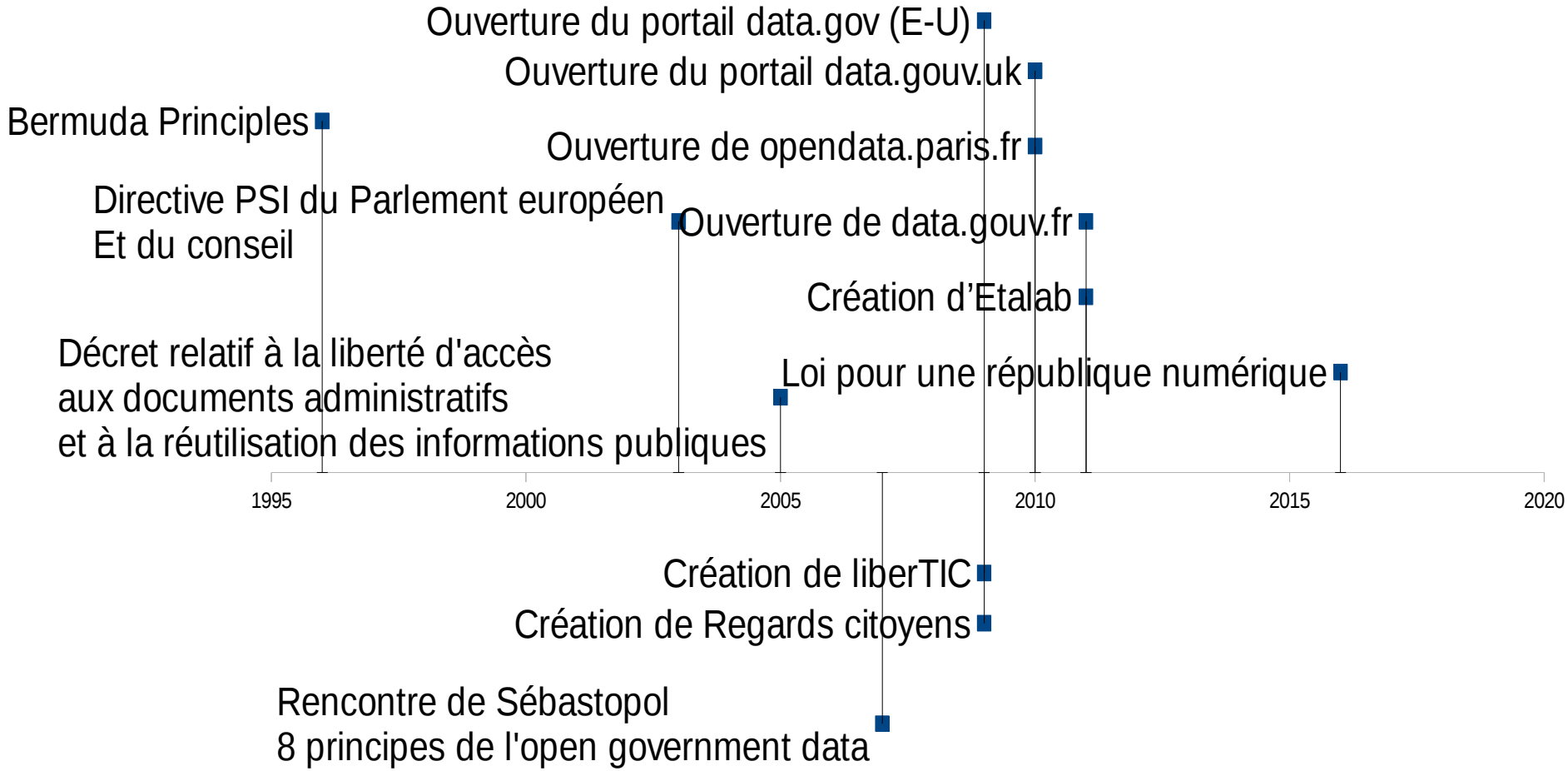
- Pas une idée neuve mais elle a pris une dimension nouvelle avec le Web 2.0 et les initiatives d'*open government*
- S'inscrit dans le concept d'*Open Knowledge* (voir *Open Knowledge Foundation*)

Définition (2)

8 principes de Sébastopol (2007) sur l'*open government data* :

1) ***Complete***
2) ***Primary***
3) *Timely*
4) *Accessible*
5) ***Machine processable***
6) *Non-discrimantory*
7) ***Non-proprietary***
8) ***License-free***

Historique de l'*open data*



Utiliser les données ouvertes publiques

- Avec les citoyens et les entrepreneurs, les chercheurs sont une des cibles privilégiées des données ouvertes publiques
  - Pas de collecte des données
  - Données gratuites
  - Données directement utilisables par l'outil informatique (téléchargement ou API)
- Une aubaine pour les chercheurs ?

Utiliser les données ouvertes publiques

- Données parfois incomplètes (http://www.data.gouv.fr/fr/datasets/liste-des-parcs-et-jardins-donnees-geographiques-ods/)
- Pas de mises à jour planifiées (Lieux de tournage de films à Paris 2002-2010)
- Pas de versionnage
- Faible volume (*open data* / *big data*)
- Assez peu de méta-données

Débats en séance publique de l'AN (1)

http://data.assemblee-nationale.fr/travaux-parlementaires/debats

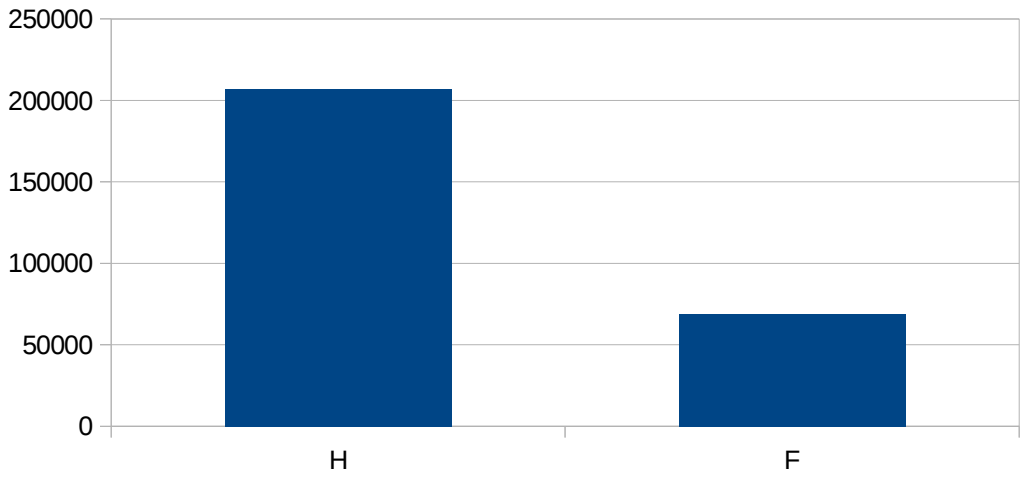
- 1 fichier XML de 427 Mo
- Pas de schéma associé
- Très peu de méta-données
- Données semi-structurées

<paragraphe ID_ACTEUR="PA700879" ID_MANDAT="-1" ID_NOMINATION_OE="PM702963" ID_NOMINATION_OP="-1" code_grammaire="QG_2_5" code_parole="PAROLE_1_2" code_style="NORMAL" id_preparation="454757" id_syceron="524586" ordinal_prise="1" ordre_absolu_seance="26" sommaire="0" valeur="" valeur_ptsODJ="1"><ORATEURS><ORATEUR><NOM>M. Harlem Désir</NOM><ID>PA700879</ID><QUALITE>secrétaire d'État</QUALITE></ORATEUR></ORATEURS><texte>La première réponse, c'est d'abord de renforcer la présence de l'Union européenne en mer Méditerranée. Ce matin, la Commission européenne a annoncé qu'elle triplerait le budget des opérations Poséidon et Triton menées par l'agence Frontex. C'est ce que nous-mêmes avions proposé. La France, vous le savez, a envoyé des navires.
Deuxième axe : la lutte contre les trafiquants. Nous demandons un mandat aux Nations-unies pour pouvoir nous attaquer à eux, en coopération avec les États de transit et les États de provenance.
Troisième axe : prévenir les flux illégaux par une politique de stabilité, de développement et de transition démocratique. <italique>(Interruptions sur les bancs du groupe UMP.)</italique></texte></paragraphe><paragraphe ID_ACTEUR="PA687214" ID_MANDAT="PM687217" ID_NOMINATION_OE="-1" ID_NOMINATION_OP="-1" code_grammaire="INTERRUPTION_1_10" code_parole="" code_style="NORMAL" id_preparation="454759" id_syceron="524695" ordinal_prise="1" ordre_absolu_seance="29" sommaire="0" valeur="" valeur_ptsODJ="1"><ORATEURS><ORATEUR><NOM>M. Sylvain Berrios</NOM><ID>PA687214</ID><QUALITE/></ORATEUR></ORATEURS><texte>Ça, ça va tout changer !</texte></paragraphe>

Prises de parole H / F

```
let $orateur-h := count(doc($debats-an)//ORATEUR[starts-with(./NOM/text(), "M.") and not(contains(./NOM/text(), "président"))])

let $orateur-f := count(doc($debats-an)//ORATEUR[starts-with(./NOM/text(), "Mme") and not(contains(./NOM/text(), "présidente"))])
```



Débats en séance publique de l'AN (4)

Quid des interruptions ?

- Réactions de groupe notées dans `<italique>`
  - `<italique>(Sourires.)</italique>`
  - `<italique>(Sourires sur les bancs du groupe SRC.) </italique>`
  - `<italique>(Applaudissements sur les bancs du groupe UMP.)</italique>`
  - `<italique>(Exclamations sur les bancs du groupe UMP.)</italique>`
  - `<italique>(Protestations sur les bancs du groupe UMP.)</italique>`
  - `<italique>(« Très bien ! » et applaudissements sur les bancs du groupe SRC.)</italique>`
  - `<italique>(Exclamations sur tous les bancs).</italique>`
  - `<italique>(Applaudissements sur les bancs du groupe socialiste, républicain et citoyen.)</italique>`
  - `<italique>(Applaudissements sur plusieurs bancs du groupe socialiste, républicain et citoyen.)</italique>`

Débats en séance publique de l'AN (5)

Quid des interruptions ?

- Réactions individuelles notées de façon plus structurée

```
<paragraphe ID_ACTEUR="PA230329" ID_MANDAT="PM645079"
ID_NOMINATION_OE="-1" ID_NOMINATION_OP="-1"
code_grammaire="INTERRUPTION_1_10" code_parole=""
code_style="NORMAL" id_preparation="50970"
id_syceron="61570" ordinal_prise="1"
ordre_absolu_seance="18" sommaire="0" type_debat="PLF"
valeur="" valeur_ptsODJ="1"><ORATEURS><ORATEUR><NOM>M.
Jean-François
Lamour</NOM><ID>PA230329</ID><QUALITE/></ORATEUR></ORAT
EURS><texte>Très bien ! Brillant cet après-midi !
</texte></paragraphe>
```

Débats en séance publique de l'AN (6)

- Document maître avant publication
- Forme et contenu mêlés
- Statut linguistique difficile à définir
  - Prise de parole publique en contexte professionnel
  - Orateurs professionnels
  - Pas un dialogue, un locuteur à la fois
  - Pourtant il y a des interactions (avec les groupes parlementaires, avec le président), notamment lors des interruptions
- Transcriptions fidèles ? pas de disfluences, peu de reprises anaphoriques
- Parole écrite ?

Open data : des données brutes ? (1)

- Beaucoup d'information sur la ré-utilisation des données ouvertes mais peu de choses sur leur production
- [Denis, Goëta 2013] "la fabrique de données brutes"
  - Les conditions de production des données ouvertes échappent à l'utilisateur
  - La protection des données personnelles empêche souvent la publication des données "brutes"
  - Les outils des systèmes d'information ne permettent pas forcément d'exporter les données dans un format exploitable : "extraction"
  - "Brutification des données"

Open data : des données brutes ? (2)

- Les données ouvertes alimentent la réflexion sur ce qu'est une donnée
- "raw data is an oxymoron" (Gitelman 2003 [Gawker 2000])

Bibliographie

COLLECTIF. Open data en SHS : Proposé par Cynthia Pedroja, Elifsu Sabuncu, Anne-Laure Stérin In : THATCamp Paris 2012 : Non-actes de la non-conférence des humanités numériques [en ligne]. Paris : Éditions de la Maison des sciences de l'homme, 2012 (généré le 01 novembre 2016). Disponible sur Internet : <http://books.openedition.org/editionsmsh/364>.

Jérôme Denis, Samuel Goëta. La fabrique des données brutes. Le travail en coulisses de l'opendata. Penser l'écosystème des données. Les enjeux scientifiques et politiques des données numériques, Feb 2013, Paris, France. <halshs-00990771>

Lisa Gitelman (dir.). « Raw data » is an oxymoron. MIT Press, Infrastructures (coll.), 2013, 182p.

Romain Lacombe, Pierre-Henri Bertin, François Vauglin, Alice Vieillefosse. Pour une politique ambitieuse des données publiques : les données publiques au service de l'innovation et de la transparence. Paris : Ecole des Ponts Paris Tech, 2011 <http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/114000407/0000.pdf>

Jacqueline Léon, Isabelle Tellier. Le data turn. Des premiers traitements statistiques du langage (1950-60) à la fouille de textes. L'information grammaticale, Peeters Publishers, 2014, pp.30-39. <hal-01132794>

http://data.assemblee-nationale.fr/travaux-parlementaires/debats

http://exist-db.org/exist/apps/homepage/index.html