

# OPEN DATA

François Bancilhon

Data Publica

[twitter.com/fbancilhon](https://twitter.com/fbancilhon)

[francois.bancilhon@data-publica.com](mailto:francois.bancilhon@data-publica.com)

[www.data-publica.com](http://www.data-publica.com)

FOSSA 11

October 27, 2011

# Open Data

- Availability of public sector information (PSI) for access and reuse by citizens and organizations
- PSI: collected, produced, maintained and used by public organizations
- Re-use: ability to change, improve, check, cross-reference, derive, integrate in applications
- Citizens and organization
  - researchers, groups, non-profits, companies, public organizations, press, etc.

# Open Data

- Data
  - Documents, data and information
  - Words and numbers
  - Structure or not
- Topics
  - Transportation, enterprises, statistics, geography sociology, ecology, environment, economy, legal, etc.
- Formats
  - xls, csv, xml, pdf, kml, etc.

# Restrictions

- State or defense secrecy
- Privacy protection
  - Some data available only after anonymisation
- IP and copyright
  - Photos and works of art

# Open Data: key drivers

- Produced with taxpayer money, they belong to the taxpayer
- Government transparency requires making public data public
- Worth money, so the State must try to monetize them
- Fuel to the emergence of new mobile and Internet apps
- They can simplify/improve citizens lives

# Open Data: a global phenomenon

- USA: data.gov initiative from the Obama administration (2009)
- UK: data.gov.uk initiative (2010)
- EU: European directive on PSI reuse (2003)
- Many countries: France, Finland, Australia, New Zealand, Northern Ireland, Kenya, etc.

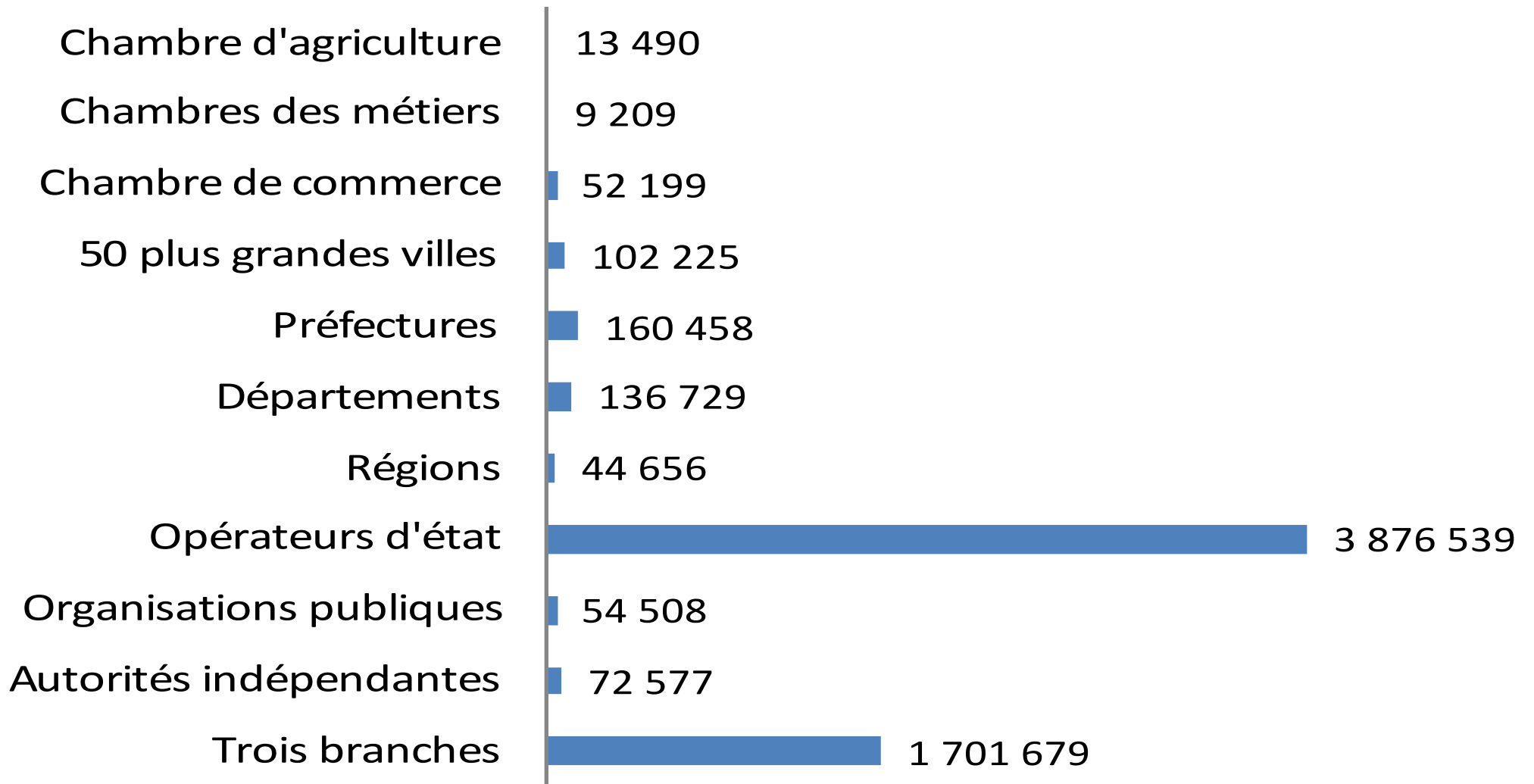
# Licenses

- Not as clear as in the Open Source world
  - No yet clear major license such as GPL
  - No clear guru such as Richard Stallman
  - Data and code are not the same thing
- Decision makers are confused
  - A multiplicity of licenses
  - Interoperability issues
- France: APIE, LIP, LO
- Int'l: CC By & CC By SA (ODbL)

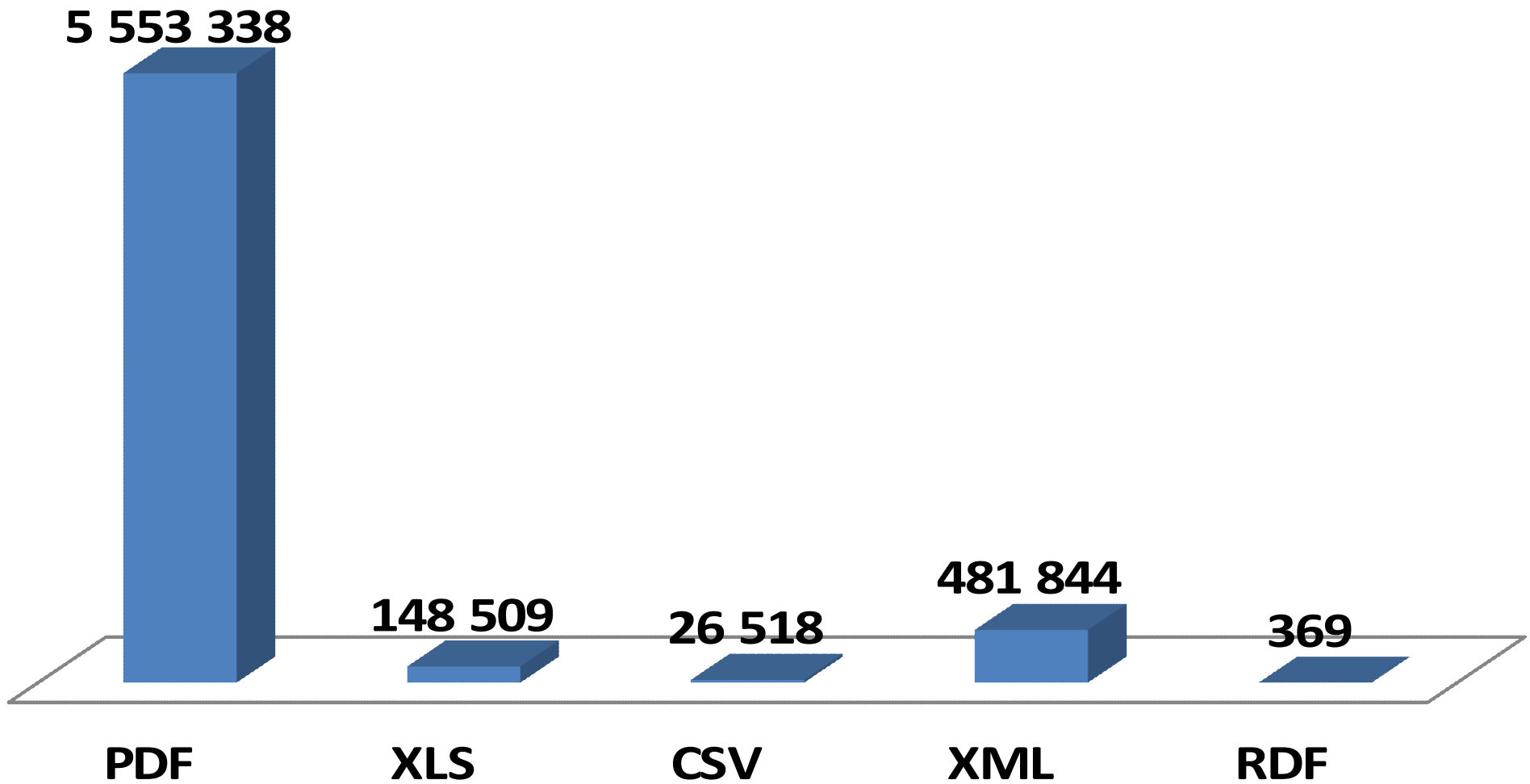
# Public Web Sites

	Listed	Sites
<b>Three Branches</b>		
Executive		
Presidency	1	1
Prime minister	1	40
Ministries	29	644
Legislative		
Chamber of deputies	1	1
Senate	1	1
Judicial		
Courts and councils	9	11
<b>Independant administrative authorities</b>	44	36
<b>State operators</b>	550	511
<b>Préfectures</b>	128	121
<b>Local Elected bodies</b>		
Regions	26	26
Departments	100	99
Cities	50	50
<b>Government controled organizations</b>	57	68
<b>Chambers of commerce</b>	166	163
<b>Chambers of craftsmen</b>	120	113
<b>Chambers of agriculture</b>	79	72
<b>Total</b>	<b>1 362</b>	<b>1 957</b>

# How much public data



# In which format?



# Traditional Data Market the French example

- 1.6 billion euros annually
- About 60% come from public data
- 170 players
- 9 verticals
- 2 horizontals
  - market intelligence and content processing

# 9 verticals

Vertical	Example	Size (M€)
Financial	Reuters	300
Press	Press Index	250
Legal	Francis Lefebvre	240
Solvability	Altarès	160
Scientific Technical Medical	Meteo France	160
Image	Sipa	60
Economy	Société.com	55
Marketing	Acxiom	55
Patents	Reuters	25

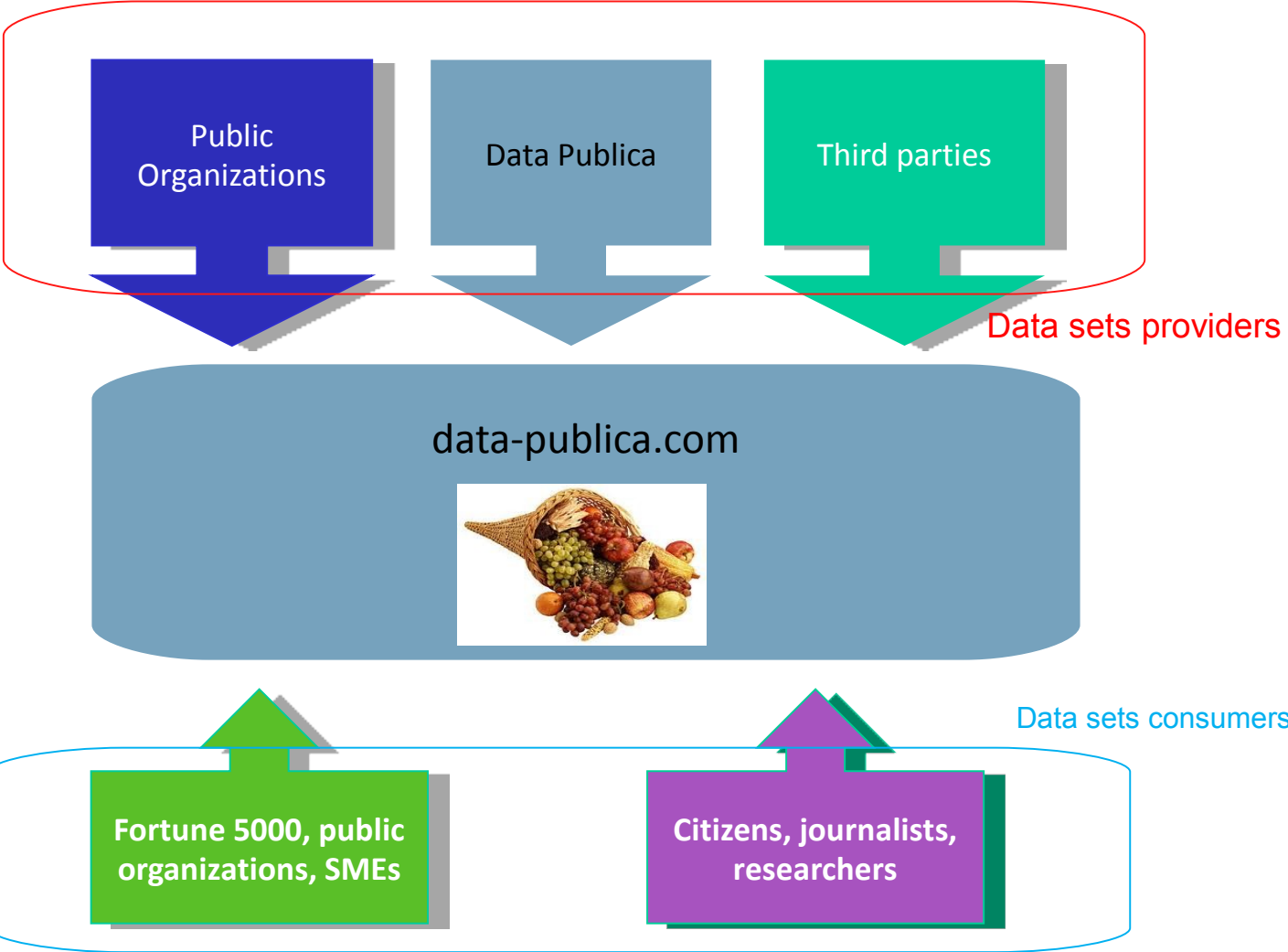
# Market: the new concepts

- DaaS (Data as a Service)
  - Subscribe to a data set and access it as you need
- Data Marketplaces
  - Buy and sell datasets
  - Sellers and buyers meet here and are treated the same
- DataStores
  - Buy datasets, provided by the operator or third parties
- Data Directories
  - Search and browse datasets

# Data Publica « Elevator Pitch »

- Development of complete and detailed knowledge of French electronic data (Free/charged, public/private, regional/national/international)
  - Operation of a data directory and associated search engine
- Two lines of revenue
  - Custom data set development
  - Data Store operation
    - Sale of Data Sets (from Data Publica and third party vendors)

# Data Publica Datastore



# Technical topics

- Extraction
  - Scraping, crawling, text mining, ETL
- Storage
  - DBMS, NoSQL
- Processing
  - Data quality & cleansing, ETL, Master Data Management, formats conversion
- Integration
  - Database, Semantic Web
- Delivery
  - API, Protocols and standards, formats
- Visualization

# Microsoft: 3 initiatives

- Azure/Datamarket/Marketplace
- Open Government Data Initiative
- OData

# Google's 3 initiatives

- DSPL
- Google Public Data Explorer
- Google Refine + Freebase

# Conclusion

- A global trend
- Many stakeholders
  - enterprises, non profits, public bodies
  - Citizens, press, researchers.
- Several dimensions
  - Technical, legal, social & political, business
- It's only a beginning



DATA PUBLICA

François Bancilhon  
[twitter.com/fbancilhon](https://twitter.com/fbancilhon)  
[francois.bancilhon@data-publica.com](mailto:francois.bancilhon@data-publica.com)  
[www.data-publica.com](http://www.data-publica.com)