UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
School of Professional & Executive Development

## DATA SHARING

**Anna Queralt**

Barcelona; January 25, 2017

1

UNIVERSITAT POLITÈCNICA DE CATALUNYA
School of Professional & Executive Development

**DATA SHARING** · **Anna Queralt** · Barcelona, Jan 2017

## About me

- Now:
  - Senior Researcher, Storage Systems Research Group
  - Big data sharing (and HPC)
- Before:
  - Teaching assistant and researcher at LSI, ESSI
  - Part-time lecturer at Estudis d'Informàtica i Multimèdia
  - Software Engineering, Knowledge Representation & Reasoning
- MSc degree in Computer Science (FIB, UPC)
- PhD in Computer Science (LSI, UPC)
- Contact:

  anna.queralt@bsc.es
  @queralt_anna

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
UPC
UOC

2

# Related Sessions

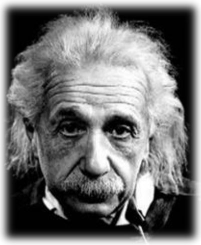
- **January 25th: Data Sharing (Teoria)**
- February 1st: Semantic Data Models (Teoria)
- February 3rd: Open Data - SPARQL (Laboratori)

# Big Data

sharing
validity
variety
velocity
vocabulary
view
vanity
visuality
value
volume

# Why sharing?

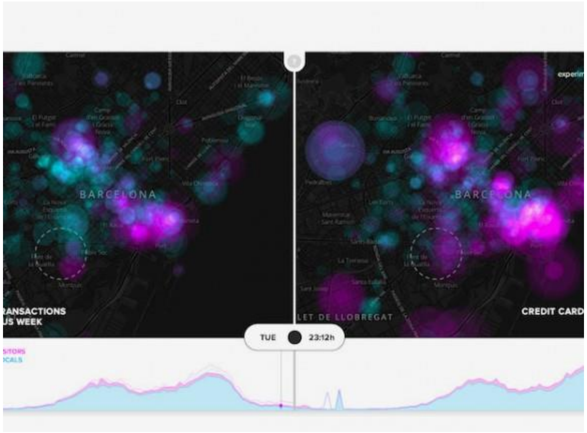
"Creativity is just connecting things"
**Steve Jobs**



"We cannot solve our problems with the same thinking we used when we created them"
**Albert Einstein**

# Example: BBVA

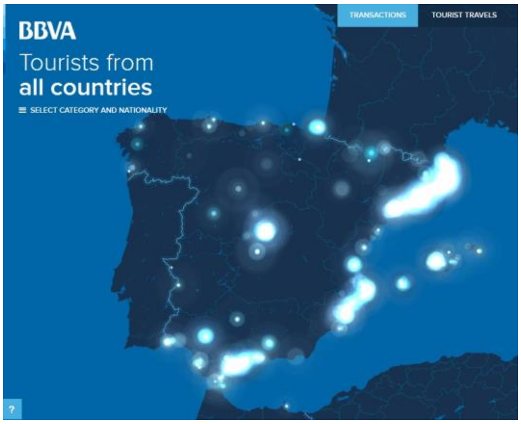
- Economic impact of the MWC 2012 in Barcelona



http://mwcimpact.com/

# Example: BBVA

- Economic impact of tourism in Spain



http://www.centrodeinnovacionbbva.com/bbvatourism

# Open Data

OPEN DATA

Open data is data that can be freely used, reused and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike.

This means:

- Availability and access
  - Must be available in a convenient and modifiable form

- Re-use and redistribution
  - Must be provided under terms that permit re-use and redistribution, including intermixing with other datasets

- Universal participation
  - No discrimination against fields or against persons or groups
    - For example, “non-commercial” restrictions that would prevent “commercial” use are not allowed

OPEN DATA

# Open Data

Why this definition? **Interoperability**

- It ensures that when you get datasets from different sources you will be able to combine them

- It allows to combine them into the larger systems where the real value lies

# Importance of Open Data in Europe

European Commission

**“Towards a thriving data driven economy”**

European strategy on data, with Open Data as a prominent element

– Infrastructure
– Analysis
– Privacy
– ...

EUROPEAN DATA PORTAL

EUDAT
European Data Infrastructure

OpenAIRE
Open Access Infrastructure for Research in Europe

pandata
Photon and Neutron Data Infrastructure

EGDI SCOPE

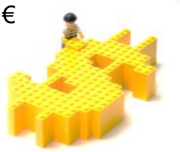
# Why?

- Makes public administration more efficient and more effective
  - Thanks to Open Data, the US government has reduced the annual costs of attending citizens from 500 M$ to 34 M$

- Open data portals stimulate innovation and economic growth
  - Research suggests that seven sectors alone could generate more than $3 trillion a year in additional value as a result of open data

    *Open Data: Unlocking Innovation And Performance With Liquid Information (*McKinsey Global Institute)
  - Big Data and open data will contribute more than 200.000M€ to the European economy by 2020

    *Big&Open Data in Europe: a growth engine or a missed opportunity? (*demosEuropa, WISE , Microsoft)



# How?

- New apps and businesses

- Vendors of support products, e.g. analysis and visualization software

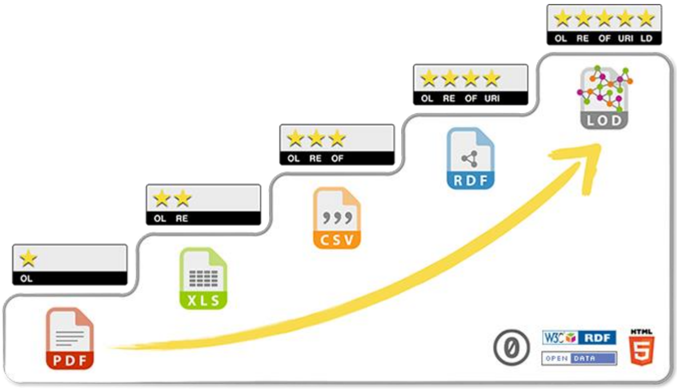
- Indirectly: better informed people and organizations

# The Goal: Linked Open Data

- From the World Wide Web to the Semantic Web
  - Published data that…
    - Is machine-readable,
    - Its meaning is explicitly defined,
    - Is linked to other external data sets,
    - And can be linked to from other external data sets

- Tim Berners-Lee outlined a set of rules:
  - Use URIs as names for things
    - Universal identifiers to represent real-world objects
  - Use HTTP URIs so that people can look up those names
    - Universally available (where to locate it)
  - When someone looks up a URI, provide useful information, using RDF and SPARQL
    - Description of the object features or characteristics
  - Include links to other URIs, so they can discover more things
    - Relationships as first-class citizens (information integration)

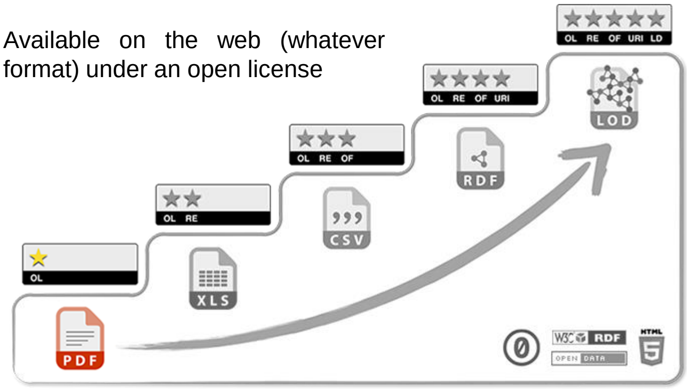
# 5-star deployment scheme



http://5stardata.info

# 5-star deployment scheme

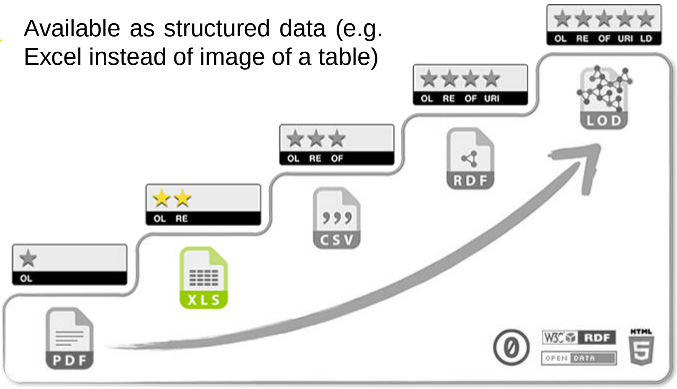
★ Available on the web (whatever format) under an open license



http://5stardata.info

# 5-star deployment scheme

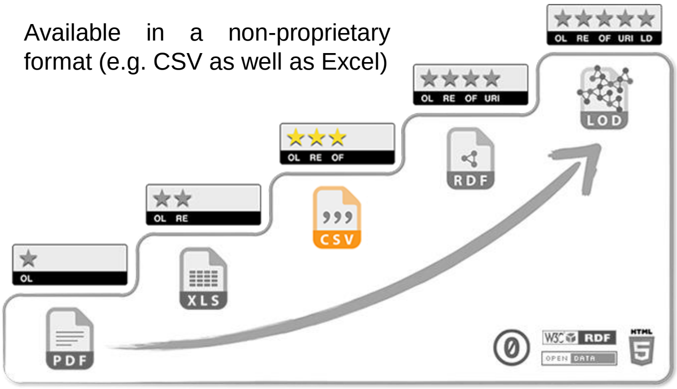
★★ Available as structured data (e.g. Excel instead of image of a table)



http://5stardata.info

# 5-star deployment scheme

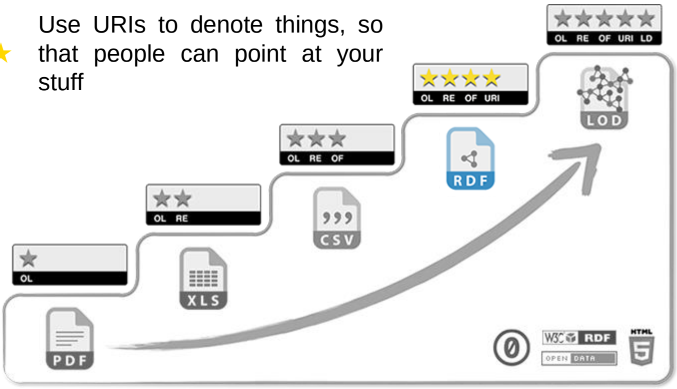
★★★ Available in a non-proprietary format (e.g. CSV as well as Excel)



http://5stardata.info

# 5-star deployment scheme

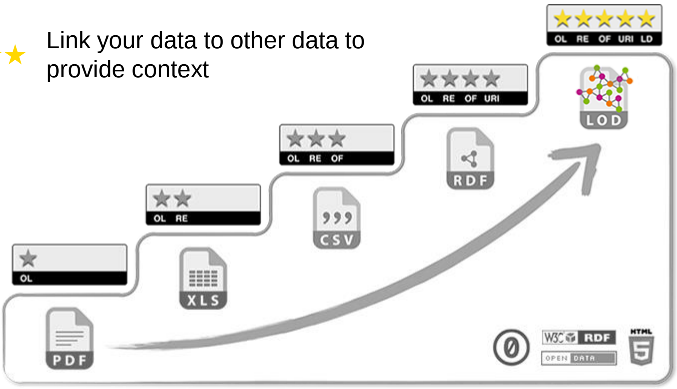
★★★★ Use URIs to denote things, so that people can point at your stuff



http://5stardata.info

# 5-star deployment scheme

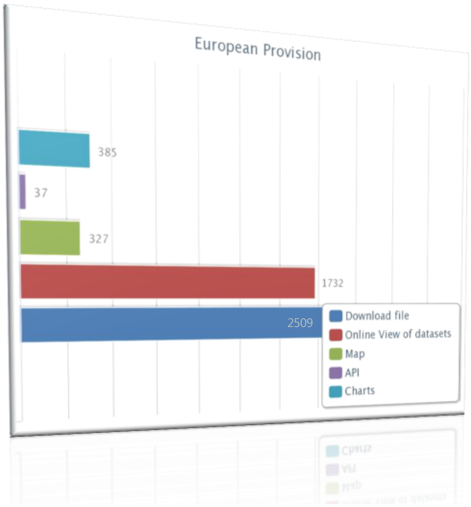
★★★★★ Link your data to other data to provide context



http://5stardata.info

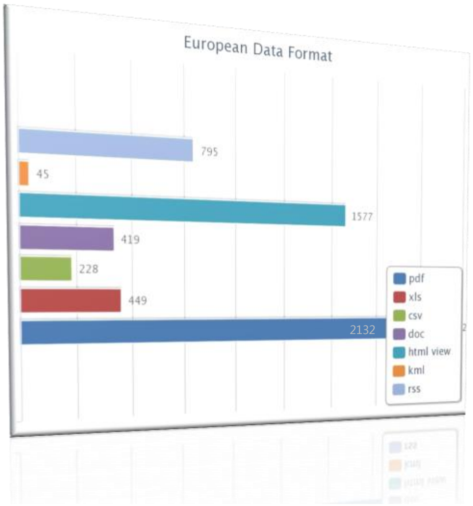
# How is data shared today?

- Most open data is available as downloadable files



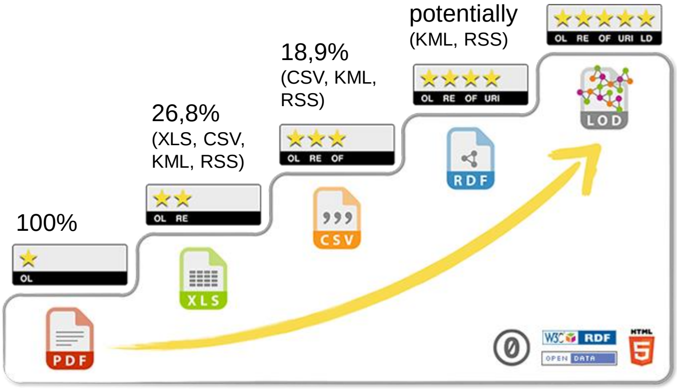
# How is data shared today?

- Only 27% of sources are provided in a processable format



engage

# Regarding the 5-star model…



100%

26,8%
(XLS, CSV, KML, RSS)

18,9%
(CSV, KML, RSS)

14,8% potentially
(KML, RSS)

14,8% potentially
(KML, RSS)

# Open data from the public sector

- Improve transparency and citizen participation
  - Global Open Data Index (http://index.okfn.org/)
  - Public Dataset Catalogs (http://datos.fundacionctic.org/)

idescat

OpenData BCN

Dades obertes gencat
Obertura de dades públiques (open data) de la Generalitat de Catalunya

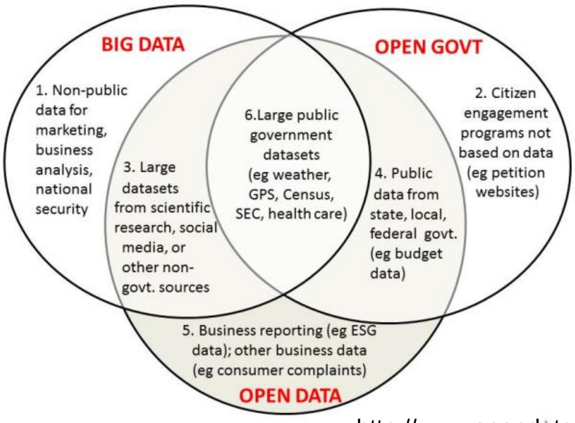
datos.gob.es
reutiliza la información pública

European Union Open Data Portal

DATA.GOV

- Some examples of promotion
  - In BCN: Apps4BCN, Apps4Transparency, Apps&Cultura, Barcelona Smart City App Hack, …
  - Worldwide: International Open Data Hackathon, Global Open Data for Agriculture and Nutrition, and many other local initiatives

# But Open Data is not just about Govt



BIG DATA

OPEN GOVT

OPEN DATA

1. Non-public data for marketing, business analysis, national security

2. Citizen engagement programs not based on data (eg petition websites)

3. Large datasets from scientific research, social media, or other non-govt. sources

4. Public data from state, local, federal govt. (eg budget data)

5. Business reporting (eg ESG data); other business data (eg consumer complaints)

6.Large public government datasets (eg weather, GPS, Census, SEC, health care)

http://www.opendatanow.com

# Open Data vs Shared Data

- **Open data** is providing unrestricted data to everyone
  - Available to all without restrictions of what they do with it
  - There cannot be legal restrictions on making it available

- **Shared data** is providing restricted data to restricted organizations or individuals
  - Restricted because
    - Provides a revenue stream
    - It is sensitive in some way (personal data, security issues,…)

- Both **public** and **private** data can be either **open** or **shared**

# What is “private data”?

- Data generated by companies when performing their activities
  - – Lists of clients or providers
  - – Sales
  - – Business processes
  - – …

- Information generated to be consumed as an independent product
  - – Polls
  - – Reports
  - – …

- Information publicly accessible on the Internet
  - – Corporate web pages
  - – Comments and likes in social networks
  - – …

# Benefits for the private sector

- Using open/shared data (combined with company data)
  - From public administrations
    - See Open Data 500
  - From social networks
  - From other private companies

- Sharing data:
  - With partners: organizations collecting other types of data, app developers, …
  - With competitors: benchmarking, common risk (insurance, pharma, apparel…)
  - With customers: transparency, concern with social responsibility, …

# Benefits for the private sector

- Or even selling data
  - Data services allow for flexible pricing models
    - Volume-based
      - Quantity-based pricing
      - Pay per call
    - Data type-based
    - Subscription
  - Publishing datasets in a Data Marketplace
    - E.g. xDayta, BDEX

# Business Models

- Data providers need some motivation other than “helping grow the economy”

- Single releases are rarely interesting
  - Data providers need that their activities are self-substaining or even profitable

- Ways to bring benefits to providers themselves, derived from models around open source
  - Cost avoidance
    - Proactively release data, and make it easy to find
    - Or avoid political/reputation cost

Open Data User Group (gov.uk)

# Business Models

- Sponsorship
  - Someone that thinks that a particular dataset should be available may pay for its publication
  - E.g. Companies that sell analysis or visualization products, data solutions…

- Freemium
  - Publish data in a basic form and offer advanced access to those who pay
    - Different formats
    - Unconstrained number of API calls
    - More sophisticated querying
    - Access to data dumps instead of through an API (or viceversa)
    - Provision of feeds of changes to the data
    - Enhancement of the data with additional information
    - Early access to data
    - …

# Business Models

- Dual licensing
  - Open license for certain purpose, and closed license for others
  - E.g. charging based on the size/revenue/kind of organization (start-ups, research centers, universities…)
- Support and services
  - Charging for support and services around the data, instead of for the data itself
    - Guarantees on data availability
    - Prioritization of bug fixes
    - Timely help for customers using the data
    - Services around visualization, analysis and mashing with other data

# Business Models

- Charging for changes
  - Charge whatever it took to support providing the data as open data, i.e. “administration costs”
- Increasing quality through participation
  - Enlisting other parties who would benefit from having the data up-to-date
  - Does not entirely cover costs, but saves effort
- Supporting primary business
  - Releasing data about the business drives the development of apps that attract new customers (e.g. Bicing, TMB, …)
  - The data provider ends up improving its own use of its data

# Some examples

- Shared data from private companies

twitter
Google Maps
The New York Times
BBVA
MARVEL
facebook
theguardian
orange
LEGO
YouTube
Google trends
BBC
BMW
ABSOLUT Country of Sweden VODKA

- Shared (open) data from communities

DBpedia
Freebase

# 2 main ways of sharing/opening data

# Downloadable files: CKAN

- CKAN is a tool to make open data websites
- Used by many national and local governments, research institutions, and other organizations
  - E.g. http://www.europeandataportal.eu
- Users can use its search features, browse and find the data, and preview it using maps, graphs and tables
- Data is published in units called *datasets*, which contain:
  - Metadata: title, publisher, formats, license, etc.
  - Resources: stored internally or as a link
    - Views: Chart, table, map…
- Each dataset is *normally* owned by an organization
  - Datasets are usually initially private, visible only to the users in the same organization

# Downloadable files: CKAN

- DataStore extension: Provides a database for storage of structured data from resources
  - Automatic data previews on the resource’s page
  - The DataStore API: search, filter and update the data without having to download and upload the entire data file
- Support for Linked Data and RDF
  - Various vocabularies can be used for describing datasets: Dublin Core, DCAT, …
- Other examples of similar products:

Socrata

The Dataverse Project

junar

DISCOVER POSSIBILITIES opengov

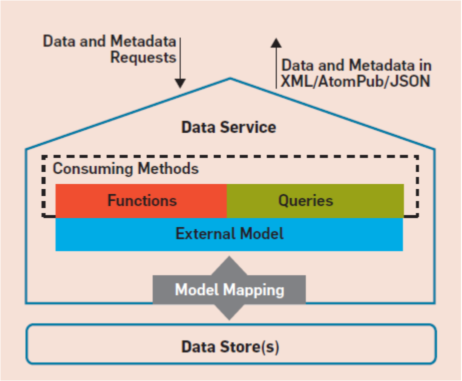
# Data services

- Motivation
  - Today’s business practices require access to enterprise data by both external and internal applications
    - Suppliers release data to retailers, health providers release data to patients, companies release data to customers
  - Data owners need to ensure access to their data is appropriately restricted and has predictable impact on their infrastructure
- They provide rich metadata, expressive languages, and APIs for consumers to access data
- They are a specialization of Web services that can be deployed on top of data stores, other services, and/or applications to encapsulate data-centric operations
- They are descendants of the stored procedures in relational database systems
- They provide Data-as-a-Service

*M.J. Carey et al. Data Services. Communications of the ACM 55(6), 2012.*

# Data services

- Data service architecture



Any kind of data store, including RDF triplestores:
- GraphDB (formerly OWLIM)
- Virtuoso as a triplestore
- …

# Data services

- Technologies
  - REST
    - Representational State Transfer
    - Architecture for scalable web services developed by W3C
    - Based on HTTP communications (POST, GET, PUT, DELETE…)
    - Web service APIs that follow this standard are called RESTful APIs:
      - Base URI
      - Internet media type for the data: JSON, XML, AtomPub…
      - Standard HTTP methods
      - Hypertext links to reference state and related resources
    - Example: Recent tweets containing hashtag #openData

```
https://api.twitter.com/1.1/search/tweets.json?q=%23openData&
result_type=recent
```

# Pros and cons of current approaches

- **Downloadable files:** Sharing bulk datasets
  - Provider point of view
    - Easy from a technical perspective
    - …But loses any kind of control on data once it is downloaded
  - Consumer point of view
    - Easy to build an app, flexible
    - No dependency on the original provider
    - …But data is never up to date
    - …But has to adapt to different data formats

# Pros and cons of current approaches

- **Data services:** Sharing data (and functionality) at a finer granularity
  - Provider point of view
    - He keeps full control
    - Allow for different business models around data
    - …But more difficult to build
  - Consumer point of view
    - Data is always up-to-date, and does not need to manage it locally
    - …But restricted to the interface offered by the provider

# So the picture is…

- Provider wants control, while consumer wants flexibility

- Provider
  - Must prepare the datasets or build a data service
  - Does not get much benefit from releasing data
    - Unlike (potentially) the apps built on top…
- Consumer
  - Depends on the datasets/API that the provider releases, which may not satisfy his needs
  - Has to adapt his applications to the data available and its format

- Essentially, the problem is that **control depends on the applications**

# Do we have a problem?

- We are constantly sharing our data as individuals (or selling it in exchange for services or goods):
  - What Google knows about you (and probably Apple, and Facebook, and …):
    - Where you have been (maps.google.com/locationhistory)
    - What you have searched (history.google.com/history)
    - What devices you use (security.google.com/settings/security/activity)
  - On-line shops and portals
  - Loyalty cards
  - …

- People are starting to be worried
  - Companies that eliminate your digital identity: Eliminalia, Red Points, …
  - Companies that buy your personal data: Datacoup
- But you always depend on the data holder and his applications
  - New ideas and research towards managing and controlling your digital self:
    - “In practical terms, a person’s data would be equivalent to their money” (World Economic Forum)
    - PIMS, Personal Information Management System (Abiteboul)
    - OpenPDS, Open Personal Data Store (MIT)

# CONCLUSIONS

# Current situation

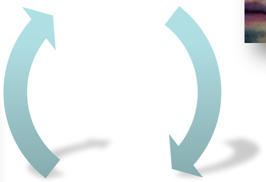
- Opening data is cool!
  - A lot of open data portals, but most data is not structured/annotated
  - This is quite ok for a single dataset and as a transparency exercise, but of little help for generating value

- The practice of releasing data is still in its infancy
  - Little experience in sharing private data
  - Little experience in selling data
  - Little knowledge on the benefits of exploiting corporate data as a product

- No solutions for sharing data in a really convenient way
  - Secure for the owner
  - Flexible for the consumer

# Sharing data is key to innovation (and economy)

See data shared by others from
a different perspective



Build new knowledge or services on top

# EXERCISE

# Part I: Sharing data

- Provider role
  - Take the company/organization where you work and think about which data it could share so that others can re-use it. Specify:
    a) Name or kind of company and brief description of the data to be shared
    b) Format. Choose between the following, and justify your decision:
       - Downloadable files, specifying the format (DOC, PDF, RDF, …)
       - Data service
    c) Concrete semantics / API
       - Downloadable files: what the dataset(s) contain(s)
       - Data service: functions offered and what they return
    d) Explain the benefits for the company, and the business model to sustain this activity, and justify your decision.

# Part II: Reusing data

- Consumer role
  - Propose 2 different apps, services, products… based on this data. For each of them explain:
    a) Name or kind of consumer (app developer, partner, competitor …)
    b) The app/service/product/business proposed using the shared data
    c) Which fields/functions from the shared data you will use, and how
    d) Which other data sources (public or private) you will use, and how