Is Open Data Really at Developer's Fingertips?

Vadim Savenkov
savenkov@dbai.tuwien.ac.at

Database and Artificial Intelligence Group
Institut für Informationssysteme
Technische Universität Wien

November 27, 2014

Open Data

- Global trend started as governmental initiative
  - *Promoting transparency and accountability*
  - *Empowering citizens to drive public sector reform*
  - *Releasing the economic and social value of information*

Andrew Stott, UK's Director of Digital Engagement
on the governmental Open Data initiative in 2009

- Gains momentum among private enterprises
  - Provide data at open (state-funded) platforms, e.g. `opendataportal.at`
  - Sponsor independent platforms `datahub.io` by Open Knowledge Foundation
- Open source software for publishers available: CKAN
  - powering `data.gov.at`, `data.gov.uk` and many others

Five stars of open data [Tim Berners-Lee, 06] [1]

- ⋆ On the Web
- ⋆ ⋆ Machine-readable data
- ⋆ ⋆ ⋆ Non-proprietary format
- ⋆ ⋆ ⋆ ⋆ RDF, use URLs to identify data items so that the can be referenced.
- ⋆ ⋆ ⋆ ⋆ ⋆ Linked RDF: reference other datasets, provide context.

As of now, most data is ⋆⋆⋆: CSV, XML (or JSON/BSON). RDF is being adopted (single “sparql” dataset at `data.gov.at`, already hundreds on `data.gov.uk` and `data.gov`)

Also, complex special purpose data, e.g.:

- Satellite imagery (e.g., LANDSAT by NASA the US Geological Survey)
- Genomics (e.g., `www.openpgx.org`)
- CERN LHC experiments (CMS online, more under embargo yet)

[1] via Axel Polleres and inkdroid.org

Making Open Data Work

- Key success factor: Apps.
- Already 257 at `data.gv.at`. Not that much fewer than on `data.gov.uk` (357) or `data.gov`(341)!
- In the spirit of "empowering citizens"
- How can public research support this?

> *Base assumption: Added value comes from **comparable** Open datasets being **combined**.*
>
> *Axel Polleres, talk at ICD DataHub 2014*

Leverage extensive research on data integration.

Server-side middleware

- Data integration systems are often designed either as specialized P2P applications or as web services
- Data-oriented web services can bring great value, e.g.:
  - Data linkage and cleaning
  - Explicit schema description
  - Harmonization of formats of different sources
  - Uniform powerful query capabilities (XQuery, SPARQL) against distributed sources
- Many systems available either as commercial services (e.g., `28.io`) or as open-source code.

Server-side middleware

- Data integration systems are often designed either as specialized P2P applications or as web services
- Data-oriented web services can bring great value, e.g.:
  - Data linkage and cleaning
  - Explicit schema description
  - Harmonization of formats of different sources
  - Uniform powerful query capabilities (XQuery, SPARQL) against distributed sources
- Many systems available either as commercial services (e.g., `28.io`) or as open-source code.

Downside: no grassroots effect

- Resource consuming (hardware, maintenance, keeping up-to-date).
- Opening source does not bring much for a typical private user or even app developer: often too complex to adopt.
- Many research projects result in papers + never adopted prototypes.

Are there other options, too?

Grassroots middleware

Learn from grown-ups:

- CERN provides its open data in the form of virtual machines with custom software already included.

Grassroots middleware

Learn from grown-ups:

- CERN provides its open data in the form of virtual machines with custom software already included.
- For simpler cases, browser can suffice (a typical Firefox can run Windows 3.1 in the meantime)

- **Provide whatever is possible as open-source JS libraries.**
- Authors of middleware must get chance to publish their work on Open Data portals
  - something like developer's hub every large software project has

Grassroots middleware

Learn from grown-ups:

- CERN provides its open data in the form of virtual machines with custom software already included.
- For simpler cases, browser can suffice (a typical Firefox can run Windows 3.1 in the meantime)

- **Provide whatever is possible as open-source JS libraries.**
- Authors of middleware must get chance to publish their work on Open Data portals
  - something like developer's hub every large software project has

Pros:

- Motivate developers
- Can work both ways: E.g., format checker by an app developer may tell data provider if their CSV is good enough (and what to fix if not)

Cons:

- Limited scalability and complexity.

Solicit, Build, Use, Give credit to

Open Data Middleware!