Open Data Ecosystems

raphael.volz@hs-pforzheim.de

Nov 30, 2017

Contents

0.1 About Me

Figure 1:

- Professor for Applied Computer Science
- Faculty of Engineering
- Research Foci:
  - Data Science
  - Cloud Computing

0.2 Talk Outline

- Introduction
  - What is open data ?
  - What is a data ecosystem ?
- Community-driven open data ecosystems
  - Wikidata
  - OpenStreetMap
- Principles for successful Open Data Ecosystems

Chapter 1

Introduction

1.1 Burger King Ad

1.2 What is open data?

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

Source: http://opendefinition.org/

1.3 What is a data ecosystem?

A community of interacting organizations and individuals that produce, use and reuse a set of data. The dataset is the keystone around which applications and services provide value and thereby become part of the data ecosystem.

Ecosystem members can have various roles. Common roles are contributor, supplier, aggregator, enabler, enricher, developer as well as the common user.

Source: Own definition

1.4 High-level view

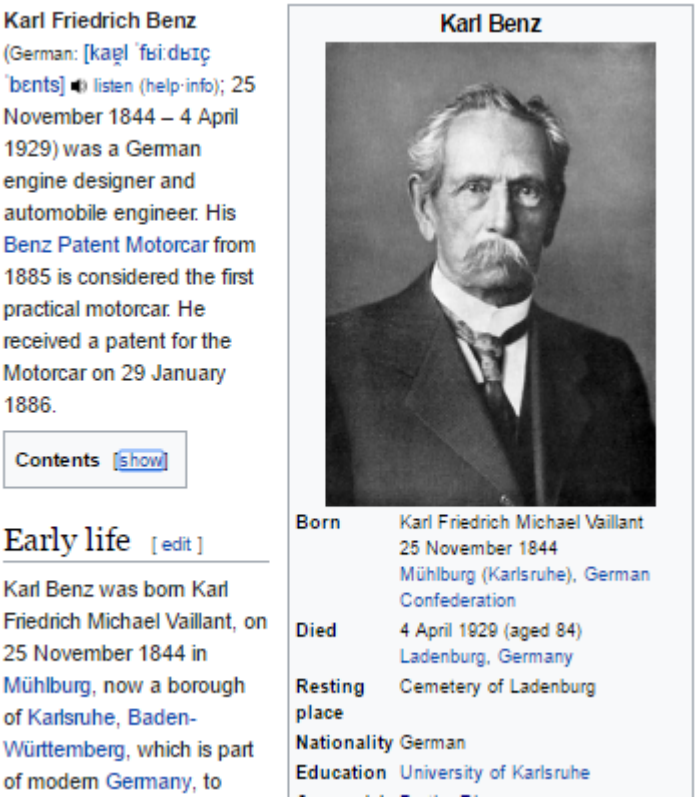
Figure 1.1: *Source: WHO eHealth*

Chapter 2

Community-Driven Open Data Ecosystems

2.1 Wikidata

2.2 Wikipedia Infoboxes



Karl Friedrich Benz
(German: [kaɐ̯l ˈfʁiːdʁɪç ˈbɛnts] listen (help·info); 25 November 1844 – 4 April 1929) was a German engine designer and automobile engineer. His Benz Patent Motorcar from 1885 is considered the first practical motorcar. He received a patent for the Motorcar on 29 January 1886.

Contents [show]

Early life [edit]

Karl Benz was born Karl Friedrich Michael Vaillant, on 25 November 1844 in Mühlburg, now a borough of Karlsruhe, Baden-Württemberg, which is part of modern Germany, to

Karl Benz

Born	Karl Friedrich Michael Vaillant 25 November 1844 Mühlburg (Karlsruhe), German Confederation
Died	4 April 1929 (aged 84) Ladenburg, Germany
Resting place	Cemetery of Ladenburg
Nationality	German
Education	University of Karlsruhe

Figure 2.1: Source: Karl Benz on Wikipedia

2.3 Wikidata Purpose

- Centralize the facts from Wikipedia info boxes
- For reuse across 300 Wikipedia languages
- e.g. 78 articles about Zika had different infoboxes
- For querying and use by third party apps
- Improve interwiki links

2.4 Wikidata properties

- a knowledge graph based on items
- free and open
- collaborative
- multilingual
- manually curated (unlike DBpedia)

2.5 Knowledge graph

People filmed with Jim Carry

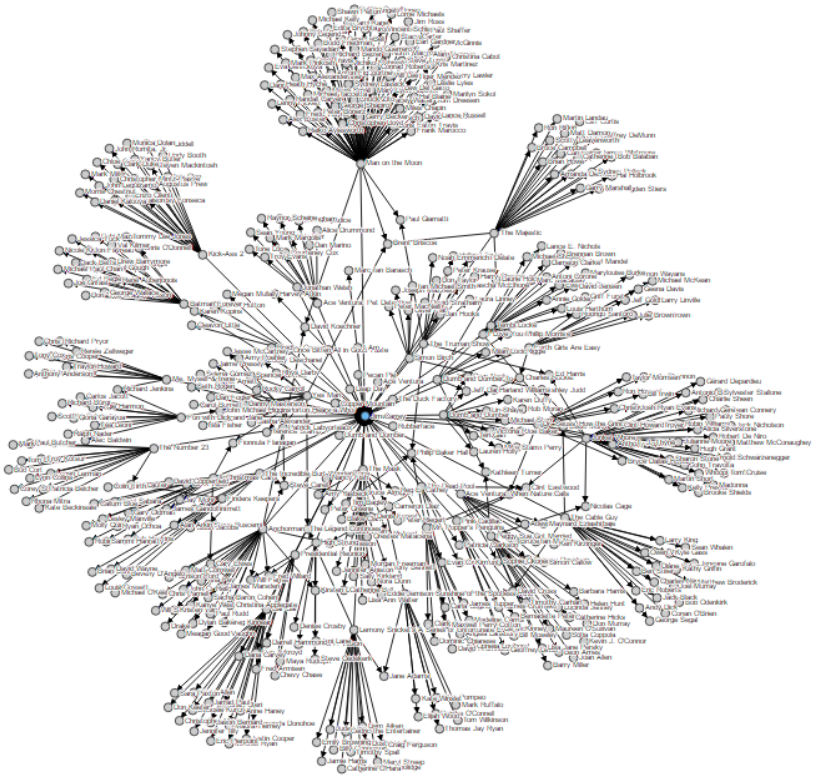


Figure 2.2: Source: Wikidata Graph Builder

2.6 Items have properties

Karl Benz

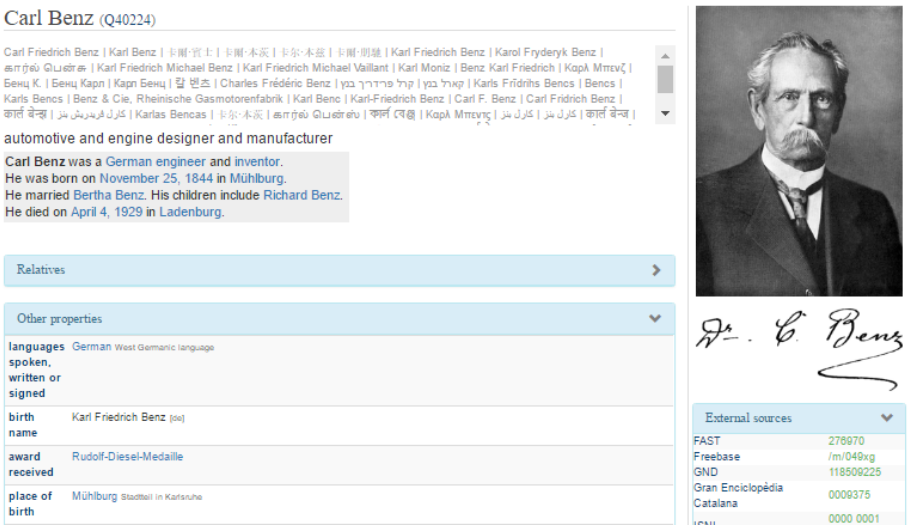


Figure 2.3: Source: Reasonator

2.7 Wikidata data model

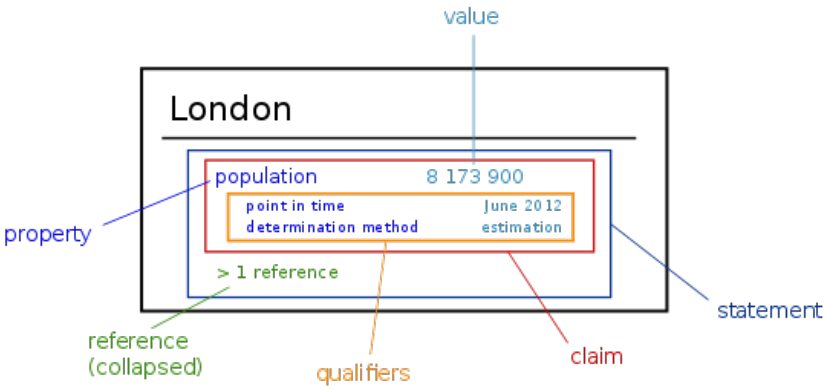


Figure 2.4: Source: Wikidata Data Model Primer

2.8 Wikidata size ~~June~~ November 2017

- ~~26.3~~ 38.6 Mio. items
- ~~150~~ 326 Mio. statements about items

- ~~500~~ 597 Mio. edits have been made since launch
- Currently >~~17~~ 18.6 Tsd. active users
- Observe growth in detail statistics

2.9 Wikidata ecosystem

- Various companies using the data
- Example Google:
  - Directly answers search requests, e.g. How high is the Eiffel tower?
  - Powers info boxes right next to search results
  - Donated data set Freebase to Wikidata
- Growing number of Applications and Services by third party developers

2.10 Wikidata Queries - Tables

What are the 10 largest cities with a female mayor ?

See Result. Modify Query.

```
SELECT DISTINCT  ?cityLabel ?population ?mayorLabel
WHERE
{
    ?city wdt:P31/wdt:P279* wd:Q515 .  # find instances of subclasses of city
    ?city p:P6 ?statement .            # with a P6 (head of goverment) statement
    ?statement ps:P6 ?mayor .           # ... that has the value ?mayor
    ?mayor wdt:P21 wd:Q6581072 .       # ... where the ?mayor has P21 (sex or gender) i
    FILTER NOT EXISTS { ?statement pq:P582 ?x }  # ... but the statement has no P582 (e

    # Now select the population value of the ?city
    # (wdt: properties use only statements of "preferred" rank if any, usually meaning
    ?city wdt:P1082 ?population .
    # Optionally, find English labels for city and mayor:
    SERVICE wikibase:label {
        bd:serviceParam wikibase:language "en" .
    }
}
ORDER BY DESC(?population)
LIMIT 10
```

2.11 Wikidata Queries - Images

List of space probes with pictures.

```
SELECT ?link ?date ?picture
 WHERE {
   ?link wdt:P31 wd:Q26529 ;
                     wdt:P18 ?picture ;
                     wdt:P619 ?date .
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "fr,en" .
   }
 }
ORDER BY ?date
LIMIT 100
```

See Result

2.12 Wikidata Queries - Maps

Map of Lighthouses in Norway

```
#defaultView:Map
SELECT DISTINCT ?item ?itemLabel ?coor ?image
{
  ?item wdt:P31 wd:Q39715 .
  ?item wdt:P17 wd:Q20 .
  OPTIONAL { ?item wdt:P625 ?coor }
  OPTIONAL { ?item wdt:P18 ?image }
  SERVICE wikibase:label
          {
    bd:serviceParam wikibase:language "nb,nn,en,fi" .}
}
ORDER BY ?itemLabel
```

See Result

2.13 OpenStreetMap (OSM)

2.14 OSM Background

- OpenStreetMap (OSM) was created in July 2004 by Steve Coast (a UCL student)
- He did not understand why the Ordnance Survey created massive geographical datasets but did not freely distribute them to those who had paid to create them

- GeoData only freely available in some countries, e.g. the US and the Netherlands

2.15 OSM Properties

- Collaborative
  - maintained by individual contributors
  - Wikipedia principle, everyone can edit and contribute
- Donated data sets imported in bulk (particularly Eastern Europe)
- Automated robots cleaning data
- **Open Data**, free to use (under OdBL license)
- **Not a map, but a database**

2.16 OSM hypergraph

- *Nodes*: basic geographic point.
  - Geographic point: latitude & longitude (WGS84)
  - Point Of Interest (POIs)
- *Ways*: ordered interconnection of nodes
  - open ways = linear features (roads, railways. . .)
  - closed ways = areas
- *Relations*: group of any primitive with associated roles
  - Relate nodes, ways and potentially other relations to each other,
  - thereby forming complex objects (multipolygons)
- Nodes, ways, relations are versioned and user attributed

2.17 OSM Elements

Each OSM entity (node, way, relation) has:

- a numeric identifier: OSM ID
- a set of generic attributes present for every element
  - uid, user: user id and user name
  - timestamp: time of the last modification
  - visible: if false then the element should only be returned by history calls
  - version: edit version of the object (starts from 1)
  - changeset: the changeset (group of edits made within a certain time by one user) in which the object was created or updated
- a set of tags (key-value pairs)

2.18 OSM Tags / Ontology

- key-value pairs
- e.g. highway=residential
- use of tags and values is not restricted
- defines the basic ontology of OSM
- see taginfo

2.19 OSM statistics

Metric	July 2016	June 2017	Nov 2017
Users	2,867,221	3,954,309	4,402,229
Nodes	3,463,959,970	3,926,828,147	4,197,365,421
Ways	360,469,340	416,654,804	454,113,805
Relations	4,387,699	5,043,226	5,390,806
GPS traces	5,280,183,660	5,715,425,150	5,953,688,363

Top user has contributed 326,511,847 (6%) GPS traces.

Source

2.20 OSM Open Data Ecosystem

- Many applications using the data
- Many services based on the data
- Many (open source) tools for handling the data
- Primary application areas
  - Map Rendering (One Dataset, several renderings)
  - Geo Search (POI, (Reverse) Name Resolution)
  - Routing
  - Geographic Database
  - Data Editors

2.21 Map Rendering

Map Compare

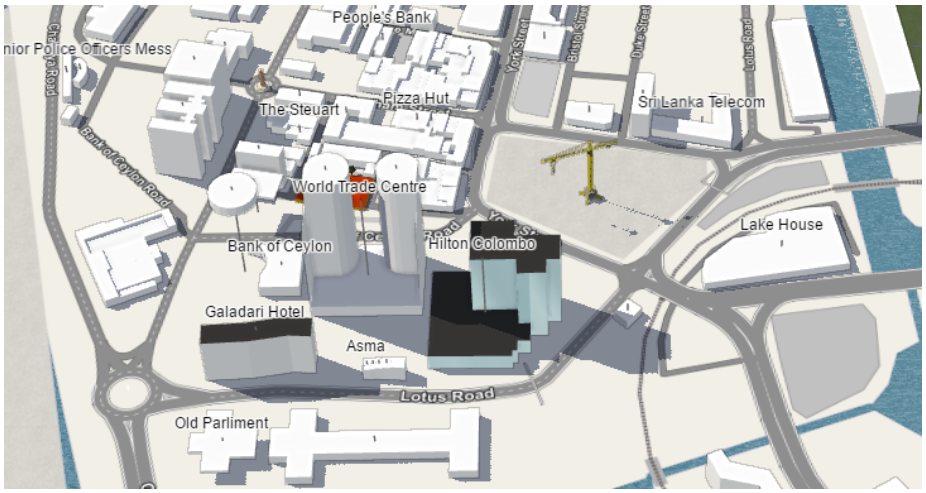


Figure 2.5: Source: F4 Map

2.22 Geo Search

Nominatim

- search for a name or address (forward search)
- look up data by its geographic coordinate (reverse search)
- Each result comes with a link to a details page where you can inspect the data
- Debug info to investigate how the address of the object has been computed.
- Available Open Source
- In Production on main site and available via APIs

2.23 Routing

- Many (open source) applications and services available
- My favorite: Graphhopper
  - Including route optimization and distance matrix services
  - Isochrone (reach calculation)
  - Matrix API
  - Route Optimization (Travelling Salesman Problem)
  - Snap to Road
  - Open Source (with commercial version and hosting service)

2.24 Data Service - Overpass

Example Query: Chinese Restaurants on the map

```
node
  [amenity=restaurant]
  [cuisine=chinese]
  ({{bbox}});
out;
```

Run Overpass Query.

2.25 Overpass Query: Streets on the map

```
way({{bbox}})
  [highway]
  [name];
out;
```

Run Overpass Query.

2.26 Case Study WheelMap.org

Figure 2.6: Source: http://wheelmap.org

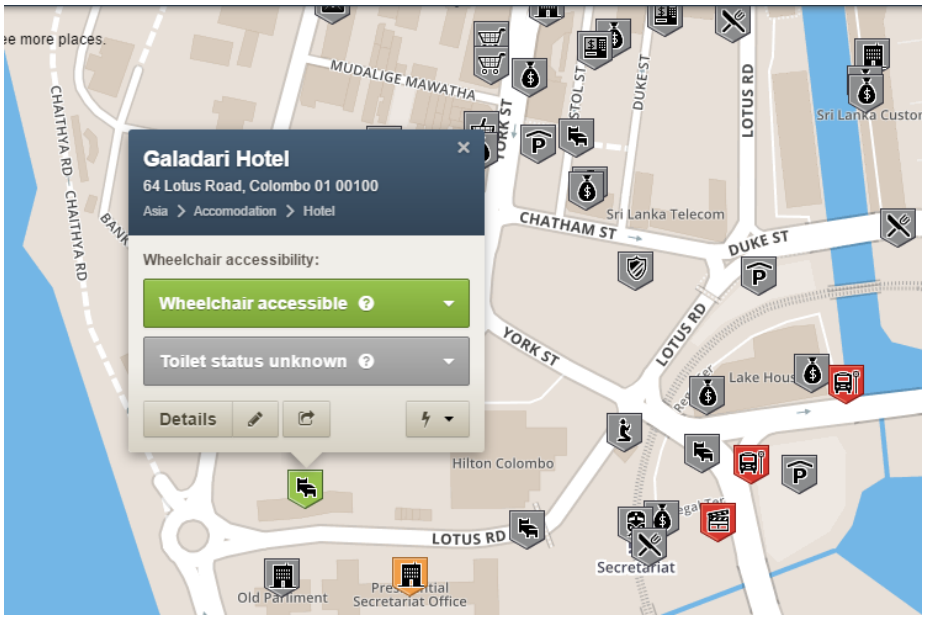


Figure 2.7: Source: http://wheelmap.org

Removed level=14 tag from building.

Edited 3 months ago by Sergey Kozyr
Version #10 · Changeset #46438782

Tags

addr:city	Colombo 01
addr:housenumber	64
addr:postcode	00100
addr:street	Lotus Road
building	yes
building:levels	16
building:material	traditional
building:obm	263768173

height	48
name	Galadari Hotel
phone	+94 11 2 544544
roof:colour	#333333
roof:material	asbestos
roof:shape	flat
smoking	separated
tourism	hotel
website	http://www.galadarihotel.lk/
wheelchair	yes

Figure 2.8: Source: OpenStreetMap

2.27 Conclusion

- Community-driven data ecosystems thriving
- High and increasing partipation
- Contribution to data sets at the core of the ecosystem
- Major issues:
  - Incompatibility of licenses (ODbl vs. CC)
  - No global identifiers (linking data sets is still hard), Wikidata providing a basis for bridging ids
  - New community-driven data ecosystems will probably be domain-focused (winner takes all)

Chapter 3

Success Factors for Open Data Ecosystems

3.1 Success Factors

- Adopt emerging best practises, see [1, 2, 3, 4, 5, 6, 7]
- Define a priority domain
- Create **linkable data** with global identifiers (URIs)
- Allow **extensions** by third party contributions
- Use **feedback cycles** including versioning
- **Track progress** with statistics of agreed on metrics
- **Manage the community**
- **Disseminate!**

3.2 Data ecosystems are all about people

Recognition [edit]

Manske is recognized as the creator of the first article in the German Wikipedia, which was on the polymerase chain reaction, first written by him in 2001.[30][31][32]

Jimmy Wales in 2002 named 25 January as Magnus Manske day in honor of his contributions to Wikipedia, proclaiming that "Tonight at dinner, every Wikipedian should say a toast to Magnus and his many inventions."[33] Larry Sanger, in his memoir on the early history of Wikipedia, highlighted the contributions of Manske to the project and attributes the eventual success of Wikipedia to a core group of actors, with Manske playing an important role:

> Wikipedia started with a handful of people, many from Nupedia. The influence of Nupedians was crucial early on. I think, especially, of the tireless Magnus Manske (who worked on the software for both projects), our resident stickler Ruth Ifcher, and the very smart poker-playing programmer Lee Daniel Crocker—to name a few...Still, because the project started with these good people, and we were able to adopt, explain, and promote good habits and policies to newer people, the Nupedian roots of the project helped to develop a robust, functional and successful community.[34]

Manske at a Wikidata event in 2015.

Manske, along with others, was recognized as a major contributor to MediaWiki by the USENIX Advanced Computing Technical Association in 2010, when MediaWiki and the Wikimedia Foundation were honoured with a STUG award (Software Tools User Group).[35]

Figure 3.1: *Source: https://en.wikipedia.org/wiki/Magnus_Manske*

Chapter 4

Thank you for your attention! Questions ?