Open Data on the Web as the fuel for Cognitive Computing

Das Potenzial von Open Data für Cognitive Computing

Axel Polleres

web: http://polleres.net  twitter: @AxelPolleres
Why are we here today?

23.3.2017 16:30 bis 23.03.2017 18:30
KICK-OFF: DIGITAL INFORMATION MANAGEMENT COMMUNITY AUSTRIA

Datum 23.03.2017 16:30 - 23.03.2017 18:30
Ort Impact Hub Vienna

künstliche Intelligenz
Cognitive Computing

Mit welchen Themen sollten wir uns befassen?

Welche Lösungen sind im Einsatz?
Cognitive computing (CC) describes technology platforms that, broadly speaking, are based on the scientific disciplines of artificial intelligence and signal processing. These platforms encompass machine learning, reasoning, natural language processing, speech and vision, human–computer interaction, dialog and narrative generation, among other technologies.[1][2]

Contents
- Definition
- Use cases
- See also
- References
- Further reading

Definition
At present, there is no widely agreed upon definition for cognitive computing in either academia or industry.[3][4]
1. What is Cognitive Computing?

Dr. Chris Welty, now Google, formerly IBM Research (Watson team)

“Help computers perform tasks we can do really quickly, at a **scale** that humans can’t”

“Computational tasks that require **inexact** solutions that **combine multiple methods** in unpredictable ways.”

“Machine intelligence is not human intelligence, the difference lies in the errors it makes.”

https://www.wu.ac.at/wutv/show/clip/20141024-welty/
John Launchbury, the Director of DARPA's Information Innovation Office

Video, published on Feb 15, 2017
2. What is Artificial Intelligence?

John Launchbury, the Director of DARPA's Information Innovation Office
Video, published on Feb 15, 2017
2. What is Artificial Intelligence?

John Launchbury, the Director of DARPA's Information Innovation Office Video, published on Feb 15, 2017

“learn from data, like spreadsheets on steroids”
2. What is Artificial Intelligence?

John Launchbury, the Director of DARPA's Information Innovation Office

Video, published on Feb 15, 2017

The (future) third wave of AI

Contextual Adaptation

Systems construct explanatory models for classes of real world phenomena
2. What is Artificial Intelligence?

John Launchbury, the Director of DARPA's Information Innovation Office Video, published on Feb 15, 2017

1. **needs** – if you want a combination of first and second wave AI – rules/explanations to model common sense and act in new contexts

2. **Again, data is at it’s heart!**
Which solutions exist now on the Web?

Welche Lösungen sind im Einsatz?

Example 1: Google’s Knowledge Graph
Which solutions exist now on the Web?

Welche Lösungen sind im Einsatz?

Example 2: FB’s Social Graph & News Recommendations

Also uses a knowledge graph...
Which solutions exist now on the Web?

Example 3: IBM Watson!

Also uses a knowledge graph...

https://youtu.be/P0Obm0DBvwI?t=951
This is the knowledge Graph that IBM Watson used:

http://lod-cloud.net/
3. What is Open Data?

**Availability and Access**: data must also be available in a convenient and modifiable form, at no more than a reasonable reproduction cost.

**Reuse and Redistribution**: data must be provided under terms that permit reuse and redistribution and must be **machine-readable**.

**Universal Participation**: everyone must be able to use, reuse and redistribute it (**no discrimination**)

See more at: [http://opendefinition.org/okd/](http://opendefinition.org/okd/)
Open Data is a global trend:

- Cities, International Organizations, National and European Portals, Int'l. Conferences:

Open Data Portals

CKAN … http://ckan.org/

• almost „de facto“ standard for Open Data Portals
• facilitates search, metadata (publisher, format, publication date, license, etc.) for datasets

• http://opendataportal.at/
• http://data.gv.at/

• machine-processable? … partially
Our research: data.wu.ac.at

- **What is the status of Open Data and what are the challenges using Open Data?**
  - OpenData PortalWatch – a project at WU

- **How can Open Data be used by enterprises?**
  - Open City Data Pipeline – a joint project with Siemens on using Open Data in an Enterprise context!

- **What's next?**
  - Improving Open Data Quality and Access: ADEQUATE (FFG - project)
  - Making Open Data Searchable
  - Building an Open Data **Knowledge Graph**!

- A striving **Data Economy** needs no silos... re-democratise the Web by Cognitive Intelligence based on Open Data?
**Open Data Portal Watch**

http://data.wu.ac.at/portalwatch/

- Periodically monitoring a list of Open Data Portals
  - 260 CKAN powered Open Data Portals worldwide
- Quality assessment
- Evolution tracking
  - Meta data
  - Data
  - Formats, growth
What have we learnt?
Open Data also has the "Vs"

- **Volume:**
  - It's growing! (we currently monitor 90 CKAN portals, 512543 resources/ 160069 datasets, at the moment (statically) ~1TB only CSV files...)

- **Variety:**
  - different datasets (from different cities, countries, etc.), only partially comparable, partially not.
  - Different metadata
  - Different data formats

- **Velocity:**
  - Open Data changes regularly (fast and slow)
  - New datasets appear, old ones disappear
Portalwatch Example:

Portal: open.whitehouse.gov

Portal Evolution

Portal Size

In reply to maxwell ogden

Axel Polieres @AxelPolieres · Feb 17
@denormalize documentation of the quite drastic the white house' #opendata policy change visualized: data.wu.ac.at/portalwatch/… pic.twitter.com/6BQwxTWAeP
Our research: data.wu.ac.at

- **What is the status of Open Data and what are the challenges using Open Data?**
  - OpenData PortalWatch – a project at WU

- **How can Open Data be used by enterprises?**
  - Open City Data Pipeline – a joint project with Siemens on using Open Data in an Enterprise context!

- **What's next?**
  - Improving Open Data Quality and Access: ADEQUATE (FFG - project)
  - Making Open Data Searchable
  - Building an Open Data **Knowledge Graph**!

- A striving **Data Economy** needs no silos... re-democratise the Web by Cognitive Intelligence based on Open Data?
Use Case: City Data – Important for Infrastructure Providers & for City Decision Makers

• City Assessment and Sustainability reports

• Tailored offerings by Infrastructure Providers

→ Needs up-to-date City Data and calculates City KPIs in a way that allows to display the current state and run scenarios of different product applications.

  e.g. towards a “Dynamic” Green City Index:

... however, these are often outdated before even published!

Goal (short term):
- Leverage Open Data for calculating a city’ performance from public sources on the Web automatically

Goal (long term):
- Define and Refine KPI models to assess specific impact of infrastructural investments and gather/check input automatically
City Data Pipeline (started 2012)

- http://citydata.wu.ac.at/

Open City Data Pipeline

We present the City Data Pipeline – a system for gathering city performance indicators published as Open Data in order to ease the compilation of studies and reports used within Siemens. Under the assumption that Open Data provides means to automate tedious data research tasks, we have built a system that integrates basic indicators for cities from various Open Data sources. The architecture is flexible, extensible, and natively based on RDF & SPARQL.

Launch Open City Data Pipeline
City Data Pipeline: could be viewed as a cognitive computing use case

Goal: understanding, combining, enriching, different open datasets using both first-wave and second wave AI methods
City Data Pipeline

citydata.wu.ac.at

- Search for indicators & cities
- obtain results incl. sources
- Integrated data served as Linked Open Data
- Predicted values AND estimated error rates for missing data...

Vienna

Municipal waste (1000 t)

- **2004**: 778.905392176222 1000 t (from [http://citydata.wu.ac.at/ns#Prediction](http://citydata.wu.ac.at/ns#Prediction), predicted by with an estimated error of %RMSE)
- **2005**: 813.77643147163 1000 t (from [http://citydata.wu.ac.at/ns#Prediction](http://citydata.wu.ac.at/ns#Prediction), predicted by with an estimated error of %RMSE)
- **2006**: 813.889824195497 1000 t (from [http://citydata.wu.ac.at/ns#Prediction](http://citydata.wu.ac.at/ns#Prediction), predicted by with an estimated error of %RMSE)
- **2007**: 811.538914636665 1000 t (from [http://citydata.wu.ac.at/ns#Prediction](http://citydata.wu.ac.at/ns#Prediction), predicted by with an estimated error of %RMSE)
- **2008**: 811.010344391444 1000 t (from [http://citydata.wu.ac.at/ns#Prediction](http://citydata.wu.ac.at/ns#Prediction), predicted by with an estimated error of %RMSE)
- **2009**: 811.172539873368 1000 t (from [http://citydata.wu.ac.at/ns#Prediction](http://citydata.wu.ac.at/ns#Prediction), predicted by with an estimated error of %RMSE)

...it's not finished, but: assumption: Predictions get better, the more Open data we integrate...

Open Data: The more, the merrier!
What is the status of Open Data and what are the challenges using Open Data?
- OpenData PortalWatch – a project at WU

How can Open Data be used by enterprises?
- Open City Data Pipeline – a joint project with Siemens on using Open Data in an Enterprise context!

What's next?
- Improving Open Data Quality and Access: ADEQUATE.at
- Making Open Data Searchable
- Building an Open Data Knowledge Graph!

A striving Data Economy needs no silos... re-democratise the Web by Cognitive Intelligence based on Open Data?
Why is Search in Open Data a problem?

Structured Data in Web Search by Alon Halevy

Open Data Search is hard...

a) No natural language "cues" like in Web tables...

b) Existing knowledge graphs don’t cover the domain of "Open Data"

c) Open Data is not properly geo-referenced
Some starting points:

- First baby steps on building an Open Data Knowledge Graph:
- Ongoing work to make Open Data geo-searchable e.g. in our project comunidata.at:

International Semantic Web conference 2016:

**Multi-level semantic labelling of numerical values**

Sebastian Neumair1, Jürgen Umbrich1, Josiane Xavier Parreiras2, and Axel Polleres1

1 Vienna University of Economics and Business, Vienna, Austria
2 Siemens AG Österreich, Vienna, Austria

Abstract. With the success of Open Data a huge amount of tabular data sources became available that could potentially be mapped and linked into the Web of (Linked) Data. Most existing approaches to “semantically label” such tabular data rely on mappings of textual information to classes, properties, or instances in RDF knowledge bases in order to link – and eventually transform – tabular data into RDF. However, as we will illustrate, Open Data tables typically contain a large portion of numerical columns and/or non-textual headers; therefore solutions that solely focus on textual “cues” are only partially applicable for mapping such data sources. We propose an approach to find and rank candidates of semantic labels and context descriptions for a given bag of numerical values. To this end, we apply a hierarchical clustering over information taken from DBpedia to build a background knowledge graph of possible “semantic contexts” for bags of numerical values, over which we perform a nearest neighbour search to rank the most likely candidates. Our evaluation shows that our approach can assign fine-grained semantic labels, when there is enough supporting evidence in the background knowledge graph. In other cases, our approach can nevertheless assign high level contexts to the data, which could potentially be used in combination with other approaches to narrow down the search space of possible labels.
Ongoing Projects (data.wu.ac.at)

**WU Open Data Portal**
WU lectures, rooms and organizations data.wu.ac.at is an Open Data portal where you can find data about lectures, rooms and organizations at WU.
- 121 datasets

**Open Data Portal Watch**
Monitoring & exposing portals' metadata
Open Data Portal Watch assesses the evolution of the (meta) data quality of about 260 Open Data portals over since September 2014.
- 259 portals

**DBpedia Wayback Machine**
Extract past DBpedia versions
The DBpedia Wayback Machine aims at providing the wayback functionality for DBpedia based on the revisions of their Wikipedia article.

**Jupyter Notebook Server**
Programming & Documentation
Notebook documents are documents which contain both computer code (e.g., python) and human-readable rich text elements.
- Only available within local WU Vienna network

**CSV Engine**
Search & enrich CSVs
The CSV Engine is a collection of tools and services for processing and enriching CSV files.

**Open Data AT Assistant**
Search chatbot for Austrian datasets
The assistant will help you to explore the content of the austrian open data portals: data.gv.at and opendataportal.at.
Our research: data.wu.ac.at

- **What is the status of Open Data and what are the challenges using Open Data?**
  - OpenData PortalWatch – a project at WU

- **How can Open Data be used by enterprises?**
  - Open City Data Pipeline – a joint project with Siemens on using Open Data in an Enterprise context!

- **What's next?**
  - Improving Open Data Quality and Access: ADEQUATE.at
  - Making Open Data Searchable
  - Building an Open Data **Knowledge Graph**!

- A striving **Data Economy** needs no silos... re-democratise the Web by Cognitive Intelligence based on Open Data?
"Both our knowledge graph and google’s have roots in Wikipedia and freebase" – but none of Google and FB make their knowledge graphs freely and openly available again as Open Data!
This is a fundamental threat to the Web itself:


1) We’ve lost control of our personal data

2) It’s too easy for misinformation to spread on the web

3) Political advertising online needs transparency and understanding
Open-Data-Fueled Cognitive Computing to the rescue!

Open-Data-Fueled Cognitive Computing to the rescue!

Roland Ledinger @Roland_Ledinger · 3. März
Business Treff Open Data Day Statement
von Staatssekretärin Duzdar offene Daten sind für Alle wichtig

Axel Polleres @AxelPolleres · Mar 3
Staatssekretärin Muna #Duzdar at #opendatataday: #fakenews? #opendata to the rescue!
Bottomline: Let’s build Open Knowledge Graphs as a basis for Cognitive Computing for all businesses!

23.3.2017 16:30 bis 23.3.2017 18:30
KICK-OFF: DIGITAL INFORMATION MANAGEMENT COMMUNITY AUSTRIA

Datum 23.3.2017 16:30 - 23.3.2017 18:30
Ort Impact Hub Vienna

künstliche Intelligenz
Cognitive Computing

Mit welchen Themen sollten wir uns befassen?
Welche Lösungen sind im Einsatz?

Danke!

#OpenData
Eine Einladung...

Wenn Sie mehr zu den Themen **Semantic Web, Cognitive Computing, Linked Open Data** erfahren wollen:

![ISWC 2017 Conference Logo](https://iswc2017.semanticweb.org/home)


[https://iswc2017.semanticweb.org/](https://iswc2017.semanticweb.org/)
Things I did NOT have time to talk about:

- Open Data and licences → DALICC
- Open Data adoption barriers → see our recent paper to be presented at CEDEM2017
- Privacy and data on the Web → http://privacylab.at