

Propojená data a příklady jejich využití

Jakub Klímek, Martin Nečaský



MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova



Propojená data jsou:

- Sada best practices pro
 - Publikaci dat
 - Sdílení dat
 - Propojování dat
- Sada webových standardů pro
 - Reprezentaci dat
 - Modelování dat
 - Přístup k datům
 - Dotazování v datech

**BEST
PRACTICE**



Propojená data jsou:

- Grafová data
- Sémanticky popsaná data
- Další (nadějný) kandidát na řešení problému datové integrace



Propojená data se hodí tam, kde:

- Existuje sada heterogenních datových zdrojů
 - Je potřeba v nich udělat pořádek
 - Je potřeba s nimi jednotně pracovat



- S velkou výhodou: Princip **Pay-as-you-go**
 - nízká iniciační investice
 - lze začít s něčím malým, levným a jednoduchým a postupně rozšiřovat dle potřeby
 - není třeba dopředu znát cílové datové schéma

Propojená data se hodí tam, kde:

- Jsou data publikována na Internetu



<http://www.doublejdesign.co.uk>

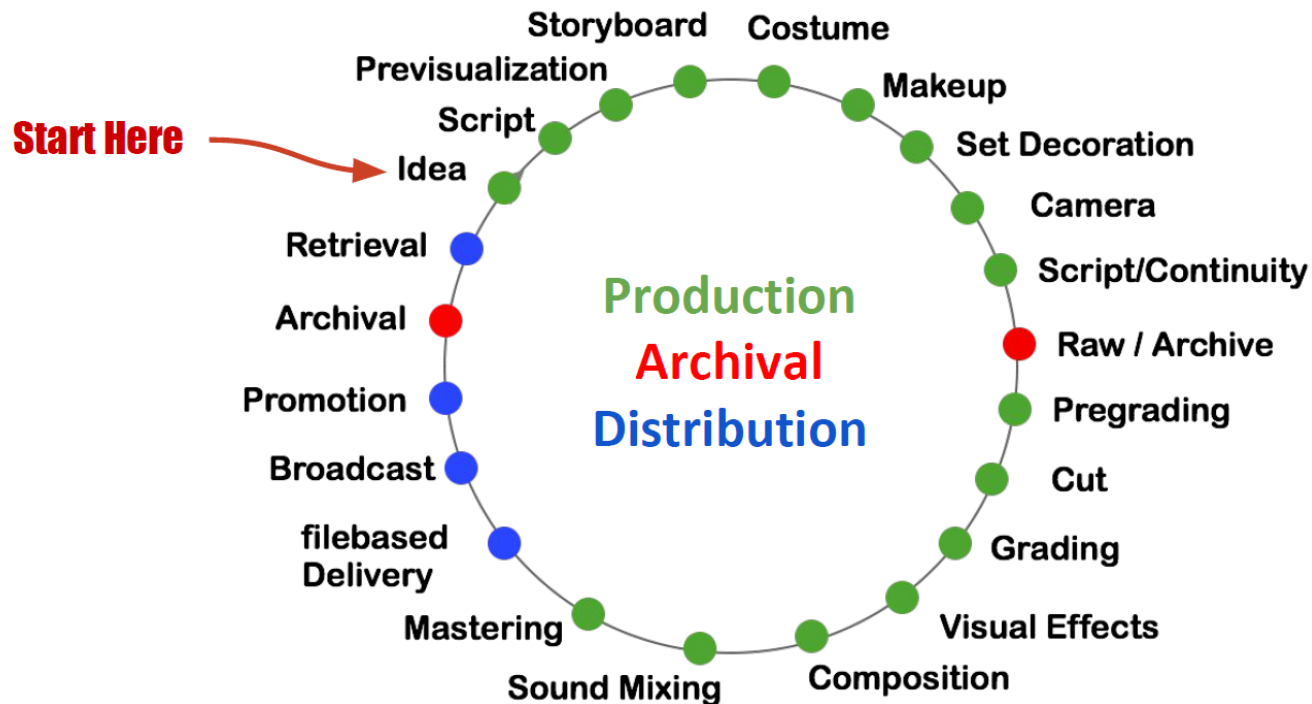
LD: Otevřená i uzavřená

- Propojená **uzavřená** data
 - Linked Closed Data (LCD), Linked Enterprise Data (LED)
 - Enterprise Knowledge Graphs (EKG)
 - Obdoba dnešních Intranetů - stránek pro lidi uvnitř organizace
- Propojená **otevřená** data
 - Linked Open Data (LOD)
 - Web of Data
 - Obdoba dnešního Webu dokumentů - stránek pro lidi na Internetu

Enterprise Knowledge Graphs

- Studia [Babelsberg](#) - projekt [dwerft](#)

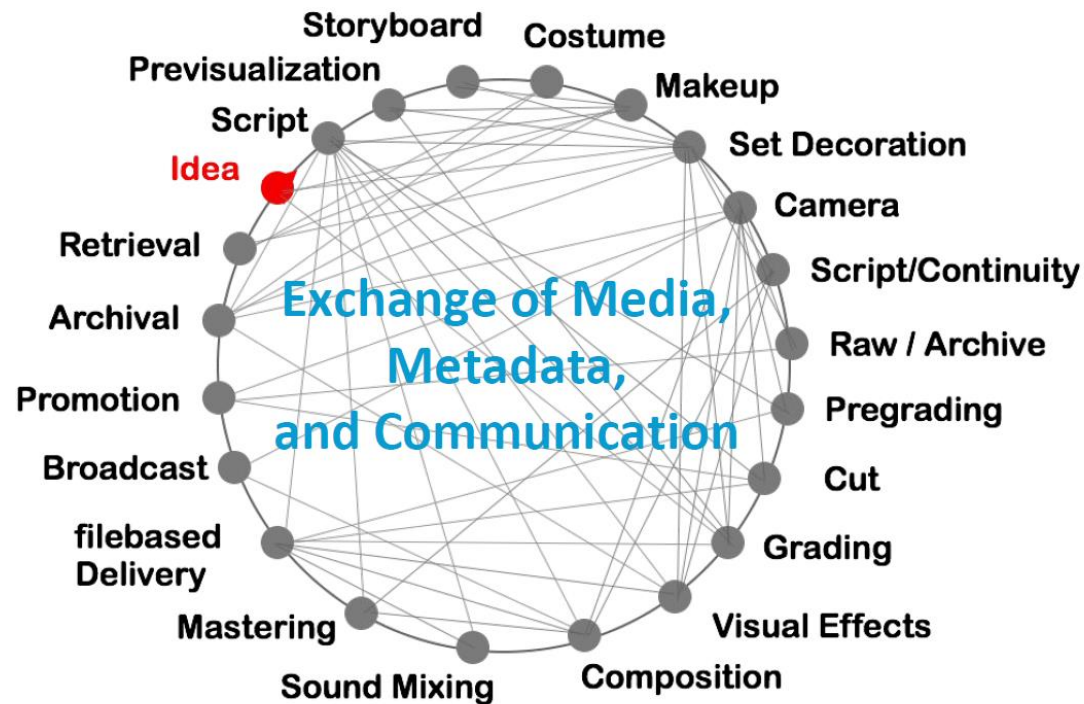
The Media Value Chain



Enterprise Knowledge Graphs

- Studia [Babelsberg](#) - projekt [dwerft](#)

The Media Value Chain

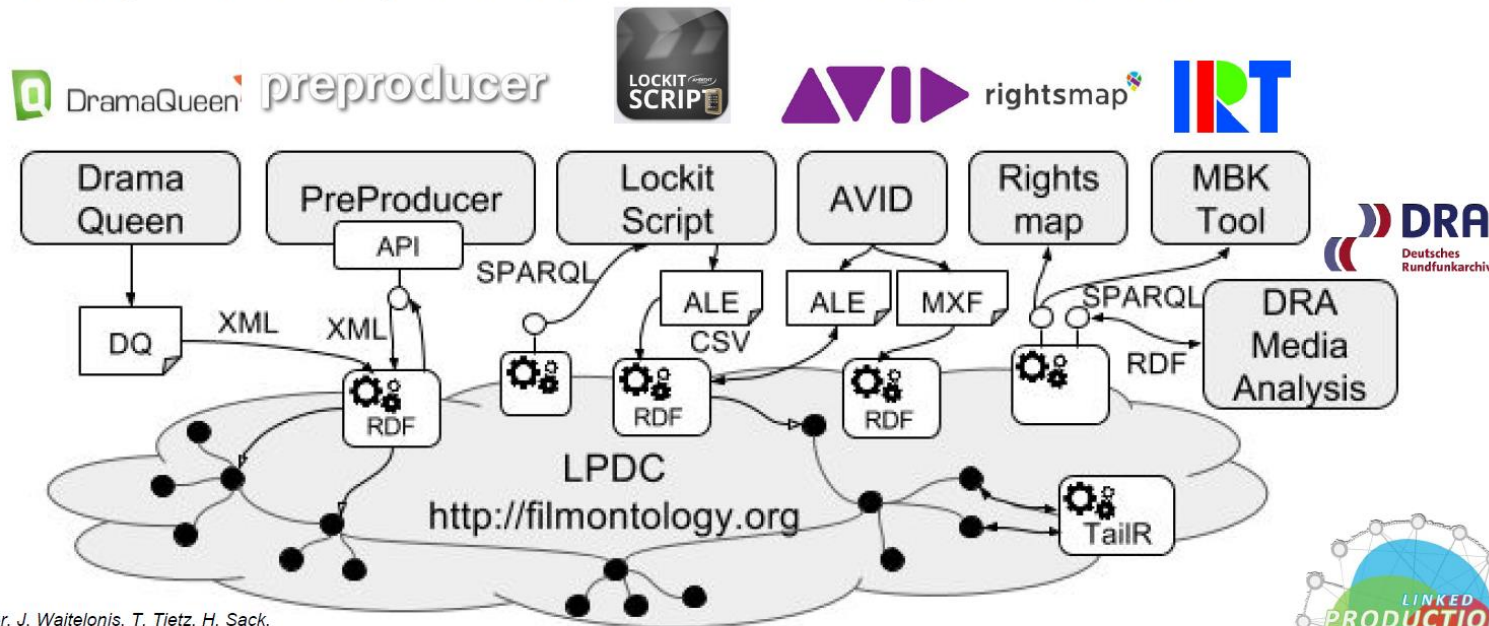


Enterprise Knowledge Graphs

- Studia Babelsberg - projekt dwerft

dwerft Master Usecase

Data Integration for Pre-production, Production & Post-production tools



H. Agt-Rickauer, J. Waitelonis, T. Tietz, H. Sack,
Data Integration for the Media Value Chain, ISWC 2016

CC BY-NC-SA 4.0

Harald Sack, 7th DBpedia Community Meeting 2016, Sep. 15, 2016, Leipzig



Enterprise Knowledge Graphs

The image shows a Google search for "da vinci". The search bar at the top contains "da vinci" and the Google logo. Below the search bar, the "All" tab is selected, showing "About 75,200,000 results (0.82 seconds)". A Knowledge Graph for Leonardo da Vinci is displayed on the left, featuring a network of interconnected nodes representing his various roles and achievements. The main search results area shows a grid of images of Leonardo da Vinci, including portraits and his Vitruvian Man drawing. Below the images, the name "Leonardo da Vinci" is displayed, followed by the title "Mathematician". A brief biography states: "Leonardo di ser Piero da Vinci, more commonly Leonardo da Vinci or simply Leonardo, was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, ... Wikipedia". Below the biography, key facts are listed: "Born: April 15, 1452, Anchio", "Died: May 2, 1519, Amboise, France", "Buried: Chapel of Saint-Hubert", and "Structures: Milan Cathedral".

Google

da vinci

Google Inside Search

Home How Search Works Tips & Tricks **Features** Search Stories Playground

All Images Videos Maps News More ▾ Search tools

About 75,200,000 results (0.82 seconds)

Leonardo da Vinci

Mathematician

Leonardo di ser Piero da Vinci, more commonly Leonardo da Vinci or simply Leonardo, was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, ... [Wikipedia](#)

Born: April 15, 1452, Anchio

Died: May 2, 1519, Amboise, France

Buried: Chapel of Saint-Hubert

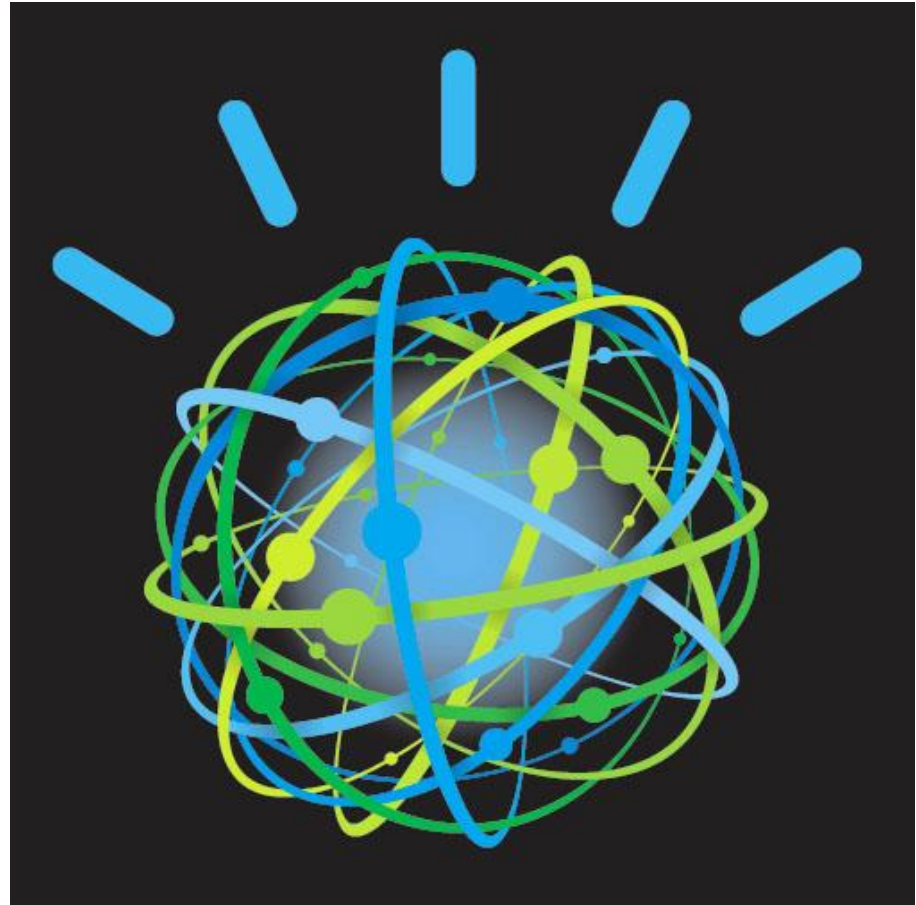
Structures: Milan Cathedral

The Knowledge Graph

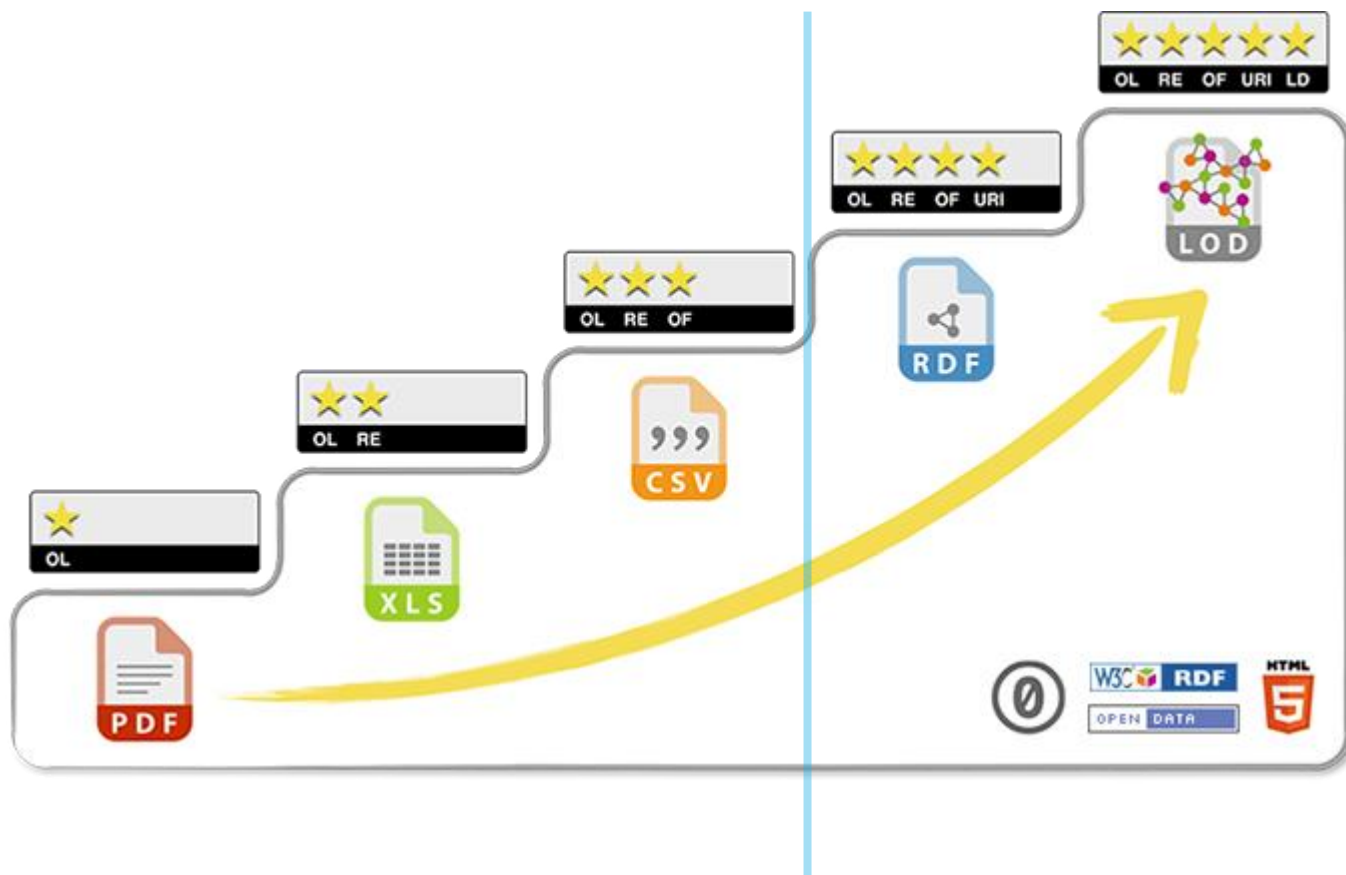
Learn more about one of the key breakthroughs behind the future of search.

Enterprise Knowledge Graphs

- IBM Watson
 - Soutěžil ve hře Jeopardy (Riskuj)
 - Využívá (mimo jiné) sémantických technologií a propojených dat



Co jsou otevřená propojená data?



Zdroj : <http://5stardata.info>

Propojená otevřená data ve Velké Británii

- v experimentálním provozu řada zdrojů otevřených propojených dat
 - <http://environment.data.gov.uk> (Ministerstvo životního prostředí)
 - <http://landregistry.data.gov.uk> (Katastr nemovitostí)
 - <http://statistics.data.gov.uk/> (Statistický úřad)
 - <https://datahub.io/dataset/reference-data-gov-uk> (referenční data, např. časové úseky nebo orgány veřejné správy)
- BBC Things
 - <http://www.bbc.co.uk/things/ed9d1ef3-eded-4f81-b158-be49cfc1ea8f>

Propojená otevřená data ve Velké Británii

- <http://environment.data.gov.uk> (Ministerstvo životního prostředí)
 - [API](#) pro sledování kvality vody na koupání



Bathing Water Quality API

Open data from the Environment Agency



The Environment Agency collects water quality data each year from May to September, to ensure that designated bathing water sites on the coast and inland are safe and clean for swimming and other activities. We are happy to make this data reusable and accessible to developers and to members of the public, by publishing it as linked data.

Propojená data (už i) v ČR

- Česká správa sociálního zabezpečení (ČSSZ)



ČESKÁ SPRÁVA SOCIÁLNÍHO ZABEZPEČENÍ
Otevřená data

Základní informace

Katalog otevřených dat

Podmínky užití

← Počet nově přiznaných důchodů v České republice dle měsíční výše důchodu

Počet nově přiznaných důchodů v České republice podle roku, druhu důchodu, měsíční výše důchodu a pohlaví

Periodicita aktualizace	Jednou ročně
URI souvisejícího geografického území	http://ruian.linked.opendata.cz/resource/staty/1
Související geografické území	Česká republika
Klasifikace dle EuroVoc	důchodové pojištění podmínky pro odchod do důchodu
Jméno a e-mail kurátora dat	Ing. Jiří Šunka
Odkaz na dokumentaci datové sady	https://data.cssz.cz/documentation/nove-priznane-duchody-dle-vyse-duchodu
Dotčené časové období od-do	01.01.2008 - 31.12.2015
Název a URL poskytovatele dat	Česká správa sociálního zabezpečení

Datové zdroje

Data	Schéma	Podmínky užití	Dotčené časové období
Data (RDF/Trig) 11,9 MB	Schéma (RDF/Trig) 16,6 kB	podmínky užití	01.01.2008 - 31.12.2015
Data (CSV) 1,1 MB	Schéma (JSON) 1,5 kB	podmínky užití	01.01.2008 - 31.12.2015

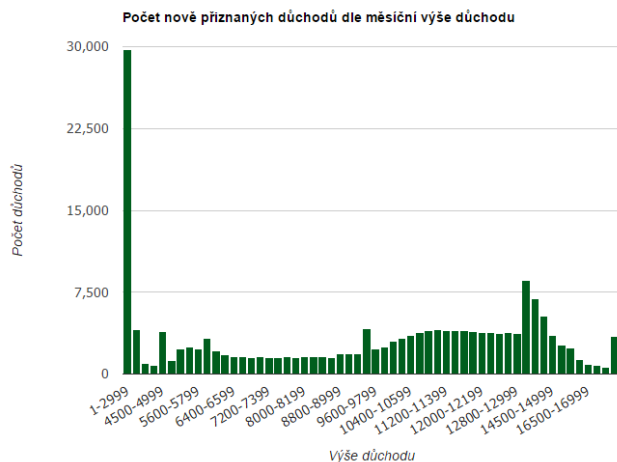
Data v RDF jsou přístupná i přes SPARQL endpoint, který je na URL <https://data.cssz.cz/sparql>

Pro získání výběrů dat z datových sad je připraven jednoduchý **formulář**, který umožňuje zadávání dotazů ve SPARQL jazyku.

Metadata o datové sadě v RDF (Trig) 7,6 kB

Vizualizace dat - grafy

Počet nově přiznaných důchodů dle měsíční výše důchodu



Propojená data (už i) v ČR

- Ministerstvo financí (MFČR) a Generální finanční ředitelství (GFŘ)
- Centrální evidence dotací (CEDR III)
- <http://cedr.mfcr.cz/cedr3internetv419/OpenData/DocumentationPage.aspx/>

The screenshot displays the IS CEDR III website. The header includes the logo of the Ministry of Finance (F. FINANČNÍ SPRÁVA) and the title IS CEDR III. There are links for English and Czech, a login button, and a contact link. The main content area is titled 'Otevřená data' and features a sidebar with navigation links. The main text explains that the General Financial Directorate, in accordance with a government decision, is publishing data from the Central Evidence of Grants (CEDR III) in a machine-readable format (Open Data). It mentions that the data is available for use, modification, and distribution, and that it is published in RDF format. A note at the bottom states that the data is available for use by specialists (programmers, data analysts, journalists, etc.) and that it is published in RDF format.

English Český

Login
Nepřihlášený uživatel

Kontakty

Dnes je 21.9.2016

Otevřená data

Úvod
Info IS CEDR
Info pro poskytovatele dotací
Příjemci pomoci
Poskytnuté dotace
Tvorba sestav
Uložené sestavy
Osobní údaje
Otevřená data
Seznam zkratk
Nápověda
Často kladené dotazy

Open Data CEDR III Licenční ujednání Technická specifikace Datová sada

Open Data CEDR III nahoru ▲

Generální finanční ředitelství v souladu s usnesením vlády České republiky ze dne 4. dubna 2012 č. 243, které schválilo plnění Akčního plánu Partnerství pro otevřené vládnutí, na těchto stránkách **zveřejňuje** údaje obsažené v Centrální evidenci dotací (CEDR III) ve strojově čitelném formátu (tzv. „otevřená data“).

Světová nezisková organizace Open Knowledge Foundation definuje otevřená data (také je možné používat termín „open data“) jako data, která může kdokoliv užívat, modifikovat a sdílet za jakýmkoliv účelem. Třemi základními požadavky na otevřená data je dostupnost a přístupnost, opětovné využití a předání a dále neomezuji přístup.

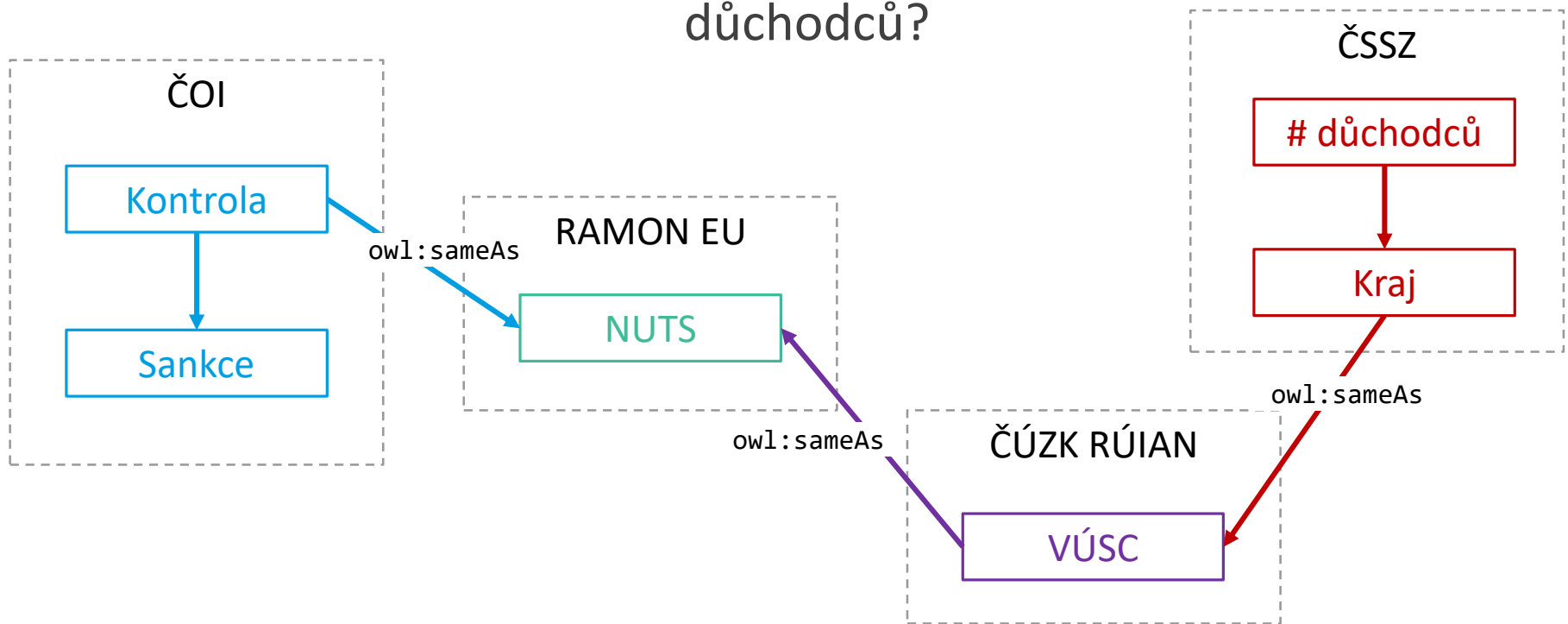
„Otevřená data jsou určena pro specialisty (programátory, datové analytiky, novináře, vědce atd.), kteří s jejich pomocí tvoří zajímavé softwarové aplikace či odborné analýzy. Ty pak slouží široké veřejnosti.“[1]

Datové sady jsou, na základě doporučení Metodiky publikace otevřených dat veřejné správy ČR vydané v roce 2012 Ministerstvem vnitra, zveřejňovány ve formátu RDF.

Upozornění k obsahu otevřených dat

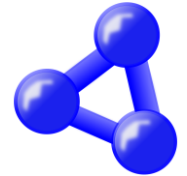
Práce s více datovými sadami

Kolik udělila ČOI v krajích sankcí a v jaké celkové výši za nekalé obchodní praktiky v porovnání s počtem starobních důchodců?



Webové standardy

- Jak webová data zapisovat
 - RDF 1.1
 - (Turtle, TriG, N-Triples, N-Quads, RDF/XML, JSON-LD, RDFa)
 - Aktuální 2014, první 2004
- Jak se dotazovat a jak měnit
 - SPARQL 1.1 - Query Language, Update, Service Description
 - Aktuální 2013, první 2008
- Jak webová data získat
 - (2012) Z relační databáze
 - Automaticky: Direct Mapping, Mapování: R2RML
 - (2015) Z CSV: Generating RDF from Tabular Data on the Web
 - (2014) Z JSON: JSON-LD
 - Z XML: (obecně) XSLT 2.0
 - (2015) XSLT 3.0 je W3C Candidate Recommendation



Co jsou otevřená propojená data?



Kontrolní akce

CISLO	PREDMET	ZAHAJENI
2012/33	Peněžní prostředky určené ...	2012/11
2012/34	Účetní závěrka a finanční ...	2012/11

Kontrolované osoby

OSOBA	IC	OKRES	KA
Ministerstvo obrany	60162694	Praha	2012/33
ČSSZ	6963	Praha	2012/34



Principy propojených dat

1. princip: Pojmenovávajíte věci pomocí IRI.

Kontrolní akce

CISLO	PREDMET	ZAHAJENI
2012/33	Peněžní prostředky určené ...	2012/11
2012/34	Účetní závěrka a finanční ...	2012/11

<http://data.nku.cz/akce/2012/33>

<http://data.nku.cz/akce/2012/34>

Kontrolované osoby

OSOBA	IC	OKRES	KA
Ministerstvo obrany	60162694	Praha	2012/33
ČSSZ	6963	Praha	2012/34

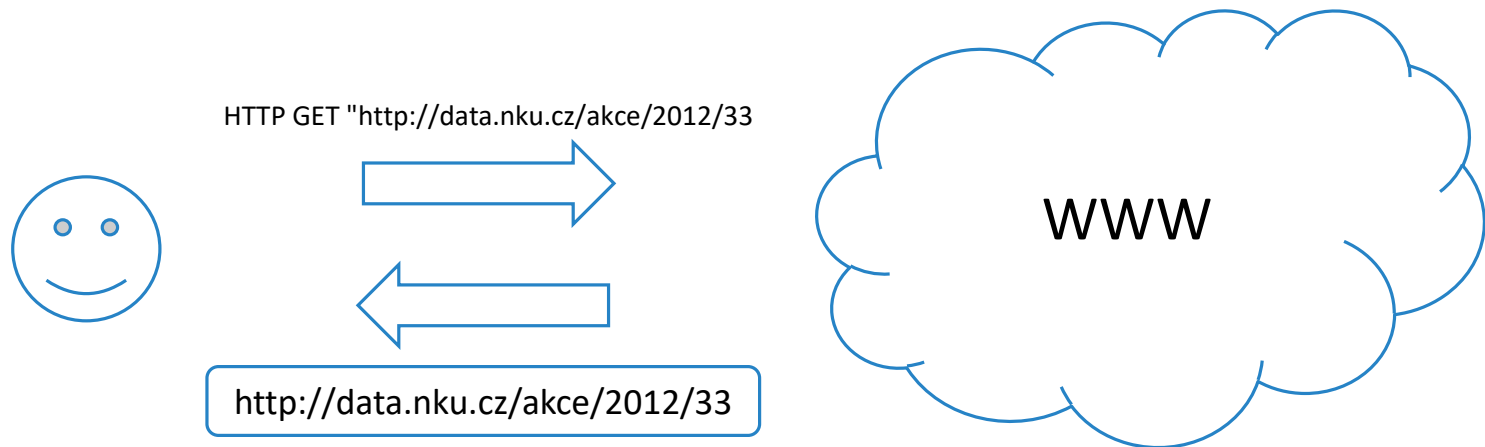
<http://data.nku.cz/osoba/60162694>

<http://data.nku.cz/osoba/6963>

<http://data.nku.cz/okres/praha>

Principy propojených dat

2. princip: Používejte HTTP(S) IRI a umožněte tak věci vyhledávat v prostředí WWW.



Principy propojených dat

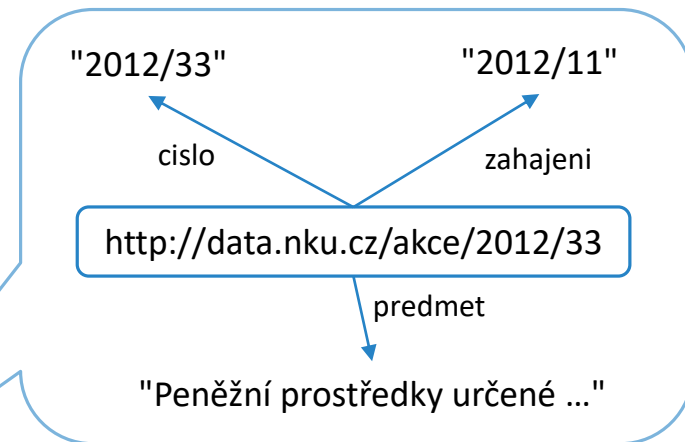
3. princip: Pokud někdo vyhledá IRI, poskytněte mu údaje o příslušné věci ve formátu RDF.

Kontrolní akce

CISLO	PREDMET	ZAHAJENI
2012/33	Peněžní prostředky určené ...	2012/11

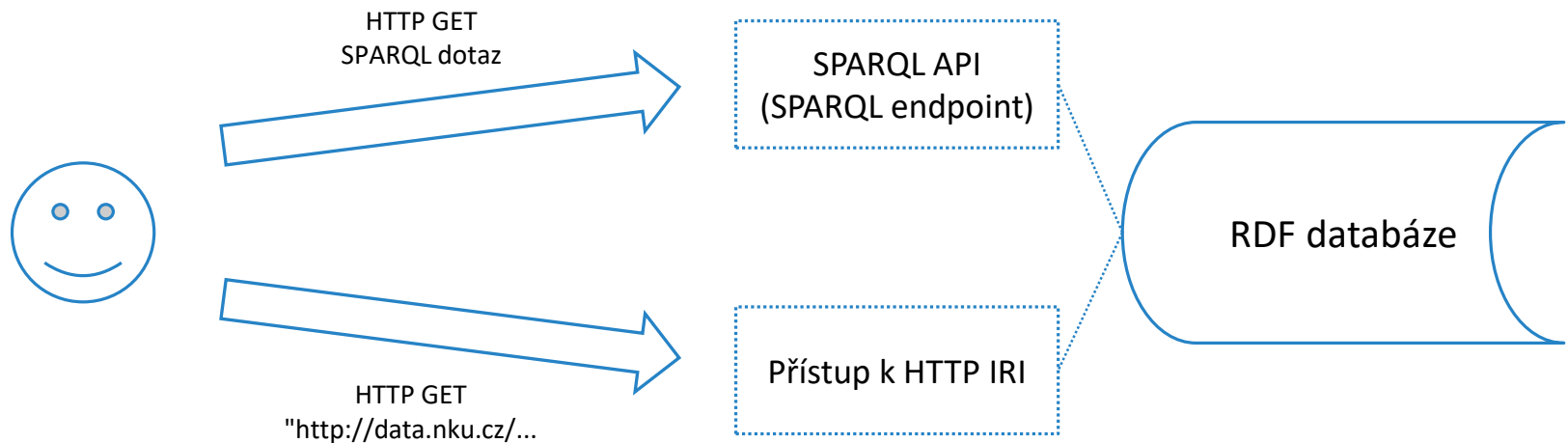
Vyjádření v RDF formátu
(notace Turtle, N-Triples)

```
<http://data.nku.cz/akce/2012/33> cislo "2012/33" .  
<http://data.nku.cz/akce/2012/33> predmet "Peněžní prostředky určené ..." .  
<http://data.nku.cz/akce/2012/33> zahajeni "2012/11" .
```



Principy propojených dat

3. princip: Pokud někdo vyhledá IRI, poskytněte mu údaje o příslušné věci ve formátu RDF. **Volitelně nabídněte SPARQL API.**



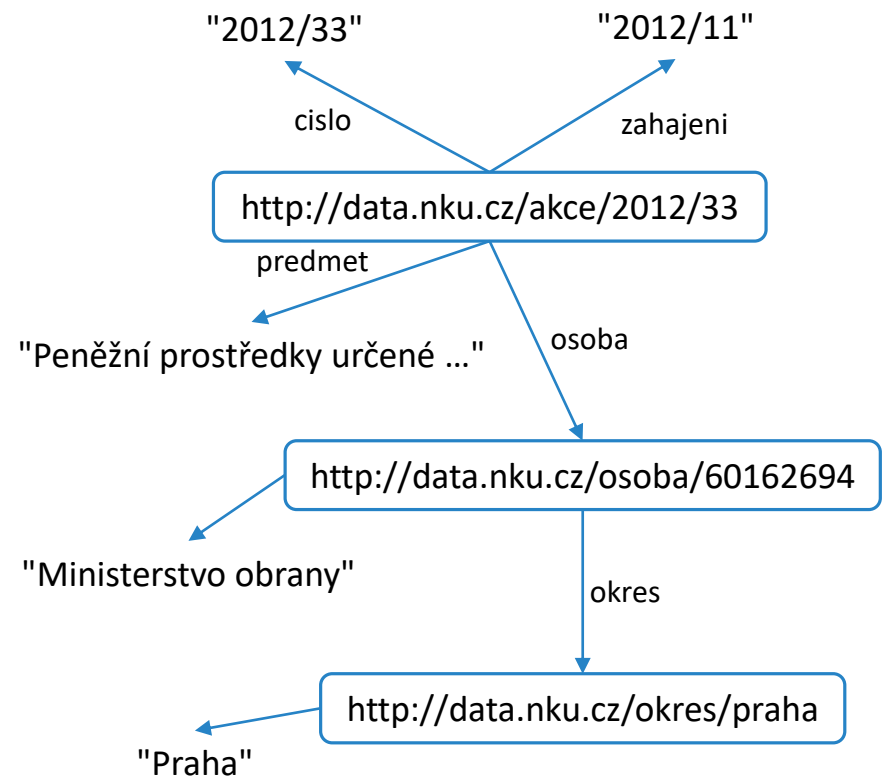
Principy propojených dat

4. princip: Mezi údaji poskytněte i odkazy na IRI souvisejících věcí.

```
<http://data.nku.cz/akce/2012/33>  
  cislo "2012/33" ;  
  predmet "Peněžní prostředky určené ..." ;  
  zahajeni "2012/11" ;  
  osoba <http://data.nku.cz/osoba/60162694> .
```

```
<http://data.nku.cz/osoba/60162694>  
  nazev "Ministerstvo obrany" ;  
  okres <http://data.nku.cz/okres/praha> .
```

```
<http://data.nku.cz/okres/praha>  
  nazev "Praha".
```



4 ★ vs 5 ★



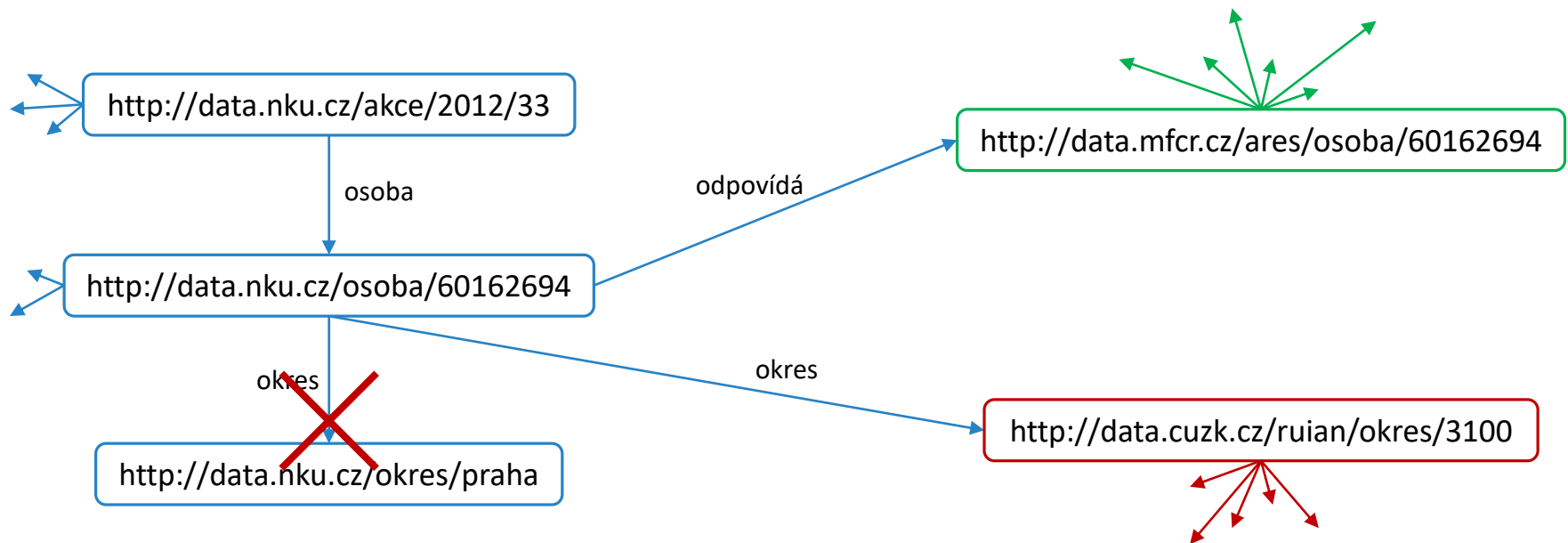
- IRI pro pojmenování věcí
- Data poskytována ve formátu RDF



- Propojená data

Principy propojených dat

4. princip: Mezi údaji poskytněte i odkazy na IRI souvisejících věcí
(včetně IRI jiných poskytovatelů).

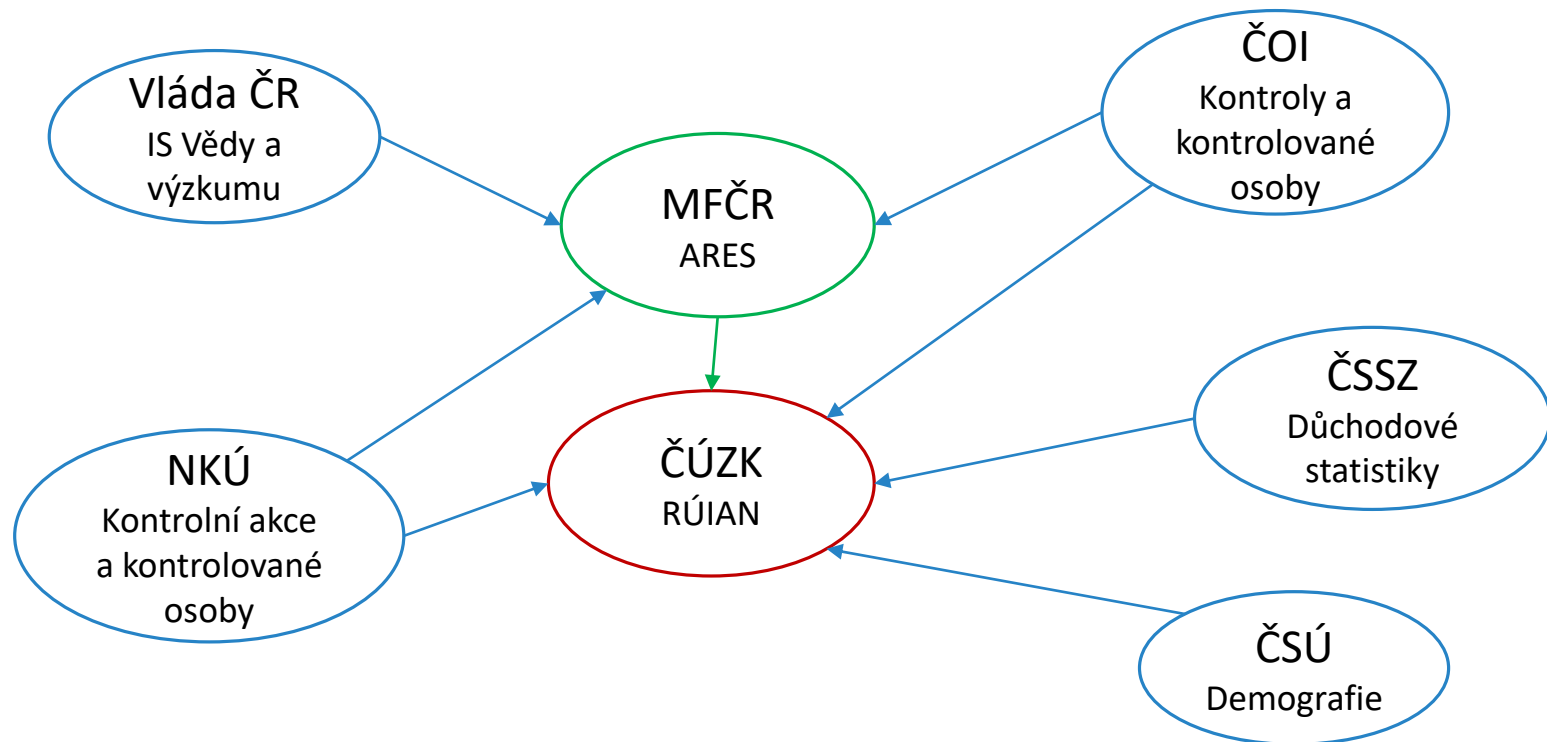


Web dokumentů vs. Web dat

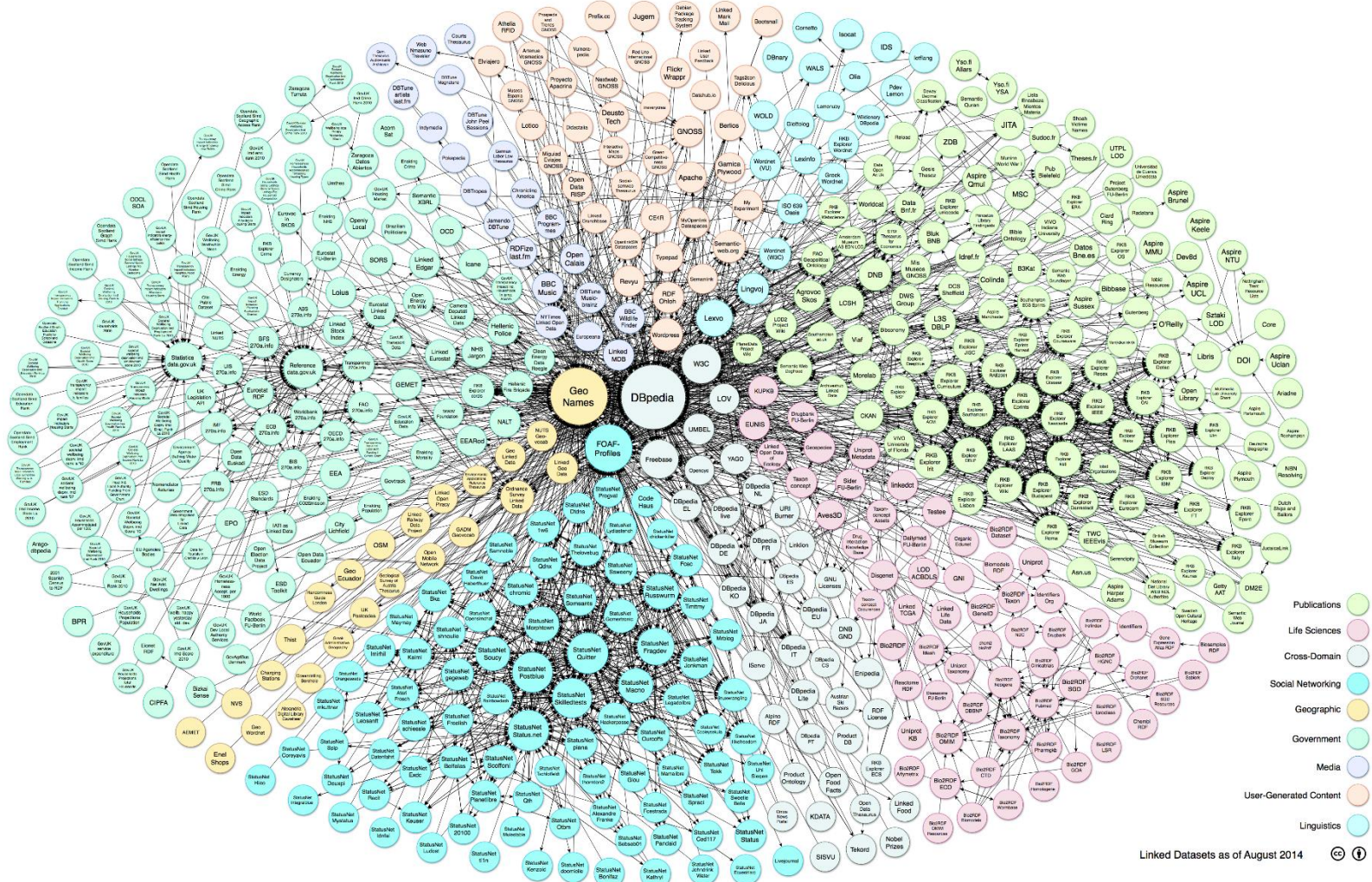
- Propojeným datům se také říká webová data
- Web dokumentů – WWW
 - Základní entita: **webová stránka** – dokument čitelný pro člověka
 - Identifikátory: **IRI**
 - Protokol pro přenos: **HTTP**
 - Jednotný jazyk: **HTML**
- Web dat – Linked Data
 - Základní entita: **datová entita** – data čitelná strojově
 - Identifikátory: **IRI**
 - Protokol pro přenos: **HTTP**
 - Jednotný jazyk: **RDF**

Principy propojených dat

4. princip: Mezi údaji poskytněte i odkazy na IRI souvisejících věcí
(včetně IRI jiných poskytovatelů).



LOD Cloud diagram 2014



Slovníky a ontologie


<http://data.nku.cz/akce/2012/33>
cislo "2012/33" ;
predmet "Peněžní prostředky určené ..." ;
zahajeni "2012/11" ;
osoba <http://data.nku.cz/osoba/60162694> .


<http://data.nku.cz/akce/2012/33>
a schema:CheckAction, nku:KontrolniAkce ;
adms:identifier "2012/33" ;
schema:object "Peněžní prostředky určené ..." ;
schema:startDate "2012/11" ;
nku:osoba <http://data.nku.cz/osoba/60162694> .

- Vlastnosti nejsou určeny řetězcem, ale pomocí HTTP IRI, stejně jako ostatní věci.
- Věci jsou přiřazovány do tříd (= kategorií věcí), které jsou také určeny pomocí HTTP IRI.
- Třídy a vlastnosti jsou definovány pomocí tzv. *slovníků* (někdy zvaných *ontologie*)
 - Dublin Core Vocabulary
 - Schema.org
 - Data Cube Vocabulary
 - ...

schema:object
=
<http://schema.org/object>

nku:KontrolniAkce
=
<http://data.nku.cz/slovník/KontrolniAkce>

Standardy pro modelování

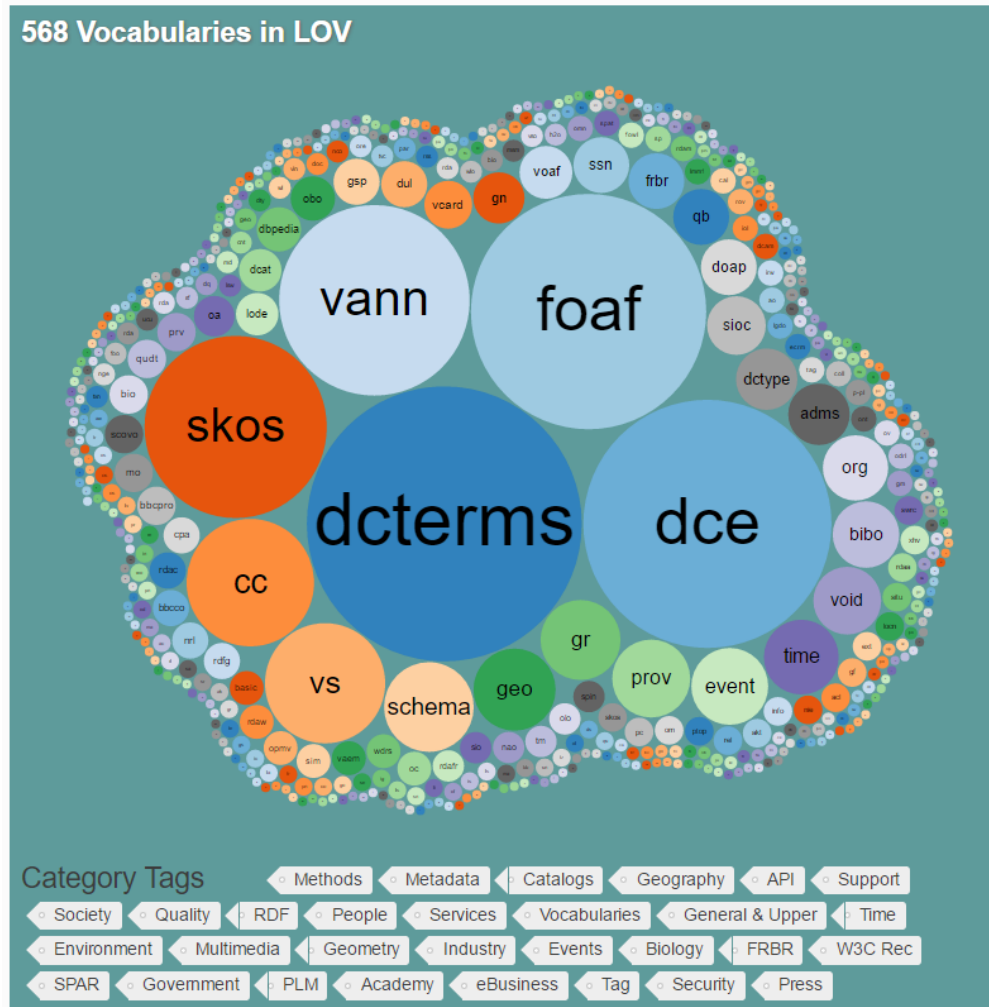
- W3C Recommendation
 - (2014) Organizace a jejich struktury - [The Organization Ontology](#)
 - (2014) Statistická data - [The RDF Data Cube Vocabulary](#)
 - (2014) Datové katalogy - [Data Catalog Vocabulary \(DCAT\)](#)
 - (2009) Hierarchie, číselníky - [SKOS Simple Knowledge Organization System Reference](#)
- W3C Group Note
 - (2014) Kontaktní údaje - [vCard Ontology - for describing People and Organizations](#)
 - (2013) Registrované organizace - [Registered Organization Vocabulary](#)
 - (2013) Schémata, modely, slovníky - [Asset Description Metadata Schema \(ADMS\)](#)

Katalogizace otevřených dat

- EU standard DCAT-AP je RDF slovník založený na principech propojených dat.
- Katalogy, datové sady a jejich distribuce jsou reprezentovány dle principů propojených dat, tj.
 - mají svá IRI
 - údaje o nich jsou reprezentovány v RDF dle slovníku DCAT-AP
 - nad údaji z katalogů se lze dotazovat pomocí SPARQL endpointu

<http://www.europeandataportal.eu/sparql-manager/en/>

Linked Open Vocabularies



Děkuji za pozornost

Jakub Klímek, Martin Nečaský

klimek@opendata.cz

[@jakub klimek](#)



MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

