

Automatizovatelná aktualizace Wikidata z veřejných databází

Jakub Klímek

Veřejné databáze, Otevřená data

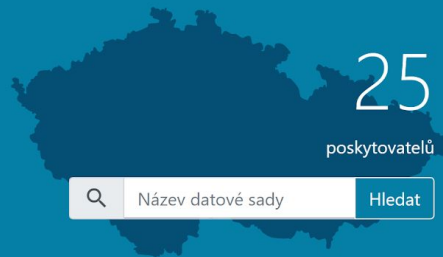
OTEVŘENÁ DATA

Novinky Datové sady Poskytovatelé Klíčová slova Další 

Otevřená data

Vše, co potřebujete vědět o otevřených datech v České republice

Chci data otevřít snadno a rychle



Zájemci o otevírání dat

Co mi otevřená data přinesou a proč bych je měl chtít? Jak pomohou mé obci či mému úřadu?

Více informací »

Poskytovatelé dat

Jak postupovat při otevírání dat? Jaké jsou příklady dobré a špatné praxe? A jaké jsou právní aspekty otevírání dat?

Více informací »

Programátoři

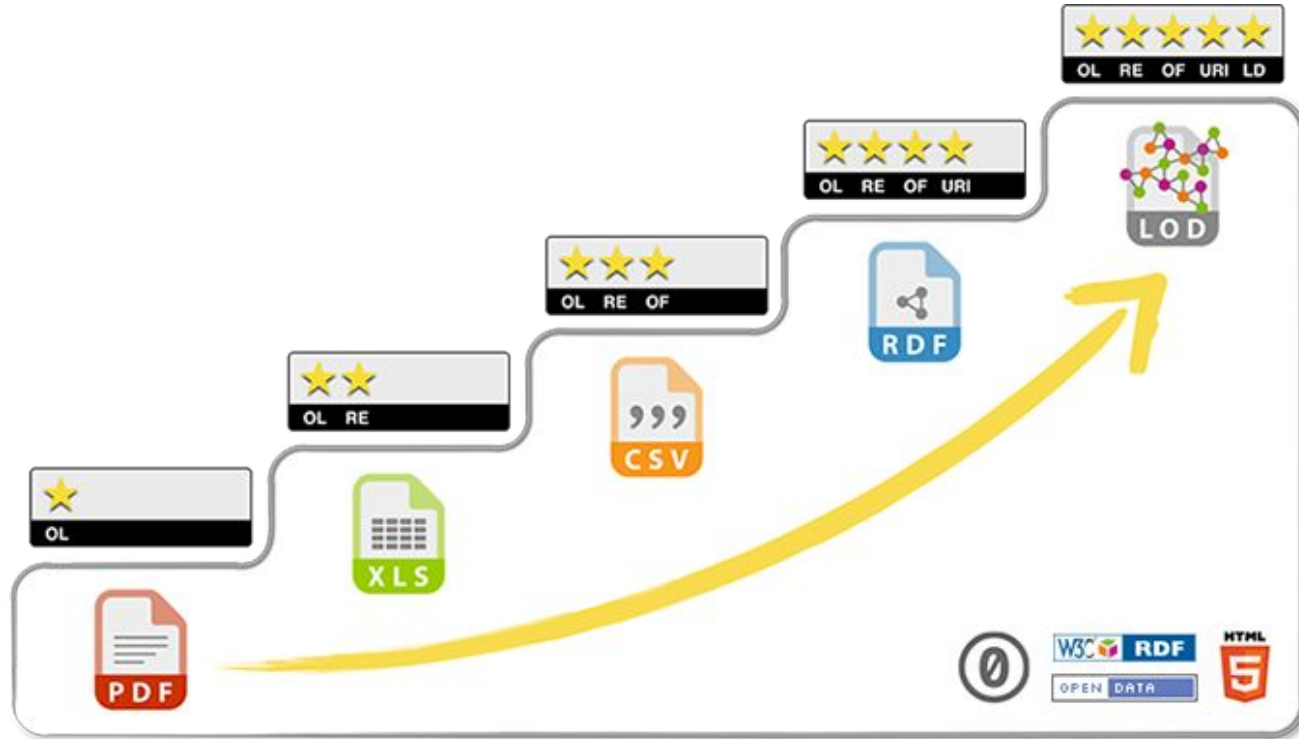
Kde najdu otevřená data? Co jsou otevřené formáty XML, CSV či RDF? Kde si mohu říct o další otevřená data nebo nahlásit chybu v datech?

Více informací »

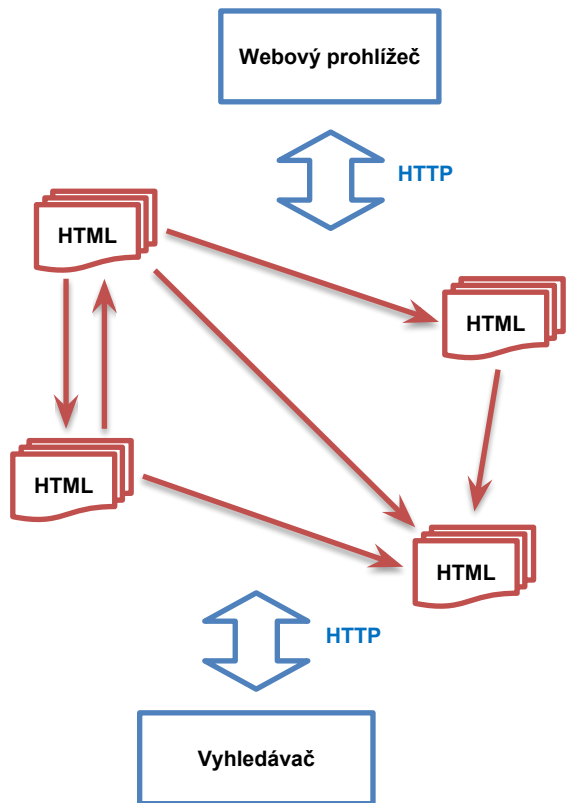
Veřejné databáze, Otevřená data - zajímavé zdroje

- Český úřad zeměměřický a katastrální
 - Registr územní identifikace, adres a nemovitostí (RÚIAN)
 - Kraje, obce, okresy, budovy, adresní místa
- Český statistický úřad
 - Počty obyvatelstva
 - Výsledky voleb
- Česká správa sociálního zabezpečení
 - Statistiky o důchodech a důchodcích

Veřejné databáze, Otevřená data - různá kvalita



Odbočka: WWW ~ Web dokumentů - všichni známe



Sdílený globální prostor dokumentů

Vybudován na pár jednoduchých principech:

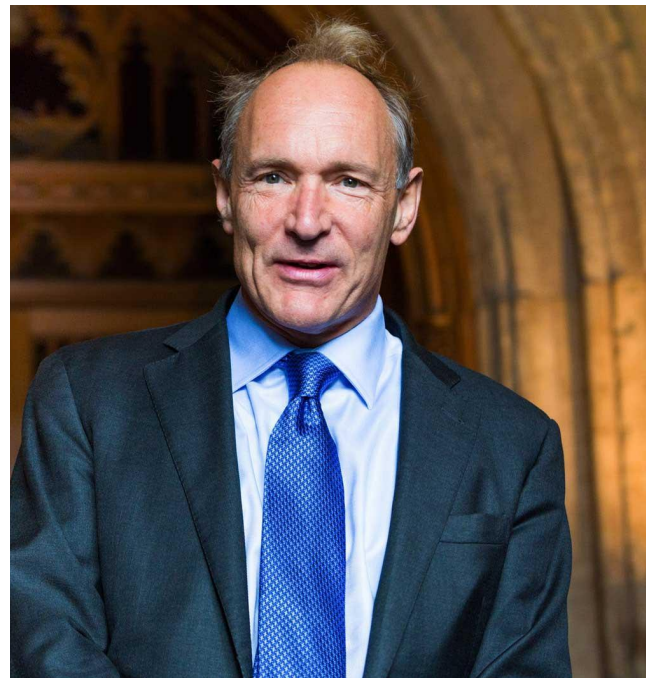
1. HTML jako jazyk pro dokumenty
2. URL jako unikátní globální identifikátory
3. HTTP pro lokalizaci a přístup k dokumentům na základě jejich URL
4. odkazy mezi dokumenty

Existují dva druhy aplikací pracujících v **tomto prostoru dokumentů**:

- **webové prohlížeče** (lokalizace a prohlížení stránek)
- **vyhledávače** (indexace a vyhledávání stránek)

Principy propojených dat

1. Pro pojmenování věcí používejte IRI
2. Používejte HTTP IRI, aby se o věcech dala dohledat data.
3. Když někdo vyhledá IRI, poskytněte užitečné informace, a to pomocí standardů (RDF, SPARQL)
4. V datech poskytněte odkazy na jiná IRI, aby uživatelé mohli objevovat další věci.



<https://5stardata.info/>

RDF - Resource Description Framework

RDF

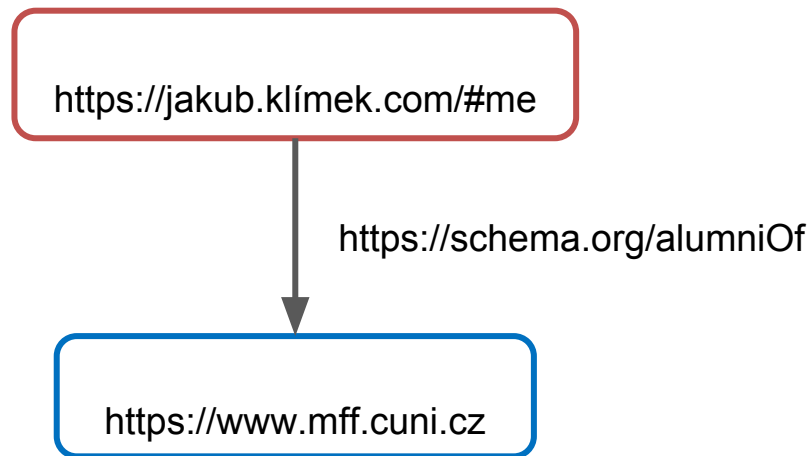
- grafový datový model
- množina trojic



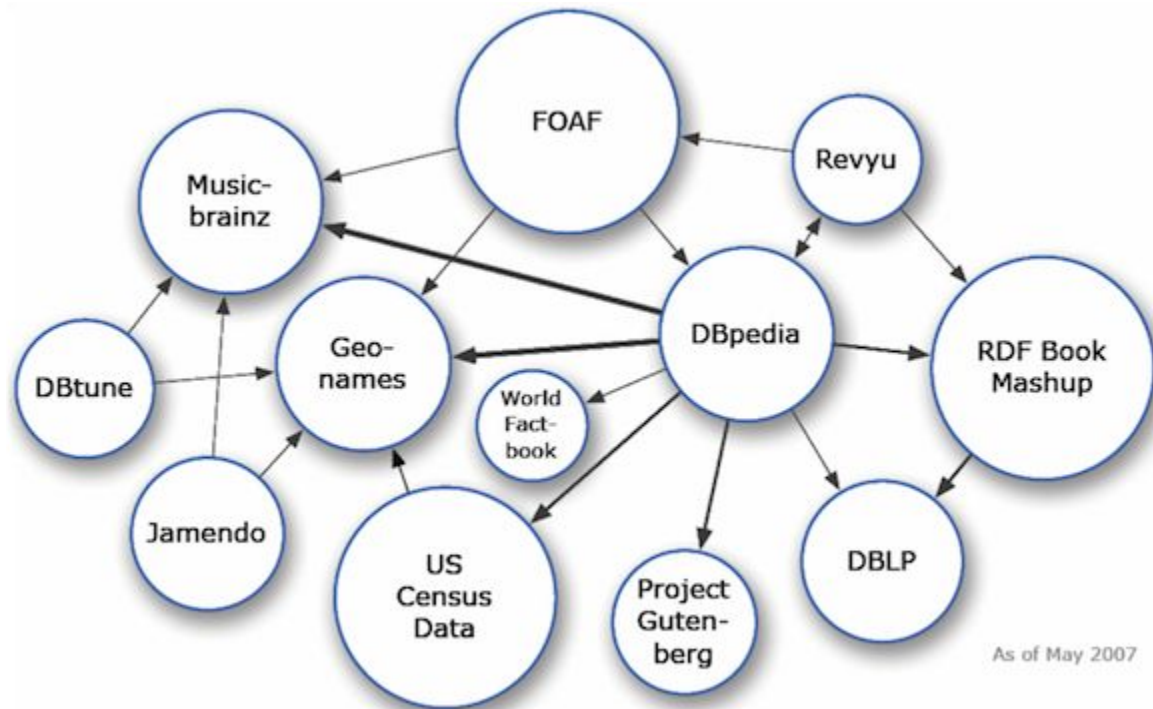
Trojice popisuje vztahy jako:

subjekt predikát objekt

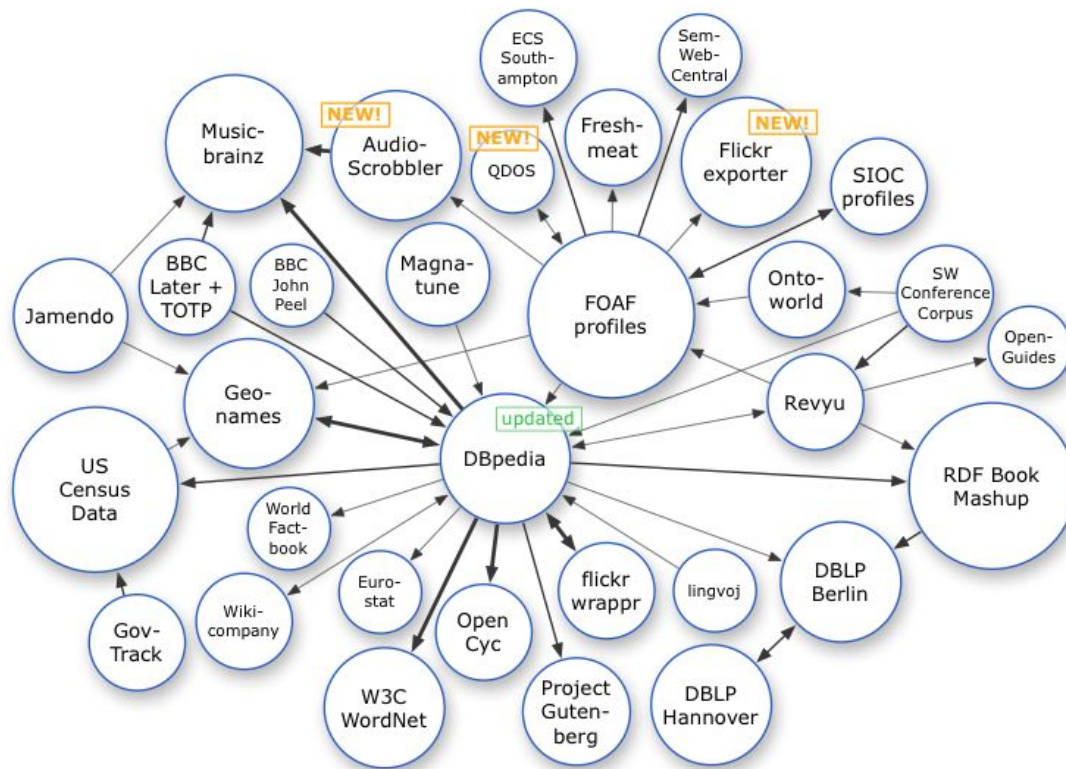
2004 & 2014 W3C Recommendations



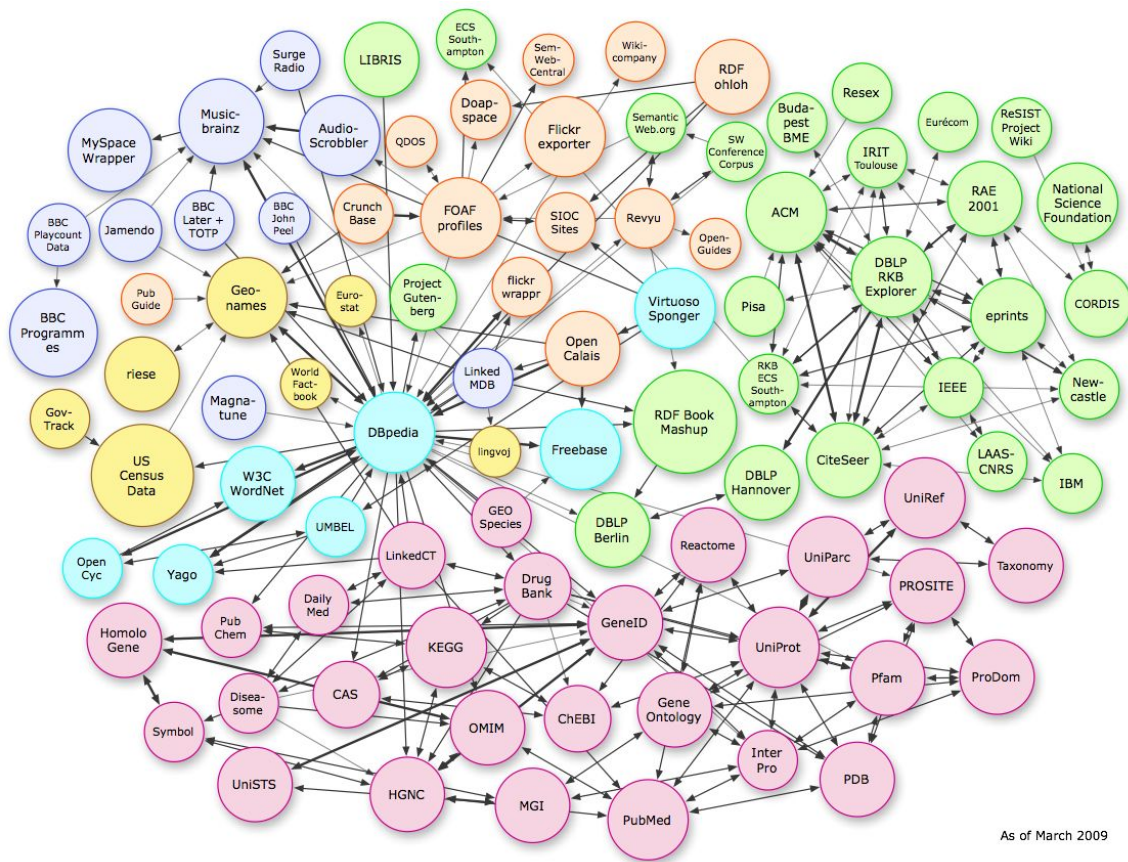
Cloud propojených dat - 2007 - 12 datových sad



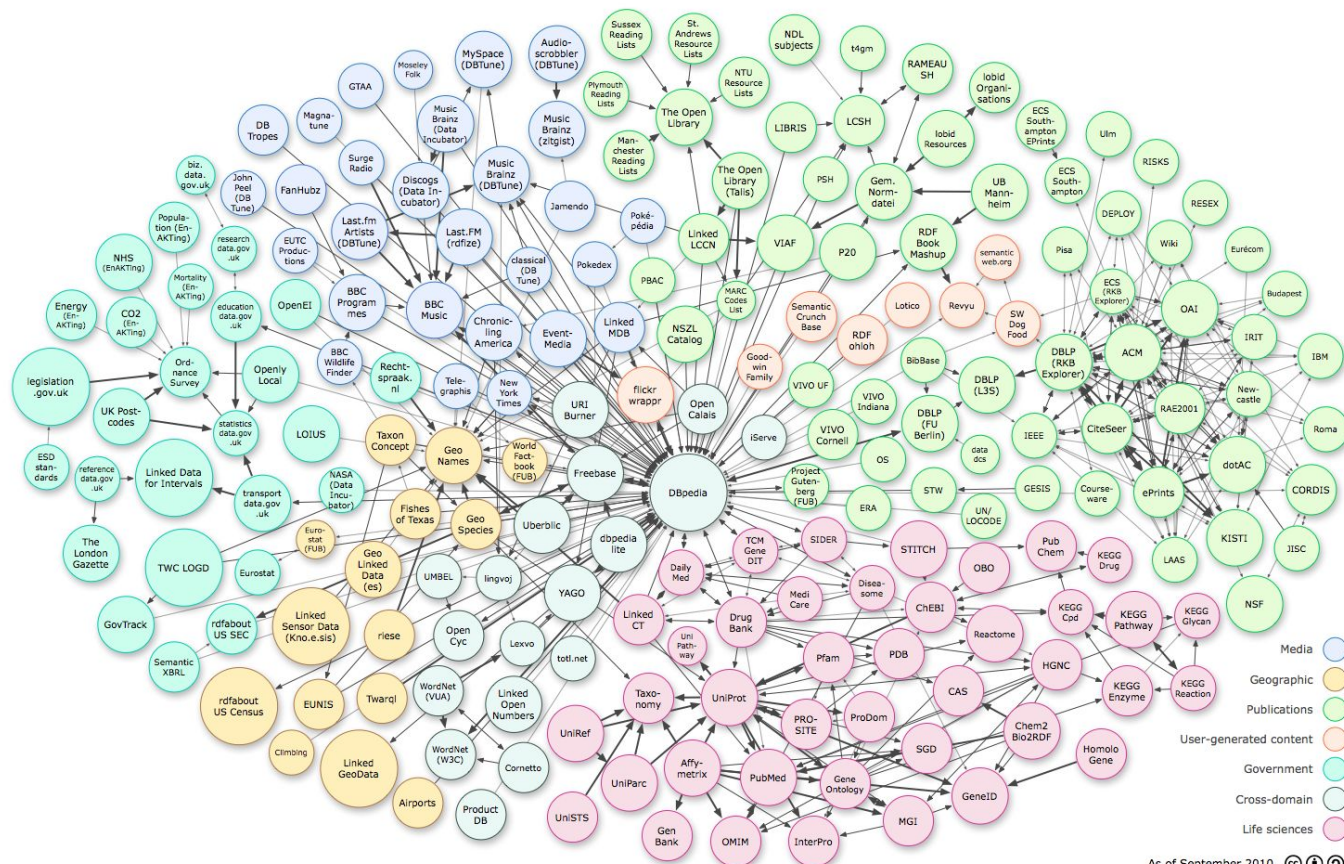
Cloud propojených dat - 2008 - 32 datových sad



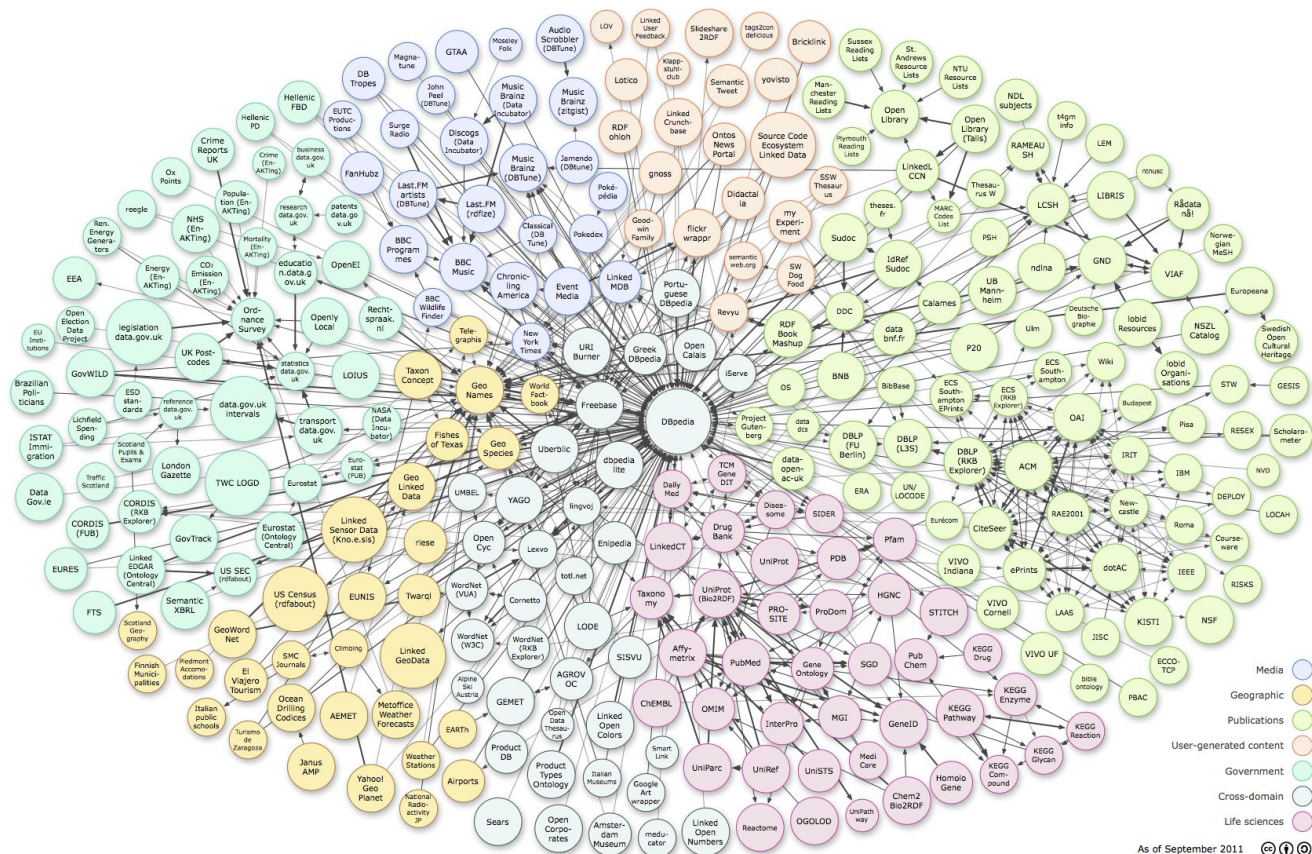
Cloud propojených dat - 2009 - 89 datových sad



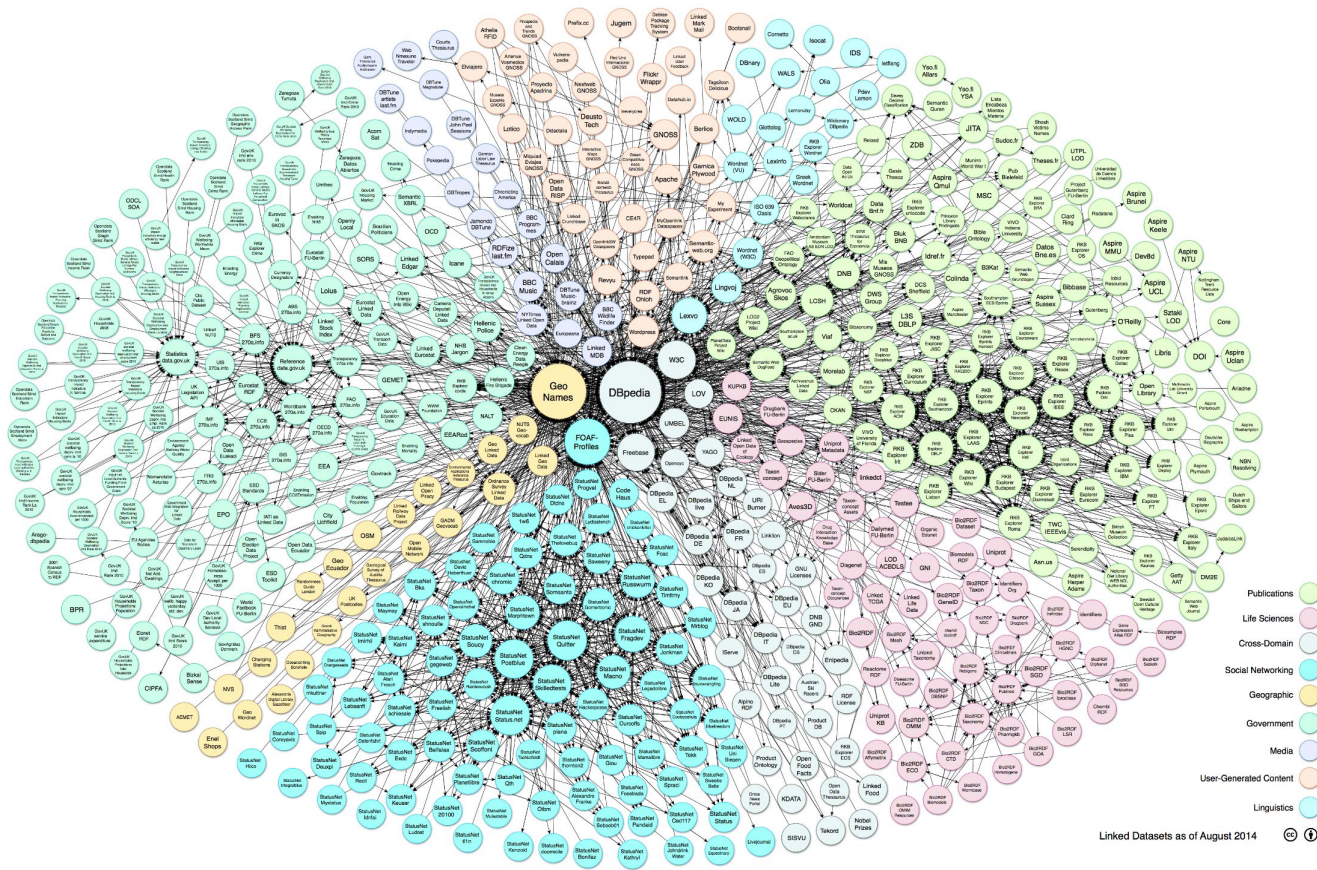
Cloud propojených dat - 2010 - 203 datových sad



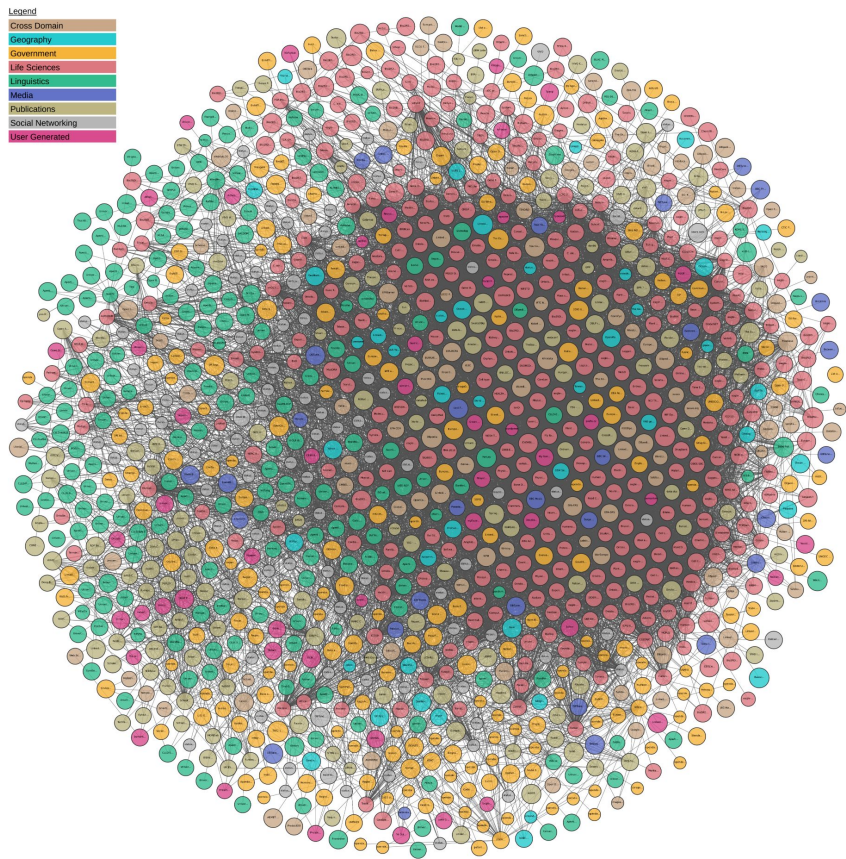
Cloud propojených dat - 2011 - 295 datových sad



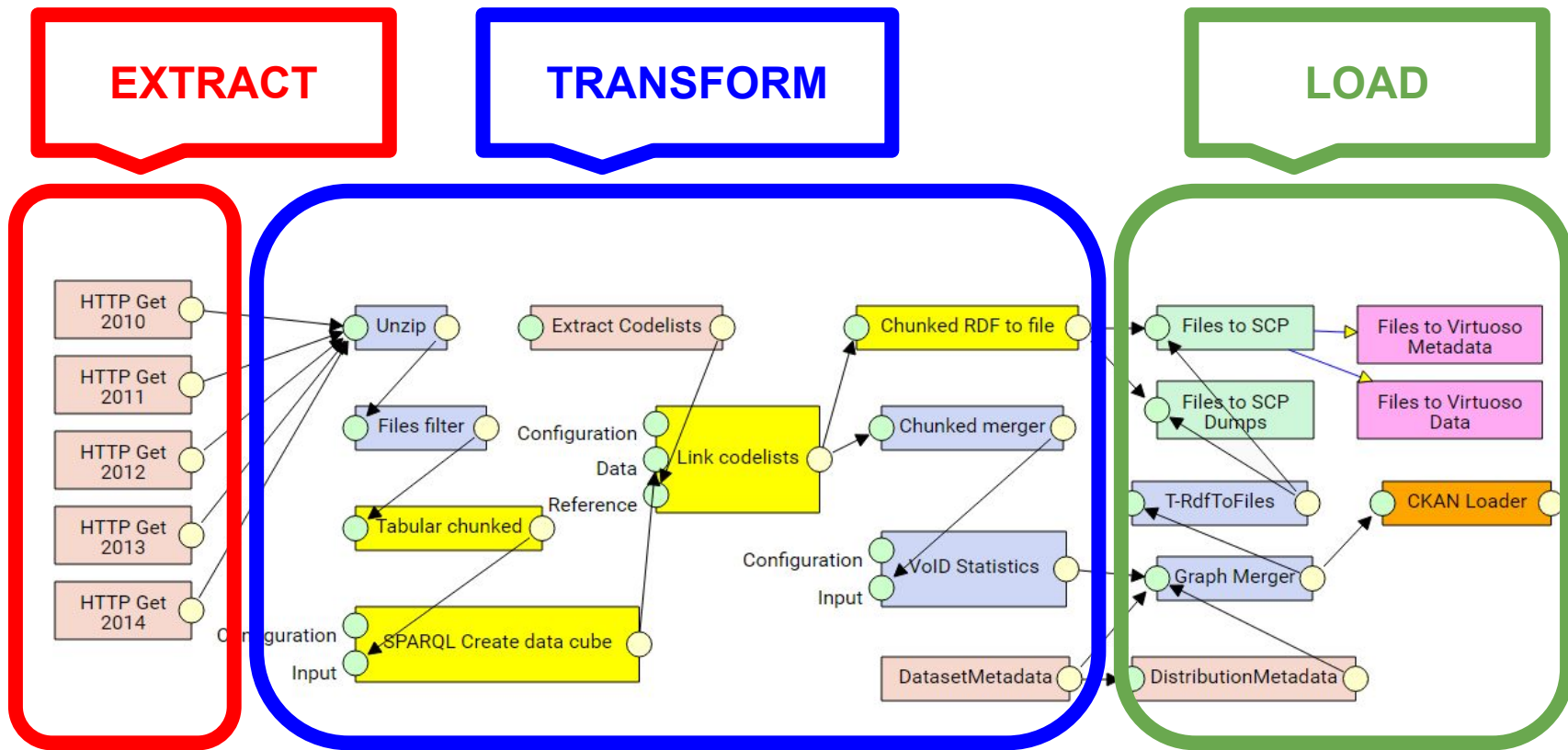
Cloud propojených dat - 2014 - 570 datových sad



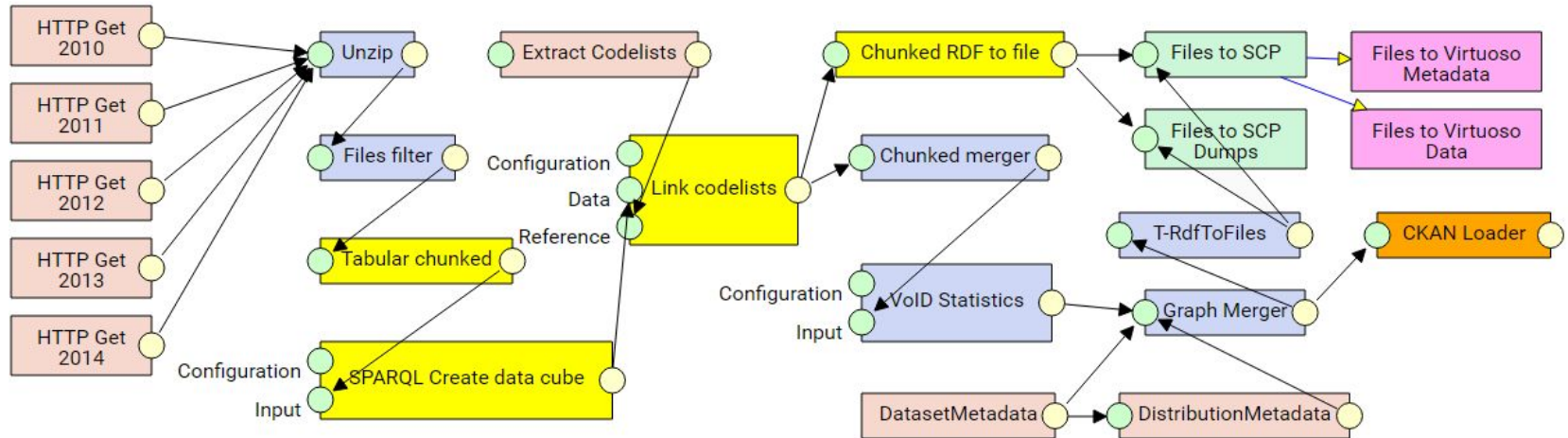
Cloud propojených dat - 2018 - 1229 datových sad

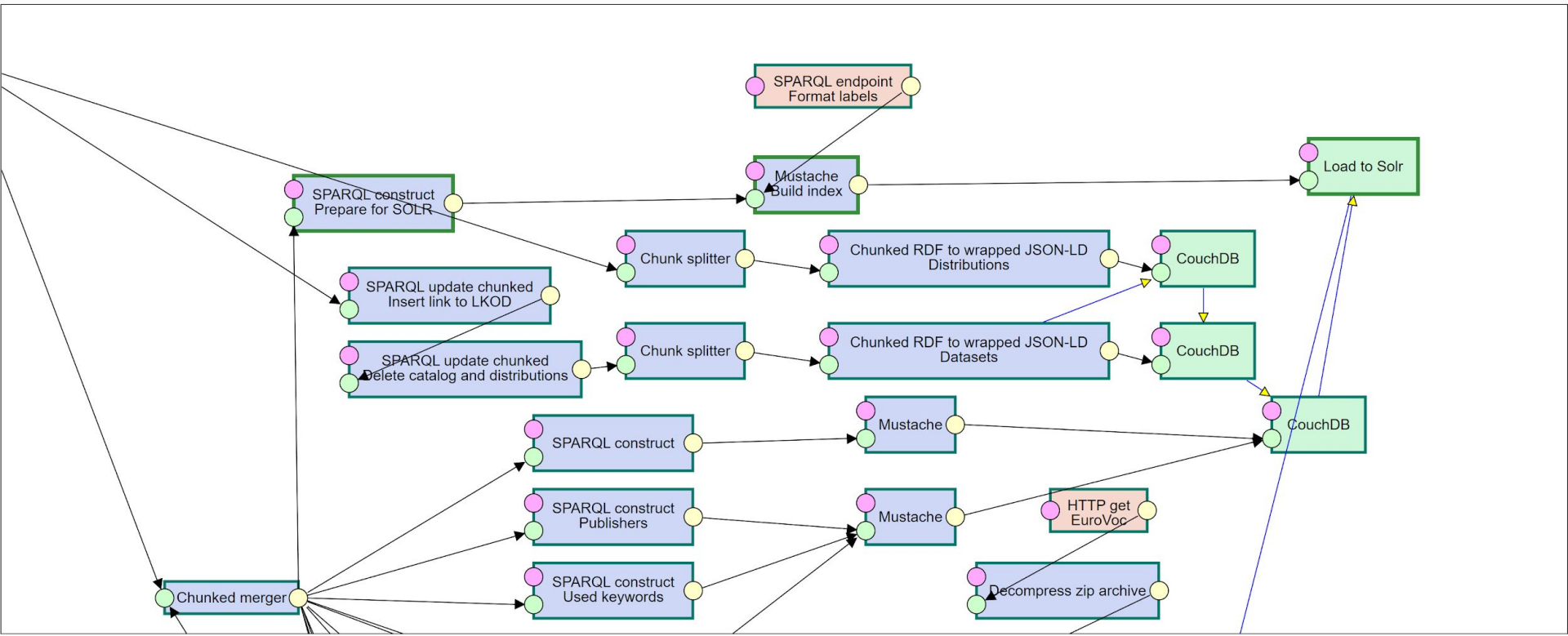


LinkedPipes ETL - **E**xtract **T**ransform **L**oad pro LOD komponenty v pipeline



LinkedPipes ETL - **E**xtract **T**ransform **L**oad pro LOD komponenty v pipeline

























LP-ETL - přes 60 znovupoužitelných komponent

- Extraktory
 - **HTTP** GET, **SPARQL** endpoint (and variants), Text holder, Extract from **FTP**, Files from **local**
- Transformery
 - SPARQL Construct, **SPARQL** Update
 - Files to RDF, **RDF** to file
 - **XSLT**, **JSON** to JSON-LD, Tabular, **Excel** to **CSV**
 - Bing translator, Decode base64, File hasher, Compress, Decompress, Mustache
- Loadery
 - Files to **SCP**
 - **SPARQL** Update, SPARQL Graph Store Protocol
 - Files to **local**

<https://etl.linkedpipes.com/components/>

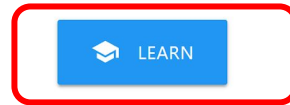
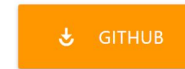
LP-ETL - Pipeliny (někde též “datovody”)

Label search

	Statistika dostupnosti distribucí, schémat, podmínek užití a dokumentace - HEAD 2018-11-27 03:21:29, 01:40:56 Full execution (No working data) Size: 248.51 mB	  
	Kvalita metadatových záznamů v NKOD 2018-11-27 03:17:00, 00:04:29 Full execution (No working data) Size: 63.22 mB	  
	PVS & LKODs 2 NKOD 2018-11-26 21:00:29, 05:39:44 Full execution (No working data) Size: 18405.7 mB	  
	Statistika dostupnosti distribucí, schémat, podmínek užití a dokumentace - HEAD 2018-11-26 03:21:53, 01:37:14 Full execution (No working data) Size: 248.56 mB	  
	Kvalita metadatových záznamů v NKOD 2018-11-26 03:15:55, 00:05:56 Full execution (No working data) Size: 63.21 mB	  
	PVS & LKODs 2 NKOD	

LP-ETL Tutoriály a návody

ETL: Extract Transform Load for Linked Data

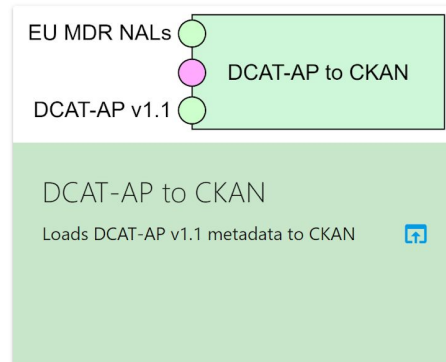


What's new

- 2018-07-04: LinkedPipes ETL featured @ ISWC 2018 Demo Session!
- 2017-10-30: LinkedPipes ETL @ iiWAS 2017!
- 2017-10-16: LinkedPipes ETL @ ODBASE 2017!
- 2017-08-17: A set of simple how-tos added
- 2017-07-31: Geocoding with nominatim tutorial available



Featured component



Wikidata

Databáze

- Volně dostupná
- Kolaborativní
- Vícejazyčná

Obsahuje

- Položky
- Tvrzení
 - Vlastnosti
 - Hodnoty
 - **Kvalifikátory**

The image shows a Wikidata profile for Douglas Adams (Q42) with several annotations:

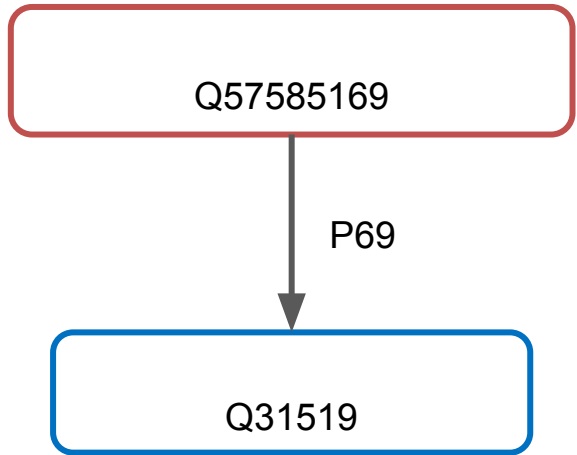
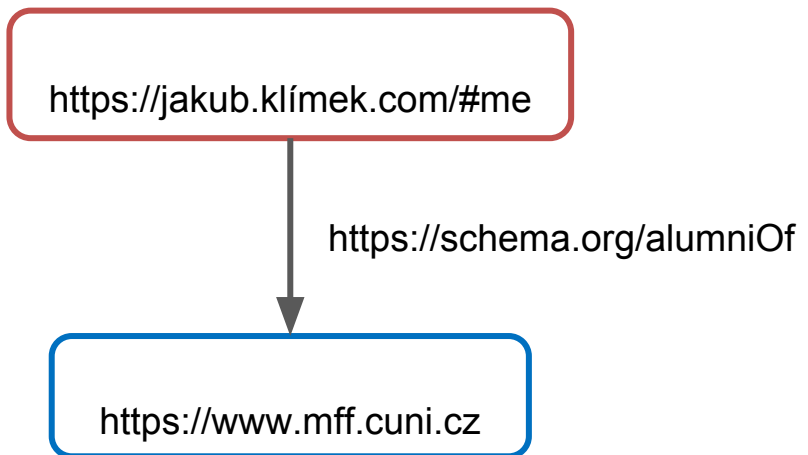
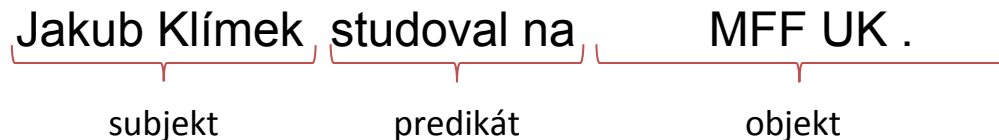
- label**: Douglas Adams (Q42) - item identifier
- description**: English writer and humorist
Douglas Noël Adams | Douglas Noel Adams - aliases
- property**: educated at - value
- rank**: points to the rank indicator (triangle) next to the statement group.
- statement group**: A green box highlights the entire statement group.
- value**: St John's College - value
- qualifiers**: A blue box highlights the qualifiers table for St John's College.

end time	1974
academic major	English literature
academic degree	Bachelor of Arts
start time	1971
- opened references**: A red box highlights the reference table for St John's College.

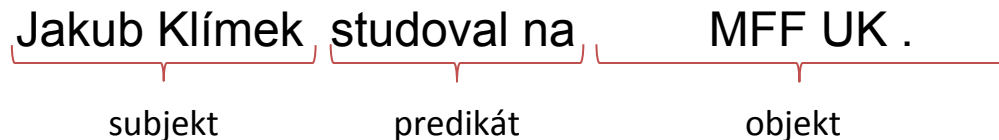
stated in	Encyclopædia Britannica Online
reference URL	http://www.nndb.com/people/731/000023662/
original language of work	English
retrieved	7 December 2013
publisher	NNDB
title	Douglas Adams (English)
- collapsed reference**: A red box highlights the reference table for Brentwood School.

0 references	
--------------	--

RDF vs. (zjednodušený) datový model Wikidata



RDF a (zjednodušený) datový model Wikidata



<https://jakub.klimek.com/#me>

<https://schema.org/alumniOf>

<https://www.mff.cuni.cz>

<https://www.wikidata.org/wiki/Q57585169>

<https://www.wikidata.org/wiki/Property:P69>

<https://www.wikidata.org/wiki/Q12036042>

Stávající problémy Wikidata s nahráváním dat v dávkách

- Převážně jednorázové skripty
 - Jen někdy používají již existující knihovny
 - Není počítáno s jejich pravidelným spouštěním
- Nejednotný přístup napříč autory
 - Autoři si pak těžko mohou navzájem skripty udržovat
 - Skripty se funkčně překrývají

Wikidata & ETL

- Projektový návrh k financování Wikimedia Foundation
- Úprava a aplikace LinkedPipes ETL pro práci s datovým modelem Wikidata
- Tvorba tutoriálu, jak na tvorbu pipeline pro Wikidata
- Úspěch záleží zejména na zájmu komunity

Prosíme o podporu formou “endorsementu”

- Účet na Wikipedii
- Jít na https://meta.wikimedia.org/wiki/Grants:Project/MFFUK/Wikidata_%26_ETL
- Přidat endorsement
 - uživatelské jméno
 - důvod, proč si myslíte že to má smysl řešit

