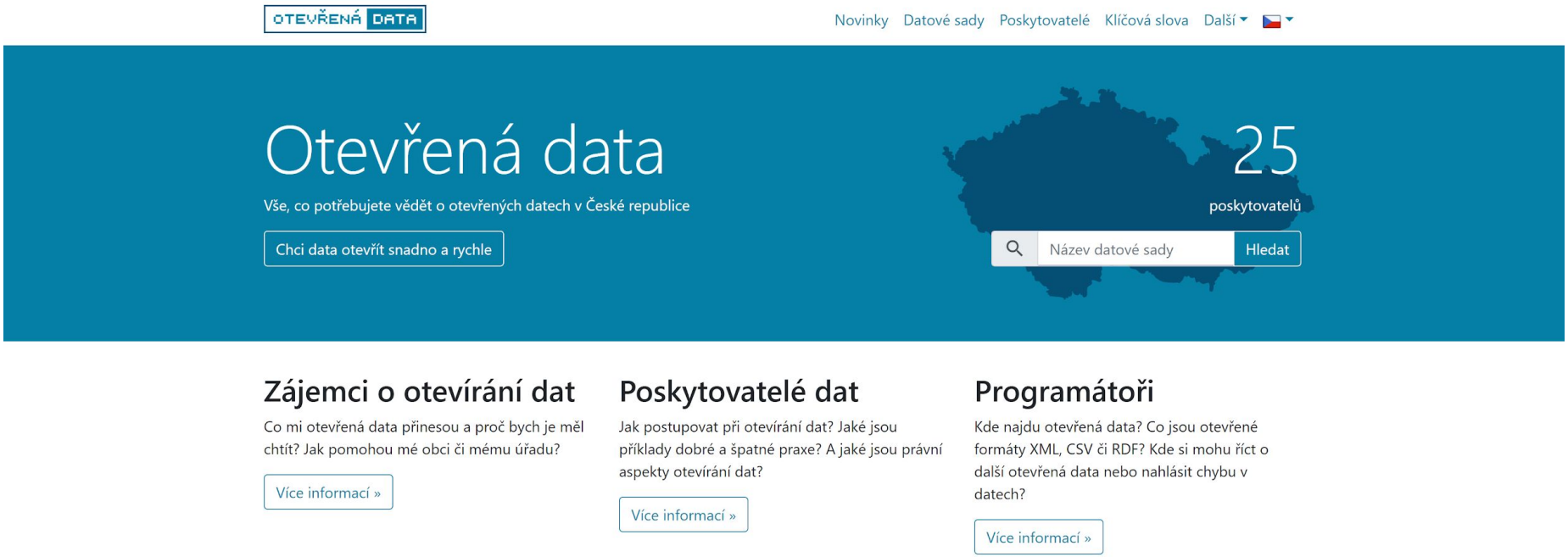
Automatizovatelná aktualizace Wikidata z veřejných databází

Jakub Klímek

Veřejné databáze, Otevřená data

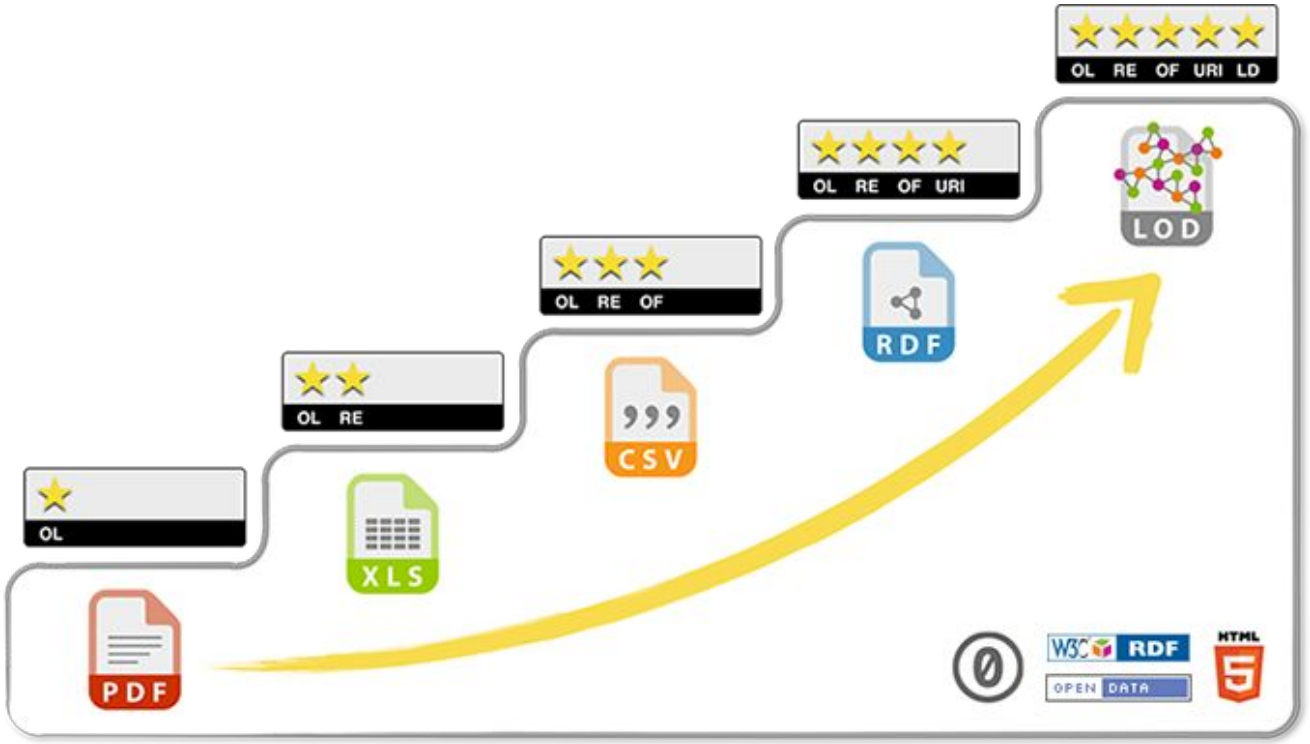


https://data.gov.cz

Veřejné databáze, Otevřená data - zajímavé zdroje

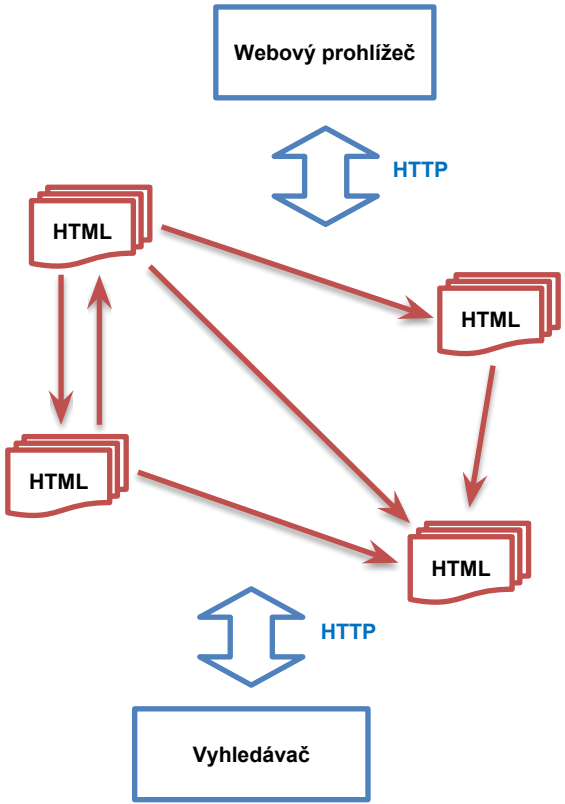
- Český úřad zeměměřický a katastrální
  - Registr územní identifikace, adres a nemovitostí (RÚIAN)
    - Kraje, obce, okresy, budovy, adresní místa
- Český statistický úřad
  - Počty obyvatelstva
  - Výsledky voleb
- Česká správa sociálního zabezpečení
  - Statistiky o důchodech a důchodcích

Veřejné databáze, Otevřená data - různá kvalita



https://5stardata.info

Odbočka: WWW ~ Web dokumentů - všichni známe



Sdílený globální prostor dokumentů

Vybudován na **pár jednoduchých** principech:

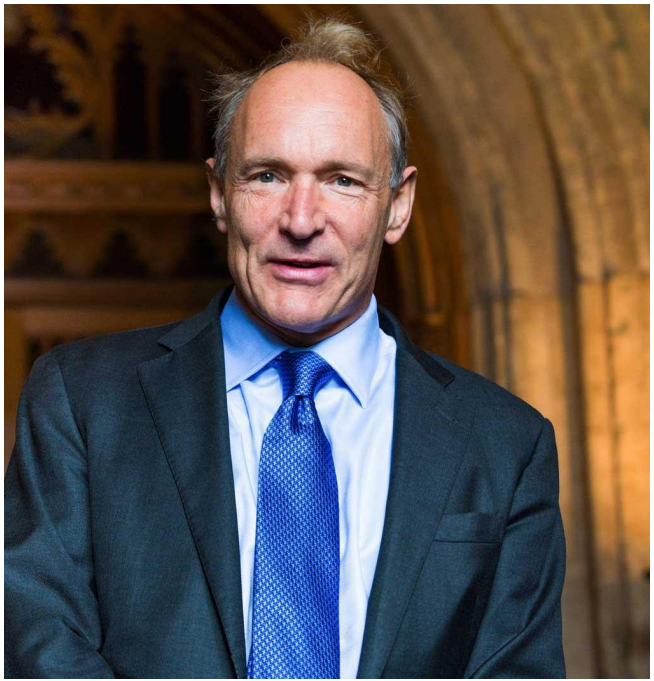
1. HTML jako jazyk pro dokumenty
2. URL jako unikátní globální identifikátory
3. HTTP pro lokalizaci a přístup k dokumentům na základě jejich URL
4. odkazy mezi dokumenty

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

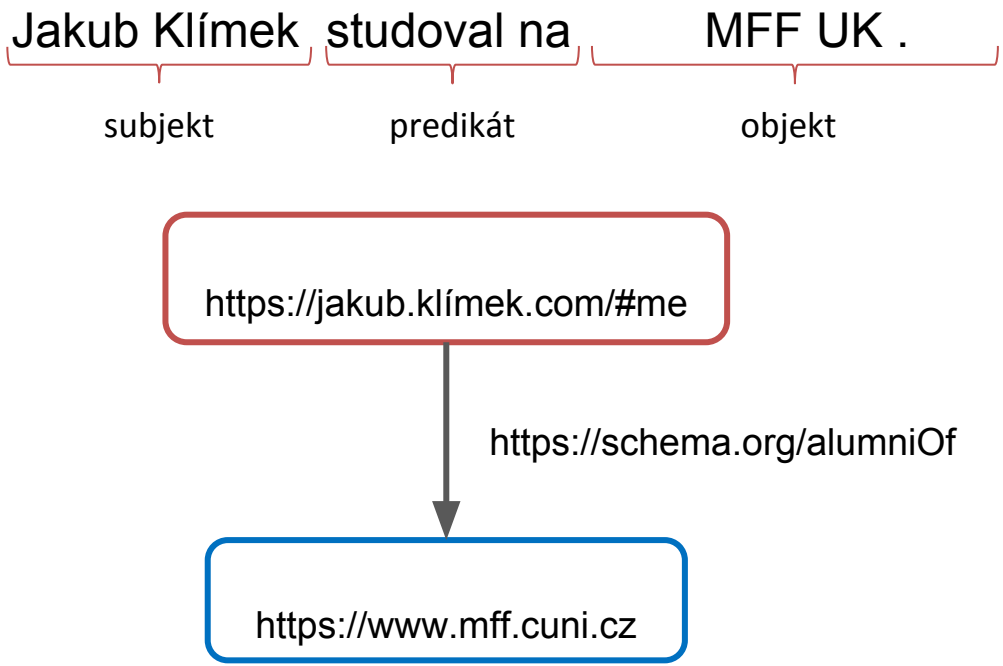
RDF

- grafový datový model
- množina trojic

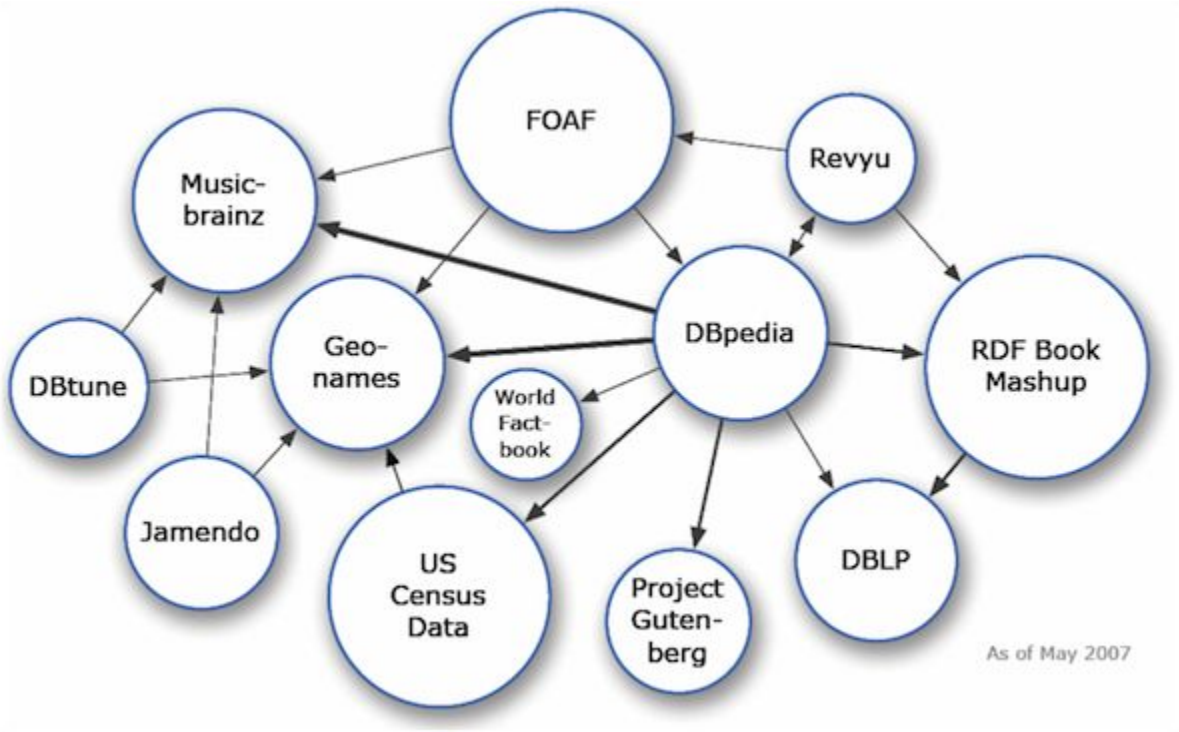
Trojice popisuje vztahy jako:

subjekt predikát objekt

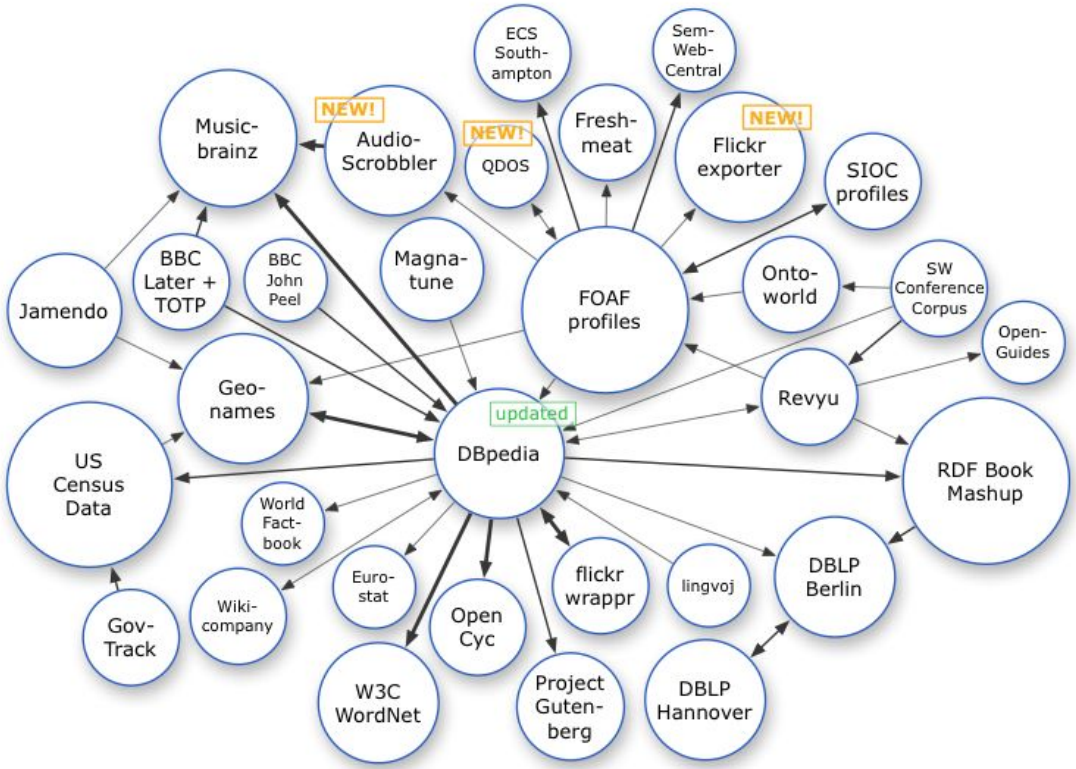
2004 & 2014 W3C Recommendations



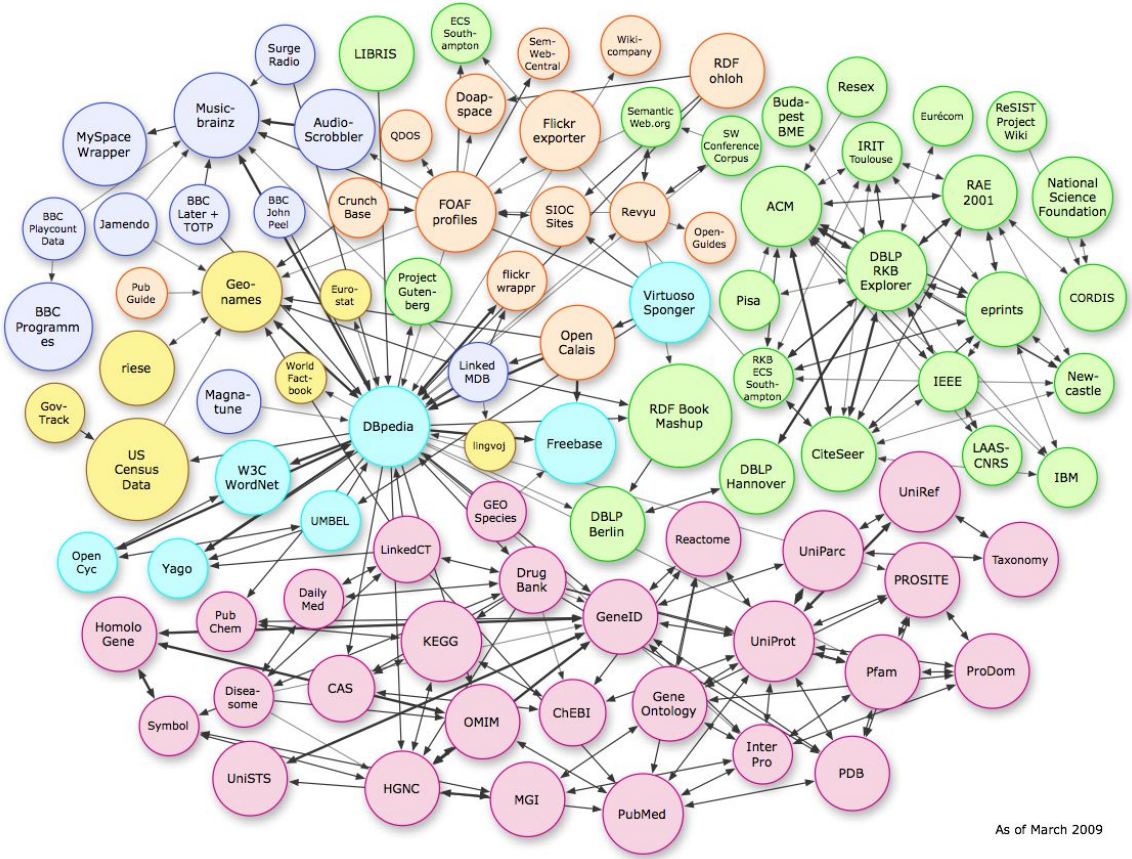
Cloud propojených dat - 2007 - 12 datových sad



Cloud propojených dat - 2008 - 32 datových sad

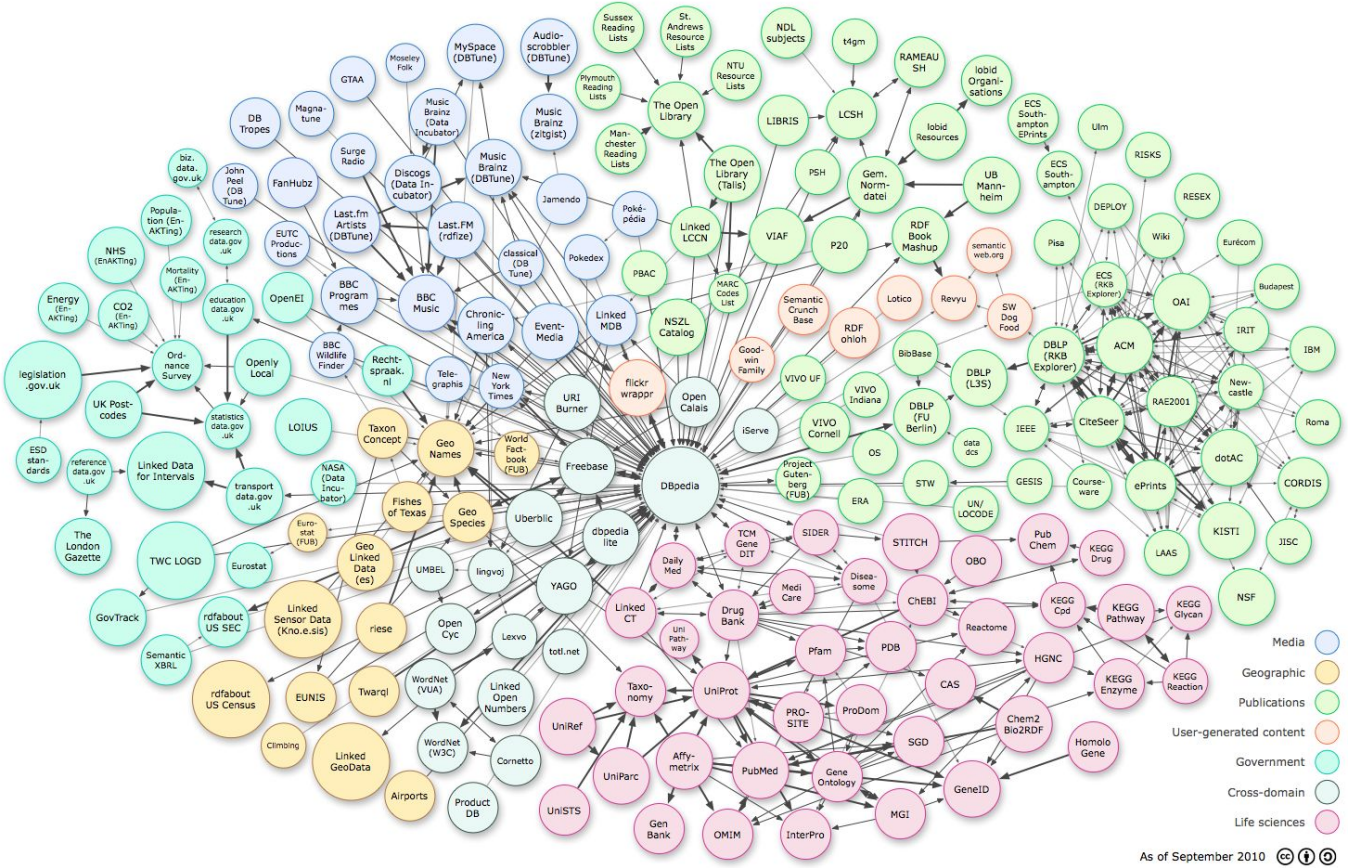


Cloud propojených dat - 2009 - 89 datových sad

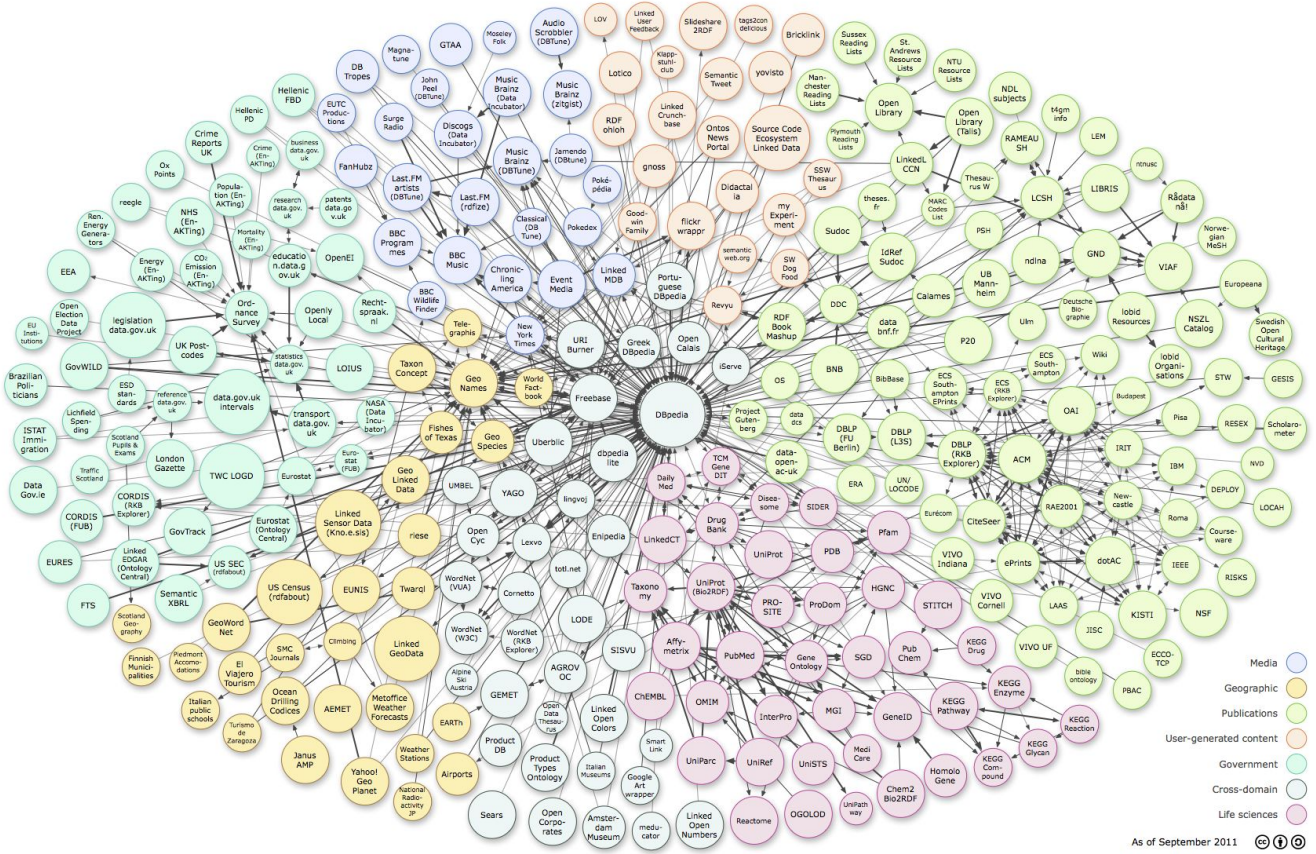


As of March 2009

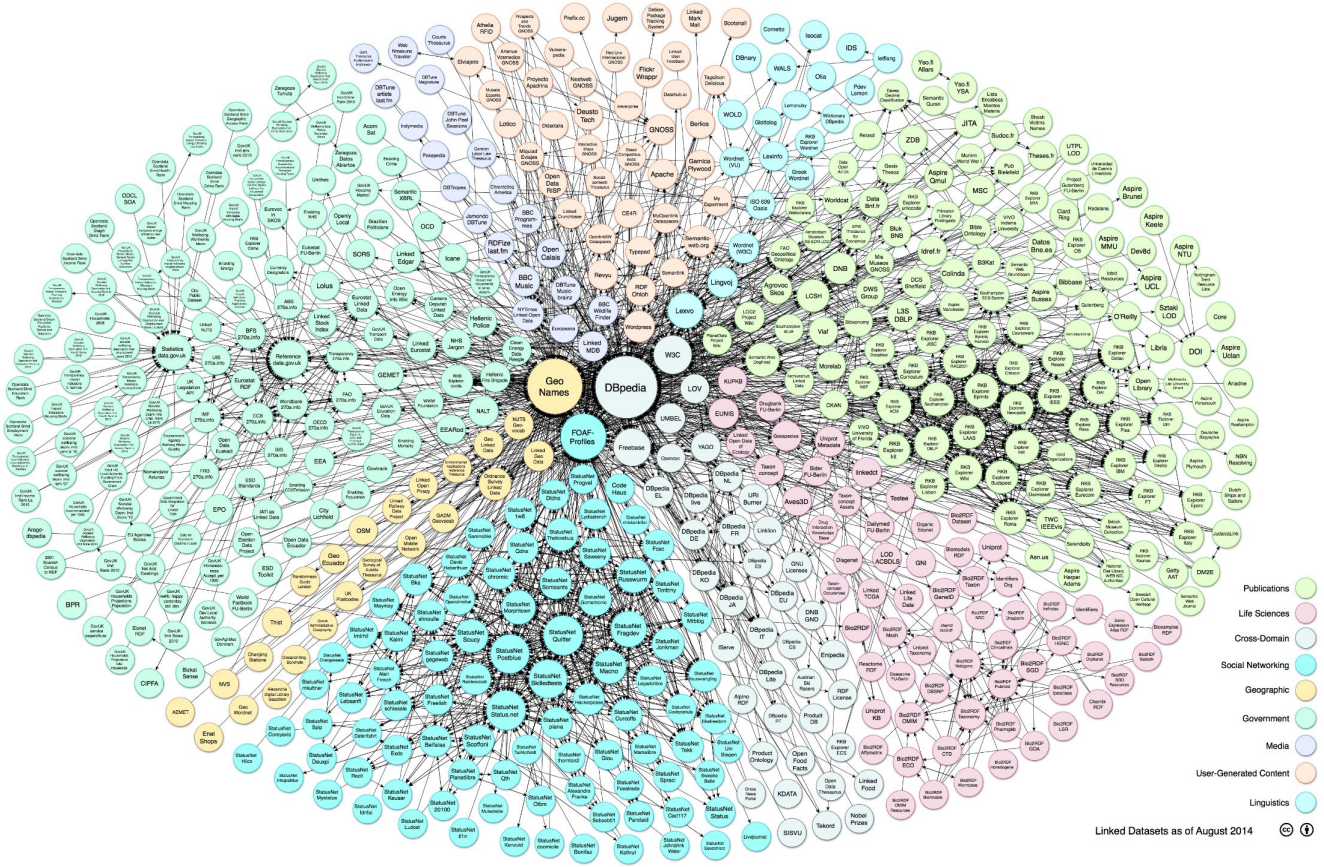
Cloud propojených dat - 2010 - 203 datových sad



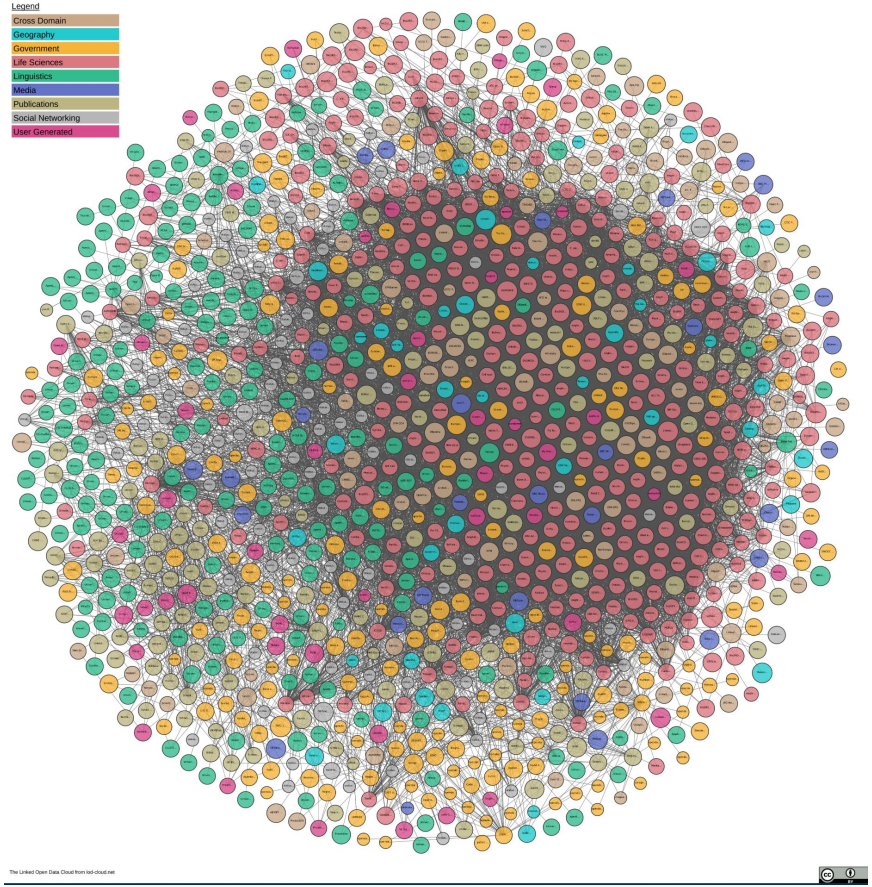
Cloud propojených dat - 2011 - 295 datových sad



Cloud propojených dat - 2014 - 570 datových sad

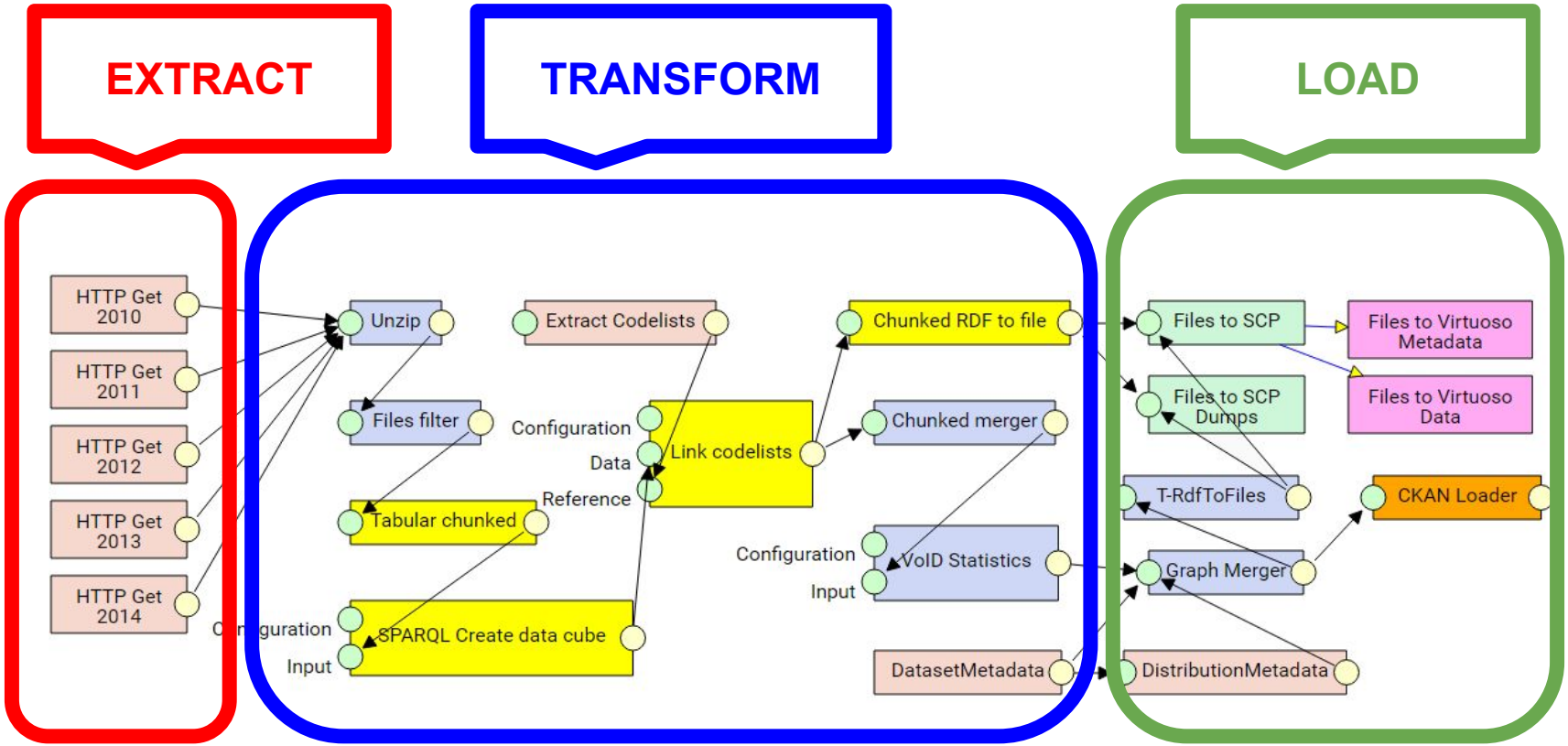


Cloud propojených dat - 2018 - 1229 datových sad



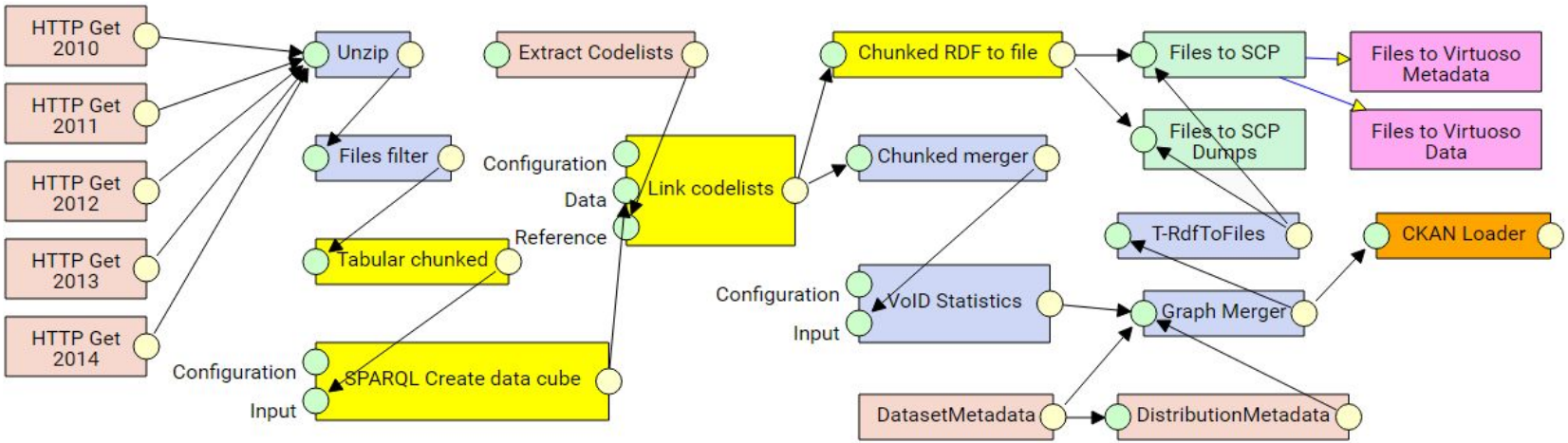
LinkedPipes ETL - Extract Transform Load pro LOD

komponenty v pipeline

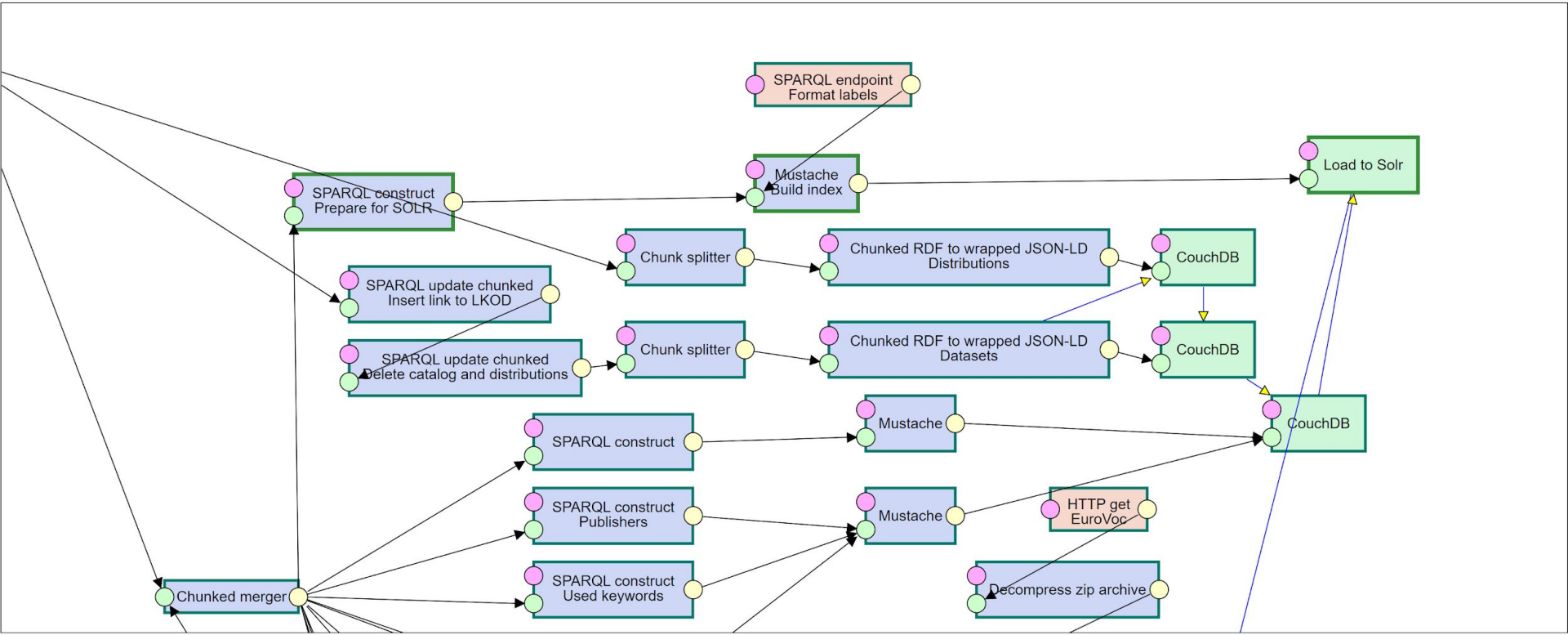


LinkedPipes ETL - Extract Transform Load pro LOD

komponenty v **pipeline**



DCAT-AP PVS, DCAT-AP LKOD, CKAN API LKOD 2 NKOD



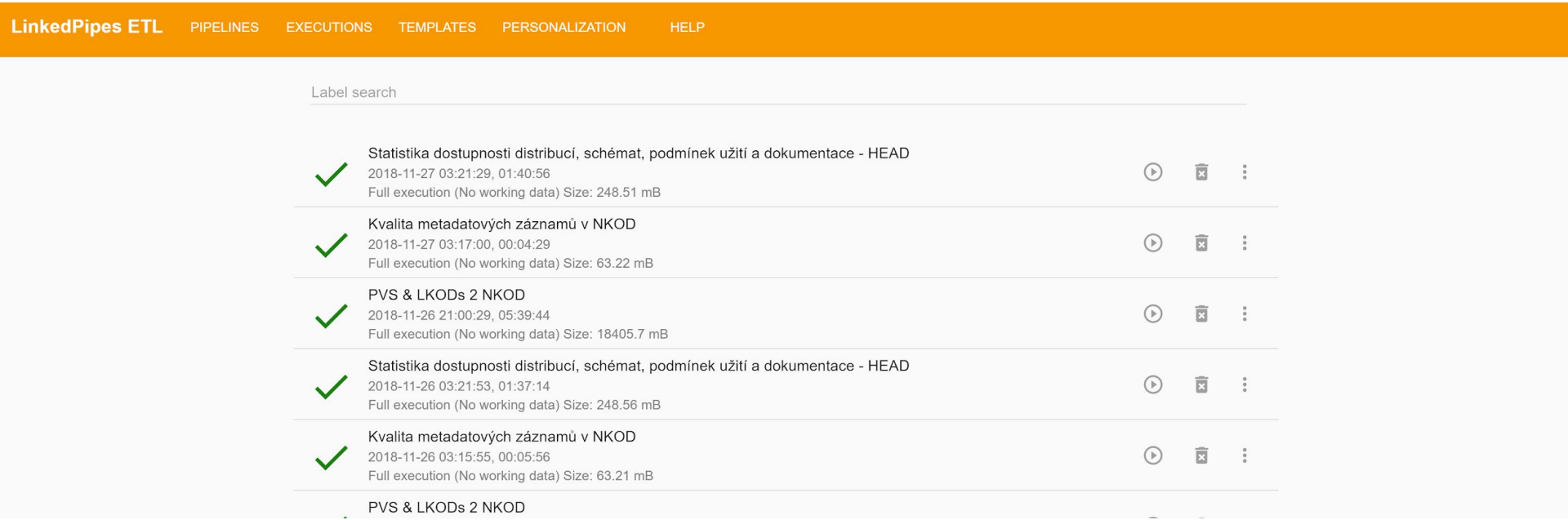
CLOSE

LP-ETL - přes 60 znovupoužitelných komponent

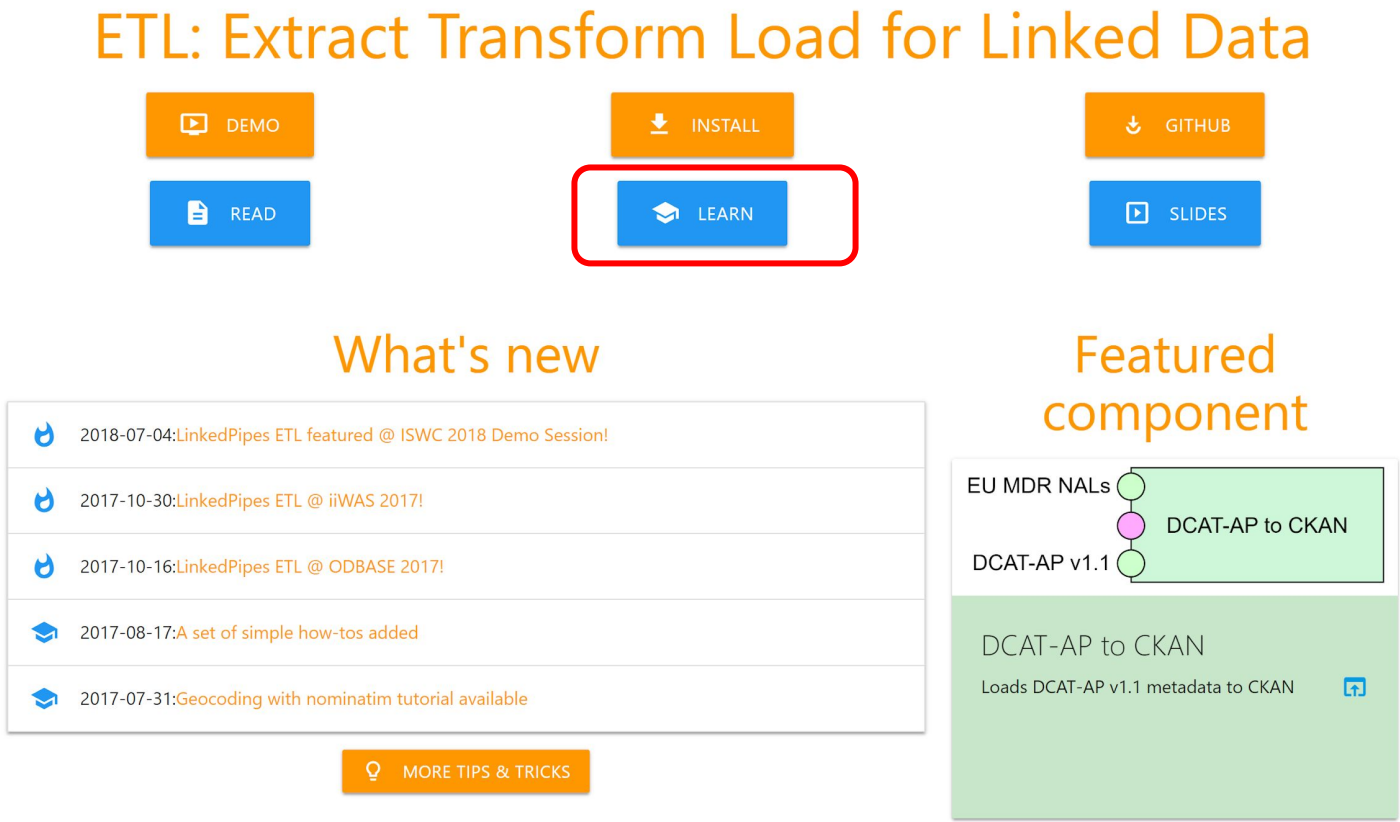
- Extraktory
  - **HTTP** GET, **SPARQL** endpoint (and variants), Text holder, Extract from **FTP**, Files from **local**
- Transformery
  - SPARQL Construct, **SPARQL** Update
  - Files to RDF, **RDF** to file
  - **XSLT**, **JSON** to JSON-LD, Tabular, **Excel** to **CSV**
  - Bing translator, Decode 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”)



LP-ETL Tutoriály a návody



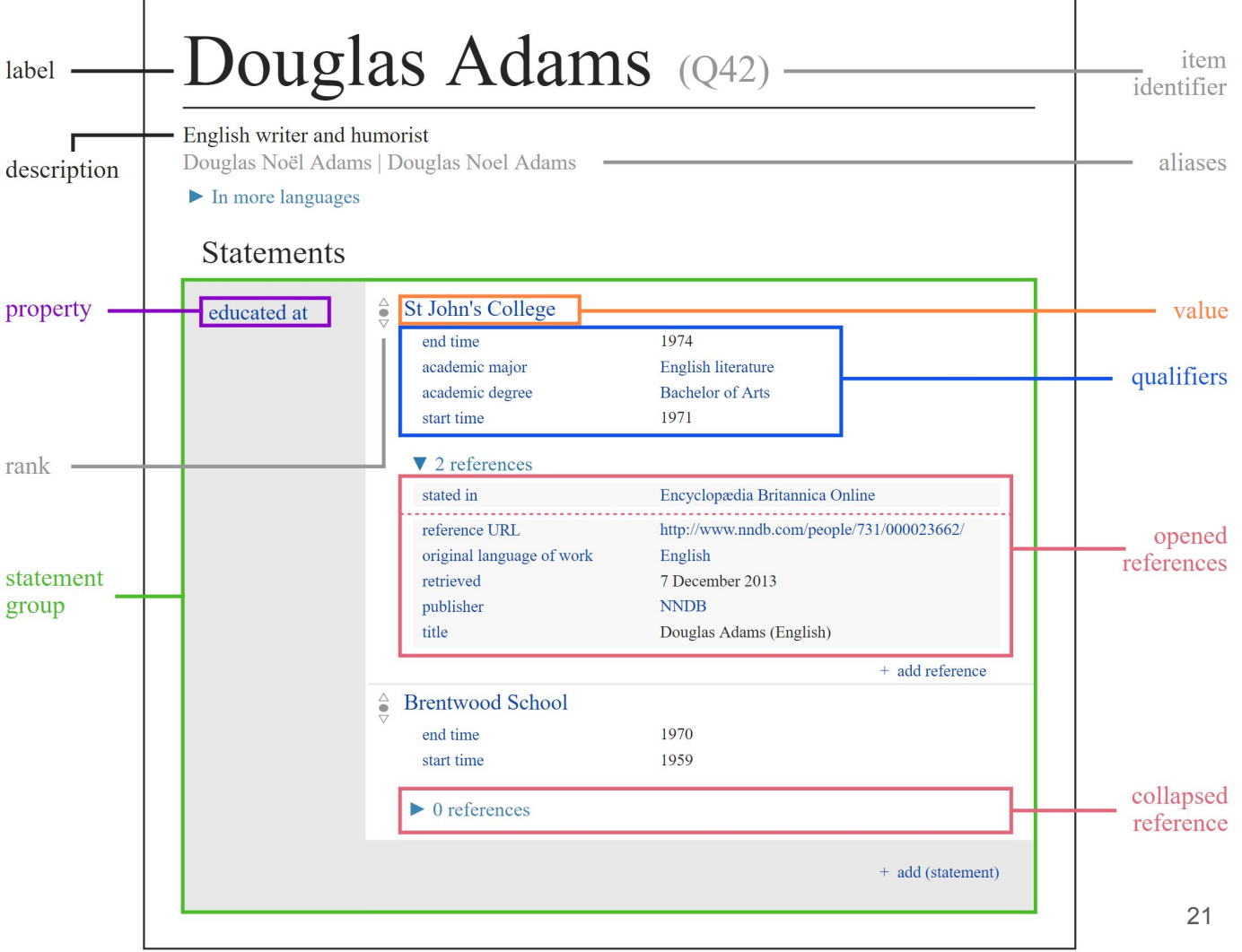
Wikidata

Databáze

- Volně dostupná
- Kolaborativní
- Vícejazyčná

Obsahuje

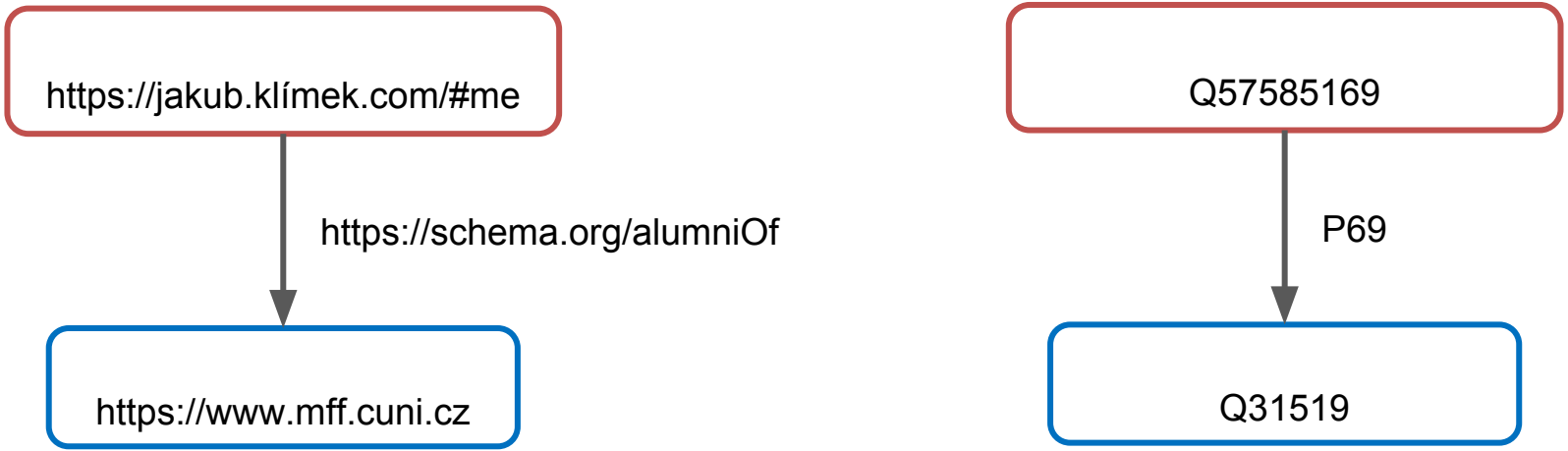
- Položky
- Tvrzení
  - Vlastnosti
  - Hodnoty
  - **Kvalifikátory**



RDF vs. (zjednodušený) datový model Wikidata

Jakub Klímek | studoval na | MFF UK .

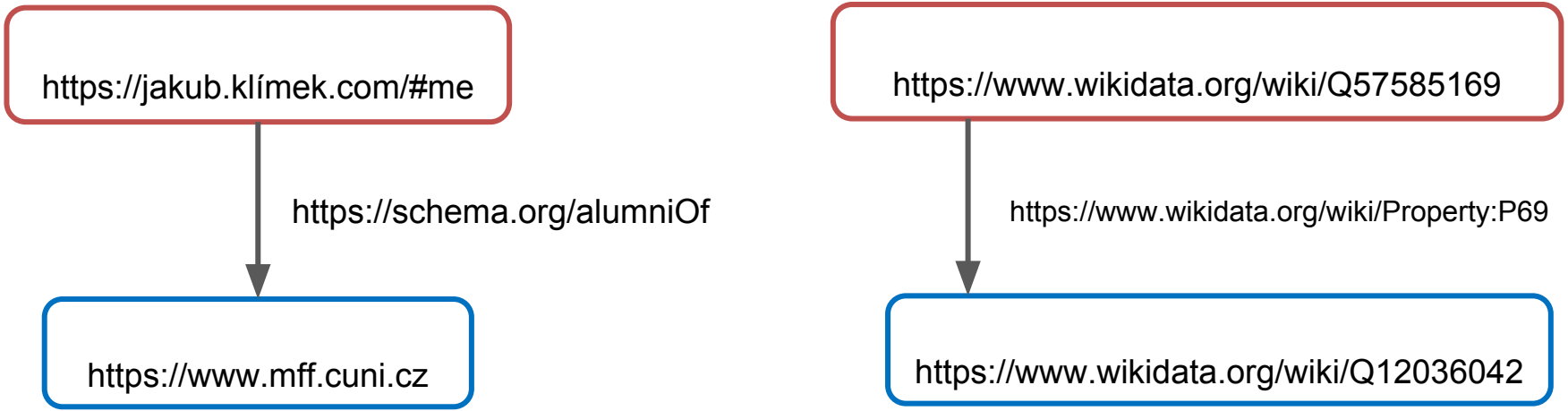
subjekt | predikát | objekt



RDF a (zjednodušený) datový model Wikidata

Jakub Klímek studoval na MFF UK .

subjekt predikát objekt



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 Wikidat
- 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