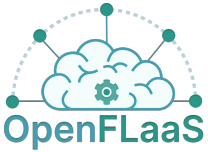
Live Fusion and Seamless Integration

Unlocking the Best-of Wikipedia, Wikidata, and Beyond with DBpedia Enterprise

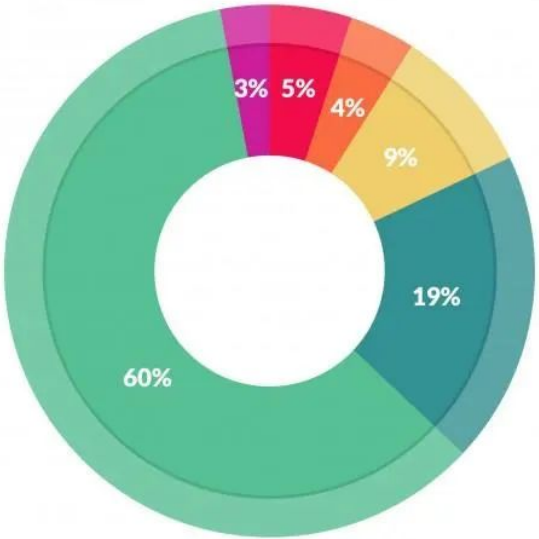
Slides: https://tinyurl.com/dbpedia-fusion-sneak

InfAI®
Institute for Applied Informatics



Bundesministerium
für Wirtschaft
und Klimaschutz

Data Acquisition Bottleneck

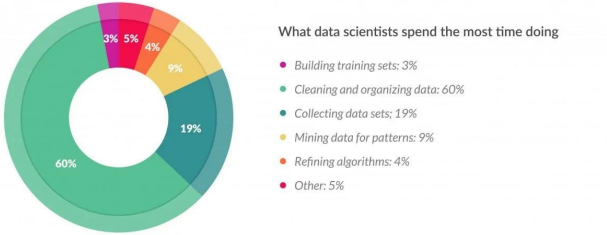


What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Source: Forbes 2016

Data Acquisition Bottleneck



Source: Forbes 2016

… i.e. one work year is spent:

80% Pre-processing

→ *January until October*

20% Actual Work

→ *November & December*

High Delay, High Opportunity Cost

DBpedia (since 2007)

- **Extracting and cleaning** data from Wikipedia and Wikidata (2 sources)
- **Organizing** data in the DBpedia Knowledge Graph

Now with fusion also **collecting** new datasets

Slides: https://tinyurl.com/dbpedia-fusion-sneak

Product: DBpedia Fusion Knowledge Graph

Powered by 18 years of specialized experience in knowledge graph pre-processing

- **Ready-to-Use Knowledge Graph:** Fully prepared for most frequent use cases
  → *Start-In-January Package*

- **Scalable Knowledge Graph Extension:** Efficiently integrate additional data on demand, more cost-effective, faster, and higher quality than in-house solutions
  → *Start-In-February Package*

- **Convenient and reliable:** Monthly updates, bug fixes, hosting, and ongoing support

Slides: https://tinyurl.com/dbpedia-fusion-sneak

Current fusion “2025-06-04-BETA”

DBpedia (EN): 14M + **Wikidata:** 117M ⟶ *122M distinct entities after **Fusion***

Partition	# Facts
Multilingual labels and descriptions	3.4 B
Relations (DBpedia Ontology)	300 M
Classes and Categories	300 M
SameAs Links	900 M
Images (Wikimedia Commons)	80 M
Source data	1.7 B
Other	1.1 B
Sum	**7.8 B**

Slides: https://tinyurl.com/dbpedia-fusion-sneak

Current fusion “2025-06-04-BETA”

Class	Entities
Person	12,181,167
Place	6,538,849
TopicalConcept	3,763,842
Work (Art & Literature)	2,421,921
TimePeriod	2,370,547
CelestialBody	2,355,783
Biomolecule	2,238,517
PopulatedPlace	2,129,303
Organisation	1,074,075
ArchitecturalStructure	1,480,287
Gene	1,224,791
Athlete	562,165
Film	326,705
Software	132,111
Dam	80,668
ResearchProject	71,047

Slides: https://tinyurl.com/dbpedia-fusion-sneak

Current fusion “2025-06-04-BETA”

- akka streaming pipeline
- postgres as entity hot cache
- neo4j as linking engine (transitive closure) - Linkmaster 3000

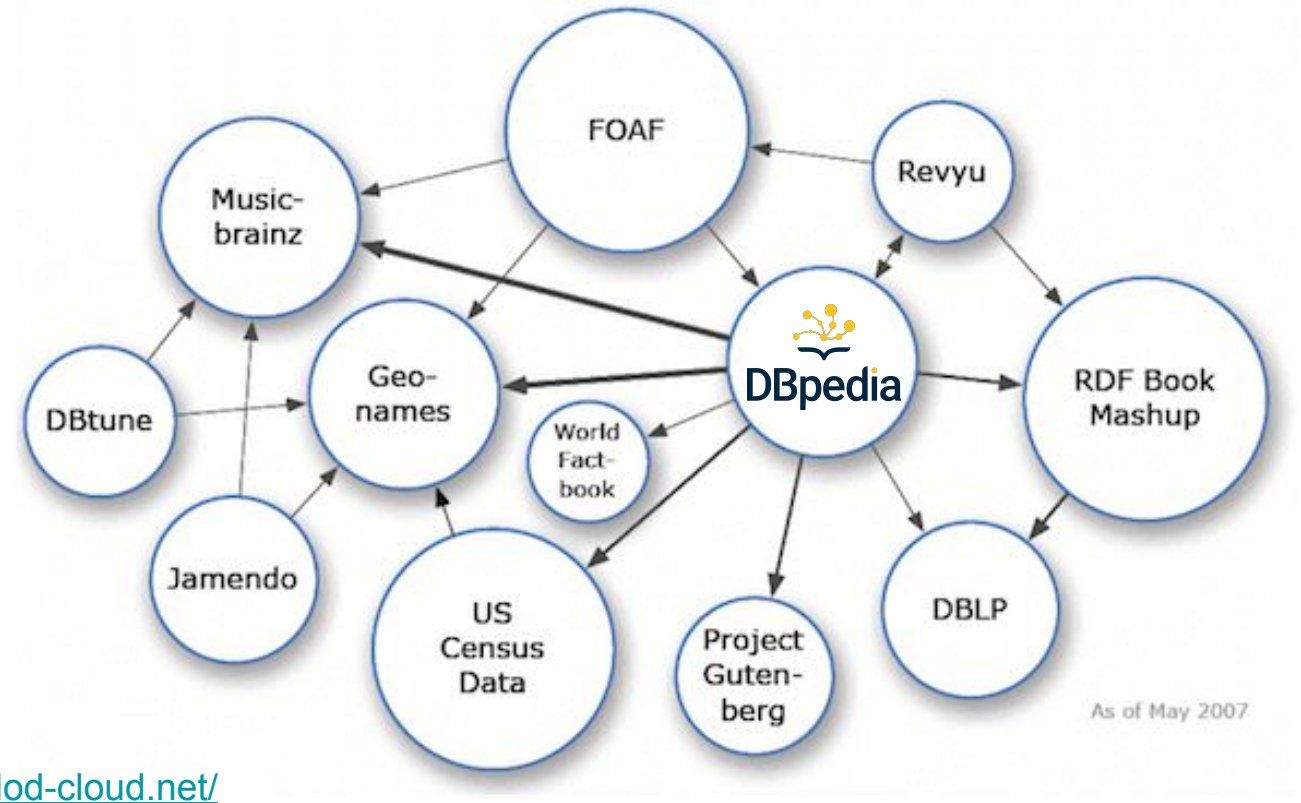
Current speed (40 cores), batch mode:

- 12h download data
- 24h extract RDF and links
- 24h fusion (2000 clusters/s, EN Wikipedia has 2 updates/s)
- 12h validation and quality control

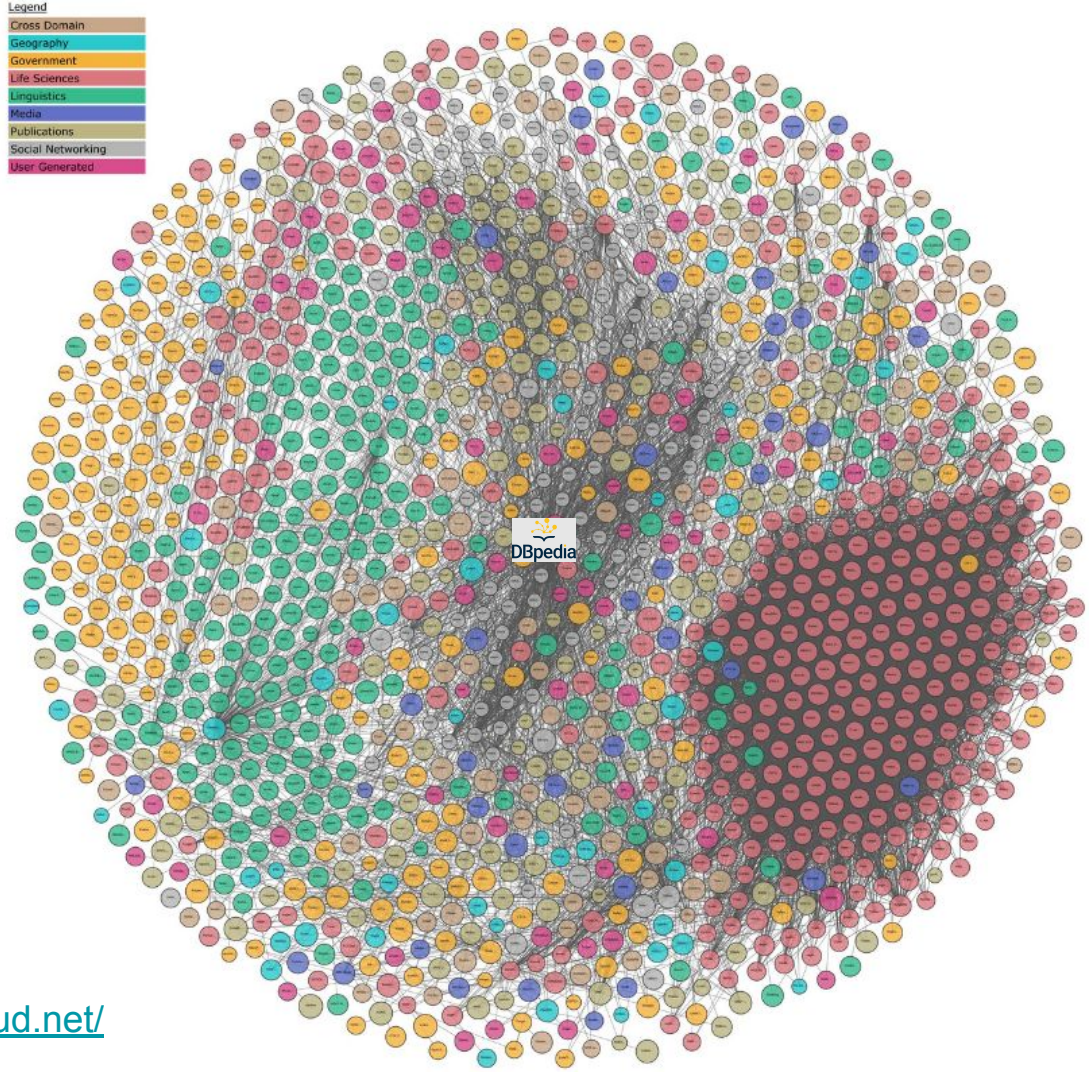
= 3 days

Slides: https://tinyurl.com/dbpedia-fusion-sneak

Growing the DBpedia Fusion KG from Linked Open Data

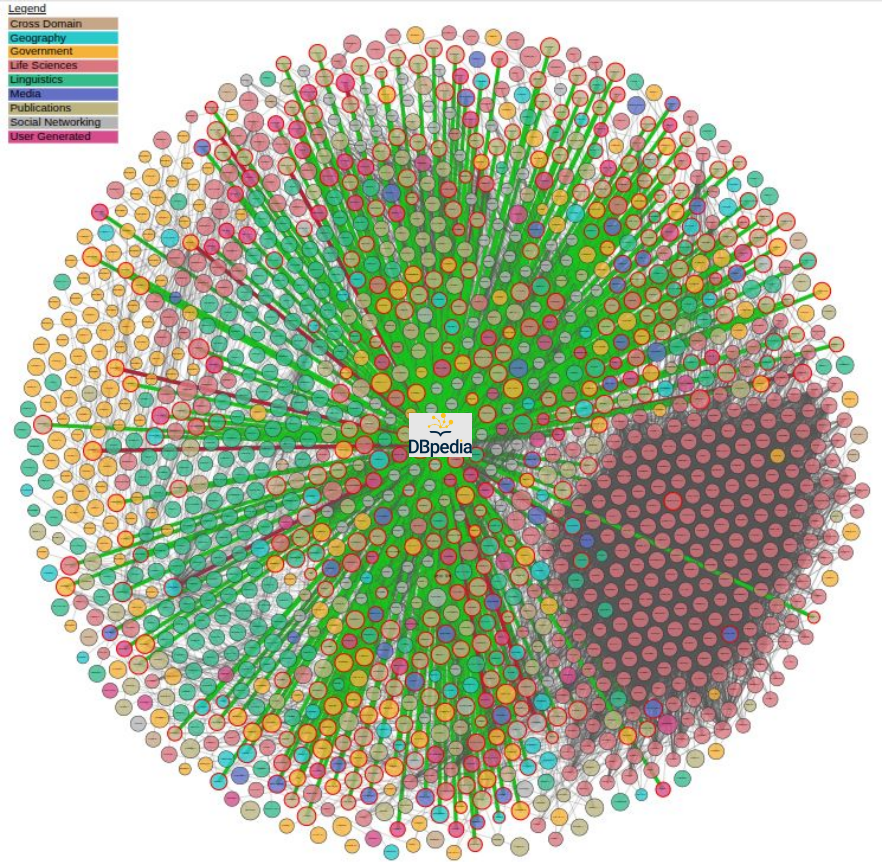


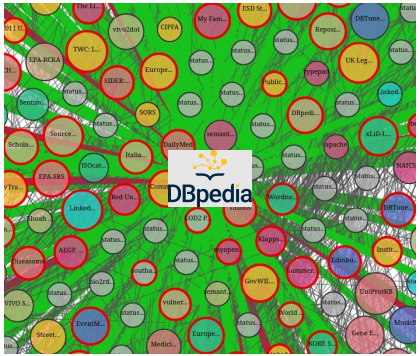
May 2007
Source: https://lod-cloud.net/



May 2025
Source: https://lod-cloud.net/

Growing the DBpedia Fusion KG from Linked Open Data





DBpedia is the most linked data set

Hub & Authority

DBpedia Fusion does the hard lifting:

80% collecting, cleaning and organizing

Growing the DBpedia Fusion KG from Linked Open Data

Example:

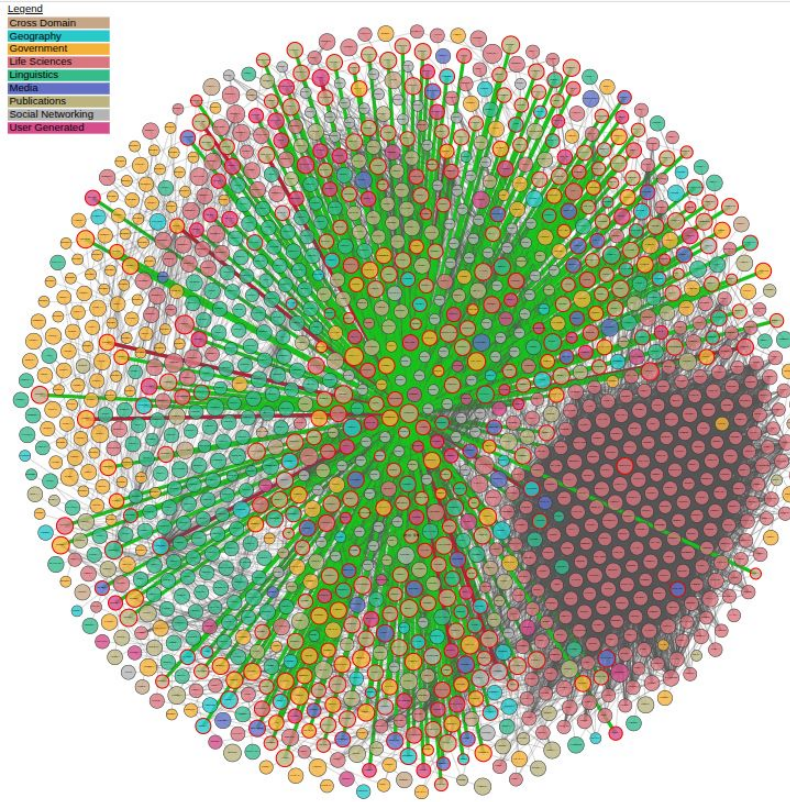
DNB (German National Library, CC0)
https://data.dnb.de/opendata/

12.2 M persons in current fusion

6 M persons in DNB, 2 M with DBpedia links

12.2 M + 6 M = 16.2 M (+ 4 M gain)

- Fusion KG grows by 4 M entities
- Data from 2 M overlapping persons is used to
  - increase density (complementary data)
  - fact check existing data



DBpedia Use Cases

track record for 18 years, elevated by fusion:

- Bootstrap: import DBpedia Fusion KG as a basis to build your own KG
  - no need to start from scratch
- Upgrade your enterprise data
  - use DBpedia Fusion KG to connect databases and analyse documents
- Background knowledge and structure for machine learning
- Enrich your data:
  - images, textual descriptions, relations
- Terminology grounding
  - use DBpedia as a vocabulary to ground your terminology
- Search enrichment: Google KG, Zalando users searched for Beyonce, when searching for her fashion label Ivy Park <Beyonce> dbo:founded <Ivy_Park>

Slides: https://tinyurl.com/dbpedia-fusion-sneak

LLM vs. factual knowledge

List head of government of the 5 countries with the highest GDP as table, no googleing

ChatGPT June 7th, 2025

Based on data up to my last update (mid-2024), here is a table listing the **Heads of Government** (usually Prime Ministers or Presidents, depending on the country's system) for the **five countries with the highest nominal GDP**:

Rank	Country	Head of Government	Title
1	United States	Joe Biden	President
2	China	Li Qiang	Premier
3	Japan	Fumio Kishida	Prime Minister
4	Germany	Olaf Scholz	Chancellor
5	India	Narendra Modi	Prime Minister

LLM need fresh and validated data, DBpedia fusion can be used for fact-checking

DBpedia Fusion KG Sneak Preview (Available Now)

Business Source License (BSL) 1.1

- Free for Research & **Non-Production Use** (Full dumps available)
- Additional license for **Production Use** required (Price 3 PM, save 10 PM)

Slides: https://tinyurl.com/dbpedia-fusion-sneak

DBpedia Fusion KG Sneak Preview (Available Now)

Business Source License (BSL) 1.1

- Free for Research & **Non-Production Use** (Full dumps available)
- Additional license for **Production Use** required (Price 3 PM, save 10 PM)

For Research & DBpedia Community

- BSL 1.1 no barriers for research papers and prototyping, full data accessible
- Open Core: open datasets will stay open
- new interesting data for interesting problems: linking, mapping, data quality, scalability

Slides: https://tinyurl.com/dbpedia-fusion-sneak

DBpedia Fusion KG Sneak Preview (Available Now)

Business Source License (BSL) 1.1

- Free for Research & **Non-Production Use** (Full dumps available)
- Additional license for **Production Use** required (Price ~3 PM, save 10 PM)

For industry

- ready-to-use knowledge graph
- free up to 80% time of your data engineers to concentrate on their actual work (value generation)
- convenience and reliability: monthly updates, bug fixes, hosting, and ongoing support
- steer the expansion of fusion into domains relevant to your company (health, business intelligence, industry 4.0, etc.)

Slides: https://tinyurl.com/dbpedia-fusion-sneak

DBpedia Fusion KG Sneak Preview (Available Now)

Register, download and try:

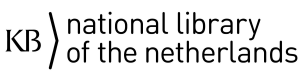
- https://www.dbpedia.org/resources/knowledge-graphs/
- https://data.dbpedia.io/databus.dbpedia.org/dbpedia-enterprise/dev/fusion-sneak-preview

Contact us for a demo session and tell us what data you need: dbpedia@infai.org

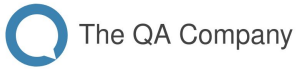
Slides: https://tinyurl.com/dbpedia-fusion-sneak

HTWK
Leipzig University
of Applied Sciences

UNIVERSIDAD
POLITÉCNICA
DE MADRID



INSTYTUT INFORMATYKI GOSPODARCZEJ /
BUSINESS INFORMATION SYSTEMS INSTITUTE



Slides: https://tinyurl.com/dbpedia-fusion-sneak

Backup Slides

Backward compatibility how to upgrade to fusion

SELECT ?birthDate WHERE {

<http://dbpedia.org/resource/Friedrich_Merz> dbo:birthDate ?birthDate .

<https://dbpedia.io/id/4PgFKDCQCtWKSnkpR4JGNcT#e> dbo:birthDate ?birthDate .

Upgrading a DBpedia query to DBpedia Fusion
?s owl:sameAs <http://dbpedia.org/resource/Friedrich_Merz> .
?s dbo:birthDate ?birthDate .

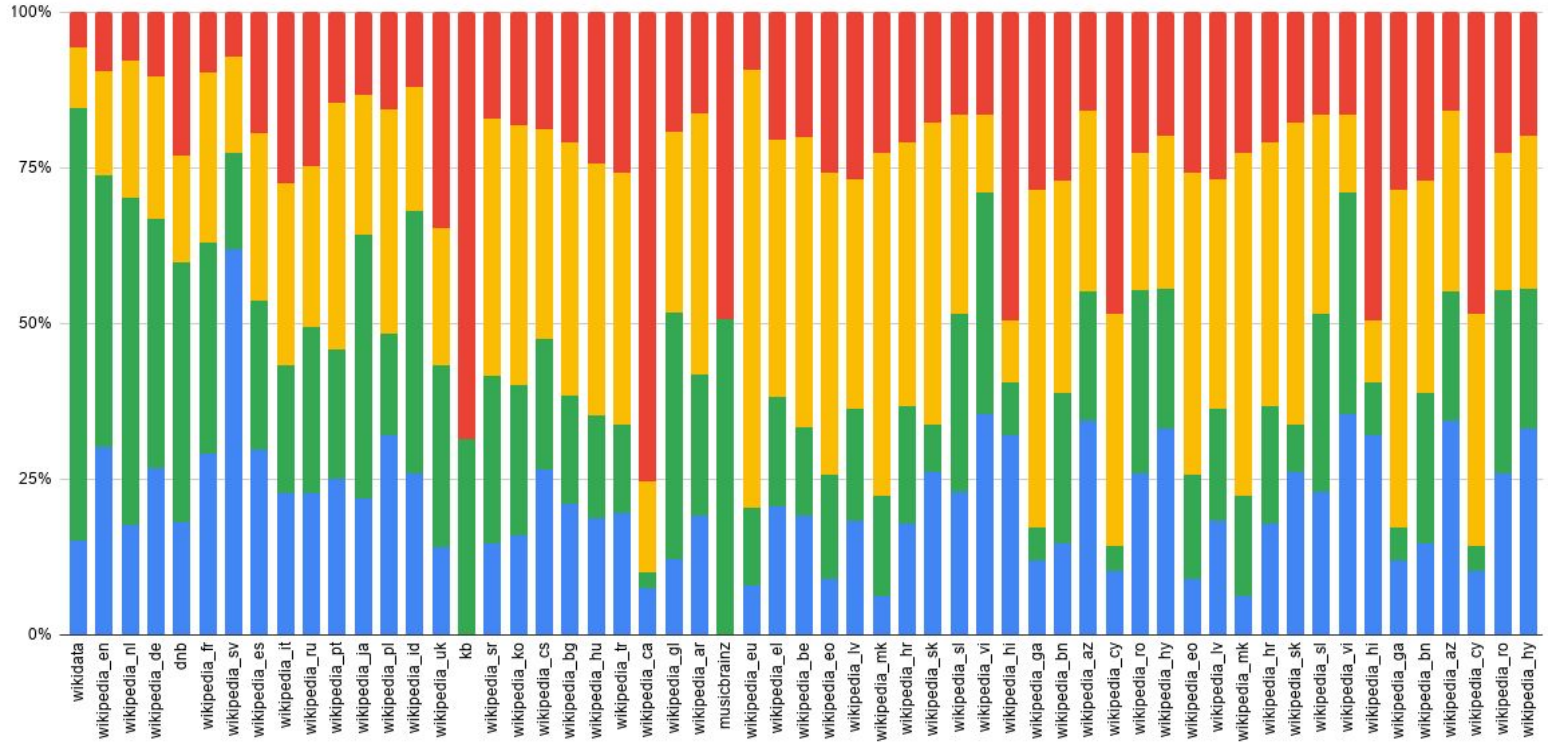
Upgrading a Wikidata query to DBpedia Fusion
?s owl:sameAs <https://www.wikidata.org/wiki/Q566257> .
?p owl:equivalentProperty <https://www.wikidata.org/wiki/Property:P569>
?s ?p ?birthDate .

Best-of Fusion

not competition, DBpedia is a stakeholder of Wikipedia and Wikidata

	EN Wikipedia	Wikidata	DBpedia Fusion
Size	14M entities	117M entities	122M entities
Data Overlap / Gain	25% overlap, 25% true complement,, 25% partial complement, 25% in conflict		Best-of
Taxonomy	DBpedia Ontology + additional Schemas	Class Tagging	Consolidated schema
Links	sameAs / seeAlso (external websites)	unclear semantics	Consolidated links + additional links
Freshness/Quality	high freshness, but extraction errors	less freshness, but no extraction errors	Best-of
License	Open	Open	Mixed (BSL 1.1)
Industry-Ready?	as-is in-house cleaning required	as-is in-house cleaning required	support & cleaning as a service

Slides: https://tinyurl.com/dbpedia-fusion-sneak



blue=synched, green=complementary, yellow=partially synched, red=conflicts

Sebastian Hellmann et al: Towards a Systematic Approach to Sync Factual Data across Wikipedia, Wikidata and External Data Sources. Qurator 2021

Slides: https://tinyurl.com/dbpedia-fusion-sneak

Link Quality Methodology

Traditional: linking tool is run once on snapshot dumps, precision is preferred over recall because of strong semantics and large connected components

Our methodology:

- not one-of, but evolving, live system (entities change)
- number of links is reduced to one per entity (the masterlink), less inferred error and easier to debug
- strong focus on error detection using novel methods
- live architecture
  - make it easy to fix errors Link Master 3000
  - leveraging of AI to assist humans

Slides: https://tinyurl.com/dbpedia-fusion-sneak

Call to Action

1. Register, download and try the DBpedia Fusion Knowledge Graph (beta):

   TODO LINK

   DBpedia Fusion is available under BSL 1.1: unrestricted academic and non-productive use

1. Contact us to speed up your data projects and data infrastructure and validated data to fact-check your LLMs

💡 *“We did the hard integration work. You don’t need to.”*

Slides: https://tinyurl.com/dbpedia-fusion-sneak

Statistics

18 years experience in Knowledge Graph Engineerin

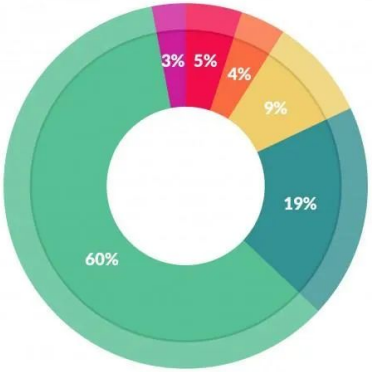
Open Core (CC-BY-SA, no registration)

- DBpedia EN (10M Entities, 1B Facts)
- DBpedia Multilingual (66M Entities, 16B Facts
- DBpedia Wikidata (90M Entities, 5B Facts)
- Open Services: SPARQL, Spotlight, Lookup with 8M daily requests

Fused Knowledge Graphs

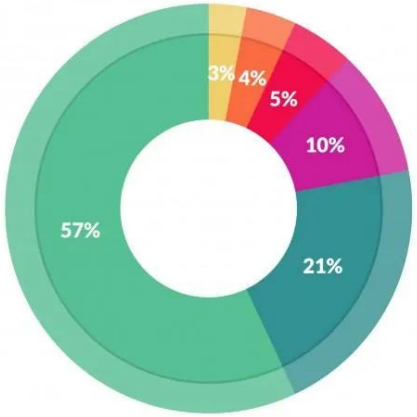
- DBpedia Best-Of-Fusion

Pains



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*



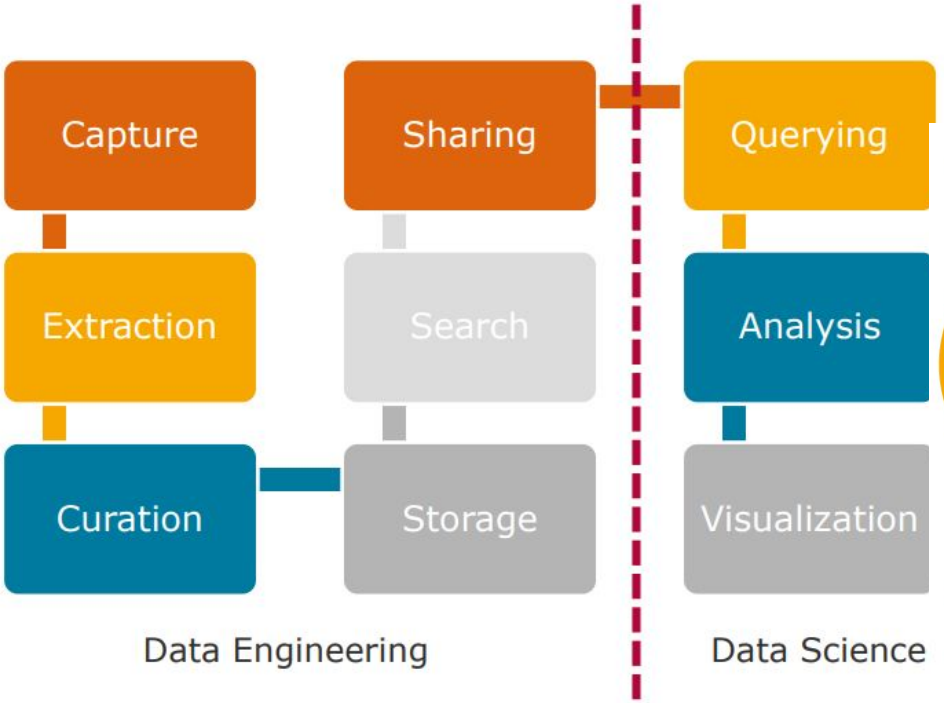
What's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

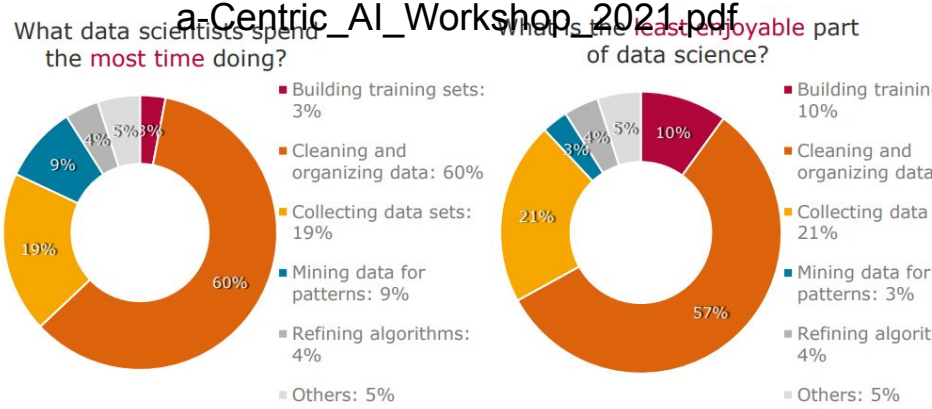
DBpedia Fusion – Licensed under BSL 1.1

1. **Free for Research & Non-Production Use (Full dumps available)**
   The license allows unrestricted use for academic, scientific, and non-production purposes—ideal for open research and community collaboration.
2. **Open Source After Change Date**
   The code becomes open under a more permissive license (e.g., CC-BY-SA) after a fixed time period (typically 4 years or sooner if announced).
3. **Commercial Use Requires License (Value generation)**
   For production or commercial deployment before the Change Date, a separate commercial license is required.
4. **Protects Innovation, Enables Ecosystem**
   BSL balances protecting innovation with fostering a vibrant open-source ecosystem by inviting contribution and experimentation.

Bad Data, Bad files



https://hpi.de/oldsite/fileadmin/user_upload/fachgebiete/naumann/Talks/BadFilesBadDataBadResults_Data-Centric_AI_Workshop_2021.pdf



"Cleaning Data: Most Time-Consuming, Least Enjoyable Data Science Task", Gil Press, Forbes, March 23rd, 2016
http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

2007
formation of the Linked Data Cloud

2010
Open editing of DBpedia Ontology
A new type of Cyc?

2012-2016
Covering all 140 Wikipedias, Commons, Wikidata
14.4 B facts extracted

2018

2020 - 22 B facts per month

Huge Linked Data - derived Open Knowledge Graphs (OKG)

2007 2015 2020

2007
first Wikipedia extraction, SPARQL Linked Data

2009
Major boost in KG and Linking Research

2011
Industry adoption

2014
Foundation of DBpedia Association Leipzig

2017
SHACL W3C Standard by Uni Leipzig
Test-driven KG development

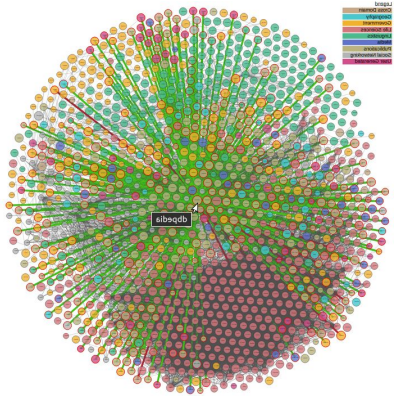
2019
DBpedia Innovation Platform - Central hub for Linked Data Technology and Ecosystem

2020 - FAIR Linked Data

Findable
Accessible
Interoperable
Reusable

CA23147 - Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs (GOBLIN)

Downloads



https://www.cost.eu/actions/CA23147/