```
######
#     #   ##   #####   ##   #####  #    #  ####
#     #  #  #    #    #  #  #    # #    # #
#     # #    #   #   #    # #####  #    #  ####
#     # ######   #   ###### #    # #    #      #
#     # #    #   #   #    # #    # #    # #    #
######  #    #   #   #    # #####   ####   ####
```

DBpedia and the Custody of Linked Open Data

Sebastian Hellmann

http://dbpedia.org

InfAI®
Institut für Angewandte Informatik

Outline

DBpedia History and Challenges

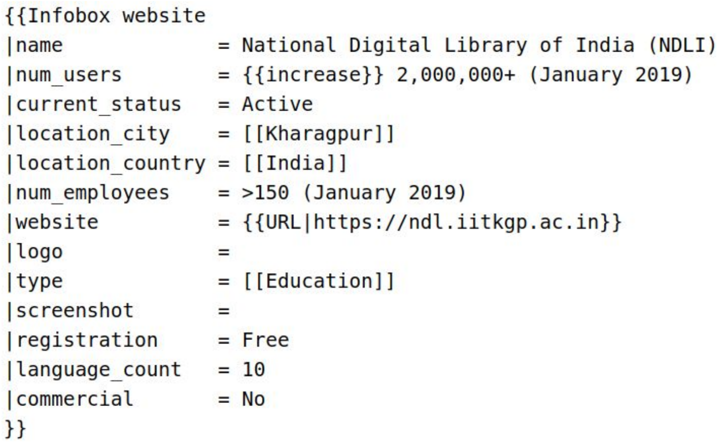
- The beginning phase (2007 - 2011)
- The manifestation phase (2011 - 2016)
- Glass ceiling phase (2016 - 2019)
- Lessons learned

The future brings solutions

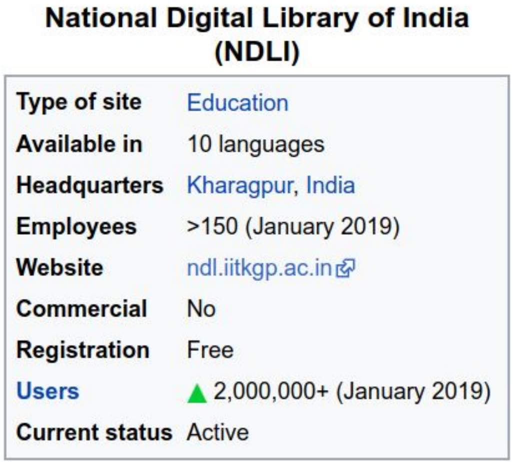
- Skyrocketing phase (2020 - 2025)
  - Databus
  - Knowledge Library
  - FlexiFusion

Beginning Phase of DBpedia 2007-2011

2007 - First Extraction of Wikipedia’s Infoboxes



```
{{Infobox website
|name             = National Digital Library of India (NDLI)
|num_users        = {{increase}} 2,000,000+ (January 2019)
|current_status   = Active
|location_city    = [[Kharagpur]]
|location_country = [[India]]
|num_employees    = >150 (January 2019)
|website          = {{URL|https://ndl.iitkgp.ac.in}}
|logo             =
|type             = [[Education]]
|screenshot       =
|registration     = Free
|language_count   = 10
|commercial       = No
}}
```



National Digital Library of India (NDLI)

Type of site	Education
Available in	10 languages
Headquarters	Kharagpur, India
Employees	>150 (January 2019)
Website	ndl.iitkgp.ac.in
Commercial	No
Registration	Free
Users	▲ 2,000,000+ (January 2019)
Current status	Active

- https://en.wikipedia.org/wiki/National_Digital_Library_of_India
- (en) https:/dbpedia.org/resource/National_Digital_Library_of_India
- (global) https://global.dbpedia.org/?s=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FNational_Digital_Library_of_India
- (global) https://global.dbpedia.org/id/3FdBE

Beginning Phase of DBpedia 2007-2011

2007 - First Extraction of Wikipedia’s Infoboxes

Query Wikipedia Like a Database:

soccer players, who are born in a country with more than 10 million inhabitants, who played as goalkeeper for a club that has a stadium with more than 30.000 seats and the club country is different from the birth country

Starting members: OPENLINK® SOFTWARE UNIVERSITÄT MANNHEIM InfAI® Institute for Applied Informatics TIB LEIBNIZ-INFORMATIONSZENTRUM TECHNIK UND NATURWISSENSCHAFTEN UNIVERSITÄTSBIBLIOTHEK

- Uni Leipzig, Open Link Software, FU Berlin (affiliation change of key persons)

Beginning Phase of DBpedia 2007-2011

Exceptional boost of research and industrial innovation

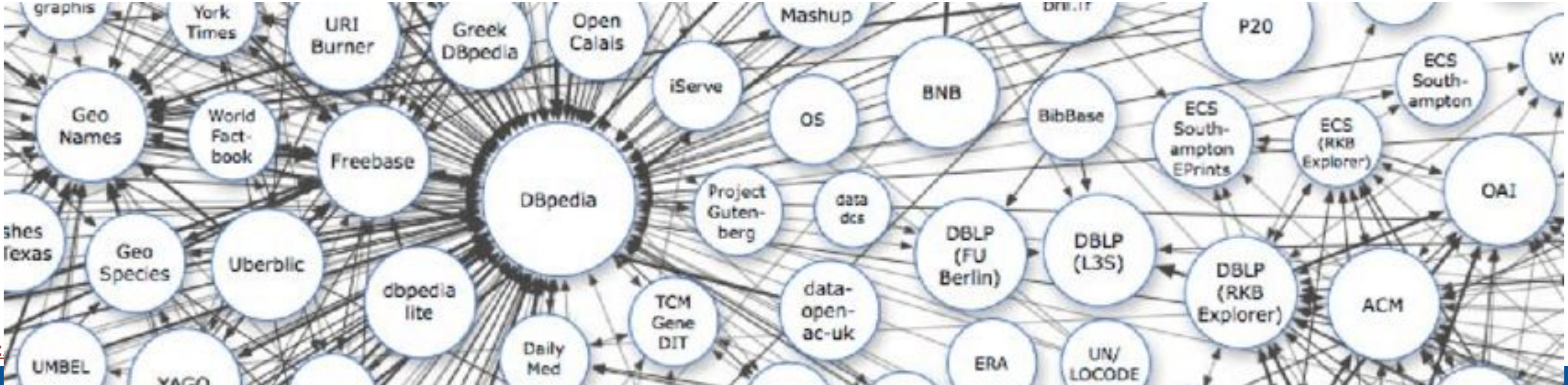
- DBpedia Dataset became the foundation for around 25000 scientific papers
- High industry adoption: BBC, New York Times, Yahoo, Watson (Jeopardy)
- Emergence of semantic technologies:
  - Knowledge Extraction
  - Entity Linking (Databases)
  - Entity Linking (Natural Language Processing) - DBpedia Spotlight
  - Graph Databases

Beginning Phase of DBpedia 2007-2011

Linked Data Best Practices

https://www.google.com/search?q=lod+cloud
https://slidewiki.org/presentation/661/_/661/5475-3/#/slide-5475-3

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. so that they can discover more things. **[Identifier Linking Paradigm]**



https:

Manifestation phase 2011-2016

Growth phase

- Internationalization
    - Covering all 140 Wikipedia languages
    - 22 language chapters, worldwide DBpedia embassies and institutional collaborators
- http://mappings.dbpedia.org
    - ~300 editors
    - DBpedia Ontology incl. links to other ontologies
    - Data cleaning rules
- Data extensions such as Wikimedia Commons, Wikidata, NIF
    - 14 Billion statements
- Ontology contributions YAGO, SUMO, Umbel, LHD, DBTax
    - 8 taxonomies to query DBpedia

Manifestation phase 2011-2015

DBpedia

Foundation of the DBpedia Association in 2014:

- Reliable core of organisational supporters for sustainability and network multiplication
- A network around a non-profit coordinator (DBpedia Association, Primus inter pares)

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Uniwersytet Ekonomiczny w Poznaniu 1926

Universität Mannheim

eccenca
command your data!

Semantic Web Company
linking data to knowledge

Fraunhofer IAIS

TIB
Leibniz-Informationszentrum Technik und Naturwissenschaften Universitätsbibliothek

OpenLink® Software

diffbot

Open Knowledge International

Network Institute

huygens ing

KB
Koninklijke Bibliotheek
Nationale bibliotheek van Nederland

wpsemantix
The SQL Knowledge Graph Company

ontotext

imec

InfAI®
Institute for Applied Informatics

Springer Nature

Sound and Vision

Linked Open Data Initiative
LinkedOpenData.jp ★★★★★

Glass ceiling phase 2016 - 2019

Glass ceiling is a term from gender inequality

Appropriate metaphor:

- Women have to work harder than men for their career
- Pushing harder yields less and less results

https://wiki.ubc.ca/Glass_Ceiling

Glass ceiling phase 2016 - 2019

Glass ceiling is a term from gender inequality

Appropriate metaphor:

- Women have to work harder than men for their career
- Pushing harder yields less and less results

Totally applies to **all** data projects

https://wiki.ubc.ca/Glass_Ceiling

Glass ceiling phase 2016 - 2019

What happened in the world?

- Fierce competition over editorial workforce
  - Orcid
  - Wikidata
  - Microsoft Academic
  - Google
  - Thousands more
- Identifier wars (Identifier Linking Paradigm)
  - everybody wants you to include their IDs in order to gain attention and workforce
- Plethora of data graveyards (project end, running out of funding)

What did DBpedia do?

- We formed a think tank with our members and community (engineers and technology leaders) and discussed and re-designed

Results:

- Open data needs a **scalable** business model
- Identified problems of decentralization and innovated to provide solutions to boost decentralized approaches (LOD)

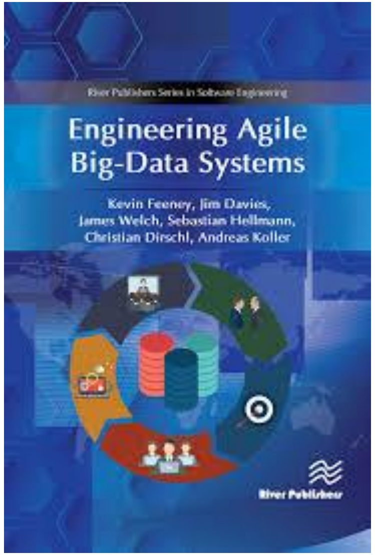
Only **coordinated decentralisation** is able to break the glass ceiling

Foundations

ALIGNED: Aligning Software & Data Engineering 2015 - 2018
http://aligned-project.eu

Engineering Agile Big-Data Systems
defines three dimensions to evaluate systems:

- productivity
- quality
- agility

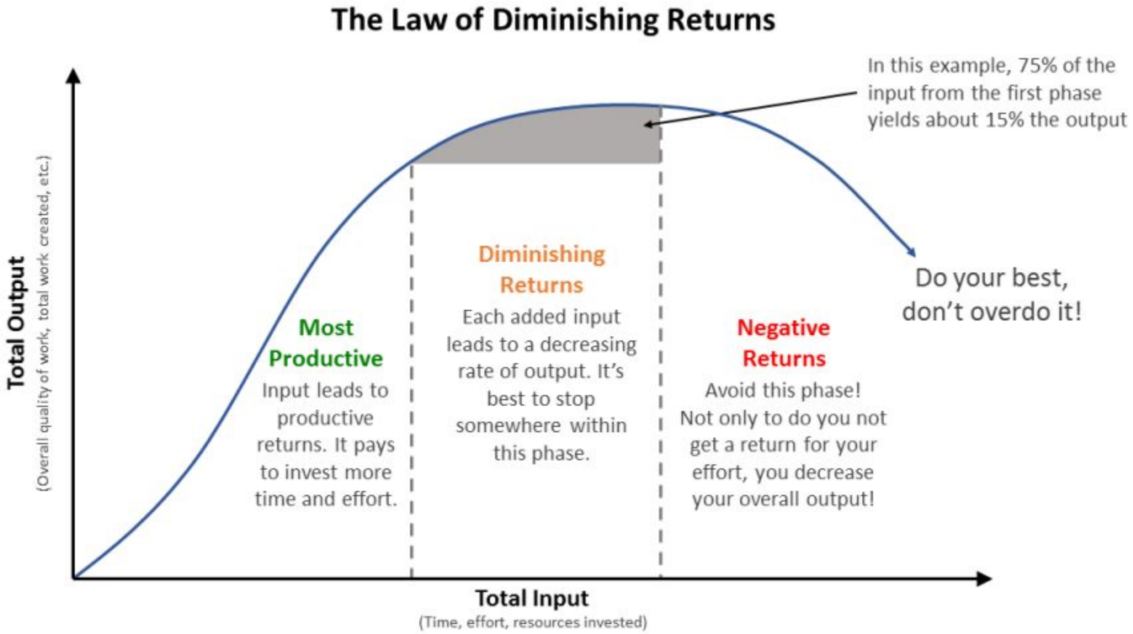

River Publishers Series in Software Engineering
Engineering Agile Big-Data Systems
Kevin Feeney, Jim Davies, James Welch, Sebastian Hellmann, Christian Dirschl, Andreas Koller
River Publishers

https://www.riverpublishers.com/book_details.php?book_id=659

4 riders of the datacalypse

1. **Law of Diminishing Returns**
   -> intrinsic to data quality, the glass ceiling
2. Copying data
   -> Duplication of effort by copy
3. Non-collaboration of data publishers
   -> Multiplication of effort by individual non-synced data sets
4. Network Disaster of Linking / Mapping
   -> Multiplication of effort at consumer side

Law of Diminishing Returns


The Law of Diminishing Returns
Total Output
(Overall quality of work, total work created, etc.)
Total Input
(Time, effort, resources invested)
In this example, 75% of the input from the first phase yields about 15% the output
Most Productive
Input leads to productive returns. It pays to invest more time and effort.
Diminishing Returns
Each added input leads to a decreasing rate of output. It's best to stop somewhere within this phase.
Negative Returns
Avoid this phase! Not only to do you not get a return for your effort, you decrease your overall output!
Do your best, don't overdo it!

Data Quality is pareto-efficient 20/80 rule

Law totally applies
-> No exception to manual curation or AI-generated datasets

More data (coverage) means lower quality

More quality means remaining errors are harder to fix

Update is more difficult than **Create**

https://tinyurl.com/dbpedia-kedl-2019

Law of Diminishing Returns

Online communities aka users are great

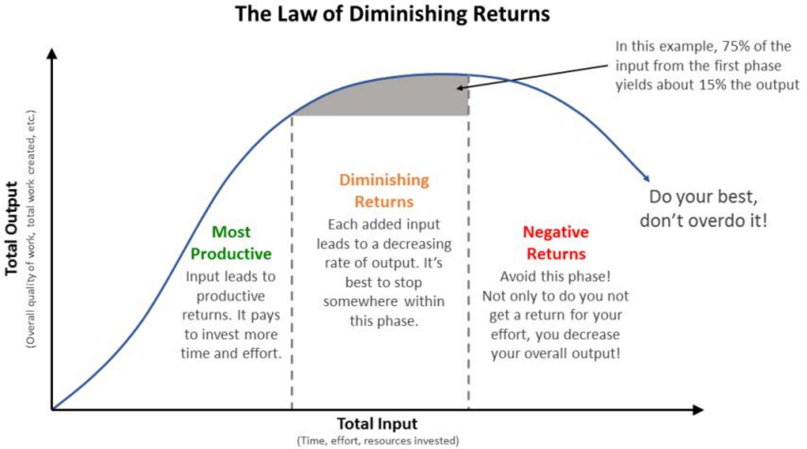
-> they are easy to motivate by the “greater good”, “better data”

-> they work for free, i.e. zero or low cost on the budget

If you are already in the “Diminishing Returns” phase, what happens if you double community activity (workforce)?

Twice as much data?
-> No, maybe 25% more

Better overall quality?
-> Maybe not, if you increase data size



The Law of Diminishing Returns

In this example, 75% of the input from the first phase yields about 15% the output

Most Productive
Input leads to productive returns. It pays to invest more time and effort.

Diminishing Returns
Each added input leads to a decreasing rate of output. It’s best to stop somewhere within this phase.

Negative Returns
Avoid this phase! Not only to do you not get a return for your effort, you decrease your overall output!

Do your best, don’t overdo it!

Total Output
(Overall quality of work, total work created, etc.)
Total Input
(Time, effort, resources invested)

Law of Diminishing Returns

No exceptions.

Test-driven Evaluation of Linked Data Quality by Dimitris Kontokostas et al. (2014, WWW) in a joint project with Dutch Libraries (Enno Meijers)

Dimitris Kontokostas was the former CTO of DBpedia and editor of the W3C standard **Shapes Constraint Language (SHACL) in 2017**, implemented as OS in RDFUnit

Test-driven data engineering follows the 20/80 rule

Small set of initial test cases is efficient, then you hit the glass ceiling

Rider no. 2: Copying data

- 10,000 downloads of a dataset dump
  -> If one unparseable line needs 15 minutes to find and fix, we are talking about 104 days of work
- publishers are struggling with data quality, but their consumers have invested 50-5000 times their effort in cleaning
- The ~600k yearly file downloads and 20 million API hits daily of DBpedia re-incarnate as local data quality problems

Step 1: Download data
Step 2: Clean and integrate

If we could just capture consumer-invested time …

Rider no. 3: Non-collaboration

What have the following organisations in common (could be thousands more)?

DEUTSCHE NATIONAL BIBLIOTHEK

KB
Koninklijke Bibliotheek
Nationale bibliotheek van Nederland

THOMSON REUTERS®

SPRINGER NATURE

GeoNames

dblp
computer science bibliography

LEI

Microsoft Academic

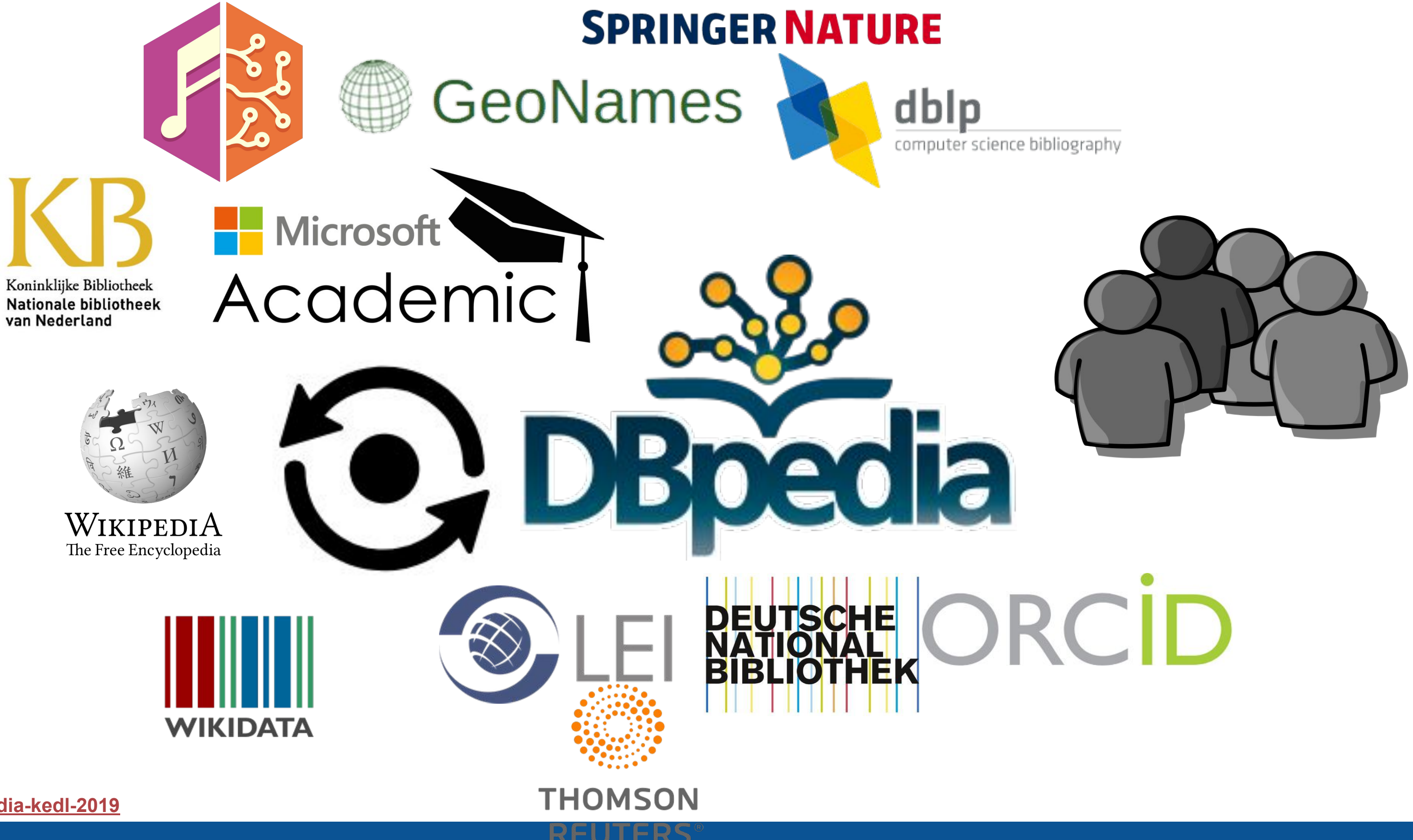
ORCID

WIKIPEDIA
The Free Encyclopedia

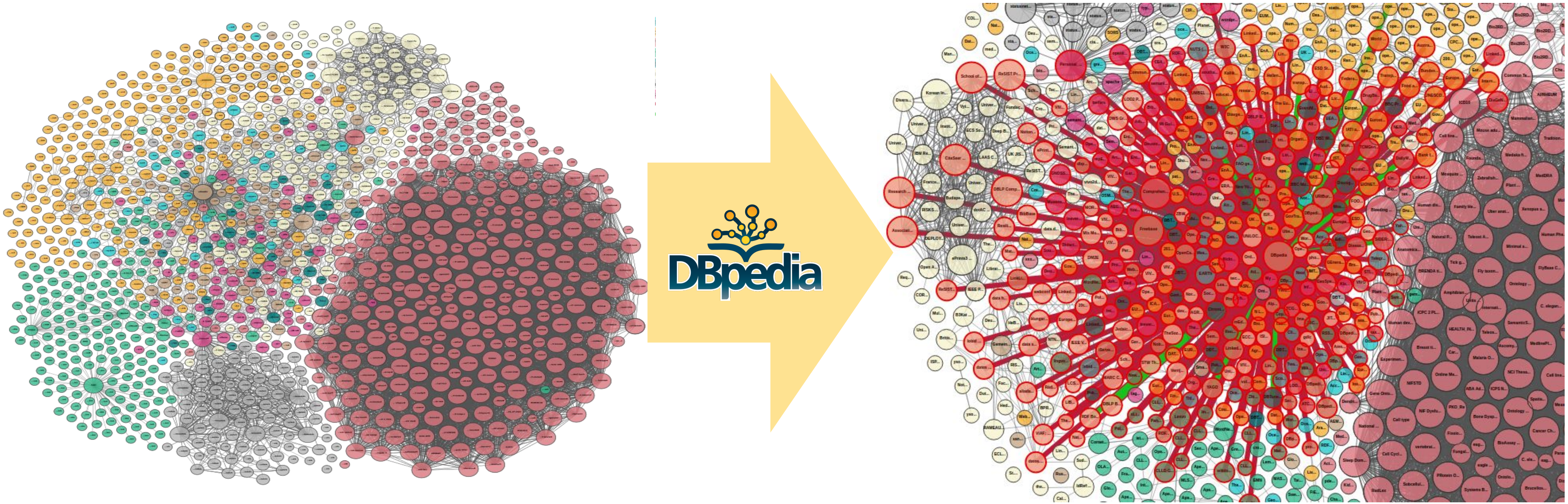
WIKIDATA

- Overlapping dataspaces
- Open licenses (compatible)
- Each organisation/project pushes against their own glass ceiling
- Wikipedia/Wikidata create yet another glass ceiling, since they aggregate from above sources

Rider no. 2 & 3: Syncing upstream solves it



Rider no. 4: Network Disaster of Linking / Mapping



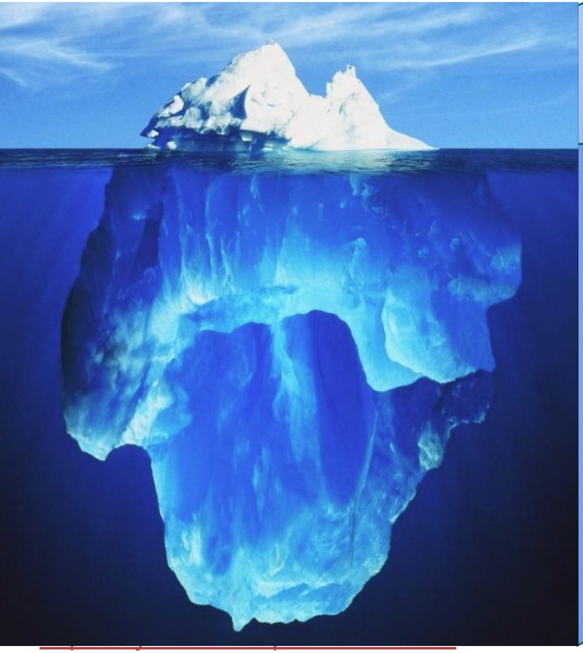
$O(n^2/2)$ -> O(n) with identifier linking paradigm, but actually
O (n) + Client-side created links + work for crawling (no standards)
Solved by Re-use via DBpedia Knowledge Library

https://tinyurl.com/dbpedia-kedl-2019

DBpedia

Less work, more benefits

“Open Data” is too altruistic
(for others)



Benefits aka Incentive Models

- Effective Third-Party contributions
- Open Data business models
- Attention / attribution

Less work by innovation

- Reusability / Deduplication of effort
- AI assistance (human in the loop)
- Off-the-shelf apps
- Power tools

Source: http://artpictures.club/shans-february-5-13.html

DBpedia Strategy Overview

Starting point

DBpedia is the most successful open knowledge graph (OKG), established 2008

take DBpedia to a global level

Global DBpedia Platform

- Communication & collaboration
- Share efforts and results
- Maximise societal value

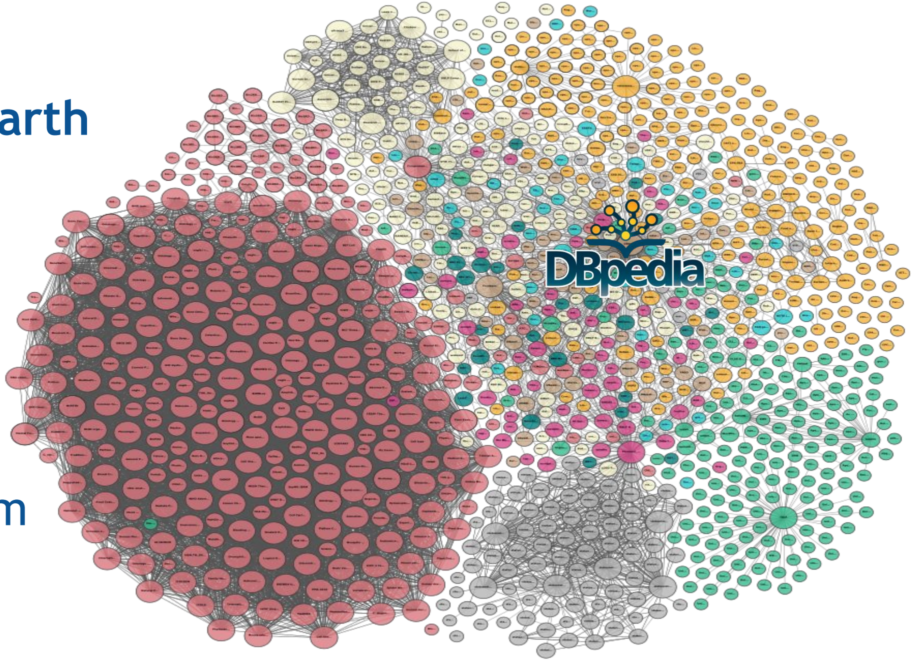
Medium term goals

- 200 orgs share value via platform
- 10% of public IT projects curate data
- 1 million user, high contribution rate
- thousands of new businesses and initiatives around the platform

DBpedia Strategy Overview

DBpedia bootstrapped Linked Open Data

LOD is the largest knowledge graph on earth



Migration from P2P to an efficient platform

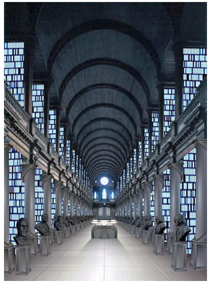
- lower entry barriers
- improved discovery and cooperation

DBpedia LOD Custody

```
######
#     #   ##   #####   ##   #####  #    #  ####
#     #  #  #    #    #  #  #    # #    # #
#     # #    #   #   #    # #####  #    #  ####
#     # ######   #   ###### #    # #    #      #
#     # #    #   #   #    # #    # #    # #    #
######  #    #   #   #    # #####   ####   ####
```

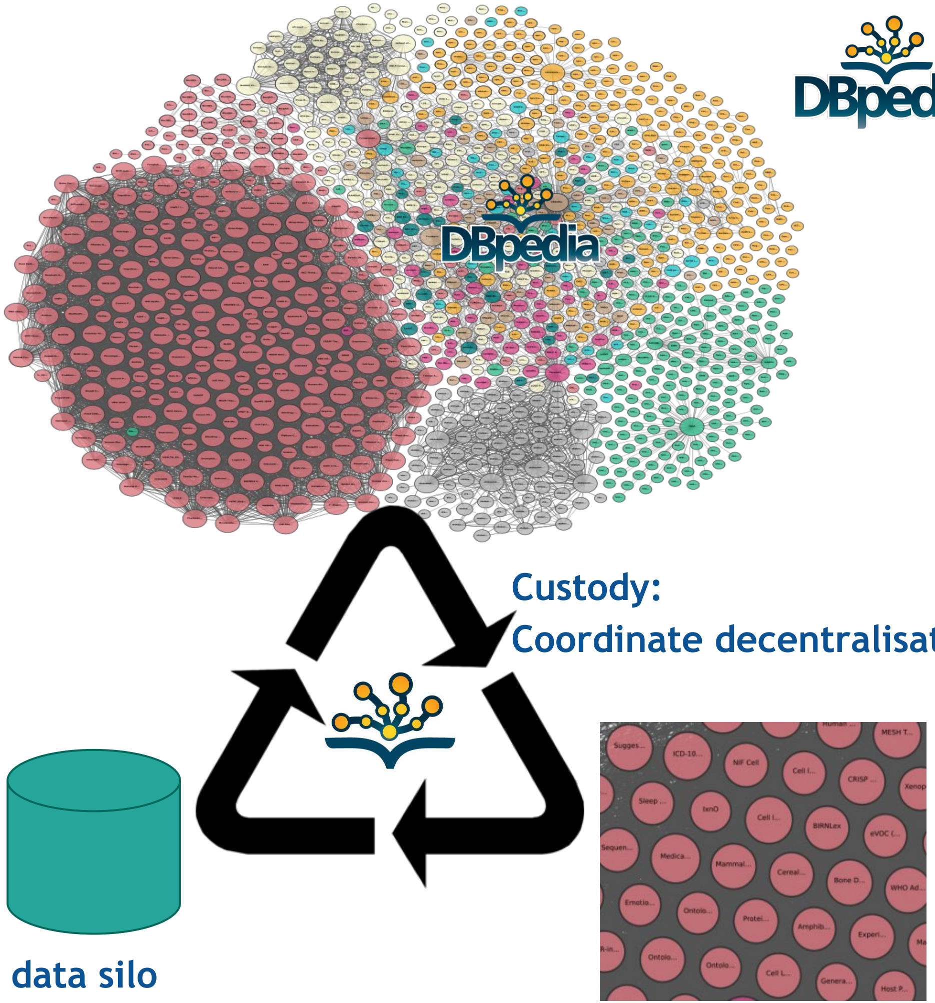
Open Decentral Platform

- file access and processing in network
- FAIR principles and W3C standards (DCAT)
- agile data engineering (Gitflow)
- high degree of automation (Maven)



Knowledge Library

- Semantic layer on data
- Semantic interoperability
- AI assistance



Custody:
Coordinate decentralisation

data silo

```
#######
#      #   ##     ######   ##    ######  #     #  ####
#      #  #   #     #     #   #  #     # #     # #
#      # #    #    #    #     # ######  #     #  ####
#      # #######    #    ####### #     # #     #      #
#      # #     #    #    #     # #     # #     # #    #
#######  #     #    #    #     # ######    ####   ####
```

Digital Factory Platform

https://databus.dbpedia.org/

https://databus.dbpedia.org/repo/sparql

https://databus.dbpedia.org/yasgui

Inspired by

MVNREPOSITORY

STEAM®

GitHub

Databus - Digital Factory Platform

Registry of files on the Web

- Virtual file warehouse
- Decentralised file storage
- Storage is cheap
- Downloadable
- File format doesn’t matter
  - RDF
  - CSV, XML
  - PDF or PDF collection

Databus - Digital Factory Platform

… but very strict metadata

- Provenance (who? - you!)
- Machine-readable licenses
- Private key signature, X509 (Trust)
- Dataset identity
- Versioning

Rejected

ZERO TO ONE PETER THIEL
EGO IS THE ENEMY RYAN HOLIDAY
RYAN HOLIDAY THE OBSTACLE IS THE WAY The Timeless Art of Turning Trials into Triumph
EXPONENTIAL ORGANIZATIONS ISMAIL, MALONE & VAN GEEST
Competing Against Luck
Value Proposition Design
THE STARTUP OWNER'S MANUAL Steve Blank Bob Dorf
The Corporate Startup Tendayi Viki Dan Toma Esther Gons

Approved

Minimal metadata core

Everybody can add more metadata systematically (no user tagging)

Databus - Digital Factory Platform

Time versioned: 2018.04.10

Build automation tool based on Maven

- Dataset Identity (ArtifactId)
  - Variance in content/format/compression
- Optimized for re-releasing the same files
  - ~ 3 days to learn and setup the tool (once)
  - 10 minutes to publish an update

https://github.com/dbpedia/databus-maven-plugin

Databus Demo

Download all data on the bus

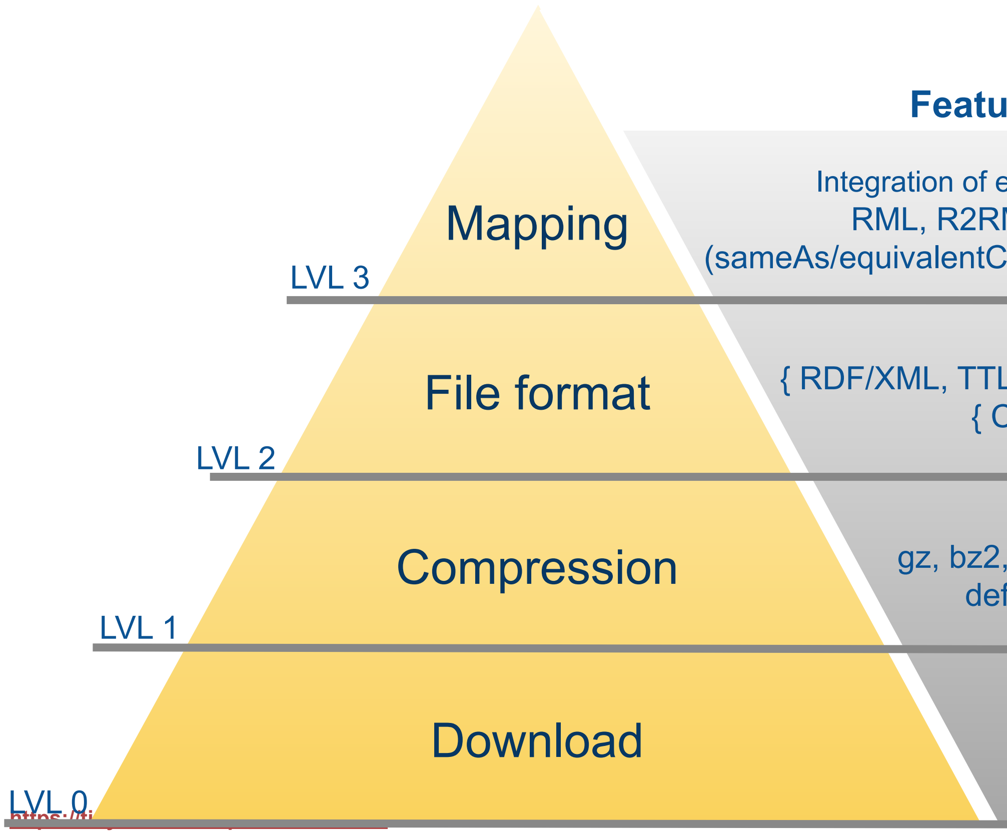
Stable Ids for collections with fixed or dynamic versions:

https://databus.dbpedia.org/dbpedia/collections/pre-release-2019-08-30

Dynamic versions:

- Latest
- Passed test suite x
- Most popular / other criteria

Databus-Client (alpha)

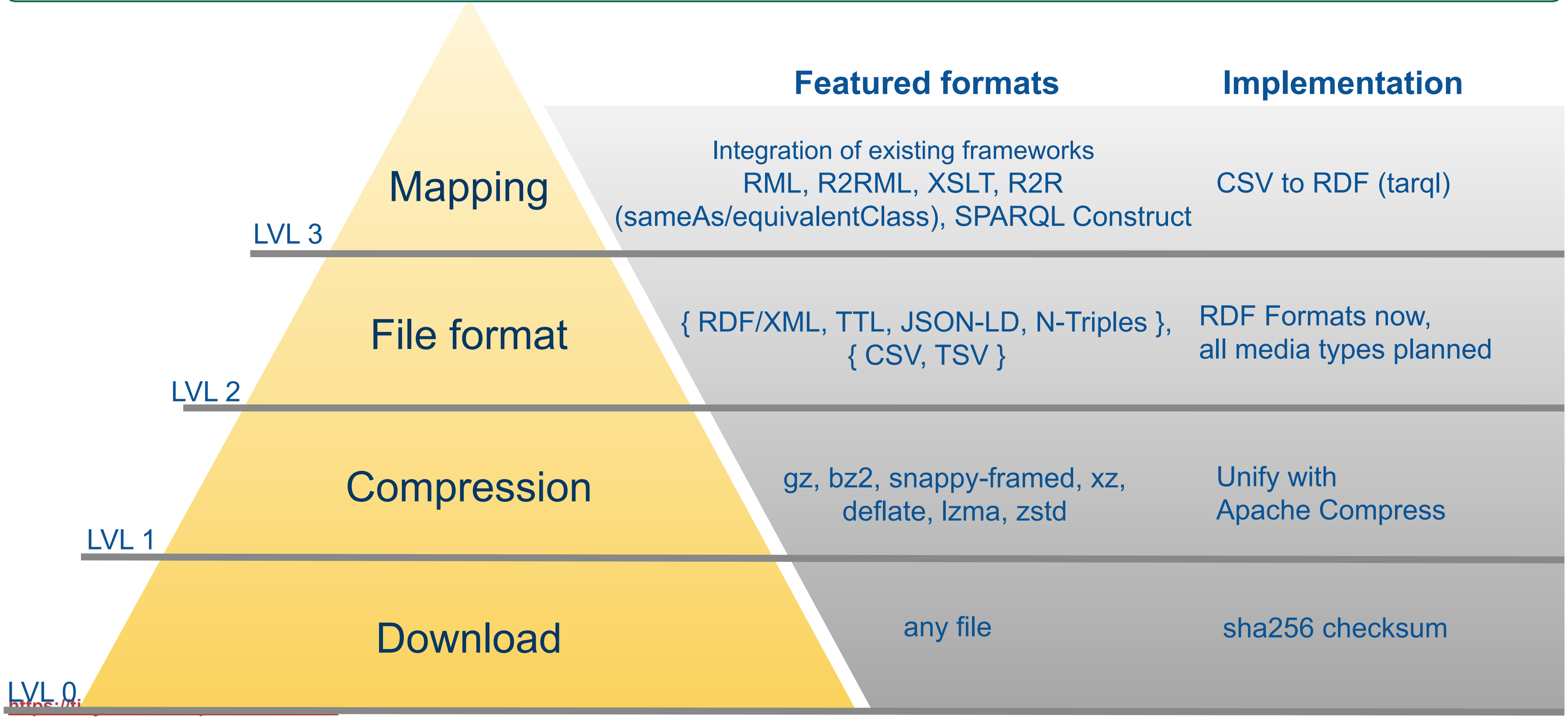


		Featured formats	Implementation
LVL 3	Mapping	Integration of existing frameworks RML, R2RML, XSLT, R2R (sameAs/equivalentClass), SPARQL Construct	CSV to RDF (tarql)
LVL 2	File format	{ RDF/XML, TTL, JSON-LD, N-Triples }, { CSV, TSV }	RDF Formats now, all media types planned
LVL 1	Compression	gz, bz2, snappy-framed, xz, deflate, lzma, zstd	Unify with Apache Compress
LVL 0	Download	any file	sha256 checksum

https://ti

Databus-Client (alpha)

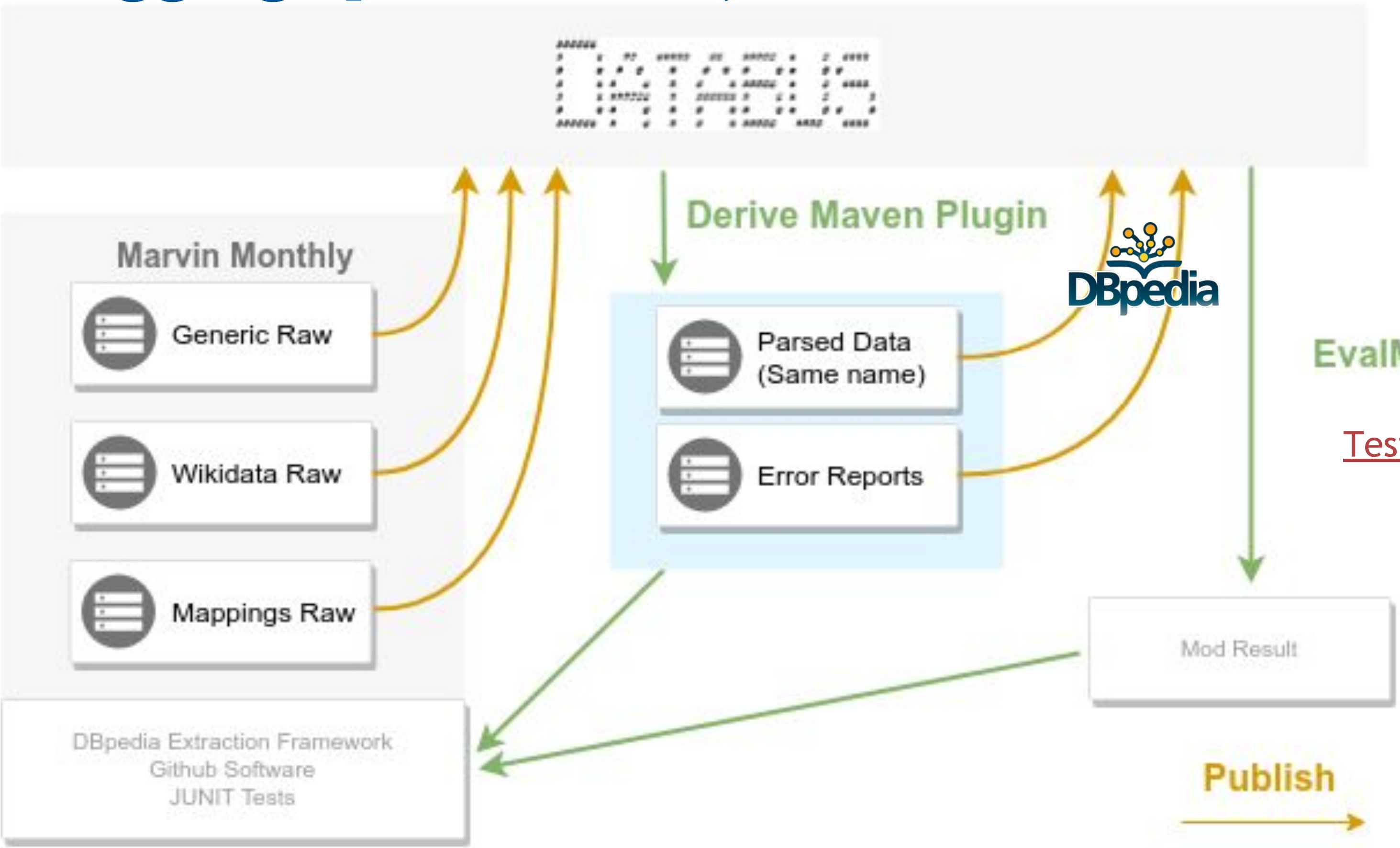
Application Replication and Deployment Layer (load data into software or databases automatically)



		Featured formats	Implementation
LVL 3	Mapping	Integration of existing frameworks RML, R2RML, XSLT, R2R (sameAs/equivalentClass), SPARQL Construct	CSV to RDF (tarql)
LVL 2	File format	{ RDF/XML, TTL, JSON-LD, N-Triples }, { CSV, TSV }	RDF Formats now, all media types planned
LVL 1	Compression	gz, bz2, snappy-framed, xz, deflate, lzma, zstd	Unify with Apache Compress
LVL 0	Download	any file	sha256 checksum

https://ti

Debugging - publish first, then test



Marvin Monthly

Generic Raw

Wikidata Raw

Mappings Raw

Derive Maven Plugin

Parsed Data (Same name)

Error Reports

EvalMod

Mod Result

DBpedia Extraction Framework
Github Software
JUNIT Tests

Publish

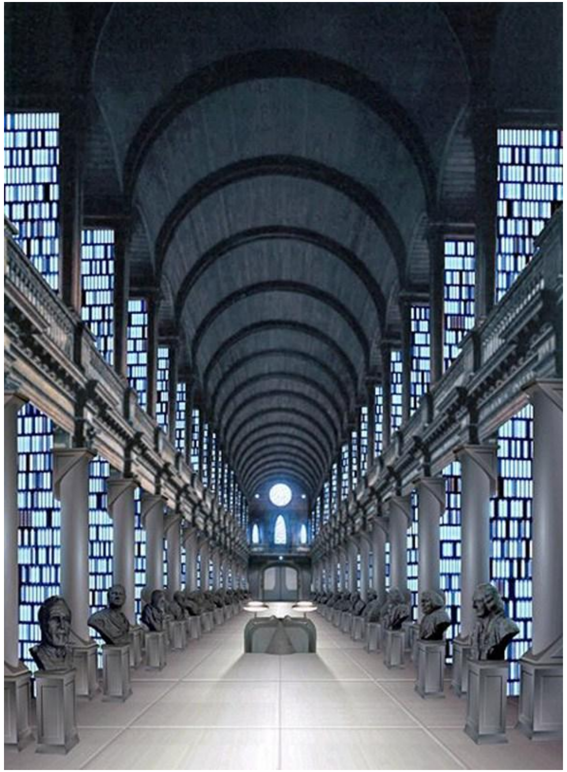
Test results

https://tinyurl.com/dbpedia-kedl-2019

DBpedia Knowledge Library

Mission is to improve external data

- Caches the core fraction of digital data (master records)
  - Aggregated from external data
  - Centrally indexed, linked and structured
- Curative focus
  - Ontologies → yes, **all** ontologies
  - Mappings → manage connection between all schema
  - Links → global view on entities
- AI assistance
  - Guides effective curation (Symbolic ILP & active learning)
  - Mitigates the gap between stability and evolution
  - Powerful discovery

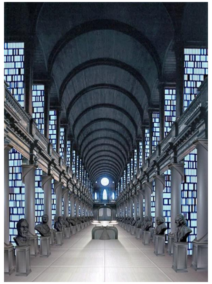


Disambiguation of Master Records

Tangible Entities (clear identification):	**Elusive Entities (vague Identification)**
Persons, Video games, Power Plants, Bibliographic References	Books, Songs, Products, Events
Tangible Facts:	**Elusive Facts:**
Height of Basketball Players	Leipzig has 600,000 citizens
Barack Obama was the 44th president of the United States from 2009 to 2017	Trump is the current US president
Excellent cost/benefit ratio Useful for 90% of use cases	Burn thousands of hours of discussion with minimal results

OntoGrate Methodology will produce global and ultimate lists for tangible entities and facts

Use Cases - Flexi Fusion



Third Party effort by

DBpedia Data Wranglers & Users

Frey et al. (ISWC 2019):
DBpedia FlexiFusion the Best of Wikipedia > Wikidata > Your Data

PreFusion	Fusion	Source Enrichment	Export
Aggregation, incl. Provenance Comparison of Data	“Best of” Like DBpedia KG, just bigger & better	Improve sources (sync) Lower curation effort	Custom fusion for use cases and applications Applications

https://tinyurl.com/dbpedia-kedl-2019

Demo of Fusion Result

https://databus.dbpedia.org/vehnem/flexifusion/fusion/2019.11.15

3.8 Million birthdates

Three targets at the moment:

- Ultimate lists of European authors (French, German, Dutch, Swiss library data)
- Power plants and energy sector
- Company data

Scalable global id management will allow growth into the second largest open knowledge graph

Summary and outlook

- 4 riders of datacalypse
  - rethink data processes
- Databus as a technical platform to version, access and process files in an automated manner
  - own file server required
- Knowledge library as a central tool of the integration platform
  - Third-party consumers
- FlexiFusion - PreFusion
  - Synchronisation and comparison of tangible data records
- FlexiFusion - Fusion
  - Second largest open knowledge graph

DBpedia

Next steps

We are looking for two female (or other) PhD students to help us build AI-Assistance

We are looking for strong partners to help us set up national data infrastructure that sync with the global DBpedia infrastructure

Stay informed via http://forum.dbpedia.org and http://blog.dbpedia.org

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

UNIWERSYTET EKONOMICZNY W POZNANIU 1926

UNIVERSITÄT MANNHEIM

eccenca
command your data!

SEMANTIC WEB COMPANY
linking data to knowledge

OPEN KNOWLEDGE INTERNATIONAL

Fraunhofer
IAIS

TIB LEIBNIZ-INFORMATIONSZENTRUM TECHNIK UND NATURWISSENSCHAFTEN UNIVERSITÄTSBIBLIOTHEK

OPENLINK® SOFTWARE

diffbot

NETWORK INSTITUTE

huygens ing

KB
Koninklijke Bibliotheek
Nationale bibliotheek van Nederland

wpsemantix
The SQL Knowledge Graph Company

ontotext

imec

InfAI®
Institute for Applied Informatics

SPRINGER NATURE

SOUND AND VISION

Linked Open Data Initiative
LinkedOpenData.jp ★★★★★

-kedl-2019