

A Library Data Management Platform Based on Linked Open Data



25 November, 2014
Jens Mittelbach | Robert Glaß



Europa fördert Sachsen.



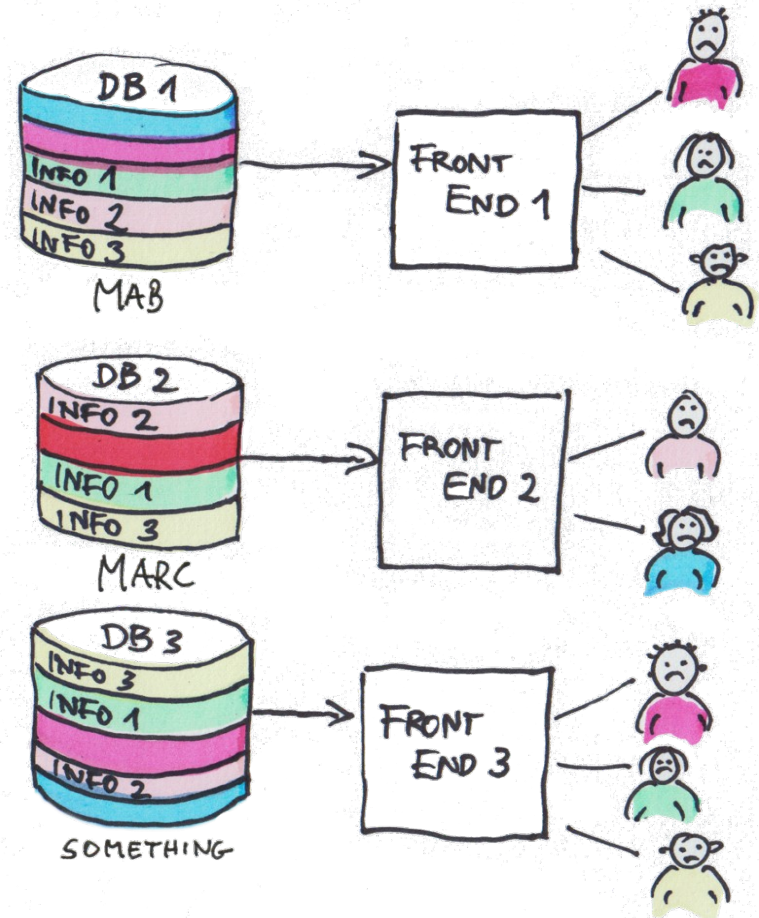
D:SWARM

A Library Data Management Platform Based on Linked Open Data

- ❖ Back in Those Days
- ❖ The Age of Discovery
- ❖ Library Data Management
- ❖ Qualify, Link and Free Your Data: D:SWARM
- ❖ Live Demo

Back in Those Days ... Data Heterogeneity

- ❖ Multiple individual data silos
 - ILS, document repositories, databases, ...
- ❖ Data saved in heterogeneous formats
 - MAB, MARC21, ...
- ❖ Each data silo gets processed individually
 - Multiple admin interfaces
 - Multiple search interfaces
 - Data unrelated to one another
- ❖ Comprehensive view of resources

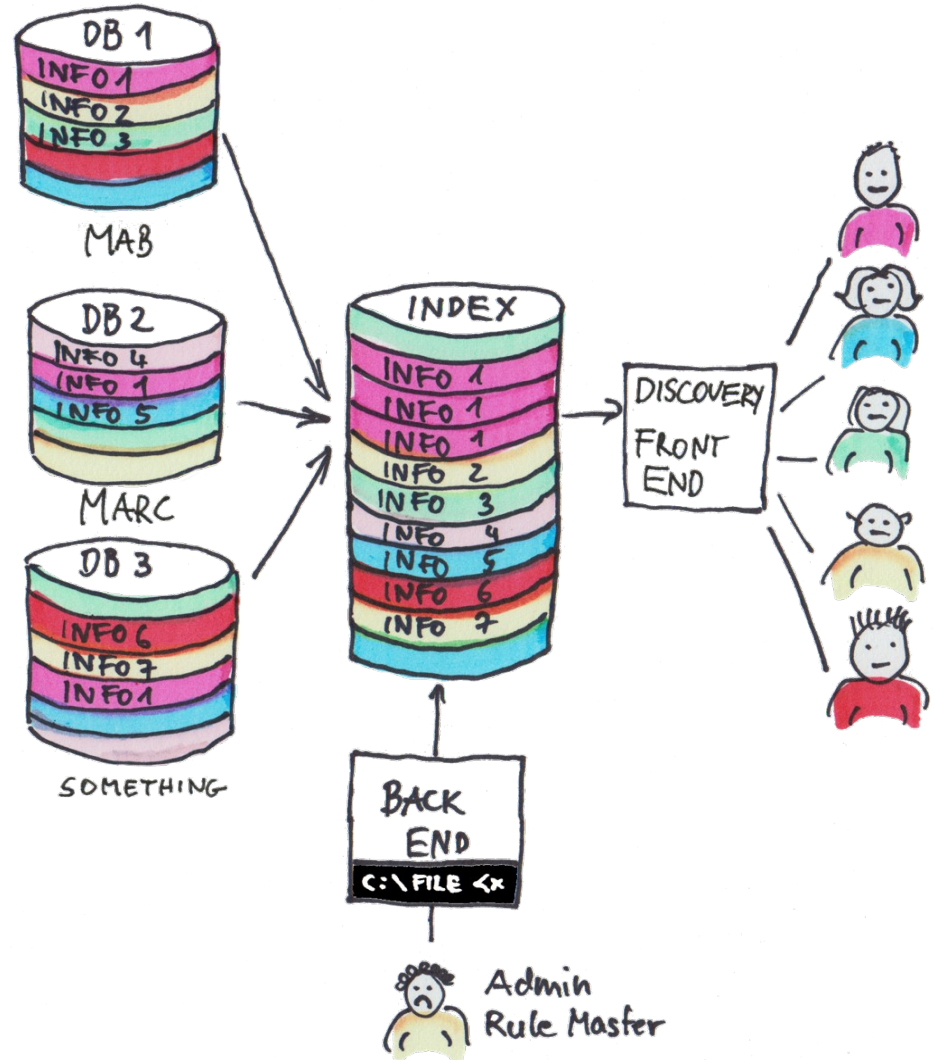


almost impossible (for users *and*



The Age of “Discovery” Data Normalization

- ❖ More comprehensive view of resources for users, but no real discovery/exploration
- ❖ Data gets normalized into one storage but not integrated
- ❖ Data available in record-oriented structures
 - External data (e.g. GND) has to be squeezed in the record
 - Metadata records are independent of each other
 - No explicit semantic quality of data



Library Data Management

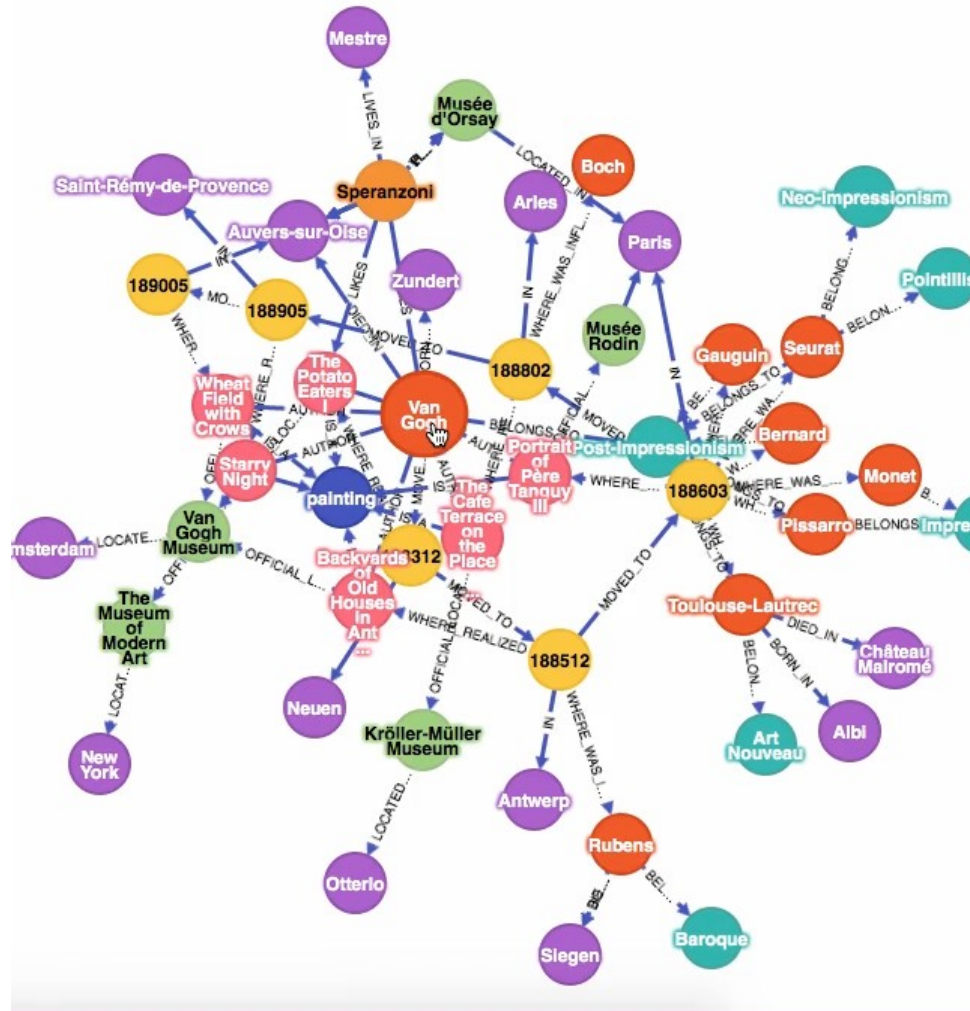
What Libraries Actually Need

★★★★★ Library Data

- ❖ Get rid of data silos
 - Open formats for exchange
- ❖ Lossless data integration instead of reductive normalization
- ❖ Data integration with entity level granularity
 - Get rid of pre-compiled data records
- ❖ Focus on linking entities/objects:
 - Graph structures creating the knowledge graph
- ❖ Stick to quality policy of libraries
 - Versioning and provenance of data



What Should Library Data Actually Look Like?



Library Data Management

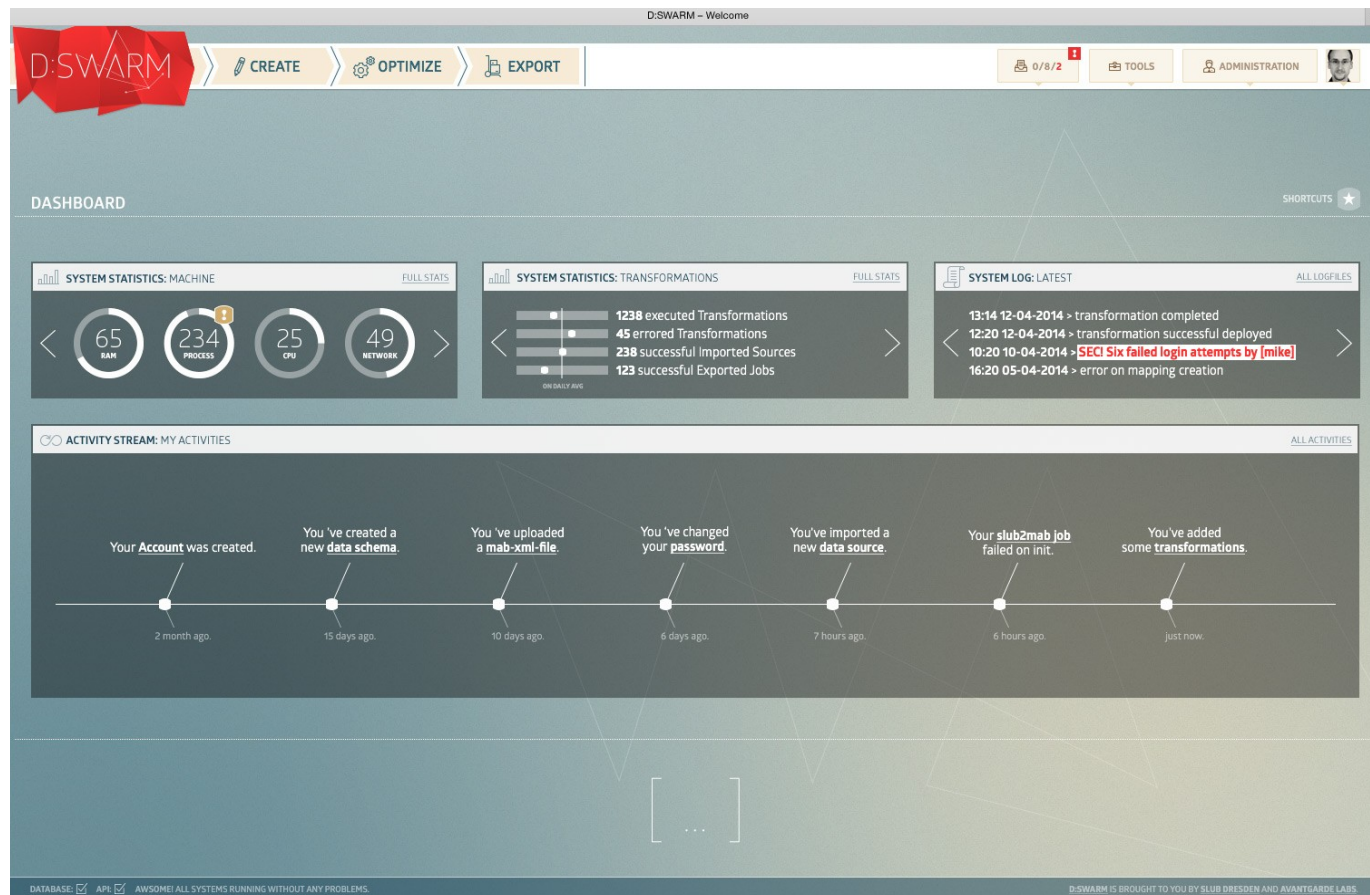
Whose Job Is Library Data Integration?

- ❖ Data integration should be done by **domain experts**
 - Librarians, not IT staff (IT always understaffed)
 - Programming skills should not be a requirement
 - Good user experience is a prerequisite for adoption
- ❖ Example driven modelling approach
- ❖ Value created in the community should be reusable

Library Data Management

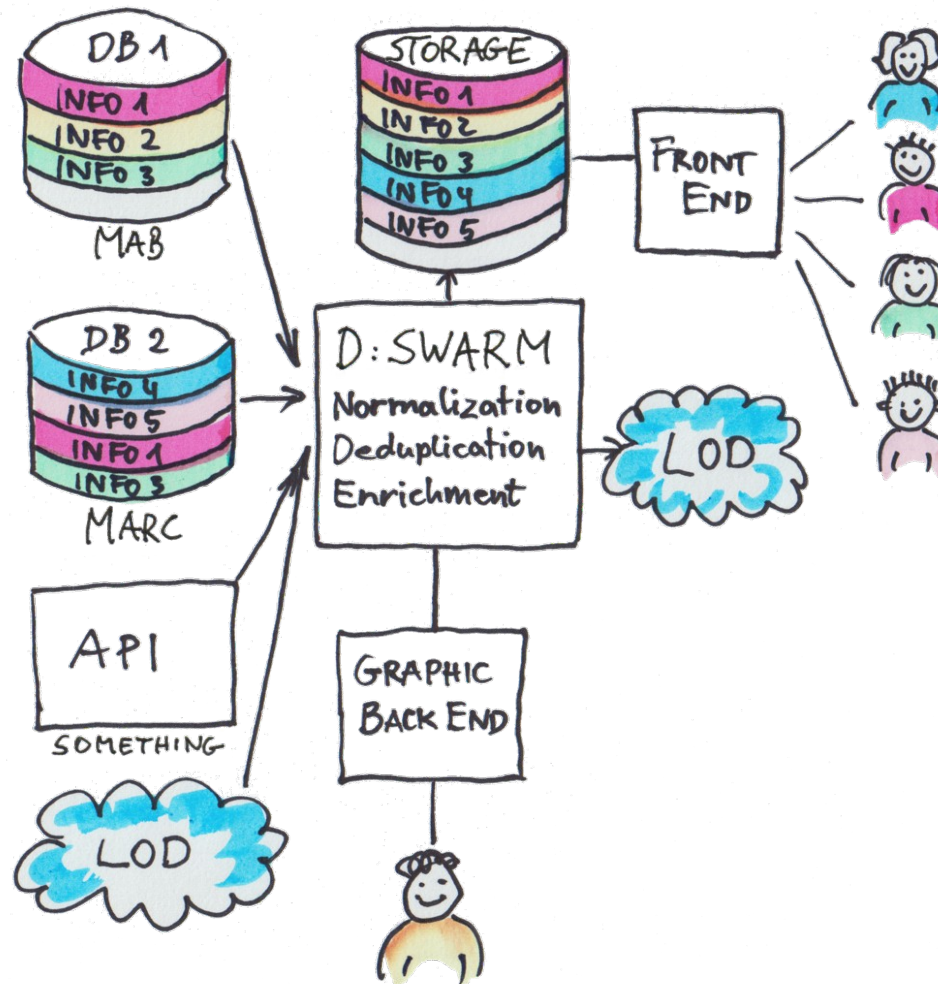
What Tools Do We Need?

Our Approach: An Open Source Data Management Platform



Library Data Management

How Can Data Integration Be Done?



Qualify, Link and Free Your Data: D:SWARM

Who's behind this Project?

- ❖ Collaborative development team of SLUB Dresden and Avantgarde Labs GmbH
- ❖ Started work in June 2013
- ❖ Funded from the European Regional Development Fund (ERDF)



Qualify, Link and Free Your Data: D:SWARM

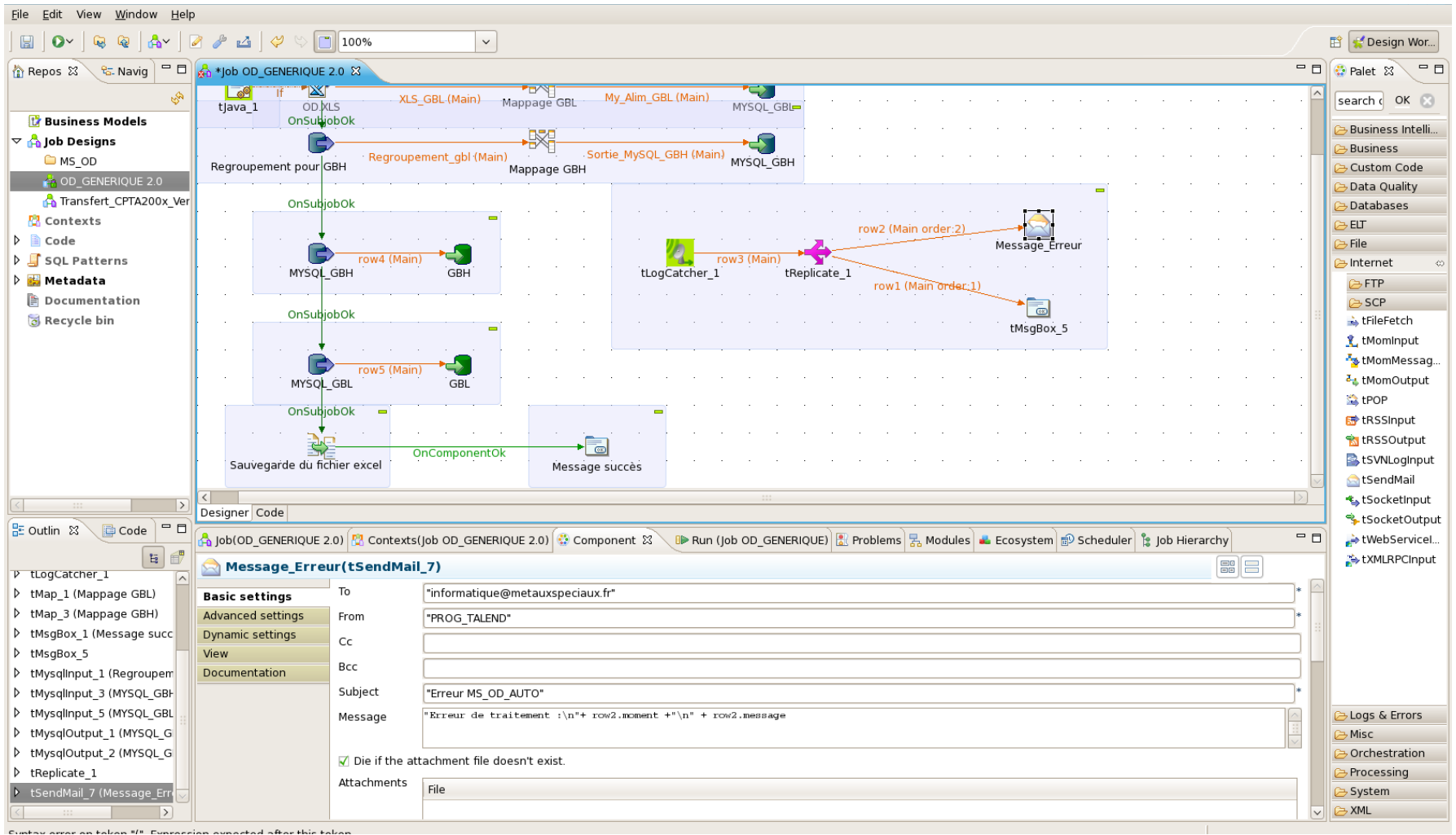
Our Challenge: Existing Data Formats: MAB, MARC

- „selection of keywords“
 - Relevant MAB fields are 902x, 907x, 912x, 917x, 922x.
 - These fields have subfields a, b, c, ... coded with further information (type of keyword, person, time, place, concept...)
 - From field 902x to field 922x we have to check
 - If in subfield "a" there is one of these strings (800|801|820|830|845|850|860|870|880)?
 - If so, is there one of these strings (c|g|k|p|s|t|z) in subfield "b"?
 - If so, the value in subfield "c" qualifies as a keyword
 - Keyword needs to be trimmed (which is the easiest part)

```
403 <datafield tag="902" ind1="x" ind2="">
404   <subfield code="_">814422</subfield>
405   <subfield code="a">040</subfield>
406   <subfield code="b">b</subfield>
407   <subfield code="c">616.07545</subfield>
408 </datafield>
409 <datafield tag="902" ind1="x" ind2="">
410   <subfield code="_">814422</subfield>
411   <subfield code="a">068</subfield>
412   <subfield code="b">c</subfield>
413   <subfield code="c">s</subfield>
414 </datafield>
415 <datafield tag="902" ind1="x" ind2="">
416   <subfield code="_">814422</subfield>
417   <subfield code="a">080</subfield>
418   <subfield code="b"></subfield>
419   <subfield code="c">SWD</subfield>
420 </datafield>
421 <datafield tag="902" ind1="x" ind2="">
422   <subfield code="_">814422</subfield>
423   <subfield code="a">800</subfield>
424   <subfield code="b">s</subfield>
425   <subfield code="c">Fiberendoskopie</subfield>
426 </datafield>
427 <datafield tag="902" ind1="x" ind2="">
428   <subfield code="_">814422</subfield>
429   <subfield code="a">808</subfield>
430   <subfield code="b">a</subfield>
431   <subfield code="c">Zetkin unter Fiberbronchoskopie
432 </datafield>
433 <datafield tag="902" ind1="x" ind2="">
434   <subfield code="_">814422</subfield>
435   <subfield code="a">850</subfield>
436   <subfield code="b">s</subfield>
437   <subfield code="c">|Endoskopie</subfield>
438 </datafield>
```

Qualify, Link and Free Your Data: D:SWARM

Our Challenge: Existing Tools: Talend



Qualify, Link and Free Your Data: D:SWARM

Our Challenge: Existing Tools: Open Refine

google refine sparqlproxy.php Permalink

Open... Export... Hi

icet / Filter Undo / Redo 0

11405 matching rows (25193 total)

Extensions: Freebase

fresh Reset All Remove All

Show as: rows records Show: 5 10 25 50 rows

« first < previous 1 - 50 next > la

status change invert reset

choices Sort by: name count Cluster

sed 1205

en 11405 exclude

en, but proposed to close 215

posed to open 59

it by choice counts

type change invert reset

choices Sort by: name count Cluster

idemies 57

mmunity School 11405 exclude

mmunity Special School 145

Setting 26

ndation School 805

ndation Special School 2

cap change reset

0.00 — 1,200.00

ript: {}

11405 matching rows (25193 total)

▼ All ▼ school ▼ label ▼ status ▼ type

48. http://education.data.gov.uk/id/school/101185

49. http://education.data.gov.uk/id/school/101187

50. http://education.data.gov.uk/id/school/101188

51. http://education.data.gov.uk/id/school/101189

53. http://education.data.gov.uk/id/school/101192

54. http://education.data.gov.uk/id/school/101193

55. http://education.data.gov.uk/id/school/101196

56. http://education.data.gov.uk/id/school/101197 Northbury Infants' School

57. http://education.data.gov.uk/id/school/101198 Ripple Junior School

59. http://education.data.gov.uk/id/school/101202 Beam Primary School

60. http://education.data.gov.uk/id/school/101206 Grafton Junior School

61. http://education.data.gov.uk/id/school/101207 Grafton Infants' School

62. http://education.data.gov.uk/id/school/101211 Marsh Green Primary School

63. http://education.data.gov.uk/id/school/101212 Rush Green Junior School

64. http://education.data.gov.uk/id/school/101213 Rush Green Infants' School

Facet Text facet

Text filter Numeric facet

Edit cells Timeline facet

Edit column Scatterplot facet

Transpose Custom text facet...

Sort... Custom numeric facet...

View Customized facets

Reconcile

Qualify, Link and Free Your Data: D:SWARM

What Is D:SWARM?

- ❖ Graphical web based **ETL modelling tool** that serves to:
 - import data from heterogeneous sources with different formats
 - map input to output schemata and design transformation workflows
 - load transformed data into property graph database
- ❖ With additional functionalities:
 - Exporting of data models as RDF
 - Sharing mappings and transformation workflows

Qualify, Link and Free Your Data: D:SWARM

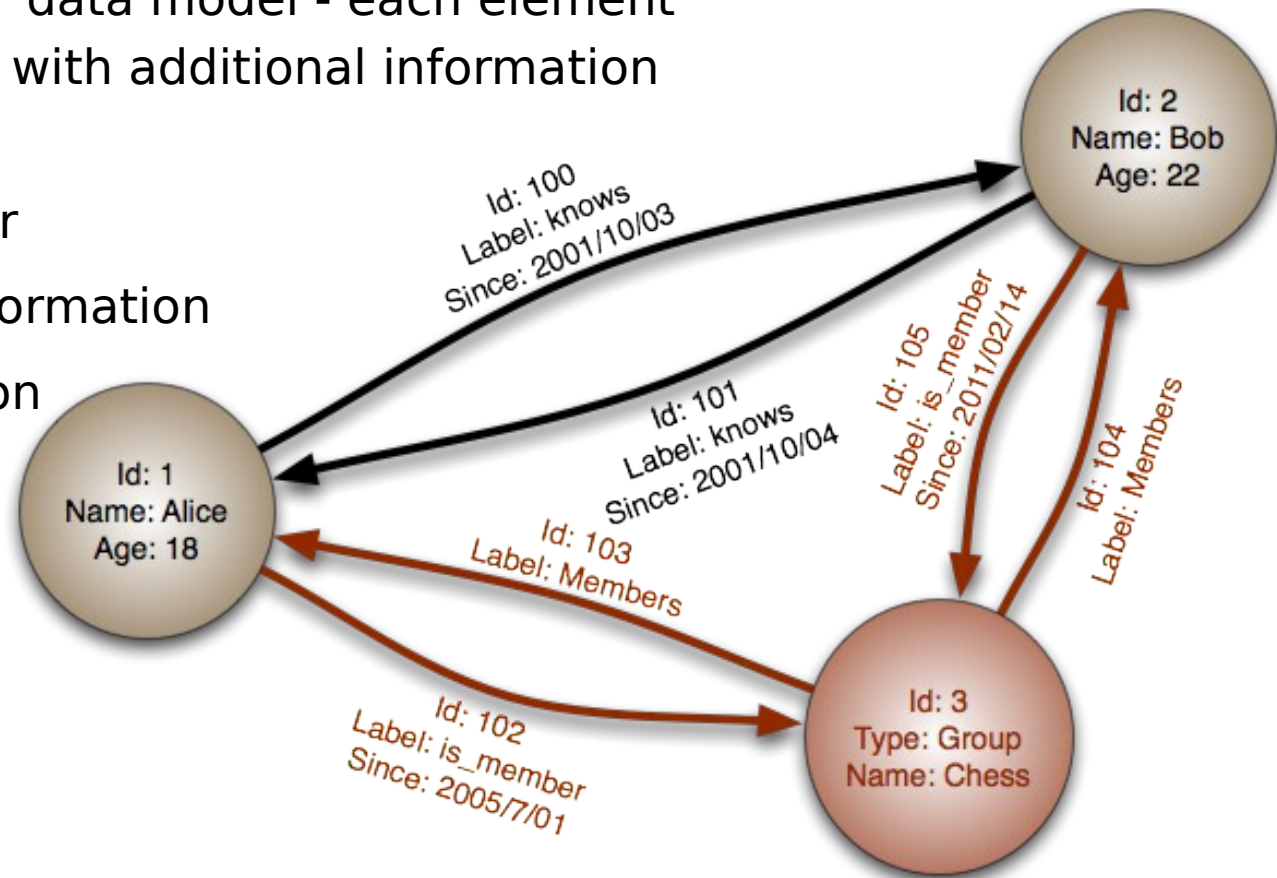
How Does D:SWARM Work?

- ❖ Modelling GUI and job repository
- ❖ Execution environment
 - Operational data from heterogeneous data sources (ILS, OAI-PMH, CSV ...) get processed according to the transformation logics defined in modelling GUI
- ❖ Admin centre
 - Scheduling & execution planning
 - Monitoring of system (data ingest, processing, errors)

Qualify, Link and Free Your Data: D:SWARM

Why a Property Graph?

- ❖ Node (S) – Edge (P) – Node (O)
- ❖ Extension of RDF data model - each element can be endowed with additional information (key : value)
 - Version number
 - Provenance information
 - Type information



Qualify, Link and Free Your Data: D:SWARM Intermediate Results as of November 2014

- ❖ Modelling GUI in 2nd version
 - Available file importer: XML, CSV, MABXML
 - Simple schema editor & graphic schema mapper
 - Transformation workflow designer & filter (Metafactory)
- ❖ Execution of mappings and transformations in modelling GUI
- ❖ Persistence in graph database (Neo4J)
- ❖ Exporter: Turtle, N-Quads, N3, ...

- ❖ Publication under Open Source licence (Apache 2):
<https://github.com/dswarm>

Qualify, Link and Free Your Data: D:SWARM Live Demo

<http://demo.dswarm.org>

Qualify, Link and Free Your Data: D:SWARM

Our Next Steps

- ❖ Provision of URI templates for resource matching and linking
- ❖ Scalable execution engine for production mode
- ❖ Extension of transformation function set
- ❖ Extension of importers
- ❖ Implementation of an administration centre
- ❖ Deduplication and FRBRization
- ❖ Integration of SLUBsemantics Enrichment Service
- ❖ Implementation of sharing features

Qualify, Link and Free Your Data: D:SWARM

Your Next Steps

- ❖ Follow us on twitter.com/dswarm or www.dswarm.org or github.com/dswarm
- ❖ Try it out and get in contact with us
 - <http://demo.dswarm.org>
 - <https://github.com/dswarm/dswarm-documentation/wiki>
 - team@dswarm.org
- ❖ Help us prioritize our backlog
 - <https://jira.slub-dresden.de/>
- ❖ Fork us on github.com/dswarm