DBpedia FlexiFusion: Best of Wikipedia and Wikidata

Johannes Frey, Marvin Hofer, Daniel Obraczka, Jens Lehmann and Sebastian Hellmann

Knowledge Integration and Linked Data Technologies (AKSW/KILT) Center @ Institute for Applied Informatics (InfAI), Leipzig Germany

DBpedia Project
Agenda

1. Motivation & Use Cases
2. DBpedia FlexiFusion
   a. DBpedia Identity Management
   b. Workflow Details
3. Evaluation of FlexiFusion
4. Related Work
5. Future Work
DBpedia Problems, Data Fusion Motivation & Use Cases

Current Wiki/DBpedia Problems:

- **Infoboxes** significantly vary between Wikipedia versions w.r.t. quality, comprehensiveness and up-to-dateness
- **Mappings** coverage, completeness and correctness varies between languages (chapters)
- **Extractor(s)** accuracy varies between languages

2 use cases to tackle this and to simplify querying & data usage of DBpedia

**Fused DBpedia** Dataset:

- Merge 40 separate DBpedia chapters into one knowledge graph
  - Increase coverage of resources/entities
  - Fuse properties and values from different sources to increase coverage of properties and increase data quality via resolution of contradictuous data
- Integrate additional data from other sources (e.g. Wikidata)
Data Fusion Use Case #2

**Enriched DBpedia Chapter Datasets:**

- Enrich each chapter with Fused DBpedia
  - No additional entities are added
  - Only properties and values from existing entities will be enriched
  - Existing values from the chapter are prioritized (overwrite protection) for functional properties
Why FlexiFusion

Variety of fusion strategies exists and appropriate settings are highly dependent on use case

- Datasource selection
- Conflict resolution
- Individual property and entity selection/reduction for domain ("slice and dice")

Idea / Goal of FlexiFusion: make it easier to mass produce custom KGs / DBpedias

- **PreFusion Dataset:** Pre-compute a merged global view of all triples from the sources with statement level provenance
- **FlexiFusion Workflow:** Customized KG’s can be derived with 2 flexible methods of the workflow being applied on PreFusion data
DBpedia FlexiFusion Workflow

- DBpedia Databus
  - Input Data Groups
    - DBpedia
    - Wikidata
    - ... Your Data
  - DBpedia FlexiFusion
    - 1. ID rewriting
    - 2. Prefuse
    - 3. Reduce & resolve
  - Mapping Management
    - Future Work
      - DBpedia Global ID Management
        - collect sameAsLinks and IRIs
        - compute clustering
        - assign Global IDs
  - DBpedia Databus
    - FlexiFusion Groups
      - Global IDs
      - Prefusion
      - Enriched DBpedia
      - FusedDBpedia

Consume → Publish
Source Selection and Results Persistence

- Based on DBpedia Databus (DBpedia’s “Maven+Github” for data)
  - Register data by uploading structured metadata (DataID)
  - Data is organized using the concepts publisher / group / artifact / version / file
- Allows flexible input data selection via SPARQL query [http://databus.dbpedia.org](http://databus.dbpedia.org)
- Automatic release of results via mvn plugin
DBpedia Identity Management

DBpedia Databus
- Input Data Groups
  - DBpedia
  - Wikidata
  - ... Your Data

Mapping Management
- Future Work

DBpedia Global ID Management
- collect sameAsLinks and IRI
- compute clustering
- assign Global IDs

DBpedia FlexiFusion
1. ID rewriting
2. PreFuse
3. Reduce & resolve

DBpedia Databus
- FlexiFusion Groups
  - Global IDs
  - PreFuse
  - Enriched DBpedia
  - FusedDBpedia

Consume → Publish
FlexiFusion Workflow and Dataflow on the Databus

DBpedia Databus
- Input Data Groups
  - DBpedia
  - Wikidata
  - .... Your Data

Mapping Management
Future Work
- DBpedia Global ID Management
  - collect sameAsLinks and IRIs
  - compute clustering
  - assign Global IDs

DBpedia FlexiFusion
1. ID rewriting
2. PreFuse
3. Reduce & resolve

FlexiFusion Groups
- Global IDs
- PreFusion
- Enriched DBpedia
- FusedDBpedia
FlexiFusion Workflow: PreFusion Dataset Creation

1. ID Rewriting
Input from DBpedia Databus:
- Custom File selection with *normalized* triples (predicates and literals)
- Snapshot of Global ID Assignment (based on sameAs Clustering)
- Replace all IRIs representing local identifiers with their sameAs cluster IRI

![Diagram with DBpedia entities and their relationships]

- `http://fr.dbpedia.org/resource/Tour_Eiffel`
- `https://global.dbpedia.org/id/12HpzV`
- `https://global.dbpedia.org/id/53y2b [fr]`
FlexiFusion Workflow: PreFusion Dataset Creation

2. PreFuse

- Derive preFused entities by grouping all triples first by same subject and then by their predicate value.
DBpedia PreFusion Dataset Format

- JSON-LD based Dataset tracing down the origin of the triples from all input files
- compact representation of (alternative) values grouped by subject-predicate pair
- Unified access to all sources: rewritten global IDs, and properties mapped to DBpedia Ontology

```json
{
    "@id": "fc4ebb0fed3c3171578c299b3ce21f411202ff2afc93568a54b4db7a75",
    "subject": { "@id": "https://global.dbpedia.org/id/12HpzV" },
    "predicate": { "@id": "http://dbpedia.org/ontology/floorCount" },
    "objects": [
      { "object": {
        "@value": "4",
        "@type": "http://www.w3.org/2001/XMLSchema#positiveInteger",
        "source": [ {
          "@id": "d0:lang=fr.ttl.bz2",
          "iHash": "cbdcb"
        } ]
      }, {
        "object": {
          "@value": "3",
          "@type": "http://www.w3.org/2001/XMLSchema#positiveInteger",
          "source": [ {
            "@id": "d0:lang=en.ttl.bz2",
            "iHash": "1e7d4"
          }, {
            "@id": "d0:lang=es.ttl.bz2",
            "iHash": "eb41e"
          } ]
        }
      }
    ],
    "@context": "sources=dbpw_context.jsonld"
}
```
FlexiFusion Workflow: 3. Fuse (Reduce & Resolve)

DBpedia Databus
Input Data Groups
- DBpedia
- Wikidata
- .... Your Data

DBpedia FlexiFusion
1. ID rewriting
2. PreFuse
3. Reduce & resolve

DBpedia Global ID Management
- collect sameAsLinks and IRIs
- compute clustering
- assign Global IDs

Mapping Management
Future Work

DBpedia Databus
FlexiFusion Groups
- Global IDs
- PreFusion
- Enriched DBpedia
- FusedDBpedia

Consume → Publish
3. Fuse: Reduce

- Customizable function applied for every subject-predicate pair to reduce or filter the amount of entities and/or the amount of information for each entity
- **Purpose:**
  - primary: remove irrelevant or bad data
  - primary: reduce fusion decisions for resolve
  - secondary: apply transformation of the data (fine tuning)
- **Examples**
  - pass only dbo:birthplace and dbo:birthdate property
  - pass only entities available in the Catalan chapter
  - pass only entities of a type
  - remove values from a source to be ignored
  - remove untyped literals
  - ...

3. Fuse: Resolve

- Function which picks a number of objects from the list of each reduced subject-predicate pair which are handed over to final fused dataset.
- Purpose: resolve conflicts to improve data quality.
- Example:
  - Select all values for union-compatible properties
  - Pick one value via majority voting for functional properties
  - Return no value if too many options are available
  - Source preference
  - ...
FlexiFusion Configuration for Evaluation Scenarios

- **FusedDBpedia (use case 1) Configuration:**
  - reduced to: 6 sources, i.e. Wikidata, the English (EN), German (DE), French (FR), Dutch (NL) and Swedish (SV) chapter
  - resolved via: select 1 object value based on language preference (Wikidata, EN, DE, FR, NL, SV) iff $\text{PMOD}(p)=1$ else take all values.
  - $\text{PMOD}(p) =$ predicate median out degree for property $p$ measured over all input sources for entities having at least one value for it

- **Enriched Catalan (use case 2) Configuration:**
  - reduced to: $sp$-pairs where $s$ is a subject from the Catalan DBpedia data
  - resolved via: select all values iff $\text{PMOD}>1$ else Catalan value has preference (if exists), otherwise use preference list of FusedDBpedia
Evaluation: Fusion - Coverage

→ increased vocabulary usage
→ increased average number of distinct properties per entity

<table>
<thead>
<tr>
<th></th>
<th>Wikidata</th>
<th>English</th>
<th>German</th>
<th>French</th>
<th>Dutch</th>
<th>Swedish</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall triples</td>
<td><strong>436,808,402</strong></td>
<td>124,994,586</td>
<td>42,630,107</td>
<td>39,438,426</td>
<td>36,924,058</td>
<td>37,942,711</td>
<td><strong>558,597,215</strong></td>
</tr>
<tr>
<td>Entities</td>
<td><strong>45,649,373</strong></td>
<td>17,576,432</td>
<td>5,020,972</td>
<td>5,429,710</td>
<td>3,638,110</td>
<td>5,862,430</td>
<td><strong>66,822,365</strong></td>
</tr>
<tr>
<td>Distinct predicates</td>
<td>166</td>
<td><strong>1,412</strong></td>
<td>598</td>
<td>1,052</td>
<td>979</td>
<td>415</td>
<td><strong>2,292</strong></td>
</tr>
<tr>
<td>Distinct subject-predicates (sp-pairs)</td>
<td><strong>179,789,022</strong></td>
<td>77,368,237</td>
<td>26,086,747</td>
<td>26,049,036</td>
<td>24,339,480</td>
<td>29,062,921</td>
<td><strong>465,018,956</strong></td>
</tr>
<tr>
<td>AVG properties per entity</td>
<td>3.938</td>
<td>4.402</td>
<td>5.196</td>
<td>4.798</td>
<td><strong>6.690</strong></td>
<td>4.957</td>
<td><strong>6.959</strong></td>
</tr>
</tbody>
</table>
Evaluation: Fusion - Type Coverage

→ type knowledge gain from ~ 10 - 33 %
→ smaller chapters also contribute novel/unique type information to the same instances which are untyped

<table>
<thead>
<tr>
<th>Type</th>
<th>Wikidata</th>
<th>English</th>
<th>German</th>
<th>French</th>
<th>Dutch</th>
<th>Swedish</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbo:Person</td>
<td>4,197,564</td>
<td>1,757,100</td>
<td>627,353</td>
<td>491,304</td>
<td>188,025</td>
<td>62,814</td>
<td>4,612,463</td>
</tr>
<tr>
<td>only in Source</td>
<td>2,246,879</td>
<td>350,137</td>
<td>26,896</td>
<td>6,498</td>
<td>4,506</td>
<td>316</td>
<td>+9.88%</td>
</tr>
<tr>
<td>dbo:Company</td>
<td>188,107</td>
<td>70,208</td>
<td>25,208</td>
<td>14,889</td>
<td>4,446</td>
<td>3,291</td>
<td>209,433</td>
</tr>
<tr>
<td>only in Source</td>
<td>80,443</td>
<td>4,038</td>
<td>834</td>
<td>548</td>
<td>89</td>
<td>121</td>
<td>+11.34%</td>
</tr>
<tr>
<td>dbo:Location</td>
<td>3,952,788</td>
<td>839,987</td>
<td>406,979</td>
<td>276,096</td>
<td>449,750</td>
<td>1,480,627</td>
<td>5,293,969</td>
</tr>
<tr>
<td>only in Source</td>
<td>2,451,306</td>
<td>27,430</td>
<td>25,804</td>
<td>14,979</td>
<td>101,422</td>
<td>33,425</td>
<td>+33.93%</td>
</tr>
<tr>
<td>dbo:Animal</td>
<td>8,307</td>
<td>228,319</td>
<td>145</td>
<td>0</td>
<td>675,337</td>
<td>437</td>
<td>784,808</td>
</tr>
<tr>
<td>only in Source</td>
<td>2,963</td>
<td>2,302</td>
<td>1</td>
<td>0</td>
<td>2,029</td>
<td>5</td>
<td>+16.21%</td>
</tr>
</tbody>
</table>
## Evaluation: Fusion - Property Coverage

<table>
<thead>
<tr>
<th>Property</th>
<th>Wikidata</th>
<th>English</th>
<th>German</th>
<th>French</th>
<th>Dutch</th>
<th>Swedish</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbo:releaseDate</td>
<td>15,346,053</td>
<td>212,290</td>
<td>10,849</td>
<td>38,262</td>
<td>12,417</td>
<td>57</td>
<td>15,435,211</td>
</tr>
<tr>
<td>distinct entities</td>
<td><strong>15,320,967</strong></td>
<td>195,359</td>
<td>10,329</td>
<td>32,573</td>
<td>6,971</td>
<td>50</td>
<td>~</td>
</tr>
<tr>
<td>only in source</td>
<td>15,220,722</td>
<td>92,648</td>
<td>2,920</td>
<td>7,713</td>
<td>997</td>
<td>19</td>
<td>+0.75%</td>
</tr>
<tr>
<td>dbo:city</td>
<td>5,401,684</td>
<td>112,483</td>
<td>36,145</td>
<td>68,080</td>
<td>6,398</td>
<td>496</td>
<td>5,136,055</td>
</tr>
<tr>
<td>distinct entities</td>
<td><strong>5,069,535</strong></td>
<td>87,306</td>
<td>34,830</td>
<td>56,559</td>
<td>4,977</td>
<td>431</td>
<td>~</td>
</tr>
<tr>
<td>only in source</td>
<td>4,961,095</td>
<td>30,673</td>
<td>6,262</td>
<td>27,736</td>
<td>521</td>
<td>39</td>
<td>+1.31%</td>
</tr>
<tr>
<td>dbo:birthDate</td>
<td>3,044,381</td>
<td>1,740,614</td>
<td>639,851</td>
<td>623,055</td>
<td>246,102</td>
<td>606</td>
<td>3,096,767</td>
</tr>
<tr>
<td>distinct entities</td>
<td><strong>3,031,415</strong></td>
<td>1,216,106</td>
<td>639,281</td>
<td>449,742</td>
<td>175,587</td>
<td>606</td>
<td>~</td>
</tr>
<tr>
<td>only in source</td>
<td>1,376,942</td>
<td>25,272</td>
<td>33,540</td>
<td>4,852</td>
<td>1,330</td>
<td>7</td>
<td>+2.16%</td>
</tr>
<tr>
<td>dbo:scientificName</td>
<td>0</td>
<td>0</td>
<td><strong>241,998</strong></td>
<td>0</td>
<td>890,644</td>
<td>1,329,536</td>
<td>1,691,734</td>
</tr>
<tr>
<td>distinct entities</td>
<td>0</td>
<td>0</td>
<td><strong>43,974</strong></td>
<td>0</td>
<td>890,567</td>
<td><strong>1,329,535</strong></td>
<td>~</td>
</tr>
<tr>
<td>only in source</td>
<td>0</td>
<td>0</td>
<td>7,171</td>
<td>0</td>
<td>351,990</td>
<td>780,555</td>
<td>+27.24%</td>
</tr>
</tbody>
</table>
Evaluation: Fusion - Data Quality

- Idea: Use RDFUnit Framework (Data Quality Unit Test Case for KGs)
  - Automatic generation of test instances based on schema knowledge (ontology) of the used vocabulary
  - Contains manually defined plausibility tests
  - Reporting for number of failed test instances and prevalence values

- Quality Indicator: lower number of failure → better data quality
- In total 12,250 generated and 14 manual tests were used for evaluation
## Evaluation: Fusion - Data Quality

### Overall summary for failed test cases

<table>
<thead>
<tr>
<th>Source Description</th>
<th>Wikidata</th>
<th>English</th>
<th>German</th>
<th>French</th>
<th>Dutch</th>
<th>Swedish</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smaller fail rates in source</td>
<td>86</td>
<td>288</td>
<td>163</td>
<td>221</td>
<td>285</td>
<td>115</td>
<td>-</td>
</tr>
<tr>
<td>Equal fail rates in source</td>
<td>5</td>
<td>84</td>
<td>8</td>
<td>74</td>
<td>32</td>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td>Greater fail rates in source</td>
<td>214</td>
<td>643</td>
<td>229</td>
<td>406</td>
<td>306</td>
<td>297</td>
<td>-</td>
</tr>
<tr>
<td>Not failed in fusion</td>
<td>20</td>
<td>40</td>
<td>18</td>
<td>21</td>
<td>24</td>
<td>12</td>
<td>-</td>
</tr>
<tr>
<td>Overall failed</td>
<td>325</td>
<td>1,055</td>
<td>418</td>
<td>722</td>
<td>647</td>
<td>432</td>
<td>1,755</td>
</tr>
<tr>
<td>Prevalence greater zero</td>
<td>531</td>
<td>5,002</td>
<td>1,992</td>
<td>3,560</td>
<td>3,332</td>
<td>1,486</td>
<td>8,060</td>
</tr>
<tr>
<td>Tendency of improvement</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
Evaluation: Enriched Catalan DBpedia - Boost

→ improved inter- and intra-linking and information density

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Enriched</th>
<th>Boost</th>
</tr>
</thead>
<tbody>
<tr>
<td>overall triples</td>
<td>4,631,162</td>
<td>31,200,104</td>
<td>6.74</td>
</tr>
<tr>
<td>distinct entities</td>
<td>981,795</td>
<td>981,795</td>
<td>1.00</td>
</tr>
<tr>
<td>properties distinct</td>
<td>111</td>
<td>2,275</td>
<td>20.50</td>
</tr>
<tr>
<td>sp-pairs</td>
<td>200,094</td>
<td>4,125,355</td>
<td>20.62</td>
</tr>
<tr>
<td>avg pred. outdegree</td>
<td>0.20</td>
<td>4.20</td>
<td>20.62</td>
</tr>
<tr>
<td>avg indegree</td>
<td>0.23</td>
<td>2.58</td>
<td>11.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Enriched</th>
<th>Boost</th>
</tr>
</thead>
<tbody>
<tr>
<td>edge to non-Ca IRI</td>
<td>248,685</td>
<td>5,725,446</td>
<td>23.02</td>
</tr>
<tr>
<td>edge to Global IDs</td>
<td>-</td>
<td>858,551</td>
<td>-</td>
</tr>
<tr>
<td>Global ID targets</td>
<td>-</td>
<td>254,515</td>
<td>-</td>
</tr>
<tr>
<td>ext. non-Ca targets</td>
<td>22,464</td>
<td>2,210,614</td>
<td>98.41</td>
</tr>
<tr>
<td>ext. non-DBp targets</td>
<td>22,464</td>
<td>1,358,754</td>
<td>60.49</td>
</tr>
<tr>
<td>ext. DBpedia targets</td>
<td>0</td>
<td>597,045</td>
<td>-</td>
</tr>
</tbody>
</table>
Evaluation: Enrichment - Value sync classification

- Sole Source Criterion (SSC): true for an sp-pair of source $d$ if all extracted values contributed from $d$ are only originated in $d$
- Alternative Choices Available Criterion (ACC): at least one different extracted value from a source other than $d$ is available in $sp$

Blue: value(s) from Source $d$ are synced / not “challenged”
Green: information only in Source $d$ (erroronous / novel??)
Red: all values from Source $d$ are not synced (unique)
Yellow: partially synced but also unique value(s) in Source $d$ or $d$ is incomplete
### Availability - GFS Data Browser

**About:** Etanôl | Etanol | Ethanol | etanol | etanolo | etanols | ethanol | Éthanol | Étanol | éthanol | Αλκοόλη | Εθανόλη | Εθανόλη | इथनल | इथाल | इथाल | エタノール | 에탄올

- [https://global.dbpedia.org](https://global.dbpedia.org)
- [https://databus.dbpedia.org/dbpedia/prefusion](https://databus.dbpedia.org/dbpedia/prefusion)

<table>
<thead>
<tr>
<th>Value</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etanôl @sv</td>
<td>sv W &amp; wikidata</td>
</tr>
<tr>
<td>Ethanol @de</td>
<td>de W &amp; wikidata</td>
</tr>
<tr>
<td>Etanol @vi</td>
<td>vi wikidata</td>
</tr>
<tr>
<td>Etanol @az</td>
<td>az wikidata</td>
</tr>
<tr>
<td>etanol @uk</td>
<td>uk wikidata</td>
</tr>
<tr>
<td>etanol @ru</td>
<td>ru wikidata</td>
</tr>
<tr>
<td>etanol @hr</td>
<td>hr wikidata</td>
</tr>
<tr>
<td>etanol @ht</td>
<td>ht wikidata</td>
</tr>
<tr>
<td>Etanol @tr</td>
<td>tr wikidata</td>
</tr>
<tr>
<td>Ethanol @en</td>
<td>en W &amp; wikidata</td>
</tr>
<tr>
<td>エタノール @ja</td>
<td>ja wikidata</td>
</tr>
<tr>
<td>جنر @ar</td>
<td>ar wikidata</td>
</tr>
</tbody>
</table>

42 different value/s in 6 source/s
Related Work

HumMer [1]: framework for fusing heterogeneous relational data in three steps: schema mapping, duplicate detection, and conflict resolution.
- pairwise similarity measurements are used to detect duplicated entities which are then extended by a uniform objectID
- conflict resolution based on user defined aggregation functions in SQL (e.g. choose source, first or last, vote, concatenate, most recent value)

Sieve [2] uses scoring functions to rank content and context-based (e.g. release date of the triple) quality indicators to calculate quality metrics.
- fusion process relies on a configuration
  - Definition of 1 fusion function for each property of a class
  - Specification of relevant quality metrics to select the best value(s)
- more fine-grained and selective
- high complexity and maintenance for config, not pragmatic for WD and DBp
Contributions / Summary

- **PreFusion Dataset** - (one of the) largest, open, general domain / purpose knowledge graphs with statement-level provenance
- Flexible and **scalable Workflow** for mass production of custom Knowledge Graphs (DBpedia’s)
- Have shown the usefulness of the dataset and the workflow with a qualitative and quantitative evaluation given 2 usage scenarios with simple configurations
Future Work

- Integration of other datasets (e.g. Musicbrainz)
- Silver standard evaluation of FlexiFusion for a specific domain
- More sophisticated reduce/resolve functions:
  - quality driven (use RDFUnit /SHACL to filter low quality triples)
- Mapping Management similar to ID Management
  - First step owl:equivalentClass and owl:equivalentProperty
- ID Management: Clustering Validation
- Iterative FlexiFusion: enable feedback loops between mapping, linking and fusion
References


Thank you

Any https://global.dbpedia.org/id/1pmvN???

{  "global": "https://global.dbpedia.org/id/1pmvN",
   "locals": [
      "http://dbpedia.org/resource/Question",
      "http://cs.dbpedia.org/resource/Otázka",
      "http://www.wikidata.org/entity/Q189756",
      "http://nl.dbpedia.org/resource/Vraag_(taal)",
      "http://fr.dbpedia.org/resource/Interrogation_(linguistique)",
      "http://de.dbpedia.org/resource/Frage",
      "http://pl.dbpedia.org/resource/Pytanie",
      "http://ta.dbpedia.org/resource/தேகள்வி",
      "http://simple.dbpedia.org/resource/Question",
      "http://da.dbpedia.org/resource/Sparsgsmål",
      "http://it.dbpedia.org/resource/Quesito",
      "http://is.dbpedia.org/resource/Spurning",
      "http://qu.dbpedia.org/resource/Tapuy",
      "http://id.dbpedia.org/resource/Pertanyaan",
      "http://ky.dbpedia.org/resource/Cyrpo",
      "http://cv.dbpedia.org/resource/Њйтую",
      "http://hi.dbpedia.org/resource/प्रश्न",
      "http://war.dbpedia.org/resource/Paki-ana",
      "http://pt.dbpedia.org/resource/Pergunta",
      "http://he.dbpedia.org/resource/שאלה",
      "http://th.dbpedia.org/resource/คำถาม",
      "http://es.dbpedia.org/resource/Pregunta",
      "http://ru.dbpedia.org/resource/Вопрос",
      "http://ro.dbpedia.org/resource/Întrebare",
      "http://uk.dbpedia.org/resource/Питання",
      "http://eo.dbpedia.org/resource/Demando",
      "http://fa.dbpedia.org/resource/پرسش",
      "http://kk.dbpedia.org/resource/Сұрақ",
      "http://hu.dbpedia.org/resource/Kérdés",
      "http://sco.dbpedia.org/resource/Quaistien",
      "http://az.dbpedia.org/resource/Sual",
      "http://ar.dbpedia.org/resource/سؤال",
      "http://sw.dbpedia.org/resource/Swali",
      "http://ca.dbpedia.org/resource/Pregunta",
      "http://ja.dbpedia.org/resource/疑問文",
      "http://ko.dbpedia.org/resource/질문",
      "http://sk.dbpedia.org/resource/Otázka",
      "http://tr.dbpedia.org/resource/Soru",
      "http://zh.dbpedia.org/resource/疑问句",
      "http://tl.dbpedia.org/resource/Tanong"  ]
}
Backup
# DBpedia PreFusion Dataset Factsheet

## Table 1. PreFusion dataset factsheet, dbpedia/prefusion/$artifact/2019.03.01

<table>
<thead>
<tr>
<th>artifact</th>
<th>distinct objects</th>
<th>source triples</th>
<th>subjects</th>
<th>sp-pairs</th>
<th>wikipedias</th>
<th>size (bz2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>labels</td>
<td>266,633,208</td>
<td>297,345,045</td>
<td>91,146,077</td>
<td>91,146,077</td>
<td>139+wd</td>
<td>7.2G</td>
</tr>
<tr>
<td>instance-types</td>
<td>191,702,603</td>
<td>293,261,187</td>
<td>25,230,546</td>
<td>25,230,546</td>
<td>40+wd</td>
<td>2.1G</td>
</tr>
<tr>
<td>mapping-based-objects</td>
<td>150,955,259</td>
<td>263,677,844</td>
<td>45,063,398</td>
<td>98,388,770</td>
<td>40+wd</td>
<td>6.1G</td>
</tr>
<tr>
<td>mapping-based-literals</td>
<td>94,111,662</td>
<td>100,049,794</td>
<td>36,500,856</td>
<td>71,427,960</td>
<td>40+wd</td>
<td>4.0G</td>
</tr>
<tr>
<td>geo-coordinates</td>
<td>41,313,484</td>
<td>51,178,574</td>
<td>8,517,009</td>
<td>34,099,723</td>
<td>140+wd</td>
<td>1.8G</td>
</tr>
<tr>
<td>specific-mapping-based</td>
<td>2,198,020</td>
<td>2,548,485</td>
<td>1,083,961</td>
<td>1,568,804</td>
<td>40</td>
<td>82M</td>
</tr>
</tbody>
</table>
Scalability

- Workflow design can scale up to huge amounts of data sources and entities/properties given an appropriate cluster and Big Data Framework.
- Prototype horizontally scalable via Apache Spark.
- Example runtime on single machine cluster with AMD Opteron 6376 @ 64x 2.6GHz, 252 GiB RAM for 2019.03.01 (Pre)Fusion datasets:

<table>
<thead>
<tr>
<th>artifact</th>
<th>ID rewriting</th>
<th>PreFuse</th>
<th>Fuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>labels</td>
<td>02:16 h</td>
<td>52:45 min</td>
<td>25:45 min</td>
</tr>
<tr>
<td>instance-types</td>
<td>02:41 h</td>
<td>38:18 min</td>
<td>15:03 min</td>
</tr>
<tr>
<td>mappingbased-objects</td>
<td>01:49 h</td>
<td>39:22 min</td>
<td>13:12 min</td>
</tr>
<tr>
<td>mappingbased-literals</td>
<td>01:33 h</td>
<td>20:37 min</td>
<td>08:08 min</td>
</tr>
<tr>
<td>geo-coordinates</td>
<td>03:38 h</td>
<td>14:41 min</td>
<td>04:17 min</td>
</tr>
<tr>
<td>specific-mappingbased</td>
<td>51:14 min</td>
<td>02:20 min</td>
<td>00:49 min</td>
</tr>
</tbody>
</table>