GlobalFactSync

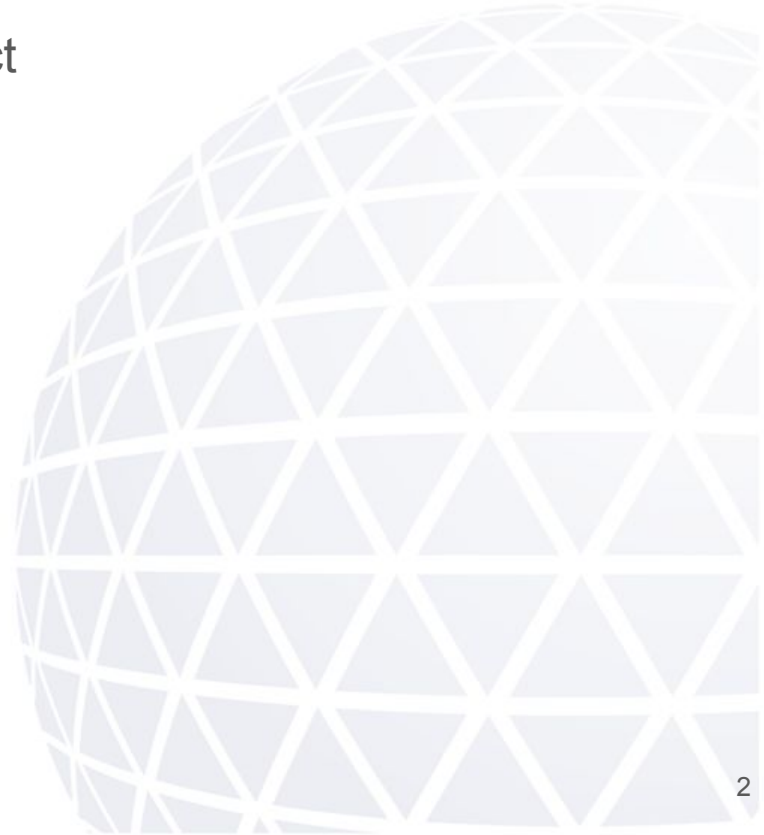
Wikimania 2019

Sebastian Hellmann, Tina Schmeißner, Marvin Hofer
Johannes Frey

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

Agenda

1. Introduction to GlobalFactSync (GFS) Project
2. GFS UI
3. GFS under the hood
   a. DBpedia in a nutshell
   b. Extractions



GlobalFactSync(RE)

- Project supported by Wikimedia Grant and DBpedia Association
  https://meta.wikimedia.org/wiki/Grants:Project/DBpedia/GlobalFactSyncRE
- Motivation:
  - Infoboxes significantly vary between different Wikipedia language versions w.r.t. quality, comprehensiveness and up-to-dateness
  - Wikidata is (still) not adopted in a wide range of infobox templates

- Problem statements:
  1) (Facts from) Wikipedia Infoboxes are not managed on a global scale
  2) Manually created and verified facts from Wikipedia infoboxes are not included in Wikidata

GlobalFactSync(RE)

- Project Goals
  - 1) enable Wikipedia-to-Wikipedia *synchronization* of Infobox facts
  - 2) enable Import/Upstream of facts from Wikipedia to Wikidata **with references**
  - 1) & 2) for the comparison of contradictious values or enrichment of missing information also take other primary sources / Knowledge Bases into account
  - Following a “Human(s)-in-the-loop” *synchronization*
    - final decision whether to synchronize a value is made by an editor who understands consensus and the implications
    - no automatic imports; focus is to drastically reduce the time for editors to research all references for individual facts and make infobox maintenance faster & more effective

GlobalFactSync(RE)

Project Relevance for Wikimania 2019 theme

Sustainable Development Goals

- Goal 4 - quality education:
  - increased data quality of infoboxes allowing high quality education
- Goal 10 - Reduce inequality within and among countries:
  - GFS tools aim to reduce the gap between infoboxes in different languages
- Goal 8 and 9 - economic growth and industry / innovation / infrastructure:
  - Data is the new oil, knowledge is key

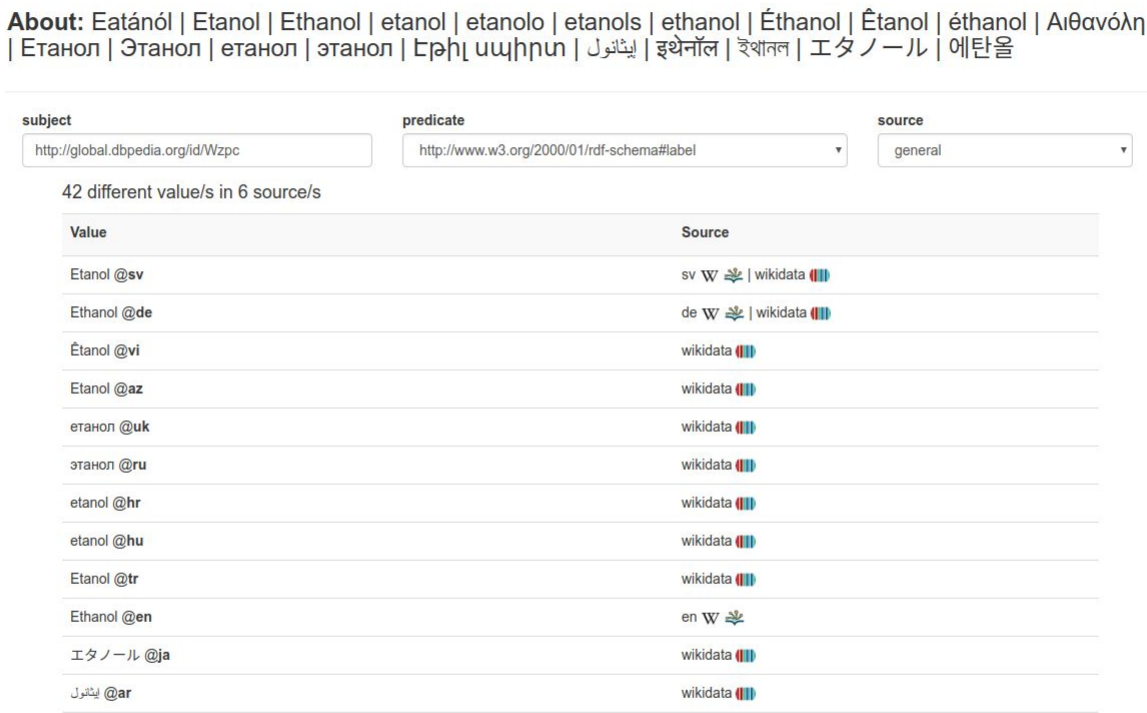
Current GFS User Story

@Jc86035: Boys Don't Cry (Q3020026) seems like a good starting point. There is several conflicting information, i.e. publication date is either 15/3 or 16/3 [...] We can already detect this partially with the prototype [...] SebastianHellmann (talk) 13:22, 29 April 2019 (UTC)

For "Boys Don't Cry"'s release date specifically, the discrepancy seems to be because the Japanese Wikipedia article (15 March) was edited to match the English and French Wikipedia articles (16 March), and Wikidata was not updated at the same time. In the Japanese Wikipedia article, "15 March" was introduced in 2012 and replaced in 2017 by two different unregistered users. I've updated Wikidata so that all five sources are in agreement. Jc86035 (talk) 13:57, 29 April 2019 (UTC)

GFS UI Prototype

- **GFS Data Browser** (https://global.dbpedia.org)
- Shows aggregated view of values (and their sources) for 1 attribute given any Wikimedia Article URL
- Links to source pages



GFS UI Wikipedia Integration Draft

- Embedded sync symbols in Infobox view (mockup) with direct link to show (alternative) values for an entry in GFS Data Browser
  (https://meta.wikimedia.org/wiki/Grants_talk:Project/DBpedia/GlobalFactSyncRE/Archive_1#New_prototype_and_new_ideas)
- User Script to include symbol on top of Wikipedia Article page linking to GFS Data Browser

user script: https://meta.wikimedia.org/wiki/User:JohannesFre/global.js
add to your https://meta.wikimedia.org/wiki/special:MyPage/global.js

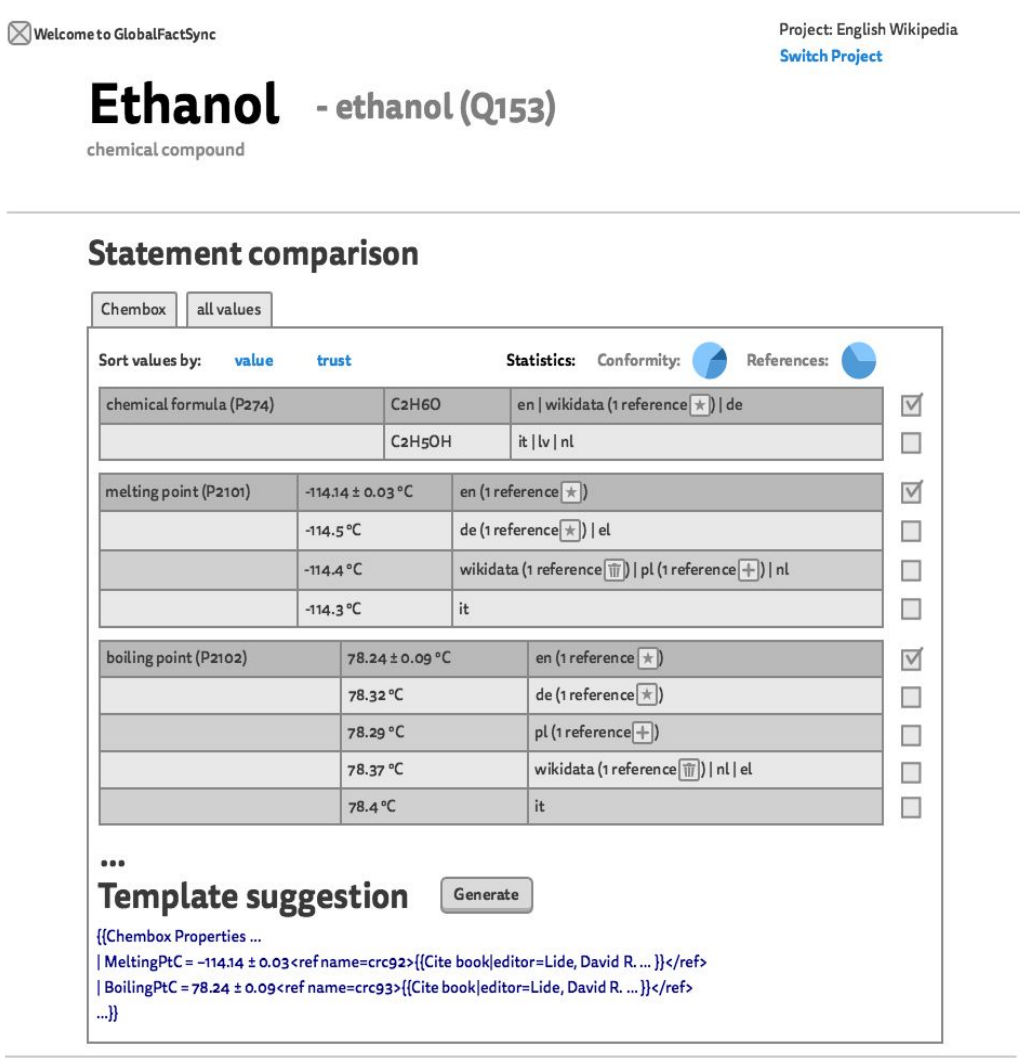
Eiffel Tower [Collapse]

Tour Eiffel

General information	
Type	Observation tower Broadcasting tower
Location	7th arrondissement, Paris, France
Coordinates:	48.858222°N 2.294500°E
Construction started	28 January 1887
Completed	15 March 1889
Opening	31 March 1889 (129 years ago)
Owner	City of Paris, France
Management	*Société d'Exploitation de la Tour Eiffel* (SETE)
Height	
Architectural	300 m (984 ft)[1]
Tip	324 m (1,063 ft)[1]
Top floor	276 m (906 ft)[1]
Technical details	
Floor count	3[2]
Lifts/elevators	8[2]
Design and construction	
Architect	Stephen Sauvestre

GFS UI Vision / Ideas

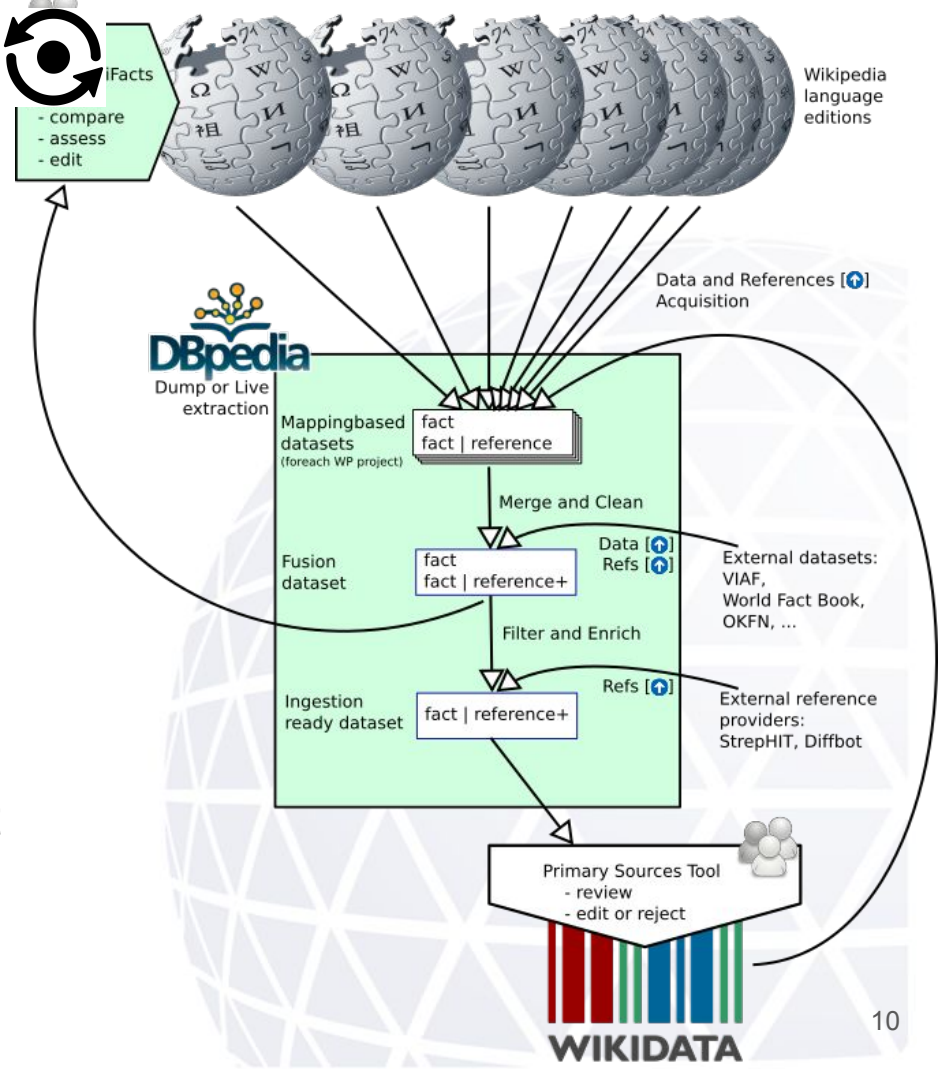
- Feedback for (in)correct values
- Order values by trustworthiness (references, conformity), edit date, ...
- Infobox template generator based on selection of synced values



DBpedia AKSW InfAI Institute for Applied Informatics UNIVERSITÄT LEIPZIG

GFS Dataflow

- DBpedia (Infobox) & reference extraction (dump or live) for Wikipedias
- cleansing and merging (including fusion of external datasets & Wikidata)
- Import into MongoDB and GFS Data Browser
- Handover of Wikipedia editor consent to Wikidata



DBpedia in a nutshell

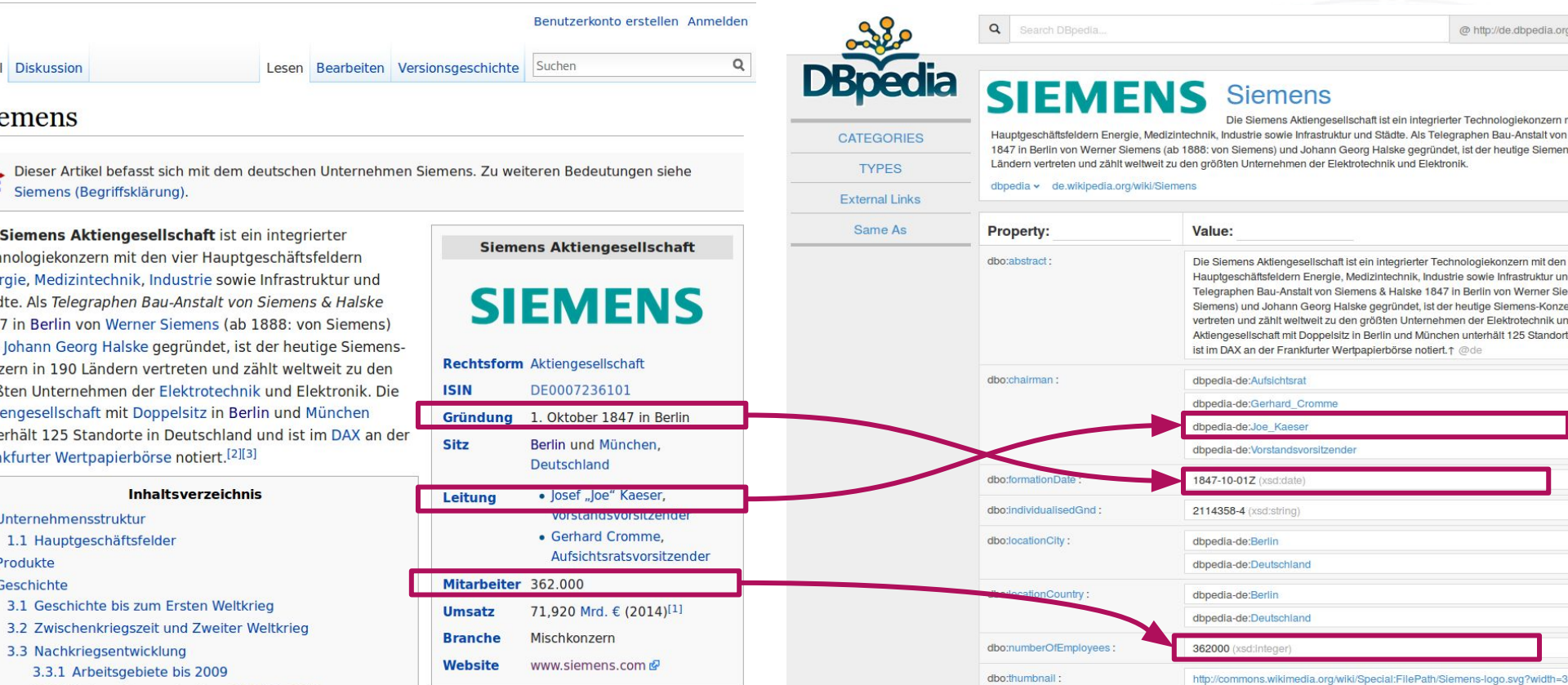
« A large-scale, multilingual Knowledge Base »

- Started in 2007 as a crowd-sourced community effort to semi-automatically extract structured (RDF) information from Wikipedia to make this information queryable on the Web (SPARQL)

« Global and unified access to knowledge graphs »

- new mission since 2018; original definition still holds true
- Databus Platform to integrate your data with other data

DBpedia: A large-scale, multilingual Knowledge Base



Extraction: Infobox ⇒ RDF

```
{{Infobox Unternehmen
| Name                = Siemens
| Logo                = Siemens-logo.svg
| Unternehmensform = [[Aktiengesellschaft
(Deutschland)|Aktiengesellschaft]]
| ISIN                = DE0007236101
| Gründungsdatum    = 1. Oktober 1847
| Sitz                = [[Berlin]] und
[[München]]
| Leitung             =
* [[Joe Kaeser]], [[Vorstandsvorsitzender]]
* [[Gerhard Cromme]], [[Aufsichtsrat]]
| Mitarbeiterzahl   = 362.000
| Umsatz              = 75,636
[[Milliarde|Mrd.]] [[Euro|€]]
| Stand               = 2015-12-31
| Branche             = [[Mischkonzern]]
| Homepage            = www.siemens.com
}}
```

DBpedia Extraction Framework &
DBpedia Mappings

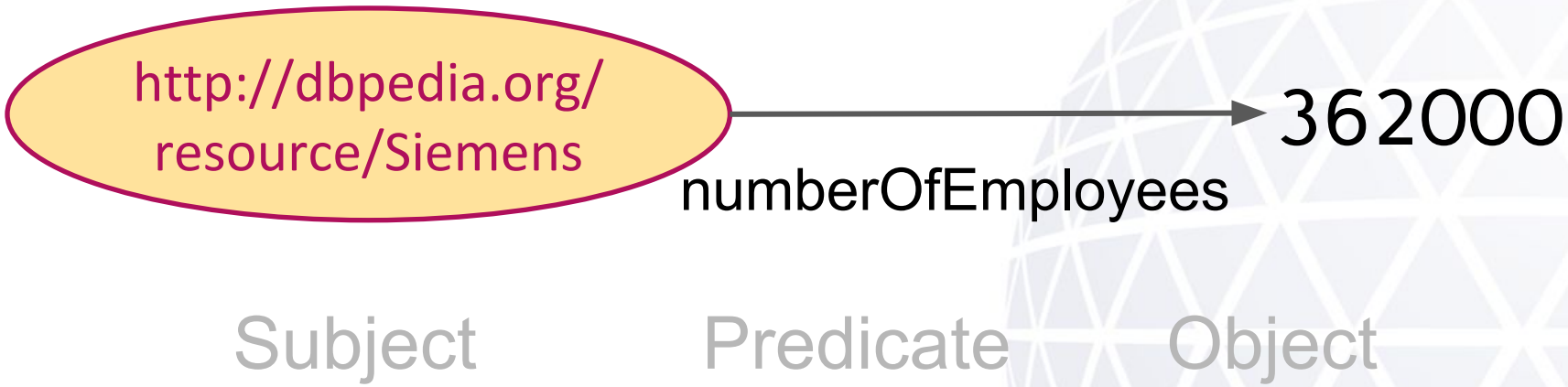
```
dbpedia-de:Siemens
      a                          dbo:Company ;
      rdfs:label                 "Siemens"@de ;
      dbo:chairman               dbpedia-de:Joe_Kaeser;
      dbo:numberOfEmployees      362000 ;
      dbo:formationDate          "1847-10-01"^^xsd:date

            ...

      :wikiPageID                  2679650 ;
      :wikiPageRevisionID          148144260 ;
```

RDF - Resource Description Framework

- Statements of subject > predicate > object
- Similar to Wikidata truthy statements in RDF/SPARQL export



Mapping-based Infobox extraction

- Crowd-sourced DBpedia mappings Wiki (mappings.dbpedia.org) contains Infobox mappings for ~ 40 Wikipedia languages
- Values are normalized (e.g. square feet to square km)
- Infobox parameters are mapped to uniform, language independent http://dbpedia.org/ontology/* properties

```
{{Infobox settlement
|official_name     = Stockholm ...
|latd=59
|latm=19
|lats = 46
|area_urban_km2    = 381.63 ...
```

Property Mapping (help)	
template property	area_urban_km2
ontology property	areaUrban
unit	squareKilometre

Geocoordinates Mapping (help)	
coordinates template property	coordinates

Geocoordinates Mapping (help)	
longitude degrees template property	longd
longitude minutes template property	longm
longitude seconds template property	longs
longitude direction template property	longEW
latitude degrees template property	latd
latitude minutes template property	latm
latitude seconds template property	lats
latitude direction template property	latNS

Wikidata extraction

- Similar approach as for Wikipedia:
- Mappings in JSON to DBpedia Ontology

+ Allows unified access over DBpedia and Wikidata
+ Wikidata has no ontology, DBpedia has 8 (DBO, Yago, Umbel,..)

```
"P279": [
    {
        "rdfs:subClassOf": "$getDBpediaClass"
    }
],
"P625": [
    {
        "rdf:type": "http://www.w3.org/2003/01/geo/wgs84_pos#SpatialT
    },
    {
        "geo:lat": "$getLatitude"
    },
    {
        "geo:long": "$getLongitude"
    },
    {
        "georss:point": "$getGeoRss"
    }
],
```

Reference Extraction

- Extracts citations from Infoboxes
- parser works for: de, en, es, fr, it, nl, pl, pt, ru, sv at the moment


```
"Infobox_name": "Infobox website",
  "Parameter_name": "programming_language",
  "Parameter_value": "http://dbpedia.org/resource/C++",
  "Reference_code": "{{cite news |url=http://www....location
  "Reference_name": "<noname_ref>",
  "Reference_parameters": [
    {
      "citation_id": "http://drdobbs.com/mobile/facebook-ado
      "citation_template": "cite news",
      "date": "October 16, 2013",
      "first": "Adrian",
      "last": "Bridgwater",
      "location": "San Francisco",
      "title": "Facebook Adopts D Language",
      "url": "http://www.drdobbs.com/mobile/facebook-adopts-
      "work": "Dr Dobb's"
    }
  ],
  "Wikipedia_article": "Facebook",
  "Wikipedia_language": "en",
  "triple": {
    "o": "http://dbpedia.org/resource/C++",
    "p": "http://dbpedia.org/property/programmingLanguage",
    "s": "http://dbpedia.org/resource/Facebook"
  }
},
```

Services for Developers and “Pro Users”

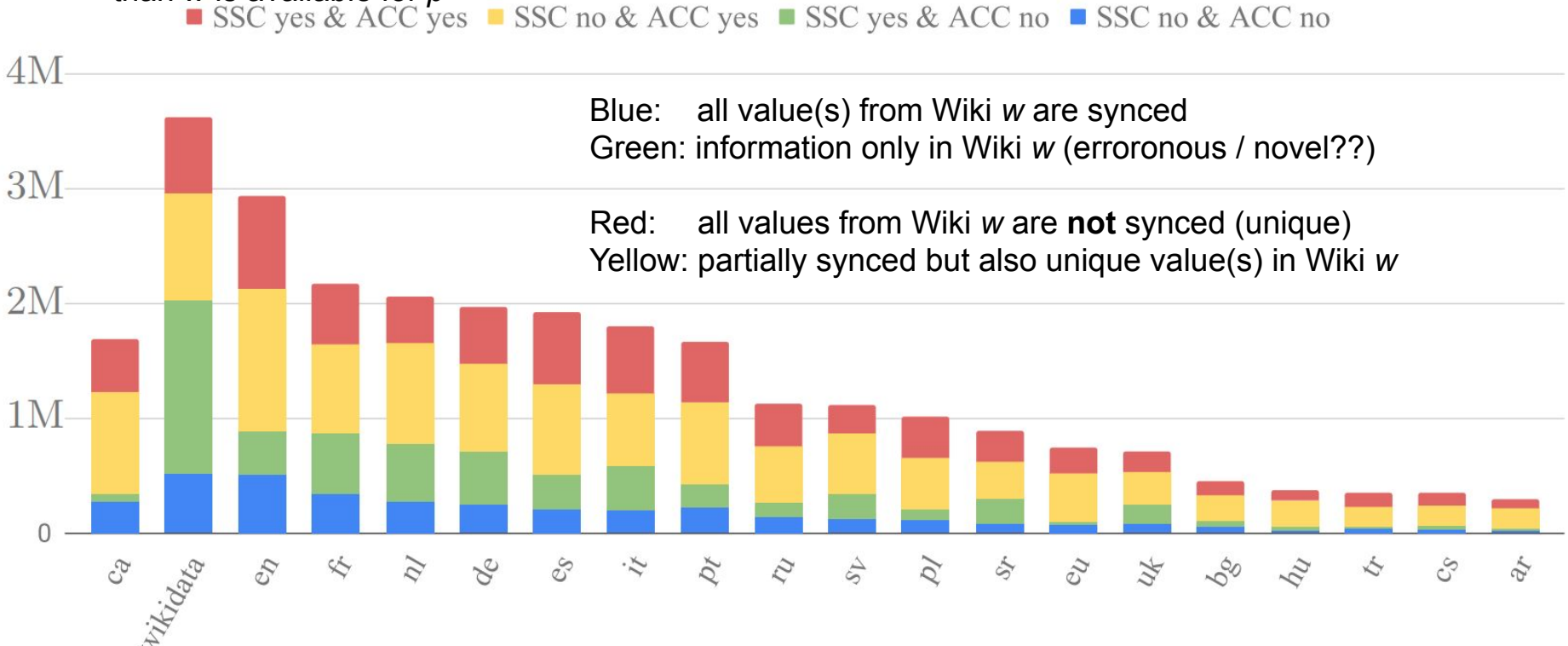
- Adhoc DBpedia Fact Extraction (RDF)
  - http://dbpedia.informatik.uni-leipzig.de:9998/server/
- Adhoc Reference Extraction (JSON or TSV)
  - Tied to DBpedia Facts
  - Standalone References
- MongoDB query endpoint with entire GFS data for lookups and simple analytical queries
  - http://global.dbpedia.org:8990 (user: read, pw: gfs)

    the number of locations having at least one population value

    locations with more than 10 population values
- GFS Data Browser as JSON
  - https://global.dbpedia.org/**raw**/?s=https%3A%2F%2Fwww.wikidata.org%2Fw%2Findex.php%3Ftitle%3DQ268%26oldid%3D617866885

Classification of GFS data focussed on Catalan Wiki

- Sole Source Criterion (SSC): all extracted values contributed from Wiki *w* for an infobox property *p* in one article are only originated in *w*
- Alternative Choices Available Criterion (ACC): at least one different extracted value from a Wiki other than *w* is available for *p*



Open Questions to Wikimedia Community

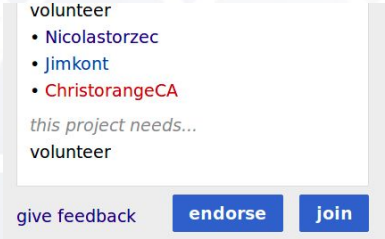
- How are changes ingested into Infobox?
  - Fully manual?
  - Copy'n'paste with generator support?
  - Visual editor extension?
- How is GFS data presented to Wikipedians?
  - Userscript approach → Users decide, only user affected
  - Template modification approach → template maintainers decide, all viewers affected
  - Visual Editor → Users decide
- How can Editors find "unsynced" values in Articles?
  - External Service with List of Articles with potentially badly synced Infoboxes
  - Pings of Watchlist members (e.g. if sth. gets outdated)
  - Notification in (Visual) Editor when new revision is saved (Userscript/editor Plugin)

Next steps

- Improve User Script
- Modify NBA template to include direct links as shown in Mockup
- Add Live Refresh to GFS Data Browser
- Add view for references into GFS Data Browser
- Select a list of Infobox types/domains as sync targets
- “Friendly fork” of Harvest Templates tool based on DBpedia Technology allowing import to Wikidata including references

Feedback requested

- Mailing List
  - gfs@infai.org
- Project discussion
  - https://meta.wikimedia.org/wiki/Grants_talk:Project/DBpedia/GlobalFactSyncRE
- **Become member of the feedback squad**
  - visit https://meta.wikimedia.org/wiki/Grants:Project/DBpedia/GlobalFactSyncRE and click on the join button at the bottom of the Project box
- On Github
  - https://github.com/dbpedia/gfs/issues
- After the talk & in the break



Thank you

Any "http://sv.dbpedia.org/resource/**Fråga**" ?


```
{
  "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/கேள்வி",
    "http://simple.dbpedia.org/resource/Question",
    "http://da.dbpedia.org/resource/Spørgsmål",
    "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/Суроо",
    "http://cv.dbpedia.org/resource/Ыйту",
    "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/Quaisten",
    "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",
```