



Wikidata as universal thesaurus

Theo van Veen, WikidataCon 2017, 28-10-2017

www.kb.nl

Overview

- The idea and motivation
- Historical newspapers as motivating use case
- Possible approaches, different environments
- Pros and objections
- Conclusions

Why do we use a thesaurus?

- Unique identification of an entity
- Providing context information
- Enabling search using the entity's properties as captured in the thesaurus (including name variants)

For what reason do we link to Wikidata?

- We use Wikidata as knowledge base and as a hub to other knowledge bases (>2000 external identifiers) to link identifiers and get more properties.

The idea: gradually adopt Wikidata as a universal thesaurus because ...

- Libraries and other institutions from different disciplines increasingly want to get connected
- Connecting to a central hub is more efficient than connecting everything to everything
- Using the same identifier for the same resource is more efficient than linking resources
- Inventing yet another identifier is less efficient than using a (rich) existing one: the Wikidata identifier
- Libraries may want to share responsibility for Wikidata as a common thesaurus

Create trusted links

KB Catalogue | **rushdie**

title: De duivelsverzen / Salman Rushdie ; [vert. Salman Rushdie]

creator: <http://data.kb.nl/thesaurus/068941056>
<http://viaf.org/viaf/29540187>
<http://www.wikidata.org/entity/Q44306>

KB thesaurus

<http://data.kb.nl/thesaurus/068941056>

VIAF

<http://viaf.org/viaf/29540187/>

WIKIDATA

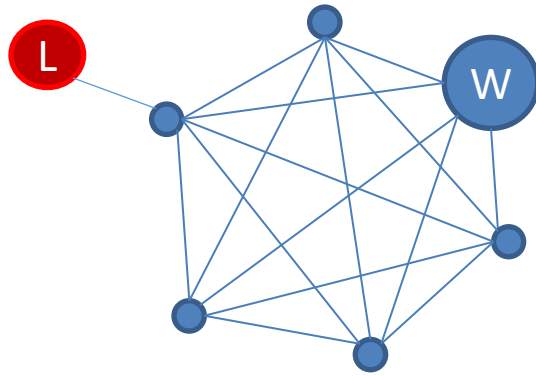
<https://www.wikidata.org/wiki/Q44306>

ISNI

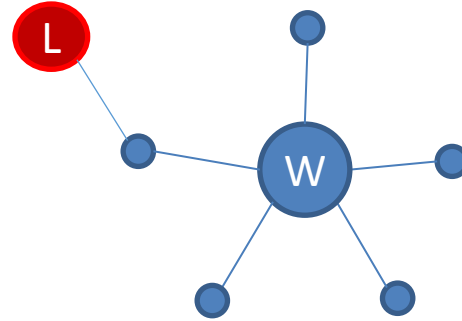
<http://www.isni.org/isni/0000000122778487>

- For bibliographic data many trusted links are available: from thesaurus to VIAF, from VIAF to ISNI and from ISNI to Wikidata
- In many situations links for persons, events and locations the links have to be created

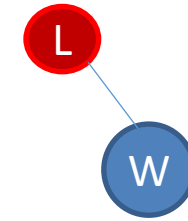
Wikidata as hub, as central hub and as universal thesaurus



“Everything links to everything”



“Wikidata as central hub”



“Everything links directly to Wikidata”

● Thesaurus

● L Library catalogue

● W Wikidata

Bibliographic record

<creator rdf:Resource="**<http://data.kb.nl/thesaurus/068350767>**">Albert Einstein</creator>

<creator rdf:Resource="**<http://bnb.data.bl.uk/id/concept/person/lcsh/EinsteinAlbert1879-1955>**">Albert Einstein</creator>

<creator rdf:Resource="**http://data.bnf.fr/11901607/albert_einstein/**">Albert Einstein</creator>

<creator rdf:Resource="**<http://id.loc.gov/authorities/names/n79022889>**">Albert Einstein</creator>

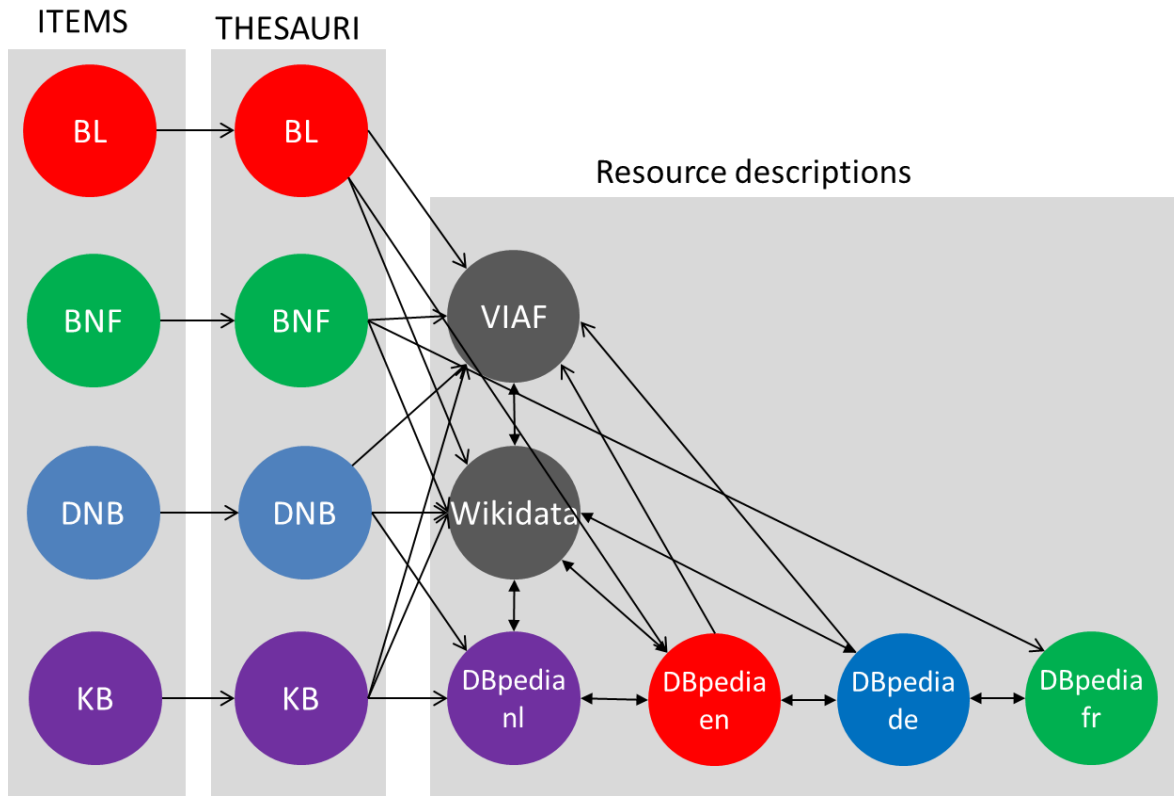
all become:

<creator rdf:Resource="**<https://www.wikidata.org/wiki/Q937>**">Albert Einstein</creator>

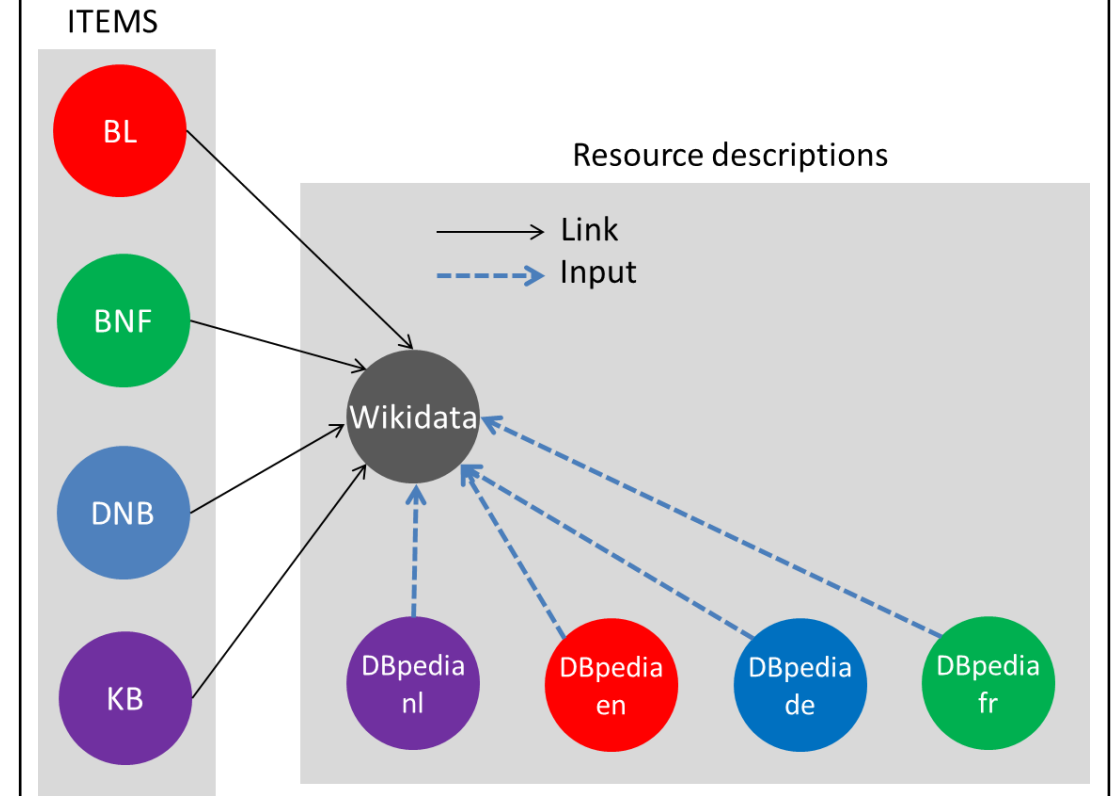
With index field: **wd_id= Q937**

Wikidata as universal thesaurus for libraries

Current situation: many to many links
(many identifiers for single resource)



Proposed: everything links to Wikidata
(same identifier for single resource)

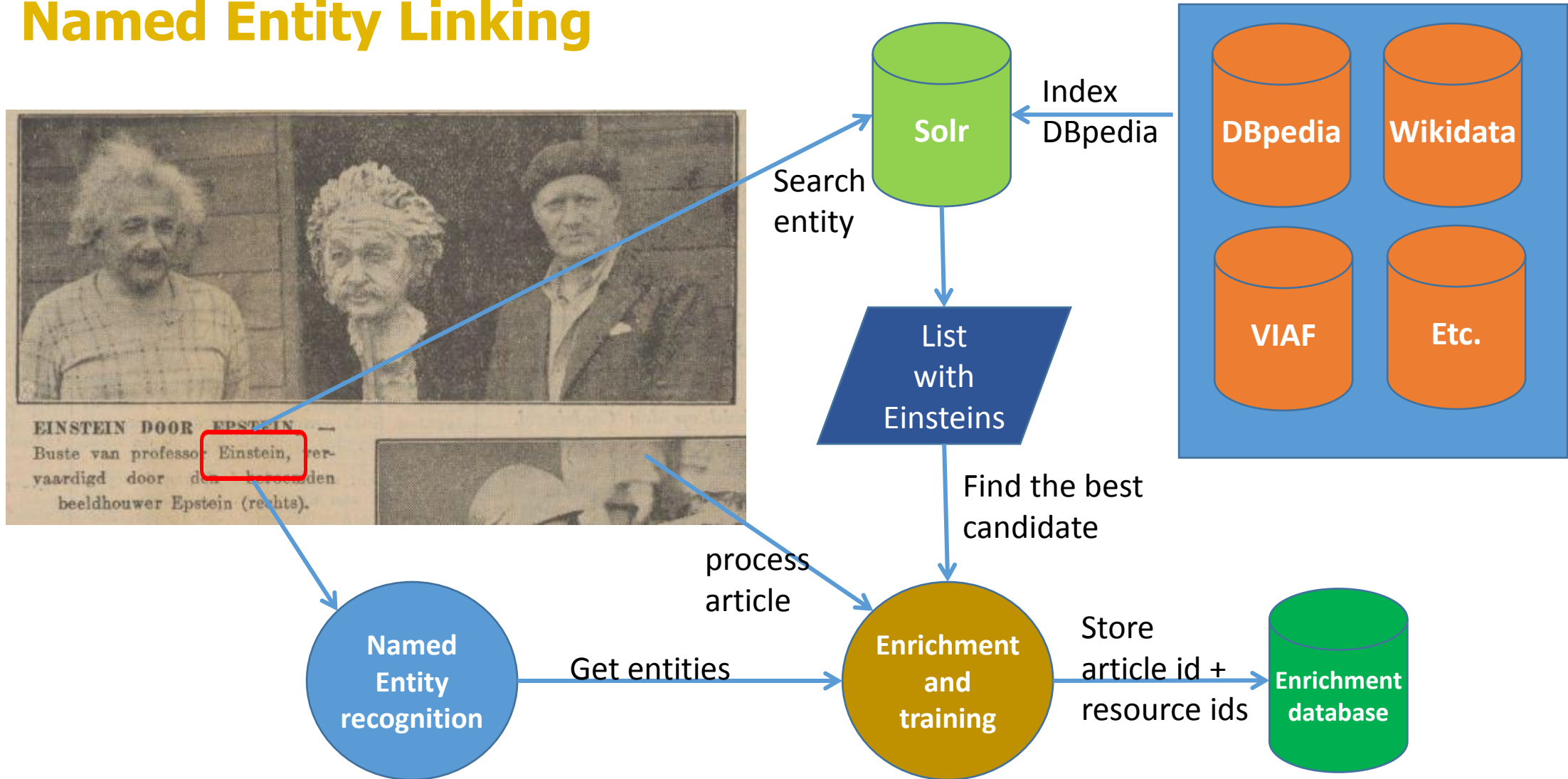


Lower barriers

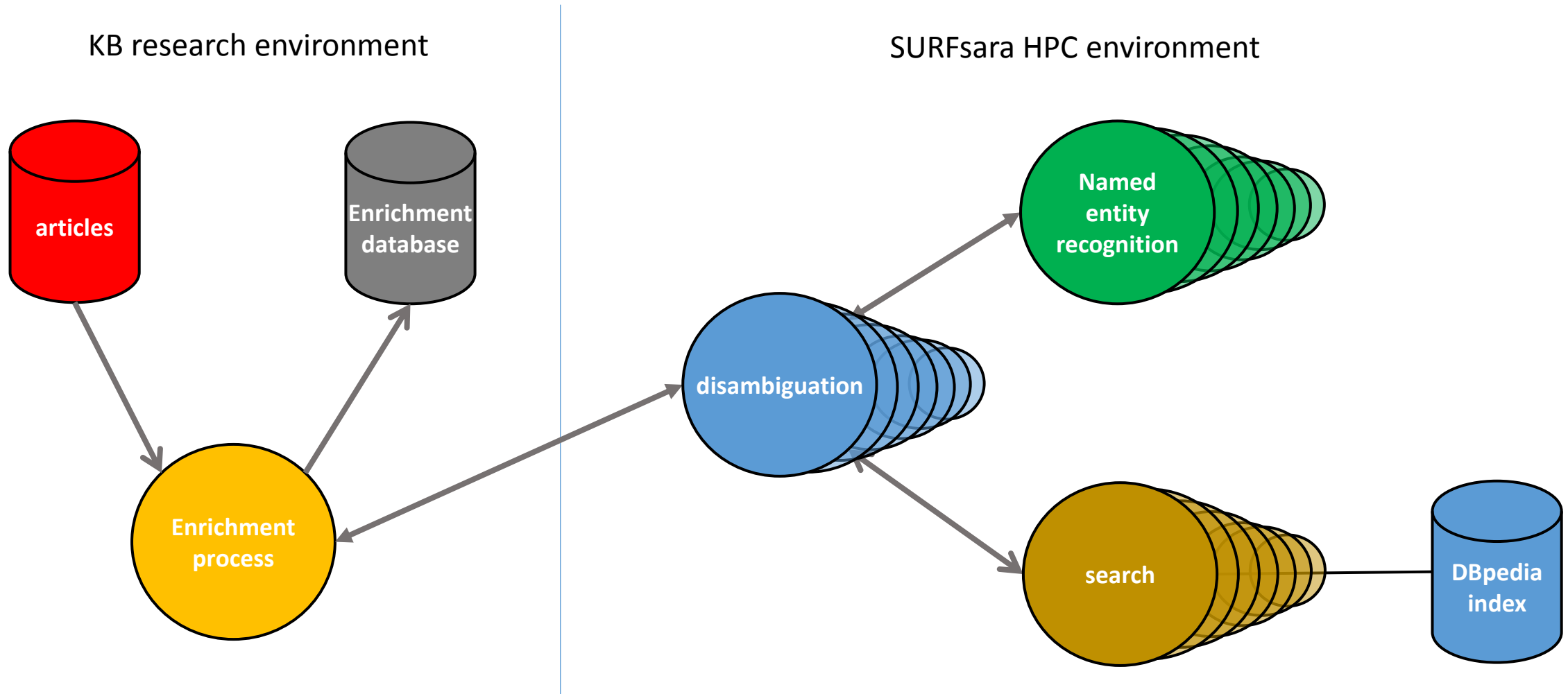
- We use a name authority thesaurus for **unique identification** of resources. Sharing a single identification across institutions will make such identification **globally unique** and usable
- Minimizing the **number of hubs**, minimizing the **number of variations** for identical queries in different databases and minimizing the **required knowledge** will lower the barrier to connect to external sources
- Sharing a global identifier makes it easier for institutions to connect without dramatically changing their infrastructure
- Why standardizing “everything” but not resource identifiers?

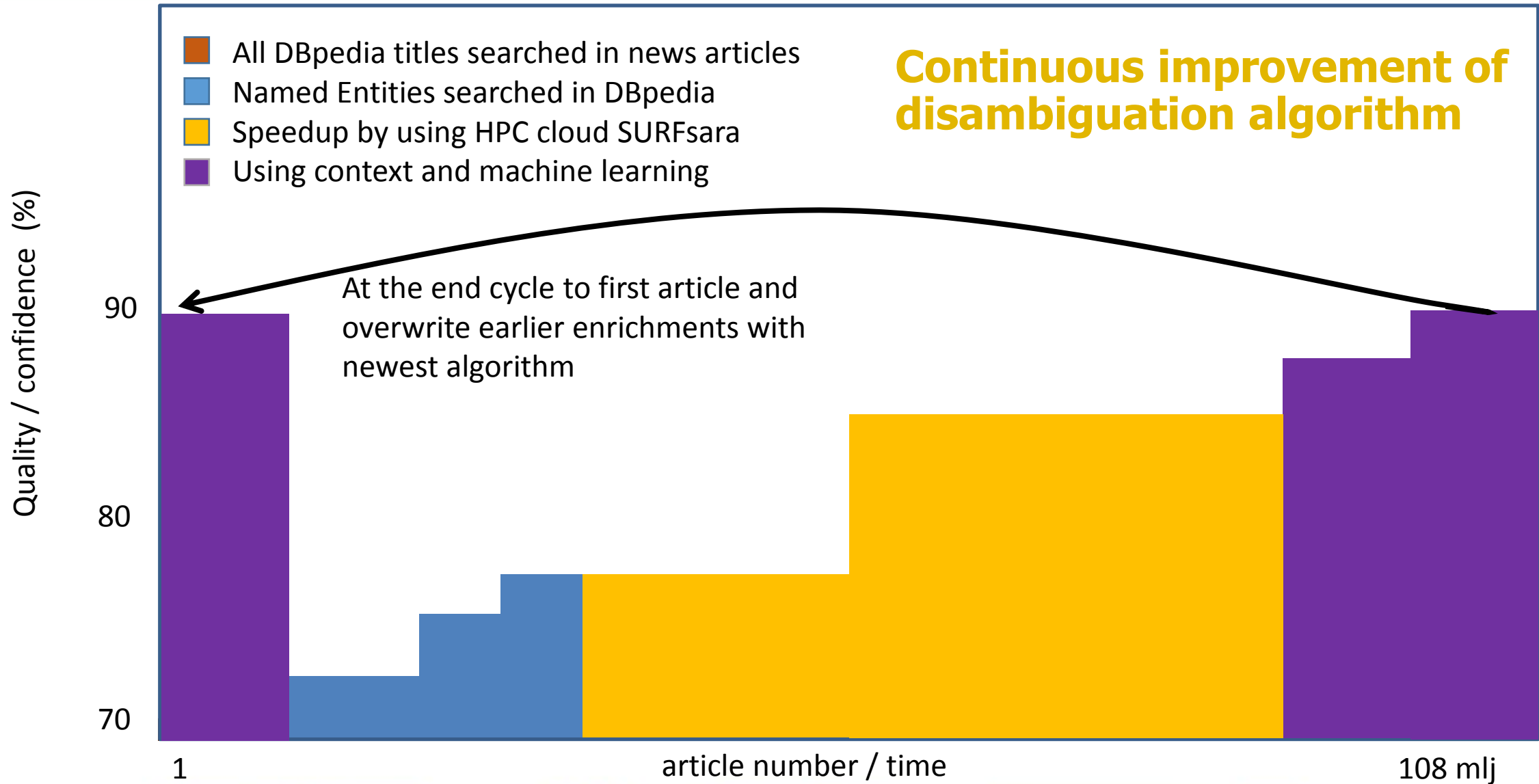
Motivating use case: KB Historical newspapers

Named Entity Linking



Enrichment infrastructure using SURFsara HPC cloud





From conventional entity linking to deep learning and beyond

algorithm	accuracy	link recall	link precision	link F-measure
Rule based	.76	.76	.65	.70
Machine learning (SVM)	.84	.76	.83	.79
Neural network	.84	.73	.87	.79
Extra features e.g. word embedding	.85	.81	.82	.82
Extra Wikidata data, more training data	.87	.81	.86	.84
Entity embedding	.88	.86	.85	.85
New architecture	?	?	?	?

Semantic search: index resource identifiers



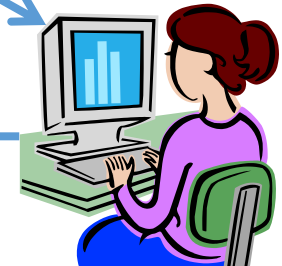
Get text for article X



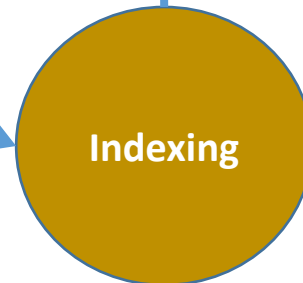
Semantic search (SPARQL) providing wikidata id's



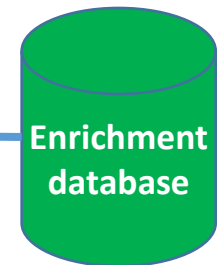
search articles with wikidata id's



Text + Viaf id + Wikidata id etc.



Get enrichments for article X





KB

Newspapers +

SELECT ?p WHERE {?p wdt:P39 wd:Q18887908 . ?p wdt:P19 ?

1 of 66488 results for SELECT ?p WHERE {?p wdt:P39 wd:Q18887908 . ?p wdt:P19 ?place . ?place wdt: >

- | | | |
|---|--|------------|
| 1 | Nederlandsche muziek en musici in Amerika.
Amerika · New York · Mengelberg · Amerikaan · Nederland · Christoffer Boe ... | 1920/10/18 |
| 2 | NEDERLAND.
Veenhuizen · Tweede Kamer der Staten-Generaal · Financiën · Hannover · Nederland · Antwerpen ... | 1861/02/28 |
| 3 | UITSLAG DER VERKIEZINGEN voor Leden van de Tweede Hamer der Sinten-Generaal.
Amsterdam · School Is Cool · Jan van Swieten · Jan Heemskerk · Joseph John Thomson · Zeventien Provinciën ... | 1864/06/17 |
| | INTÉRIEUR. LA HAYE, 27 août. LA HOLLANDE A L'EUROPE. COUP D'OEIL SURLARÉVOLUTION BELGE.
België · Frankrijk · Erasme Louis Surllet de Chokier · Charles de Brouckère · Léhon · Goswin de Stassart ... | 1831/08/28 |
| | Zomer- Vergadering DER PROVINCIALE STATEN.
Klaas Dijkhoff · Gilze · Willem Frederik Hermans · Jan van Swieten · Carl Verheijen · Provinciale Staten ... | 1858/07/08 |
| 6 | TWEEDE KAMER.
Rochussen · Jan Karel van Goltstein · Zuylen · Jean-Jacques van Zuylen van Nyevelt · Van Foreest · Kek ... | 1864/11/25 |
| 7 | TWEEDE KAMER. | 1865/11/30 |

Articles mentioning members of parliament not born in the Netherlands

```
SELECT ?p WHERE {
  ?p wdt:P39 wd:Q18887908 .
  ?p wdt:P19 ?place .
  ?place wdt:P17 ?country .
  FILTER NOT EXISTS {
    ?place wdt:P17 wd:Q55 .
  }
}
```




For the same query in the catalogue the Wikidata identifier is converted to the local thesaurus identifier

- 1 **Verklaring / van Goswin Joseph Augustin baron de Stassart (1780-1854) en Etienne Constantin de Gerlache (1785-1871), Tweede-Kamerleden, over de herverdeling door de Staten-Generaal van lasten over de provincies**
- 2 **Brieven van Goswin Joseph Augustin baron de Stassart (1780-1854)**
- 3 **Die Niederländische Ost-Kompagnie A.G. : die Pionierarbeit von Dr. M.M. Rost van Tonningen im Osten / M.M. Rost van Tonningen & F.S. Rost van Tonningen-Heubel**
- 4 **Pour ou contre Rome ou le comte de Zuijlen de Nyevelt à la recherche d'un parti politique**
- 5 **Dekking onzer buitengewone uitgaven**
- 6 **Verkeert Nederland in gevaar?**
- 7 **Moet het Westen zeggen dat het nooit als eerste kernwapens zal inzetten?**
- 8 **Hydrographische kaart der zeegaten van de monden der Schelde, met de reeden van Vlissingen en Veere**
- 9 **Briefe aus der Vergangenheit. Bd. 1**
- 10 **Briefe aus der Vergangenheit. Bd. 2**

Navigation example

- **Semantic query between [], in this case expand to all Roman Emperors**
- **Select “newspaper+” collection**
- Select a result
- Click on a linked named entity for more information
- Click on “More info” for properties of this entity
- Click on a property for searching more articles about resources with that property
- And see the result: all articles mentioning persons that have been married to Elizabeth Taylor

KB Newspapers + speltakelfilm and [romeinse keizer]

KB Catalogue
Memory of the Netherlands
Books
Journals
1. A Newspapers
2. In Newspapers +
3. C e-Depot
4. S Parliamentary papers
Note: T ANP newsbulletins
therefor Watermarks
Please request: Medieval manuscripts
Have fur Flamish lit. journals
Pootman collection
LITERAT
Alba Amicorum
Chess affiches
Manuscripts

Using square brackets the software tries a few Wikidata SPARQL queries and replaces this string by the Wikidata results.

Navigation example

- Semantic query between [], in this case expand to all Roman Emperors
- Select “newspaper+” collection
- **Select a result**
- Click on a linked named entity for more information
- Click on “More info” for properties of this entity
- Click on a property for searching more articles about resources with that property
- And see the result: all articles mentioning persons that have been married to Elizabeth Taylor

The screenshot shows a search results page from the Koninklijke Bibliotheek. At the top, there is a search bar with the query 'spektakelfilm and [romeinse keizer]' and a search icon. Below the search bar, it indicates '1 of 28 results for spektakelfilm and [romeinse keizer]'. The results are listed in a table with 10 entries. Each entry includes a number, a title, a list of related entities, and a date. A red arrow points to the 6th result, 'Spektakelfilm „De mantel“ uit 1953', which is highlighted. Below the list, there are two buttons: 'Show timeline' and 'Download timeline'.

Number	Title	Related Entities	Date
1	„De val van Rome” film waarin heel wat te beleven is	Rome · Commodus · Nero · Caligula · Samuel Bronston · Sophia Loren ...	1965/01/02
2	Luxor, Sittard DE VAL VAN ROME	Luxor · Sittard · Commodus · Livius · Kelten · Anthony Mann ...	1965/01/29
3	Quo Vadis	Quo Vadis · Amerikaanse · Robert Taylor · Deborah Kerr · Peter Ustinov · Patricia Laffan ...	1984/12/28
4	Astoria, Brunssum City, Nieuwenhagen VAL VAN ROME	Astoria · Romeinse · Marcus Aurelius · Livius · Lucilla · Commodus ...	1965/01/22
5	...EN DAT VAN SAMUEL BRONSTON	Samuel Bronston · Komes · Anthony Mann · Cleopatra · Rome · Charles Romes ...	1964/07/03
6	Spektakelfilm „De mantel” uit 1953	Richard Burton · Rome · Caligula · Tiberius Claudius Nero · Palestina · The Robe ...	1975/05/28
7	In Den Haag	Bloedgeld · Flora · Misdaad · Theodosius I · Frans · Bourvil ...	1962/01/26
8	In Den Haas	Bloedgeld · Flora · Tortuga · Misdaad · Theodosius I · Frans ...	1962/01/26
9	Maastricht	Maastricht · David Lean · Robert Mitchum · Trevor Howard · Zweedse · Rome ...	1971/09/10
10	SSVC	Symphony · Richard Hickox · Europe · Amsterdam · Demetrius and the Gladiators · Romeinse ...	1988/08/20

[Show timeline](#) [Download timeline](#)

Navigation example

- Semantic query between [], in this case expand to all Roman Emperors
- Select “newspaper+” collection
- Select a result
- **Click on a linked named entity for more information**
- Click on “More info” for properties of this entity
- Click on a property for searching more articles about resources with that property
- And see the result: all articles mentioning persons that have been married to Elizabeth Taylor

The screenshot shows the Koninklijke Bibliotheek website interface. At the top, there is a search bar with the query "spektakelfilm and [romeinse keizer]" and a dropdown menu for "Newspapers +". Below the search bar, it indicates "6 of 28 results for query spektakelfilm and [romeinse keizer]".

The "Enrichments" section is expanded, showing a list of related entities: "Richard Burton", "Rome", "Caligula", "Tiberius Claudius Nero", "Palestina", "The Robe", "Victor Mature", and "Capri (eiland)". A red arrow points to the "Caligula" entity.

Below the enrichments, there are sections for "Services", "Text", and "Details".

The main content area displays a newspaper clipping titled "Spektakelfilm „De mantel” uit 1953". The text of the clipping reads: "Van 20.40 tot 22.50 uur gaat vanavond over het tweede net Henry Koster's speelfilm uit 1953 „The robe” (De mantel) met Richard Burton, Jean Simmons, Victor Mature e.a. Het verhaal speelt zich af in het oude Rome, waar Caligula heer en meester is. Zijn aanstaande schoonvader, de oude keizer".

At the bottom of the page, there is a JavaScript request: `javascript:jsonrequest('http://tomcat.kbresearch.nl/links/nir?id=DBP:Richard_Burton_(acteur)&callback=showResource')`.

Navigation example

- Semantic query between [], in this case expand to all Roman Emperors
- Select “newspaper+” collection
- Select a result
- Click on a linked named entity for more information
- **Click on “More info” for properties of this entity**
- Click on a property for searching more articles about resources with that property
- And see the result: all articles mentioning persons that have been married to Elizabeth Taylor

The screenshot shows the Koninklijke Bibliotheek website interface. At the top, there is a search bar with the query "spektakelfilm and [romeinse keizer]" and a search button. Below the search bar, a navigation menu displays "6 of 28 results for query spektakelfilm and [romeinse keizer]". The main content area features a list of search results, with the first result highlighted. This result includes a thumbnail image of a man in Roman attire (Richard Burton) and a snippet of text from a newspaper article: "akelfilm mantel” 53 tot 22.50 uur gaat van- avond over het tweede net Henry Kos- ter's speelfilm uit 1953 „The robe” (De mantel) met Richard Burton, Jean Simmons, Victor Mature e.a. Het ver- haal speelt zich af in het oude Rome, waar Caligula heer en meester is. Zijn aanstaande schoonvader, de oude keizer Mar- rius, regeert vanuit Capri. Mar-". A red arrow points to the "More info" button located below the thumbnail image. The "More info" button is highlighted in blue. A dropdown menu is visible, listing various search engines and databases: "DBpedia", "VIAF", "Wikidata", "Catalogus", "KB thesaurus", and "DBpedia". The "More info" button is also highlighted in blue.

Navigation example

- Semantic query between [], in this case expand to all Roman Emperors
- Select “newspaper+” collection
- Select a result
- Click on a linked named entity for more information
- Click on “More info” for properties of this entity
- **Click on a property for searching more articles about resources with that property**
- And see the result: all articles mentioning persons that have been married to Elizabeth Taylor

KB Newspapers + spektakelfilm and [romeinse keizer]

6 of 28 results for query spektakelfilm and [romeinse keizer]

Enrichments ^

Richard Burton Rome Caligula Tiberius Claudius Nero Palestina The Robe Victor Mature Capri (eiland)

Search this entity

- > DBpedia
- > VIAF
- > Wikidata
- > Catalogus
- > KB thesaurus
- > DBpedia

Click on one of the property values below to search for newspaper articles about entities with the same property value.

religie	atheïsme
land van nationaliteit	Verenigd Koninkrijk
is een	mens
taalbeheersing	Engels
taalbeheersing	Welsh
beroep	acteur
achternaam	Burton
echtgenoot	Elizabeth Taylor
moedertaal	Welsh

JavaScript:wdExpansion("P26",Q34851)

akelfilm
mantel”
53
tot 22.50 uur gaat van-
et tweede net Henry Kos-
uit 1953 „The robe” (De
Richard Burton, Jean
tor Mature e.a. Het ver-
ich af in het oude Rome,
heer en meester is. Zijn
nvader, de oude keizer
eert vanuit Capri. Mar-

Navigation example

- Semantic query between [], in this case expand to all Roman Emperors
- Select “newspaper+” collection
- Select a result
- Click on a linked named entity for more information
- Click on “More info” for properties of this entity
- Click on a property for searching more articles about resources with that property
- **And see the result: all articles mentioning persons that have been married to Elizabeth Taylor**

KB Newspapers + [P26=Q34851] 1 of 4566 results for ID26...

1	Elizabeth Taylor weer eens gescheiden Elizabeth Taylor · Amerikaanse · Senator · Marc War...	
2	Liz Taylor weer gescheiden Liz Taylor · Elizabeth Taylor · Larry Fortensky · Kelly Taylor · New York Post · Richard Burton ...	1995/09/01
3	Elizabeth Taylor verloofd Elizabeth Taylor · Mexicaanse · Philadelphia · Kelly Taylor · Santiago Luna · Britse ...	1983/08/11
4	Het heilige monsten van Hollywood Hollywood · Garbo · Charles Boyer · Cleopatra · Liz Taylor · Zwitserse ...	1978/02/10
5	Cannes bloeit weer Filmfestivalkaartjes voor 250 gulden! Cannes · Amerikaanse · Mike Todd · Franse · Festival ·	1957/05/03
6	Voor oog en oor „The old maid and the thief”: amusante opera van Menotti Gian Carlo Menotti · KRO · NBC · Linda de Vries · Laetitia · Michael Todd ...	1961/10/12
7	PASSAGIERS. Batavia · Jan Brewer · Universiteit van Californië · Berkeley · Bradley Cooper · Vitamine E · Heavy metal ...	1935/06/07
8	Scheepsberichten. Op ten Noort · Soerabaia · Semarang · Calcutta · Ned · Josefina ...	1935/05/13
9	Scheepsberichten Pieter Mijer · Dries van Agt · Akkerman · Sibolga · Tetje Mierendorf · Hamburg ...	1935/04/03
10	„Feestje” met 18.000 gasten in New York Gastheer-miljonair Mike Todd maakte er dolle avond van Feestje · New York · Mike Todd · Madison Square Garden · Spirited Away · Sweeney Todd ...	1957/10/18

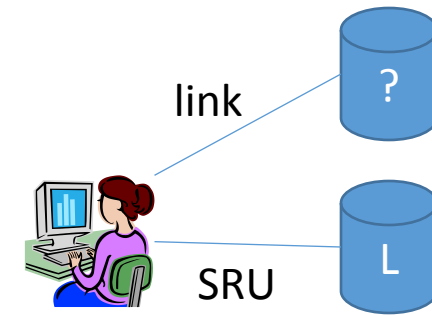
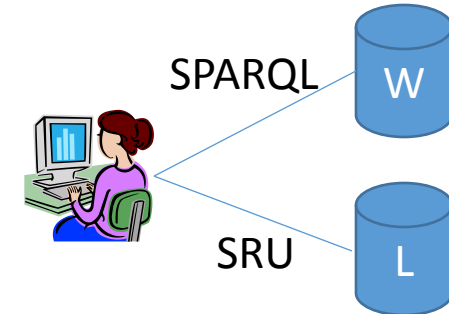
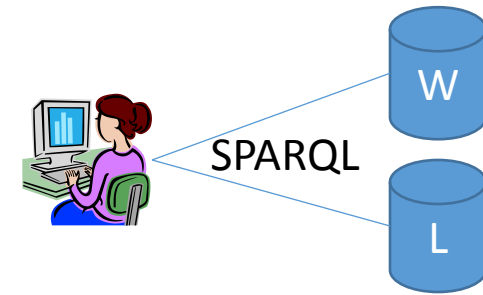
Show timeline Download timeline

spouse=Elizabeth Taylor

Coverage, approaches and environments

Usage with different infrastructural impact

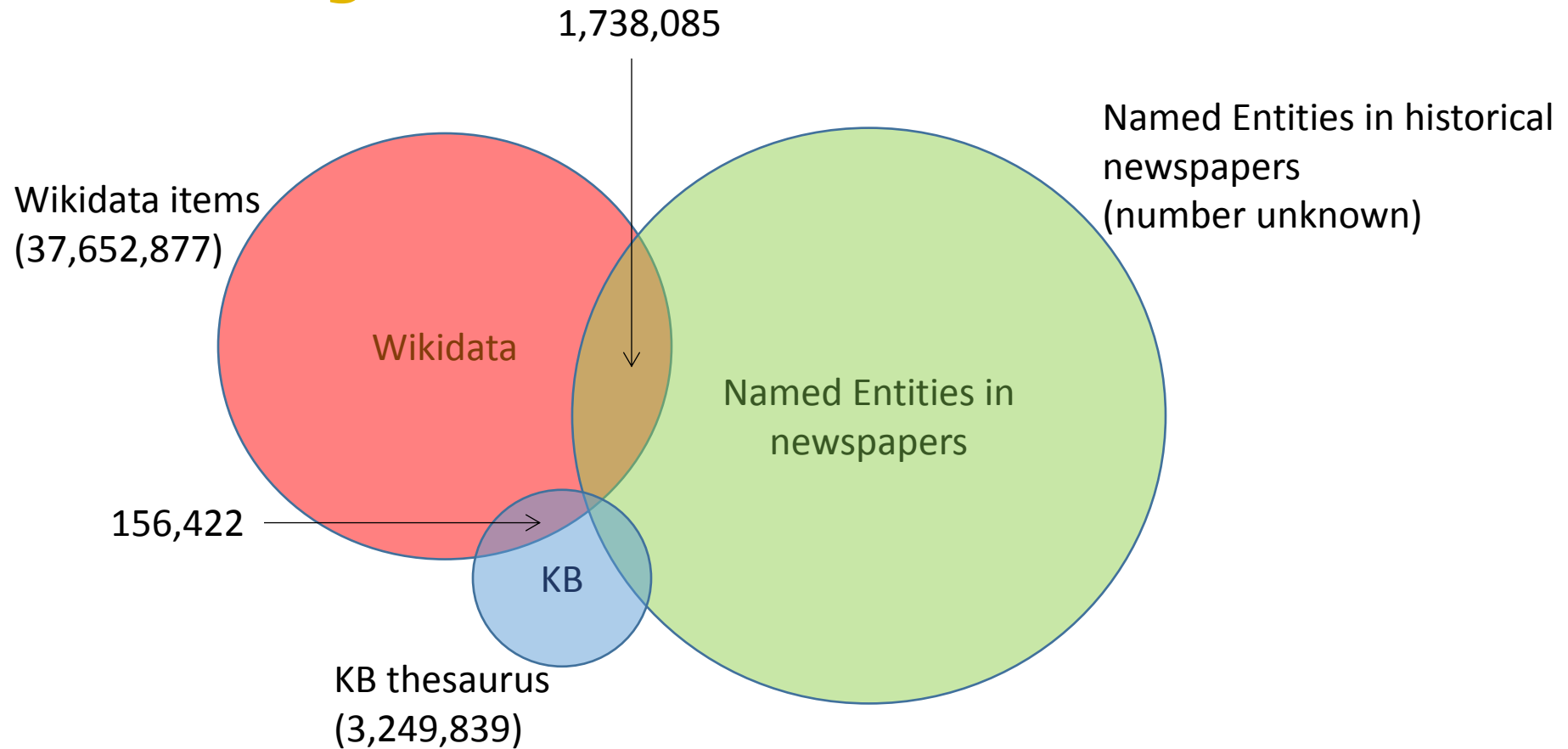
- Federated SPARQL queries to search in a local database and Wikidata
- Mixed use of conventional queries and SPARQL: using the output of SPARQL as conventional query input using Wikidata identifiers
- Generate “just in case links” to construct a query for an external database (lowest implementation barrier)



Mixed use of resource identifiers

- All items are in a local thesaurus and in some cases the Wikidata item contains the local identifier
- Some items in the local thesaurus contain a link to Wikidata
- Mixed use: Items in bibliographic records link to Wikidata or to the local thesaurus
- All items in bibliographic records link to Wikidata
- The local thesaurus is kept for administrative purposes or as backup. New items might be pre-staged in the local administrative thesaurus

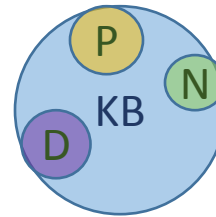
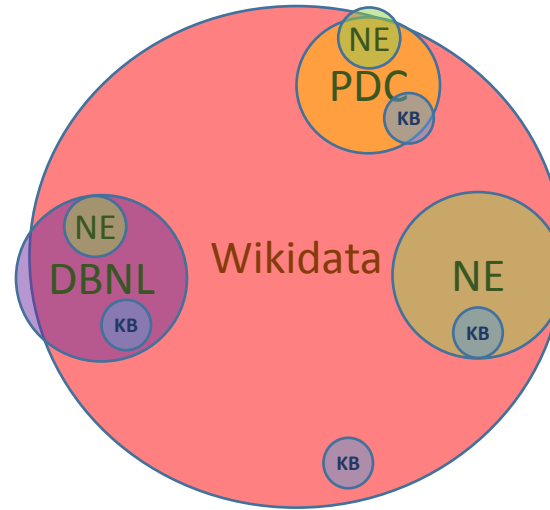
Current coverage

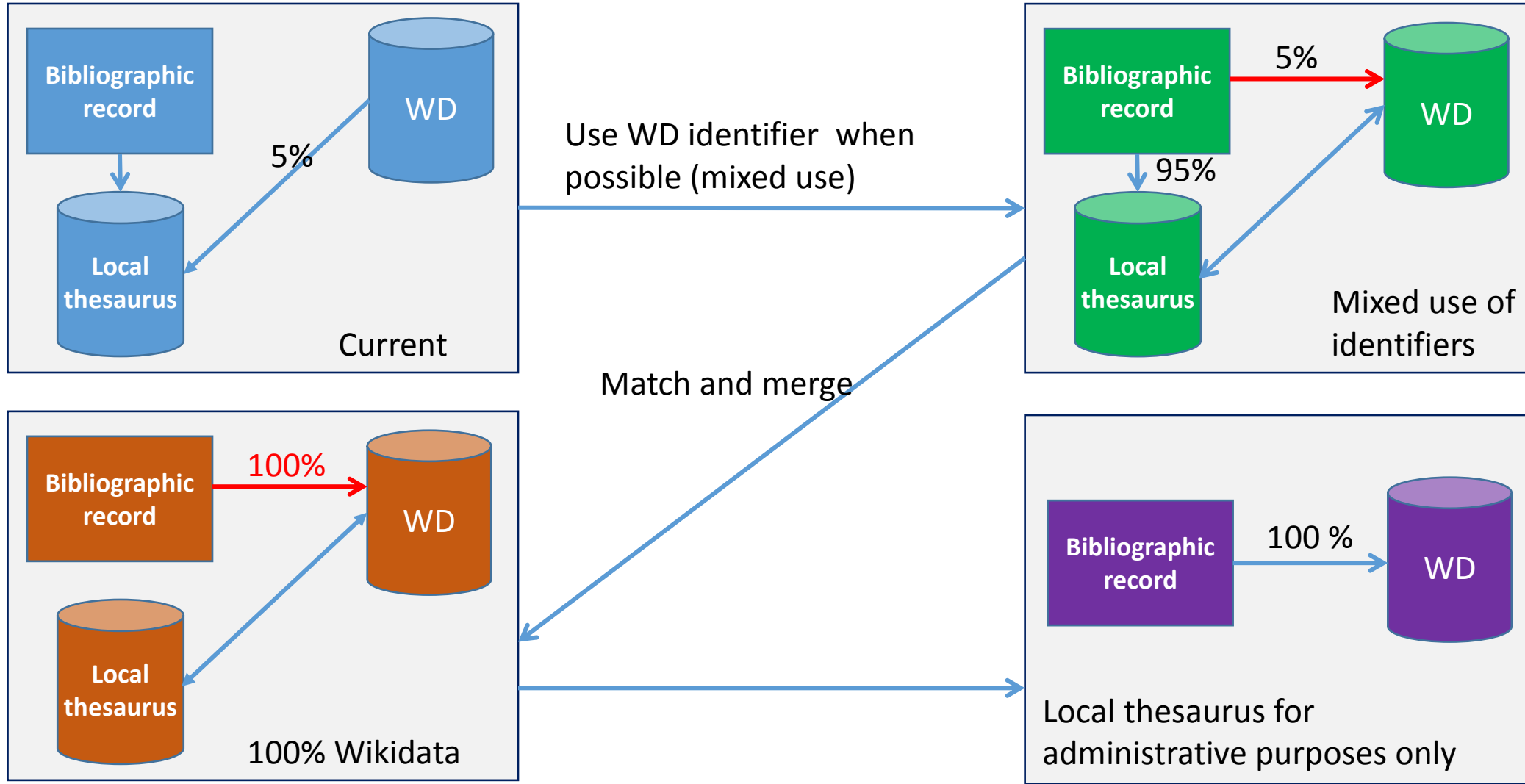


Fragmented landscape of linked and unlinked resources

Institutions maintain links to more than one collection with different coverage. In case of the Koninklijke Bibliotheek, for example:

- Library catalogue
- Newspaper collection
- Parliamentary Papers
- DBNL
- Etc...





Pros and cons of Wikidata as universal thesaurus

Pros

- Lower barrier to connect different databases
- It doesn't require an advanced infrastructure to benefit
- Less maintenance
- Coverage of more domains
- Potentially richer set of properties

Possible objections

- Libraries perceive it as losing control over their thesaurus: unauthorized users may change items
- Different organizations or countries may have a different view on specific items
- There is a risk of duplicate entries for new items
- What if Wikidata disappears?

Summary and conclusions

- Wikidata can serve as universal library thesaurus
- Using Wikidata as single universal thesaurus facilitates identification of entities across organizations and sharing properties
- Replacing thesaurus identifiers with Wikidata identifiers in bibliographic records can be done gradually
- When the transition is complete thesauri can be kept for administrative purposes
- The use of Wikidata identifiers is not restricted to SPARQL: identifiers can also be indexed and used in conventional queries

um odoratarumque nonnullarum
Purgantium historiae libri IV,
riae pemptades sive libri XXX.
et groot Cruydtboeck, hetwelk
zouden uitgeven en in gansch
Plantin had zich langzamer-
an al de houtblokken, die bij
kers voor andere herbariums
gd bij die van zijn eigen uit-
lustratiemateriaal uit zonder
het Museum Plantin Moretus
van Pieter van der Borcht, in
overwegend en niemand min-
roote kunst, die de Mechelsche
ag legde. „Het is mij”, schrijft
en tijd in het bekijken van de
uydtboeck van Dodoens ver-
ik, daarna in den tuin rond-
en bloem en heester geschaard
eidscher dos meende te zien
der takken, de levendigheid
ormatie der wuivende kronen,
voor mij geopenbaard, en ik
ren in den eeuwig afwisselen-
as. De prenten van Dodoens
epping duidelijker en dieper
as de trouwe illustratie van
nter der levende natuur ge-

tgever door de banden eener
den. Toen Dodoens in 1584
gedurende een paar jaren
was, liet Plantin dit met
emeenschappelijken vriend

, als blijk van vriendschap,
verbeterden en vermeerder-
dschen tekst van zijn ge-

schiedenis der
in 1616 do
in 1618 d
werd uit
Met Cl
Deze gel
Gent en
geweste
te Wit
ging. T
van R
de jo
plant
op de
hij te
wege
bloed
de l
ver
na
vr
tr
k
t
g

met
in
best
dijn
wafte
m
am
w
ve
van
du
ben
sijn
mijn
vuechten
deme
e steen
ten

Actueel

- KB pilotproject voor gebouwspaspoort
 - › Green Deal Circulaire Gebouwen echt van start
- 'Jeugdige overmoed'. Els Stronks over denkbeelden over jongeren in digitale teksten
 - › Lezing van de KB-fellow
- Bibliotheekcollecties in het netwerk
 - › Oproep: heeft u een digitale bibliotheekcollectie?
- Europa, niet het Universum
 - › Lees de KB-blog
- › Alle actuele artikelen

Any questions?

theo.vanveen@kb.nl

<http://www.kbresearch.nl/xportal>

145

34 plaats, in uitvoerige ac-
d (nots. Knoll 1734 6 April,
De nu volgende jaren staan in
ise in het land en op de Haag-
de jaren 1737 tot 1747 staakte
st „door ongelukkige tijden en
ppers”. Ten einde aan het vervol-
den uitweg om — gebruik ma-
n gecontroleerde „aucties onder de
erkochte fondsen onderling in veiling-
obligaties te betalen, en die vervol-
e te transporteren aan de bankiers.
nd, te transporteren van geld. Hierbij schijnt
hebben gespeeld. Hij en zijn zoon Pieter,
ing, zijn niet alleen de voornaamste koo-
ies, die tegen 1740 uitliepen op de „Compag-
ne de vier anderen benoemden om toezicht te hou-
de veilingen, maar ook duidelijk het middel-
de majeure, hun speculatiën faalden. In 1744
Gosse, Block, Swart, Beauregard, Moetjens.
e de vier anderen moesten zich laten welgeval-
meischers personen beropen om toezicht op zijn
oedel en bewerken in Duitschland te liquideeren en
gegeven zijn zaken in Duitschland te liquideeren en
cht hij het tot stand dat heel zijn bezit, huis en
ligaties natuurlijk, werd overgebracht op zijn
Compagnie. — De oudste zoon, Henri Albert was
de firma was intusschen uitgebreid
ave chez H. A.
Libraire
Nicoolaas