Climbing the Tower of Babel

Challenges and Opportunities in Multilingual Data for the Digital Humanities

7th LIDER Roadmapping Workshop
Linked Data for Digital Humanities and Linguistics

20 October 2015, Madrid

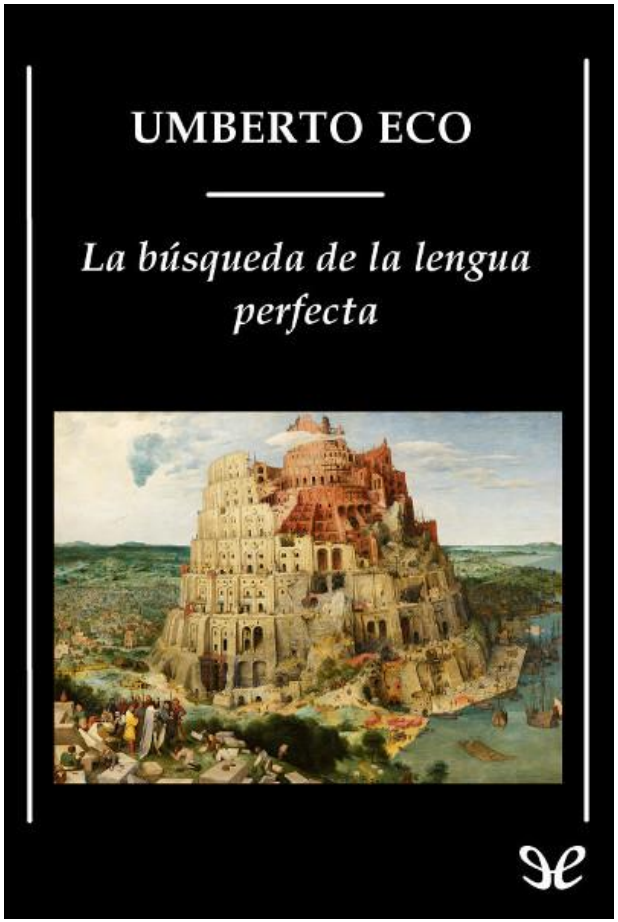
Clemens Neudecker
Staatsbibliothek zu Berlin
@cneudecker

lider

Staatsbibliothek zu Berlin
Preußischer Kulturbesitz

La búsqueda de la lengua perfecta

- Umberto Eco, 1994
- „I certainly will never advise to follow the bizarre thought presented here and dream of a universal language“



How many languages are there?

- The Holy Bible, 1. Mose 10: **72 (70)**
- Max Planck Institute for Evolutionary Anthropology: **6500 – 7000**
- ISO 639-3: **7704** (ISO 639-2: 450)
- Google Translate supported: **90**
- Europeana content: currently **50**

Metadata

- To enjoy a painting or music on Europeana, no special language skills are required?

- Wrong!
  - Cultural objects are described using metadata
  - Metadata comes in different languages (country of origin of the data provider)
  - Most often metadata does not have language information
  - How to still find what you are looking for?

Problem: Metadata

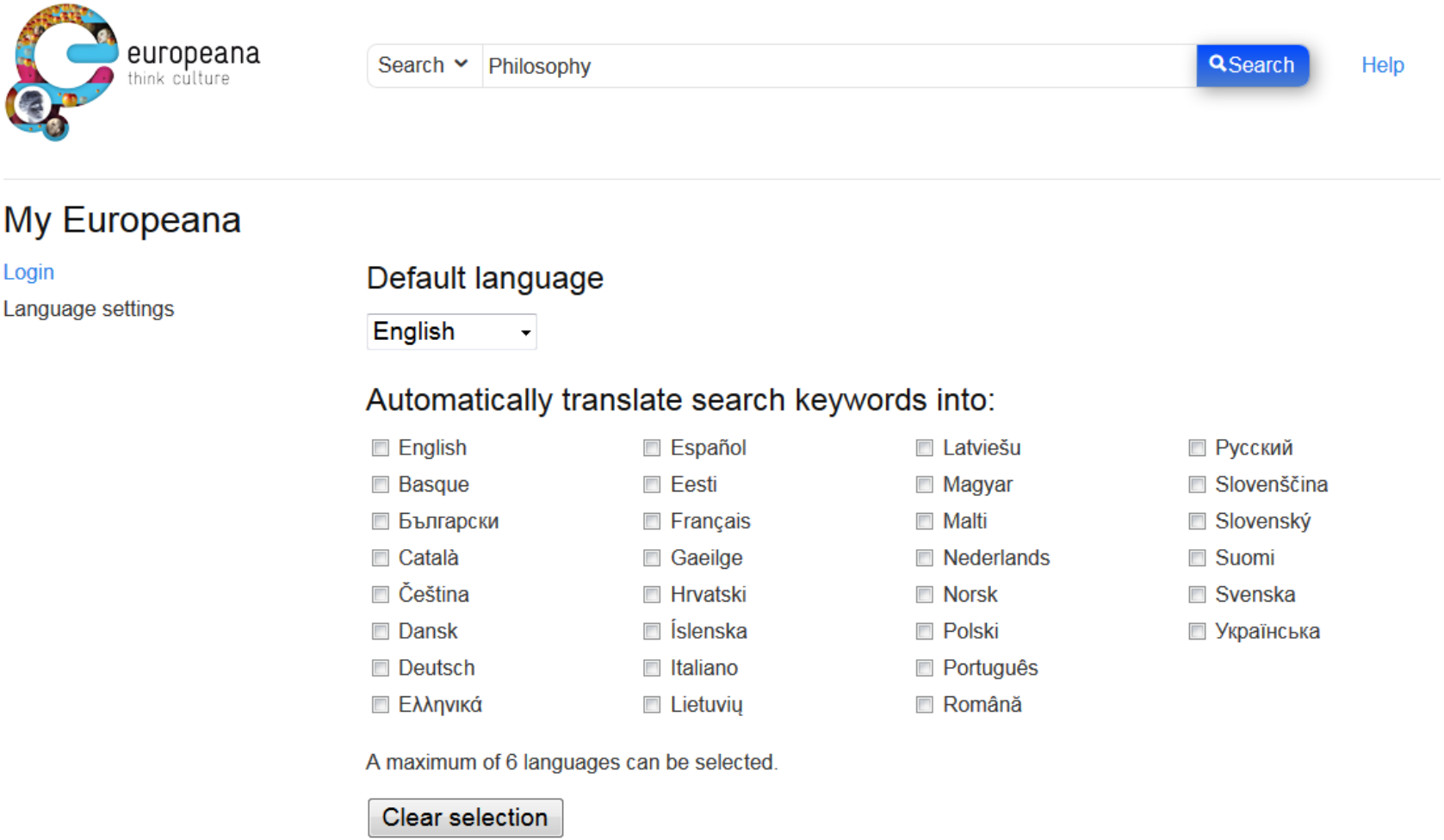
- Example: Subject „Philosophy“
  - Philosophie
  - Filosofía
  - Filosofie
  - Filosofija
  - Heimspeki
  - Филозофија
  - Etc.

Metadata: Option 1

- Indicate the language of the metadata
- This supports the use of translation or mapping tools to find the correct term in other languages/controlled vocabularies
- Example:

```
<subject language=„English“>Philosophy</subject>
```

Europeana Query Translation



Europeana Query Translation

- How it works:

`http://www.europeana.eu/portal/search.html?query=`**`Philosophy`**

Europeana Query Translation

http://**[language]**.wikipedia.org/w/api.php?action=query&prop=langlinks&format=json&titles=**[query term]**

→ {"lang":"**de**", :"**Philosophie**"},
{"lang":"**es**", :"**Filosofía**"}

Europeana Query Translation

http://www.europeana.eu/portal/search.html?query=**Philosophy&Philosophie&Filosofía**

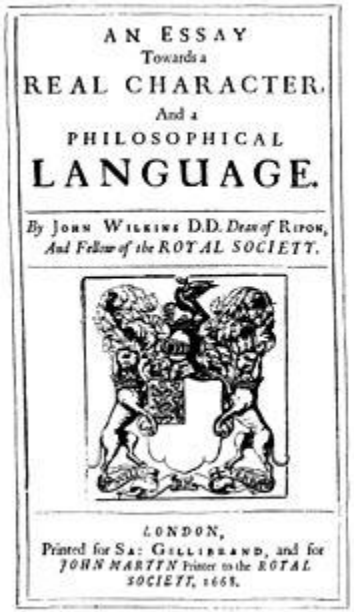
(simplified for illustration purposes –
above query does not really work,
as the query expansion is done internally)

Europeana Query Translation

- Read more:
  - Query Translation in Europeana: http://journal.code4lib.org/articles/10285
  - Improving Europeana Multilingual Search: http://blog.europeana.eu/2014/08/improving-search-across-languages/

El idioma analítico de John Wilkins

- Jorge Luis Borges, *Otras Inquisiciones*


AN ESSAY
Towards a
REAL CHARACTER,
And a
PHILOSOPHICAL
LANGUAGE.
By JOHN WILKINS D.D. *Dean of* RIPON,
And Fellow of the ROYAL SOCIETY.
LONDON,
Printed for SA: GILLIBRAND, and for
JOHN MARTYN Printer to the *ROYAL*
SOCIETY, 1668.

- „Theoretically, it is not impossible to think of a language where the name of each thing says all the details of its destiny, past and future“

Metadata: Option 2

- Even better: Use a language-independent identifier for subject classification (e.g. Library of Congress, WikiData, DDC)

- Example:

```
<subject id=„loc“>sh85100849</subject>

<subject id=„wikidata“>Q5891</subject>
```

Two examples

- Europeana 1914 – 1918
  http://www.europeana1914-1918.eu/

- Europeana Newspapers
  http://www.europeana-newspapers.eu/

Europeana 1914 - 1918

- In fact, three projects:
  - Europeana Collections 1914-1918
    400.000 digitised items from World War I
  - Europeana 1914-1918
    User generated content from World War I
  - European Film Gateway 1914
    740 hours of film related to World War I

- How to present these as a uniform collection?

Europeana 1914 - 1918

- Analysis of subject classifications available at content holding institutions, e.g. catalogues

ARK Sachkatalog (1501-1955) der Staatsbibliothek zu Berlin - im Aufbau

ARK ▸ Geschichte · Ethnographie · Geographie ▸ Krieg 1914

▸ Allgemeines	Weltkr. 1 - Weltkr. 44
▸ Ursprung · Vorgeschichte	Weltkr. 45 - Weltkr. 56 e Ungarische
▸ Kriegsverlauf	Weltkr. 57 - Weltkr. 151
▸ Der Krieg in politischer Hinsicht	Weltkr. 152 - Weltkr. 265
▸ Geistiger Krieg	Weltkr. 266 - Weltkr. 280 Schweizerische
▸ Die Kriegführenden	Weltkr. 281 - Weltkr. 345 h Zank
▸ Kriegsgebiete	Weltkr. 346 - Weltkr. 379
▸ Das deutsche Inland	Weltkr. 380 - Weltkr. 392
▸ Der Krieg in militär-technischer Hinsicht	Weltkr. 393 - Weltkr. 452
▸ Krieg und Technik	Weltkr. 453

Europeana 1914 - 1918

- Ranking of most frequent subjects

Subject Heading	Count
World War, 1914-1918--Campaigns	4307
World War, 1914-1918--Trench warfare	2990
World War, 1914-1918--Transportation	2171
World War, 1914-1918--Caricatures and cartoons	2013
World War, 1914-1918--Serbia	1755
…	…

Europeana 1914 - 1918

- Mapping subjects to LoC identifiers

Subject Heading	LoC identifier
World War, 1914-1918--Campaigns	sh85148240
World War, 1914-1918--Trench warfare	sh2008113804
World War, 1914-1918--Transportation	sh2008113817
World War, 1914-1918--Caricatures and cartoons	sh2010119466
World War, 1914-1918--Serbia	Sh2008113856
…	…

Europeana 1914 - 1918

- Enrichment of metadata with LCSH identifiers

```
▼<mods:language>
    <mods:scriptTerm authority="ISO15924" type="code">217</mods:scriptTerm>
  </mods:language>
▼<mods:subject authority="lcsh">
    <mods:topic>sh2010007643</mods:topic>
  </mods:subject>
▼<mods:subject authority="lcsh">
    <mods:topic>sh85106971</mods:topic>
  </mods:subject>
▼<mods:subject authority="lcsh">
    <mods:topic>sh85148236</mods:topic>
  </mods:subject>
  <mods:accessCondition type="use and reproduction">Public Domain</mods:accessCondition>
 </mods:mods>
</mets:xmlData>
```

Europeana 1914 - 1918

- Translation of all subjects

1	Europeana 1914-1918 concept URI	Dbpedia	LCSH	Europeana 1914-1918 concept Label			
2	skos:Concept rdf:about	owl:sameAs	owl:sameAs	skos:prefLabel xml:lang="en">	skos:prefLabel xml:lang="fr">	skos:prefLabel xml:lang="de">	skos:prefLabel xml:lang="da">
3	data.europeana.eu/concept/gwa/keyword/conscientious_objection			Conscientious Objection	Objecteurs de conscience	Kriegsdienstverweig	Militærnægtelse
4	data.europeana.eu/concept/gwa/keyword/manufacture			Manufacture	Manufacture	Produktion	Produktion
5	data.europeana.eu/concept/gwa/keyword/transport			Transport	Transport	Transport	Transport
6	data.europeana.eu/concept/gwa/keyword/medical			Medical	Médical	Gesundheitswesen	Medicinsk
7	data.europeana.eu/concept/gwa/keyword/gas_warfare			Gas warfare	Guerre chimique	Gaskrieg	Krigsførelse med gas
8	data.europeana.eu/concept/gwa/keyword/tanks_and_armoured_fighting_ve			Tanks and Armoured Fighting Vehicles	Tanks and and véhicules armés	Panzer	Tanks og pansrede køretøjer
9	data.europeana.eu/concept/gwa/keyword/trench_life			Trench life	Vie dans les tranchées	Leben im Schützengraben	Livet i skyttegravene
10	data.europeana.eu/concept/gwa/theatre/aerial_warfare			Aerial Warfare	Guerre Aérienne	Luftkrieg	Luftkrig
11	data.europeana.eu/concept/gwa/theatre/naval_warfare			Naval Warfare	Marine	Seekrieg	Søslag
12	data.europeana.eu/concept/gwa/keyword/artillery			Artillery	Artillerie	Artillerie	Artilleri
13	data.europeana.eu/concept/gwa/keyword/recruitment_and_conscription			Recruitement and Conscription	Recrutement et conscription	Rekrutierung und Wehrpflicht	Rekrutering og værnepligt
14	data.europeana.eu/concept/gwa/keyword/prisoners_of_war			Prisoners of War	Prisonniers de guerre	Kriegsgefangene	Krigsfanger

Treffer für Ihre Suche

Quelle: aus Archiven, Bibliotheken und Museen
Subject: Propaganda

Verfeinern Sie

Schlagwörter hir

Quelle
- Alle Dokument
- von Privatperso
- **Aus archiven,**

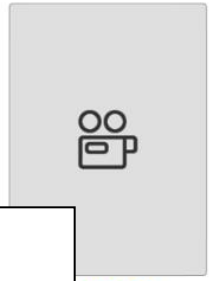
Subject
- **Propaganda**
- Krieg und film (
- Patriotismus (2
- Karikaturen un
- Krieg und gese
- Soldaten (17)
- Belgien (13)
- Kriegsanleihen
- Zivilisten im kri
- Infanterie (12)
- Nationalismus
- Massenmedien
- Ungarn (10)
- Grabenkrieg (1
- Werbung (10)
- Militärzeremoni
- Gebäude -- krie
- Rumänien (9)
- Serbien (9)
- Frauen und krie
- Autonomie- und unabhängigkeitsbewegungen (9)
- Erster weltkrieg 1914-1918 -- denkmäler (9)
- Mobilisierung der streitkräfte (8)
- Städtisches leben (8)
- Russland (7)
- Kinder und krieg (7)

Ergebnis Ihrer Suche nach (447 Treffer)

juram antului
lor Legiunii

Pro Italia, preferite sempre i prodotti nazionali

Central Institute for the



alor civile
915–1915

zione Cineteca
Italiana

The European
n Gateway

Friede wird
ensfragen und
ntworten
ankenberg und
dwigsdorf

sbibliothek zu
- Preußischer
ulturbesitz

Subject
- Battlefields (4388)
- Campaigns (4307)
- Trench warfare (2
- Transportation (21
- Caricatures and ca
- Serbia (1469)
- Religious aspects
- Literature and the
- War and society (1
- Soldiers (1252)
- Aerial operations (
- Medical care (102
- Prisoners of war (9
- Estonia -- history - independence, 19
- Campaigns -- italy
- Technology (774)

Subject
- Schlachtfelder (4388)
- Feldzüge (4307)
- Grabenkrieg (2984)
- Verkehr (2169)
- Karikaturen und cartoons (1950
- Serbien (1469)
- Der krieg in religiöser hinsicht (1429)
- Krieg und literatur (1335)
- Krieg und gesellschaft (1300)
- Soldaten (1252)
- Luftkampf (1093)
- Erster weltkrieg--medizinische versorgung (1020)
- Kriegsgefangene (948)
- Estland -- geschichte -- unabhängigkeitskrieg 1918-192 (918)
- Feldzüge -- italien (887)
- Technik (774)

Subject
- Luoghi di battaglia (4388)
- Operazioni militari (4307)
- Guerra di trincea (2984)
- Trasporti (2169)
- Caricature e fumetti (1950)
- Serbia (1469)
- Ruolo [della] religione (1429)
- Letteratura -- guerra mondiale 1914-1918 (1335)
- Società -- effetti [della] guerra mondiale 1914-1918 (1300)
- Soldati (1252)
- Operazioni aeree (1093)
- Assistenza sanitaria (1020)
- Prigionieri (948)
- Guerra di indipendenza estone, 1918-1920 (918)
- Operazioni militari - italia (887)
- Tecnologie (774)

Europeana Newspapers

- **Full text** collection of 12 million digitised newspaper pages from 23 European libraries
- Around 40 different languages overall
- Newspapers from 1618 - 1990 → historical spelling variants!
- www.theeuropeanlibrary.org/tel4/newspapers

Europeana Newspapers

- Content in Europeana Newspapers

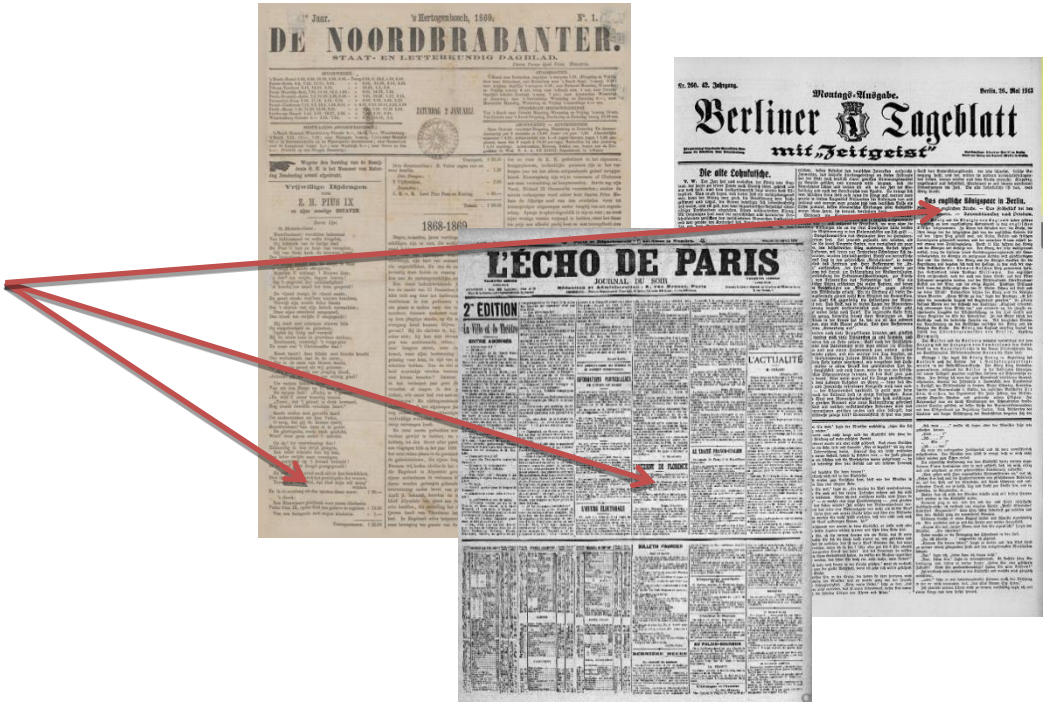
Europeana Newspapers

- 12 million newspaper pages = approximately 102,000,000,000 words!
- Impossible to translate everything to multiple languages
- But there are alternatives…

Europeana Newspapers

- What if it were possible to search for persons, locations, events, across languages?

Siege of Przemyśl



Europeana Newspapers

- Named Entity Recognition

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell-Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:
LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

- University of Stanford NER toolkit

Europeana Newspapers

- Named Entity Disambiguation

„Jordan“

- Comparison of context

Europeana Newspapers

- Named Entity Linking

Freebase

freebase.com/m/**054c1**

LIBRARY OF CONGRESS AUTHORITIES

lccn.loc.gov/**n92121379**

WIKIDATA

wikidata.org/wiki/**Q41421**

What if…

- All metadata in Europeana 1914-1918 had language-independent identifiers
- All entities in Europeana Newspapers had language-independent identifiers
- It should be possible to link the two distinct collections!

Research Questions

- This would allow for some very intersting digital humanities research questions, e.g.
  - How were World War I events covered in newspapers of different nations across Europe?
  - What were the relations between persons, places and events during World War I?

The Republic of Letters

- http://stanford.edu/group/toolingup/rplviz/rplviz.swf

Global Database of Events, Language and Tone

- http://www.gdeltproject.org/

Conclusion

- We need know-how and technologies for multilinugual linking of objects across cultural heritage organisations and digital collections

- We need guidelines and standards that support the creation and provision of metadata in cultural heritage objects as multilingual linked data

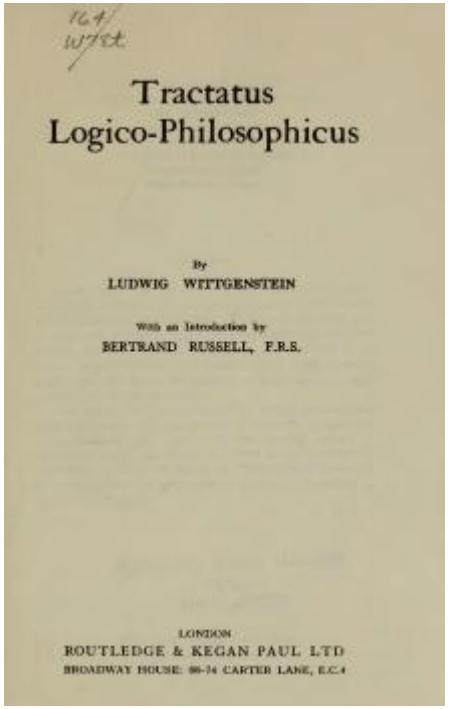
To follow up

- Europeana White Paper on Best Practices for Multilingual Access to Digital Libraries
- W3C Community Group Best Practices for Multilingual Linked Open Data
- Europeana Connect - Multilinguality

Tractatus Logico-Philosophicus

- Ludwig Wittgenstein, 1922

- Proposition 7: „Whereof one cannot speak, thereof one must be silent“



Thank you for you attention!

7th LIDER Roadmapping Workshop
Linked Data for Digital Humanities and Linguistics

20 October 2015, Madrid

Clemens Neudecker
Staatsbibliothek zu Berlin
@cneudecker

Staatsbibliothek zu Berlin
Preußischer Kulturbesitz