Open Science and Linked Open Data for Intra-Belgian Book Translations 1970-2020

Sven Lieber & Ann Van Camp

DHBenelux conference 1 june, 2022

KBR Koester de tijd Protégeons le temps

KU LEUVEN

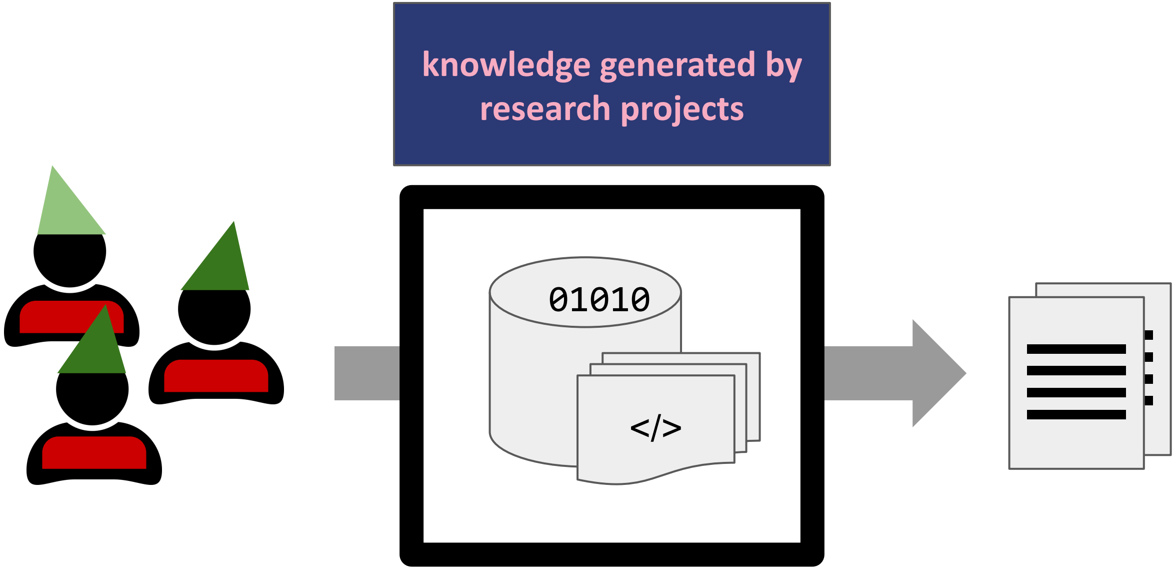
UCLouvain

It is important ***how*** something was achieved!

Underpants gnomes © South Park, Comedy Central

PHASE 1
PHASE 2
PHASE 3
Collect underpants
?
Profit

Performed work is often undocumented and thus unrecognised



knowledge generated by research projects

Researchers

Data processing and data

Research articles/ analyses

Edmond, Jennifer, and Francesca Morselli. "Sustainability of digital humanities projects as a publication and documentation challenge." *Journal of Documentation* (2020).

Use case: translation flow project BELTRANS

Different data sources as challenge

How do we integrate data? [Example] GitHub

Discussion of our approach

Take home message:

Document your valuable data processing such that it can be recognised, ideally make it transparent/open source!

Use case: translation flow project BELTRANS

Different data sources as challenge

How do we integrate data? [Example]

Discussion of our approach

BELTRANS - Intra-Belgian book translations 1970-2020

Financed by

belspo

BRAIN.be 2.0 programme
Pillar 2 “Heritage science”
National thematic project

Studying intra-Belgian translation flows and their context

4 Years project (2021-2025)

Cooperation with KU Leuven and UCLouvain

3 PhD students working on it

KBR responsible for data management

https://www.kbr.be/en/projects/beltrans/

KBR Koester de tijd Protégeons le temps

KU LEUVEN

UCLouvain

The scope of book translations in the BELTRANS project

BEL TRANS

intra-Belgian	author / illustrator / scenarist / publishing director
location	published anywhere in the world
literary	literary genres (novel, youth literature, comics, poetry) + literary non-fiction (mainly history books)
translations	FR-NL / NL-FR
time-period	1970-2020, since the creation of the Communities

.be

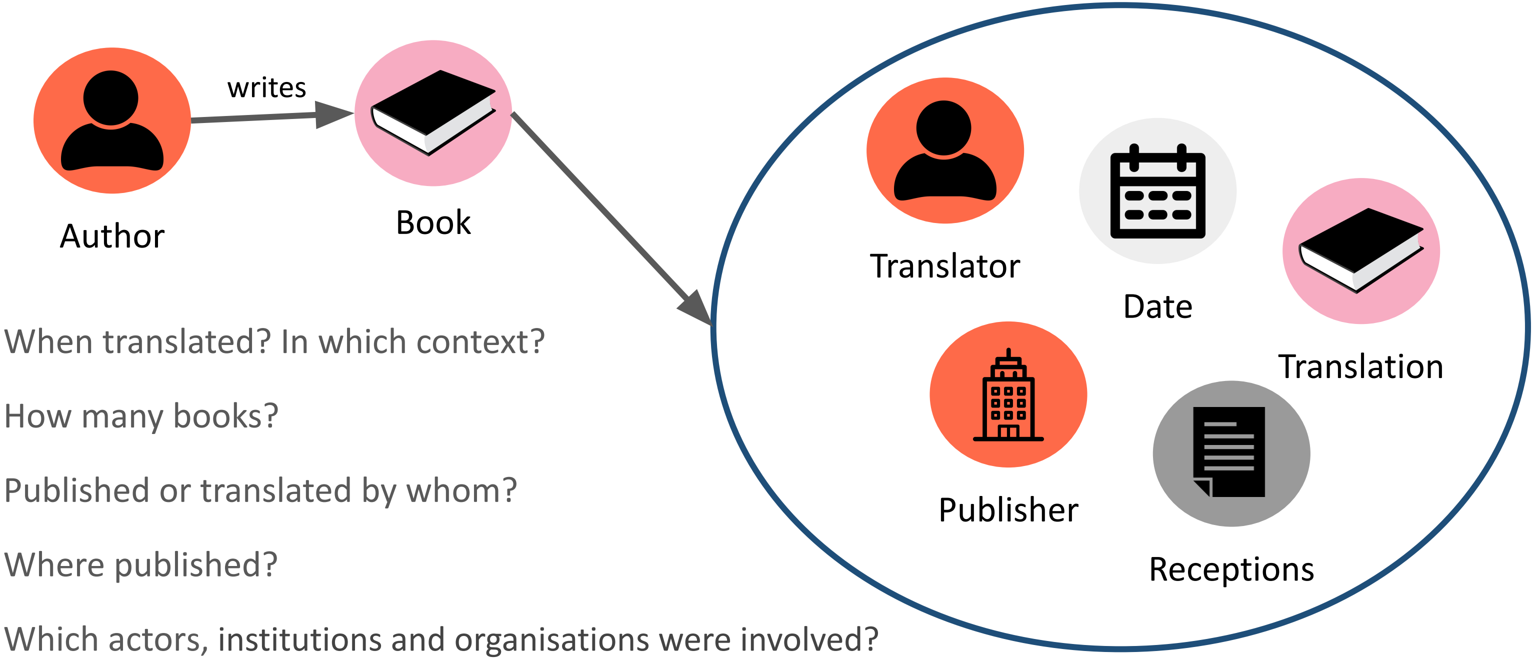
Vlaamse overheid

FÉDÉRATION WALLONIE-BRUXELLES

Ostbelgien

KBR Koester de tijd Protégeons le temps

BELTRANS is about *translation flows*



When translated? In which context?

How many books?

Published or translated by whom?

Where published?

Which actors, institutions and organisations were involved?

How were the translations received in the other language community?

...

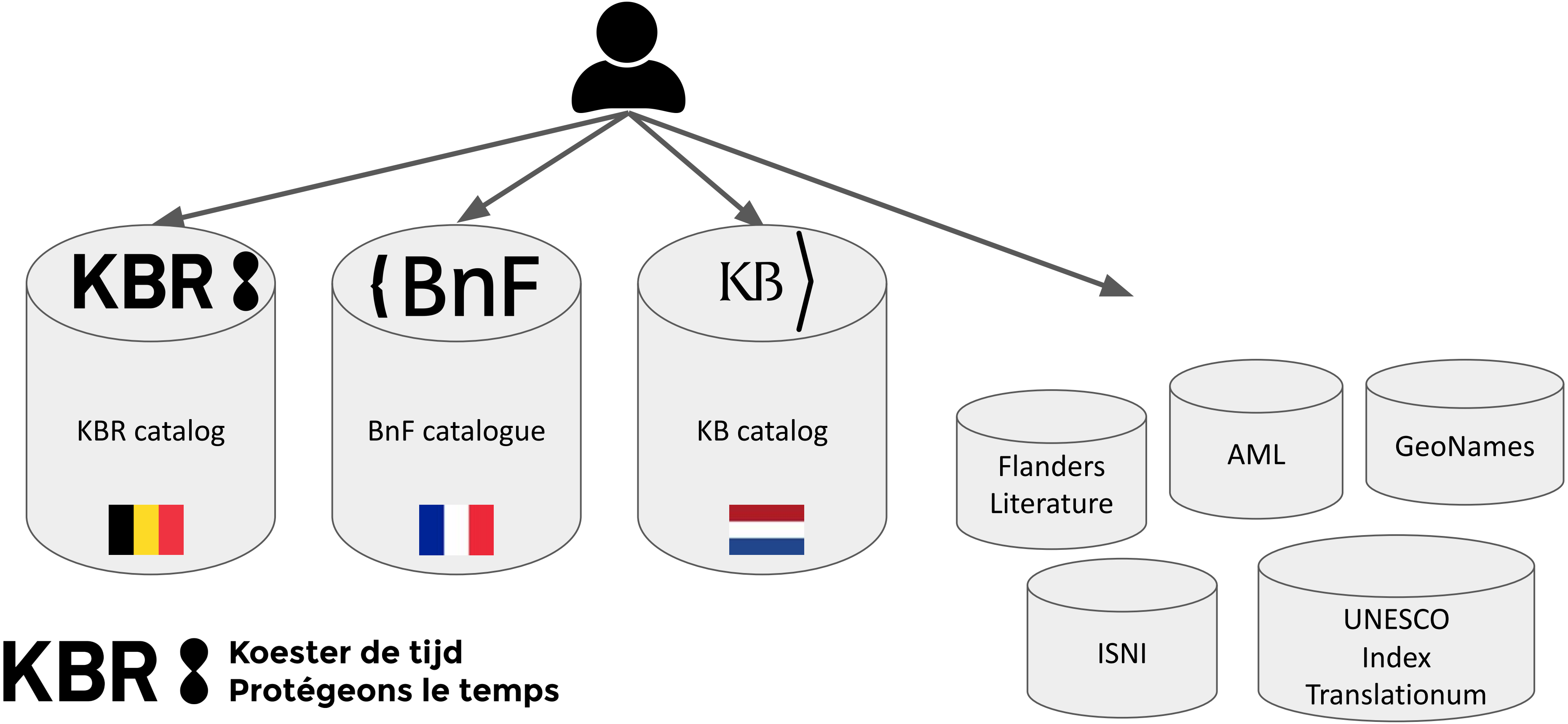
Use case: translation flow project BELTRANS

Different data sources as challenge

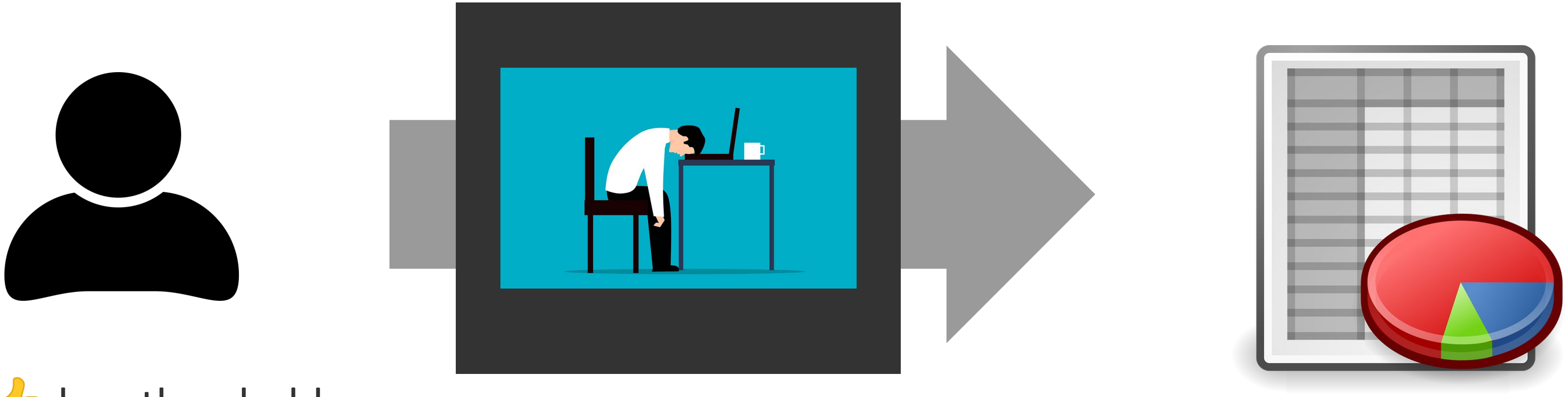
How do we integrate data? [Example]

Discussion of our approach

Different data *sources* in different *formats* of different *quality*



Manual way of integrating has many disadvantages



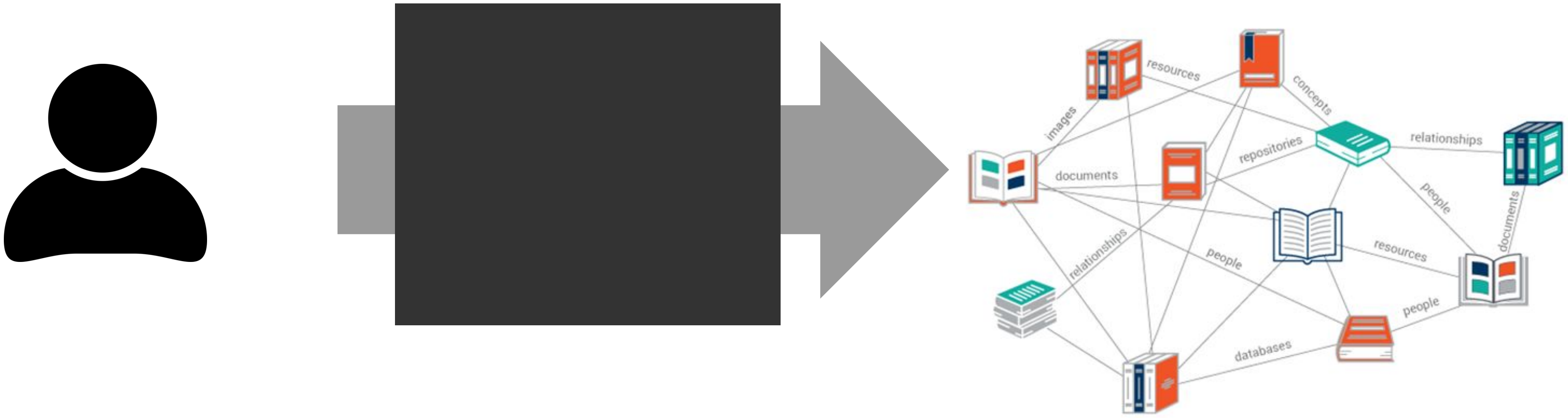
👍 low-threshold
👍 human expertise

but also

👎 time consuming
👎 difficult to reproduce
👎 difficult to scale
👎 difficult to update

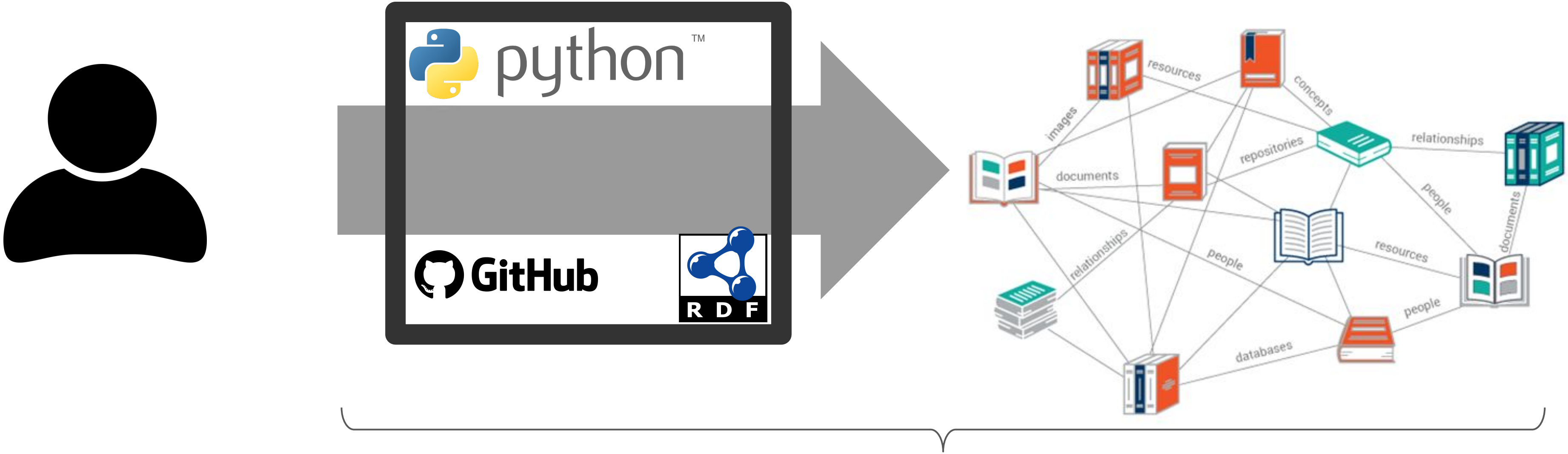
👎 subject to inconsistencies
👎 poorly documented
👎 difficult to verify
👎 difficult to reuse

We want a FAIR data corpus



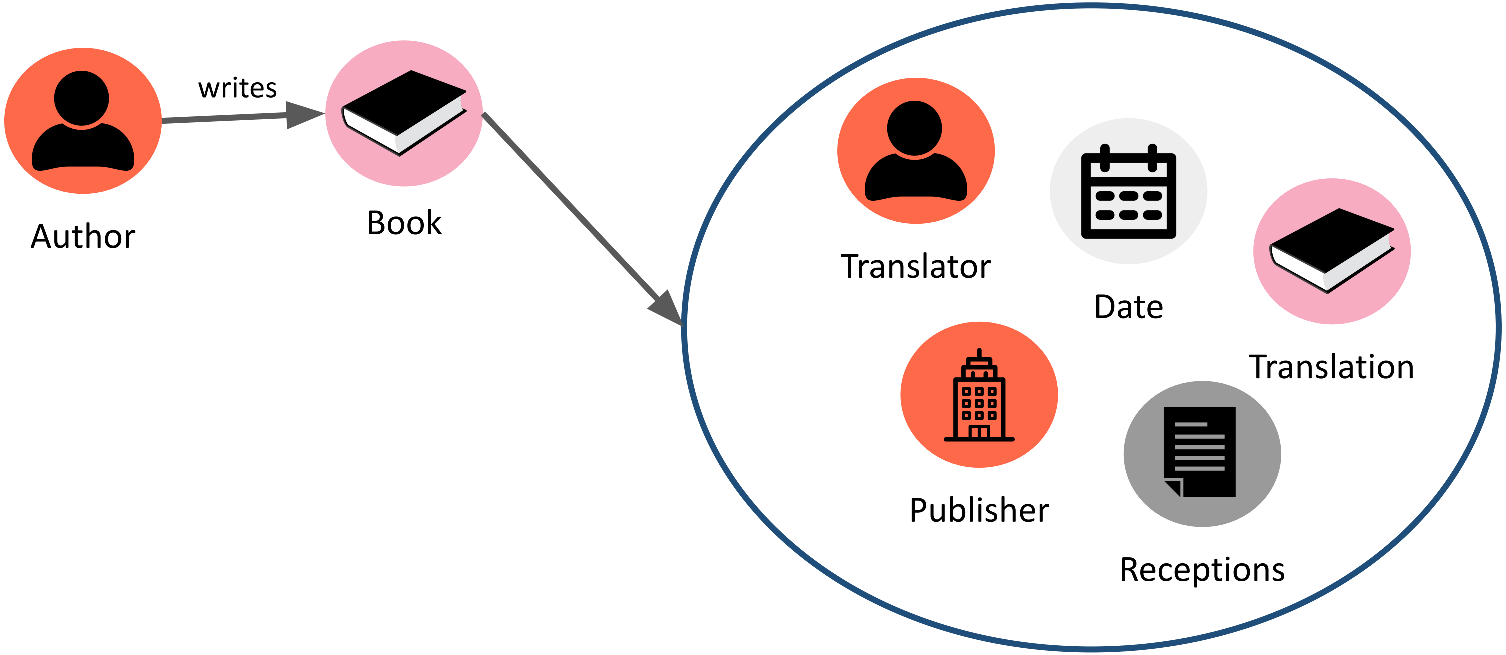
Findable **A**ccessible **I**nteroperable **R**eusable

But also open up the blackbox!
Our performed work should be documented and ideally be reused!

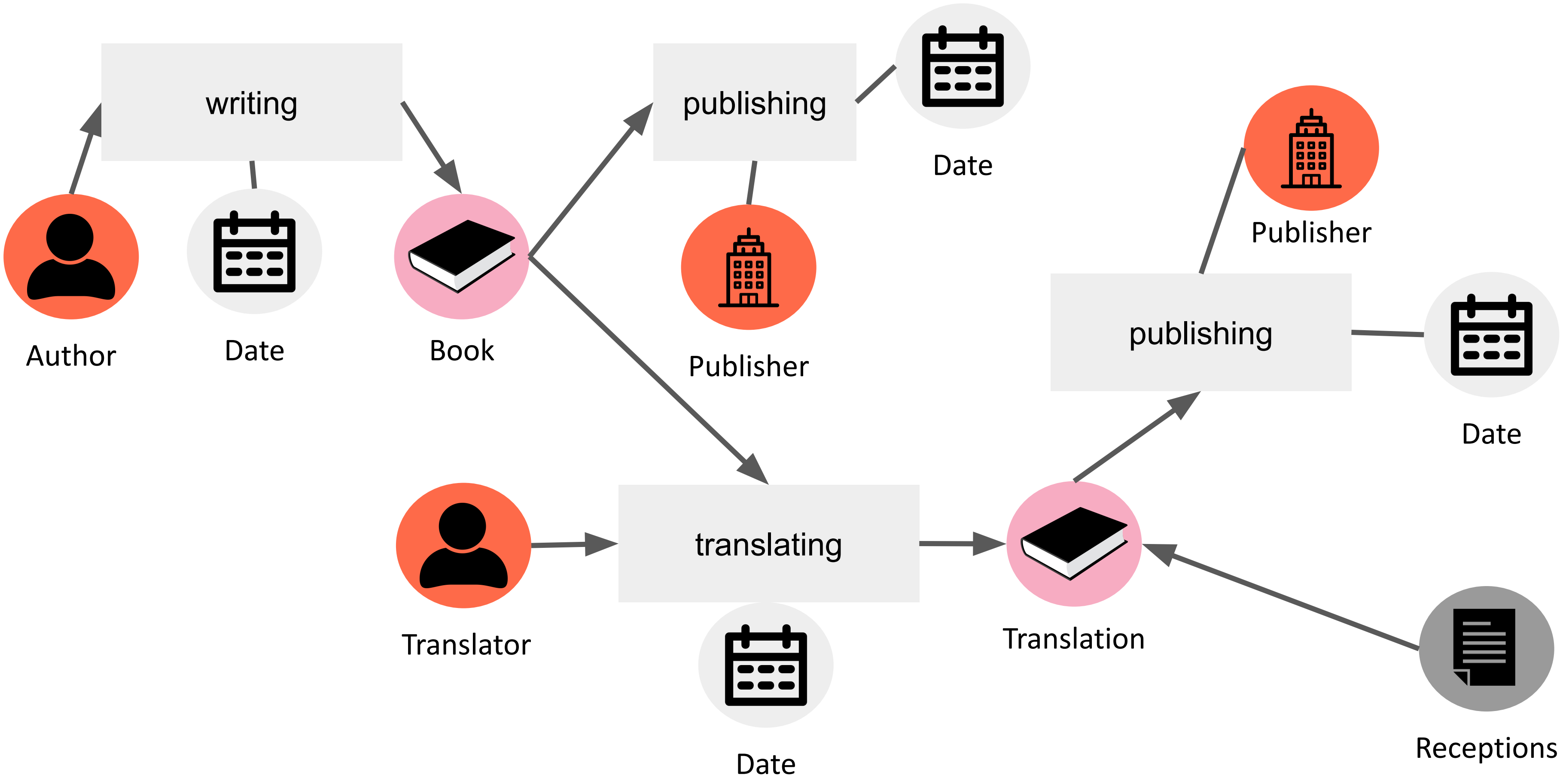


Findable Accessible Interoperable Reusable

Looking back at possible components of *translation flows* …



Linked Data - exactly what we need!



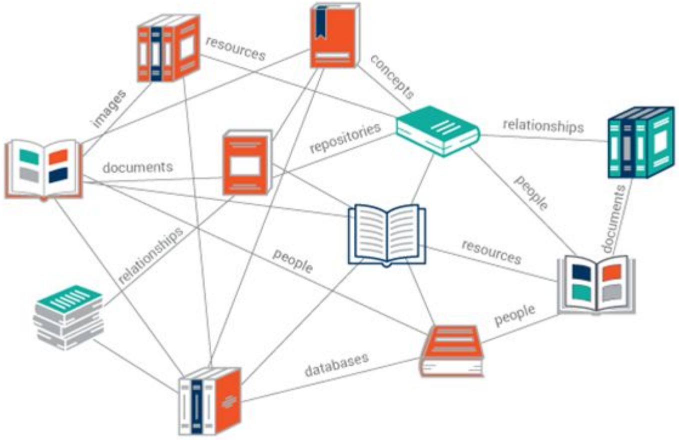
Use case: translation flow project BELTRANS

Different data sources as challenge

How do we integrate data? [Example]

Discussion of our approach

We developed a pipeline with *different* processing for *different* data sources



KBR

KBR catalog

BnF

BnF catalogue

KB

KB catalog

Flanders Literature

AML

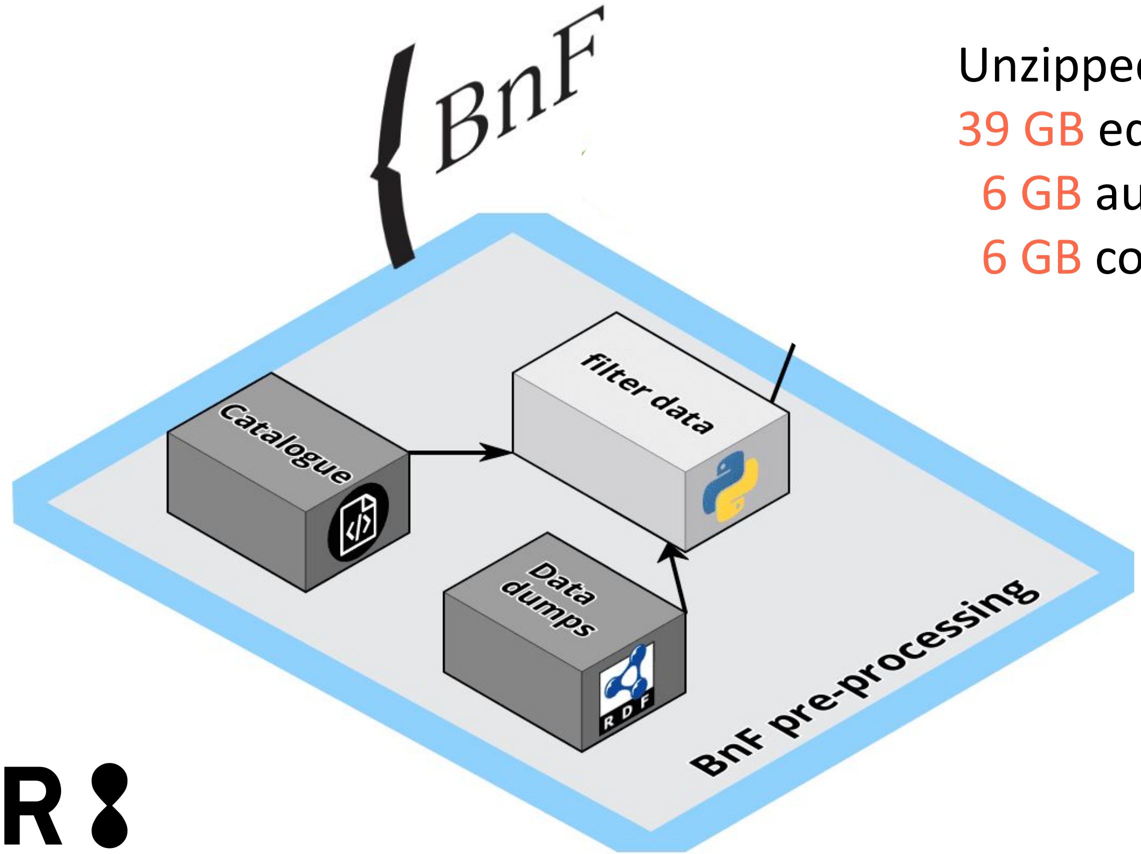
GeoNames

...

UNESCO Index Translationum

KBR Koester de tijd
Protégeons le temps

Python scripts to filter the large BnF dumps based on catalogue info



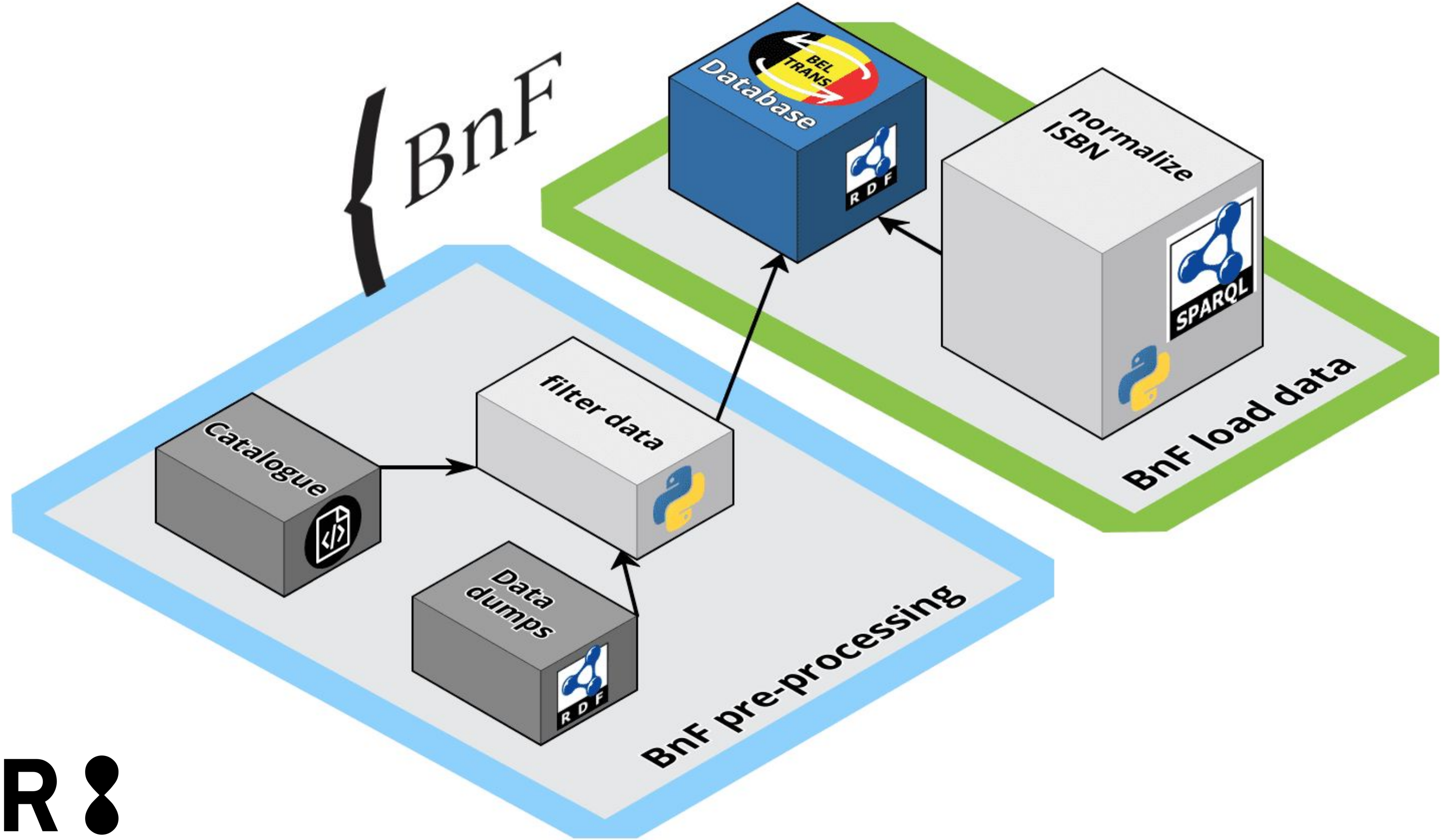
Unzipped RDF/XML dumps

39 GB editions

6 GB authors

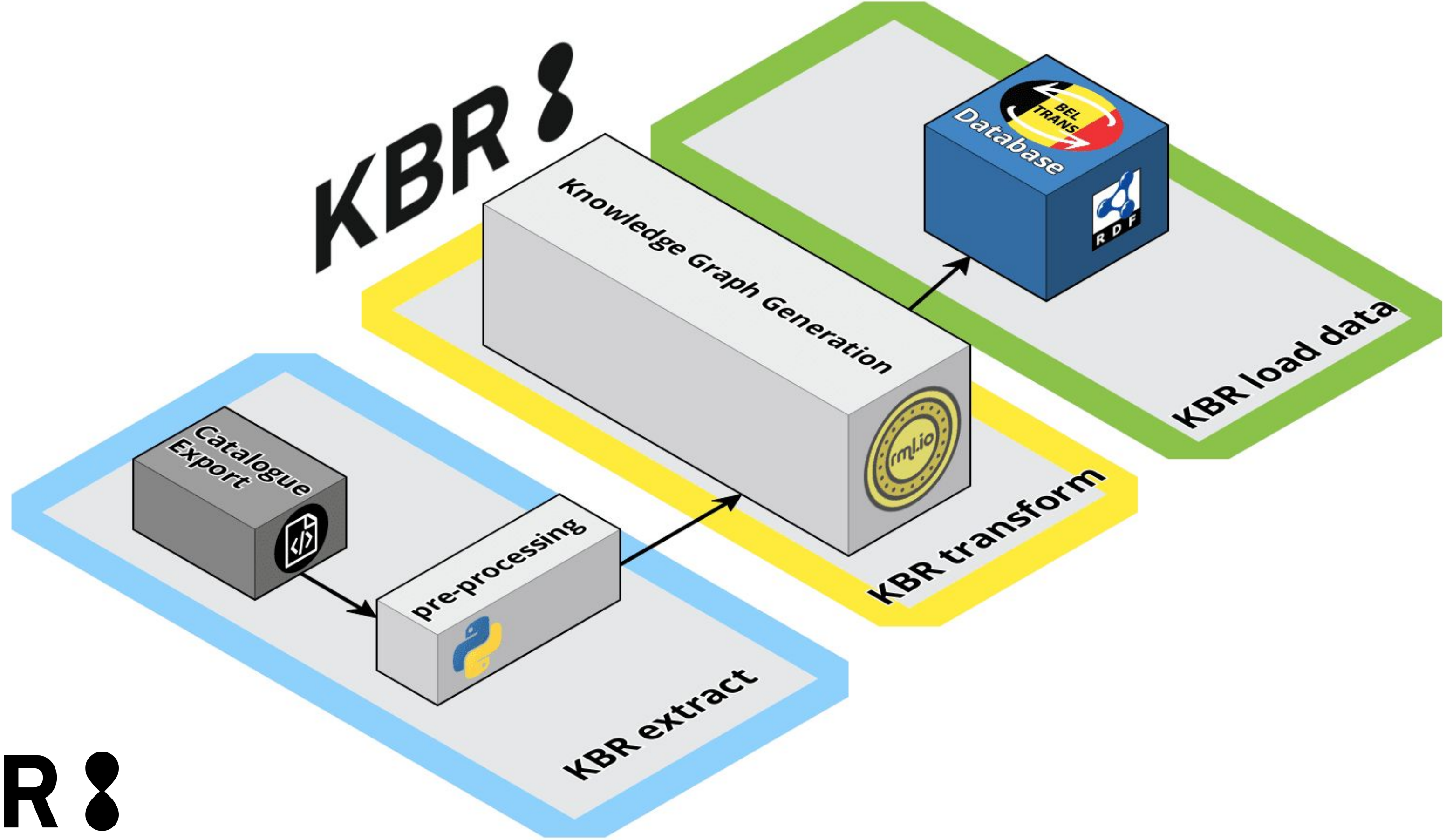
6 GB contribution links

Load data and perform SPARQL update queries to normalize ISBN



KBR

Different processing for KBR



KBR

RML mapping in YARRRML syntax to generate Linked Data

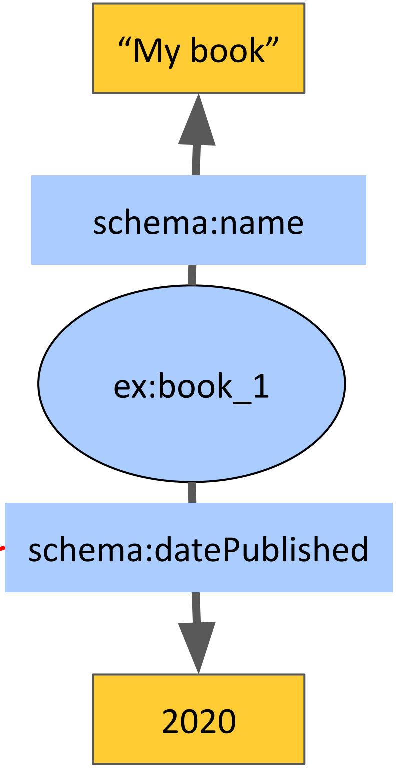
What we have (csv data) …

KBRID	title	yearOfPub
1	My book	2020
2	Your book	2019

How we get there (mapping rules)

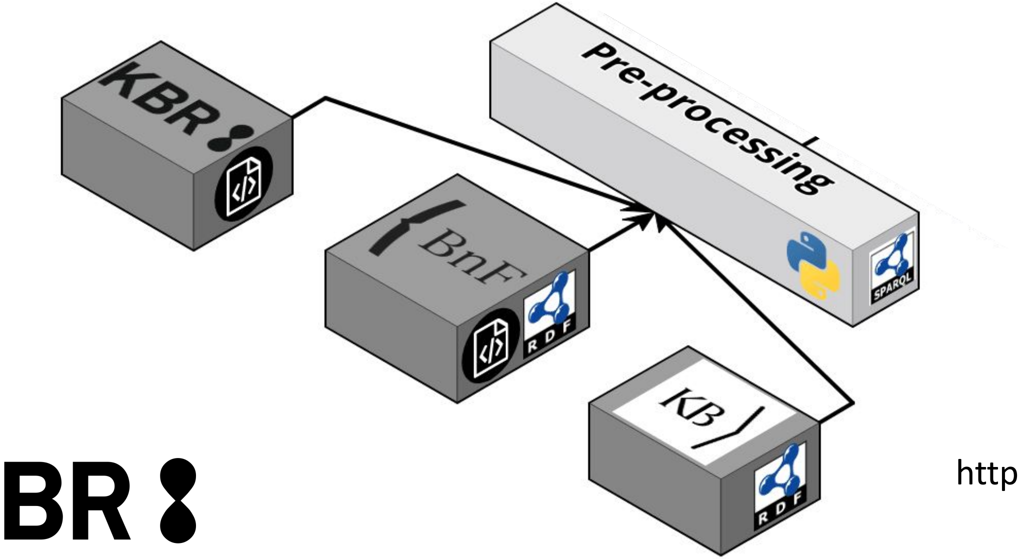
```
myMapping:
   sources:
     - access: data-export.csv
       referenceFormulation: csv
       delimiter: ','
   subjects: ex:book_$(KBRID)
   predicateobjects:
     - [a, schema:CreativeWork]
     - [schema:name, $(title) ]
     - [schema:datePublished, $(yearOfPub) ]
```

What we want! (RDF Linked Data)



KBR Koester de tijd
Protégeons le temps

In general: we pre-process data using Python and SPARQL

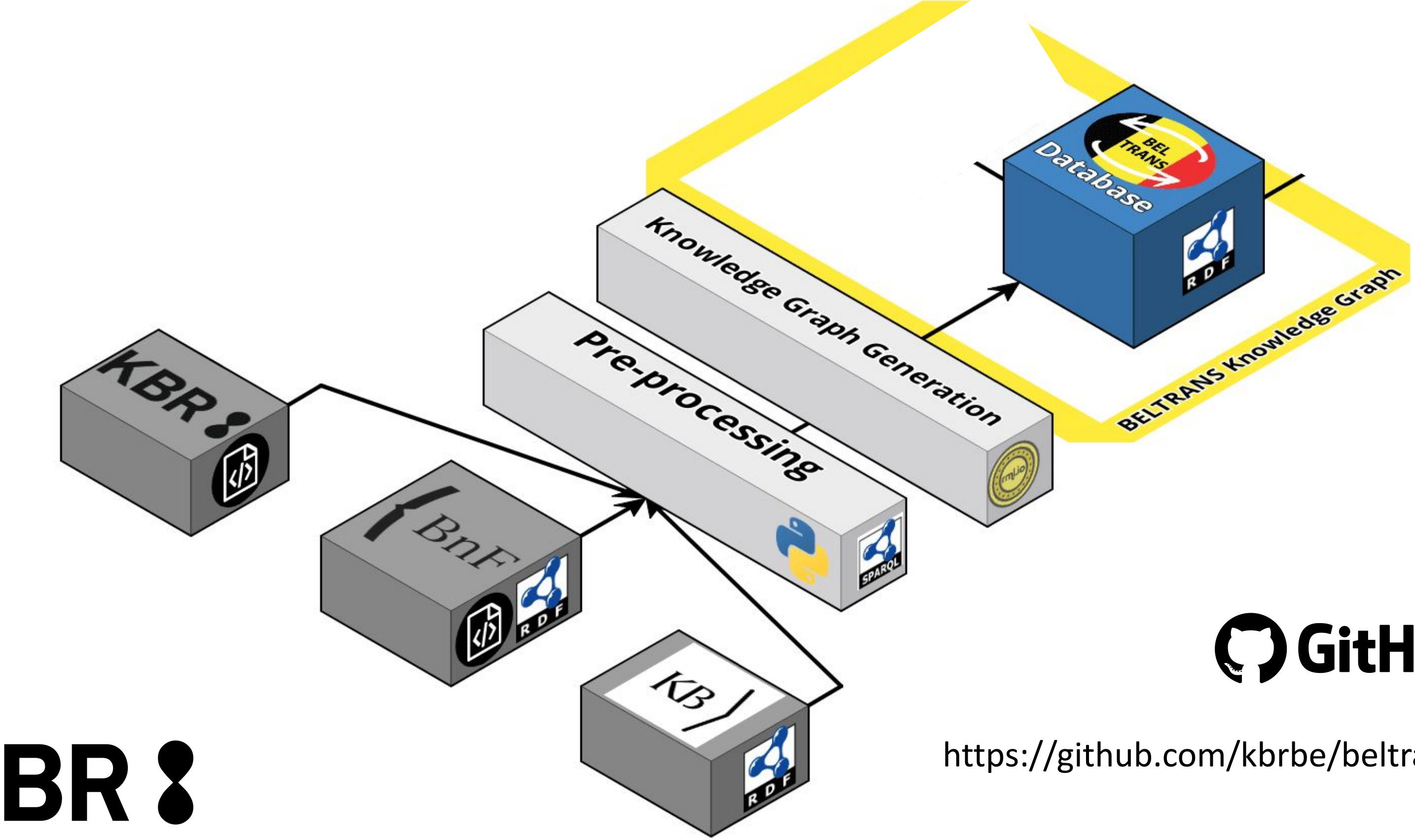


GitHub

https://github.com/kbrbe/beltrans-data-integration

KBR

We generate an RDF Knowledge Graph using RML



GitHub

https://github.com/kbrbe/beltrans-data-integration

KBR

International Standard Identifiers are very helpful

International Standard Book Number

> to find and link book editions

> using ISBN10 and ISBN13

> perhaps also to integrate publishers data

International Standard Name Identifier

> to find and link contributors

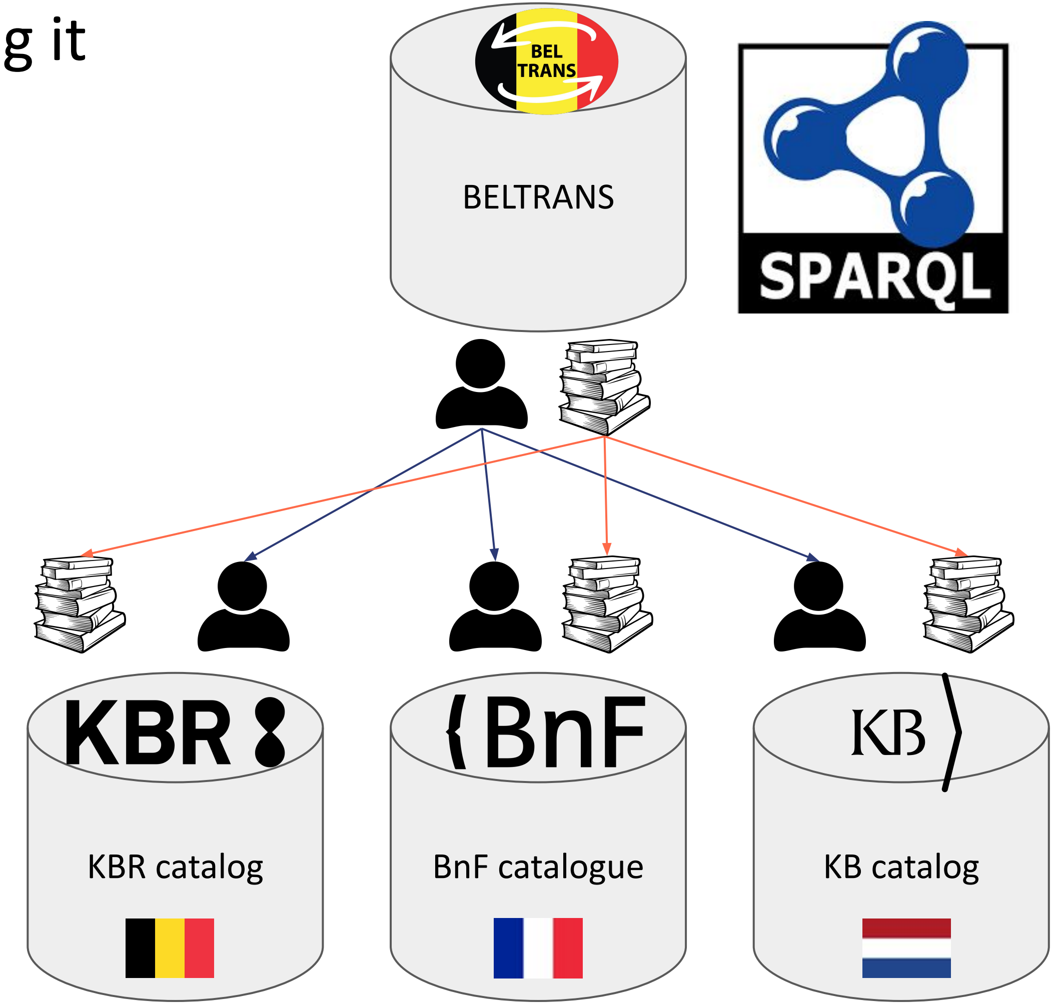
> to enrich KBR person records with Belgian nationality

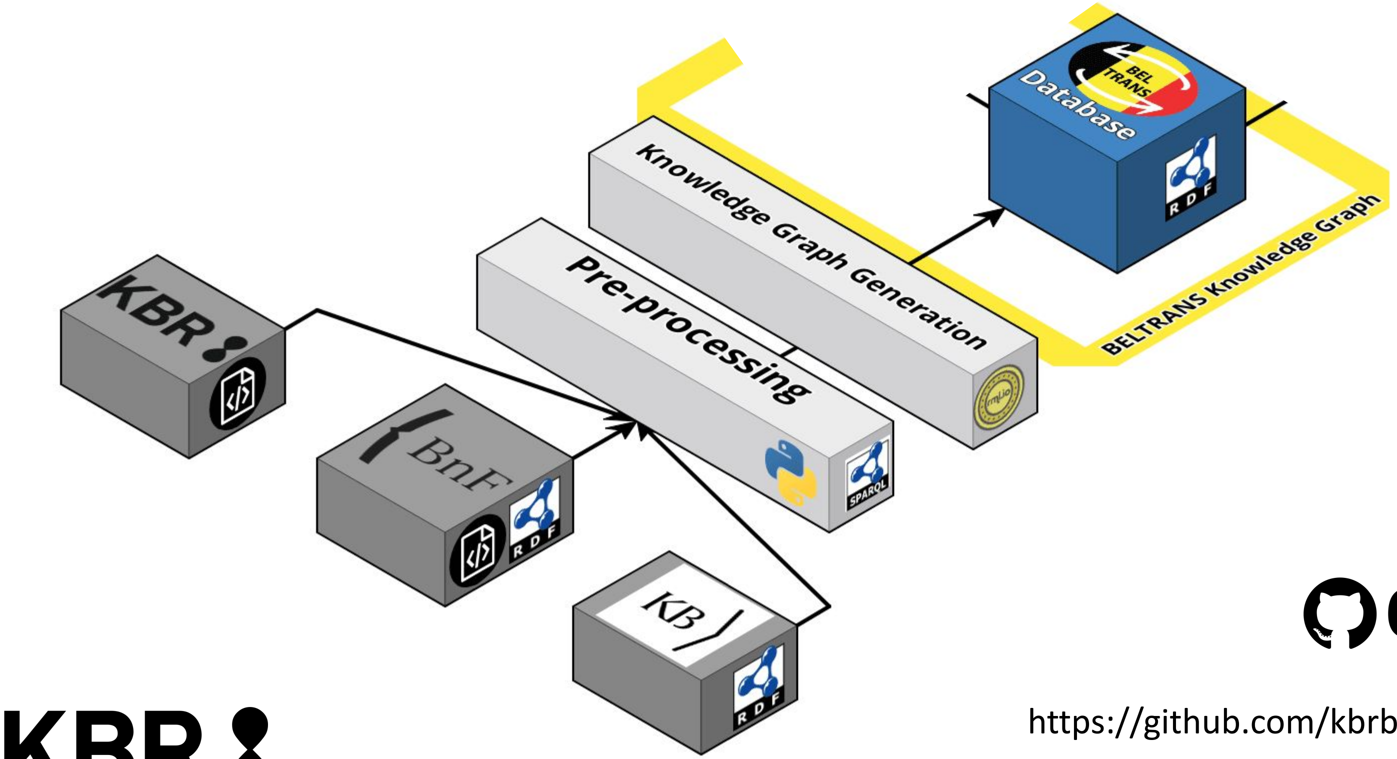
Integrate data by interlinking it

schema:sameAs

schema:sameAs

VIAF



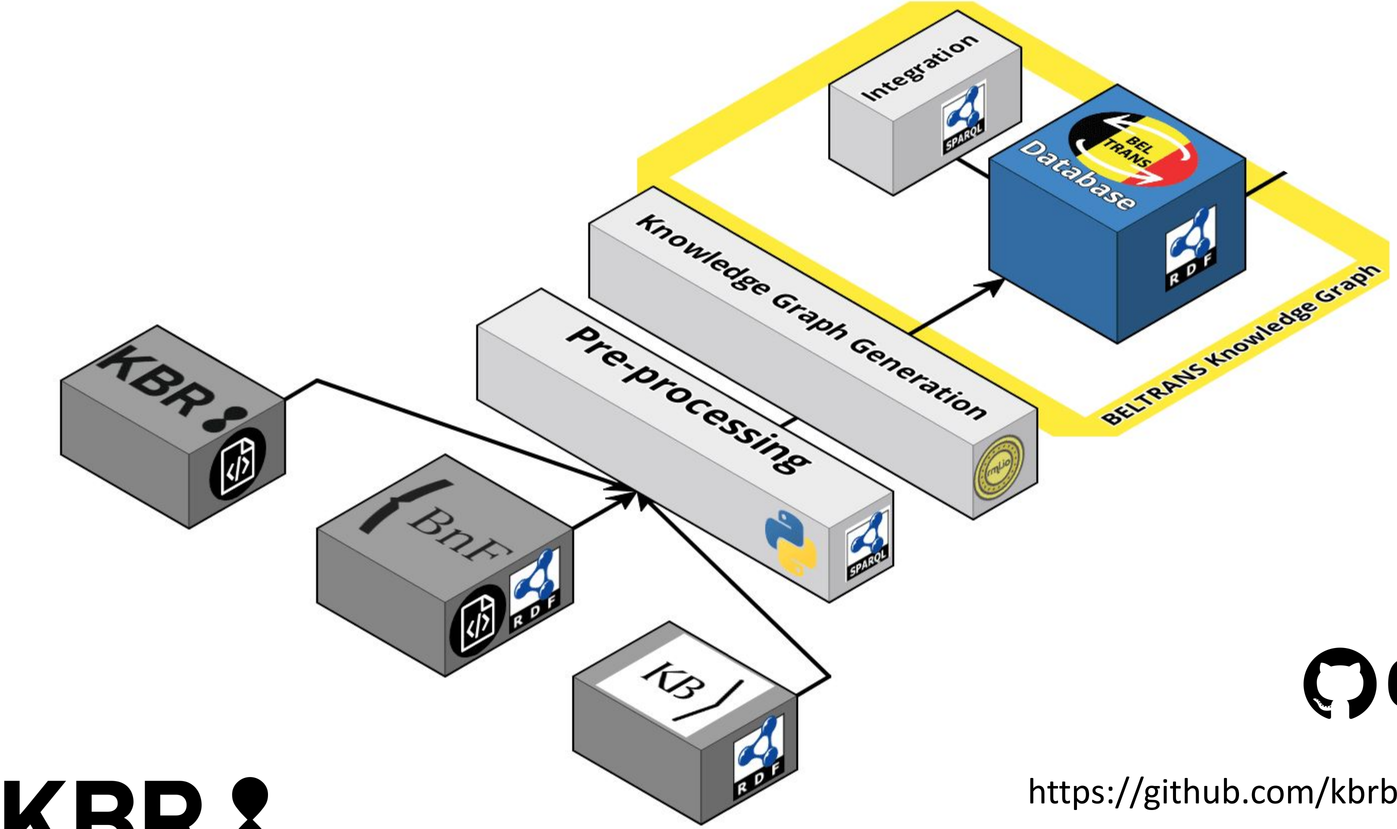


GitHub

https://github.com/kbrbe/beltrans-data-integration

KBR

We integrate the data using SPARQL Update queries



KBR

GitHub

https://github.com/kbrbe/beltrans-data-integration

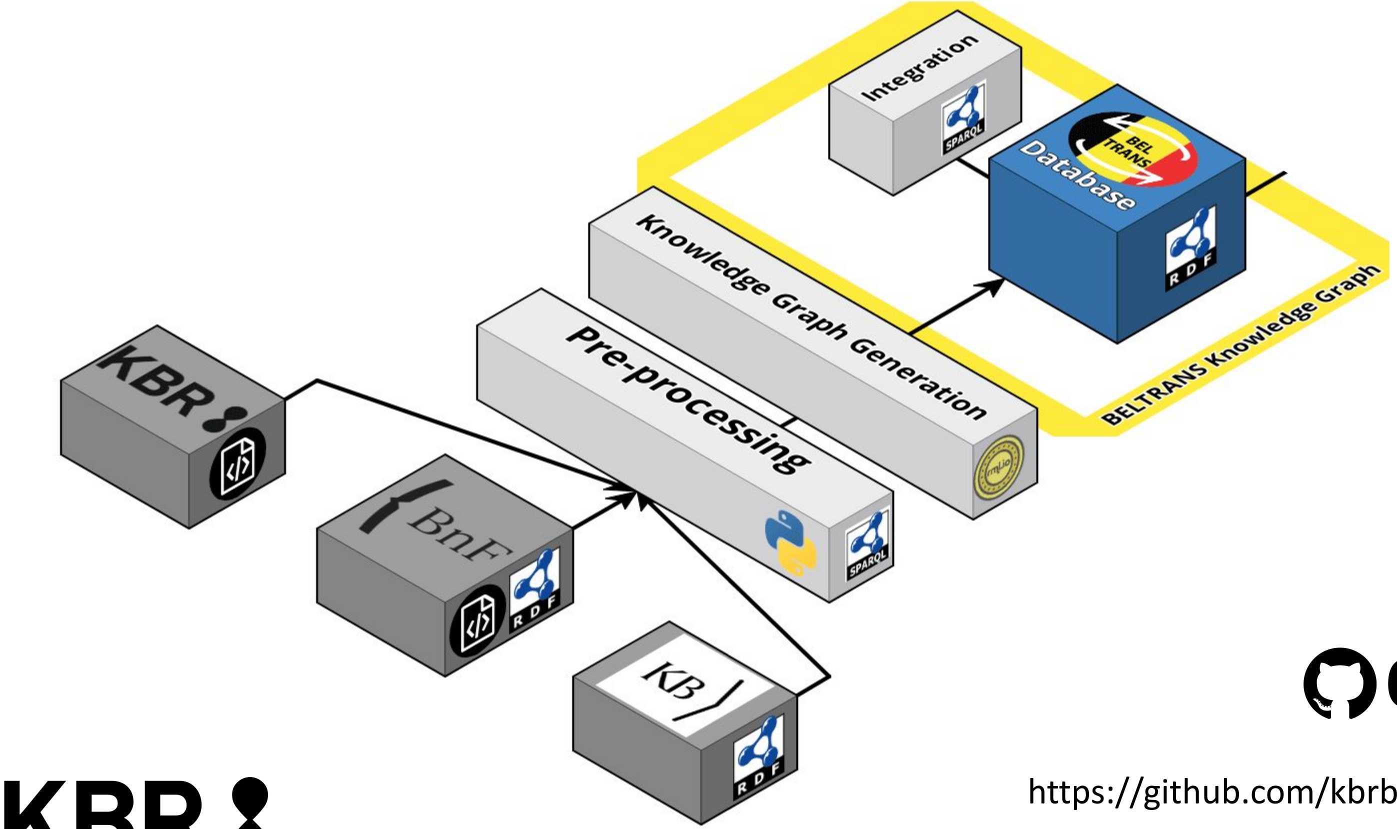
Postprocess the data and create a lightweight CSV file

ID	year Of Publication KBR	Year Of Publication BnF	Year Of Publication KB	…
1	2020	2020	2019	
2	2019		2019	

ID	year Of Publication	…
1	2019 or 2020	
2	2019	

Merge data and report inconsistencies

Koester de tijd
Protégeons le temps

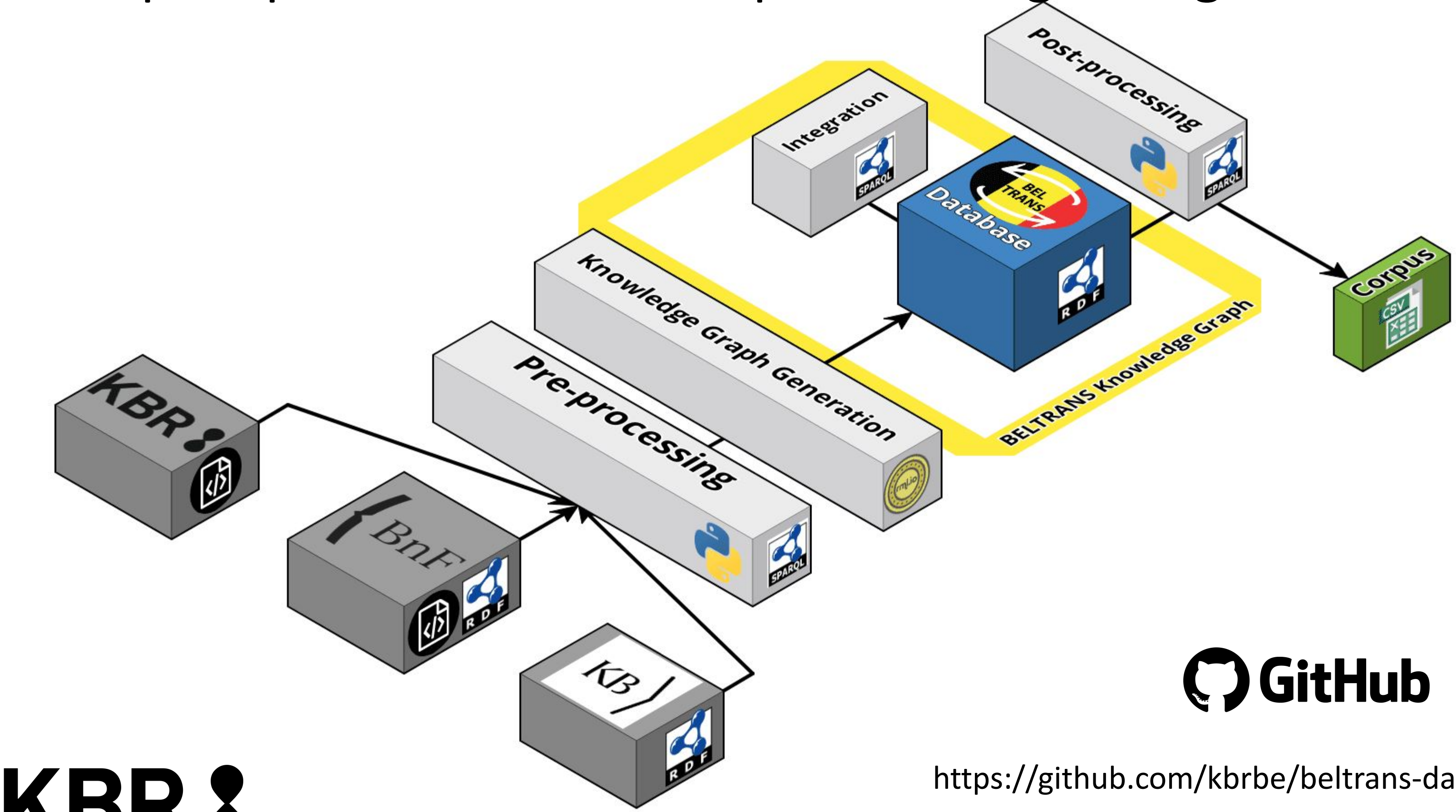


GitHub

https://github.com/kbrbe/beltrans-data-integration

KBR

We post-process the data to provide a lightweight CSV corpus



GitHub

https://github.com/kbrbe/beltrans-data-integration

KBR

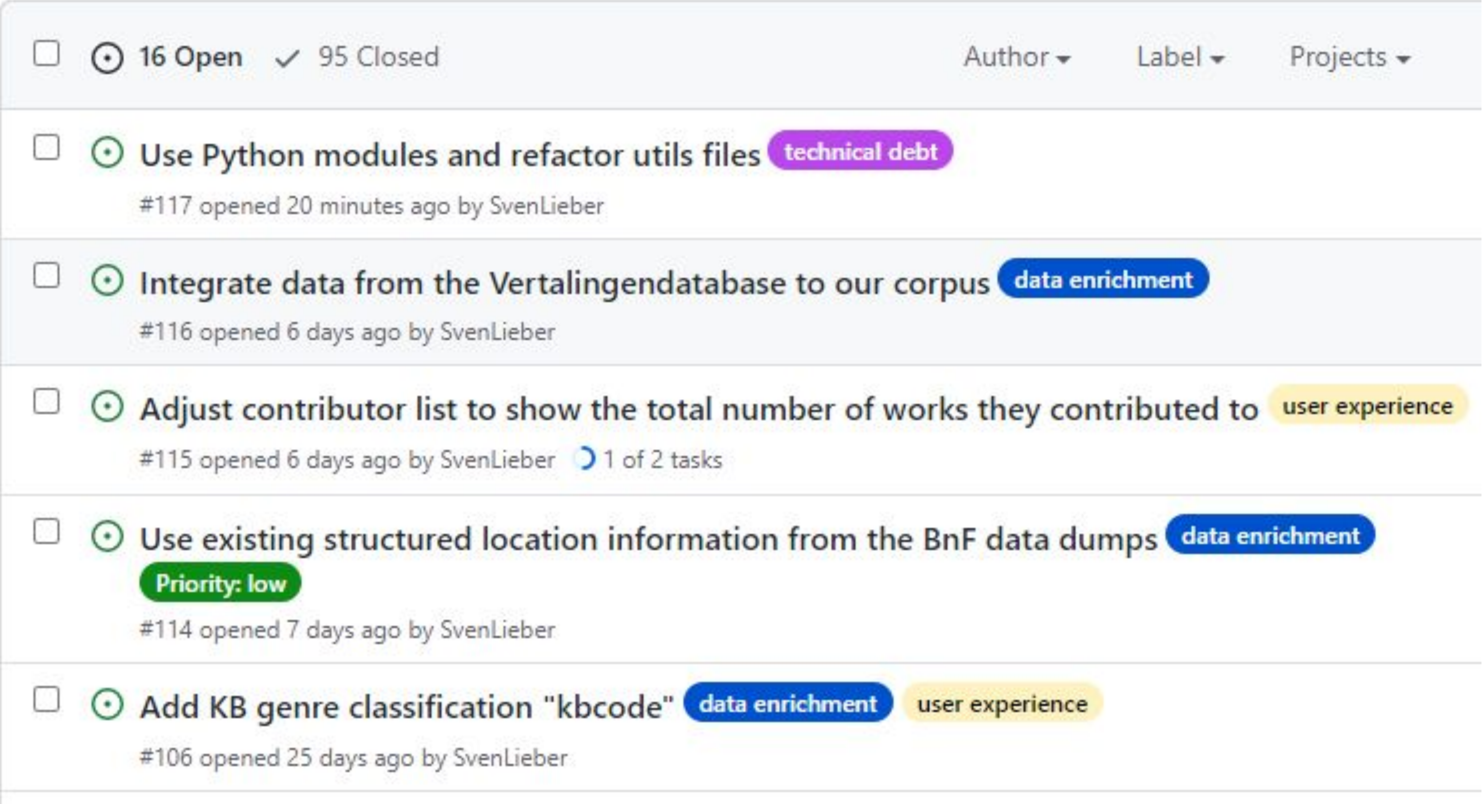
Use case: translation flow project BELTRANS

Different data sources as challenge

How do we integrate data? [Example]

Discussion of our approach

Open Science: provenance of data processing pipeline



□ ⊙ 16 Open ✓ 95 Closed Author ▾ Label ▾ Projects ▾

□ ⊙ Use Python modules and refactor utils files technical debt
#117 opened 20 minutes ago by SvenLieber

□ ⊙ Integrate data from the Vertalingendatabase to our corpus data enrichment
#116 opened 6 days ago by SvenLieber

□ ⊙ Adjust contributor list to show the total number of works they contributed to user experience
#115 opened 6 days ago by SvenLieber 1 of 2 tasks

□ ⊙ Use existing structured location information from the BnF data dumps data enrichment Priority: low
#114 opened 7 days ago by SvenLieber

□ ⊙ Add KB genre classification "kbcode" data enrichment user experience
#106 opened 25 days ago by SvenLieber

GitHub

Linked Data solution as blackbox for non-RDF experts?

Identify data issues: where in the pipeline did something go wrong?

Several points of failure, but it is transparent and can be fixed, additionally we have extensive unit/integration tests

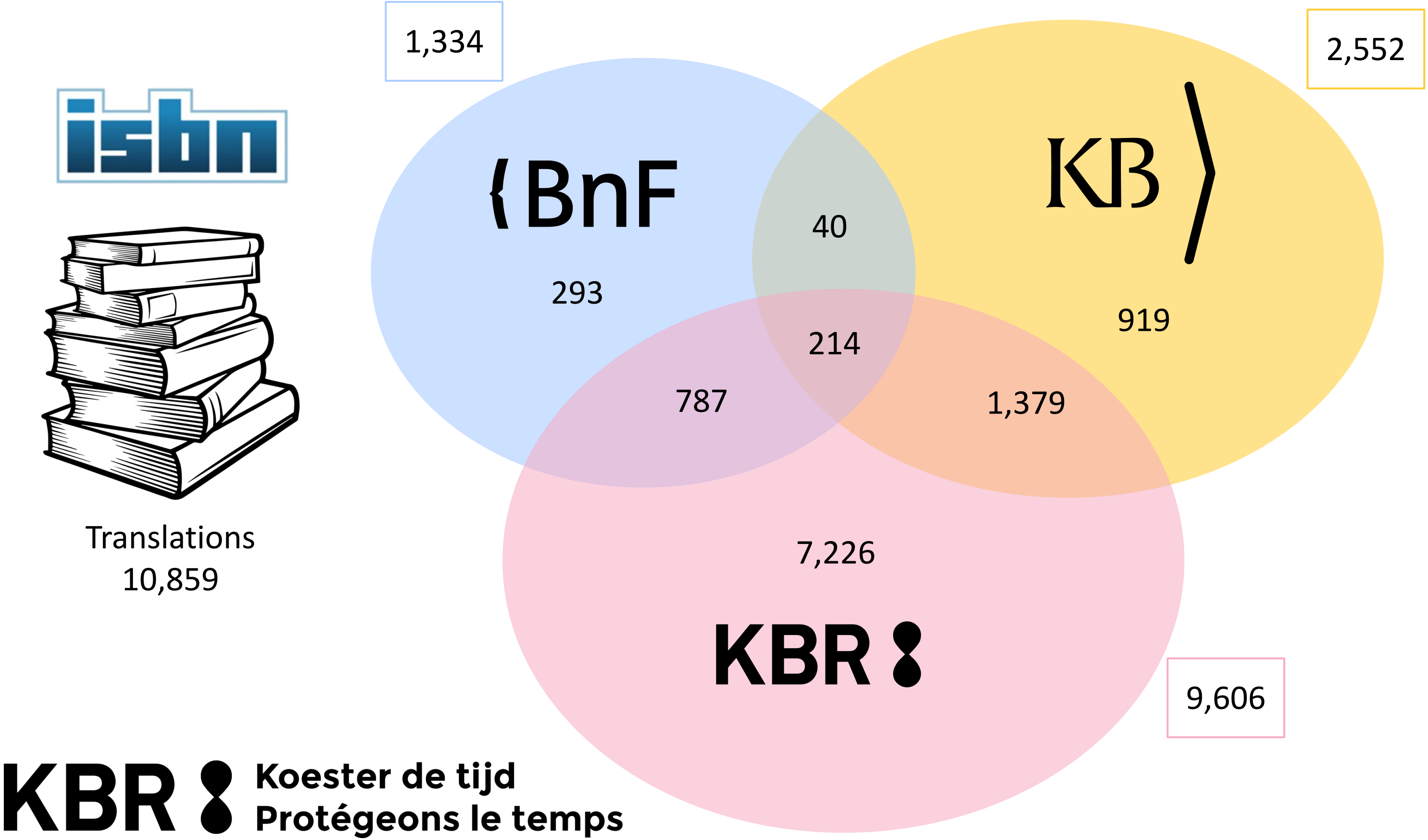
Distant reading vs close reading: iterative approach is needed

Humans are needed when data anomalies are found

Infrastructure needed: Python environment, RDF database

Requires technical skills, but eased with state-of-the-art technology (VMs, docker, Python)

Integrated data from the three big data sources KBR, BnF and KB



KBR Koester de tijd
Protégeons le temps

Questions & Answers

Sven Lieber (Sven.Lieber@kbr.be, sven-lieber.org)
Data manager BELTRANS project

Ann Van Camp (Ann.VanCamp@kbr.be)
Development of contemporary collections

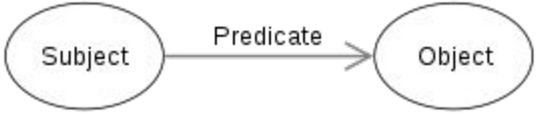
GitHub

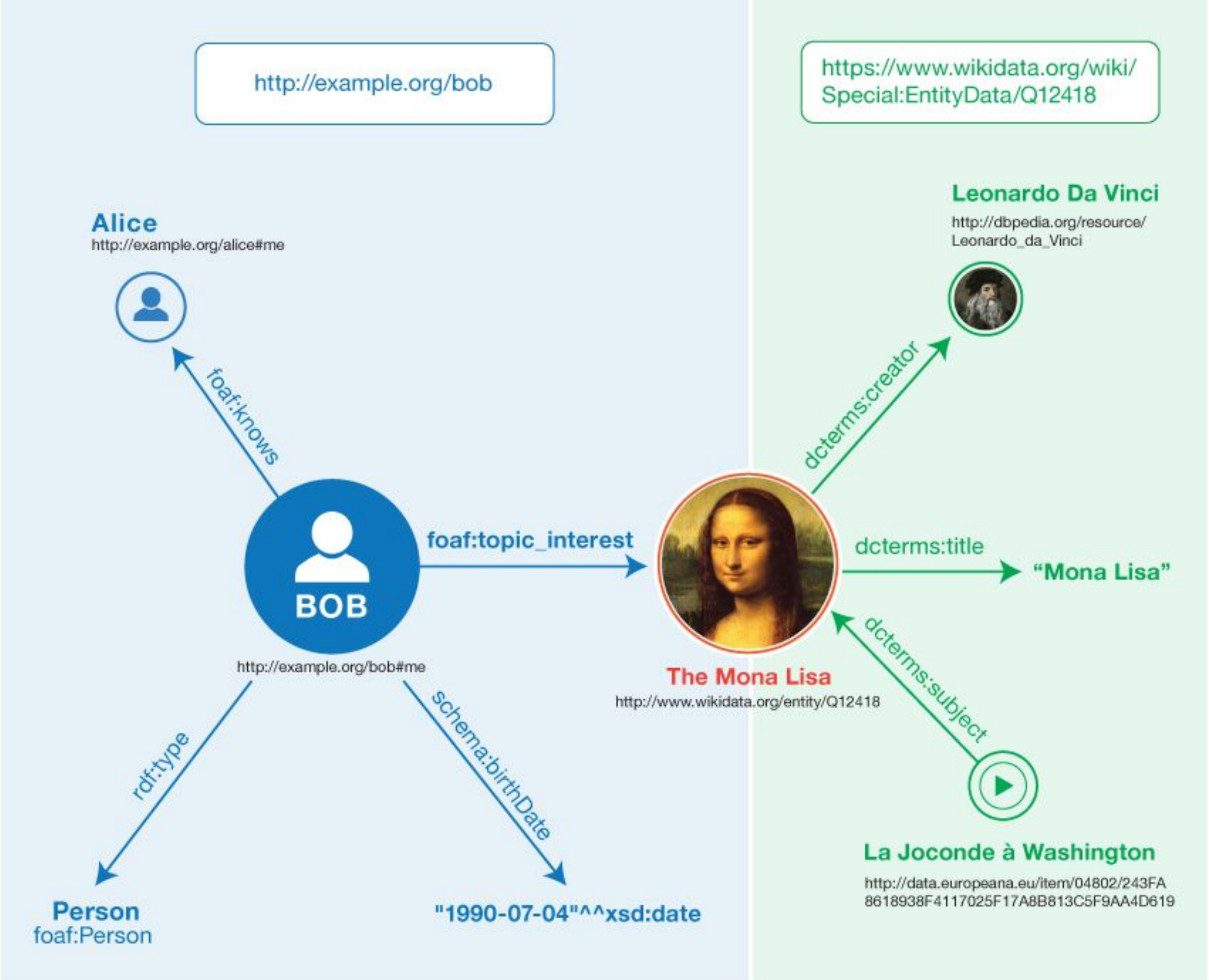
https://github.com/kbrbe/beltrans-data-integration

KBR Koester de tijd
Protégeons le temps

Backup slides for discussion

How to connect data in a meaningful way?

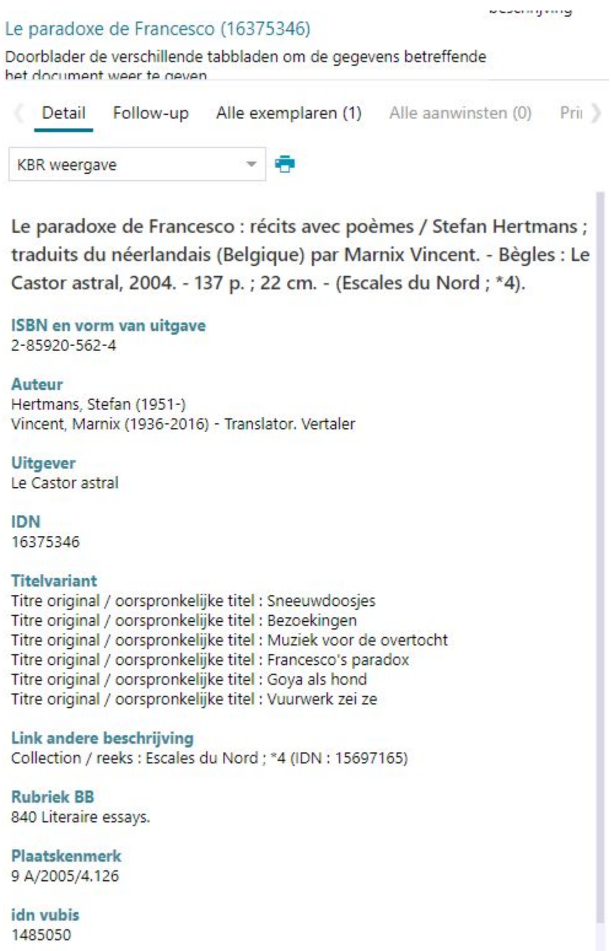



http://example.org/bob
https://www.wikidata.org/wiki/Special:EntityData/Q12418
Alice
http://example.org/alice#me
foaf:knows
Leonardo Da Vinci
http://dbpedia.org/resource/Leonardo_da_Vinci
dcterms:creator
BOB
http://example.org/bob#me
foaf:topic_interest
The Mona Lisa
http://www.wikidata.org/entity/Q12418
dcterms:title
"Mona Lisa"
dcterms:subject
rdf:type
schema:birthDate
Person
foaf:Person
"1990-07-04"^^xsd:date
La Joconde à Washington
http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619

KBR

Koester de tijd
Protégeons le temps

How can the KBR catalog help in BELTRANS?



Le paradoxe de Francesco (16375346)

Doorblader de verschillende tabbladen om de gegevens betreffende het document weer te geven

Le paradoxe de Francesco : récits avec poèmes / Stefan Hertmans ; traduits du néerlandais (Belgique) par Marnix Vincent. - Bègles : Le Castor astral, 2004. - 137 p. ; 22 cm. - (Escales du Nord ; *4).

ISBN en vorm van uitgave
2-85920-562-4

Auteur
Hertmans, Stefan (1951-)
Vincent, Marnix (1936-2016) - Translator. Vertaler

Uitgever
Le Castor astral

IDN
16375346

Titelvariant
Titre original / oorspronkelijke titel : Sneeuwdoosjes
Titre original / oorspronkelijke titel : Bezoekingen
Titre original / oorspronkelijke titel : Muziek voor de overtocht
Titre original / oorspronkelijke titel : Francesco's paradox
Titre original / oorspronkelijke titel : Goya als hond
Titre original / oorspronkelijke titel : Vuurwerk zei ze

Link andere beschrijving
Collection / reeks : Escales du Nord ; *4 (IDN : 15697165)

Rubriek BB
840 Literaire essays.

Plaatskenmerk
9 A/2005/4.126

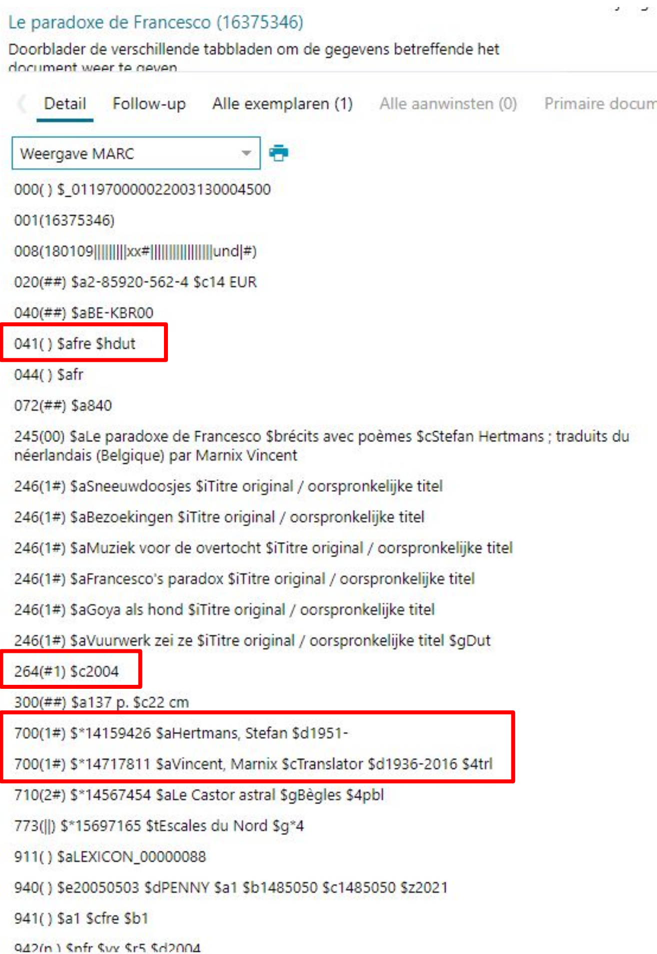
idn vubis
1485050

Original language
Translation language
041$a = "fre"
041$h = "dut"

Year of publication
264$c = "2004"

Contributors and roles
700$* = "14159426"
700$4 = "aut"

700$* = "14717811"
700$4 = "trl"



Le paradoxe de Francesco (16375346)

Doorblader de verschillende tabbladen om de gegevens betreffende het document weer te geven

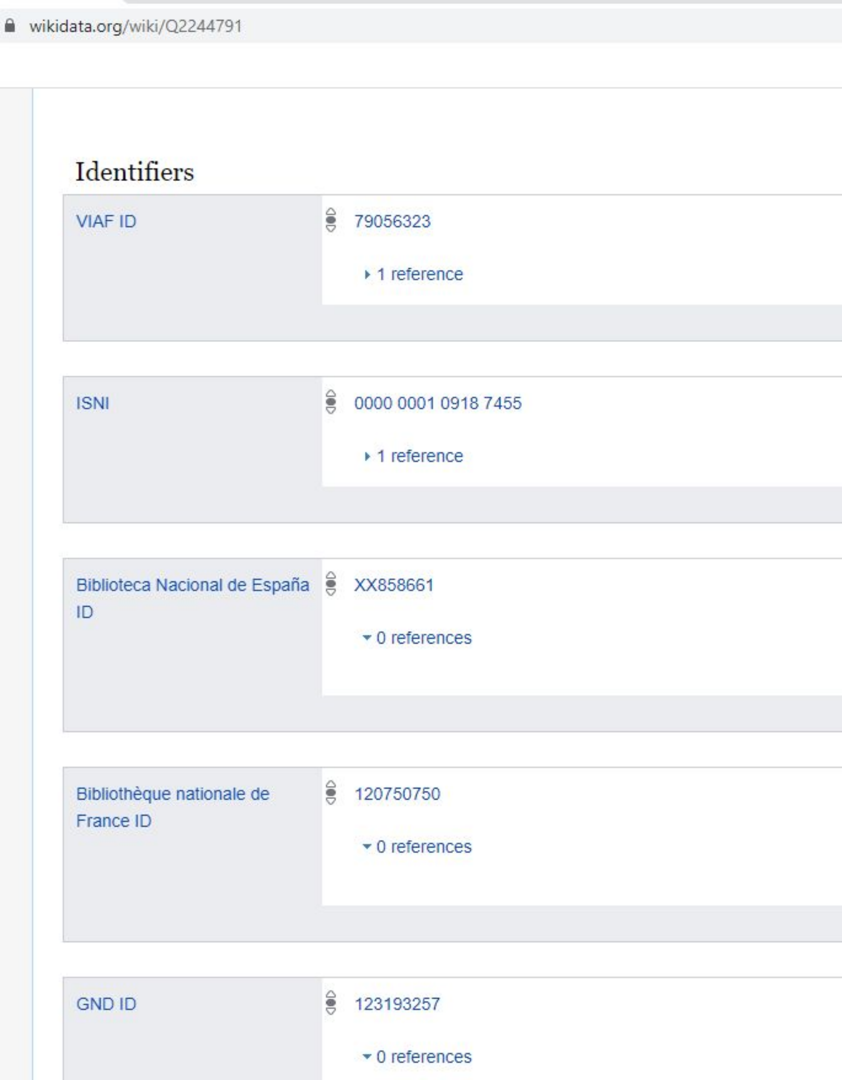
```
000( ) $_011970000022003130004500
001(16375346)
008(180109||||||||xx#|||||||||||||||und|#)
020(##) $a2-85920-562-4 $c14 EUR
040(##) $aBE-KBR00
041( ) $afre $hdut
044( ) $afr
072(##) $a840
245(00) $aLe paradoxe de Francesco $brécits avec poèmes $cStefan Hertmans ; traduits du
néerlandais (Belgique) par Marnix Vincent
246(1#) $aSneeuwdoosjes $iTitre original / oorspronkelijke titel
246(1#) $aBezoekingen $iTitre original / oorspronkelijke titel
246(1#) $aMuziek voor de overtocht $iTitre original / oorspronkelijke titel
246(1#) $aFrancesco's paradox $iTitre original / oorspronkelijke titel
246(1#) $aGoya als hond $iTitre original / oorspronkelijke titel
246(1#) $aVuurwerk zei ze $iTitre original / oorspronkelijke titel $gDut
264(#1) $c2004
300(##) $a137 p. $c22 cm
700(1#) $*14159426 $aHertmans, Stefan $d1951-
700(1#) $*14717811 $aVincent, Marnix $cTranslator $d1936-2016 $4trl
710(2#) $*14567454 $aLe Castor astral $gBègles $4pbl
773(||) $*15697165 $tEscales du Nord $g*4
911( ) $aLEXICON_00000088
940( ) $e20050503 $dPENNY $a1 $b1485050 $c1485050 $z2021
941( ) $a1 $cfre $b1
942(n ) $nfr $vx $r5 $d2004
```

Other unique identifiers for an author make it possible to retrieve more information from other sources

https://catalogue.bnf.fr/ark:/12148/cb**12075075**0

https://data.bnf.fr/en/**12075075**/stefan_hertmans/

https://data.bnf.fr/en/**12075075**/stefan_hertmans/rdf.n3



wikidata.org/wiki/Q2244791

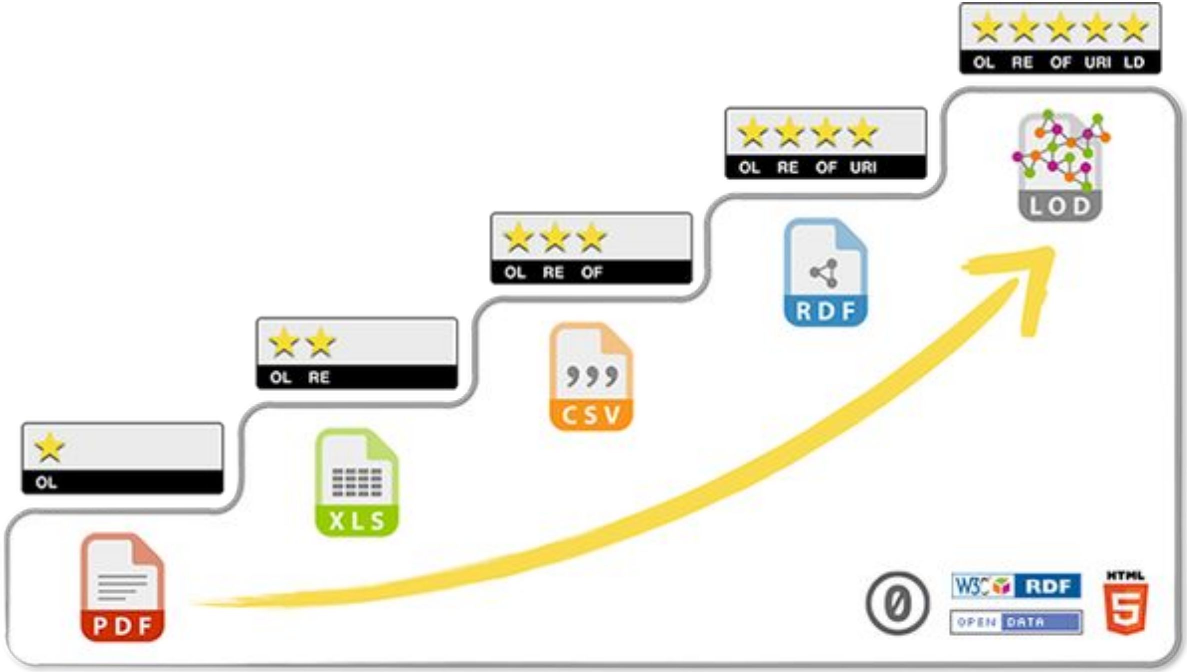
Identifiers

VIAF ID	79056323 ▸ 1 reference
ISNI	0000 0001 0918 7455 ▸ 1 reference
Biblioteca Nacional de España ID	XX858661 ▾ 0 references
Bibliothèque nationale de France ID	120750750 ▾ 0 references
GND ID	123193257 ▾ 0 references

KBR

Koester de tijd
Protégeons le temps

Data sharing -> What is Linked (Open) Data?



Open license

Machine readable structured format

Open Format

Identifier

Link your data to other data

International Standard Identifiers are very helpful

International Standard Name Identifier

> to find and link contributors

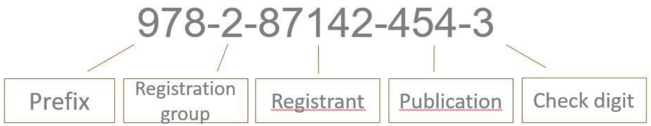
> to enrich KBR person records with Belgian nationality

International Standard Book Number

> to find and link book editions

> perhaps also to integrate publishers data

9 782871 424543


978-2-87142-454-3

Prefix | Registration group | Registrant | Publication | Check digit

⚠ Take into account both ISBN-10 and ISBN-13

Theory versus practice: different editions may have one same ISBN
co-editions have two ISBNs