A LITL more quality: improving the *correctness* and *completeness* of library catalogs with a Librarian-In-The-Loop linked data workflow

Sven Lieber, Ann Van Camp, Hannes Lowagie

SWIB conference, 28 November 2022

KBR Koester de tijd
Protégeons le temps

What is data quality, what is correct and what is valid?

Current quality procedures at KBR

The use case: BELTRANS project

Our **L**ibrarian-**I**n-**T**he-**L**oop workflow

Lessons learned and future work

Take home message:

Use your data to achieve something!
You will likely encounter
data quality issues…

What is data quality, what is correct and what is valid?

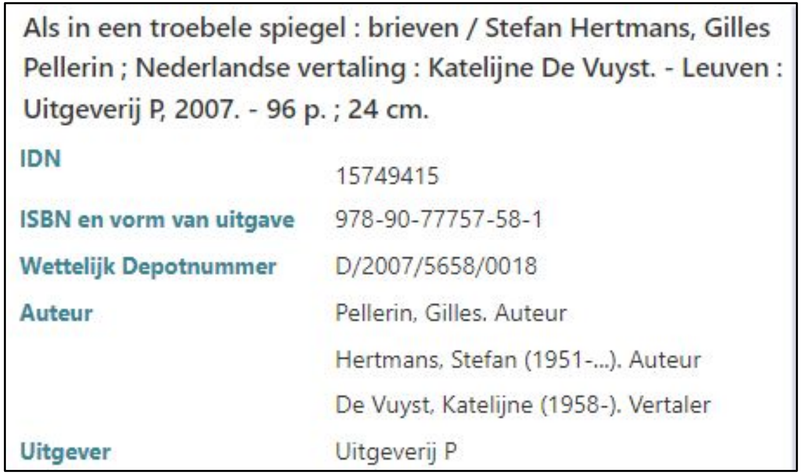
Current quality procedures at KBR

The use case: BELTRANS project

Our **L**ibrarian-**I**n-**T**he-**L**oop workflow

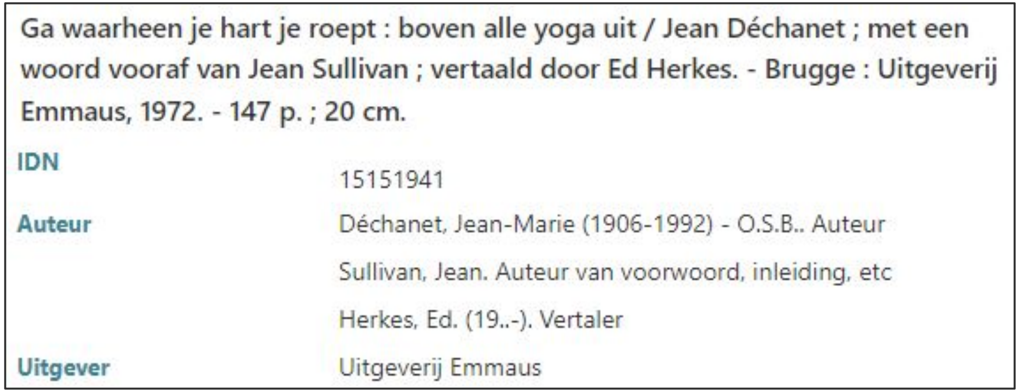
Lessons learned and future work

Data quality is use case specific



Als in een troebele spiegel : brieven / Stefan Hertmans, Gilles Pellerin ; Nederlandse vertaling : Katelijne De Vuyst. - Leuven : Uitgeverij P, 2007. - 96 p. ; 24 cm.

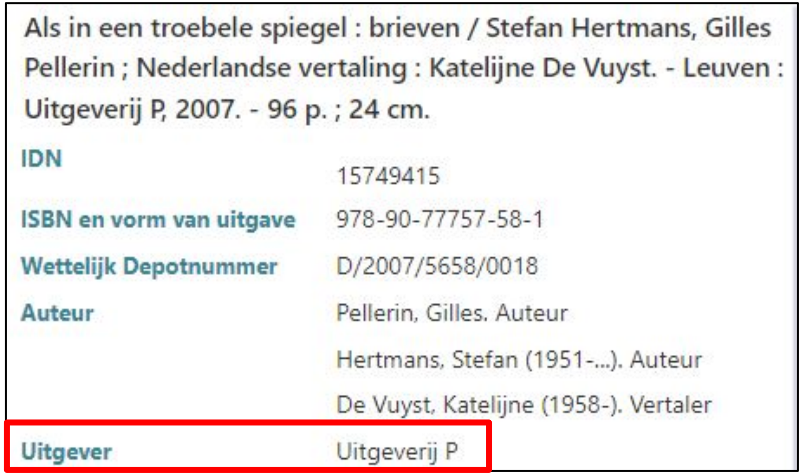
IDN	15749415
ISBN en vorm van uitgave	978-90-77757-58-1
Wettelijk Depotnummer	D/2007/5658/0018
Auteur	Pellerin, Gilles. Auteur
	Hertmans, Stefan (1951-...). Auteur
	De Vuyst, Katelijne (1958-). Vertaler
Uitgever	Uitgeverij P



Ga waarheen je hart je roept : boven alle yoga uit / Jean Déchanet ; met een woord vooraf van Jean Sullivan ; vertaald door Ed Herkes. - Brugge : Uitgeverij Emmaus, 1972. - 147 p. ; 20 cm.

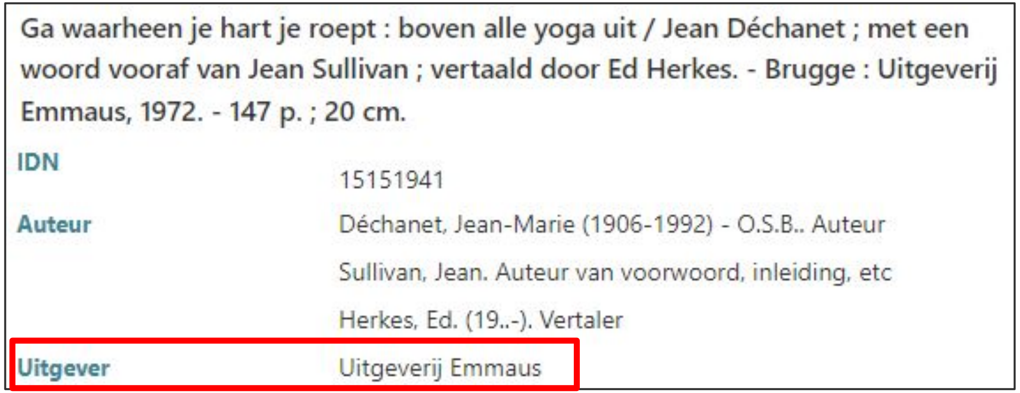
IDN	15151941
Auteur	Déchanet, Jean-Marie (1906-1992) - O.S.B.. Auteur
	Sullivan, Jean. Auteur van voorwoord, inleiding, etc
	Herkes, Ed. (19..-). Vertaler
Uitgever	Uitgeverij Emmaus

Data quality is use case specific - similar quality when viewing



Als in een troebele spiegel : brieven / Stefan Hertmans, Gilles Pellerin ; Nederlandse vertaling : Katelijne De Vuyst. - Leuven : Uitgeverij P, 2007. - 96 p. ; 24 cm.

IDN	15749415
ISBN en vorm van uitgave	978-90-77757-58-1
Wettelijk Depotnummer	D/2007/5658/0018
Auteur	Pellerin, Gilles. Auteur Hertmans, Stefan (1951-...). Auteur De Vuyst, Katelijne (1958-). Vertaler
Uitgever	Uitgeverij P



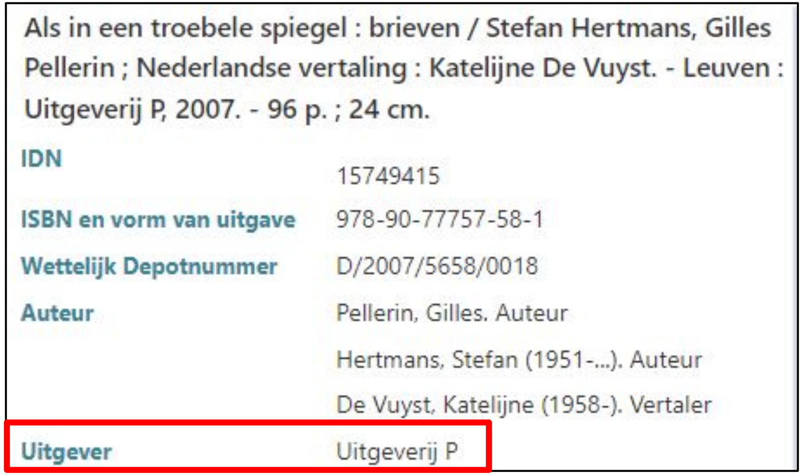
Ga waarheen je hart je roept : boven alle yoga uit / Jean Déchanet ; met een woord vooraf van Jean Sullivan ; vertaald door Ed Herkes. - Brugge : Uitgeverij Emmaus, 1972. - 147 p. ; 20 cm.

IDN	15151941
Auteur	Déchanet, Jean-Marie (1906-1992) - O.S.B.. Auteur Sullivan, Jean. Auteur van voorwoord, inleiding, etc Herkes, Ed. (19..-). Vertaler
Uitgever	Uitgeverij Emmaus

use case *“Which publisher?”*

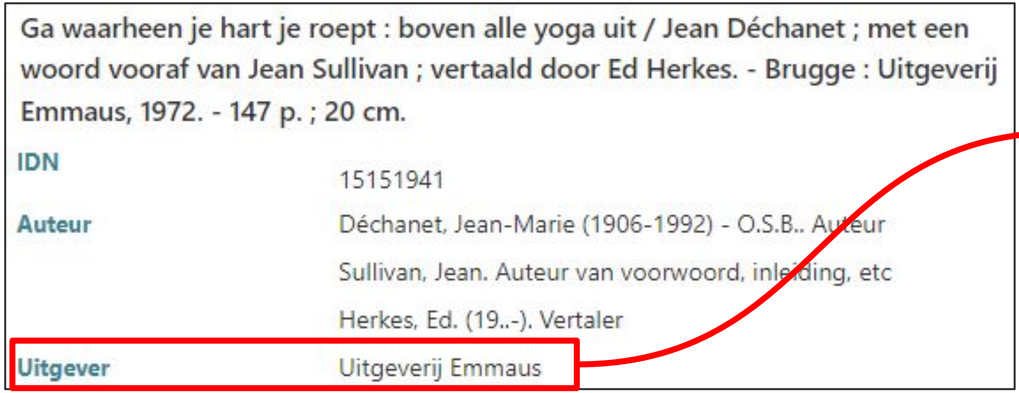
Data quality is use case specific - similar quality when viewing

use case *“Where is the publisher from?”*



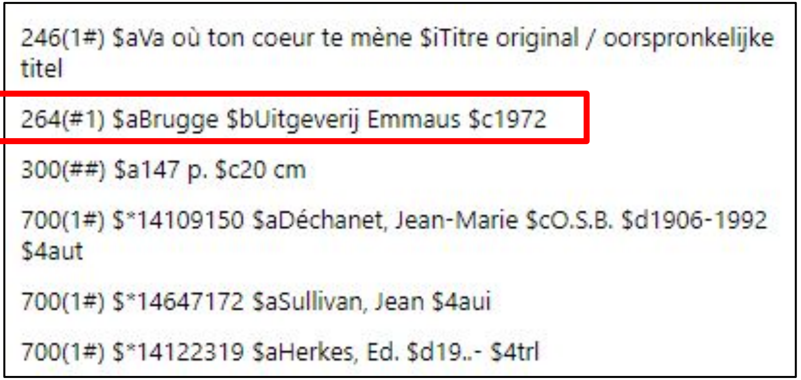
Als in een troebele spiegel : brieven / Stefan Hertmans, Gilles Pellerin ; Nederlandse vertaling : Katelijne De Vuyst. - Leuven : Uitgeverij P, 2007. - 96 p. ; 24 cm.

IDN	15749415
ISBN en vorm van uitgave	978-90-77757-58-1
Wettelijk Depotnummer	D/2007/5658/0018
Auteur	Pellerin, Gilles. Auteur Hertmans, Stefan (1951-...). Auteur De Vuyst, Katelijne (1958-). Vertaler
Uitgever	Uitgeverij P



Ga waarheen je hart je roept : boven alle yoga uit / Jean Déchanet ; met een woord vooraf van Jean Sullivan ; vertaald door Ed Herkes. - Brugge : Uitgeverij Emmaus, 1972. - 147 p. ; 20 cm.

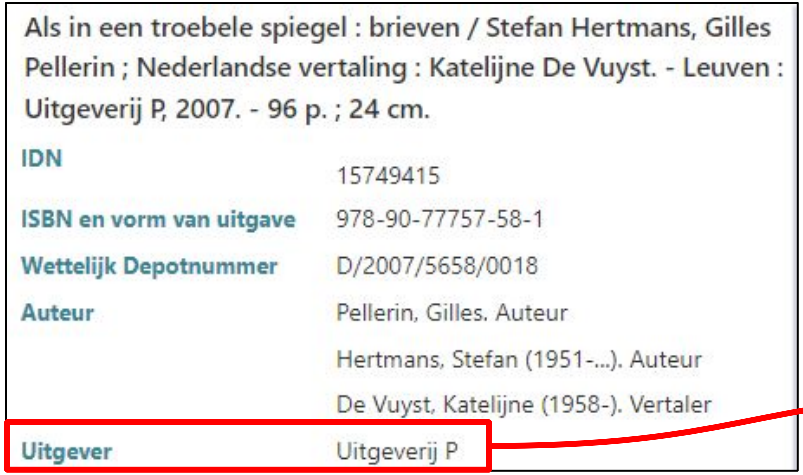
IDN	15151941
Auteur	Déchanet, Jean-Marie (1906-1992) - O.S.B.. Auteur Sullivan, Jean. Auteur van voorwoord, inleiding, etc Herkes, Ed. (19..-). Vertaler
Uitgever	Uitgeverij Emmaus



```
246(1#) $aVa où ton coeur te mène $iTitre original / oorspronkelijke titel
264(#1) $aBrugge $bUitgeverij Emmaus $c1972
300(##) $a147 p. $c20 cm
700(1#) $*14109150 $aDéchanet, Jean-Marie $cO.S.B. $d1906-1992 $4aut
700(1#) $*14647172 $aSullivan, Jean $4aui
700(1#) $*14122319 $aHerkes, Ed. $d19..- $4trl
```

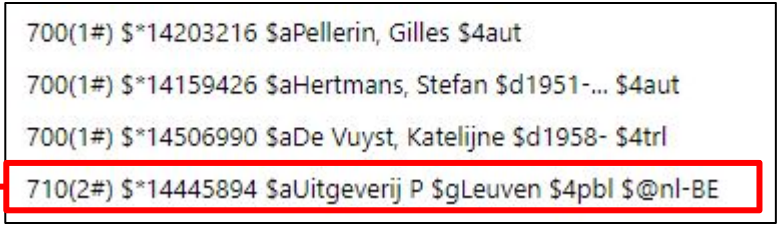
Data quality is use case specific - similar quality in MARC

use case *"Where is the publisher from?"*



Als in een troebele spiegel : brieven / Stefan Hertmans, Gilles Pellerin ; Nederlandse vertaling : Katelijne De Vuyst. - Leuven : Uitgeverij P, 2007. - 96 p. ; 24 cm.

IDN	15749415
ISBN en vorm van uitgave	978-90-77757-58-1
Wettelijk Depotnummer	D/2007/5658/0018
Auteur	Pellerin, Gilles. Auteur Hertmans, Stefan (1951-...). Auteur De Vuyst, Katelijne (1958-). Vertaler
Uitgever	Uitgeverij P

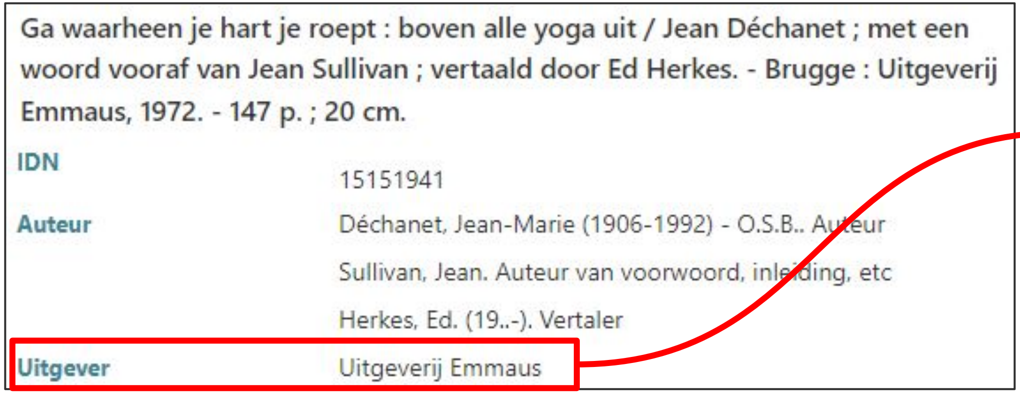


700(1#) $*14203216 $aPellerin, Gilles $4aut

700(1#) $*14159426 $aHertmans, Stefan $d1951-... $4aut

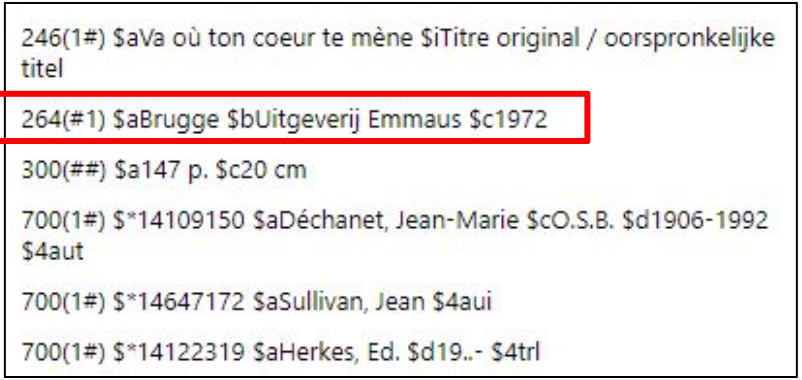
700(1#) $*14506990 $aDe Vuyst, Katelijne $d1958- $4trl

710(2#) $*14445894 $aUitgeverij P $gLeuven $4pbl $@nl-BE



Ga waarheen je hart je roept : boven alle yoga uit / Jean Déchanet ; met een woord vooraf van Jean Sullivan ; vertaald door Ed Herkes. - Brugge : Uitgeverij Emmaus, 1972. - 147 p. ; 20 cm.

IDN	15151941
Auteur	Déchanet, Jean-Marie (1906-1992) - O.S.B.. Auteur Sullivan, Jean. Auteur van voorwoord, inleiding, etc Herkes, Ed. (19..-). Vertaler
Uitgever	Uitgeverij Emmaus



246(1#) $aVa où ton coeur te mène $iTitre original / oorspronkelijke titel

264(#1) $aBrugge $bUitgeverij Emmaus $c1972

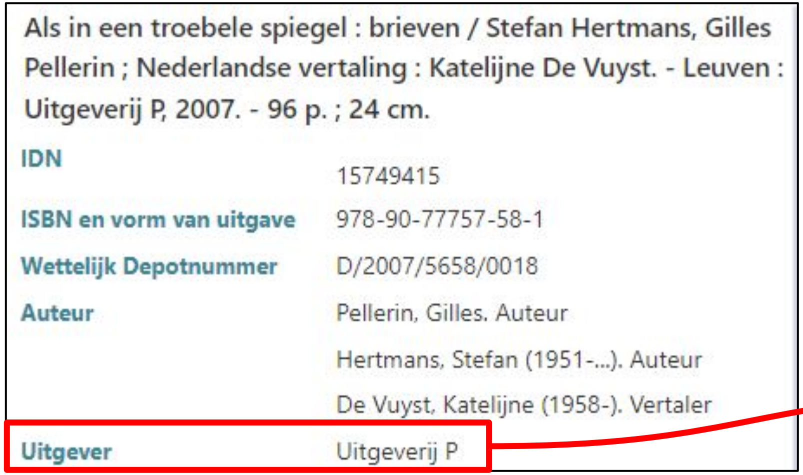
300(##) $a147 p. $c20 cm

700(1#) $*14109150 $aDéchanet, Jean-Marie $cO.S.B. $d1906-1992 $4aut

700(1#) $*14647172 $aSullivan, Jean $4aui

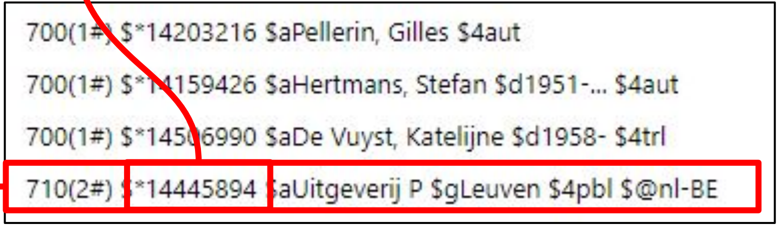
700(1#) $*14122319 $aHerkes, Ed. $d19..- $4trl

Data quality is use case specific - quality difference



Als in een troebele spiegel : brieven / Stefan Hertmans, Gilles Pellerin ; Nederlandse vertaling : Katelijne De Vuyst. - Leuven : Uitgeverij P, 2007. - 96 p. ; 24 cm.

IDN	15749415
ISBN en vorm van uitgave	978-90-77757-58-1
Wettelijk Depotnummer	D/2007/5658/0018
Auteur	Pellerin, Gilles. Auteur Hertmans, Stefan (1951-...). Auteur De Vuyst, Katelijne (1958-). Vertaler
Uitgever	Uitgeverij P

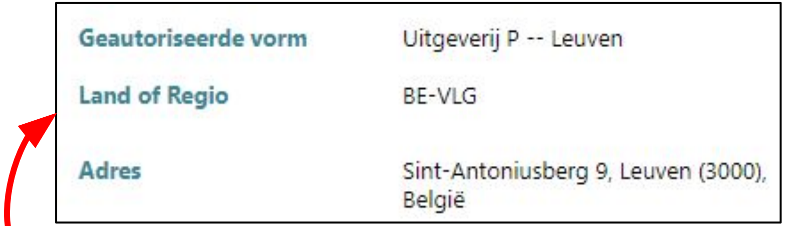


700(1#) $*14203216 $aPellerin, Gilles $4aut

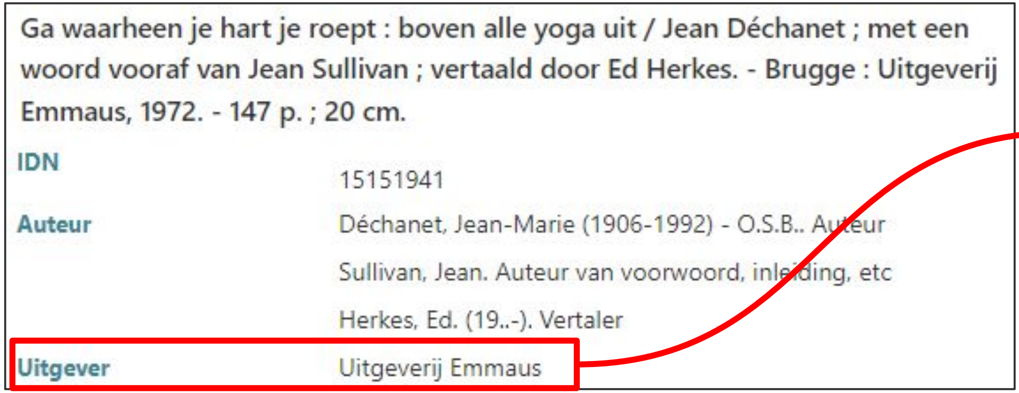
700(1#) $*14159426 $aHertmans, Stefan $d1951-... $4aut

700(1#) $*14506990 $aDe Vuyst, Katelijne $d1958- $4trl

710(2#) $*14445894 $aUitgeverij P $gLeuven $4pbl $@nl-BE

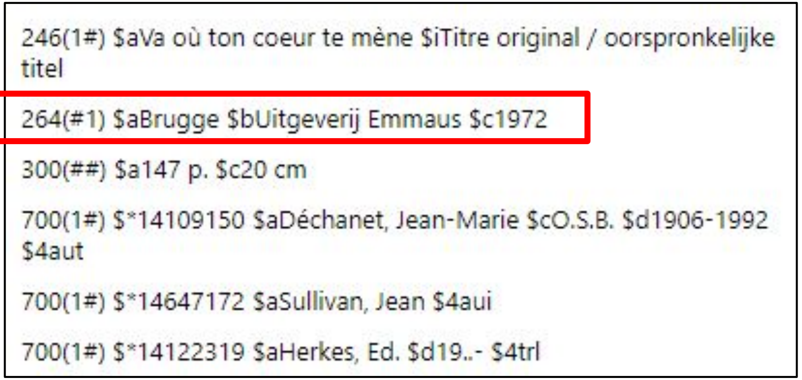


Geautoriseerde vorm	Uitgeverij P -- Leuven
Land of Regio	BE-VLG
Adres	Sint-Antoniusberg 9, Leuven (3000), België



Ga waarheen je hart je roept : boven alle yoga uit / Jean Déchanet ; met een woord vooraf van Jean Sullivan ; vertaald door Ed Herkes. - Brugge : Uitgeverij Emmaus, 1972. - 147 p. ; 20 cm.

IDN	15151941
Auteur	Déchanet, Jean-Marie (1906-1992) - O.S.B.. Auteur Sullivan, Jean. Auteur van voorwoord, inleiding, etc Herkes, Ed. (19..-). Vertaler
Uitgever	Uitgeverij Emmaus



246(1#) $aVa où ton coeur te mène $iTitre original / oorspronkelijke titel

264(#1) $aBrugge $bUitgeverij Emmaus $c1972

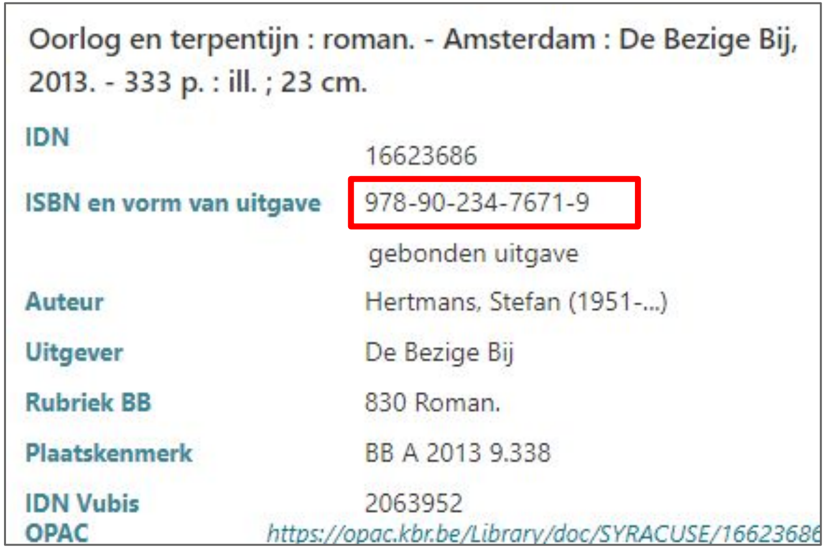
300(##) $a147 p. $c20 cm

700(1#) $*14109150 $aDéchanet, Jean-Marie $cO.S.B. $d1906-1992 $4aut

700(1#) $*14647172 $aSullivan, Jean $4aui

700(1#) $*14122319 $aHerkes, Ed. $d19..- $4trl

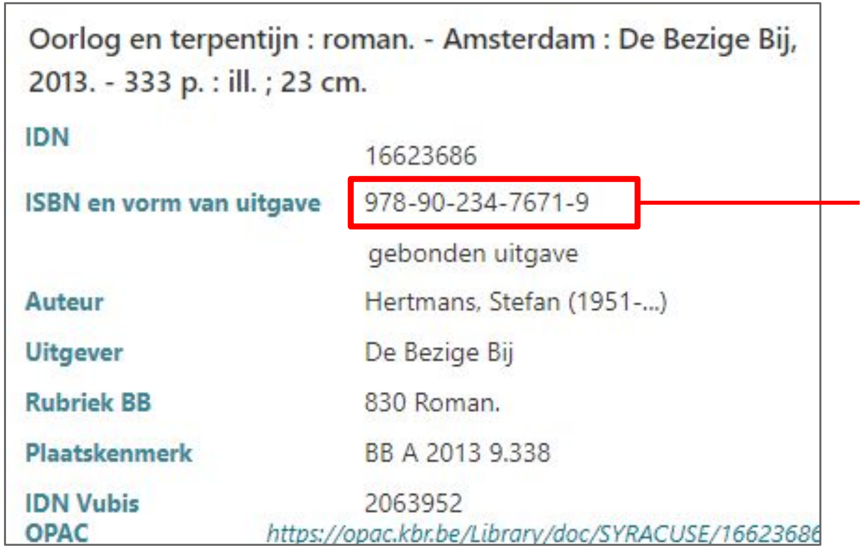
Incorrect data which can be detected based on syntax rules



Oorlog en terpentijn : roman. - Amsterdam : De Bezige Bij, 2013. - 333 p. : ill. ; 23 cm.

IDN	16623686
ISBN en vorm van uitgave	978-90-234-7671-9 gebonden uitgave
Auteur	Hertmans, Stefan (1951-...)
Uitgever	De Bezige Bij
Rubriek BB	830 Roman.
Plaatskenmerk	BB A 2013 9.338
IDN Vubis	2063952
OPAC	https://opac.kbr.be/Library/doc/SYRACUSE/16623686

Incorrect data which can be detected based on syntax rules

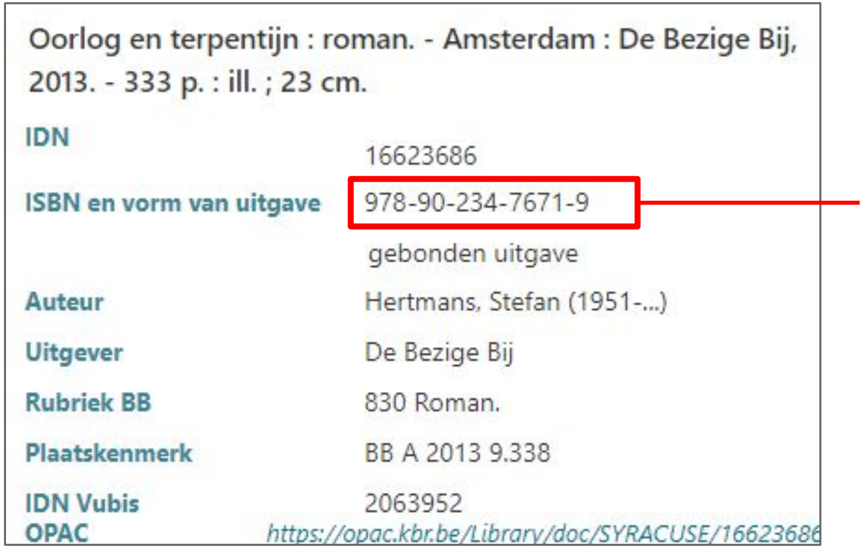


Oorlog en terpentijn : roman. - Amsterdam : De Bezige Bij, 2013. - 333 p. : ill. ; 23 cm.

IDN	16623686
ISBN en vorm van uitgave	978-90-234-7671-9 gebonden uitgave
Auteur	Hertmans, Stefan (1951-...)
Uitgever	De Bezige Bij
Rubriek BB	830 Roman.
Plaatskenmerk	BB A 2013 9.338
IDN Vubis	2063952
OPAC	https://opac.kbr.be/Library/doc/SYRACUSE/16623686

ISBN is valid

Incorrect data which can be detected based on syntax rules



Oorlog en terpentijn : roman. - Amsterdam : De Bezige Bij, 2013. - 333 p. : ill. ; 23 cm.

IDN	16623686
ISBN en vorm van uitgave	978-90-234-7671-9
	gebonden uitgave
Auteur	Hertmans, Stefan (1951-...)
Uitgever	De Bezige Bij
Rubriek BB	830 Roman.
Plaatskenmerk	BB A 2013 9.338
IDN Vubis	2063952
OPAC	https://opac.kbr.be/Library/doc/SYRACUSE/16623686

ISBN is valid

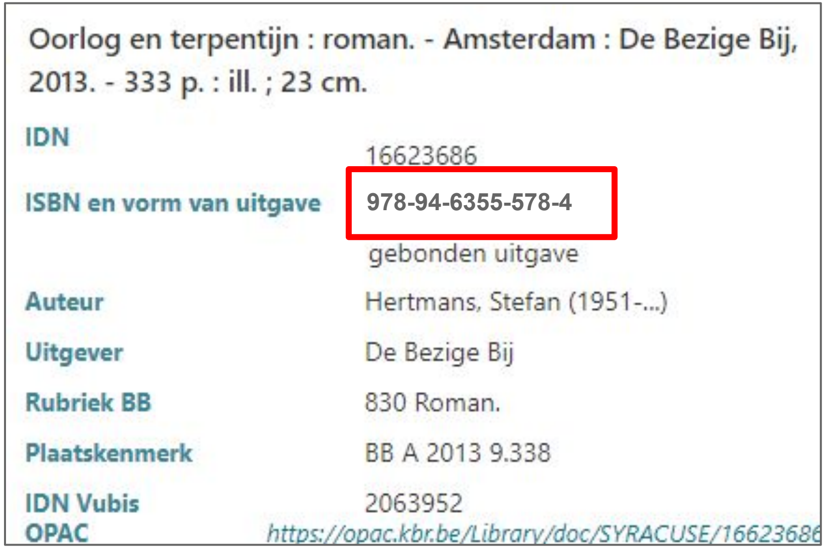
978-90-234-767**2**-9

ISBN is invalid:
wrong check digit

KBR

Koester de tijd
Protégeons le temps

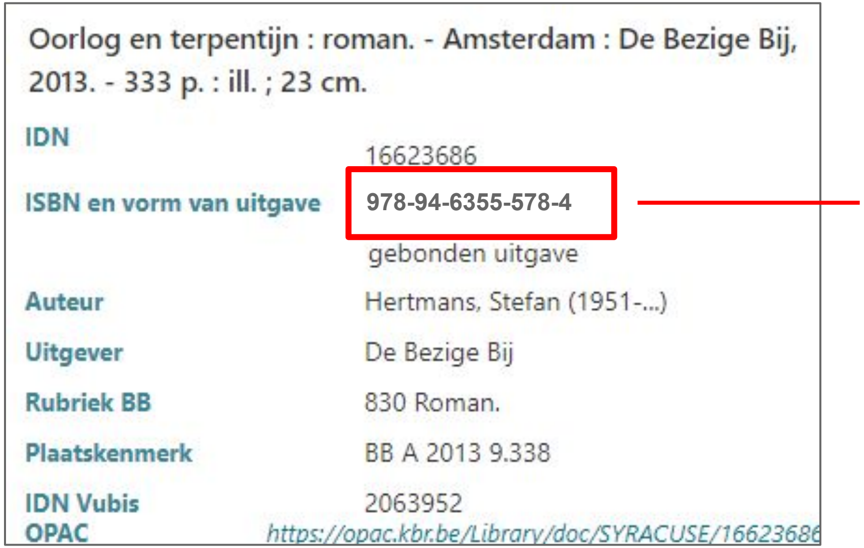
Incorrect data which needs to be checked manually



Oorlog en terpentijn : roman. - Amsterdam : De Bezige Bij, 2013. - 333 p. : ill. ; 23 cm.

IDN	16623686
ISBN en vorm van uitgave	978-94-6355-578-4 gebonden uitgave
Auteur	Hertmans, Stefan (1951-...)
Uitgever	De Bezige Bij
Rubriek BB	830 Roman.
Plaatskenmerk	BB A 2013 9.338
IDN Vubis	2063952
OPAC	https://opac.kbr.be/Library/doc/SYRACUSE/16623686

Incorrect data which needs to be checked manually



Oorlog en terpentijn : roman. - Amsterdam : De Bezige Bij, 2013. - 333 p. : ill. ; 23 cm.

IDN	16623686
ISBN en vorm van uitgave	978-94-6355-578-4 gebonden uitgave
Auteur	Hertmans, Stefan (1951-...)
Uitgever	De Bezige Bij
Rubriek BB	830 Roman.
Plaatskenmerk	BB A 2013 9.338
IDN Vubis	2063952
OPAC	https://opac.kbr.be/Library/doc/SYRACUSE/16623686

ISBN is valid

… but actually belongs to a different book

What is data quality, what is correct and what is valid?

Current quality procedures at KBR

The use case: BELTRANS project

Our **L**ibrarian-**I**n-**T**he-**L**oop workflow

Lessons learned and future work

Current quality procedures at KBR

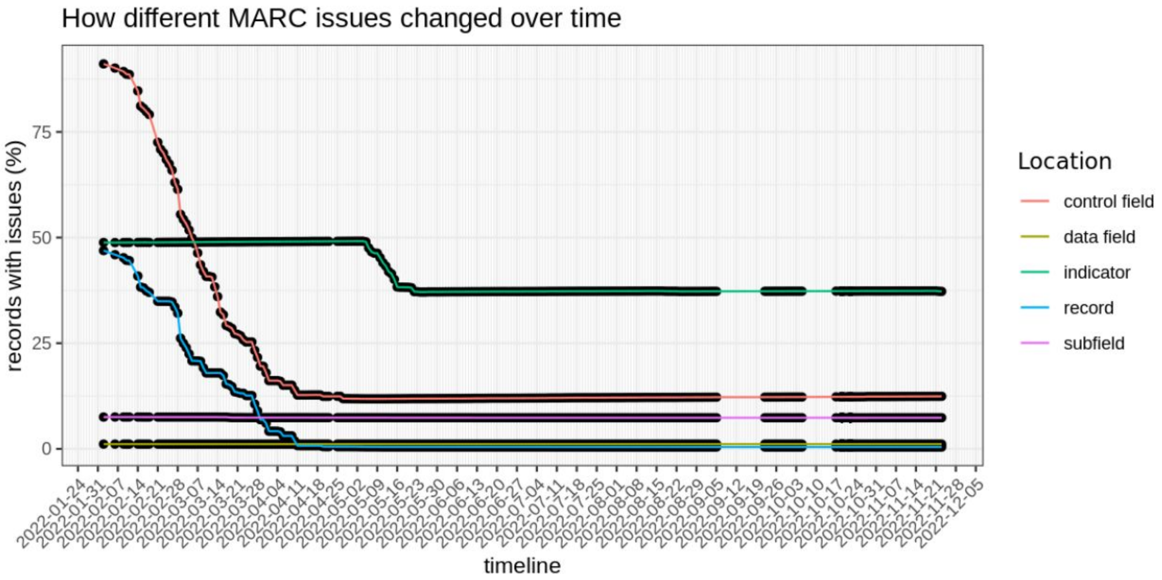
Librarians manually adding rich descriptions (+ automatic ISBN check, required fields, …)

Current quality procedures at KBR

Librarians manually adding rich descriptions (+ automatic ISBN check, required fields, …)

Conformity to the MARC standard, QA catalogue tool from Péter Király

Daily conformity checks with the MARC standard, QA catalogue tool from Péter Király



Péter Király & Hannes Lowagie. "Implementation of a daily MARC assessment with open source tools at KBR, the royal library of Belgium."
IFLA Metadata Newsletter, Volume 8, Number 1, June 2022. pp. 12-15
https://repository.ifla.org/handle/123456789/1976

KBR Koester de tijd
Protégeons le temps

Current quality procedures at KBR

Librarians manually adding rich descriptions (+ automatic ISBN check, required fields, …)

Conformity to the MARC standard, QA catalogue tool from Péter Király

Increased use of standard ISNI identifier

ISNI is a standard name identifier, increasingly used at KBR

isni

ISNI is the ISO 27729:2012 **standard** that uniquely identifies **public identities** who contributed to **creative works**

ISNI is a standard name identifier, increasingly used at KBR

isni

ISNI is the ISO 27729:2012 **standard** that uniquely identifies **public identities** who contributed to **creative works**

In 2020, KBR became an ISNI registration agency

ISNI is a standard name identifier, increasingly used at KBR

ISNI is the ISO 27729:2012 **standard** that uniquely identifies **public identities** who contributed to **creative works**

In 2020, KBR became an ISNI registration agency

In 2021 KBRs data was aligned to the ISNI database via a bulk load submission
Goal of the bulk load submission was to match KBR records with ISNIs central database: **60% of the 900k** persons were **assigned an ISNI**

Current quality procedures at KBR

Librarians manually adding rich descriptions (+ automatic ISBN check, required fields, …)

Conformity to the MARC standard, QA catalogue tool from Péter Király

Increased use of standard ISNI identifier

Working groups in our agency for bibliographic information (URI, RDA, IFLA-LRM, …)

Use URIs in our multilingual catalog

Example: an authority who was **born in Antwerp** and who **died in Brussel**

Use URIs in our multilingual catalog

Example: an authority who was **born in Antwerp** and who **died in Brussel**

MARC field 370 (associated place)

Multilingual text: we need/want a unique URI

Example: an authority who was **born in Antwerp** and who **died in Brussel**

MARC field 370 (associated place)

```
<datafield tag=”370”>
    <subfield code=”a”>Antwerpen</subfield>
    <subfield code=”b”>Bruxelles</subfield>
</datafield>
```

Not clear which URI is which place

Example: an authority who was **born in Antwerp** and who **died in Brussel**

MARC field 370 (associated place)
Contains a single repeatable subfield $1 (Real World Object URI)

```
<datafield tag=”370”>
   <subfield code=”1”>https://sws.geonames.org/2803138/</subfield>
   <subfield code=”1”>https://sws.geonames.org/2800866/</subfield>
   <subfield code=”a”>Antwerpen</subfield>
   <subfield code=”b”>Bruxelles</subfield>
</datafield>
```

?

Use repeatable 370 for place of birth and place of death

Example: an authority who was **born in Antwerp** and who **died in Brussel**

MARC field 370 (associated place)
Contains a single repeatable subfield $1 (Real World Object URI)

```
<datafield tag=”370”>
   <subfield code=”1”>https://sws.geonames.org/2803138/</subfield>
    <subfield code=”a”>Antwerpen</subfield>
</datafield>
```

```
<datafield tag=”370”>
   <subfield code=”1”>https://sws.geonames.org/2800866/</subfield>
    <subfield code=”b”>Bruxelles</subfield>
</datafield>
```

What is data quality, what is correct and what is valid?

Current quality procedures at KBR

The use case: BELTRANS project

Our **L**ibrarian-**I**n-**T**he-**L**oop workflow

Lessons learned and future work

BELTRANS project: studying book translations flows

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

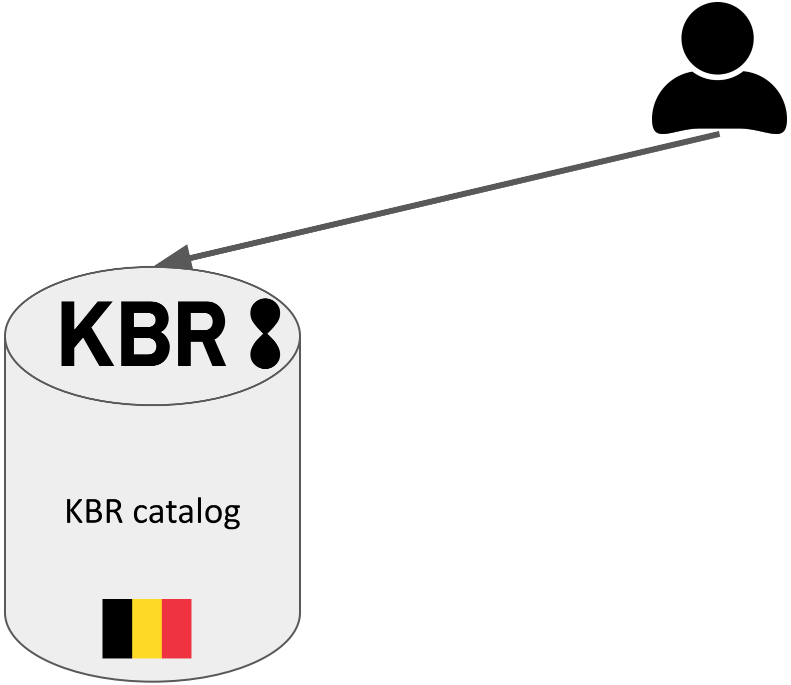
KBR Koester de tijd
Protégeons le temps

KU LEUVEN

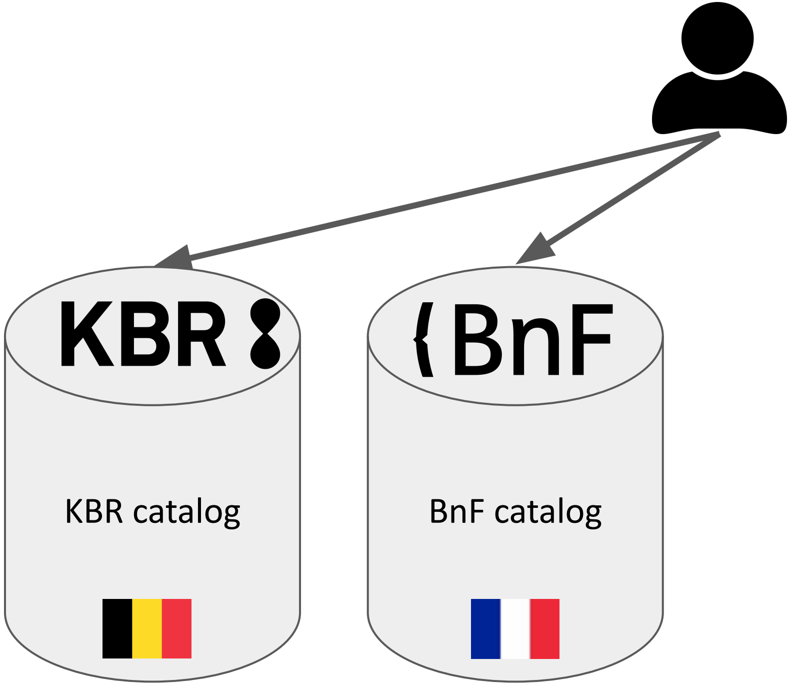
UCLouvain

belspo

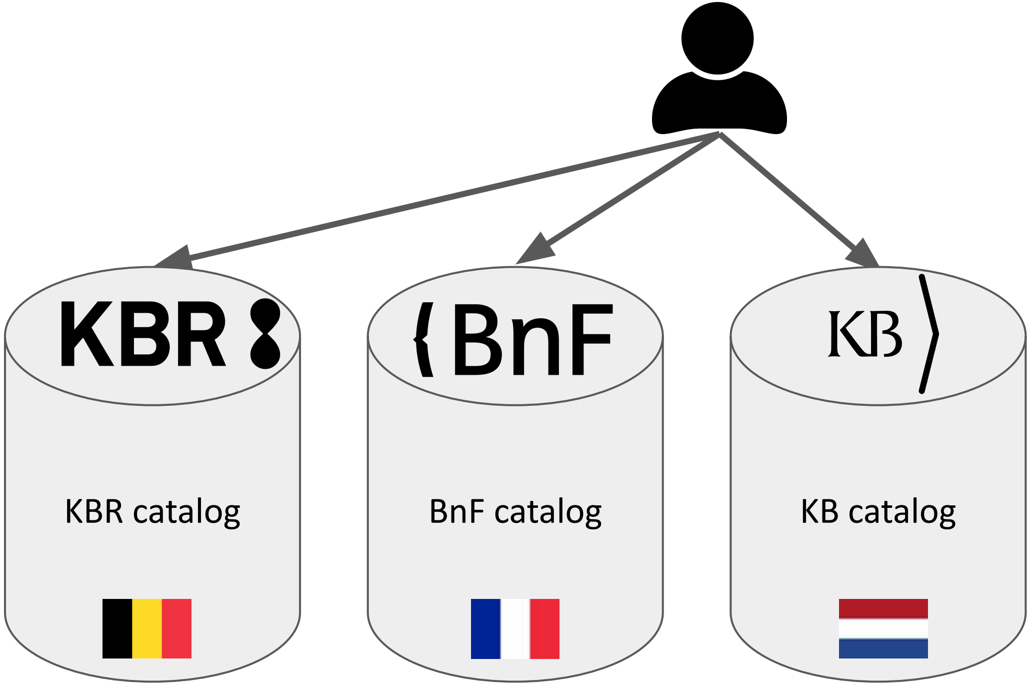
Different data *sources* in different *formats* of different *quality*



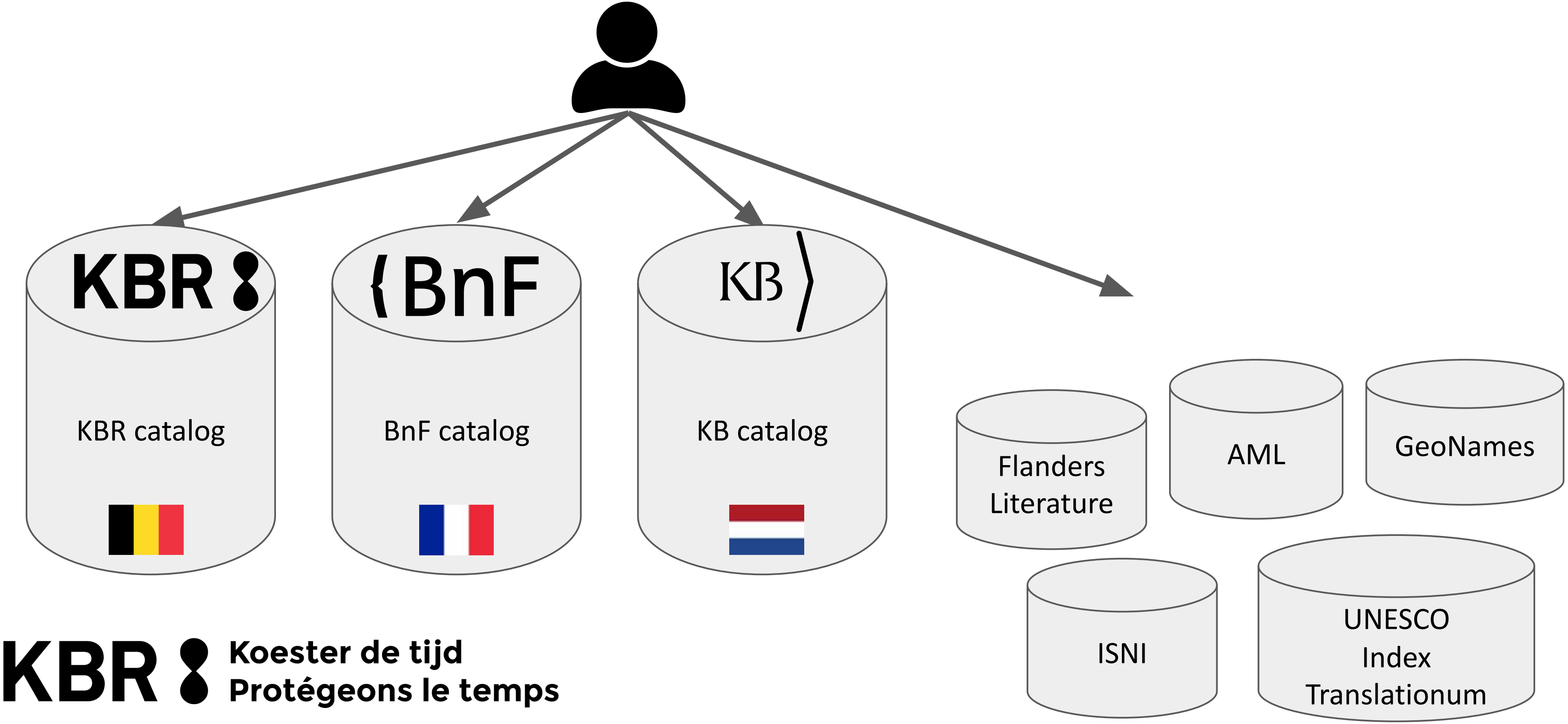
Different data *sources* in different *formats* of different *quality*



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



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

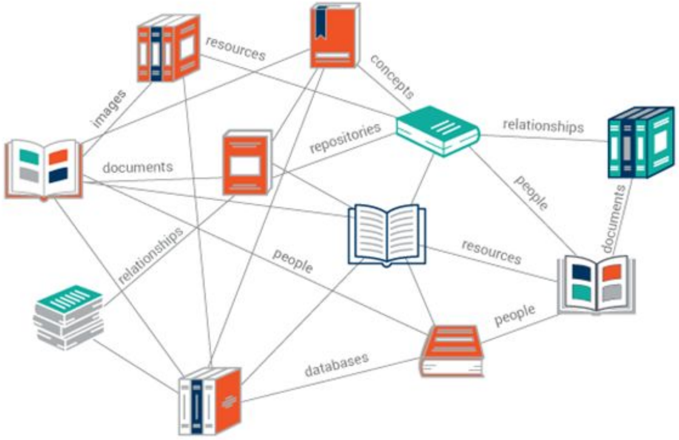


KBR Koester de tijd
Protégeons le temps

We developed a pipeline with *different* processing per data source to create integrated data

rml.io

R D F



KBR

KBR catalog

BnF

BnF catalog

KB

KB catalog

Flanders Literature

AML

GeoNames

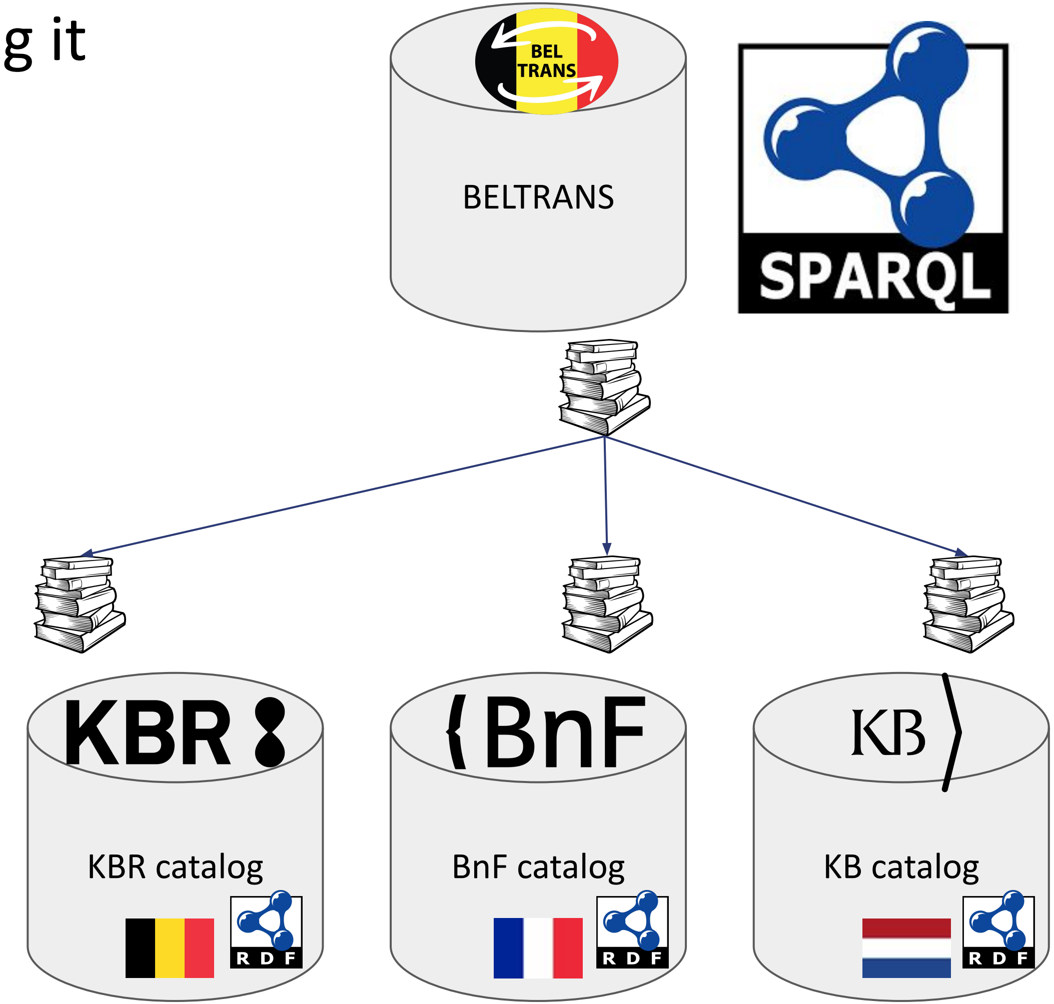
...

UNESCO Index Translationum

KBR Koester de tijd
Protégeons le temps

Integrate data by interlinking it with *standard identifiers*

schema:sameAs



KBR Koester de tijd
Protégeons le temps

Integrate data by interlinking it with *standard identifiers*

isbn

9 782871 424543

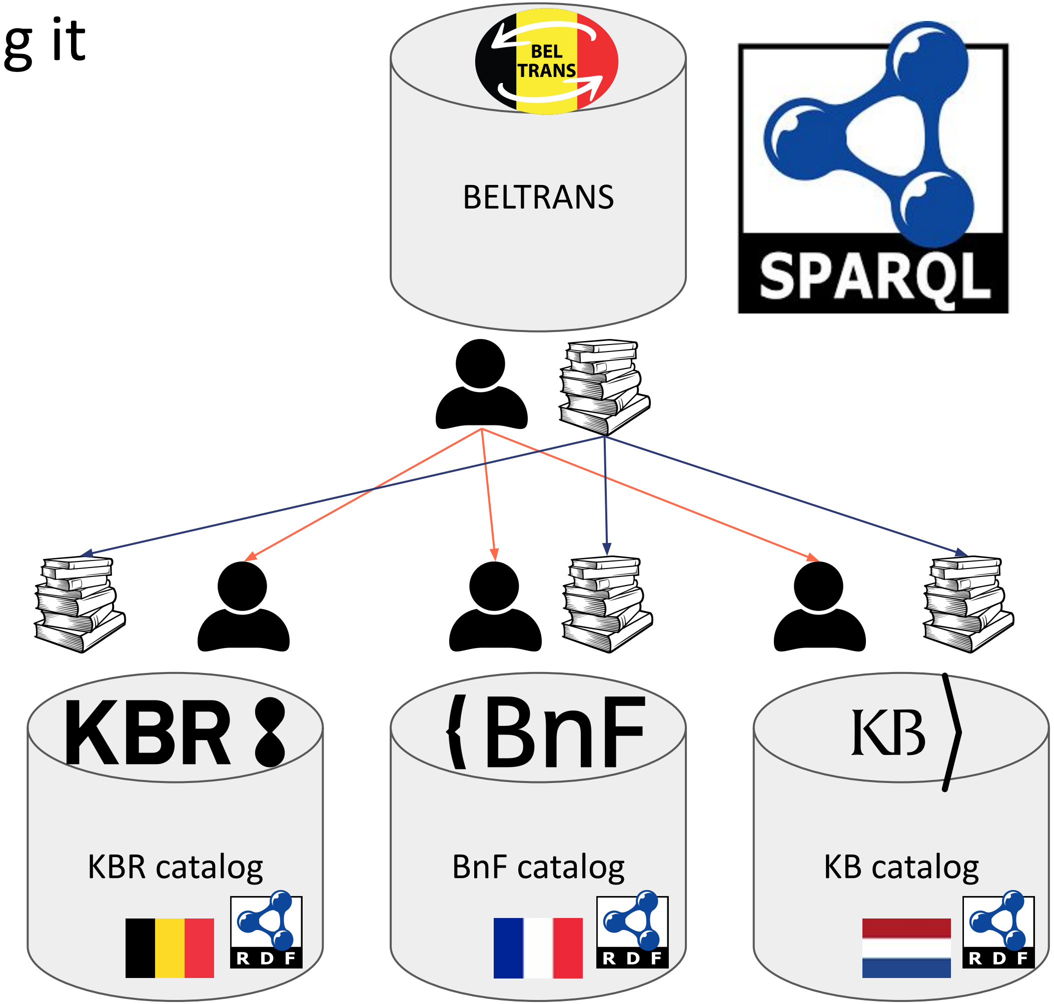
schema:sameAs

isni

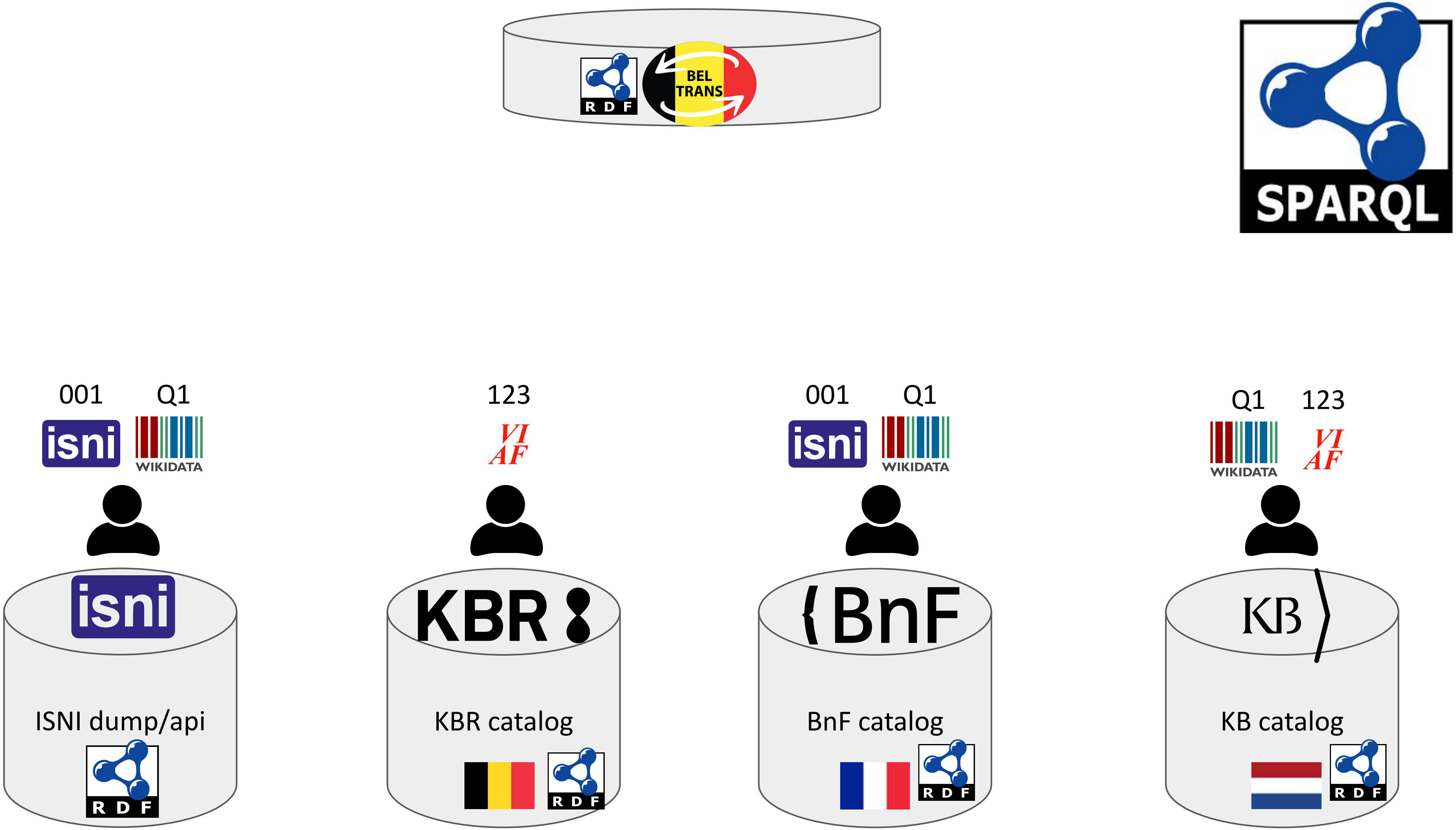
VIAF

WIKIDATA

schema:sameAs



KBR Koester de tijd
Protégeons le temps

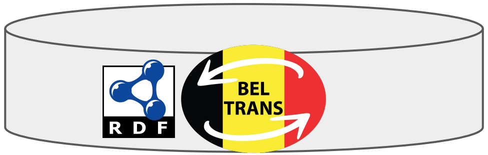


ISNI dump/api

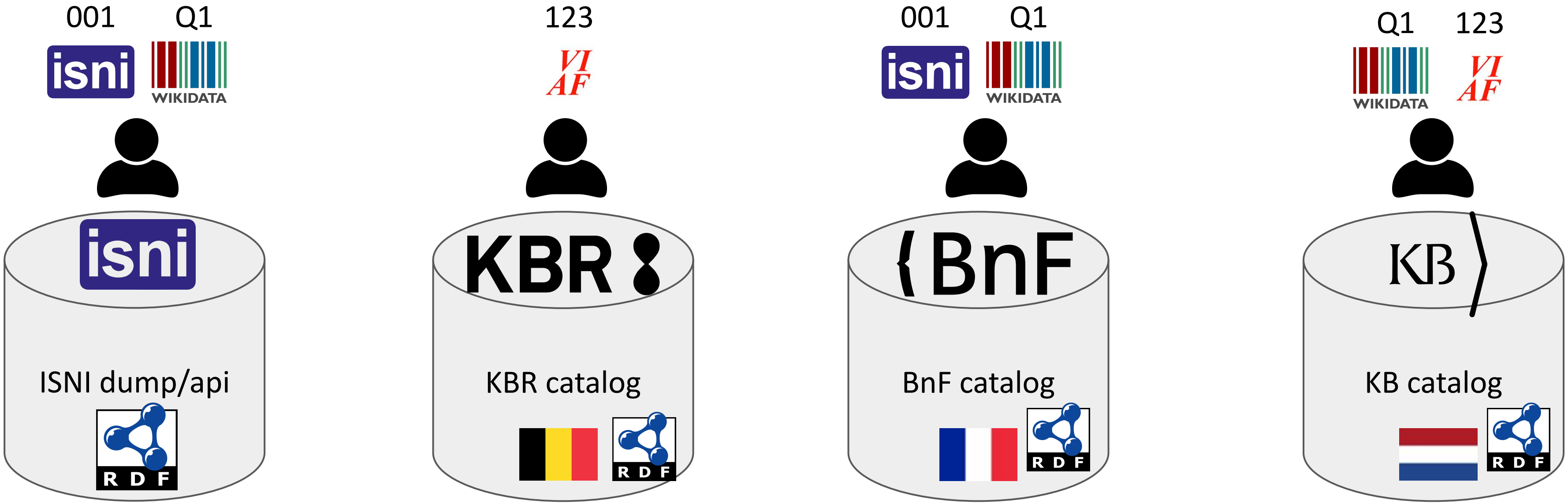
KBR catalog

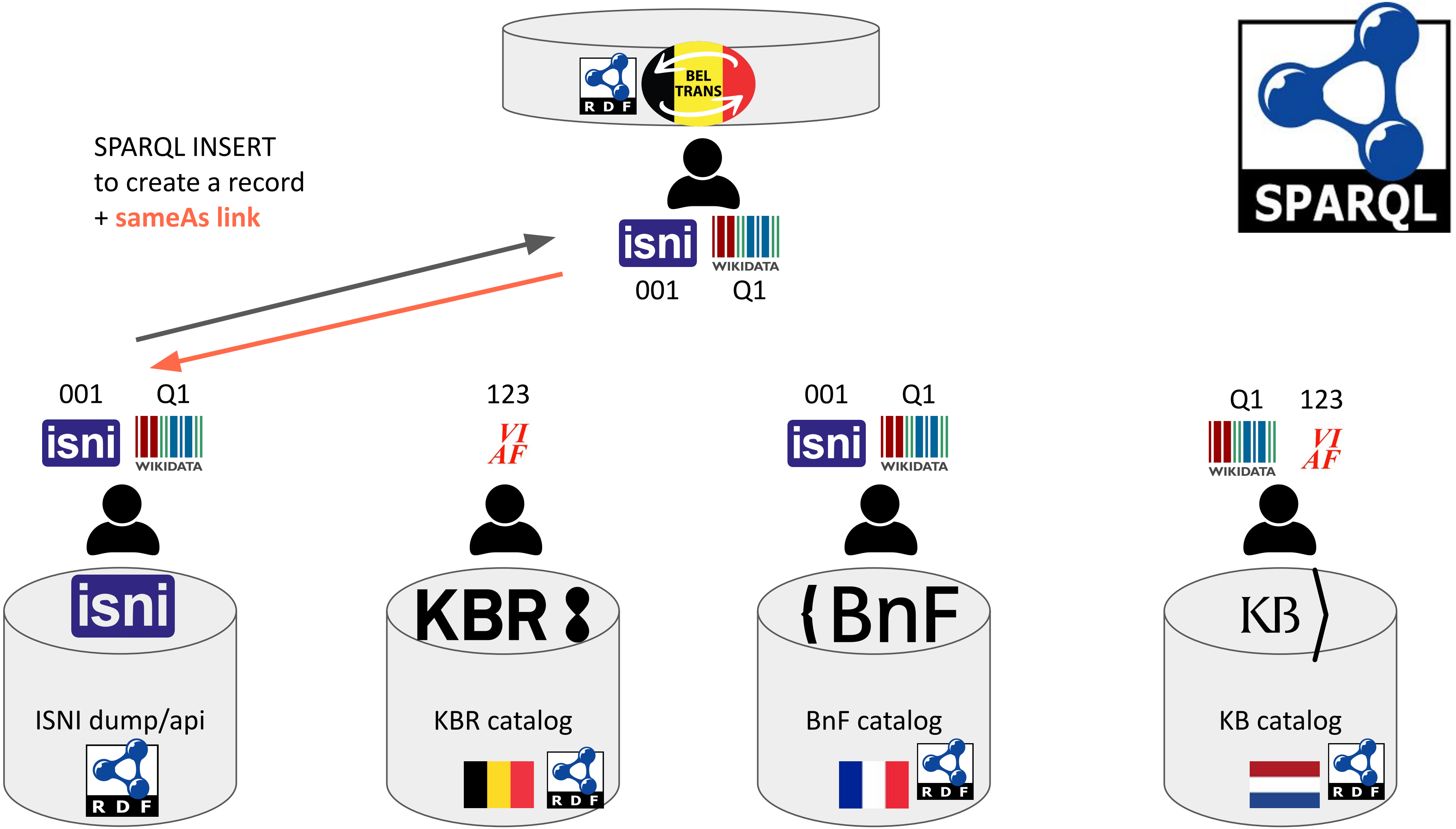
BnF catalog

KB catalog



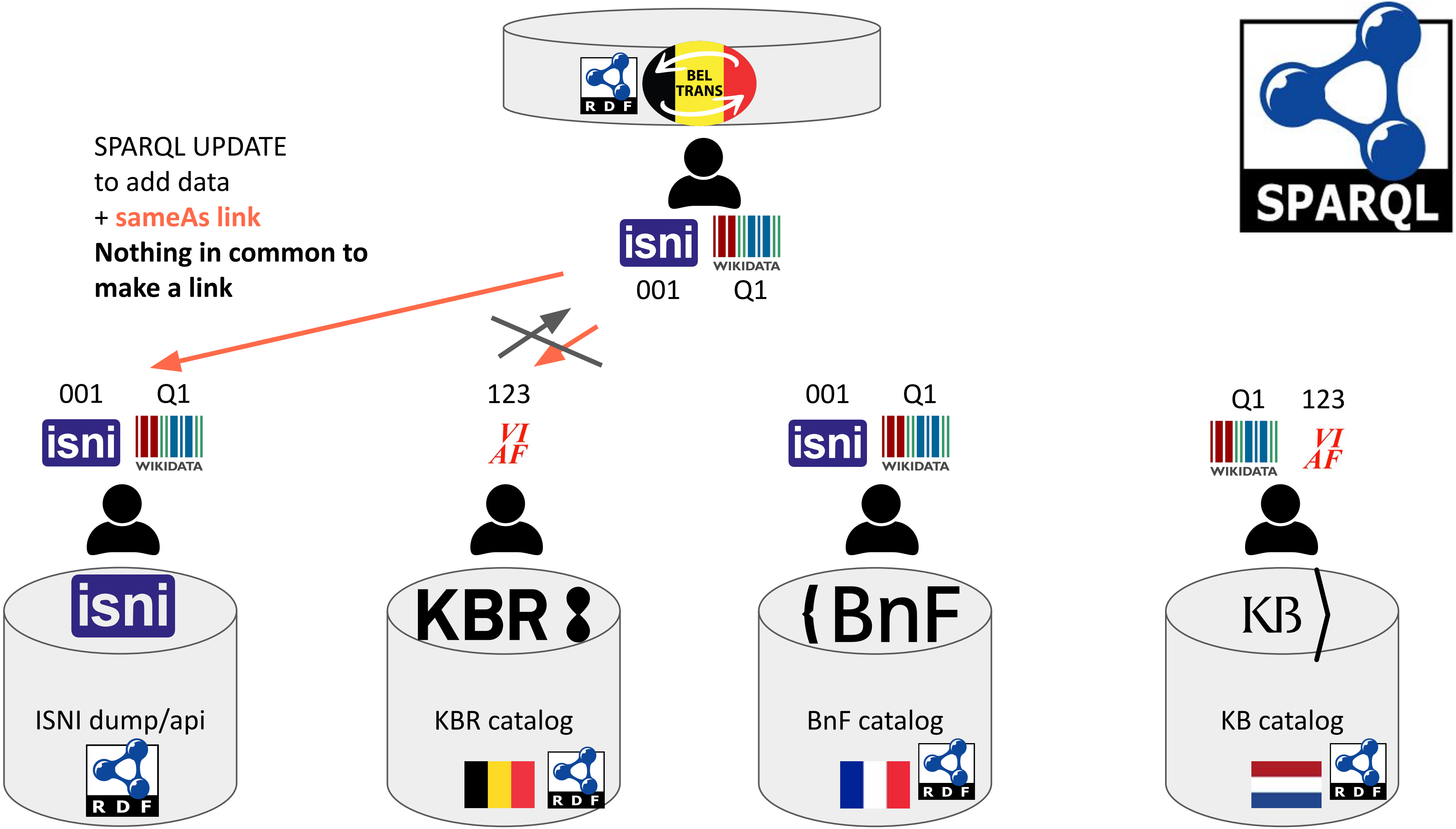
Python script performing SPARQL INSERT per data source and subsequent SPARQL UPDATE queries from other sources

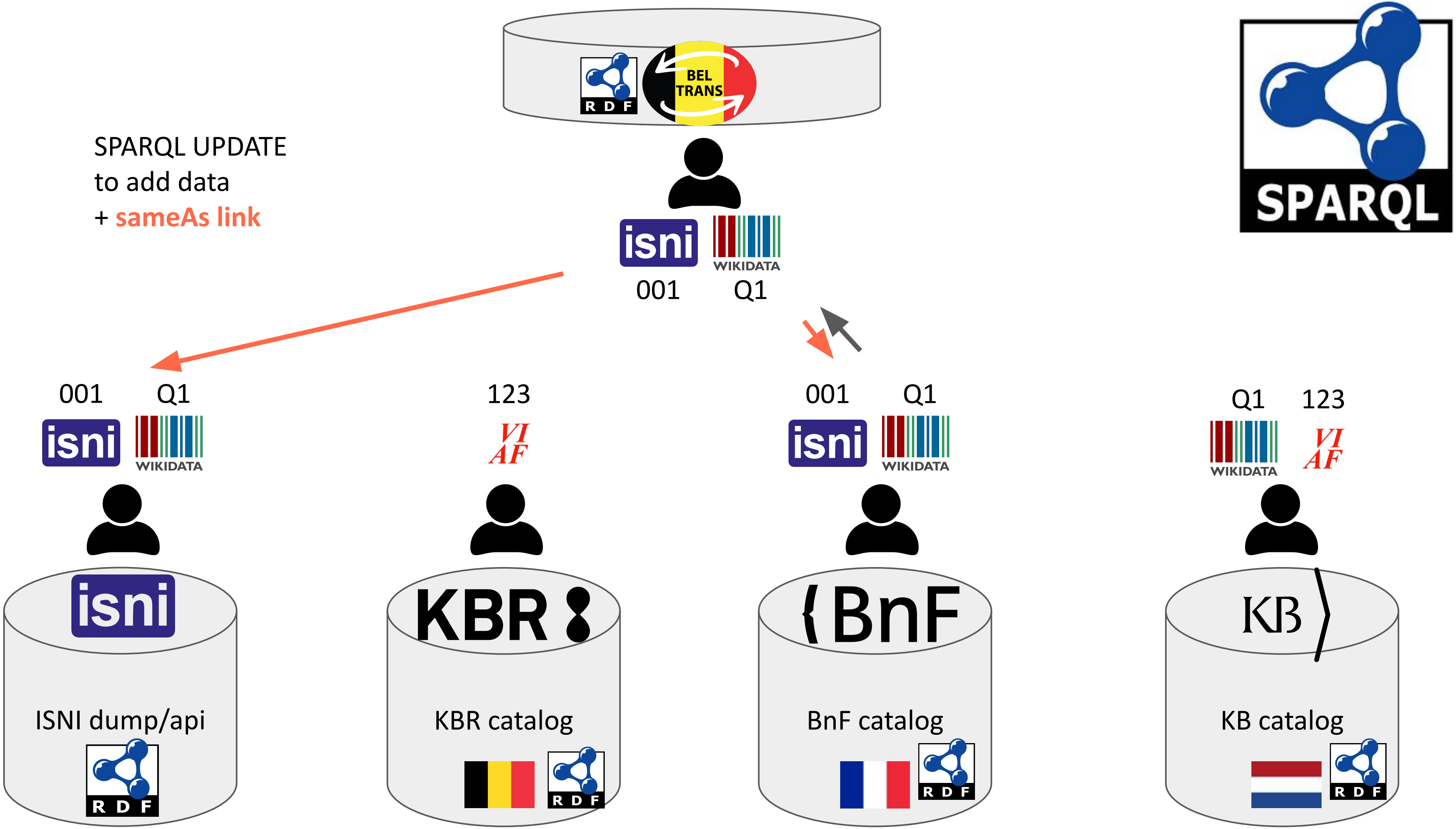




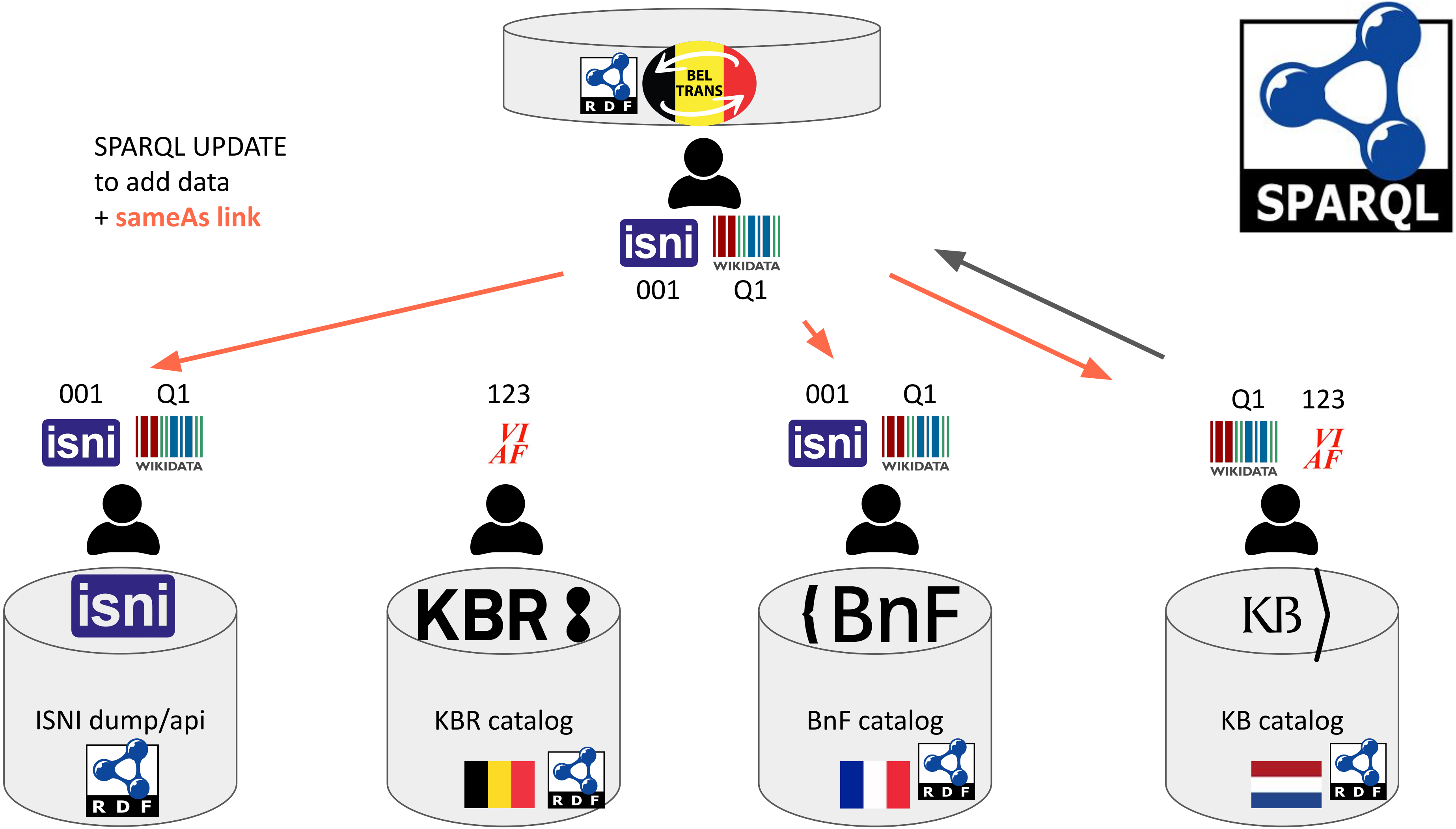
SPARQL INSERT
to create a record
+ sameAs link

SPARQL UPDATE
to add data
+ sameAs link
Nothing in common to make a link

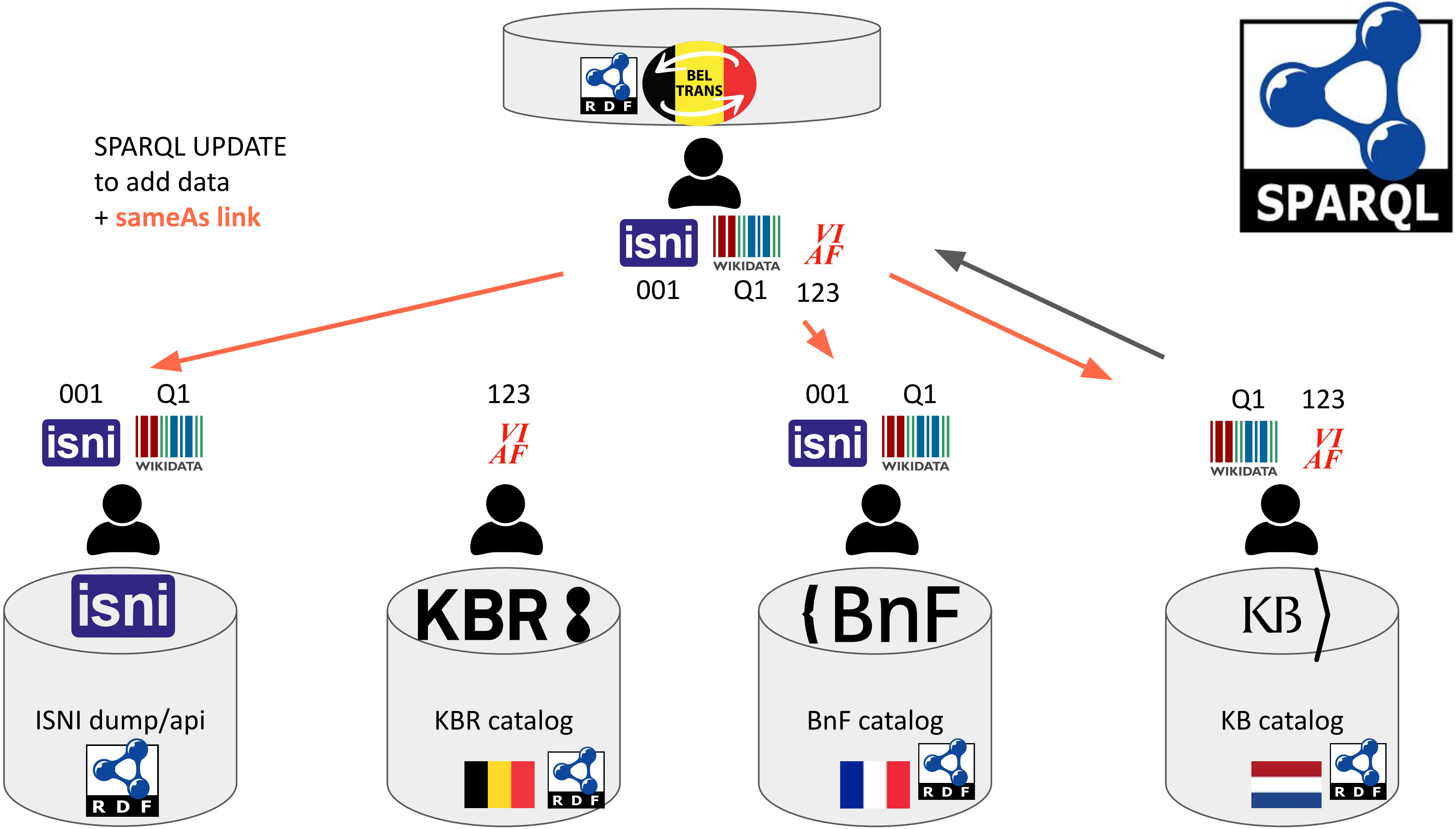




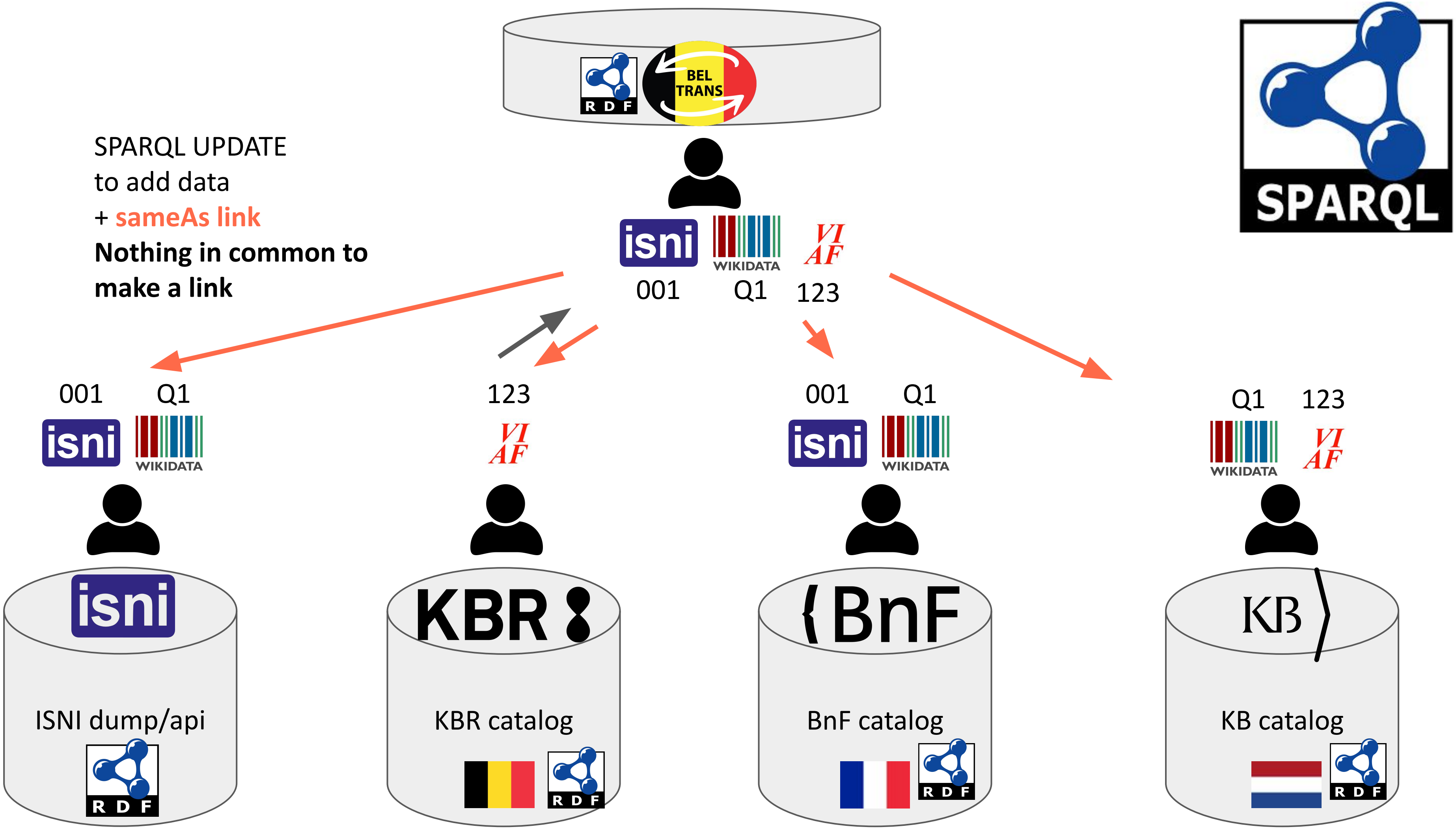
SPARQL UPDATE
to add data
+ sameAs link



SPARQL UPDATE
to add data
+ sameAs link

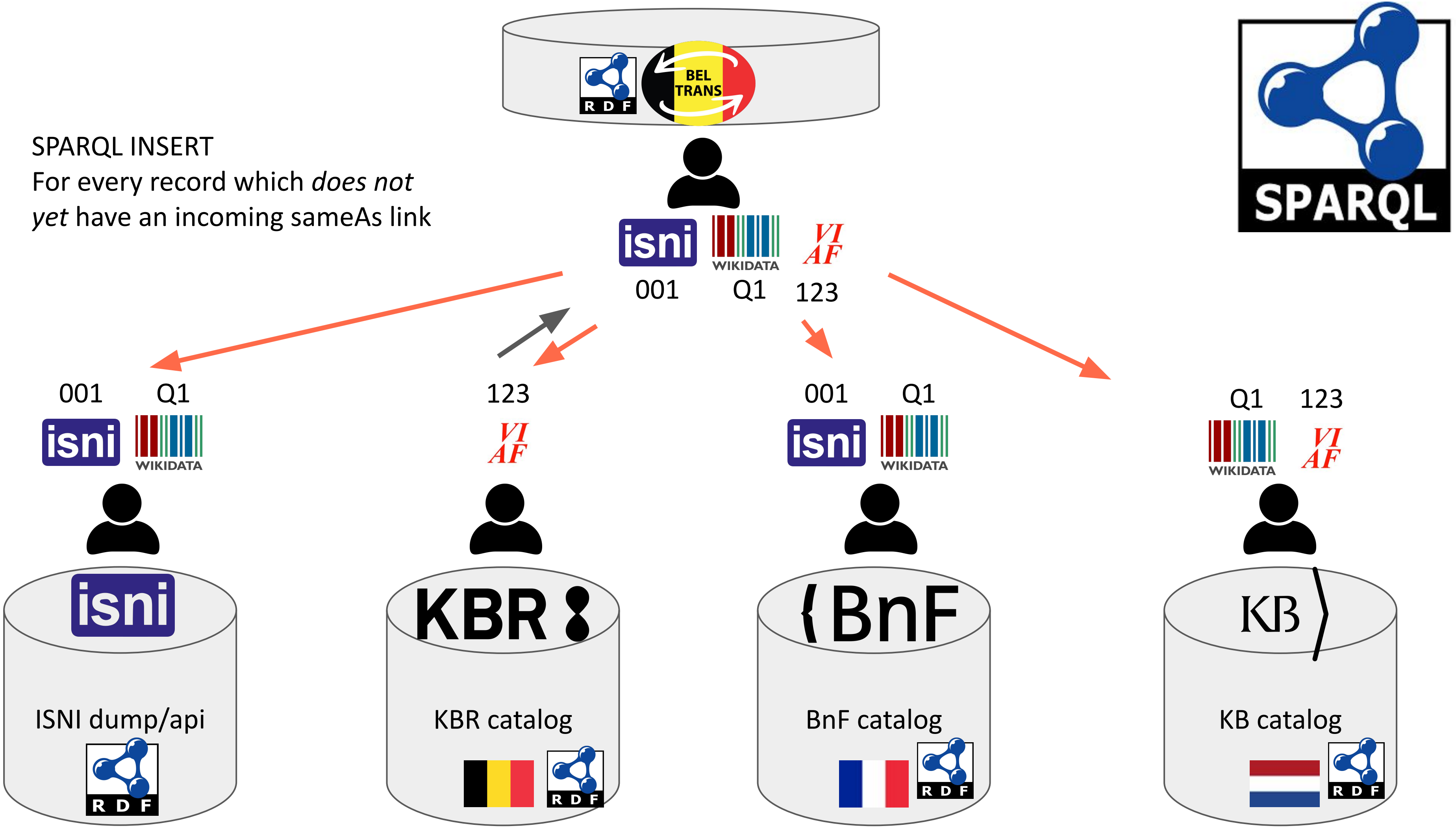


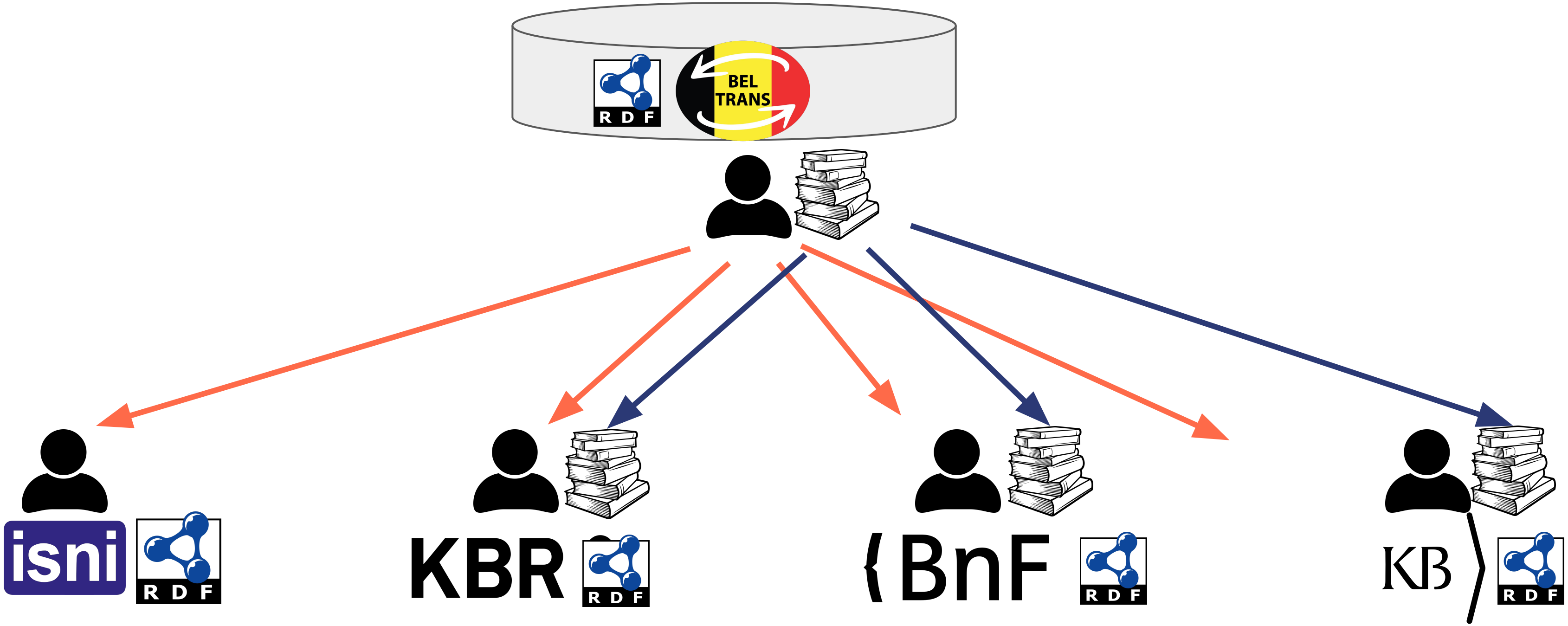
SPARQL UPDATE
to add data
+ sameAs link

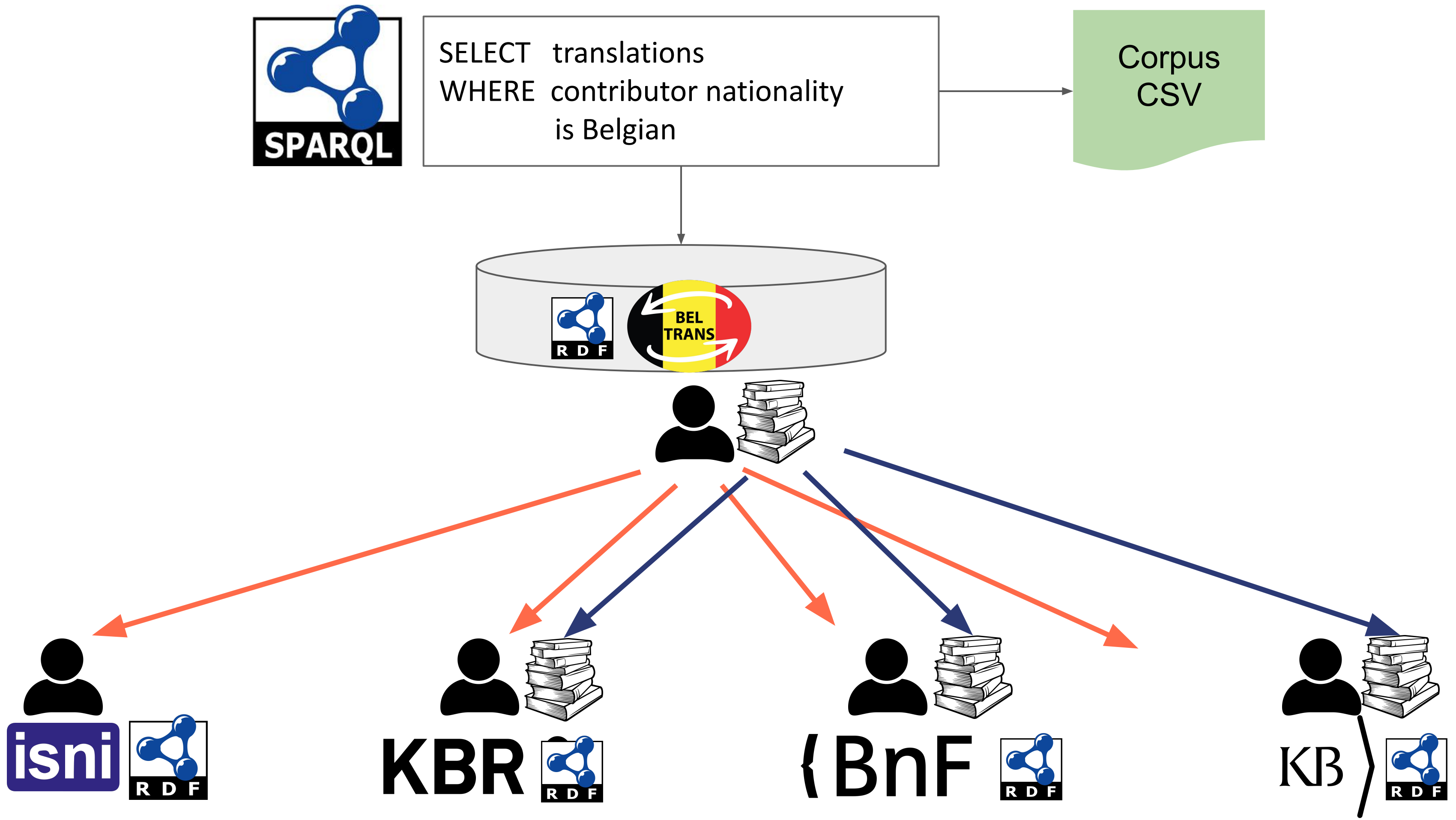


SPARQL UPDATE
to add data
+ sameAs link
Nothing in common to make a link

SPARQL INSERT
For every record which *does not yet* have an incoming sameAs link




RDF
BEL TRANS
isni
RDF
KBR
RDF
BnF
RDF
KB
RDF



SELECT translations
WHERE contributor nationality
is Belgian

Corpus
CSV

Postprocess the data and create a lightweight CSV file

ID	title	year Of Publication KBR	Year Of Publication BnF	Year Of Publication KB	…
1	Book1	2020	2020	2019	
2	Book2	2019		2019	

Postprocess the data and create a lightweight CSV file

ID	title	year Of Publication KBR	Year Of Publication BnF	Year Of Publication KB	…
1	Book1	2020	2020	2019	
2	Book2	2019		2019	

ID	title	year Of Publication	…
1	Book1	2019 or 2020	
2	Book2	2019	

Merge data and report inconsistencies

What is data quality, what is correct and what is valid?

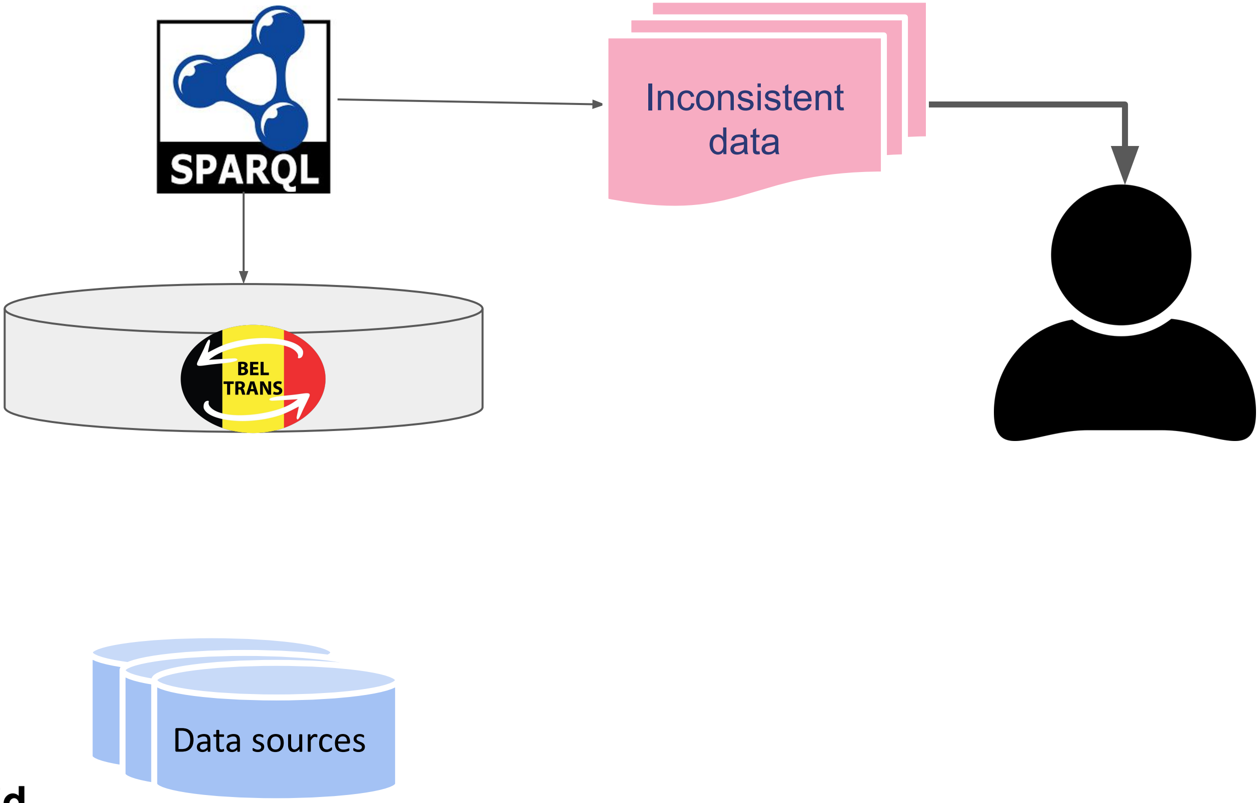
Current quality procedures at KBR

The use case: BELTRANS project

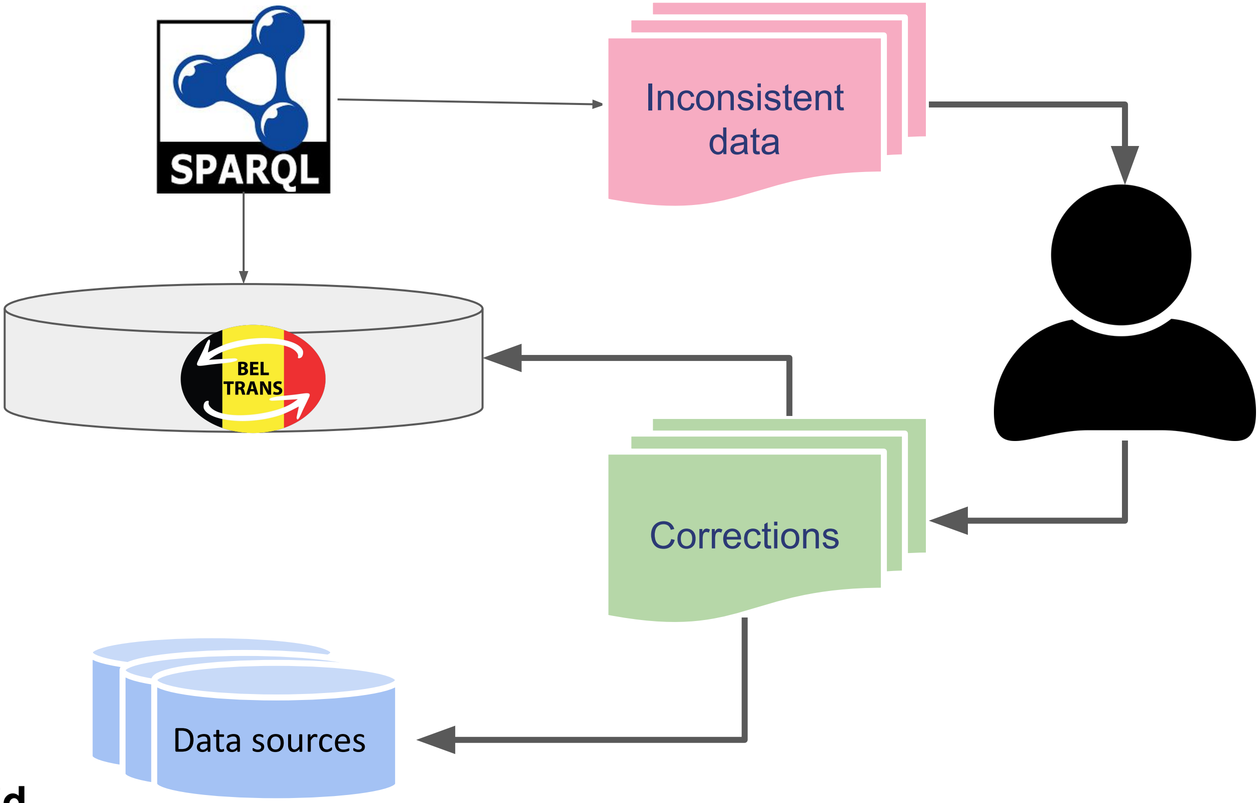
Our Librarian-In-The-Loop workflow

Lessons learned and future work

The Librarian-In-The-Loop Workflow: CSV input files



The Librarian-In-The-Loop Workflow: CSV output files



KBR Koester de tijd
Protégeons le temps

Data Quality Assessment

Phase 1: Requirements Analysis

Use Case Analysis

Phase 2: Quality Assessment

Identification of quality issues

Analysis

Phase 3: Quality Improvement

Root cause analysis

Fixing quality problems

Rula, Anisa, and Amrapali Zaveri. "Methodology for Assessment of Linked Data Quality." *LDQ@ SEMANTiCS*. 2014.

Data Quality Assessment

Phase 1: Requirements Analysis

Use Case Analysis

Phase 2: Quality Assessment

Identification of quality issues

Analysis

Phase 3: Quality Improvement

Root cause analysis

Fixing quality problems

Rula, Anisa, and Amrapali Zaveri. "Methodology for Assessment of Linked Data Quality." *LDQ@ SEMANTiCS*. 2014.

Identification of quality issues

We defined issue types and issues

ID	Name	Description	Issue detection
1	No link from an authority name to an authority record	A data source only provides the name of an authority.	data source pre-processing
2	Several links from an authority name to an authority record	A data source only provides the name of an authority, automatic linking identified several candidate records.	data source pre-processing
3	Conflicting dates	Different local records of an integrated record have a conflicting date	after data integration
4	**Duplicate identifiers**	**An integrated record (linking with sameAs to several local records) contains more than one identifier of a kind**	**after data integration**
5	Several bibliographic library identifiers	An integrated bibliograhpic record (linking wiht sameAs to several local records) refers to more than one library identifier of a kind	after data integration

Identification of quality issues

We defined issue types and issues

ID	Name	Description	Issue detection
1			
2			
3			
4	**Duplicate identifiers**	**An integrated record (linking with sameAs to several local records) contains more than one identifier of a kind**	**after data integration**
5			

Identification of quality issues

ID	Issue type	Name	Description	Detection
int1	3	conflicting birth date	Conflicting birth dates	int1-conflicting-birth-dates.py
int2	3	conflicting death date	Conflicting death dates	int2-conflicting-death-dates.py
int3	3	conflicting publication years	Conflicting publication years	int3-conflicting-publication-years.py
int4	**4**	**duplicate ISNI identifier**	**More than one ISNI identifier**	**int4-duplicate-isni.sparql**
int5	**4**	**duplicate VIAF identifier**	**More than one VIAF identifier**	**int5-duplicate-viaf.sparql**
int6	**4**	**duplicate Wikidata identifier**	**More than one Wikidata identifier**	**int6-duplicate-wikidata.sparql**
int7	**4**	**duplicate KBR authority identifier**	**More than one KBR identifier**	**int7-duplicate-kbr.sparql**
int8	**4**	**duplicate BnF authority identifier**	**More than one BnF identifier**	**int8-duplicate-bnf.sparql**
int9	**4**	**duplicate NTA authority identifier**	**More than one NTA identifier**	**int9-duplicate-nta.sparql**
int10	5	duplicate KBR bibliographic identifier	More than one KBR identifier	int10-duplicate-kbr-bib.sparql
int11	5	duplicate BnF bibliographic identifier	More than one BnF identifier	int11-duplicate-bnf-bib.sparql
int12	5	duplicate KB bibliographic identifier	More than one KB identifier	int12-duplicate-kb-bib.sparql

Identification of quality issues

ID	Issue type	Name	Description	Detection
int1	3			
int2	3			
int3	3			
int4	**4**	**duplicate ISNI identifier**	**More than one ISNI identifier**	**int4-duplicate-isni.sparql**
int5	**4**	**duplicate VIAF identifier**	**More than one VIAF identifier**	**int5-duplicate-viaf.sparql**
int6	**4**	**duplicate Wikidata identifier**	**More than one Wikidata identifier**	**int6-duplicate-wikidata.sparql**
int7	**4**	**duplicate KBR authority identifier**	**More than one KBR identifier**	**int7-duplicate-kbr.sparql**
int8	**4**	**duplicate BnF authority identifier**	**More than one BnF identifier**	**int8-duplicate-bnf.sparql**
int9	**4**			
int10	5			
int11	5			
int12	5			

Example of CSV presented to the Librarian

ID	name	Duplicate identifier	ISNI	KBR	BnF	NTA
123	**Balis, Arnout**	0000000116876507; 000000011960753X	Balis, Arnout, ISNI 0000000116876507, VIAF 22469787 BnF 12059871 Balis, Arnout, ISNI 0000000116876507, VIAF 305671655 12059871 BnF 12059871 Balis, Arnout, ISNI 0000000116876507, VIAF 54169473 BnF 12059871	14290507 Balis, Arnout, ISNI 0000000116876507	cb15754192t **Joost Vander Auwera**, ISNI 000000011960753X, VIAF 22469787	

Identification of quality issues

A common identifier led to a wrong linking of records

ID	name	Duplicate identifier	ISNI	KBR	BnF	NTA
123	**Balis, Arnout**	0000000116876507; 000000011960753X	Balis, Arnout, ISNI 0000000116876507, VIAF **22469787** BnF 12059871 Balis, Arnout, ISNI 0000000116876507, VIAF 305671655 12059871 BnF 12059871 Balis, Arnout, ISNI 0000000116876507, VIAF 54169473 BnF 12059871	14290507 Balis, Arnout, ISNI 0000000116876507	cb15754192t **Joost Vander Auwera**, ISNI 000000011960753X, VIAF **22469787**	

Data Quality Assessment

Phase 1: Requirements Analysis

Use Case Analysis

Phase 2: Quality Assessment

Identification of quality issues

Analysis

Phase 3: Quality Improvement

Root cause analysis

Fixing quality problems

Rula, Anisa, and Amrapali Zaveri. "Methodology for Assessment of Linked Data Quality." *LDQ@ SEMANTiCS*. 2014.

CSV files as a log of correct and incorrect identifiers

wrong.csv

Source	sourceID	wrongID	ID type
ISNI dump	0000000116876507	22469787	VIAF

correct.csv

Source	sourceID	correctID	ID type

CSV files as a log of correct and incorrect identifiers

wrong.csv

Source	sourceID	wrongID	ID type
ISNI dump	0000000116876507	22469787	VIAF

correct.csv

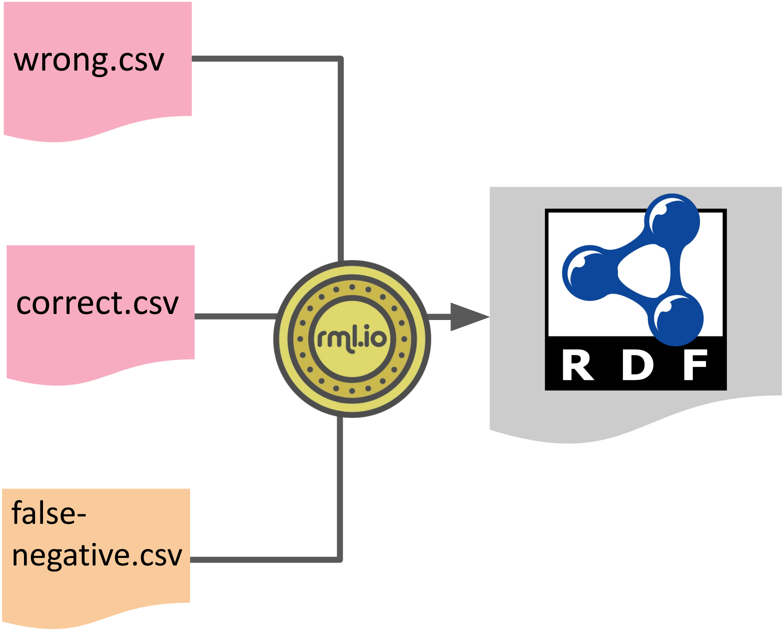
Source	sourceID	correctID	ID type

Indicate cases where a **duplicate identifier was checked by a human and judged to be valid**.
Thus this false-negative will *not be counted* as error during the *next* error-identification SPARQL query

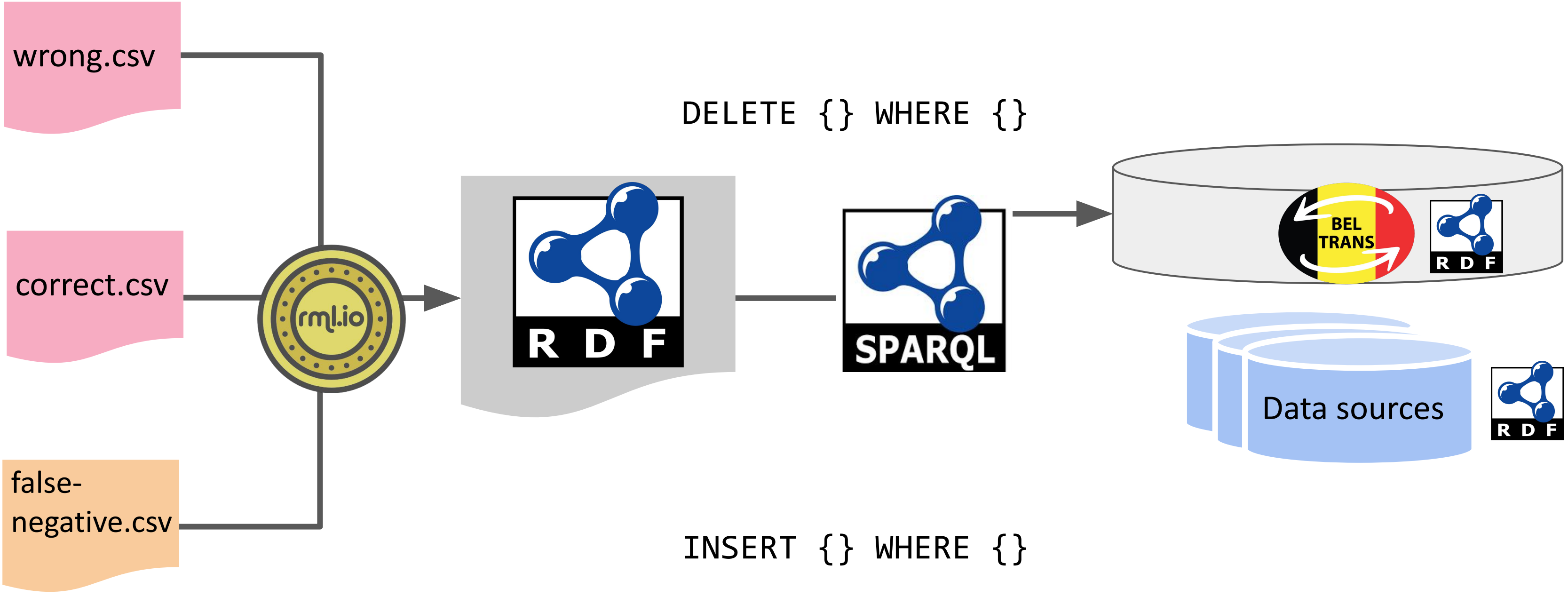
false-negative.csv

Source graph	correctIDCombinationString	ID type

Transform quality log to RDF



SPARQL UPDATES queries before the integration step (thus avoiding that a wrong link is created)



Fix the issues at the data source

In theory, the logs can be used to fix data at the source (automatically or via notification)

However, in the shown example we **did not investigate the root cause**

Preliminary quality issue results - conflicting dates

Translation corpus of **13,005 translations** with **6,371 person contributors**

Conflicting dates

- conflicting birth dates (issue id, int1): 58
- conflicting death dates (issue id, int2): 18
- conflicting publication dates issue id, int3): 187

Preliminary quality issue results - duplicate identifiers

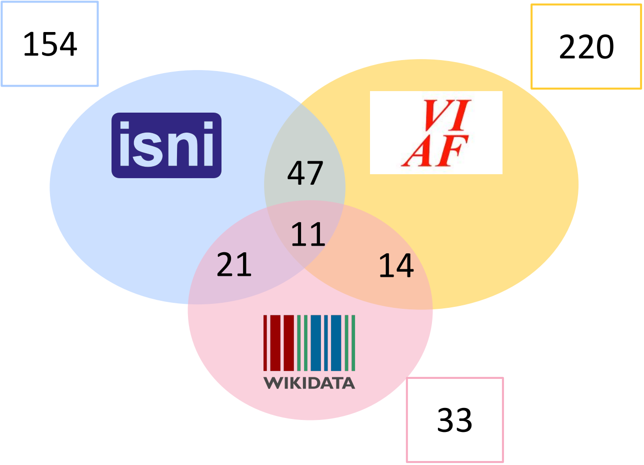
Translation corpus of **13,005 translations** with **6,371 person contributors**

Conflicting dates

conflicting birth dates (issue id, int1):	58
conflicting death dates (issue id, int2):	18
conflicting publication dates issue id, int3):	187

Duplicate identifiers detected via SPARQL queries (union is 337 persons)

154 persons with duplicate ISNI identifier
220 persons with duplicate VIAF identifier
33 persons with duplicate Wikidata identifier



Preliminary quality issue results - dependency between issues

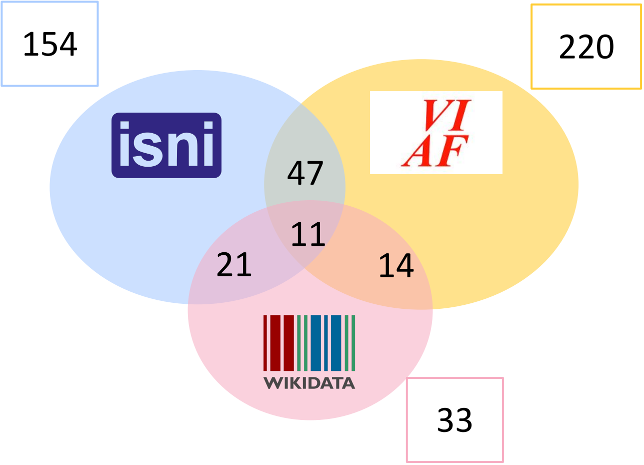
Translation corpus of **13,005 translations** with **6,371 person contributors**

Conflicting dates

conflicting birth dates (issue id, int1): 58
conflicting death dates (issue id, int2): 18
conflicting publication dates issue id, int3): 187

Duplicate identifiers detected via SPARQL queries (union is 337 persons)

154 persons with duplicate ISNI identifier
220 persons with duplicate VIAF identifier
33 persons with duplicate Wikidata identifier



Conflicting dates possible due to duplicate identifiers

Preliminary quality issue results - not all is “wrong”

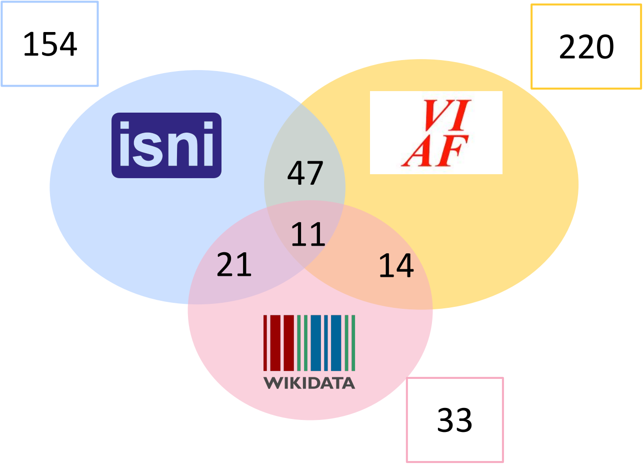
Translation corpus of **13,005 translations** with **6,371 person contributors**

Conflicting dates

conflicting birth dates (issue id, int1):	58
conflicting death dates (issue id, int2):	18
conflicting publication dates issue id, int3):	187

Duplicate identifiers detected via SPARQL queries (union is 337 persons)

- 154 persons with duplicate ISNI identifier
- 220 persons with duplicate VIAF identifier
- 33 persons with duplicate Wikidata identifier



Conflicting dates possible due to duplicate identifiers

Duplicate identifiers partially because of pseudonyms

What is data quality, what is correct and what is valid?

Current quality procedures at KBR

The use case: BELTRANS project

Our **L**ibrarian-**I**n-**T**he-**L**oop workflow

Lessons learned and future work

The pipeline relies on identifiers

👎 Without identifiers in common more sophisticated entity disambiguation techniques are needed

The pipeline relies on identifiers

👎 Without identifiers in common more sophisticated entity disambiguation techniques are needed

👎 Certain identifiers seem to lead to wrong links or contribute wrong data

The pipeline relies on identifiers

- 👎 Without identifiers in common more sophisticated entity disambiguation techniques are needed
- 👎 Certain identifiers seem to lead to wrong links or contribute wrong data
- 👍 The pipeline is configurable: we can measure integration with/without a certain identifier

The pipeline relies on identifiers

- 👎 Without identifiers in common more sophisticated entity disambiguation techniques are needed
- 👎 Certain identifiers seem to lead to wrong links or contribute wrong data
- 👍 The pipeline is configurable: we can measure integration with/without a certain identifier
- 👍 The pipeline does the “heavy lifting” enabling more detailed work afterwards and allowing to spot “wrong information” mistakes in the first place

Fixing the data integration vs fixing underlying issues

👎 Fixing the linking is not fixing the underlying issues

e.g. wrong or duplicate ISNIs which need to be corrected at ISNI

Fixing the data integration vs fixing underlying issues

👎 Fixing the linking is not fixing the underlying issues
e.g. wrong or duplicate ISNIs which need to be corrected at ISNI

👍 In case the data fields in the CSV are not sufficient for the librarian,
its creation can be adjusted (SPARQL or Python)

Fixing the data integration vs fixing underlying issues

👎 Fixing the linking is not fixing the underlying issues
e.g. wrong or duplicate ISNIs which need to be corrected at ISNI

👍 In case the data fields in the CSV are not sufficient for the librarian,
its creation can be adjusted (SPARQL or Python)

ℹ Instead of “fixing” wrong identifiers to support the automatic integration,
we can let the **human decide which records are the same**,
those should be **excluded by the automatic integration**

Fix issues directly at the source & use most recent data

👍 **Iterative approach**: identify quality issues as **automatic executable step**, possibly after every new data export (or preprocessing)

Fix issues directly at the source & use most recent data

👍 **Iterative approach**: identify quality issues as **automatic executable step**, possibly after every new data export (or preprocessing)

ℹ Fixing data sources (under control) immediately seems to be most efficient as for comparison the record might have to be opened anyway

Fix issues directly at the source & use most recent data

👍 **Iterative approach**: identify quality issues as **automatic executable step**, possibly after every new data export (or preprocessing)

ℹ Fixing data sources (under control) immediately seems to be most efficient as for comparison the record might have to be opened anyway

ℹ Most recent data should be used, but sometimes only outdated information available, e.g. BnF data continuously corrected in the BnF catalogue, but the RDF dumps are only created yearly

Future work: adapt integration workflow for **legal deposit** completeness

As the national library of Belgium, KBR collects and preserves:

all publications that are published on **Belgian territory**

all publications by authors **of Belgian nationality and domiciled in Belgium** and whose work is **published abroad**

Future work: adapt integration workflow for **legal deposit** completeness

As the national library of Belgium, KBR collects and preserves:

all publications that are published on **Belgian territory**

all publications by authors **of Belgian nationality and domiciled in Belgium** and whose work is **published abroad**

1) Select main sources for “Belgian” publications, in Belgium and abroad

Future work: adapt integration workflow for **legal deposit** completeness

As the national library of Belgium, KBR collects and preserves:

- all publications that are published on **Belgian territory**
- all publications by authors **of Belgian nationality and domiciled in Belgium** and whose work is **published abroad**

1) Select main sources for “Belgian” publications, in Belgium and abroad

2) Map data to RDF and load it into a RDF database

Future work: adapt integration workflow for **legal deposit** completeness

As the national library of Belgium, KBR collects and preserves:

all publications that are published on **Belgian territory**

all publications by authors **of Belgian nationality and domiciled in Belgium** and whose work is **published abroad**

1) Select main sources for "Belgian" publications, in Belgium and abroad

2) Map data to RDF and load it into a RDF database

3) Integrate data and filter on Belgian nationality

Future work: adapt integration workflow for **legal deposit** completeness

As the national library of Belgium, KBR collects and preserves:

all publications that are published on **Belgian territory**

all publications by authors **of Belgian nationality and domiciled in Belgium** and whose work is **published abroad**

1) Select main sources for “Belgian” publications, in Belgium and abroad

2) Map data to RDF and load it into a RDF database

3) Integrate data and filter on Belgian nationality

4) Check via ISBN10 and ISBN13 which publications are present at KBR and acquire the ones which are missing

Use your data to achieve something!
You will likely encounter data quality issues…

Sven Lieber (Sven.Lieber@kbr.be, sven-lieber.org)
Data manager, research and innovation department

Ann Van Camp (Ann.VanCamp@kbr.be)
Collection development, contemporary collections department

Hannes Lowagie (Hannes.Lowagie@kbr.be)
Head of agency for bibliographic information

KBR Koester de tijd
Protégeons le temps

GitHub kbrbe/beltrans-data-integration