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
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
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 Librarian-In-The-Loop workflow

Lessons learned and future work
Data quality is use case specific
Data quality is use case specific - similar quality when viewing use case “Which publisher?”

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IDN</td>
</tr>
<tr>
<td>Wettelijk Depotnummer</td>
</tr>
<tr>
<td>Auteur</td>
</tr>
<tr>
<td><strong>Uitgever</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IDN</td>
</tr>
<tr>
<td><strong>Uitgever</strong></td>
</tr>
</tbody>
</table>
Data quality is use case specific - similar quality when viewing use case “Where is the publisher from?”

<table>
<thead>
<tr>
<th>Data Quality</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDN</td>
<td>15749415</td>
</tr>
<tr>
<td>Wettelijk Depotnummer</td>
<td>D/2007/5658/0018</td>
</tr>
<tr>
<td>Auteur</td>
<td>Pellerin, Gilles. Auteur Hertmans, Stefan (1951-...). Auteur De Vuyst, Katelijne (1958-). Vertaler</td>
</tr>
<tr>
<td>Uitgever</td>
<td>Uitgeverij P</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Data Quality</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDN</td>
<td>15151941</td>
</tr>
<tr>
<td>Uitgever</td>
<td>Uitgeverij Emmaus</td>
</tr>
</tbody>
</table>

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

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

300(2#) $a147 p. $c20 cm

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

700(1#) $a*14647172 $bSullivan, Jean $4aut

700(1#) $a*14122319 $bHerkes, Ed. $d19-... $4trl
Data quality is use case specific - similar quality in MARC

use case “Where is the publisher from?”
Data quality is use case specific - quality difference


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
Incorrect data which can be detected based on syntax rules

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDN</td>
<td>16623686</td>
</tr>
<tr>
<td>gebonden uitgave</td>
<td></td>
</tr>
<tr>
<td>Auteur</td>
<td>Hertmans, Stefan (1951-...)</td>
</tr>
<tr>
<td>Uitgever</td>
<td>De Bezige Bij</td>
</tr>
<tr>
<td>Rubriek BB</td>
<td>830 Roman.</td>
</tr>
<tr>
<td>Plaatskenmerk</td>
<td>BB A 2013 9.338</td>
</tr>
<tr>
<td>IDN Vubis</td>
<td>2063952</td>
</tr>
<tr>
<td>OPAC</td>
<td><a href="https://opac.kbr.be/Library/doc/SYRACUSE/16623686">https://opac.kbr.be/Library/doc/SYRACUSE/16623686</a></td>
</tr>
</tbody>
</table>
Incorrect data which can be detected based on syntax rules

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IDN</td>
</tr>
<tr>
<td>ISBN en vorm van uitgave</td>
</tr>
<tr>
<td>gebonden uitgave</td>
</tr>
<tr>
<td>Auteur</td>
</tr>
<tr>
<td>Uitgever</td>
</tr>
<tr>
<td>Rubriek BB</td>
</tr>
<tr>
<td>Plaatskenmerk</td>
</tr>
<tr>
<td>IDN Vubis</td>
</tr>
</tbody>
</table>

**ISBN is valid**
Incorrect data which can be detected based on syntax rules

ISBN is valid: 978-90-234-7671-9

ISBN is invalid: wrong check digit
Incorrect data which needs to be checked manually


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

<table>
<thead>
<tr>
<th>IDN</th>
<th>16623686</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISBN en vorm van uitgave</td>
<td><strong>978-94-6355-578-4</strong></td>
</tr>
<tr>
<td></td>
<td>gebonden uitgave</td>
</tr>
<tr>
<td>Auteur</td>
<td>Hertmans, Stefan (1951-...)</td>
</tr>
<tr>
<td>Uitgever</td>
<td>De Bezige Bij</td>
</tr>
<tr>
<td>Rubriek BB</td>
<td>830 Roman.</td>
</tr>
<tr>
<td>Plaatskenmerk</td>
<td>BB A 2013 9.338</td>
</tr>
<tr>
<td>IDN Vubis</td>
<td>2063952</td>
</tr>
</tbody>
</table>

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 Librarian-In-The-Loop 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
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 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 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 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

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

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 Librarian-In-The-Loop workflow

Lessons learned and future work
**BELTRANS project: studying book translations flows**

<table>
<thead>
<tr>
<th>intra-Belgian</th>
<th>🇧🇪 author / illustrator / scenarist / publishing director</th>
</tr>
</thead>
<tbody>
<tr>
<td>location</td>
<td>🌍 published anywhere in the world</td>
</tr>
<tr>
<td>literary</td>
<td>📚 literary genres (novel, youth literature, comics, poetry) + literary non-fiction (mainly history books)</td>
</tr>
<tr>
<td>translations</td>
<td>🗣️ FR-NL / NL-FR</td>
</tr>
<tr>
<td>time-period</td>
<td>📅 1970-2020, since the creation of the Communities</td>
</tr>
</tbody>
</table>
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
We developed a pipeline with *different* processing per data source to create integrated data.
Integrate data by interlinking it with standard identifiers.
Integrate data by interlinking it with standard identifiers.

- **ISBN**: schema:sameAs
- **ISNI**: schema:sameAs
- **KBR**: Coester de tijd, Protégeons le temps

![Diagram showing interlinking of data sources](image-url)
Python script performing SPARQL INSERT per data source and subsequent SPARQL UPDATE queries from other sources.
SPARQL INSERT to create a record + sameAs link

ISNI dump/api

KB catalog

KBR catalog

{BnF}

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

ISNI dump/api
KBR catalog
BnF catalog
KB catalog
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.
SELECT translations
WHERE contributor nationality is Belgian
Postprocess the data and create a lightweight CSV file

<table>
<thead>
<tr>
<th>ID</th>
<th>title</th>
<th>year Of Publication KBR</th>
<th>Year Of Publication BnF</th>
<th>Year Of Publication KB</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Book1</td>
<td>2020</td>
<td>2020</td>
<td>2019</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Book2</td>
<td>2019</td>
<td>2019</td>
<td>2019</td>
<td></td>
</tr>
</tbody>
</table>
Postprocess the data and create a lightweight CSV file

<table>
<thead>
<tr>
<th>ID</th>
<th>title</th>
<th>year Of Publication KBR</th>
<th>Year Of Publication BnF</th>
<th>Year Of Publication KB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Book1</td>
<td>2020</td>
<td>2020</td>
<td>2019</td>
</tr>
<tr>
<td>2</td>
<td>Book2</td>
<td>2019</td>
<td></td>
<td>2019</td>
</tr>
</tbody>
</table>

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

Data sources

SPARQL

Inconsistent data

Corrections

BEL TRANS

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

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

We defined issue types and issues

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Description</th>
<th>Issue detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Duplicate identifiers</td>
<td>An integrated record (linking with sameAs to several local records) contains more than one identifier of a kind</td>
<td>after data integration</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td>Issue type</td>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
<td>-------------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>int1</td>
<td>3</td>
<td>conflicting birth date</td>
<td>Conflicting birth dates</td>
</tr>
<tr>
<td>int2</td>
<td>3</td>
<td>conflicting death date</td>
<td>Conflicting death dates</td>
</tr>
<tr>
<td>int3</td>
<td>3</td>
<td>conflicting publication years</td>
<td>Conflicting publication years</td>
</tr>
<tr>
<td>int4</td>
<td>4</td>
<td>duplicate ISNI identifier</td>
<td>More than one ISNI identifier</td>
</tr>
<tr>
<td>int5</td>
<td>4</td>
<td>duplicate VIAF identifier</td>
<td>More than one VIAF identifier</td>
</tr>
<tr>
<td>int6</td>
<td>4</td>
<td>duplicate Wikidata identifier</td>
<td>More than one Wikidata identifier</td>
</tr>
<tr>
<td>int7</td>
<td>4</td>
<td>duplicate KBR authority identifier</td>
<td>More than one KBR identifier</td>
</tr>
<tr>
<td>int8</td>
<td>4</td>
<td>duplicate BnF authority identifier</td>
<td>More than one BnF identifier</td>
</tr>
<tr>
<td>int9</td>
<td>4</td>
<td>duplicate NTA authority identifier</td>
<td>More than one NTA identifier</td>
</tr>
<tr>
<td>int10</td>
<td>5</td>
<td>duplicate KBR bibliographic identifier</td>
<td>More than one KBR identifier</td>
</tr>
<tr>
<td>int11</td>
<td>5</td>
<td>duplicate BnF bibliographic identifier</td>
<td>More than one BnF identifier</td>
</tr>
<tr>
<td>int12</td>
<td>5</td>
<td>duplicate KB bibliographic identifier</td>
<td>More than one KB identifier</td>
</tr>
</tbody>
</table>
## Identification of quality issues

<table>
<thead>
<tr>
<th>ID</th>
<th>Issue type</th>
<th>Name</th>
<th>Description</th>
<th>Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>int1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>int2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>int3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>int4</td>
<td>4</td>
<td>duplicate ISNI identifier</td>
<td>More than one ISNI identifier</td>
<td>int4-duplicate-isni.sparql</td>
</tr>
<tr>
<td>int5</td>
<td>4</td>
<td>duplicate VIAF identifier</td>
<td>More than one VIAF identifier</td>
<td>int5-duplicate-viaf.sparql</td>
</tr>
<tr>
<td>int6</td>
<td>4</td>
<td>duplicate Wikidata identifier</td>
<td>More than one Wikidata identifier</td>
<td>int6-duplicate-wikidata.sparql</td>
</tr>
<tr>
<td>int7</td>
<td>4</td>
<td>duplicate KBR authority identifier</td>
<td>More than one KBR identifier</td>
<td>int7-duplicate-kbr.sparql</td>
</tr>
<tr>
<td>int8</td>
<td>4</td>
<td>duplicate BnF authority identifier</td>
<td>More than one BnF identifier</td>
<td>int8-duplicate-bnf.sparql</td>
</tr>
<tr>
<td>int9</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>int10</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>int11</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>int12</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Example of CSV presented to the Librarian

<table>
<thead>
<tr>
<th>ID</th>
<th>name</th>
<th>Duplicate identifier</th>
<th>ISNI</th>
<th>KBR</th>
<th>BnF</th>
<th>NTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>Balis, Arnout</td>
<td>00000000116876507; 000000011960753X</td>
<td>Balis, Arnout, ISNI 0000000116876507, VIAF 22469787 BnF 12059871</td>
<td>14290507 Balis, Arnout, ISNI 000000116876507</td>
<td>cb15754192t Joost Vander Auwera, ISNI 00000011960753X, VIAF 22469787</td>
<td>12059871</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Balis, Arnout, ISNI 0000000116876507, VIAF 305671655 12059871 12059871</td>
<td></td>
<td></td>
<td>54169473, BnF 12059871</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Balis, Arnout, ISNI 0000000116876507, VIAF 54169473 12059871</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Identifying quality issues

A common identifier led to a wrong linking of records

<table>
<thead>
<tr>
<th>ID</th>
<th>name</th>
<th>Duplicate identifier</th>
<th>ISNI</th>
<th>KBR</th>
<th>BnF</th>
<th>NTA</th>
</tr>
</thead>
</table>
| 123| Balis, Arnout | 00000000116876507; 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 | 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

CSV files as a log of correct and incorrect identifiers

<table>
<thead>
<tr>
<th>Source</th>
<th>sourceID</th>
<th>wrongID</th>
<th>ID type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISNI dump</td>
<td>0000000116876507</td>
<td>22469787</td>
<td>VIAF</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>sourceID</th>
<th>correctID</th>
<th>ID type</th>
</tr>
</thead>
</table>
CSV files as a log of correct and incorrect identifiers

### wrong.csv

<table>
<thead>
<tr>
<th>Source</th>
<th>sourceID</th>
<th>wrongID</th>
<th>ID type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISNI dump</td>
<td>0000000116876507</td>
<td>22469787</td>
<td>VIAF</td>
</tr>
</tbody>
</table>

### correct.csv

<table>
<thead>
<tr>
<th>Source</th>
<th>sourceID</th>
<th>correctID</th>
<th>ID type</th>
</tr>
</thead>
</table>

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.
Transform quality log to RDF

wrong.csv

correct.csv

false-negative.csv
SPARQL UPDATES queries before the integration step (thus avoiding that a wrong link is created)

wrong.csv

correct.csv

false-negative.csv

DELETE {} WHERE {}

INSERT {} WHERE {}

Data sources

KBR  Koester de tijd  Protégeons le temps
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

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 Librarian-In-The-Loop 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.

KBR | Koester de tijd
Protégeons le temps
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)

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

KBR Koester de tijd Protégeons le temps
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)

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).

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

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.

KBR

Koester de tijd
Protégeons le temps
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