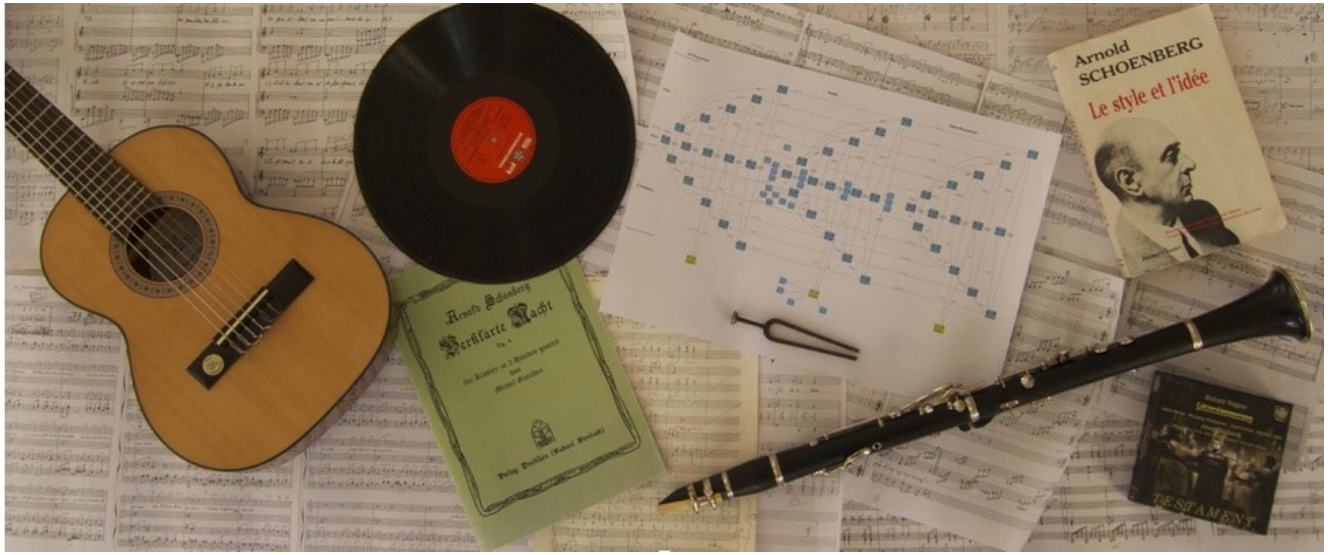
DOing
REusable
MUSical data

http://www.doremus.org



Tutorial ESWC 2016

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR
ANR

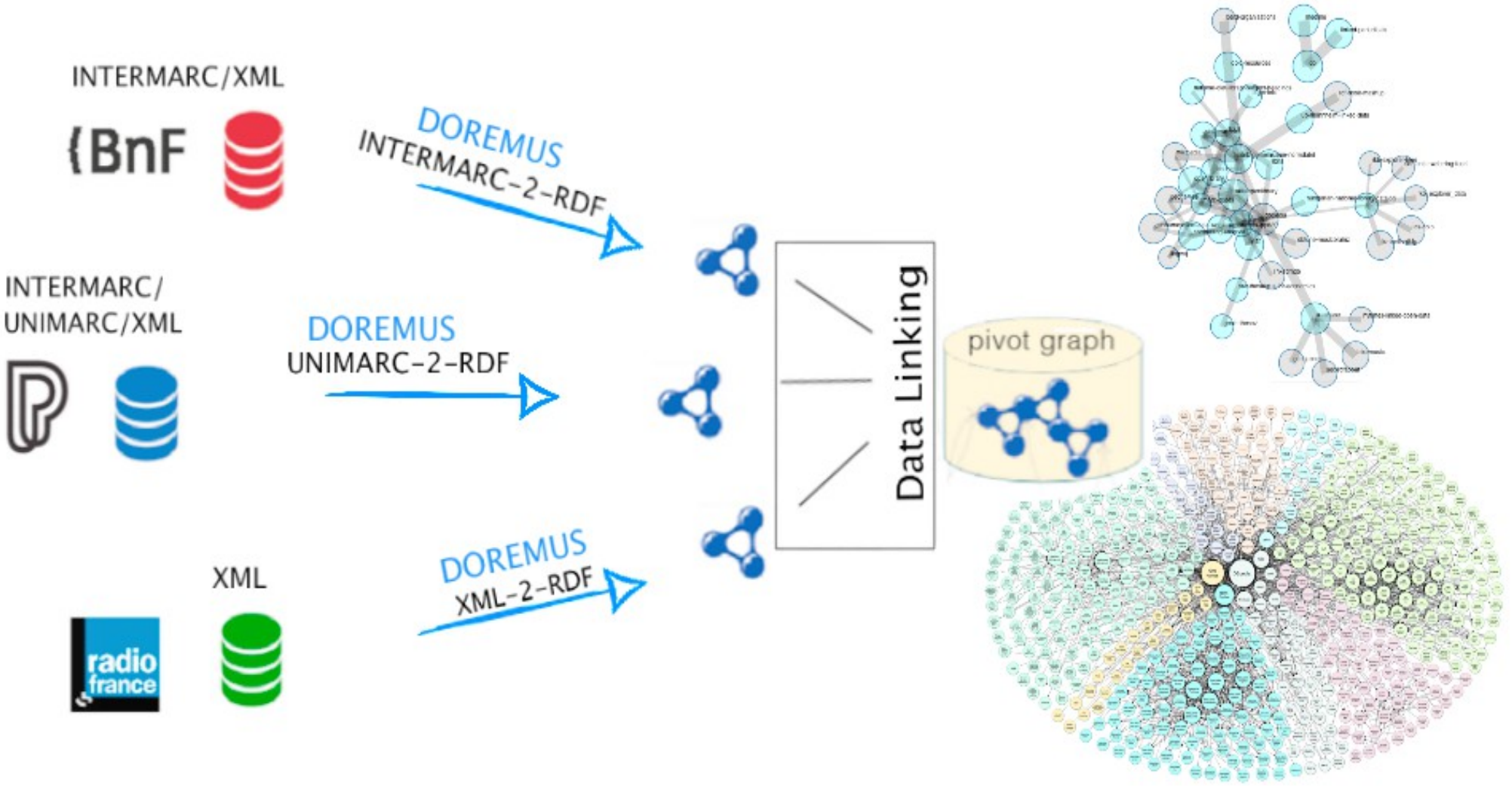
<ANR-14-CE24-0020>

Data Lifting and Linking

Konstantin Todorov,
Manel Achichi, Raphaël Troncy

Outline

1. Input Data
2. Conversion to Doremus RDF
3. Data Linking
4. Connect to the Web of Data



Outline

1. Input Data

INTERMARC/XML

UNIMARC/XML

XML

1. Input Data

	BnF	**PP** Médiathèque	**PP** Concerts	**Radio France** Disco-thèque	**Radio France** Docu-mentation musicale	**Radio France** Docu-mentation sonore	**Target entity** ↓
Format	XML/ INTER MARC	XML/ UNIMARC	XML	XML	XML	XML	
Uniform Music Titles (TUM) & work entries	135 940	6 846			62 550		**Work**
Scores	89 184	30 319			9 154		**Expression**
Books		21 035					
CD/DVD/ Vinyls		8 602		340 609			**Performance**
Concerts		2 447	2 717		7 700	1 800	

1. Input Data

Introducing the MARC family

MARC:

Machine **R**eadable **C**ataloging

a bibliographical data exchange format

```
001 FRBNF139081882
008 890821130211yy    sn        1801
048   $aka01
100   $313891295$w.0..b.....$aBeethoven$mLudwig van$d1770-1827
144 1 $w....b.fre.$aSonates$bPiano$pOp. 27, no 2$tDo dièse mineur
444 1 $w....b.fre.$aSonates$bPiano$nNo 14$tDo dièse mineur
```

A **MARC** file is

- a succession of fields of different lengths, each carrying a label (a 3 digit number)

- each field is a succession of sub-fields (also of variable lengths)

- a sub-field is delimited by the “$” symbol

- sub-fields can repeat in order to “host” data of the same kind

Different variants of MARC...

- USMARC in the United States, CANMARC in Canada, UKMARC in the UK,...

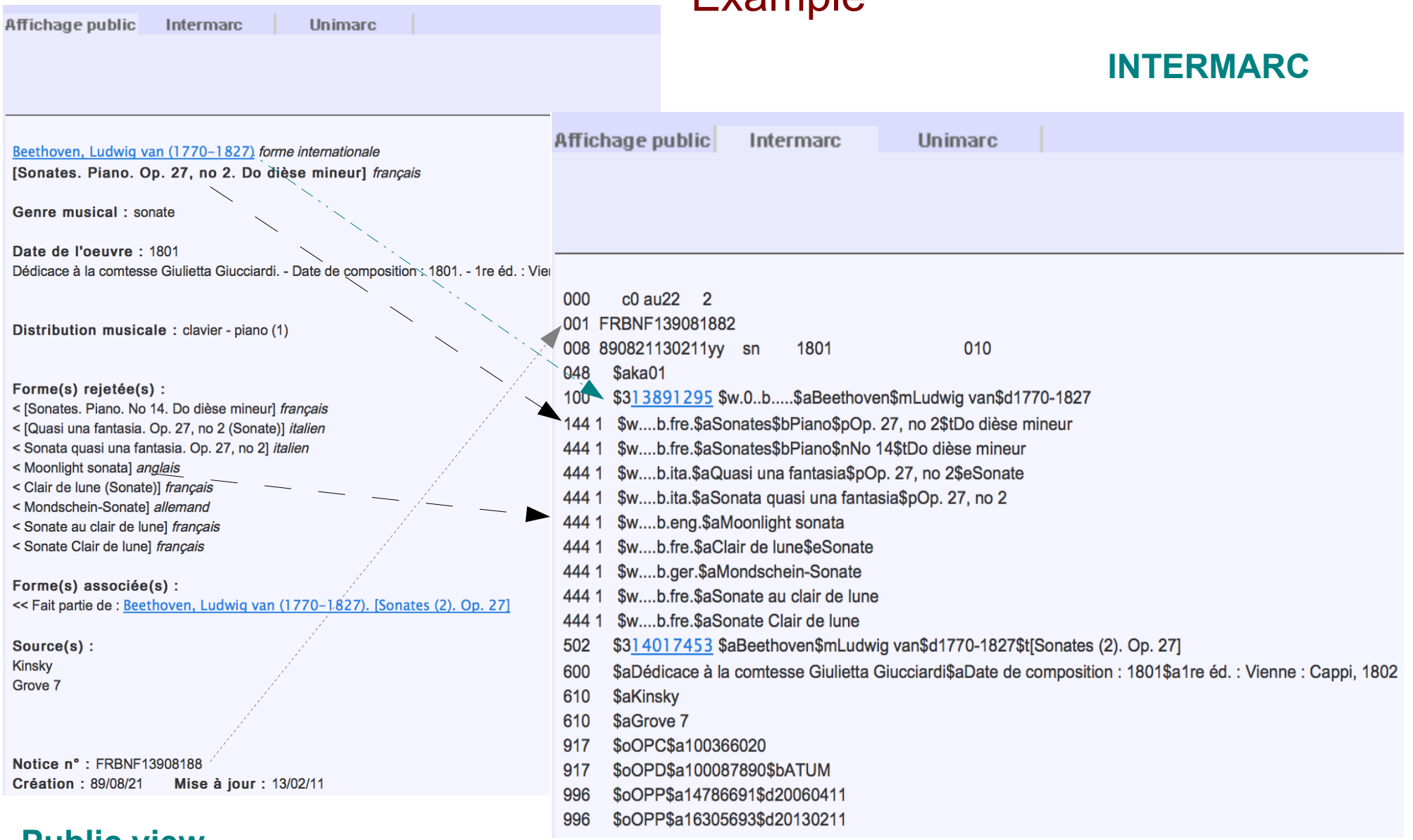
- MARC21 unifies USMARC, AUSMARC, UKMARC, CANMARC.

- INTERMARC is used by the BNF and other libraries in Paris and Lyon in France.

- UNIMARC was initially designed as a unique format for exchange between the different MARCs, it became the official french MARC format.

1. Input Data

Example



INTERMARC

Public view

1. Input Data

Different kinds of records within and across institutions

DO RE MUS
DOing REusable MUSical data

This regards music works!

title

composer

variant titles

BNF -TUM

```
001 FRBNF139081882
008 890821130211yy    sn        1801                              010
048   $aka01
100   $313891295$w.0..b.....$aBeethoven$mLudwig van$d1770-1827
144 1 $w....b.fre.$aSonates$bPiano$pOp. 27, no 2$tDo dièse mineur
444 1 $w....b.fre.$aSonates$bPiano$nNo 14$tDo dièse mineur
444 1 $w....b.ita.$aQuasi una fantasia$pOp. 27, no 2$eSonate
444 1 $w....b.ita.$aSonata quasi una fantasia$pOp. 27, no 2
444 1 $w....b.eng.$aMoonlight sonata
444 1 $w....b.fre.$aClair de lune$eSonate
444 1 $w....b.ger.$aMondschein-Sonate
444 1 $w....b.fre.$aSonate au clair de lune
444 1 $w....b.fre.$aSonate Clair de lune
502   $314017453$aBeethoven$mLudwig van$d1770-1827$t[Sonates (2). Op. 27]
600   $aDédicace à la comtesse Giulietta Giucciardi$aDate de composition : 1801$alre éd. : Vienne : Cappi, 1802
610   $aKins
```

PP - Work Record

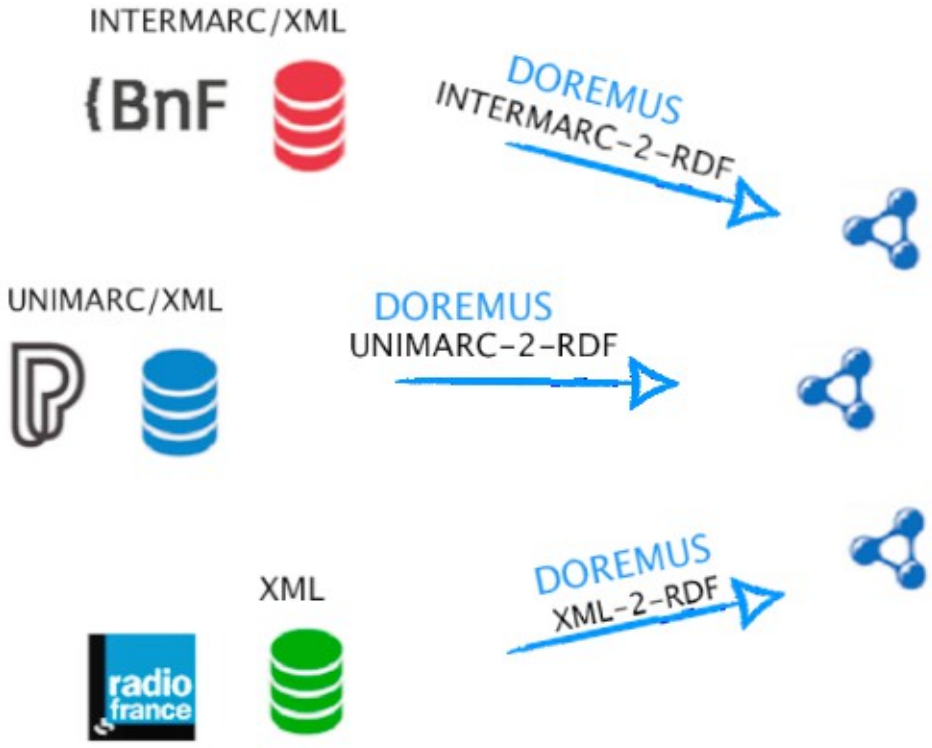
```
019 $aUNI100
100 $a20041214d||||    uuuy0frey0103    ba
200 $aSonate pour piano no 14 "Clair de lune"$fLudwig van Beethoven
600 $308042319$aSonates$rPiano$sOp. 27 no 2$uDo dièse mineur$nNo 14
610 $30068838$aSonate$b04
610 $30144424$a19 éme siècle$b02
610 $30067958$aMusique romantique$b04
610 $30144079$aPiano$b01
700 $30038954$aBeethoven$bLudwig van$f1770-1827$4230
830 $aBN OPALE-PLUS 2007/02/26. Guide de la musique de piano et de clavecin ( Fayard 1987)
909 $aDédicace à la comtesse Giulietta Giucciardi. Parue sous le nom de "Sonate pour piano quasi una fantasia en ut dièse m
919 $aPremière publication : Vienne, Cappi, 1802
937 $30069690$aLes 32 [Trente-deux$bUNI1
937 $30072431$aBeethoven piano sonatas / Denis Matthews, 56 p.$bUNI1
937 $30078731$aLes Sonates de Beethoven / Paul Badura-Skoda, Jörg Demus ; trad. de l'all. par Jean Malignon, 239 p.$bUNI1
937 $30086700$aWilhelm Kempff : Schumann : Arabeske, Papillons, Davidsbündlertänze ; Beethoven : Piano sonatas n° 14 "Moonl
937 $30094153$aKlaviersonaten : Band I, II / Beethoven ; nach Eigenschriften, Abschriften und Originalausgaben herausgegebe
937 $30183795$aMondschein Sonata, Op.27 N°2 : 1er mvt. / Ludwig van Beethoven ; édité par Emile Naoumoff, 1 partition (7 p.
937 $30817410$aBarenboim on Beethoven : the complete piano sonatas, live from Berlin / Ludwig van Beethoven ; Daniel Barenb
938 $30075908$aBeethoven : intégrale des sonates pour piano : du jeudi 8 au dimanche 11 mars 2001 / Hélène Pierrakos, 46 p.
938 $30081170$aSonate no. 14 cis-moll op. 27 n° 2 "Mondschein-Sonate" / Ludwig van Beethove
938 $30235758$aLudwig van Beethoven [compilation$bUNI2
938 $30583341$aNocturnes I - Clairs de lune : du dimanche 2 au jeudi 6 mai 2004 / [Auteurs
938 $30789054$aSonate no 14 en ut dièse mineur, op. 27 no 2 "quasi una fantasia" "Clair de
938 $30889727$aBeethoven / Debussy : dimanche 12 octobre 2008 / Denis Herlin, Hélène Pierra
938 $30890637$aIntégrale des sonates pour piano : CD 4 / Ludwig van Beethoven, composition
938 $30897214$aSonate no 14 en ut dièse mineur, op. 27 no 2 "quasi una fantasia" "Clair de
938 $30906429$aThe Complete piano sonatas on period instruments / Beethoven, composition ;
938 $31003004$aAdagio sostenuto, extrait de la "Sonate no 14 en ut dièse mineur, op. 27 no
938 $31040755$aComplete piano sonatas : CD 4 : Sonatas Opp. 26, 27, 28 / Ludwig van Beethov
938 $31042188$aJe découvre la musique classique CD 2 : S'amuser en classique / Antonio Viva
```

PP - TUM

```
017 $aATU243$oOPSYS
019 $aAIC14$b243
144 $aSonates$rPiano$nNo 14$pOp. 27 no 2$tDo dièse mineur$uClair de lune
322 $30038954$aBeethoven$mLudwig van$d1770-1827
444 $aMoon light sonata$rPiano$nNo 14$pOp. 27 no 2
444 $w||||||||||$aSonate Mondschein$rPiano$nNo 14$pOp. 27 no 2$tDo dièse mineur
444 $aSonata quasi una fantasia
444 $aSonate au clair de lune
602 $afr
909 $aFR$bCITE MUSIQUE$c20040721
```

Outline

2. Conversion to Doremus RDF



2. Data Conversion to RDF

Two Converters

- MARC *2* MARC-RDF
  - Direct extraction of the relations from the MARC file
  - Construction of a triples-based graph

- MARC *2* DOREMUS-RDF
  - Mapping rules to retrieve the values from the MARC files
  - Following and implementing the DOREMUS model
  - Aligned to the DOREMUS controlled vocabularies

2. Data Conversion to RDF

MARC 2 MARC-RDF

The semantics of the fields and sub-fields in the MARC files are described in different documents (according to the MARC variant, see the links below).

A subfield tag changes its meaning depending on the field, in which it is found!

MARC 2 MARC-RDF:

A low-level mapping from the fields and subfields semantics to RDF properties and triples.

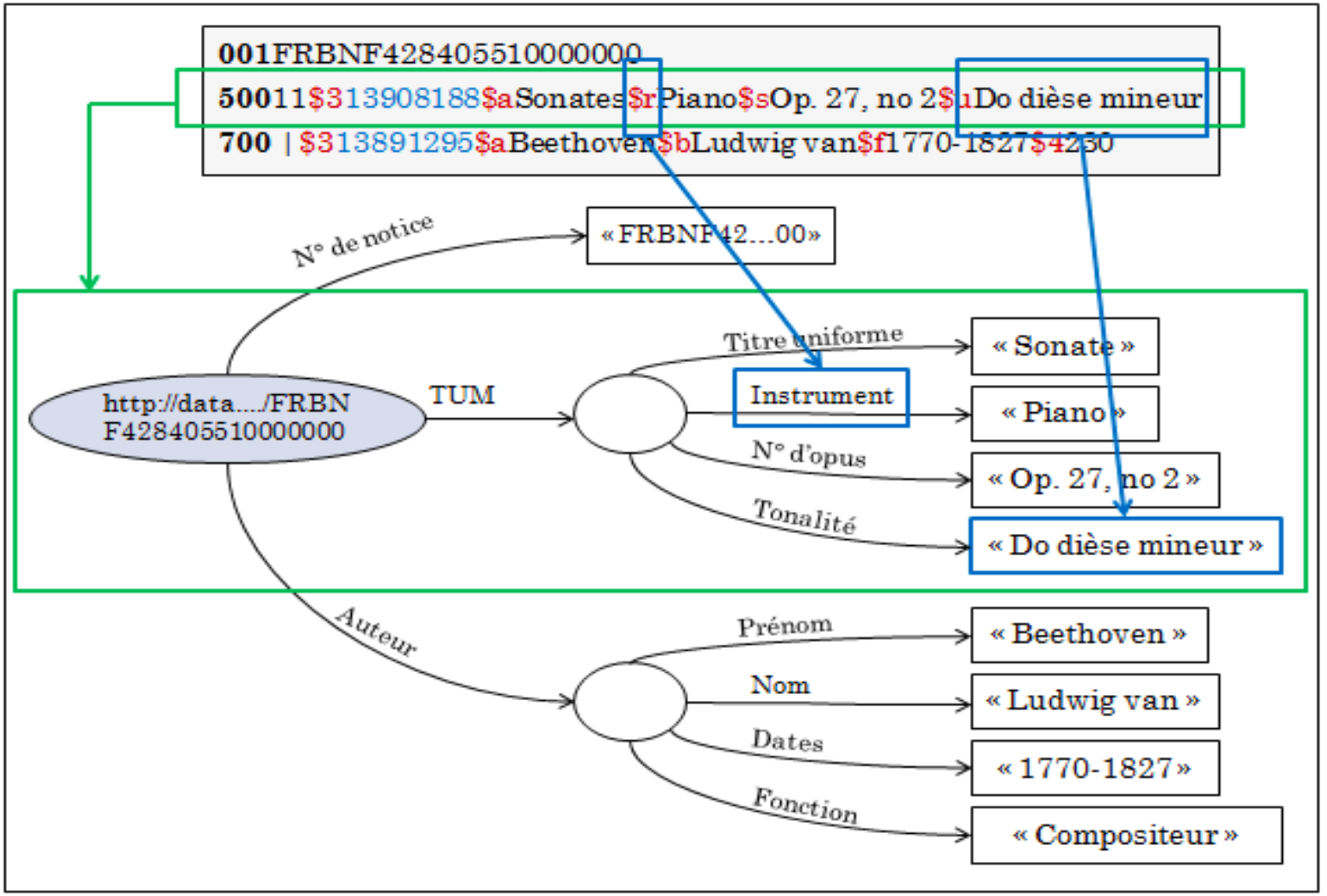
500 Uniform title
$3 ID number of the authority record
$9/a Hierarchical ID of the sub-record
$9/b Match author/title
$a Uniform title
$h Number of the part
$i Title of the part
$k Publication date
$l Rejected form
$m Language
$n Other information
$q Version
$r Distribution of execution (for music)

UNIMARC

UNIMARC (authority records):
http://www.ifla.org/publications/ifla-series-on-bibliographic-control-38
UNIMARC (bibliographical records):
http://www.ifla.org/publications/ifla-series-on-bibliographic-control-36
INTERMARC: http://www.ifla.org/node/4858

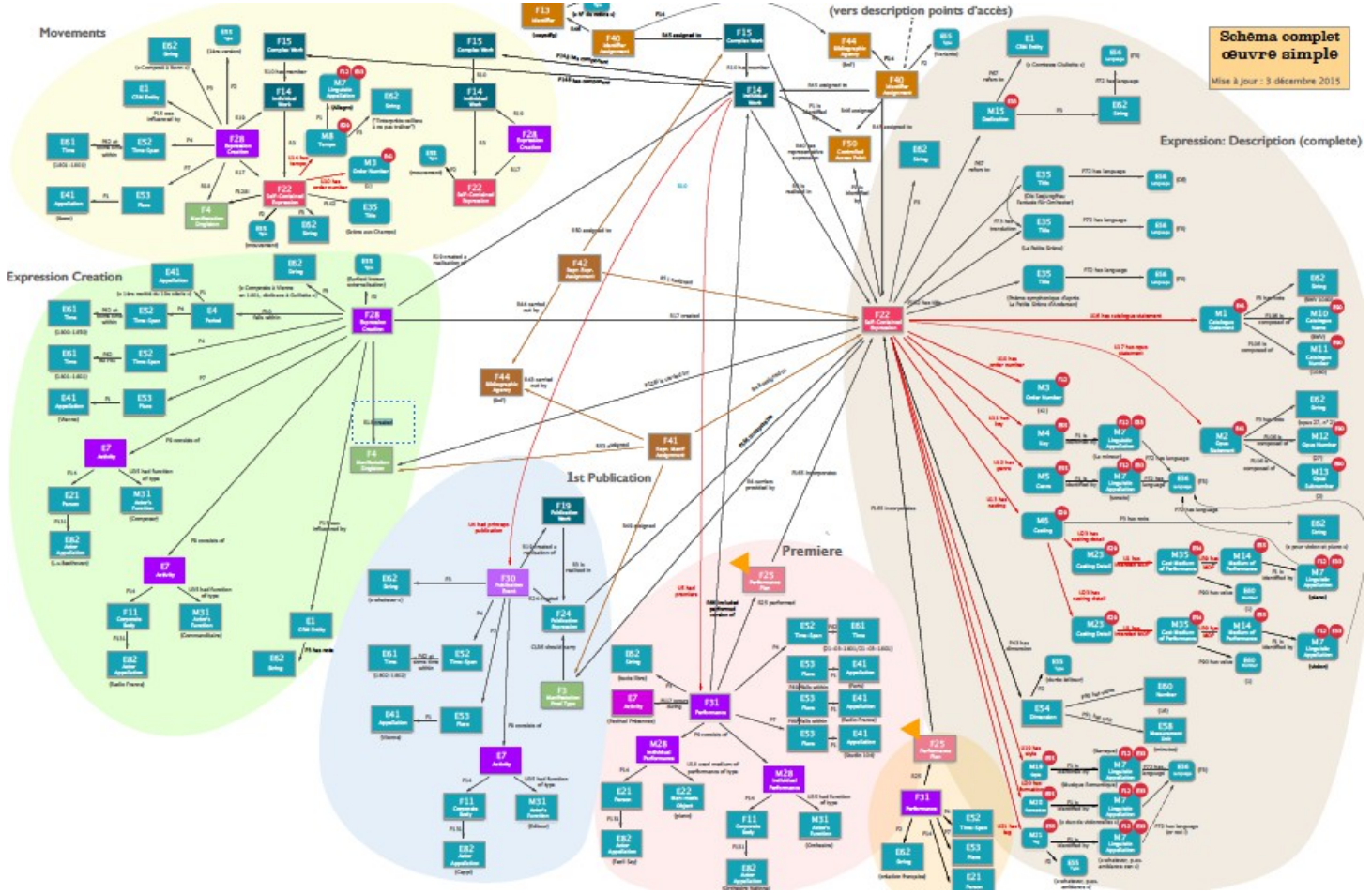
2. Data Conversion to RDF

MARC 2 MARC-RDF



2. Data Conversion to RDF

Remember the Doremus model?...



Let's do DOREMUS RDF!

2. Data Conversion to RDF

Expert-defined mapping rules

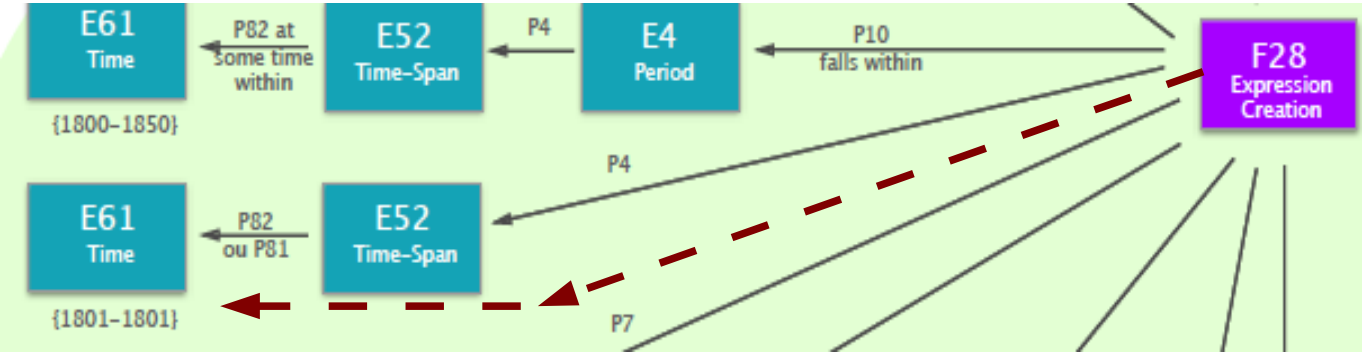
- Where to look for information and how to interpret it
- Implementing the DOREMUS model
- Reflect the practices of each institution: a mapping table *per* institution

Identifier	F28
Unit of information	Work: Date of the work
Object	Date of expression creation
Remarks	Date and machine format
Path	F28 Expression Creation P4 has time-span E52 Time-Span P82 at some time within E61 Time Primitive
Unimarc and Intermarc Philharmonie	UNI100: 909 $g $h
Transfer rules	If $h is identical to $g, keep only $g. Add a slash between $g and $h if they have different values.
Examples	UNI100: 909 $g1801 $h1801 > E52 Time-Span P81 ongoing through E61 = 1801 UNI100:909 $g1834 $h1856 > E52 Time-Span P81 ongoing through E61 = 1834/1856

2. Data Conversion to RDF

Expert-defined mapping rules

An example



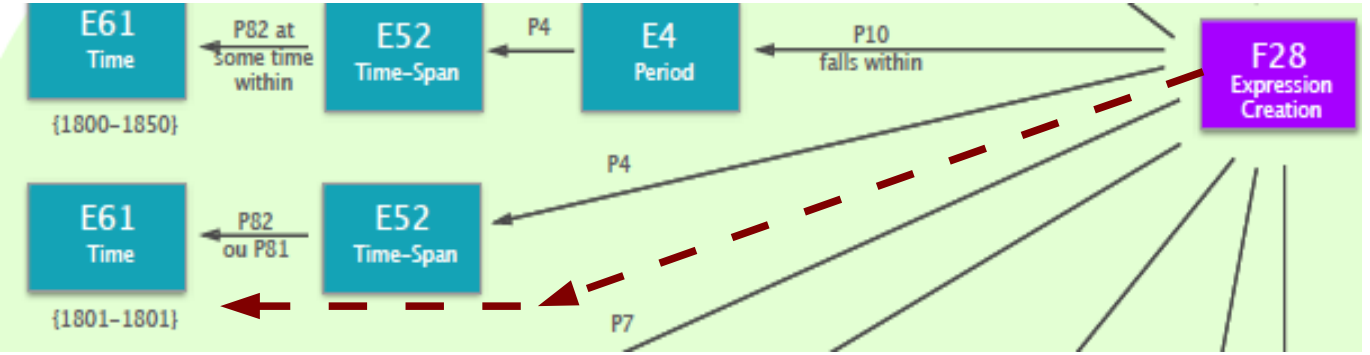
Identifier	F28
Unit of information	Work: Date of the work (representative expression)
Object	Date of expression creation
Remarks	Date and machine format
Path	F28 Expression Creation P4 has time-span E52 Time-Span P82 at some time within E61 Time Primitive
Unimarc and Intermarc Philharmonie	UNI100: 909 $g $h
Transfer rules	If $h is identical to $g, keep only $g. Add a slash between $g and $h if they have different values.
Examples	UNI100: 909 $g1801 $h1801 > E52 Time-Span P81 ongoing through E61 = 1801 UNI100:909 $g1834 $h1856 > E52 Time-Span P81 ongoing through E61 = 1834/1856

What to look for?

2. Data Conversion to RDF

Expert-defined mapping rules

An example



Identifier	F28
Unit of information	Work: Date of the work (representative expression)
Object	Date of expression creation
Remarks	Date and machine format
Path	F28 Expression Creation P4 has time-span E52 Time-Span P82 at some time within E61 Time Primitive
Unimarc and Intermarc Philharmonie	UNI100: 909 $g $h
Transfer rules	If $h is identical to $g, keep only $g. Add a slash between $g and $h if they have different values.
Examples	UNI100: 909 $g1801 $h1801 > E52 Time-Span P81 ongoing through E61 = 1801 UNI100:909 $g1834 $h1856 > E52 Time-Span P81 ongoing through E61 = 1834/1856

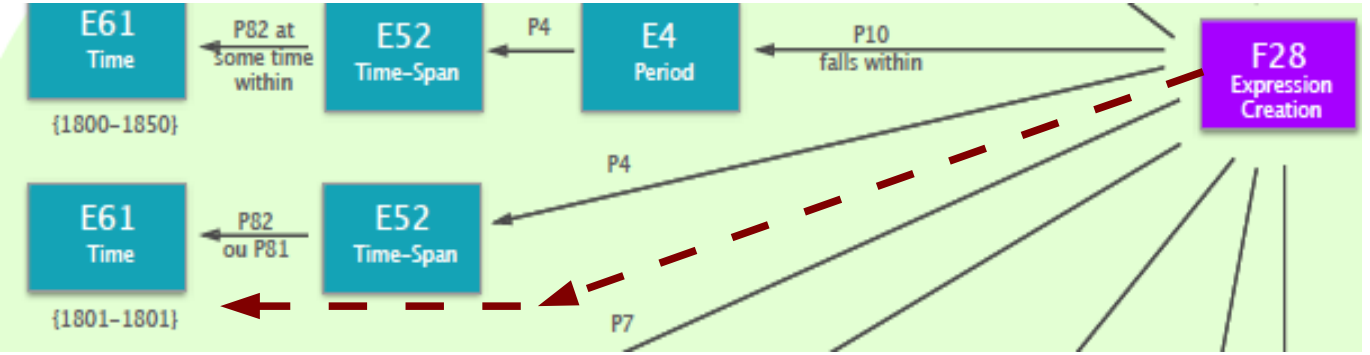
What to look for?

Where to look?

2. Data Conversion to RDF

Expert-defined mapping rules

An example



Model

Identifier	F28
Unit of information	Work: Date of the work (representative expression)
Object	Date of expression creation
Remarks	Date and machine format
Path	F28 Expression Creation P4 has time-span E52 Time-Span P82 at some time within E61 Time Primitive

What to look for?

MARC file

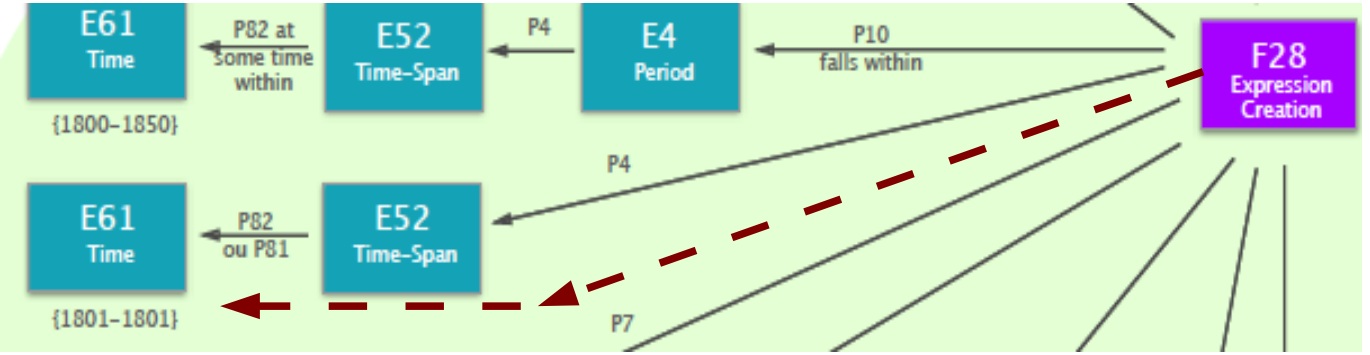
Unimarc and Intermarc Philharmonie	UNI100: 909 $g $h
Transfer rules	If $h is identical to $g, keep only $g. Add a slash between $g and $h if they have different values.
Examples	UNI100: 909 $g1801 $h1801 > E52 Time-Span P81 ongoing through E61 = 1801 UNI100:909 $g1834 $h1856 > E52 Time-Span P81 ongoing through E61 = 1834/1856

Where to look?

2. Data Conversion to RDF

Expert-defined mapping rules

An example



Identifier	F28
Unit of information	Work: Date of the work (representative expression)
Object	Date of expression creation
Remarks	Date and machine format
Path	F28 Expression Creation P4 has time-span E52 Time-Span P82 at some time within E61 Time Primitive
Unimarc and Intermarc Philharmonie	UNI100: 909 $g $h
Transfer rules	If $h is identical to $g, keep only $g. Add a slash between $g and $h if they have different values.
Examples	UNI100: 909 $g1801 $h1801 > E52 Time-Span P81 ongoing through E61 = 1801 UNI100:909 $g1834 $h1856 > E52 Time-Span P81 ongoing through E61 = 1834/1856

What to look for?

Where to look?

2. Data Conversion to RDF

DOREMUS resource URI naming convention

The DOREMUS convention combines the *best practices* (see the DataLift project [6]) with the *DOREMUS model*

the Datalift convention

http :// data. {Domain} / {Theme} / {Class}/ {Identifier}

the Doremus convention 1

http://data.doremus.org/Name/Code/UUID

*the **class** from the DOREMUS model*

Example:

http://data.doremus.org/Self_Contained_Expression/F22/b90b3b97-2526-4152-95bb-273

the Doremus convention 2 (under discussion)

http://data.doremus.org/expression/UUID

2. Data Conversion to RDF

The DOREMUS property naming convention

Properties in the DOREMUS ontology:
three namespaces

- CIDOC-CRM **cidoc-crm**: <http://www.cidoc-crm.org/cidoc-crm/>
- FRBRoo **frbroo**: <http://erlangen-crm.org/efrbroo/>
- DOREMUS **mus**: <http://data.doremus.org/ontology/>

Constructing a property URI: concatenate the namespace URI and the property identifier (code + name in the model)

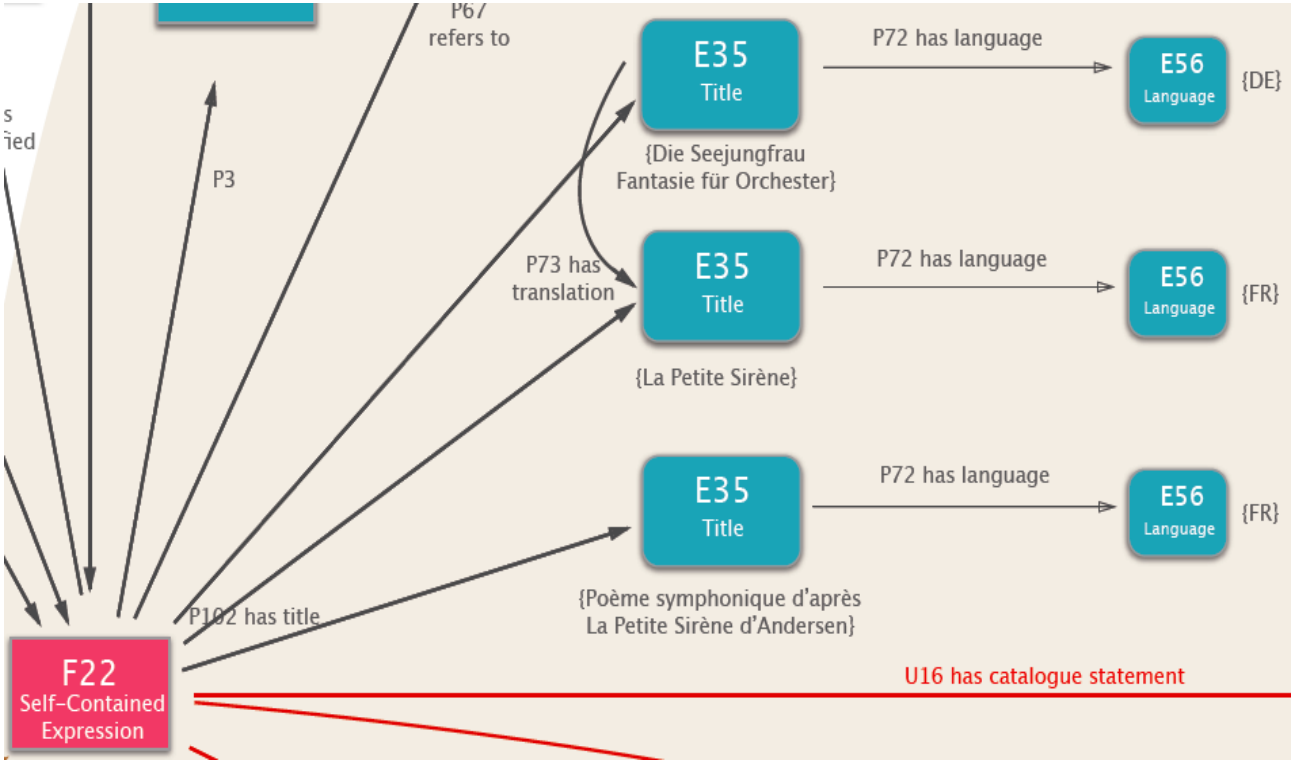
→ *see next slide.*

2. Data Conversion to RDF

The DOREMUS property naming convention

Properties are identified by their codes followed by their names.

CIDOC-CRM properties:



P102_has_title

P72_has_language

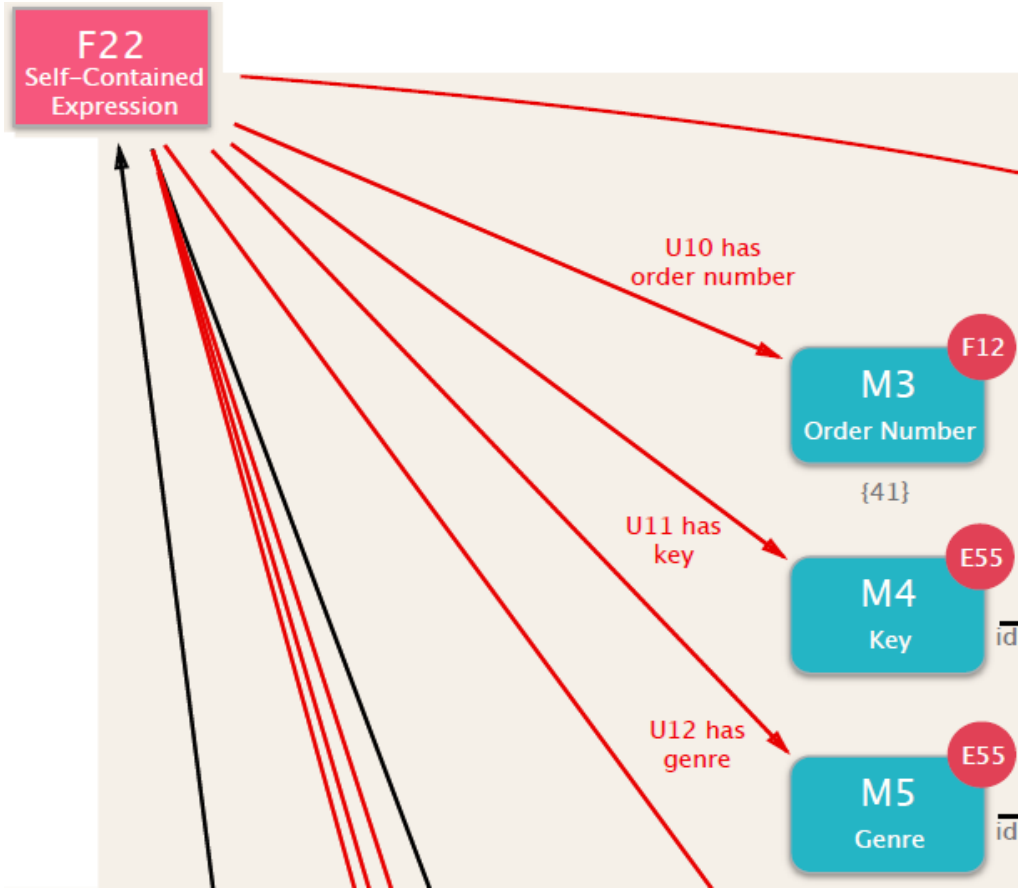
P73_has_translation

The CIDOC-CRM ns: @prefix **cidoc-crm**: <http://www.cidoc-crm.org/cidoc-crm/>

2. Data Conversion to RDF

The DOREMUS property naming convention

DOREMUS properties:



U11_has_key

U12_has_genre

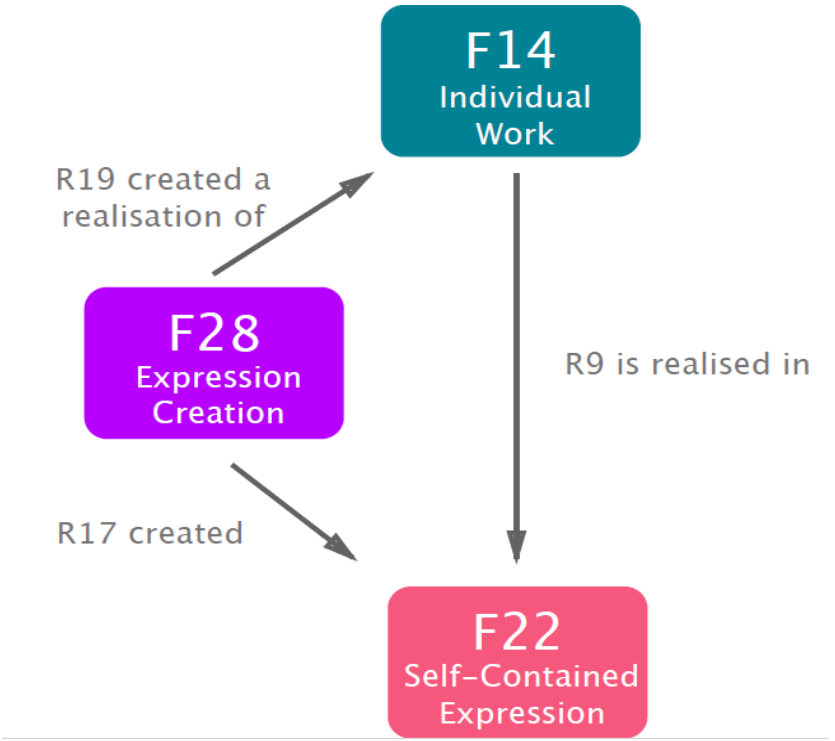
P10_has_order_number

The DOREMUS namespace: @prefix **mus**: <http://data.doremus.org/ontology/>

2. Data Conversion to RDF

The DOREMUS property naming convention

FRBRoo properties:



R17_created

R9_is_realized_in

R19_created_a_realisation_of

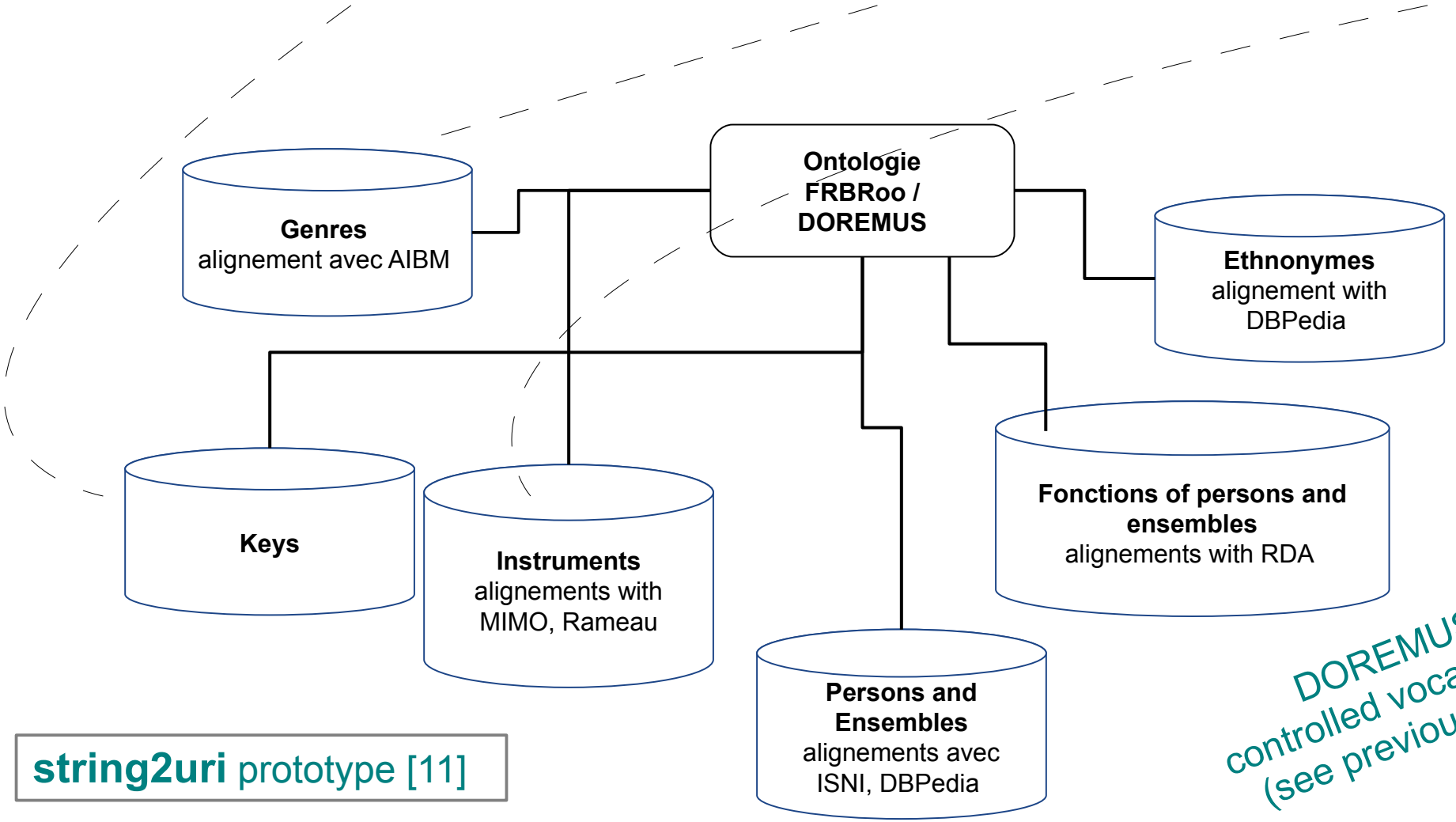
The FRBRoo namespace: @prefix **frbroo**: <http://erlangen-crm.org/efrbroo/>

2. Data Conversion to RDF

The DOREMUS properties

DOREMUS data type properties / object properties

U11_has_key “C-sharp” U12_has_genre “symphony” U13_has_casting “piano”

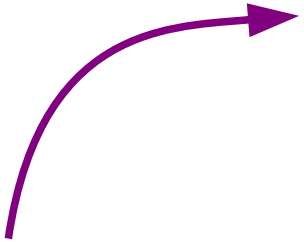


string2uri prototype [11]

DOREMUS controlled vocabularies (see previous talks).

2. Data Conversion to RDF

Example: a converted BNF TUM



```
001 FRBNF139081882
008 890821130211yy    sn        1801
048   $aka01
100   $313891295$w.0..b.....$aBeethov
144 1 $w....b.fre.$aSonates$bPiano$pO
444 1 $w....b.fre.$aSonates$bPiano$nN
444 1 $w....b.ita.$aQuasi una fantasi
444 1 $w....b.ita.$aSonata quasi una
444 1 $w....b.eng.$aMoonlight sonata
444 1 $w....b.fre.$aClair de lune$eSo
444 1 $w....b.ger.$aMondschein-Sonate
444 1 $w....b.fre.$aSonate au clair de lune
444 1 $w....b.fre.$aSonate Clair de lune
502   $314017453$aBeethoven$mLudwig van$d1770-1827$t[Sonates (2). Op. 27]
600   $aDédicace à la comtesse Giulietta Giucciardi$aDate de composition : 1801$a1re éd. : Vienne : Cappi, 1802
610   $aKinsky
610   $aGrove 7
917   $oOPC$a100366020
917   $oOPD$a100087890$bATUM
996   $oOPP$a14786691$d20060411
996   $oOPP$a16305693$d20130211
400   $w....b.....$aBeethoven$mLudwig von$d1770-1827
```

```
<http://data.doremus.org/Self_Contained_Expression/F22/061b4ccd-ac20-42ff-b571-4b8ce41e864c>
  ns0:U11_has_key [ ns1:P1_is_identified_by "Do dièse mineur"@fr ] ;
  ns0:U12_has_genre [ ns1:P1_is_identified_by "sonate"@fr ] ;
  ns0:U13_has_casting "Piano" ;
  ns0:U17_has_opus_statement [
    ns1:P106_is_composed_of "27", "2" ;
    ns1:P3_has_note "Op. 27, no 2"
  ] ;
  ns1:P102_has_title "Sonate Clair de lune"@fr ;
  ns1:P67_refers_to [ ns1:P3_has_note "Dédicace à la comtesse Giulietta Giucciardi" ] .

<http://data.doremus.org/Expression_Creation/F28/4a91d2a7-62ac-4b87-899a-406fa95efc91>
  ns2:R17_created <http://data.doremus.org/Self_Contained_Expression/F22/061b4ccd-ac20-42ff-b571-4b8ce41e864c> ;
  ns1:P4_has_time_span [
    a ns1:E52_Time_Span ;
    ns1:P82_at_some_time_within "18010101/18011231"^^ns3:terms-W3CDTF
  ] ;
  ns1:P9_consists_of [
    a ns1:E7_activity ;
    ns1:P14_carried_out_by [
      a ns1:E21_Person ;
      ns1:P131_is_identified_by "Beethoven,Ludwig van(1770-1827)"
    ] ;
    ns1:U35_had_function_of_type "compositeur"
  ] .
```

2. Data Conversion to RDF

Data describing a work in the Philharmonie de Paris have to be looked up in two different records.

PP - Work Record

```
019 $aUNI100
100 $a20041214d||||    uuuy0frey0103    ba
200 $aSonate pour piano no 14 "Clair de lune"$fLudwig
500 $30804231$aSonates$rPiano$sOp. 27 no 2$uDo dièse m
610 $30068838$aSonate$b04
610 $30144424$a19 ème siècle$b02
610 $30067958$aMusique romantique$b04
610 $30144079$aPiano$b01
700 $30038954$aBeethoven$bLudwig van$f1770-1827$4230
830 $aBN OPALE-PLUS 2007/02/26. Guide de la musique de
909 $aDédicace à la comtesse Giulietta Giucciardi. Par
919 $aPremière publication : Vienne, Cappi, 1802
937 $30069690$aLes 32 [Trente-deux$bUNI1
937 $30072431$aBeethoven piano sonatas / Denis Matthew
937 $30078731$aLes Sonates de Beethoven / Paul Badura-
937 $30086700$aWilhelm Kempff : Schumann : Arabeske, Papillons, Davidsbündlertänze ; Beethoven : Piano sonatas n° 14 "Moonl
937 $30094153$aKlaviersonaten : Band I, II / Beethoven ; nach Eigenschriften, Abschriften und Originalausgaben
937 $30183795$aMondschein Sonata, Op.27 N°2 : 1er mvt. / Ludwig van Beethoven
937 $30817410$aBarenboim on Beethoven : the complete piano sonatas, live from
938 $30075908$aBeethoven : intégrale des sonates pour piano : du jeudi 8 au d
938 $30081170$aSonate no. 14 cis-moll op. 27 n° 2 "Mondschein-Sonate" / Ludwi
938 $30235758$aLudwig van Beethoven [compilation$bUNI2
938 $30583341$aNocturnes I - Clairs de lune : du dimanche 2 au jeudi 6 mai 20
938 $30789054$aSonate no 14 en ut dièse mineur, op. 27 no 2 "quasi una fantas
938 $30889727$aBeethoven / Debussy : dimanche 12 octobre 2008 / Denis Herlin,
938 $30890637$aIntégrale des sonates pour piano : CD 4 / Ludwig van Beethoven
938 $30897214$aSonate no 14 en ut dièse mineur, op. 27 no 2 "quasi una fantas
938 $30906429$aThe Complete piano sonatas on period instruments / Beethoven,
938 $31003004$aAdagio sostenuto, extrait de la "Sonate no 14 en ut dièse mine
938 $31040755$aComplete piano sonatas : CD 4 : Sonatas Opp. 26, 27, 28 / Ludwig van Beethoven, Composition ; Maurizio Pollini
938 $31042188$aJe découvre la musique classique CD 2 : S'amuser en classique / Antonio Vivaldi ; jacques Offenbach ; Frédér
```

```
<http://data.doremus.org/Self_Contained_Expression/F22/430197c2-5a4c-416e-ba03-80f211c2dcf6>
  ns0:U10_has_order_number "14" ;
  ns0:U11_has_key [ ns1:P1_is_identified_by "Do dièse mineur"@fr ] ;
  ns0:U13_has_casting [ ns1:P3_has_note "20040721" ], [ ns1:P3_has_note "Piano" ] ;
  ns0:U17_has_opus_statement [
    ns1:P106_is_composed_of "27no2" ;
    ns1:P3_has_note "Op. 27 no 2"
  ] ;
  ns1:P102_has_title "Sonate pour piano no 14 \"Clair de lune\"", "Sonate au clair de lune" ;
  ns1:P3_has_note "FR. ", "Dédicace à la comtesse Giulietta Giucciardi. Parue sous le nom de \"Sonate pour piano quasi una fantasi
a en ut dièse mineur, alla Damigella comtessa Giuletta Guicciardi\". Le titre \"Clair de lune\" fut inventé par le poète Ludwig Ra
llstab. Comprend : 1- adagio sostenuto, 2- allegretto, 3- presto agitato. Première publication : Vienne, Cappi, 1802" .

<http://data.doremus.org/Expression_Creation/F28/6ab49882-fa9a-4db0-b3ee-98185589bc16>
  ns2:R17_created <http://data.doremus.org/Self_Contained_Expression/F22/430197c2-5a4c-416e-ba03-80f211c2dcf6> ;
  ns1:P3_has_note "1801", "CITE MUSIQUE" ;
  ns1:P4_has_time_span [
    a ns1:E52_Time_Span ;
    ns1:P82_at_some_time_within "1801"^^ns3:terms-W3CDTF
  ] ;
  ns1:P9_consists_of [
    a ns1:E7_activity ;
    ns1:P14_carried_out_by [
      a ns1:E21_Person ;
      ns1:P131_is_identified_by "Beethoven,Ludwig van(1770-1827)"
    ] ;
    ns1:U35_had_function_of_type "compositeur"
```

PP - TUM

```
017 $aATU243$oOPSYS
019 $aAIC14$b243
144 $aSonates$rPiano$nNo 14$pOp. 27 no 2$tDo dièse mineur$uClair de lune
322 $30038954$aBeethoven$mLudwig van$d1770-1827
444 $aMoon light sonata$rPiano$nNo 14$pOp. 27 no 2
444 $w||||||||||$aSonate Mondschein$rPiano$nNo 14$pOp. 27 no 2$tDo dièse mineur
444 $aSonata quasi una fantasia
444 $aSonate au clair de lune
602 $afr
909 $aFR$bCITE MUSIQUE$c20040721
```

2. Data Conversion to RDF

You can find the current version of the MRAC2DOREMUS-RDF converter here.

https://github.com/DOREMUS-ANR

```
001 FRBNF139081882
008 890821130211yy    sn        1801
048   $aka01
100   $313891295$w.0..b.....$aBeethov
144 1 $w....b.fre.$aSonates$bPiano$pO
444 1 $w....b.fre.$aSonates$bPiano$nN
444 1 $w....b.ita.$aQuasi una fantasi
444 1 $w....b.ita.$aSonata quasi una
444 1 $w....b.eng.$aMoonlight sonata
444 1 $w....b.fre.$aClair de lune$eSo
444 1 $w....b.ger.$aMondschein-Sonate
444 1 $w....b.fre.$aSonate au clair de lune
444 1 $w....b.fre.$aSonate Clair de lune
502   $3140174 53$aBeethoven$mLudwig van$d1770-1827$t[Sonates (2). Op. 27]
600   $aDédicace à la comtesse Giulietta Giucciardi$aDate de composition : 1801$a1re éd. : Vienne : Cappi, 1802
610   $aKinsky
610   $aGrove 7
917   $oOPC$a100366020
917   $oOPD$a100087890$bATUM
996   $oOPP$a14786691$d20060411
996   $oOPP$a16305693$d20130211
400   $w....b.....$aBeethoven$mLudwig von$d1770-1827
```

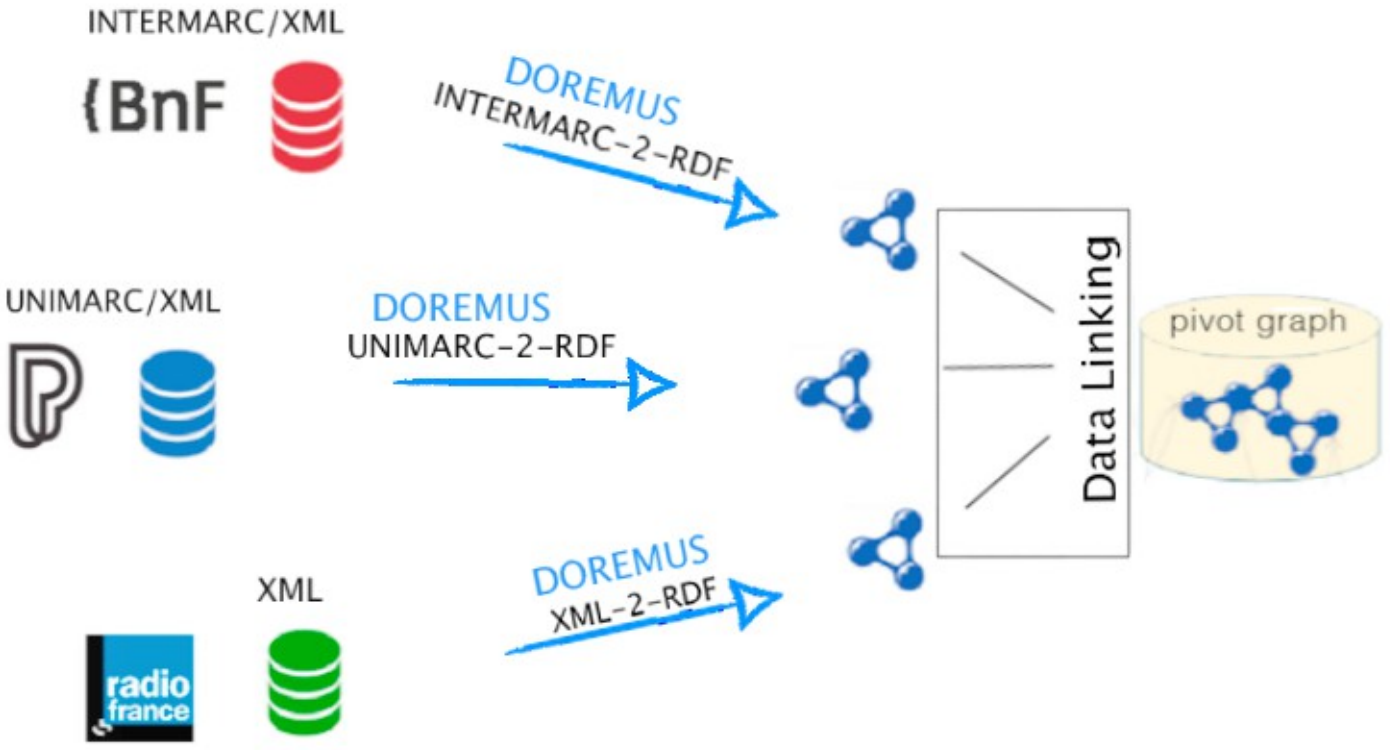
```
<http://data.doremus.org/Self_Contained_Expression/F22/061b4ccd-ac20-42ff-b571-4b8ce41e864c>
  ns0:U11_has_key [ ns1:P1_is_identified_by "Do dièse mineur"@fr ] ;
  ns0:U12_has_genre [ ns1:P1_is_identified_by "sonate"@fr ] ;
  ns0:U13_has_casting "Piano" ;
  ns0:U17_has_opus_statement [
    ns1:P106_is_composed_of "27", "2" ;
    ns1:P3_has_note "Op. 27, no 2"
  ] ;
  ns1:P102_has_title "Sonate Clair de lune"@fr ;
  ns1:P67_refers_to [ ns1:P3_has_note "Dédicace à la comtesse Giulietta Giucciardi" ] .

<http://data.doremus.org/Expression_Creation/F28/4a91d2a7-62ac-4b87-899a-406fa95efc91>
  ns2:R17_created <http://data.doremus.org/Self_Contained_Expression/F22/061b4ccd-ac20-42ff-b571-4b8ce41e864c> ;
  ns1:P4_has_time_span [
    a ns1:E52_Time_Span ;
    ns1:P82_at_some_time_within "18010101/18011231"^^ns3:terms-W3CDTF
  ] ;
  ns1:P9_consists_of [
    a ns1:E7_activity ;
    ns1:P14_carried_out_by [
      a ns1:E21_Person ;
      ns1:P131_is_identified_by "Beethoven,Ludwig van(1770-1827)"
    ] ;
    ns1:U35_had_function_of_type "compositeur"
  ] .
```

Work in progress...

Outline

3. Data Linking



3. Data Linking

...Anyone?

The 4th principle of the web of data:
when publishing data, provide links to other, already published data!

Link datasets on the web!

3. Data Linking

Links

A link-statement is a triple (as any other) that

links an instance from one dataset (*the subject*)

to an instance of another dataset (*the object*)

via a *link-predicate* coming from established vocabularies, such as owl:sameAs (meaning that the 2 instances are equivalent), but also skos:closeMatch, rdf:seeAlso, or other.

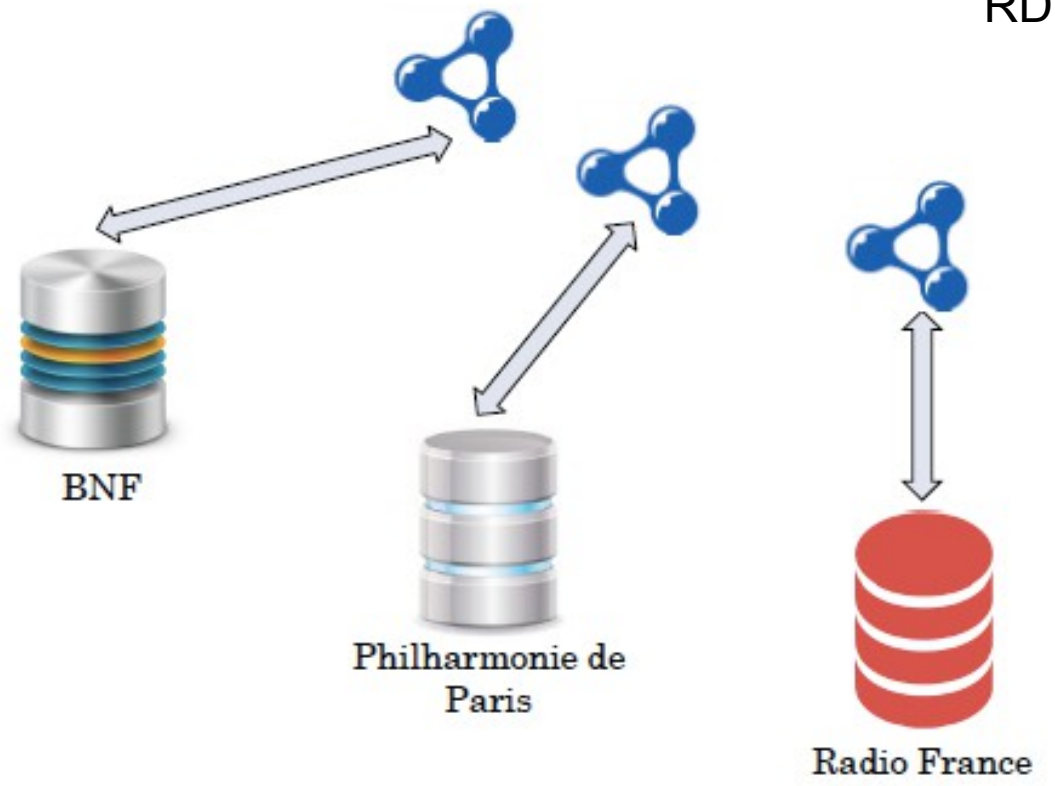
Example:

`http://yago-knowledge.org/resource/Ludwig_van_Beethoven`,
owl:sameAs, `http://dbpedia.org/resource/Ludwig_van_Beethoven`

3. Data Linking

DOREMUS: What do we have so far?

An RDF graph per institution.



A work exists potentially in each of the 3 RDF datasets identified by different URIs.

Among the reasons for this decision:

– the descriptions of a given work across institutions are not uniform (see following slides)
– not always a 1:1 correspondance
– independence of representation

So, we need to link these datasets!

3. Data Linking

Some basics:
The data linking processing chain

(1) preprocessing → (2) instance matching → (3) post-processing

– reduce the search-space, identify a set of pairs of linking candidates, identify key properties

– make instances comparable: models of representation, handling multilingualisme

See [4],[5].

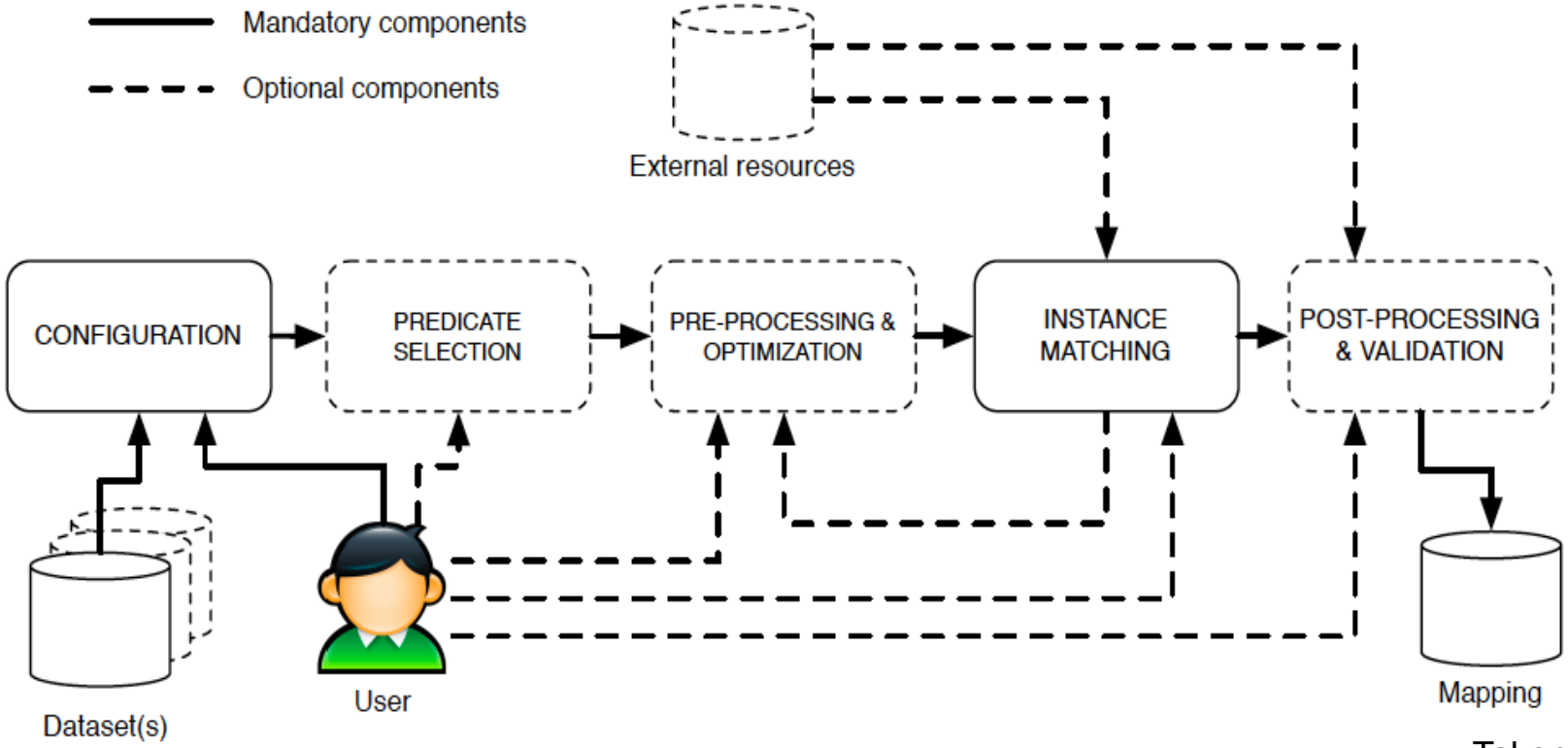
– discover a link between two resources, give it a type and a cofidence value

See [1].

– filter out erroneous matches
– infer new ones

3. Data Linking

A generic architecture



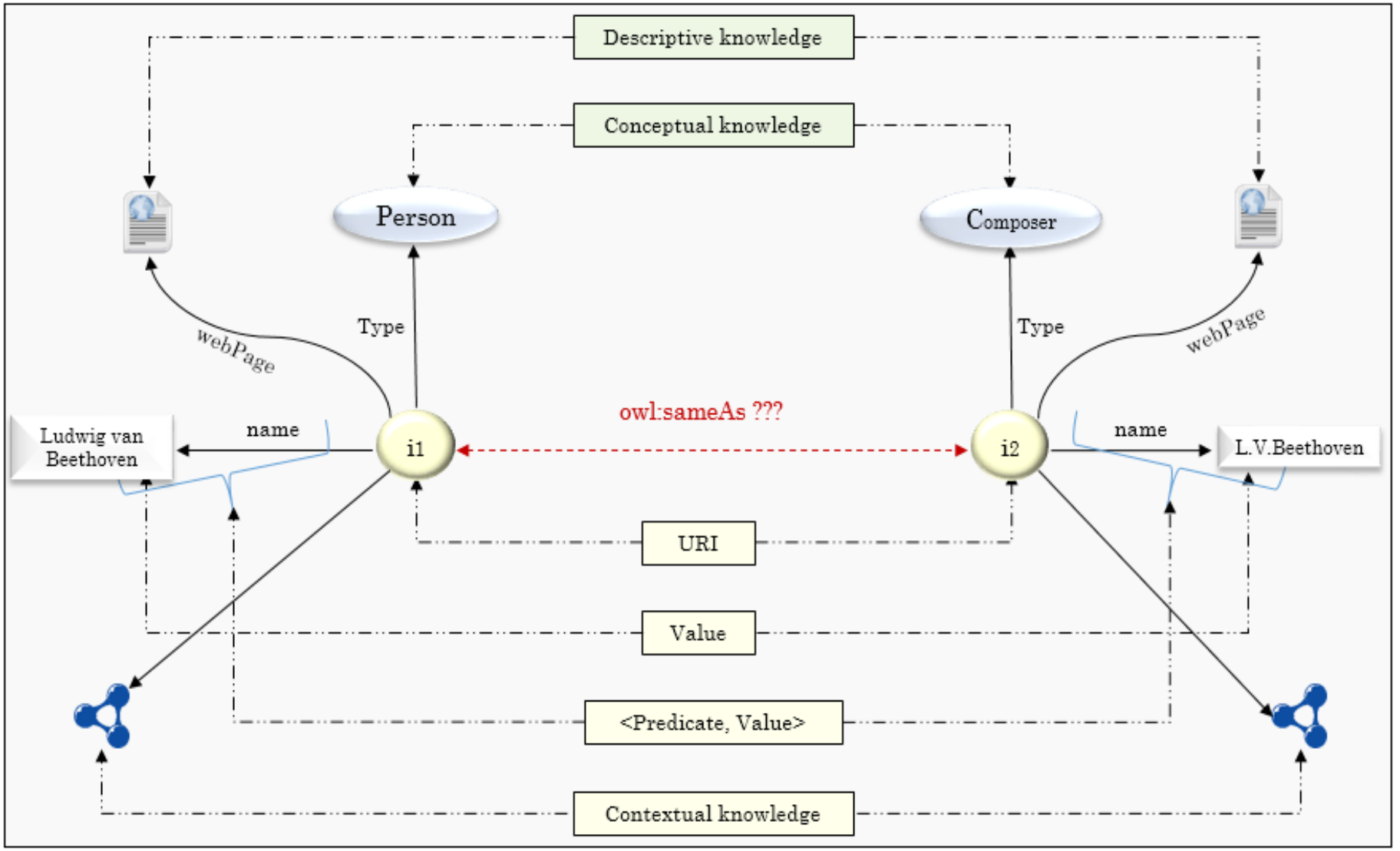
Taken from [1].

A plethora of tools:
LIMES http://aksw.org/Projects/LIMES.htm,
SILK http://silkframework.org, RiMOM, RDF-AI,...
See OAEI for more: http://oaei.ontologymatching.org/2015/im/index.html

From a user perspective, the tool configuration is 90% of the task.

3. Data Linking

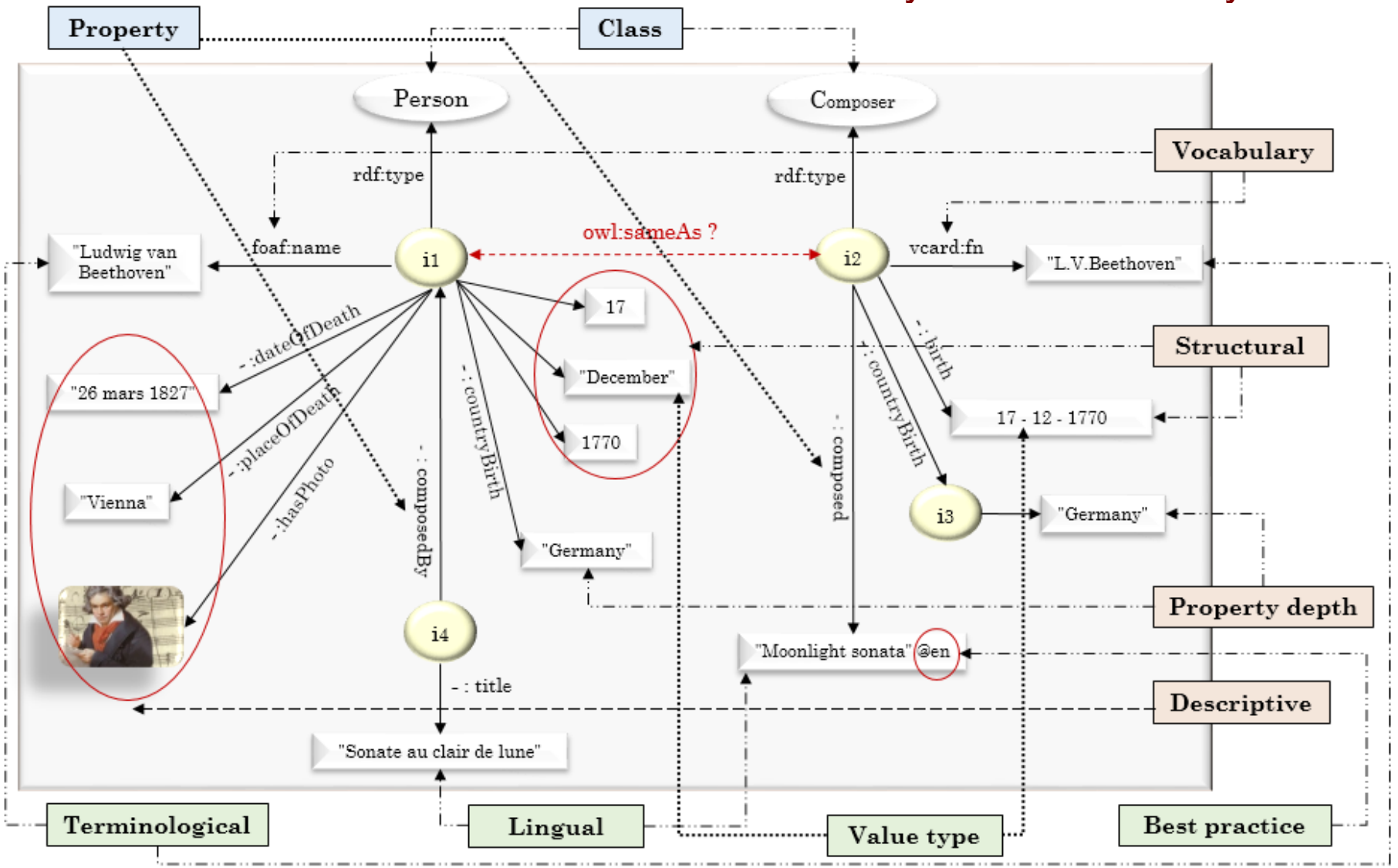
Levels of comparison



Where to look for information to compare instances?

3. Data Linking

Why is it not that easy...



Datasets can be highly heterogeneous!

3. Data Linking

Why is it not that easy...

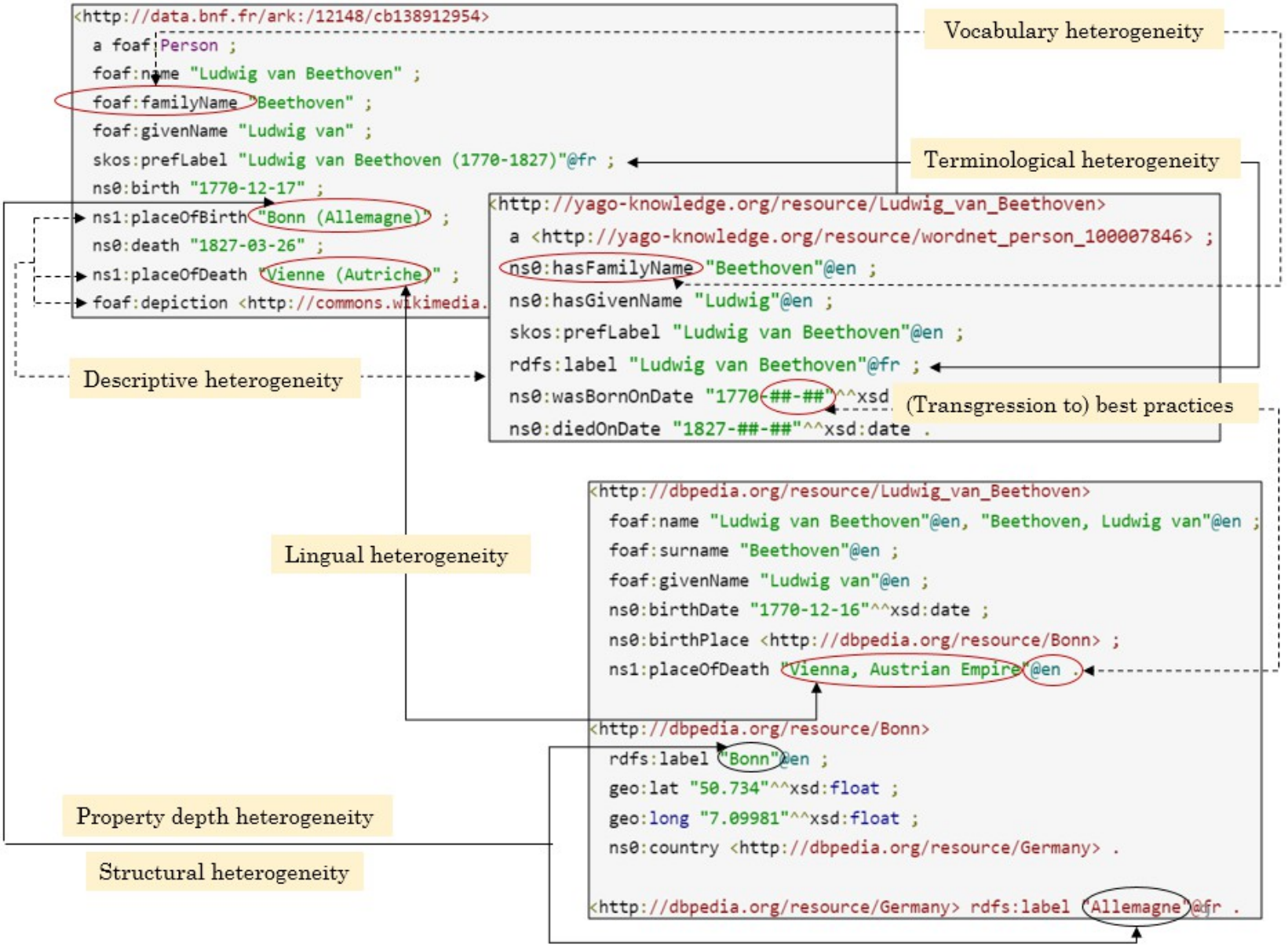
Data heterogeneity: any difference in the description or expression of equivalent resources and information.

“Moonligth sonata” “Sonate au claire de lune”

Title of a music work

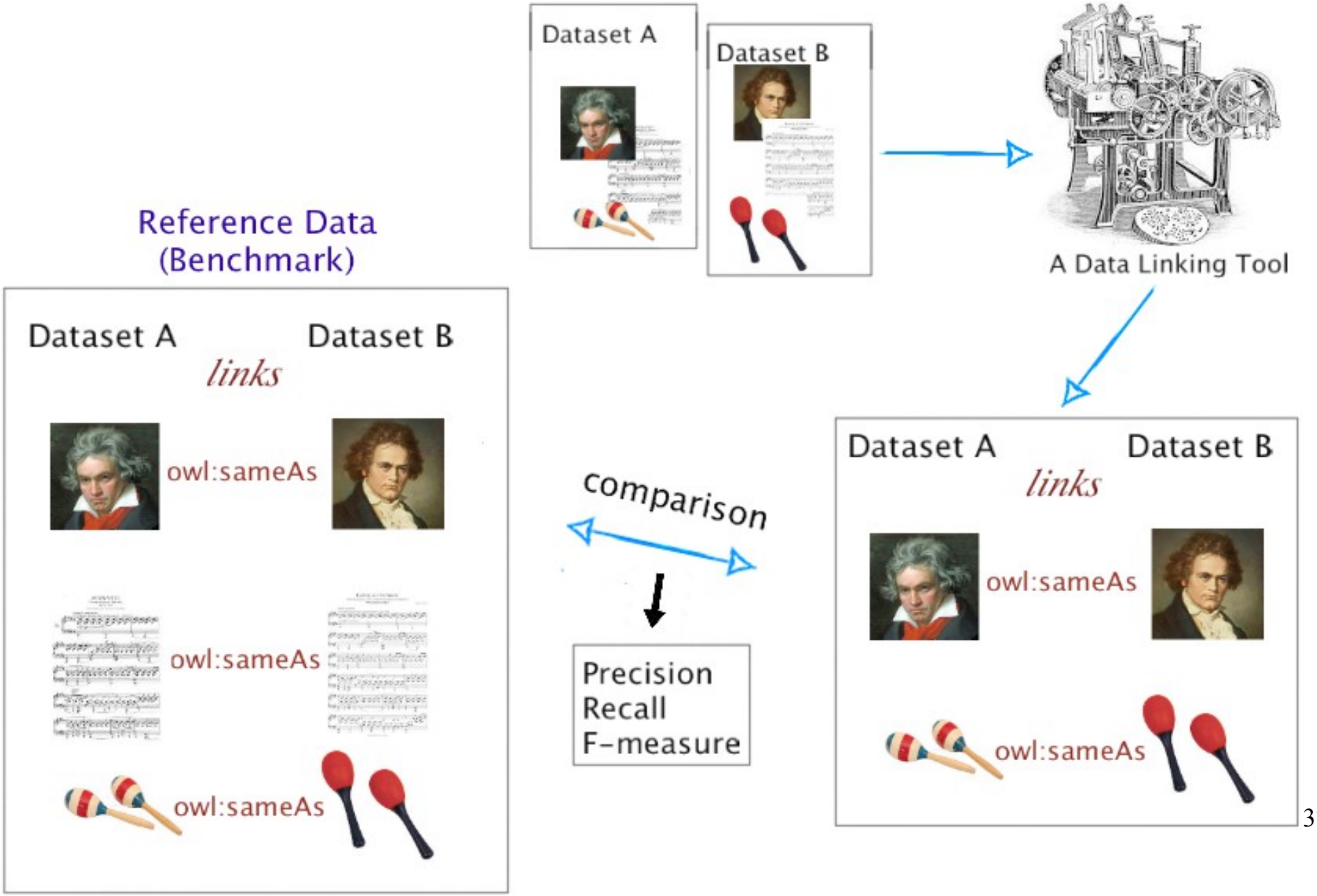
3. Data Linking

Some examples...



3. Data Linking

A common approach to develop and evaluate linking tools



3. Data Linking

The DOREMUS benchmark data
Dataset 1:
Nine heterogeneities

What are the **heterogeneities** manifested by music bibliographical data?

We asked experts to identify the most current problems that may appear.

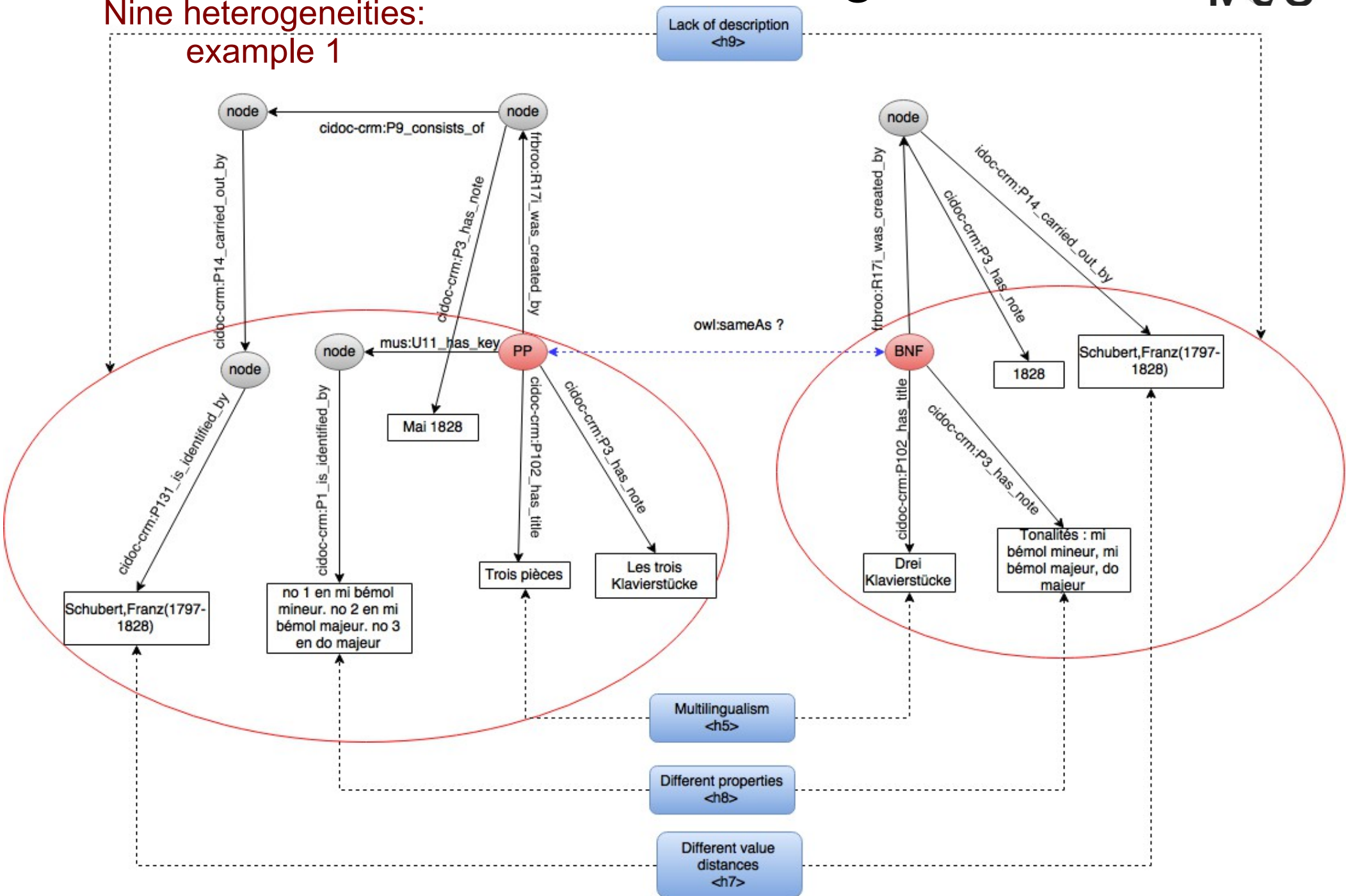
We did some tests.

H1. Letters or numbers in the property values (particularly titles)
H2. Differences in spelling (terminological heterogeneity)
H3. Missing catalog numbers and/or opus numbers
H4. Different catalogues *(no works so far)*
H5. Multilingual titles
H6. Letters with diacritical signs
H7. Different value distances
H8. Different properties describing tonalities or instruments
H9. Missing properties (lack of description)
H10. Missing titles

–- a small dataset of corresponding pairs of works from the BnF and the Philharmonie de Paris, organised per category, available here.

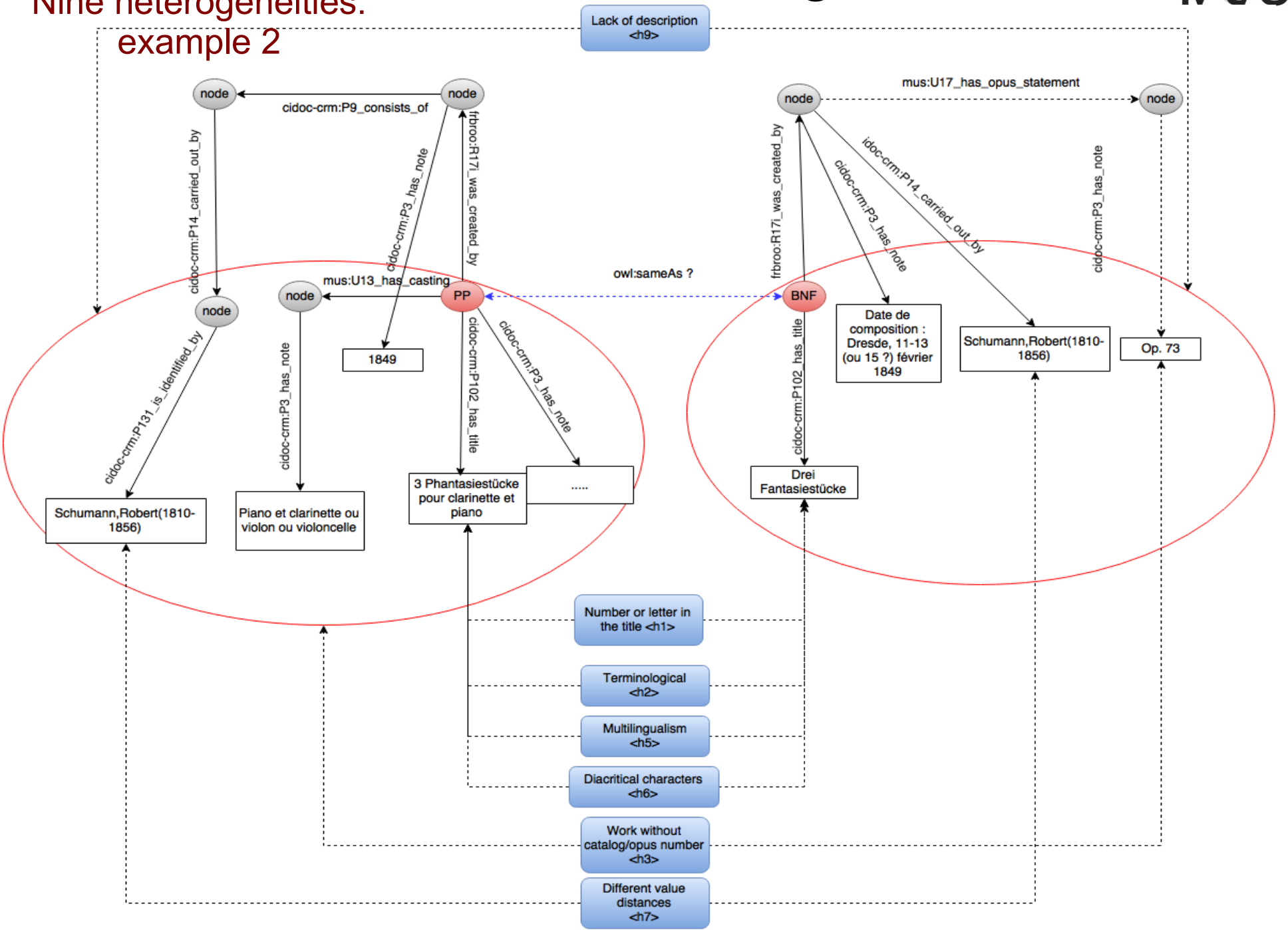
3. Data Linking

Nine heterogeneities:
example 1



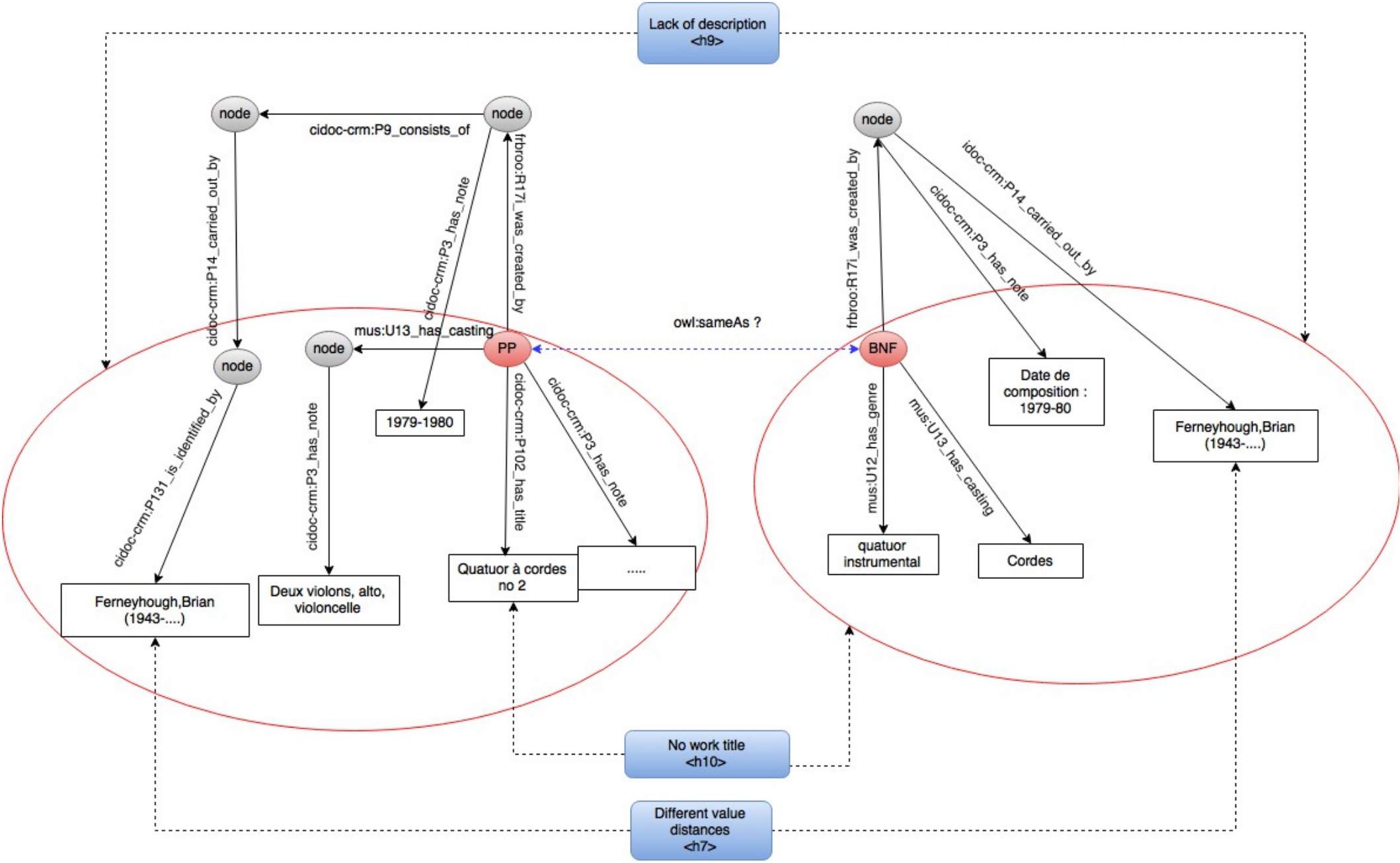
3. Data Linking

Nine heterogeneities: example 2



3. Data Linking

Nine heterogeneities: example 3



3. Data Linking

The DOREMUS benchmark data
Dataset 2:
Four Heterogeneities

SILK: the only instance matching tool that returned results.

After testing, we selected 4 groups of heterogeneities that appeared to be most problematic for the linking tool.

H2. Differences in spelling (terminological heterogeneity)
H5. Multilingual titles
H9. Missing properties (lack of description)
H10. Missing titles

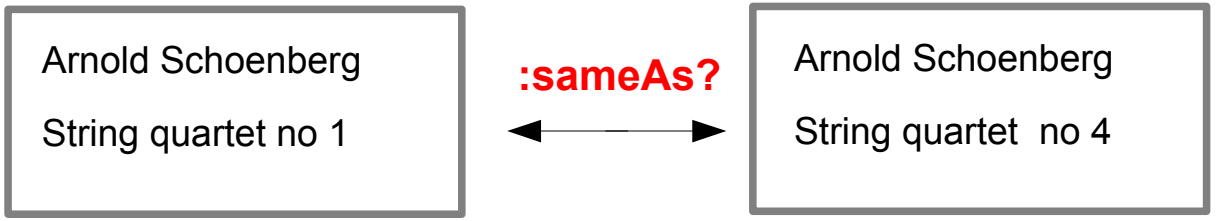
– a larger dataset of about 200 pairs of works, organised wrt the four categories, available here.

3. Data Linking

The DOREMUS benchmark data
Dataset 3:
The False Positives Trap

Again, we asked experts for help.

A dataset containing pairs of **different** musical works that are **highly similar** in their descriptions (same composer, title, key, instruments...).



Challenge the linking tools capacity to discover *difficult* discriminative properties.

A dataset of around 50 pairs of instances.

3. Data Linking

The DOREMUS benchmark data
Dataset 4:
Machine Learning

A dataset for learning automatic classifiers.

example	class
(w1, w')	*same*
(w2,w'')	*different*
...	...

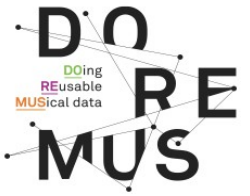
Training data: examples of pairs of works with a class label (same/different).

A standard binary classification problem setting.

Learning a *prediction rule* that allows to correctly classify an unseen example (a pair of works) to one of the two categories: *same* or *different*.

3. Data Linking

Let's take an example...



3. Data Linking

```
<http://data.doremus.org/Self_Contained_Expression/F22/061b4ccd-ac20-42ff-b571-4b8ce41e864c>
  ns0:U11_has_key [ ns1:P1_is_identified_by "Do dièse mineur"@fr ] ;
  ns0:U12_has_genre [ ns1:P1_is_identified_by "sonate"@fr ] ;
  ns0:U13_has_casting "Piano" ;
  ns0:U17_has_opus_statement [
    ns1:P106_is_composed_of "27", "2" ;
    ns1:P3_has_note "Op. 27, no 2"
  ] ;
  ns1:P102_has_title "Sonate Clair de lune"@fr ;
  ns1:P67_refers_to [ ns1:P3_has_note "Dédicace à la comtesse Giulietta Giucciardi" ] .

<http://data.doremus.org/Expression_Creation/F28/4a91d2a7-62ac-4b87-899a-406fa95efc91>
  ns2:R17_created <http://data.doremus.org/Self_Contained_Expression/F22/061b4ccd-ac20-42ff-b571-4b8ce41e864c> ;
  ns1:P4_has_time_span [
    a ns1:E52_Time_Span ;
    ns1:P82_at_some_time_within "18010101/18011231"^^ns3:terms-W3CDTF
  ] ;
  ns1:P9_consists_of [
    a ns1:E7_activity ;
    ns1:P14_carried_out_by [
      a ns1:E21_Person ;
      ns1:P131_is_identified_by "Beethoven,Ludwig van(1770-1827)"
    ] ;
    ns1:U35_had_function_of_type "compositeur"
  ] .
```

```
<http://data.doremus.org/Self_Contained_Expression/F22/430197c2-5a4c-416e-ba03-80f211c2dcf6>
  ns0:U10_has_order_number "14" ;
  ns0:U11_has_key [ ns1:P1_is_identified_by "Do dièse mineur"@fr ] ;
  ns0:U13_has_casting [ ns1:P3_has_note "20040721" ], [ ns1:P3_has_note "Piano" ] ;
  ns0:U17_has_opus_statement [
    ns1:P106_is_composed_of "27no2" ;
    ns1:P3_has_note "Op. 27 no 2"
  ] ;
  ns1:P102_has_title "Sonate pour piano no 14 \"Clair de lune\"", "Sonate au clair de lune" ;
  ns1:P3_has_note "FR. ", "Dédicace à la comtesse Giulietta Giucciardi. Parue sous le nom de \"Sonate pour piano
a en ut dièse mineur, alla Damigella comtessa Giuletta Guicciardi\". Le titre \"Clair de lune\" fut inventé par
llstab. Comprend : 1- adagio sostenuto, 2- allegretto, 3- presto agitato. Première publication : Vienne, Cappi,

<http://data.doremus.org/Expression_Creation/F28/6ab49882-fa9a-4db0-b3ee-98185589bc16>
  ns2:R17_created <http://data.doremus.org/Self_Contained_Expression/F22/430197c2-5a4c-416e-ba03-80f211c2dcf6> ;
  ns1:P3_has_note "1801", "CITE MUSIQUE" ;
  ns1:P4_has_time_span [
    a ns1:E52_Time_Span ;
    ns1:P82_at_some_time_within "1801"^^ns3:terms-W3CDTF
  ] ;
  ns1:P9_consists_of [
    a ns1:E7_activity ;
    ns1:P14_carried_out_by [
      a ns1:E21_Person ;
      ns1:P131_is_identified_by "Beethoven,Ludwig van(1770-1827)"
    ] ;
    ns1:U35_had_function_of_type "compositeur"
  ]
```

3. Data Linking

```
<http://data.doremus.org/Self_Contained_Expression/F22/061b4ccd-ac20-42ff-b571-4b8ce41e864c>
  ns0:U11_has_key [ ns1:P1_is_identified_by "Do dièse mineur"@fr ] ;
  ns0:U12_has_genre [ ns1:P1_is_identified_by "sonate"@fr ] ;
  ns0:U13_has_casting "Piano" ;
  ns0:U17_has_opus_statement [
    ns1:P106_is_composed_of "27", "2" ;
    ns1:P3_has_note "Op. 27, no 2"
  ] ;
  ns1:P102_has_title "Sonate Clair de lune"@fr ;
  ns1:P67_refers_to [ ns1:P3_has_note "Dédicace à la comtesse Giulietta Giucciardi" ] .

<http://data.doremus.org/Expression_Creation/F28/4a91d2a7-62ac-4b87-899a-406fa95efc91>
  ns2:R17_created <http://data.doremus.org/Self_Contained_Expression/F22/061b4ccd-ac20-42ff-b571-4b8ce41e864c> ;
  ns1:P4_has_time_span [
    a ns1:E52_Time_Span ;
    ns1:P82_at_some_time_within "18010101/18011231"^^ns3:terms-W3CDTF
  ] ;
  ns1:P9_consists_of [
    a ns1:E7_activity ;
    ns1:P14_carried_out_by [
      a ns1:E21_Person ;
      ns1:P131_is_identified_by "Beethoven,Ludwig van(1770-1827)"
    ] ;
    ns1:U35_had_function_of_type "compositeur"
  ] .
```

```
<http://data.doremus.org/Self_Contained_Expression/F22/430197c2-5a4c-416e-ba03-80f211c2dcf6>
  ns0:U10_has_order_number "14" ;
  ns0:U11_has_key [ ns1:P1_is_identified_by "Do dièse mineur"@fr ] ;
  ns0:U13_has_casting [ ns1:P3_has_note "20040721" ], [ ns1:P3_has_note "Piano" ] ;
  ns0:U17_has_opus_statement [
    ns1:P106_is_composed_of "27no2" ;
    ns1:P3_has_note "Op. 27 no 2"
  ] ;
  ns1:P102_has_title "Sonate pour piano no 14 \"Clair de lune\"", "Sonate au clair de lune" ;
  ns1:P3_has_note "FR. ", "Dédicace à la comtesse Giulietta Giucciardi. Parue sous le nom de \"Sonate pour piano
a en ut dièse mineur, alla Damigella comtessa Giuletta Guicciardi\". Le titre \"Clair de lune\" fut inventé par
llstab. Comprend : 1- adagio sostenuto, 2- allegretto, 3- presto agitato. Première publication : Vienne, Cappi,

<http://data.doremus.org/Expression_Creation/F28/6ab49882-fa9a-4db0-b3ee-98185589bc16>
  ns2:R17_created <http://data.doremus.org/Self_Contained_Expression/F22/430197c2-5a4c-416e-ba03-80f211c2dcf6> ;
  ns1:P3_has_note "1801", "CITE MUSIQUE" ;
  ns1:P4_has_time_span [
    a ns1:E52_Time_Span ;
    ns1:P82_at_some_time_within "1801"^^ns3:terms-W3CDTF
  ] ;
  ns1:P9_consists_of [
    a ns1:E7_activity ;
    ns1:P14_carried_out_by [
      a ns1:E21_Person ;
      ns1:P131_is_identified_by "Beethoven,Ludwig van(1770-1827)"
    ] ;
    ns1:U35_had_function_of_type "compositeur"
  ]
```

Linking tools look for equivalent properties with similar/identical values

3. Data Linking

Linking Configurations and Tests

```
<?xml version="1.0" encoding="utf-8" ?>
<Silk>
…
 <DataSources>
  <DataSource type="file" id="ontoA">
   <Param name="file" value="/pathFile/0804232.rdf" />
</DataSource>
  <DataSource type="file" id="ontoB">
   <Param name="file" value="/pathFile/13908188.rdf" />
</DataSource>
 </DataSources>
…
    <SourceDataset dataSource="ontoA" var="a">
     <RestrictTo>
      ?a cidoc-crm:P102_has_title ?r .
     </RestrictTo>
    </SourceDataset>

    <TargetDataset dataSource="ontoB" var="b">
     <RestrictTo>
      ?b cidoc-crm:P102_has_title ?t .
     </RestrictTo>
    </TargetDataset>
…
<Compare metric="levenshtein" threshold="1" required="true">
        <TransformInput function="tokenize">
          <Input path="?a/cidoc-crm:P102_has_title" />
        </TransformInput>
         <TransformInput function="tokenize">
          <Input path="?b/cidoc-crm:P102_has_title" />
        </TransformInput>
       </Compare>
…

</Silk>
```

Using only titles.

Specify the path of the two datasets

Restrict the instances to those having the properties listed here

Tune the similarity metric

Specify the pairs of properties to be compared

The two resources were interconnected with a threshold equal to 0.9.

3. Data Linking

Linking Configurations and Tests

Using all properties.

```
<?xml version="1.0" encoding="utf-8" ?>
<Silk>
…
 <DataSources>
  <DataSource type="file" id="ontoA">
   <Param name="file" value="/pathFile/0804232.rdf" />
</DataSource>
  <DataSource type="file" id="ontoB">
   <Param name="file" value="/pathFile/13908188.rdf" />
</DataSource>
 </DataSources>
…
<Compare metric="levenshtein" threshold="1" required="true">
        <TransformInput function="tokenize">
         <Input path="?a/cidoc-crm:P102_has_title" />
        </TransformInput>
         <TransformInput function="tokenize">
         <Input path="?b/cidoc-crm:P102_has_title" />
        </TransformInput>
       </Compare>
<Compare metric="levenshtein" threshold="1" required="true">
        <TransformInput function="tokenize">
         <Input path="?a/cidoc-crm:P3_has_note" />
        </TransformInput>
         <TransformInput function="tokenize">
         <Input path="?b/cidoc-crm:P67_refers_to/cidoc-crm:P3_has_note" />
        </TransformInput>
       </Compare>
…

</Silk>
```

Specify the path of the two datasets

Tune the similarity metric

Specify the properties to be compared

The two resources were interconnected with a threshold equal to 0.9.

3. Data Linking

Lessons Learned

Lessons learned:

– SILK is the only off-the-shelf tool that returns results without any data re-writing

– Heterogeneities in titles appear to be very problematic

– Multilingual information is hard to handle correctly

– Need for a specific method for linking musical data
  – combine expert knowledge with
  – automatic key-discovery

— Coming up:

DOREMUS instance matching track at IM@OAEI (ISWC 2016) in Kobe!

http://oaei.ontologymatching.org

3. Data Linking

The audience might ask...

How about representing events (concerts, recordings), expressions (scores) and linking them to works?

How about finding more complicated types of relations, other than owl:sameAs?

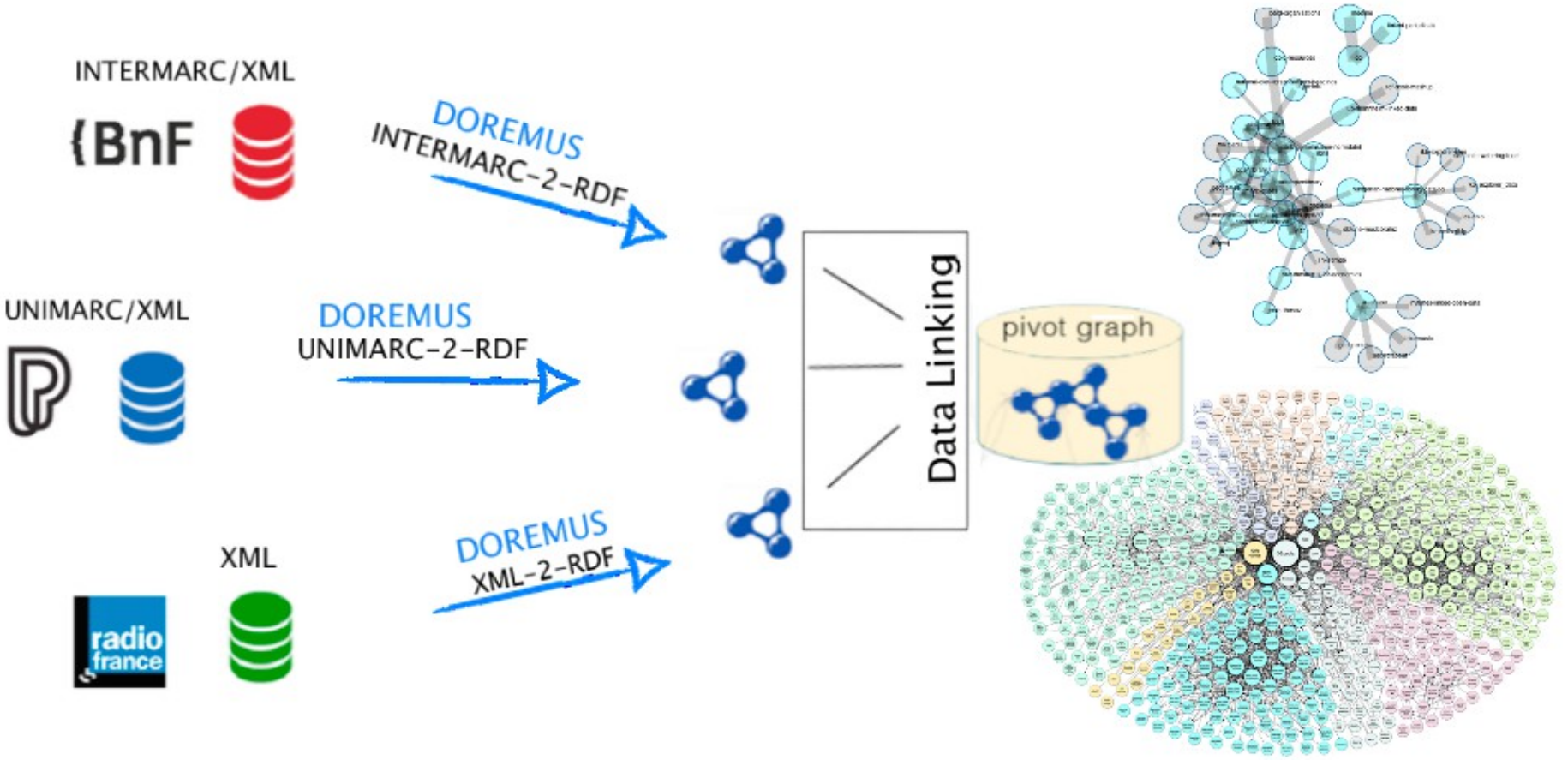
...and rightfully so.

„Wovon man nicht sprechen kann, darüber muss man schweigen.“

https://www.youtube.com/watch?v=57PWqFowq-4

Outline

4. Connecting to the web of data



The DOREMUS Playground

For those of you who would like to try all that out, check the DOREMUS Playground.
https://github.com/DOREMUS-ANR/doremus-playground

You will find a folder containing:

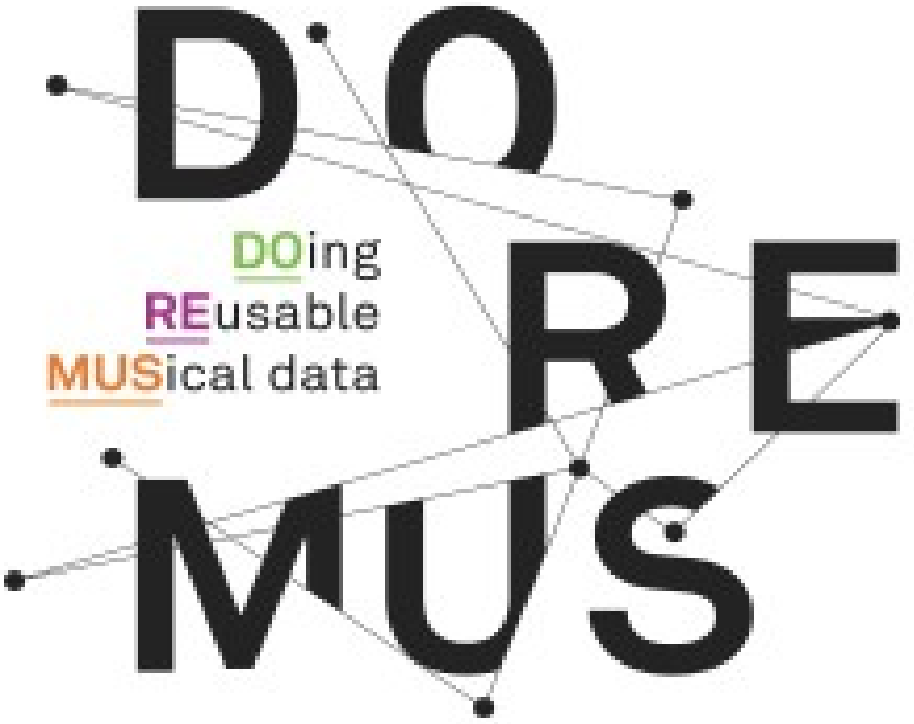
1) The dataset 1 (**DS1: nine heterogeneities**), composed of

- the original MARC data of the BnF and the PP
- the two datasets in RDF.
- the reference file, containing the correspondences between the works
- a correspondance between each pair of works and its heterogeneity type

2) Various SILK configuration files, each using different combinations of properties for the link discovery

3) A “readme” document, explaining the rules and the aim of the game and containing useful links.

Thanks for listening.



References and Links

[1] Ferrara, A., Nikolov, A., & Scharffe, F. (2013). Data linking for the semantic web. *Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications*, *169*.

[2] Achichi, M., Bailly, R., Cecconi, C., Destandau, M., Todorov, K., & Troncy, R. (2015). DOREMUS: Doing Reusable Musical Data. *ISWC2015 P&D track.*

[3] Volz, J., Bizer, C., Gaedke, M., & Kobilarov, G. (2009). Silk-A Link Discovery Framework for the Web of Data. *LDOW, 538*.

[4] Soru, T., Marx, E., & Ngonga Ngomo, A. C. (2015, May). ROCKER: A refinement operator for key discovery. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1025-1033). International World Wide Web Conferences Steering Committee.

[5] Symeonidou, D., Armant, V., Pernelle, N., & Saïs, F. (2014). SAKey: Scalable almost key discovery in rdf data. In *The Semantic Web–ISWC 2014* (pp. 33-49). Springer International Publishing.

[6] The DataLift project: http://datalift.org

[7] The DOREMUS github repository and playground: https://github.com/DOREMUS-ANR, https://github.com/DOREMUS-ANR/doremus-playground

[8] UNIMARC (authority records): http://www.ifla.org/publications/ifla-series-on-bibliographic-control-38

[9] UNIMARC (bibliographical records): http://www.ifla.org/publications/ifla-series-on-bibliographic-control-36

[10] INTERMARC: http://www.ifla.org/node/4858

[11] String2URI prototype: https://github.com/ThibWeb/stringtouri

[12] Instance matching track DOREMUS at OAEI: http://oaei.ontologymatching.org

[13] Destandau, M., Troncy, R., Todorov, K., Cecconi, C., Voisin, M., Canno; I., Leresche, F. (2016). Linked Data Approach for Structuring and Interlinking Musical Catalogs: How Three Major French Cultural Institutions Finally Came to an Agreement. *IFLA's satellite event : Data in Libraries: the big picture.*