renouvaud

Réseau vaudois
des bibliothèques

Derivaud

Importing poorly structured library metadata into a MARC21 union catalogue

Michael Hertig

ELAG 2022

June 9th 2022

canton de vaud
LIBERTÉ PATRIE

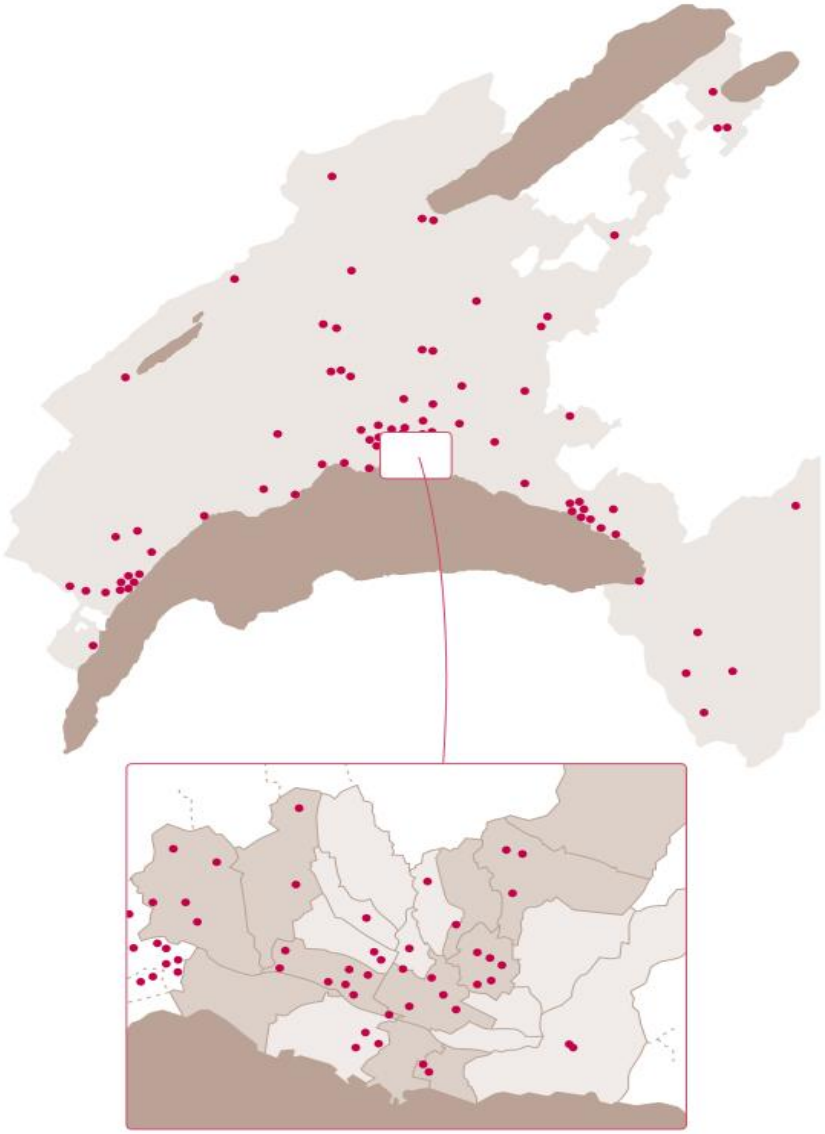
Context


rencontre
découverte
étudier
information
manuscrits
patrimoine
collections
recherche
anbisnu
sciences
document
lire
savoir
e-books
littérature
formation
films
archives
magazines
audiovisuel
digital
conservation
vaud
public
livres
expositions
culture
presse
numérique
connaissance
multimédia

The Renouvaud Library Network

- Canton de Vaud, Switzerland
- Capital city: Lausanne
- French speaking
- Bibliothèque Cantonale Universitaire Lausanne

The Renouvaud Library Network



- 140 libraries
  - Primary and high school, higher education, research, public, heritage, etc.
- 190’000 patrons
- Union catalogue
  - 4.2M physical resources
  - 1.4M electronic resources

Library integration

- Since 2016, integration of more than 30 libraries
  - Integration of 5 libraries in progress
  - 20 and more libraries to come

- Thousands of records
  - Collections from 1k to 60k titles

- Non MARC21 data
  - Bibliomaker (csv export file)
  - Oracle db
  - Filemaker db
  - MS Excel
  - Other…

Challenges

- Importing high quality metadata from poorly structured ones
  - MARC21 format
  - RDA Cataloging norms
- Automatizing record creation
- Custom mapping: not an option
- Avoiding duplication: using existing records in the Renouvaud catalogue

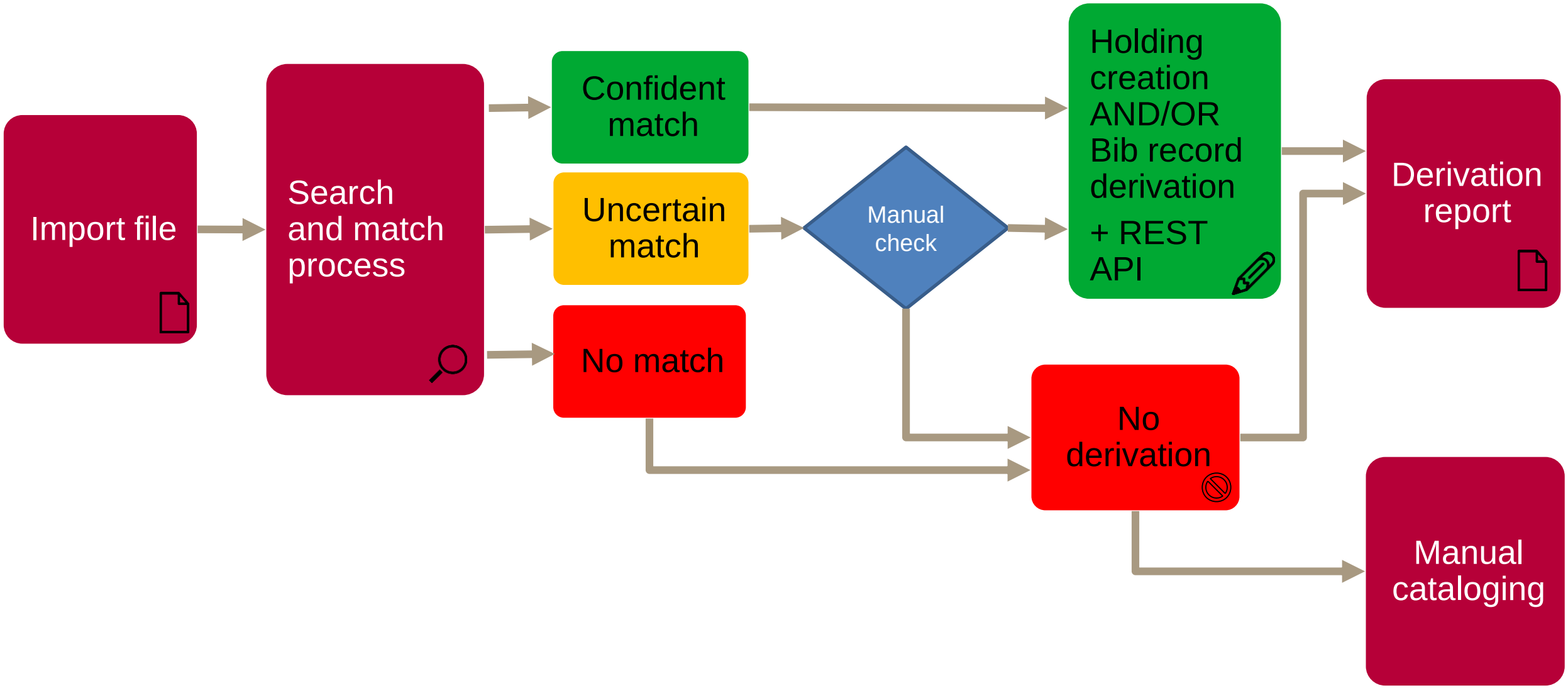
➔Attaching local holdings to records that exist in the catalogue
or
➔Record derivation: Retrieving records from existing remote repositories and importing them into the catalogue

Derivaud

- = “DERIvation” + “Canton de VAUD”
- Web application
  - Since 2019
- Collaboration with Datuman SàRL
- Programming language: Scala
- Integration with Ex Libris Alma
  - REST API

The Derivaud Workflow

The derivation process



Required data

- Bibliographic record matching
  - ISBN
  - Title
  - Author
  - Publishing year
  - Publisher
  - Document type (MARC21 LDR position 6 taxonomy)

- Creation of inventory
  - Call number
  - Location
  - Barcode
  - Loan policy

Document and format types

- Document types managed by the system
  - Monographs
  - Multivolume works
  - Sound recordings
  - Video recordings
  - Musical scores (bad results)

- Formats
  - csv (tsv)
  - MARC21
  - Unimarc Converter (Bibliothèque nationale de France)

User interface: Manual verification view

renouvaud **[Arbitrage forcé] Notices à date exacte** Rejeter

Rendez-vous avec le mal : une enquête de Samson et Delilah, les détectives du Yorkshire / Chapman, Julia	**Paris : R. Laffont, 2018**		**9782221218440**	
Rendez-vous avec le mal : une enquête de Samson et Delilah, les détectives du Yorkshire / Julia Chapman ; trad. de l'anglais par Dominique Haas et Stéphanie Leigniel	Paris : R. Laffont, 2018	391 p.	9782221218440	100 %
Rendez-vous avec le mystère : une enquête de Samson et Delilah, les détectives du Yorkshire / Julia Chapman ; trad. de l'anglais par Dominique Haas et Viviane Mikhalkov	Paris : R. Laffont, 2018	451 p.	9782221221044	82 %
Rendez-vous avec le crime : une enquête de Samson et Delilah, les détectives du Yorkshire / Julia Chapman ; trad. de l'anglais par Dominique Haas	Paris : Robert Laffont, [2018]	383 pages	9782221215494	78 %

renouvaud **[Arbitrage forcé] Notices à date inexacte** Rejeter

Voir la coupette à moitié pleine / Mathou	**Paris : Editions Delcourt, 2019**		**9782413022701**	
Et puis Colette / Sophie Henrionnet ; Mathou	[Paris] : Delcourt, 2018	201 p. : ill.	9782413011279	66 %
À volonté : Tu t'es vue quand tu manges? / Mademoiselle Caroline et Mathou	[Paris] : Delcourt, 2020	127 pages : illustrations	9782413024330	66 %
A volonté : tu t'es vue quand tu manges ? / Mademoiselle Caroline & Mathou	[Paris] : Delcourt, 2020	117 pages : illustrations	9782413024330	66 %
Peurs bleues / Mathou	[Paris] : Delcourt, [2020]	121 pages : illustrations	9782413022756	65 %

User interface: candidate record view

Comparaison automatique: Candidat fort - Score moyen: 0.90

Critère	Source	Cible	Résultat
ISBN	9782205001464	9782205001464	☺
Titre	Une aventure d'Astérix || Le tour de Gaule	Le tour de Gaule d'Astérix	⊗ 0.42
Auteur	GOSCINNY, René; UDERZO, Albert	Goscinny, René; Uderzo, Albert	☺ 1.00
Éditeur	Paris... [et. al.] : Dargaud,	Paris : Dargaud, 1989	☺ 1.00
Date	1989	1989	☺ 1.00
Type de document	BOOK	BOOK; ARTICLE; NEWSPAPER; SCORE; MAP	☺ 1.00

Exemplaire source

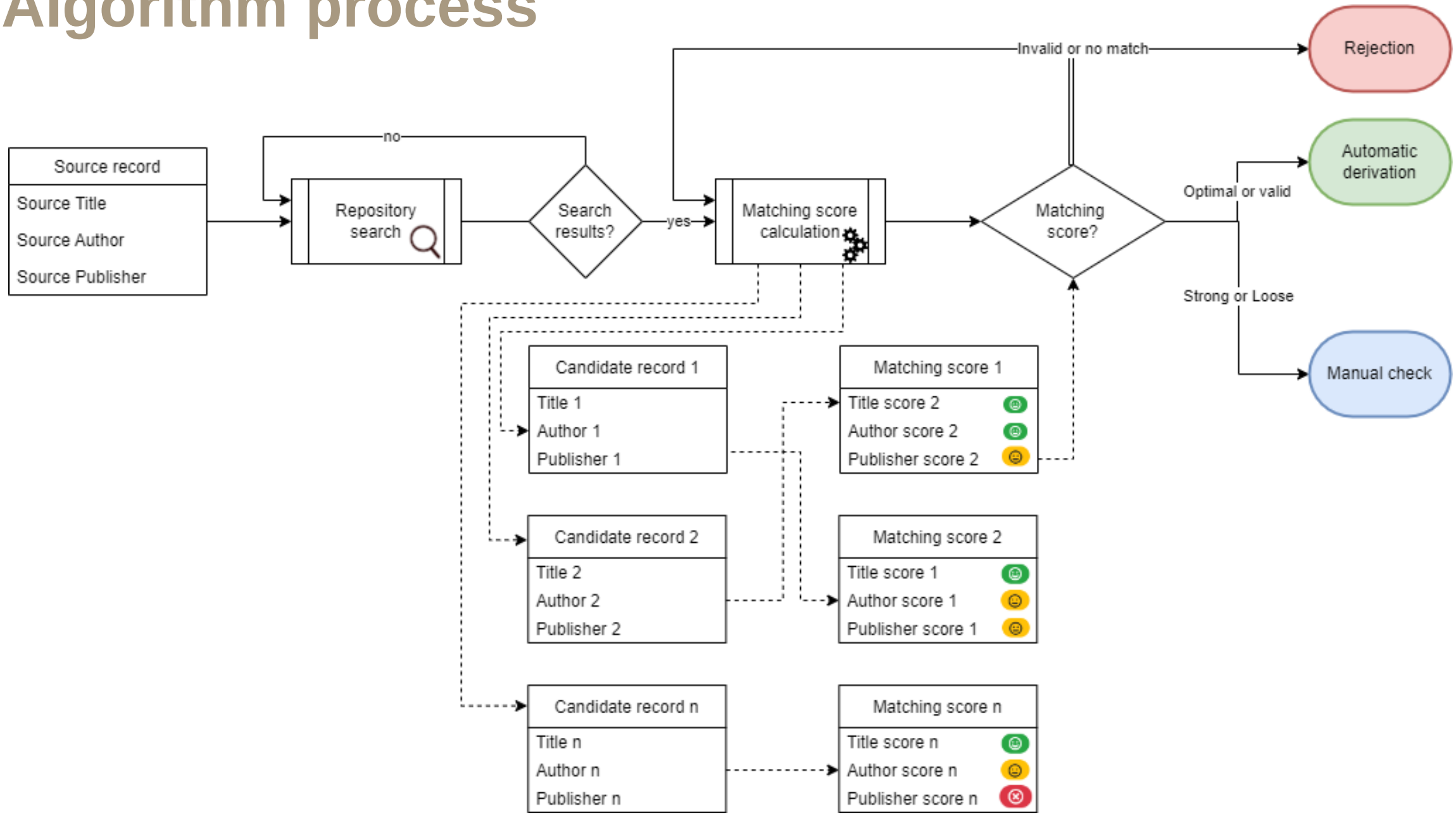
Numero	20767
force_user_validation	Le tour de Gaule
Titre_mention_responsabilite	Une aventure d'Astérix || Le tour de Gaule / texte de Goscinny ; dessins Uderzo
Edition	
Lieu_editeur_date	Paris... [et. al.] : Dargaud,
Collation	48 p.
Collection	
ISBN	9782205001464
Annee_publication	1989
Vedette_principale	GOSCINNY, René; UDERZO, Albert

Notice cible

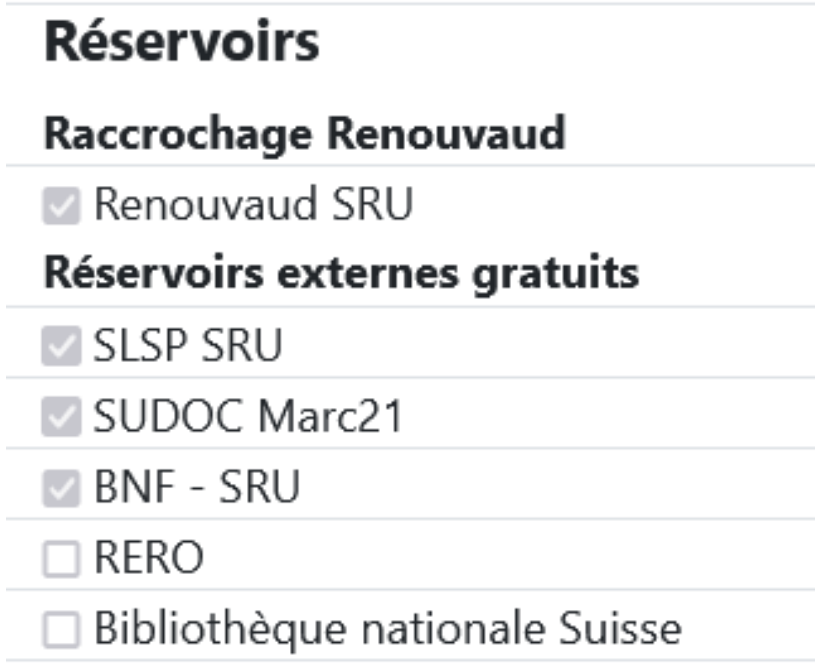
LDR		00883nam a2200289 a 4500
020		$$a 9782205001464
035		$$a (RERO)007194386 $$9 ExL
035		$$a R007194386
035		$$a (RNV_B)0002048557
039	9	$$a 201605161322 $$b VLOAD $$y 201211290930 $$z 8672
040		$$a RERO vdoron
072	7	$$a s1zz $$2 rero
100	1	$$a Goscinny, René $$4 aut
245	13	$$a Le tour de Gaule d'Astérix / $$c texte de Goscinny ; dessins de Uderzo

The algorithm

Algorithm process



Searching repositories



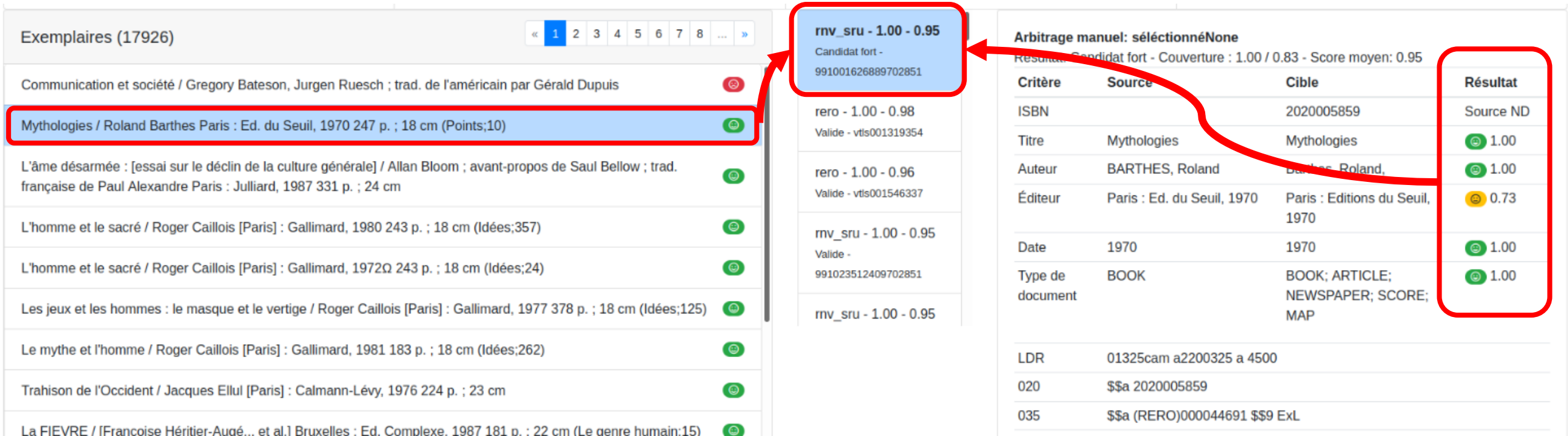
- SRU or Z39.50 Protocols
- Renouvaud union catalogue searched first
- Choice of repositories
- Ranking of *free* repositories
- Fee-based repositories searched last

- For each repository, the search process stops when:
  - There are enough candidate records (configurable minimum number of results)
  - All combinations have been tried.

Matching process

- 2 steps
  1. Comparing individual data elements and assigning matching scores
     - ISBN, Title, Publishing year, Etc.

  2. Combining the individual matching scores and calculating a global matching score



Global matching scores and output

- Depending on the global matching score, Derivaud will perform the following action:

Optimal	The record is imported automatically with no further search
Valid	The record can be imported automatically
Strong	The record must be manually checked for a match.
Loose	The record must be manually checked for a match IF there is no other strong match
Invalid	No corresponding record found
Rejected	The candidate record is not fit for comparison.

Calculating global matching scores

21 rules applied in a given order

Rejected record if:

1. No Title
2. ISBN AND (Publisher OR Publishing date) are missing
3. …

Optimal match if:

6. …

Valid match if:

7. All available criteria have a strong match
8. Valid ISBN AND all available criteria have a close match
9. …

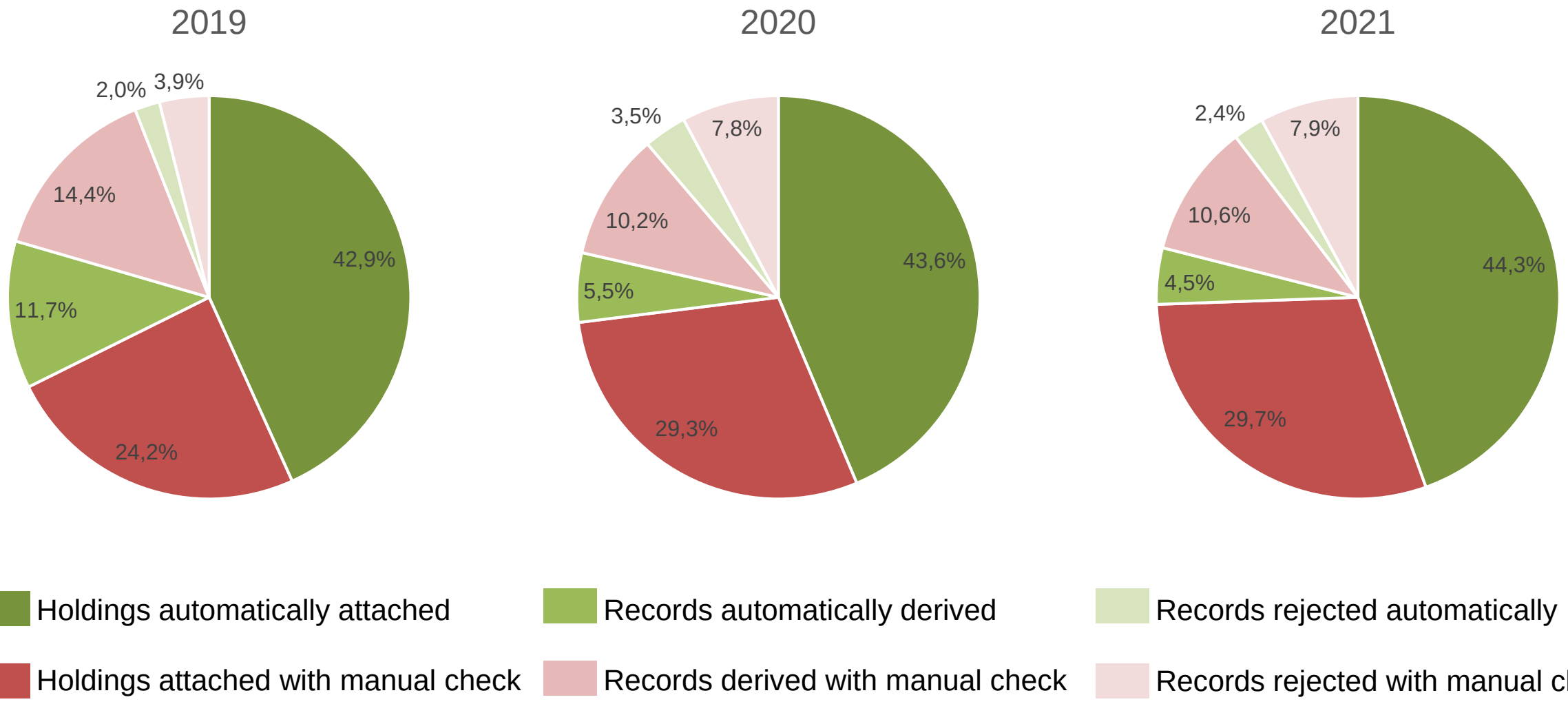
Strong match if:

10. Same ISBN
11. Nearly perfect match of title AND close match of author

…

Feedback and perspectives

Facts and figures: automation vs. manual work



Perspectives

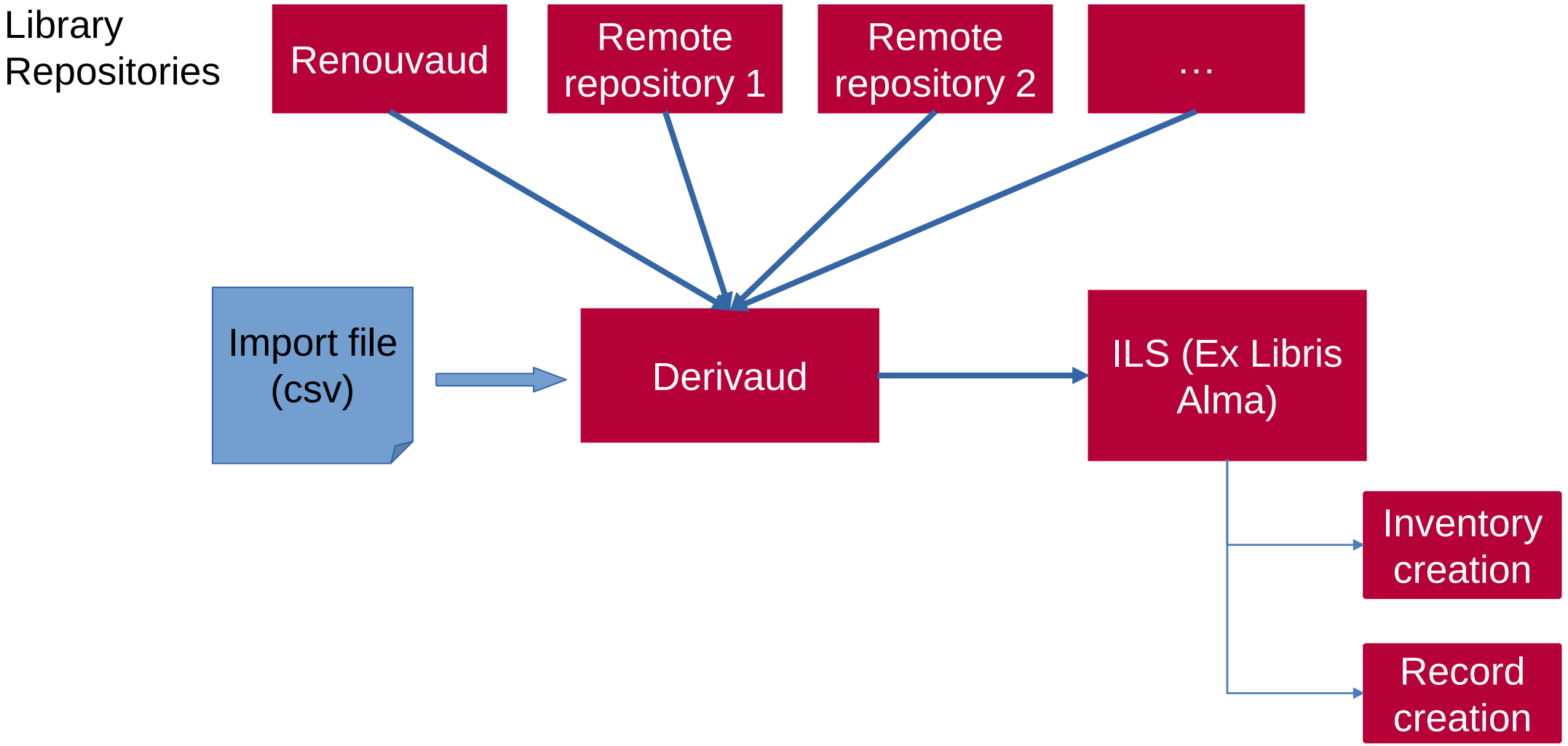
- Continuous improvements
- Name and Subject heading enrichment
- Machine learning
  - Instead of manual checking
  - Use of manual checking data as a training dataset

Thanks for your attention!

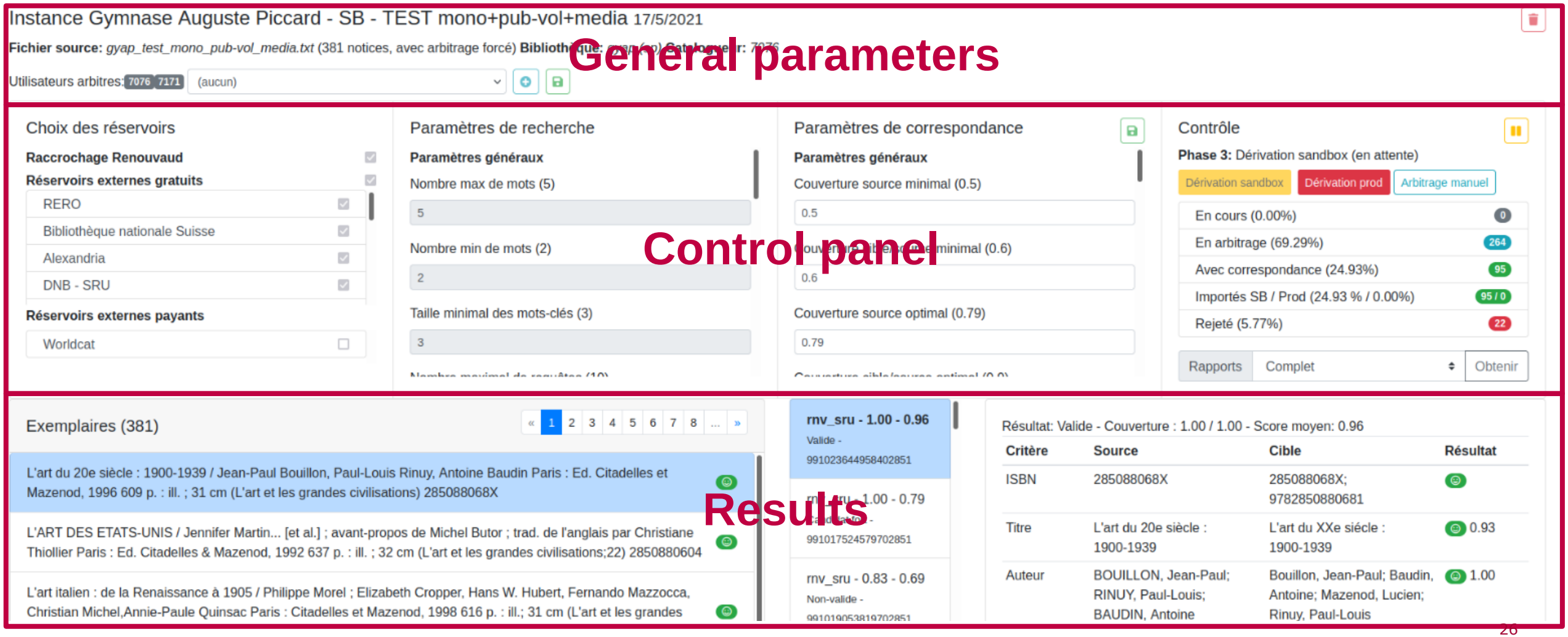
Questions?

Contact information
Renouvaud: michael.hertig@bcu.unil.ch
Datuman: info@datuman.ch

The overall Derivaud workflow

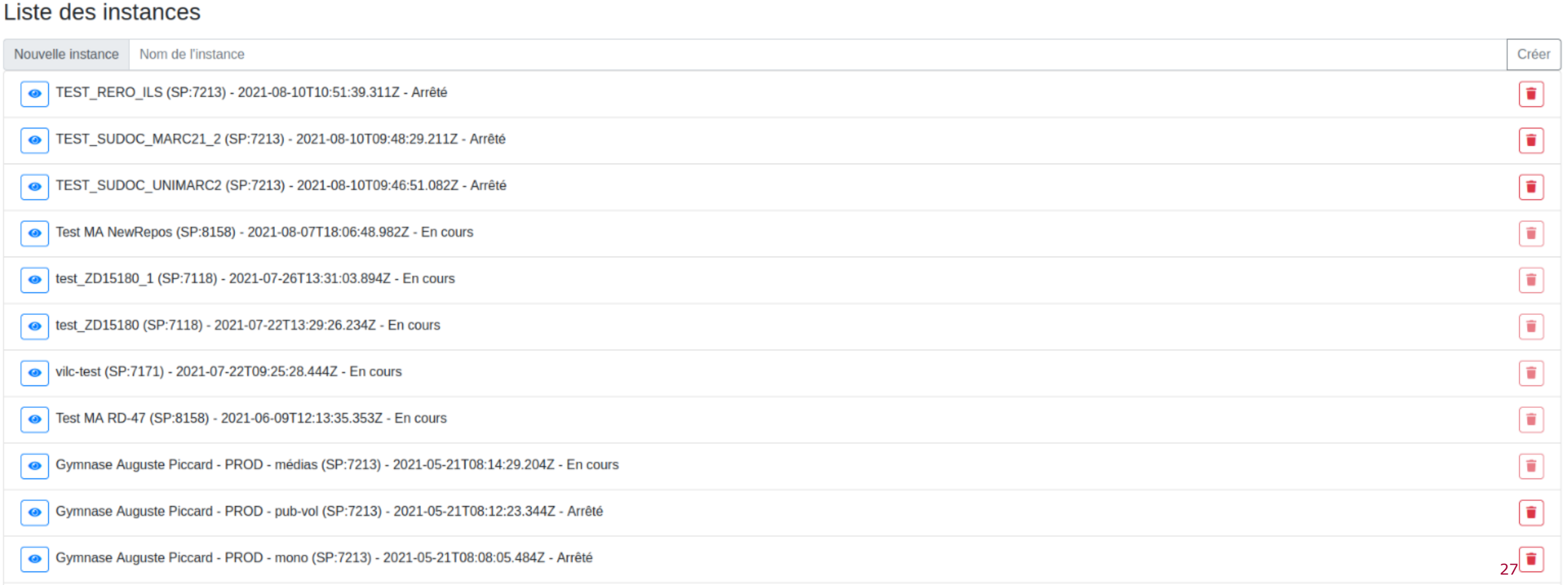


Configuration UI



User interface

- 1 import file → 1 «instance»
  - Depends on library, document type, etc.



Searching repositories: Finding candidate records

- One search process in a repository runs various distinct requests.
  - Uses SRU or Z39.50 parameters
  - For instance:
    - Full title AND ISBN
    - Main title AND Author
    - Parts of the title

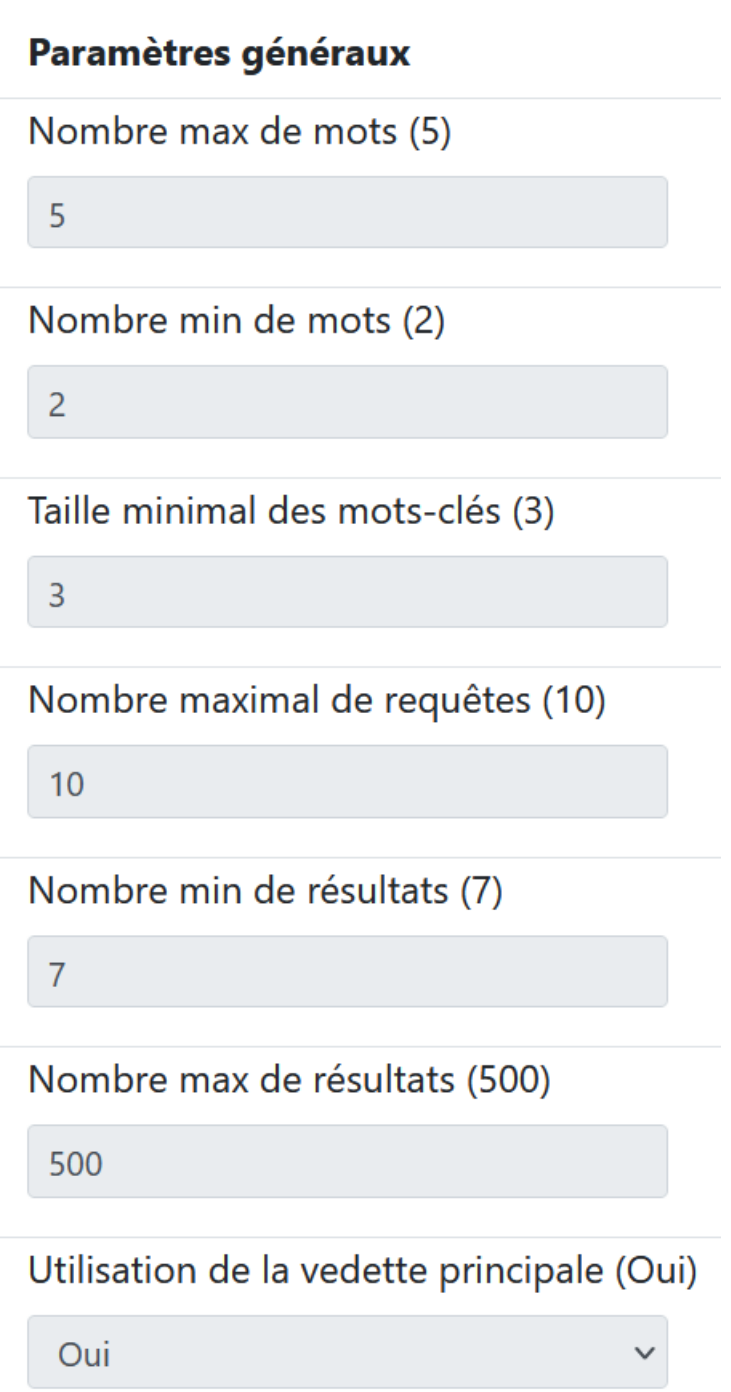
- If one request did not yield enough candidates, a new request is performed using other criteria

- For each repository, the search process stops when:
  - There are enough candidate records (configurable minimum number of results)
  - All combinations have been tried.
  - An optimal match has been found

Search parameters

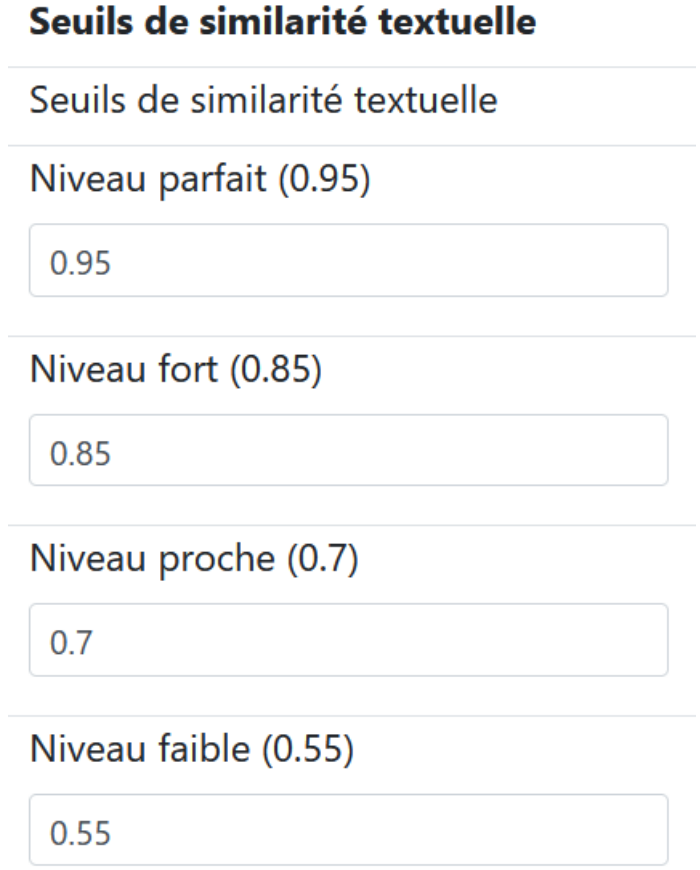
- Maximum and minimum number of words
- Minimum length of keywords (excludes stop words)
- Maximum number of requests
- Minimum and maximum number of results (up to 6000)

- Parameters are here for balancing the depth of the search with the length of the search process.



Individual matching score calculation

- Every data element gets a matching score.
- Matching values are:
  - Perfect match
  - Strong match
  - Close match
  - Loose match
  - No match
- Configurable



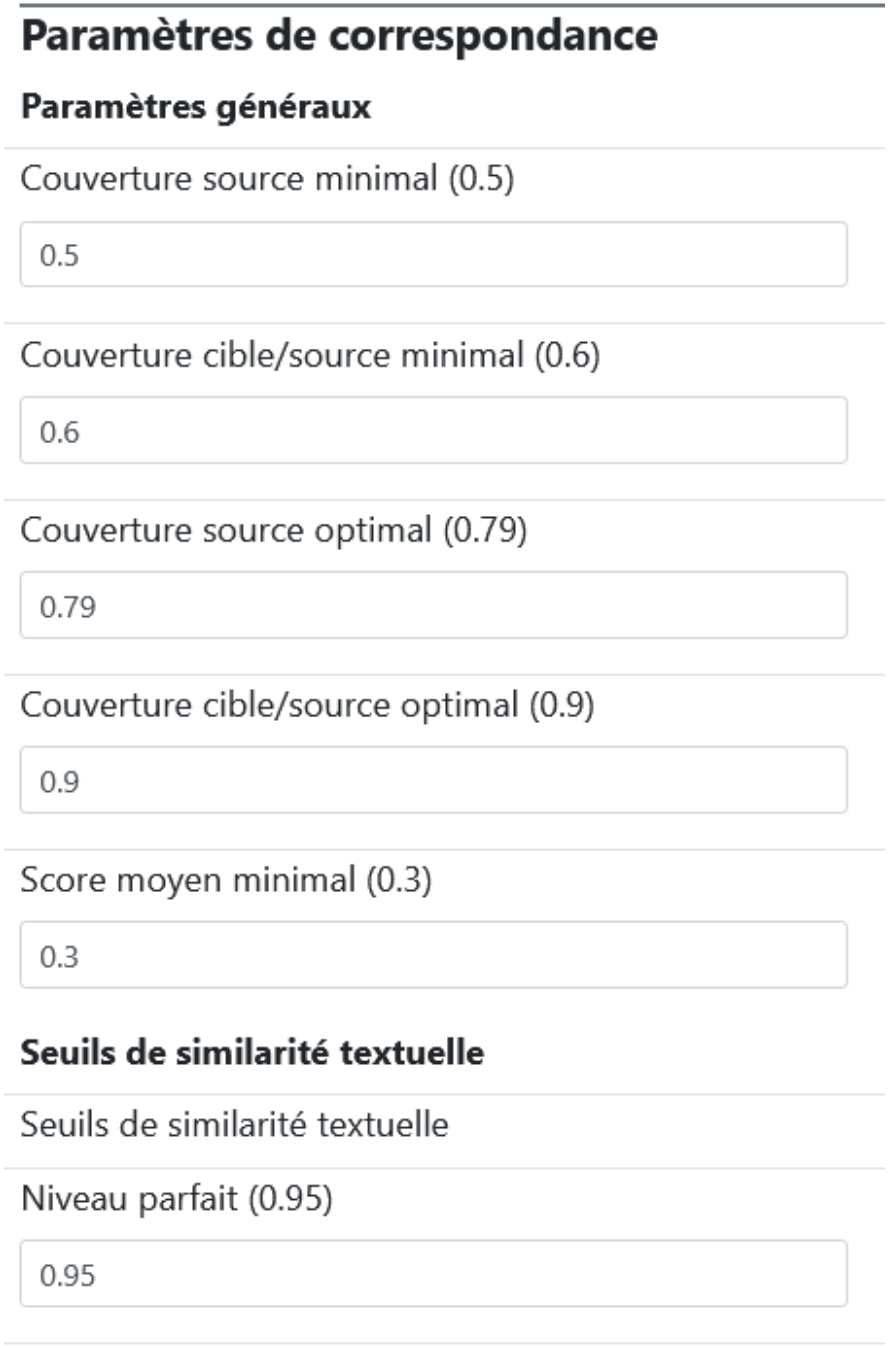
Individual matching score calculation

- Extracting and pre-processing values
- Calculating matching score with specific method

	Target fields	Pre-processing	Comparison method
ISBN	020$a, 024$a, 028$a, 020$z	Delete () with the data inbetween ISBN13 Normalization	Is equal
Titel	245	Delete [] with the data inbetween, when before / ASCII Normalization	Fuzzy string matching (Levenshtein distance)

Match parameters

- minimum/optimal source coverage
- minimum/optimal target coverage
- minimum average score
- Specific parameters
  - Whole string or string parts of Title
  - Ratio of valid authors
  - Maximal Publishing date difference
  - Etc.



Facts and figures

Records treated	2019	2020	2021
Total	88758	160411	91700
Attached holdings (raccrochage)	67,1%	72,9%	74,0%
Derived records	26,1%	15,7%	15,1%
Rejected records	5,9%	11,2%	10,3%
Records errors	0,9%	0,2%	0,6%

Facts and figures

Repository use	2019	2020	2021	2022
RERO	11573	12098	5267	1827
Worldcat	6245	8930	**1208**	**0**
British Library	2160	1477	1080	1468
DNB	990	1887	736	517
LoC (SRU)	659	562	832	1345
BnF (UNIMARC)	NA	NA	**2891**	**0**
SLSP	NA	NA	**1478**	**1114**
SUDOC (MARC21)	NA	NA	**247**	**2074**

Facts and figures (Fall 2021)

	% rejected records	% derived records	% records manually checked
Primary and Secundary school libraries	5 %	95 %	37 %
High school libraries	10 %	90 %	40 %
Heritage and special libraries	13 %	87 %	60 %
Average speed of derivation process			

Quelques améliorations prévues

- Intégration du réservoir Sudoc (UNIMARC)
- Améliorer conversion INTERMARC pour les CD et les DVD
- Paramètres de recherche et de correspondance réglables par groupes de réservoirs
- Amélioration de la recherche pour ne pas considérer les non-candidats comme des résultats