

# Renouvaud

Réseau vaudois  
des bibliothèques

## Derivaud

Importing poorly structured library  
metadata into a MARC21 union  
catalogue

Michael Hertig  
ELAG 2022  
June 9th 2022

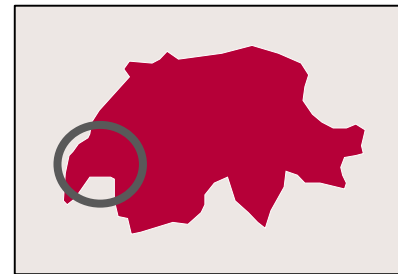


# Context

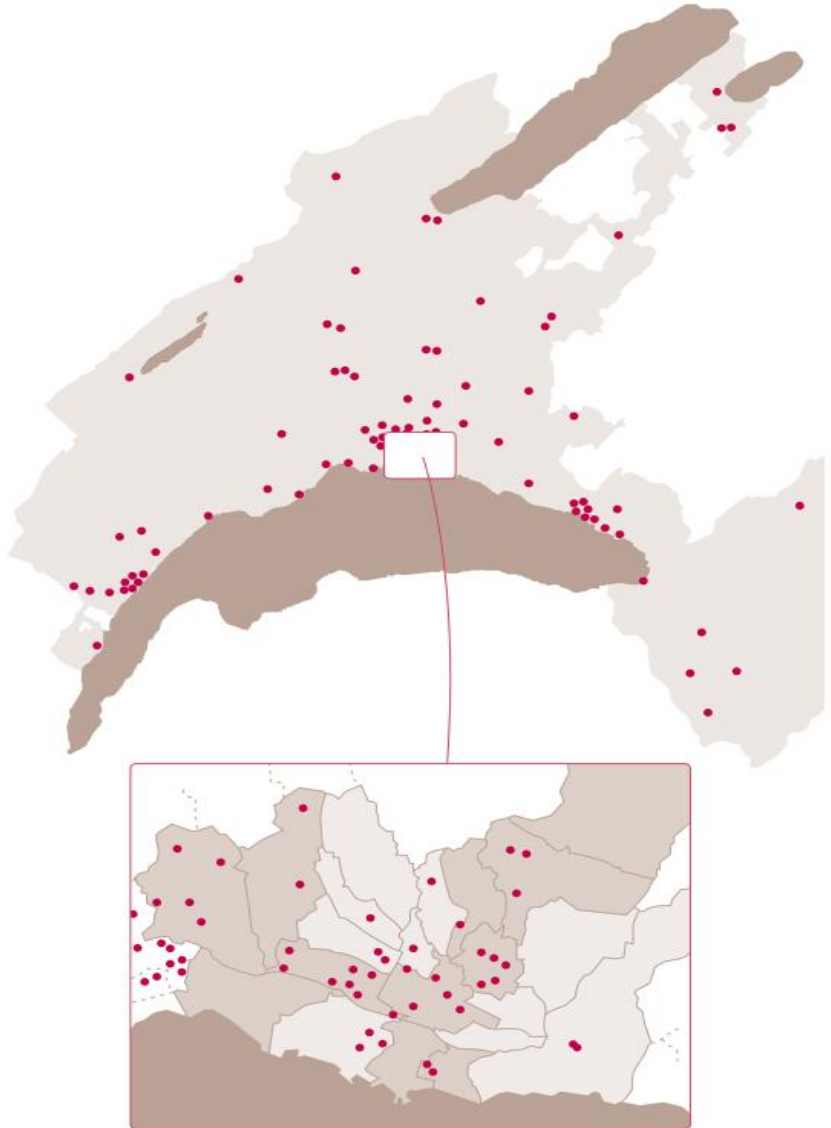


# The Renouvaud Library Network

- 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
  - Automating 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

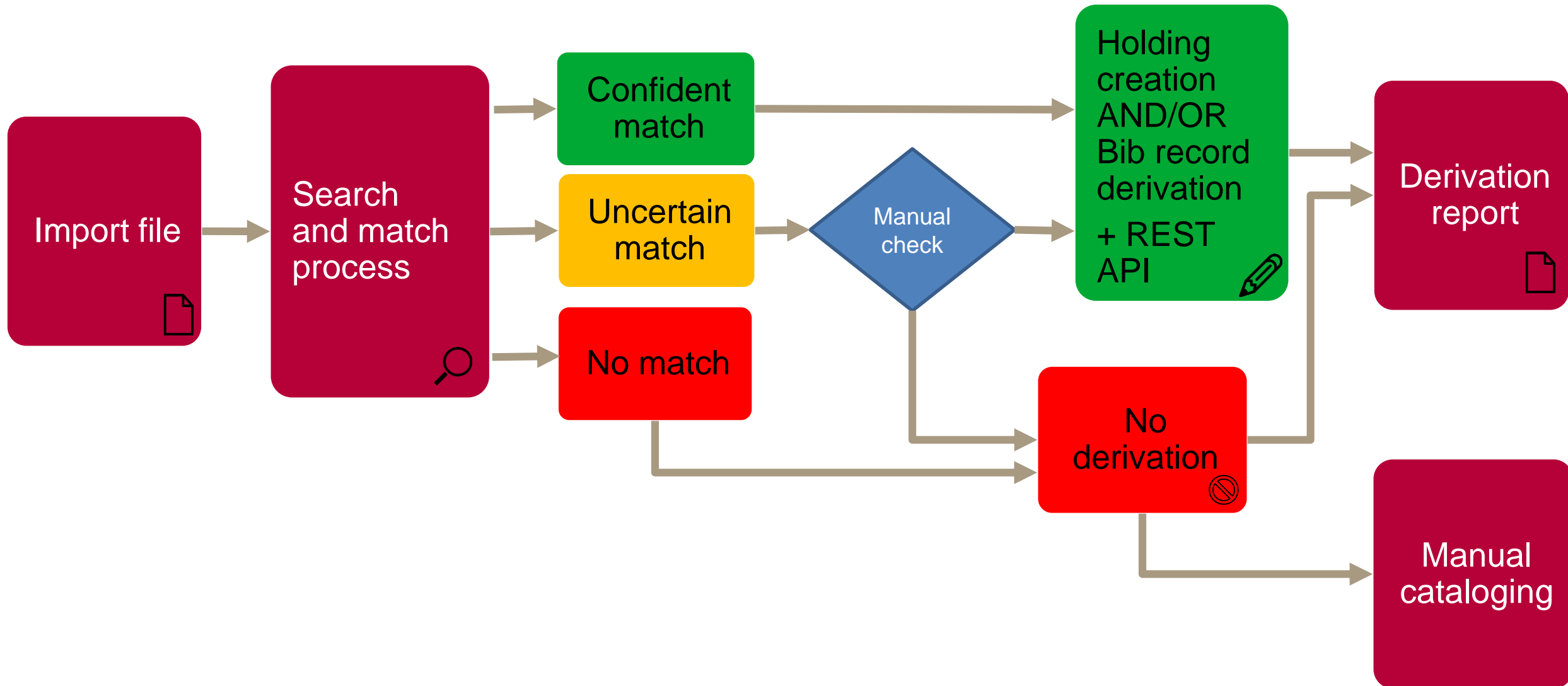
- = “DERIvation” + “Canton de VAUD”
- Web application
  - Since 2019
- Collaboration with Datuman SàRL
- Programming language: Scala
- Integration with Ex Libris Alma
  - REST API







# The 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 ▼	
<b>Rendez-vous avec le mal : une enquête de Samson et Delilah, les détectives du Yorkshire / Chapman, Julia</b>	<b>Paris : R. Laffont, 2018</b>		<b>9782221218440</b>		
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 ▼	
<b>Voir la coupette à moitié pleine / Mathou</b>	<b>Paris : Editions Delcourt, 2019</b>		<b>9782413022701</b>		
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

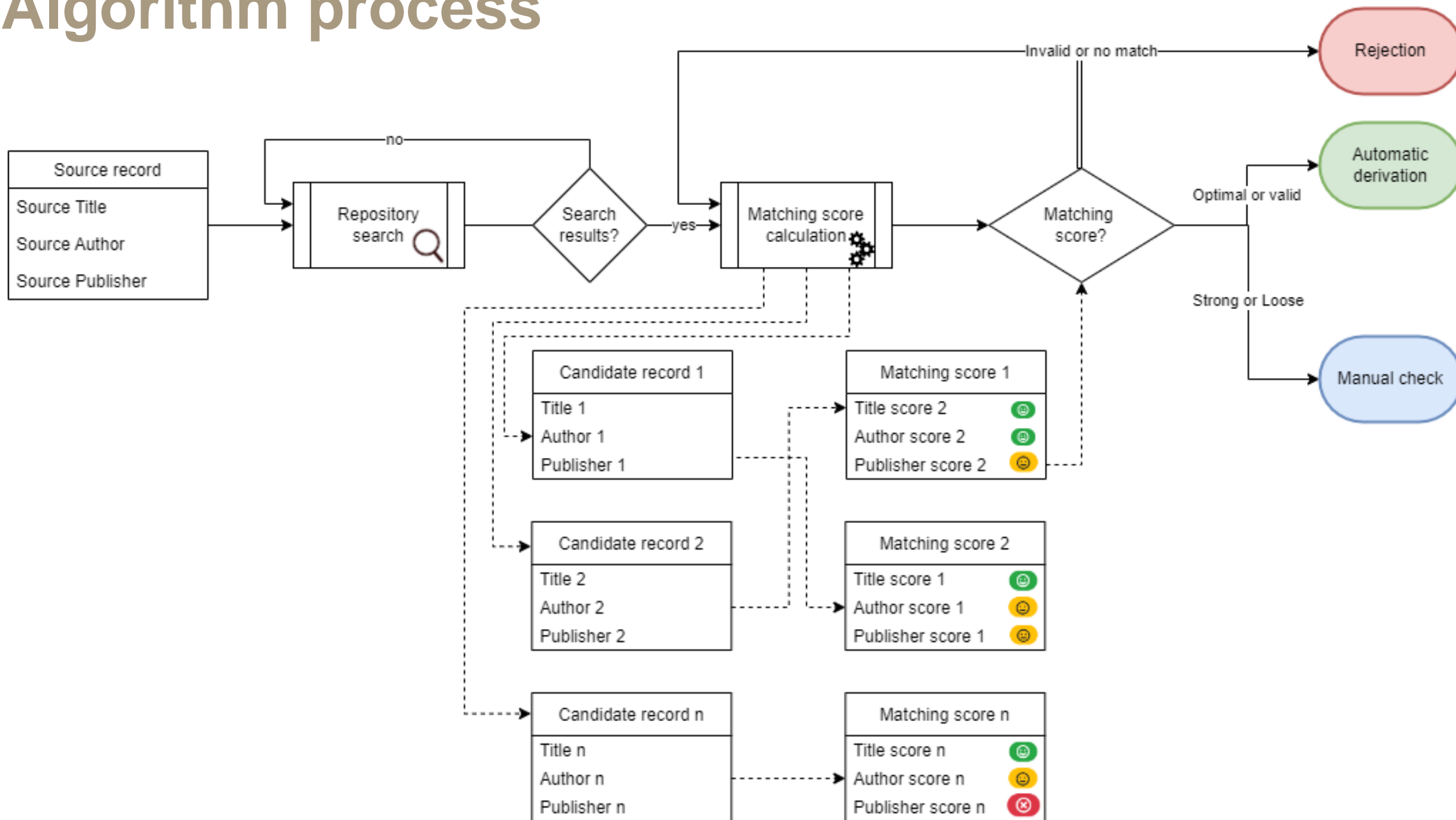
  

Exemplaire source		Notice cible	
Numero	20767	LDR	00883nam a2200289 a 4500
force_user_validation	Le tour de Gaule	020	\$\$a 9782205001464
Titre_mention_responsabilite	Une aventure d'Astérix    Le tour de Gaule / texte de Goscinny ; dessins Uderzo	035	\$\$a (RERO)007194386 \$\$9 ExL
Edition		035	\$\$a R007194386
Lieu_editeur_date	Paris... [et. al.] : Dargaud,	035	\$\$a (RNV_B)0002048557
Collation	48 p.	039 9	\$\$a 201605161322 \$\$b VLOAD \$\$y 201211290930 \$\$z 8672
Collection		040	\$\$a RERO vdonon
ISBN	9782205001464	072 7	\$\$a s1zz \$\$2 rero
Annee_publication	1989	100 1	\$\$a Goscinny, René \$\$4 aut
Vedette_principale	GOSCINNY, René; UDERZO, Albert	245 13	\$\$a Le tour de Gaule d'Astérix / \$\$c texte de Goscinny ; dessins de Uderzo

# The algorithm



# Algorithm process



# Searching repositories

---

## Réservoirs

### Raccrochage Renouvaud

Renouvaud SRU

### Réservoirs externes gratuits

SLSP SRU

SUDOC Marc21

BNF - SRU

RERO

Bibliothèque nationale Suisse

---

- 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

Exemplaires (17926) « 1 2 3 4 5 6 7 8 ... »

Communication et société / Gregory Bateson, Jurgen Ruesch ; trad. de l'américain par Gérald Dupuis

**Mythologies / Roland Barthes Paris : Ed. du Seuil, 1970 247 p. ; 18 cm (Points;10)**

L'âme désarmée : [essai sur le déclin de la culture générale] / Allan Bloom ; avant-propos de Saul Bellow ; trad. française de Paul Alexandre Paris : Julliard, 1987 331 p. ; 24 cm

L'homme et le sacré / Roger Caillois [Paris] : Gallimard, 1980 243 p. ; 18 cm (Idées;357)

L'homme et le sacré / Roger Caillois [Paris] : Gallimard, 1972Ω 243 p. ; 18 cm (Idées;24)

Les jeux et les hommes : le masque et le vertige / Roger Caillois [Paris] : Gallimard, 1977 378 p. ; 18 cm (Idées;125)

Le mythe et l'homme / Roger Caillois [Paris] : Gallimard, 1981 183 p. ; 18 cm (Idées;262)

Trahison de l'Occident / Jacques Ellul [Paris] : Calmann-Lévy, 1976 224 p. ; 23 cm

La FIEVRE / [Francoise Héritier-Augué... et al.] Bruxelles : Ed. Complexe, 1987 181 p. ; 22 cm (Le genre humain;15)

**rnv\_sru - 1.00 - 0.95**  
Candidat fort -  
991001626889702851

rero - 1.00 - 0.98  
Valide - vtis001319354

rero - 1.00 - 0.96  
Valide - vtis001546337

rnv\_sru - 1.00 - 0.95  
Valide -  
991023512409702851

rnv\_sru - 1.00 - 0.95

**Arbitrage manuel: sélectionnéNone**  
Résultat: Candidat fort - Couverture : 1.00 / 0.83 - Score moyen: 0.95

Critère	Source	Cible	Résultat
ISBN		2020005859	Source ND
Titre	Mythologies	Mythologies	🟢 1.00
Auteur	BARTHES, Roland	Barthes, Roland,	🟢 1.00
Éditeur	Paris : Ed. du Seuil, 1970	Paris : Editions du Seuil, 1970	🟡 0.73
Date	1970	1970	🟢 1.00
Type de document	BOOK	BOOK; ARTICLE; NEWSPAPER; SCORE; MAP	🟢 1.00

LDR 01325cam a2200325 a 4500

020 \$\$a 2020005859

035 \$\$a (RERO)000044691 \$\$9 ExL

# Global matching scores and output

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

<b>Optimal</b>	The record is imported automatically with no further search
<b>Valid</b>	The record can be imported automatically
<b>Strong</b>	The record must be manually checked for a match.
<b>Loose</b>	The record must be manually checked for a match IF there is no other strong match
<b>Invalid</b>	No corresponding record found
<b>Rejected</b>	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:**

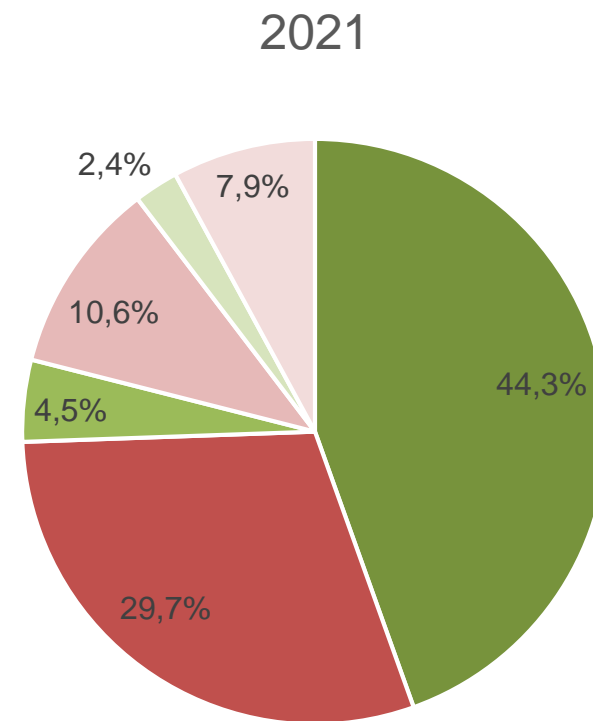
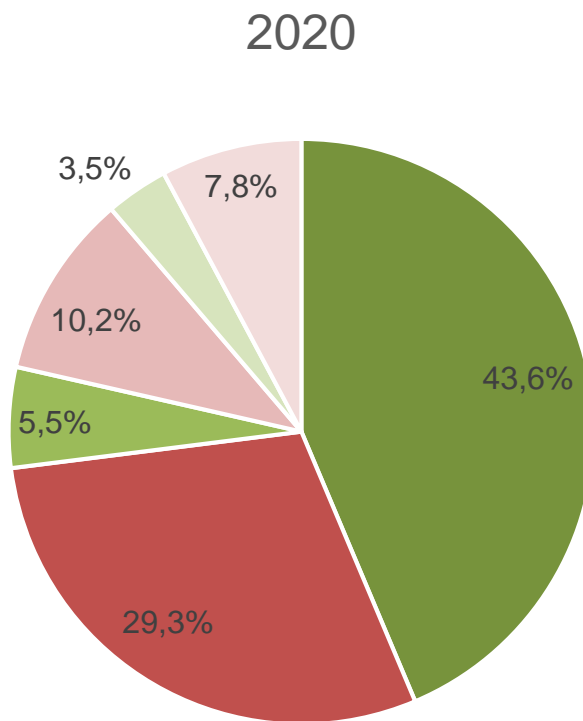
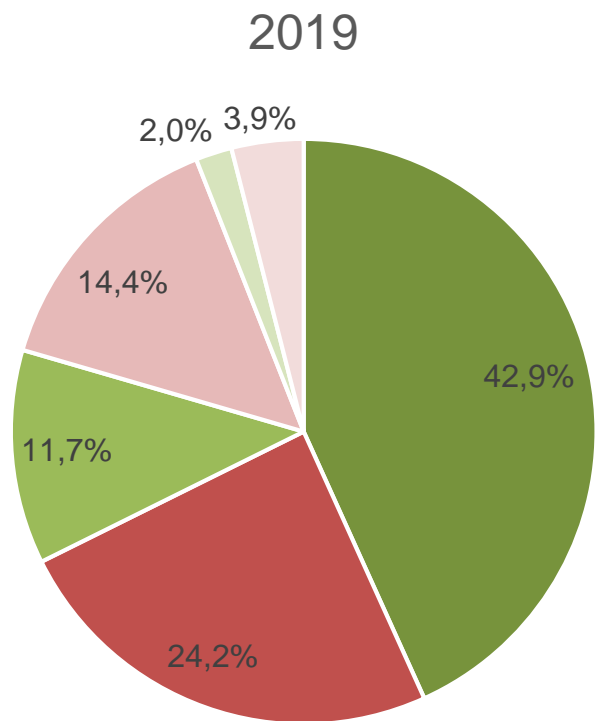
10. Same ISBN
11. Nearly perfect match of title AND close match of author

...

# Feedback and perspectives



# Facts and figures: automation vs. manual work



 Holdings automatically attached

 Records automatically derived

 Records rejected automatically

 Holdings attached with manual check

 Records derived with manual check

 Records rejected with manual check

# 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](mailto:michael.hertig@bcu.unil.ch)

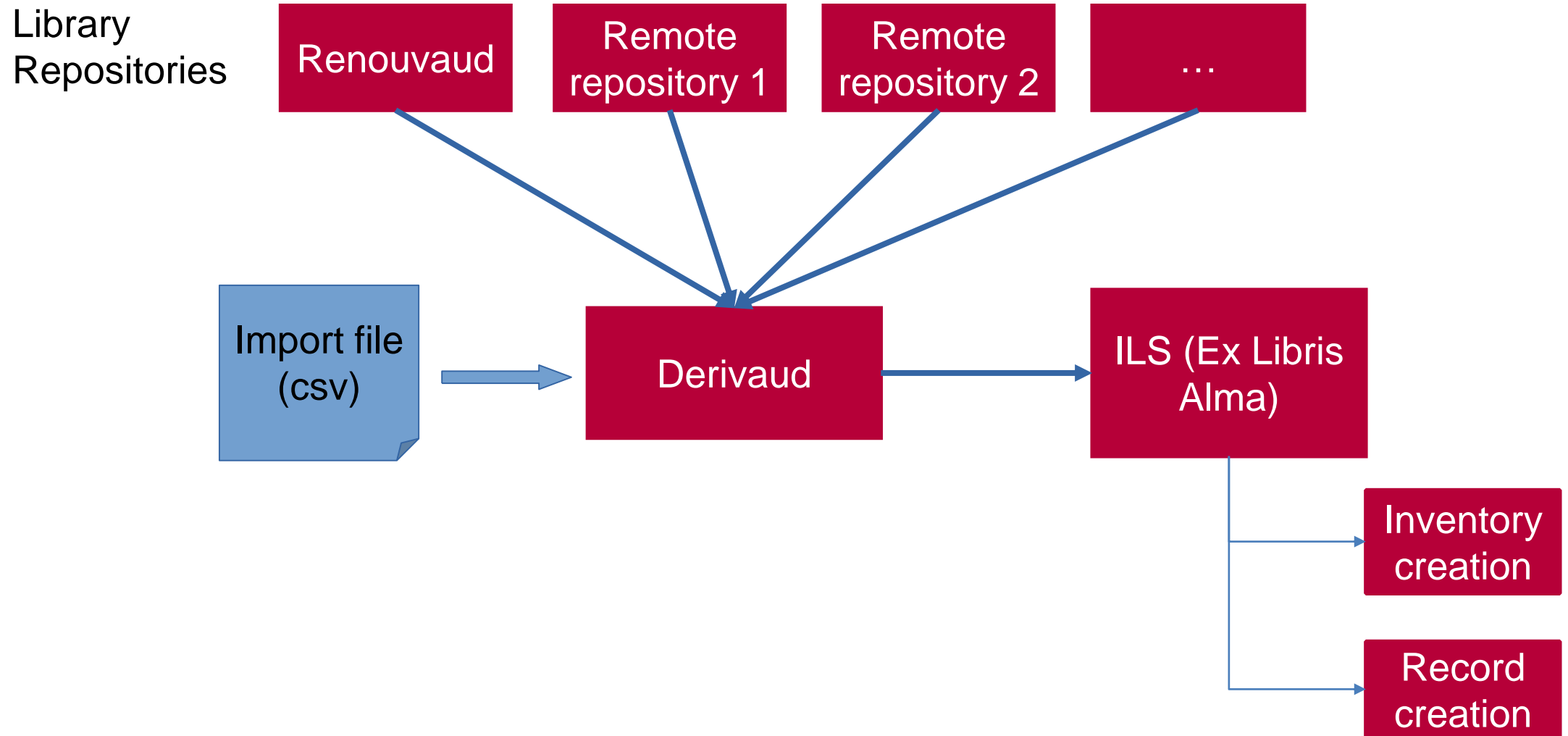
Datuman: [info@datuman.ch](mailto:info@datuman.ch)







# The overall Derivaud workflow



# Configuration UI

Instance Gymnase Auguste Piccard - SB - TEST mono+pub-vol+media 17/5/2021

Fichier source: *gyap\_test\_mono\_pub-vol\_media.txt* (381 notices, avec arbitrage forcé) Bibliothèque: *gyap (p)* Stat. logeur: *r: 7076*

Utilisateurs arbitres: **7076 7171** (aucun)

## General parameters

### Choix des réservoirs

Raccrochage Renouvaud

Réservoirs externes gratuits

RERO

Bibliothèque nationale Suisse

Alexandria

DNB - SRU

Réservoirs externes payants

Worldcat

### Paramètres de recherche

Paramètres généraux

Nombre max de mots (5)

5

Nombre min de mots (2)

2

Taille minimal des mots-clés (3)

3

Nombre maximal de requêtes (10)

### Paramètres de correspondance

Paramètres généraux

Couverture source minimal (0.5)

0.5

Couverture source cible minimal (0.6)

0.6

Couverture source optimal (0.79)

0.79

Couverture cible/source optimal (0.0)

### Contrôle

Phase 3: Dérivation sandbox (en attente)

Dérivation sandbox Dérivation prod Arbitrage manuel

En cours (0.00%) 0

En arbitrage (69.29%) 264

Avec correspondance (24.93%) 95

Importés SB / Prod (24.93 % / 0.00%) 95 / 0

Rejeté (5.77%) 22

Rapports Complet Obtenir

## Control panel

### Exemplaires (381)

« 1 2 3 4 5 6 7 8 ... »

L'art du 20e siècle : 1900-1939 / Jean-Paul Bouillon, Paul-Louis Rinuy, Antoine Baudin Paris : Ed. Citadelles et Mazenod, 1996 609 p. : ill. ; 31 cm (L'art et les grandes civilisations) 285088068X

L'ART DES ETATS-UNIS / Jennifer Martin... [et al.] ; avant-propos de Michel Butor ; trad. de l'anglais par Christiane Thiollier Paris : Ed. Citadelles & Mazenod, 1992 637 p. : ill. ; 32 cm (L'art et les grandes civilisations;22) 2850880604

L'art italien : de la Renaissance à 1905 / Philippe Morel ; Elizabeth Cropper, Hans W. Hubert, Fernando Mazzocca, Christian Michel, Annie-Paule Quinsac Paris : Citadelles et Mazenod, 1998 616 p. : ill. ; 31 cm (L'art et les grandes

rnv\_sru - 1.00 - 0.96

Valide -  
991023644958402851

rnv\_sru - 1.00 - 0.79

Non-valide -  
991017524579702851

rnv\_sru - 0.83 - 0.69

Non-valide -  
991019053819702851

Résultat: Valide - Couverture : 1.00 / 1.00 - Score moyen: 0.96



















Critère	Source	Cible	Résultat
ISBN	285088068X	285088068X; 9782850880681	
Titre	L'art du 20e siècle : 1900-1939	L'art du XXe siècle : 1900-1939	0.93
Auteur	BOUILLON, Jean-Paul; RINUY, Paul-Louis; BAUDIN, Antoine	Bouillon, Jean-Paul; Baudin, Antoine; Mazenod, Lucien; Rinuy, Paul-Louis	1.00

## Results

# User interface

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

## Liste des instances

Nouvelle instance	Nom de l'instance	Créer
	TEST_RERO_ILS (SP:7213) - 2021-08-10T10:51:39.311Z - Arrêté	
	TEST_SUDOC_MARC21_2 (SP:7213) - 2021-08-10T09:48:29.211Z - Arrêté	
	TEST_SUDOC_UNIMARC2 (SP:7213) - 2021-08-10T09:46:51.082Z - Arrêté	
	Test MA NewRepos (SP:8158) - 2021-08-07T18:06:48.982Z - En cours	
	test_ZD15180_1 (SP:7118) - 2021-07-26T13:31:03.894Z - En cours	
	test_ZD15180 (SP:7118) - 2021-07-22T13:29:26.234Z - En cours	
	vilc-test (SP:7171) - 2021-07-22T09:25:28.444Z - En cours	
	Test MA RD-47 (SP:8158) - 2021-06-09T12:13:35.353Z - En cours	
	Gymnase Auguste Piccard - PROD - médias (SP:7213) - 2021-05-21T08:14:29.204Z - En cours	
	Gymnase Auguste Piccard - PROD - pub-vol (SP:7213) - 2021-05-21T08:12:23.344Z - Arrêté	
	Gymnase Auguste Piccard - PROD - mono (SP:7213) - 2021-05-21T08:08:05.484Z - Arrêté	

# 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.

## Paramètres généraux

Nombre max de mots (5)

Nombre min de mots (2)

Taille minimal des mots-clés (3)

Nombre maximal de requêtes (10)

Nombre min de résultats (7)

Nombre max de résultats (500)

Utilisation de la vedette principale (Oui)



# 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

## Seuils de similarité textuelle

Seuils de similarité textuelle

Niveau parfait (0.95)

Niveau fort (0.85)

Niveau proche (0.7)

Niveau faible (0.55)

# 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.

---

## Paramètres de correspondance

### Paramètres généraux

Couverture source minimal (0.5)

Couverture cible/source minimal (0.6)

Couverture source optimal (0.79)

Couverture cible/source optimal (0.9)

Score moyen minimal (0.3)

### Seuils de similarité textuelle

Seuils de similarité textuelle

Niveau parfait (0.95)



# Facts and figures

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

# Facts and figures

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

# Facts and figures (Fall 2021)

	% rejected records	% derived records	% records manually checked
Primary and Secondary 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