

Luminess

Advanced Digital Business Process

**AI presentation
Luminess / WIPO**

12 March 2024

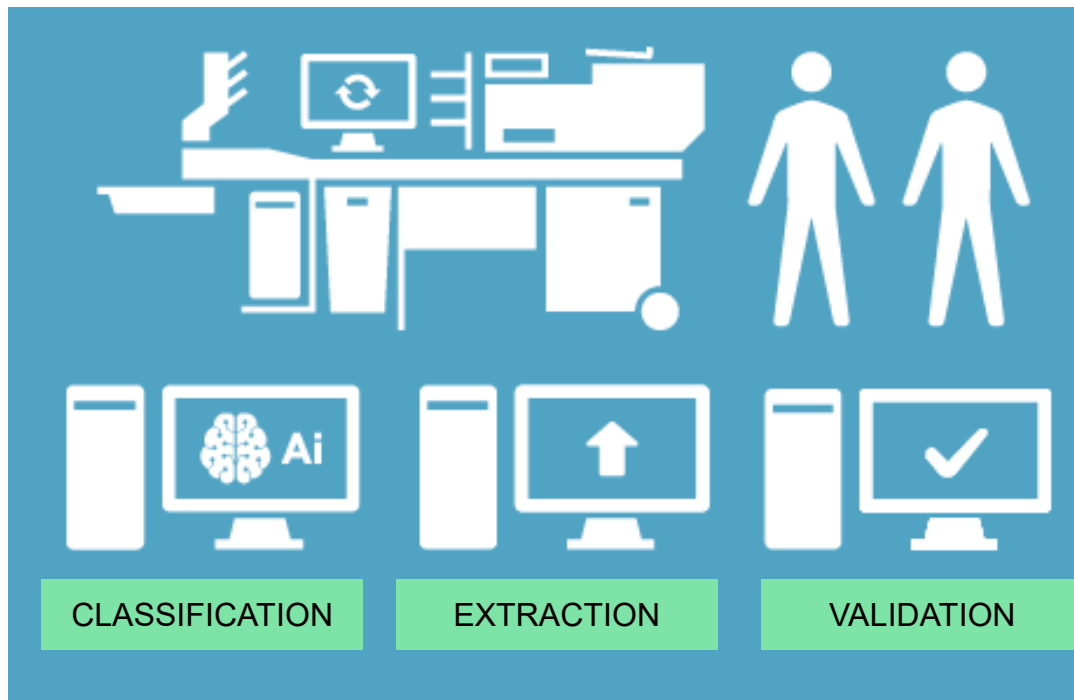
Speakers

- Clémence MARTINOT (Luminess)
- Jacques GUYOT (SimpleShift)



Luminess

Luminess,
a ubiquitous positioning



On the one hand
BPO expertise



on the other side
Technological excellence

Expert in complex process automation

With AI all along the process

SEE

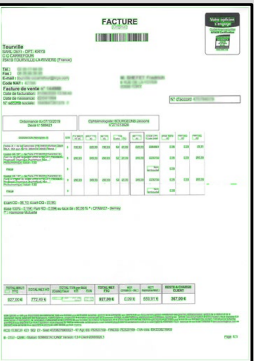
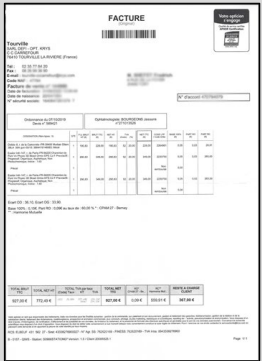
READ

UNDERSTAND

ANALYZE /
PROCESS

AUTOMATE

MONITOR



```
FACTURE
Votre
opticien
à
Riot
(original)
LA
Qualité
de
service
certifiés
AENGR
Certification
N
UN
NUM
ET
2
DE
SERVICE
te
S
HAUTE
:
Tourville
...
Client
```

```
FACTURE
Votre
opticien
à
Riot
(original)
LA
Qualité
de
service
certifiés
AENGR
Certification
N
UN
NUM
ET
2
DE
SERVICE
te
S
HAUTE
:
Tourville
...
Client
```

A green arrow points from the 'ANALYZE / PROCESS' stage to the 'UNDERSTAND' stage, indicating a feedback loop.



▶ **EXTRACTION**

▶ **UNDERSTANDING**

▶ **AUTOMATION**

MULTIMODAL APPROCH :
IMAGE, TEXTE, VIDEO, DATA

DEEP LEARNING

OCR

GENERATIVE AI, LLM, NLP

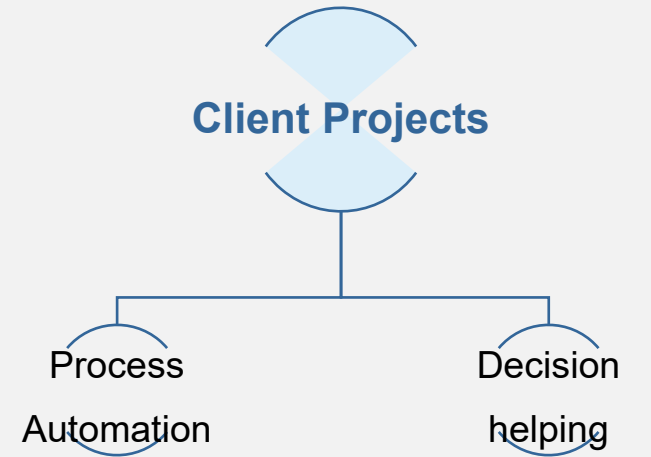
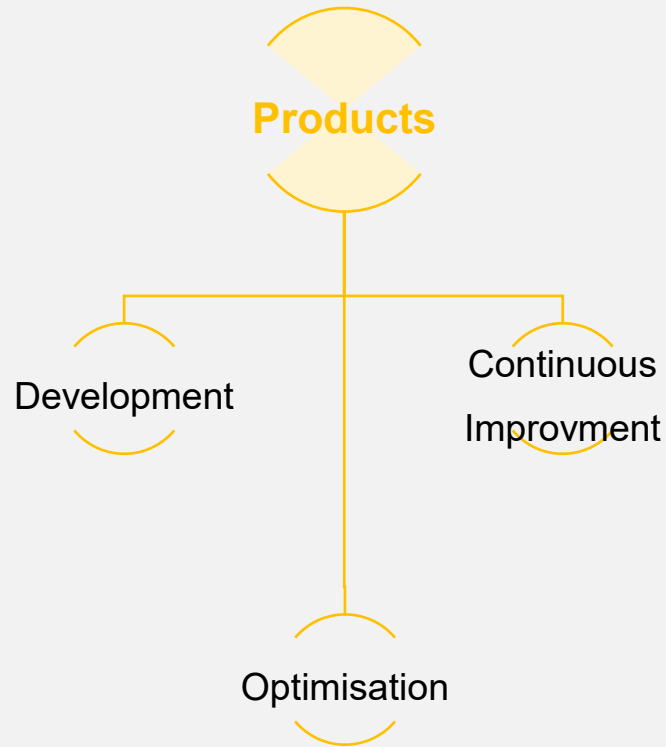
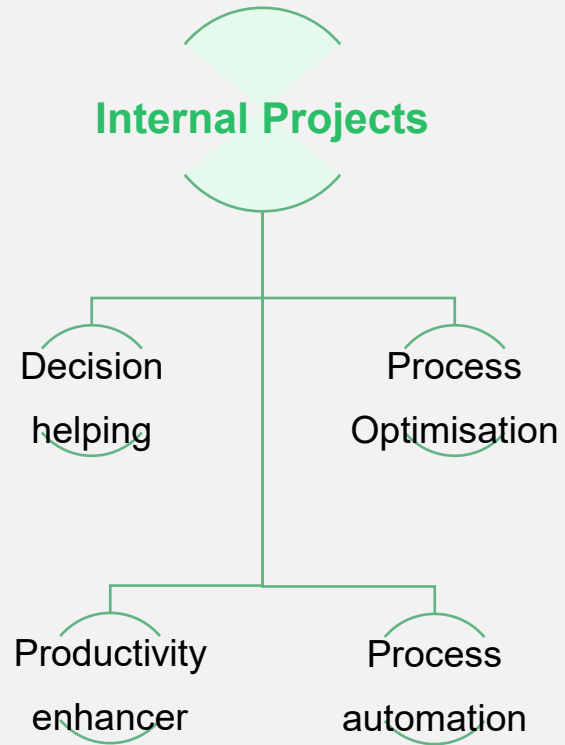
MACHINE LEARNING

**COMPORTEMENTAL
ANALYSIS**

OPTIMISATION

**DIAGNOSTIC
ANALYSIS**

Luminess AI Project types



Examples

OEB

Spreadsheet and
mathematical
equations detection

Banking, insurance

Document automatic
data extraction

DILA

Automatic abstract
from legal texts

DILA

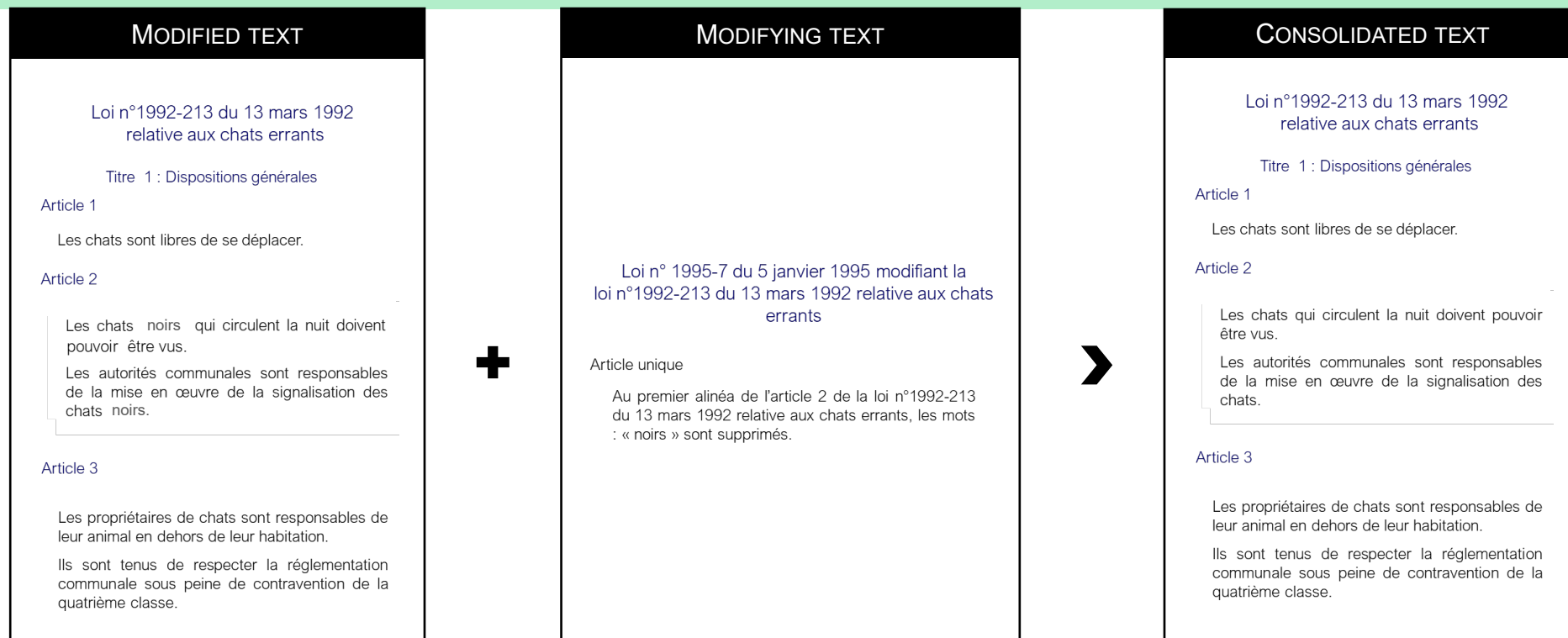
Consolidating french
legislation



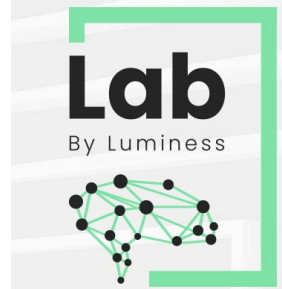
DILA - Consolidating French Legislation

DILA : Public Direction for Legislation & Administration information

Among other goals : consolidating french legislation when publication on the OJ of modifying texts

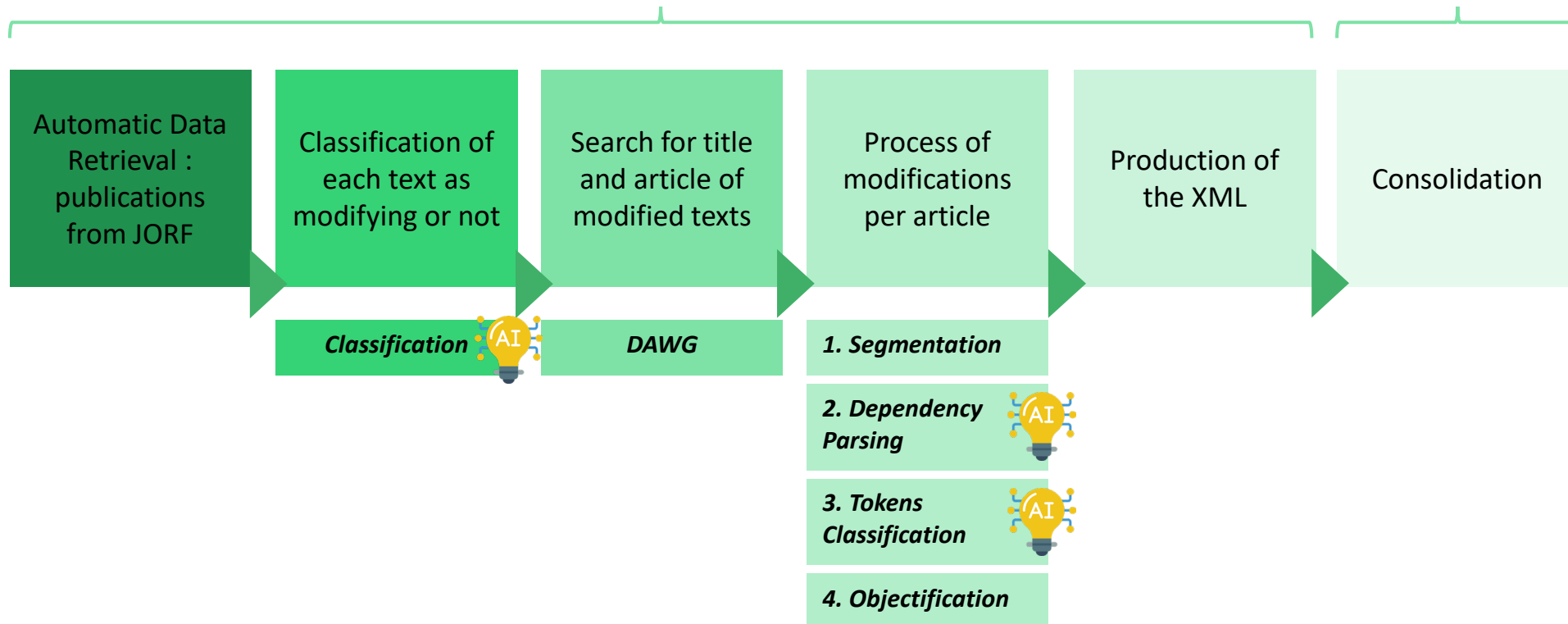


DILA - Consolidating French Legislation



App 1 : Annotation module

App 2 : Consolidation module



DILA - Consolidating French Legislation



EASY TAKE OVER BY LA DILA

- Solution based on the DILA tools
- Same work environment for DILA agents

TRIPLE PERFORMANCE IMPROVMENT

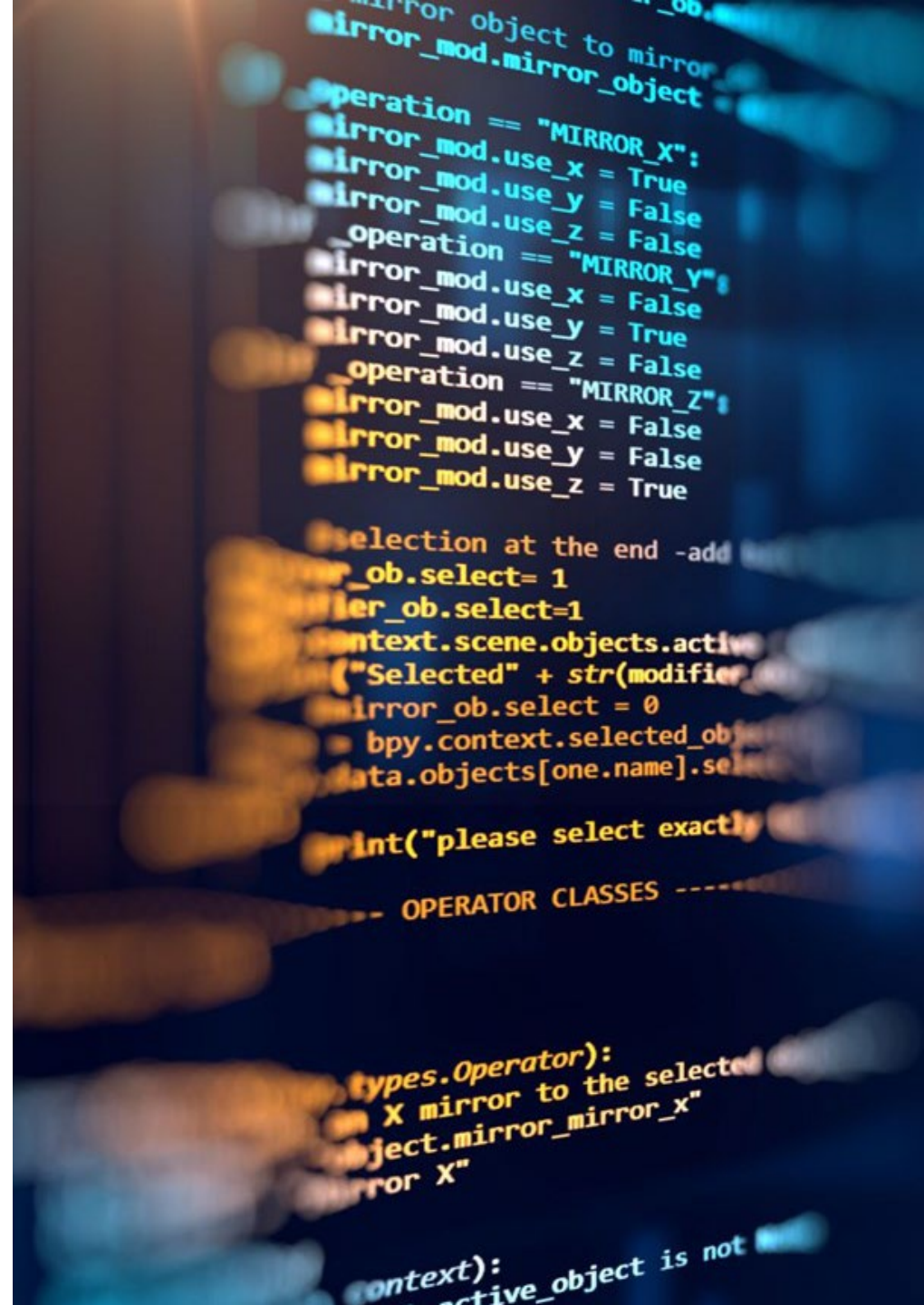
- Time Improvement
- Number of texts Improvement
- Complexity Improvement : valorization of agents work

TECHNICAL BUT NOT TECHNOPHILE

- Performance
- Optimised execution time
- Sustainable solution
- Continuous improvement

Luminess

- AI by Luminess
 - Presentation Luminess
 - Focus on NLP, LLM & Generative IA
- Field of application at Luminess
 - Internal projects
 - Products
 - Clients
 - DILA : présentation du projet de consolidation automatique de textes juridiques
 - OEB : classification de brevet
 - OEB : reconnaissance automatique de tableaux, équations mathématiques
 - DILA : résumé automatique de textes juridiques



AI reclassification Service

An approach targeting an AI-based intellectual reclassification service within IPCWLMS for patent families that have reached stage 3 of IPC reclassification as an alternative to the Default Transfer

Jacques Guyot Simple-Shift

Plan

01

**History &
objectives &
constraints**

02

AI Principle

03

**Training data for
AI**

04

Evaluation

05

**System in
production**

06

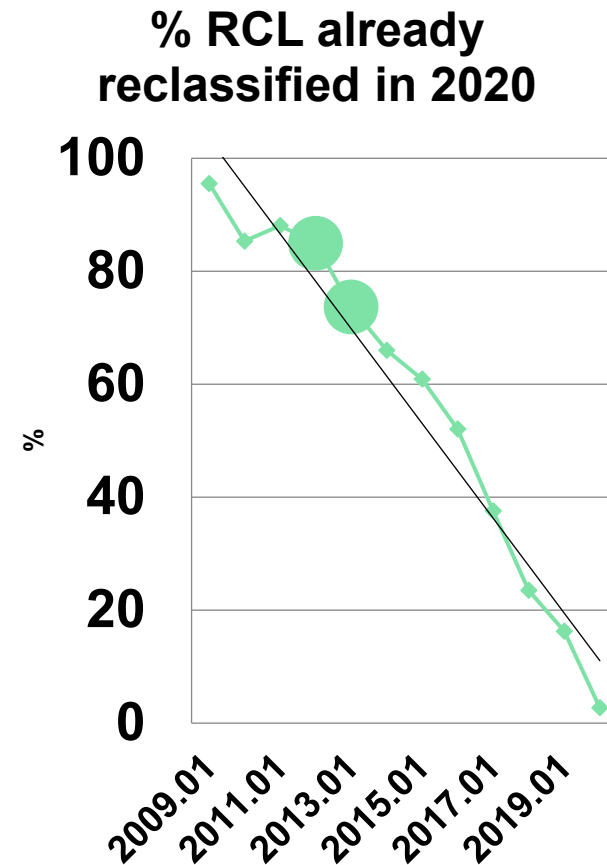
**Possible
improvements**

01

HISTORY & OBJECTIVES & CONSTRAINTS

History & Objectives & Constraints

- Objectives
 - Automatically RCL reclassification
- Why use AI?
 - Avoid Default transfer
- Key dates
 - Prototype and tests - 2020
 - System developed - 2021
 - Put into production - 2022



History & Objectives & Constraints

- Constraints

Data :

- Use manually reclassified documents to train AI
- Use the Revision Concordance List (RCL) and Working Lists (WL)

Performance evaluation :

- Control the quality and performance of AI
- Allow evaluation before execution with clear indicators

02

AI PRINCIPLE

AI Principle

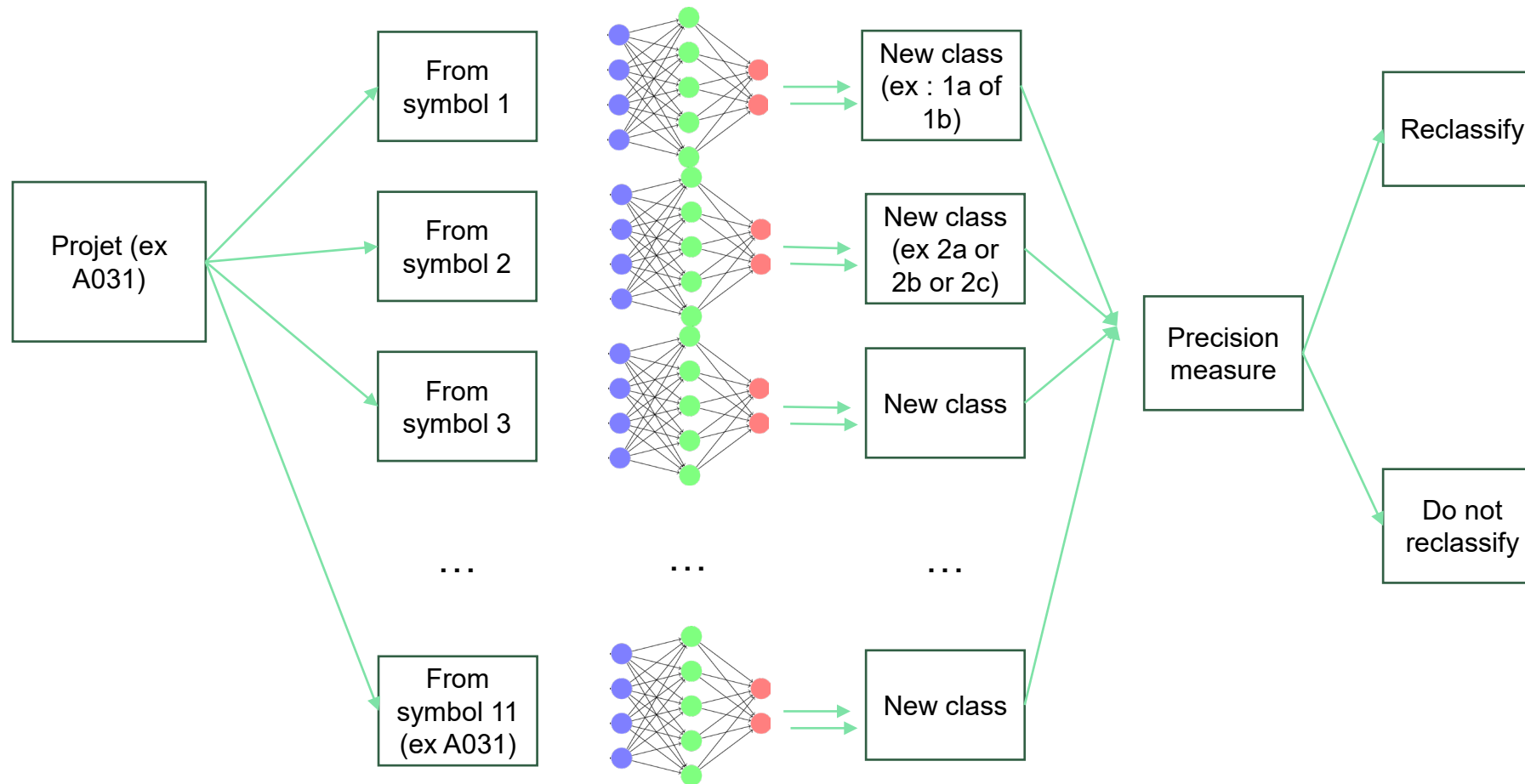
Select project

Select symbol

Calculate new classes

Evaluate

Act



03

TRAINING DATA FOR AI

Training and Test Corpus

- Based on 2020 “February” Wipo Delta EN
- The training corpus contains all the documents having:
 - A title and a summary in English, and IPC symbols
 - Belonging to a family where exists at least one member (publication) with
 - A "to" target symbol from one of the 2012.01 revision projects
 - A "R" value at position 39 (see WIPO ST.8). already been reclassified
 - A version date 201201
 - Also includes the indication of its family, so we have reconstituted the list of families with their associated documents.
 - Each family was also assigned to a symbol AI if one of its documents had at least one target symbol (To).
- The corpora have been divided into families, in order to be able to constitute consistent test or training sets. The documents of a family are either in the test set or in the training set.
- So for the 130 symbol AIs, we can build training and test sets that concern only the prediction from the source symbol "From" to the target symbol "To" of the RCL.

Luminess

Advanced Digital Business Process



04

EVALUATION

Definitions



Precision

The AI always makes a prediction.

We only examine the first prediction.

Precision is the ratio of correct predictions to all predictions.



Accuracy

Accuracy is the minimum precision required by the user



Recall

Recall determines the percentage of predictions reaching the required accuracy



AI precision for an 80/20 set

- Precision 80% on 2012
- Precision 78% on 2013

Project	% Macro Average precision for first prediction at 80/20 (R)
A023	80
A030	79
A031	86
A032	75
A033	84
A034	77
A035	81
A036	83
A037	91
A038	85
A040	76
F002	75
F003	73
F005	83
Total	80

AI Precision/Recall for 80/20



During prediction, the symbol AI gives a value that estimates the potential of the prediction. The higher this value, the bigger is the confidence in the prediction.



The value is called the "score" of the prediction. This value must be calibrated for each neural network to be translated into "accuracy".

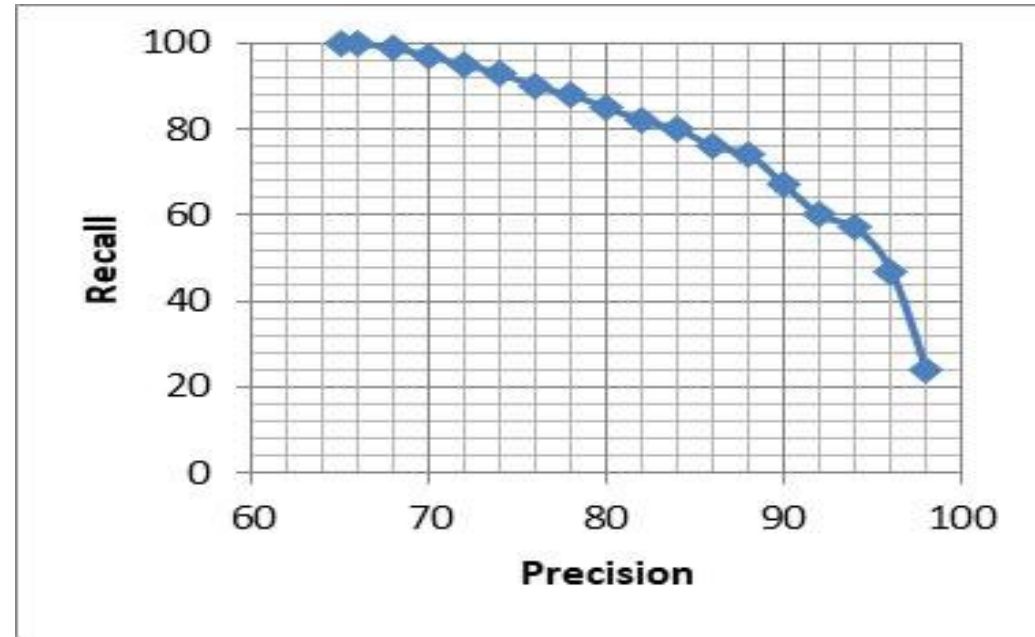


To illustrate this concept, we took a symbol AI with moderate performance

F003 source symbol	Precision 80/20 (recall 100%)	Test numbers	Number of target symbols (to)
D04H0001540000	65	1 599	16

Fixing accuracy impact recall

Accuracy %	Recall %	Score
65	100	756
66	100	789
68	99	820
70	97	845
72	95	863
74	93	881
76	90	904
78	88	924
80	85	950
82	82	969
84	80	985
86	76	1007
88	74	1025
90	67	1054
92	60	1088
94	57	1116
96	47	1164
98	24	1273



By imposing an **80% accuracy** , we impose a recall at 85%.

That is, in 15% of cases, AI will not give an answer. In these cases, a default transfer (or other symbol allocation process)

WL corpus to be reclassified for the 2012

- The “WL corpus” is defined as corpus of documents extracted from families to be reclassified. It is selected on the basis of the Wipo Delta EN
- The WL corpus contains all documents (publications) with
 - Having a title and a summary and IPC symbols
 - A "from" source symbol for the 2012 revision
 - A version date prior to 2012

Coverage of WL

- 82% of families to be reclassified have a document in English
- 18% of the families to be reclassified for the 20120101 revision would require additional processing.
- The cases are as follows:
 - If there is a text in another language (use automatic translation- discuss later)
 - If there is no text at all, then to perform automatically a Default Transfer.

Reclassification with > 80% accuracy

Families may have several symbols to be reclassified, so the results are in terms of symbols to be reclassified. The overall average of the revision is therefore 83% (Macro average).

The calculation of the family distribution (Micro average) is 75%. This leaves 25% of symbols that will require special treatment (default transfer)

(81% macro & 71.2% micro for 2013 revision)

05

SYSTEM IN PRODUCTION

Implementation flow



The user asks to simulation for a project (fix accuracy)

Collect RCL, Training, WL
Compute Simulation - % precision - % recall



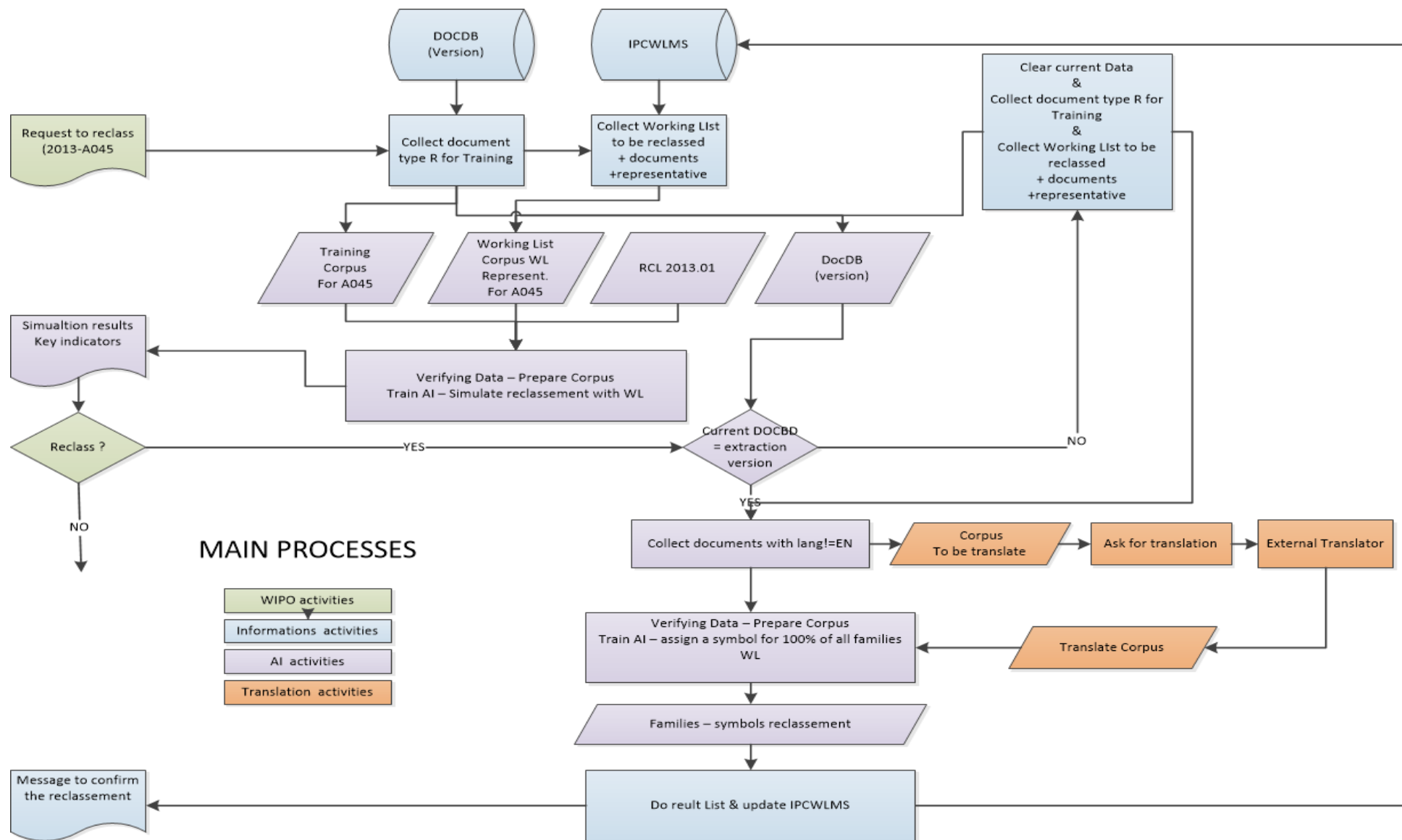
The user decides on results to wait



The user decides on results to accept

Update RCL, Training, WL (Training 100%)
Compute info for Result List (reclassification)
Modify DocDB Version (record new Symbol)

Implementation flow



06

POSSIBLE IMPROVMENTS

Improvement for reclassification

Adding documents with no english text

- Using automatic translation

Adding information to documents

- A part of description (where to find it?)
- Add cpc symbol
- Add document classified after the revision (B)

Assisted Manual Classification

- AI offers a list of documents to reclassify to accelerate learning

Change precision calculation

- If prediction is the default transfer symbol then it's ok

Luminess

Thanks for your attention

Welcome to your questions

Contact :

bdelorme@luminess.eu
11 Boulevard de Sébastopol
75001 Paris - France

Luminess

Advanced Digital Business Process

07

ANNEXES

Summary RCL, Training, precision

Revision		
% families already reclassified	85 %	74 %
RCL nb projects	14	17
RCL nb source symbols (from)	130	139
RCL nb target symbols (to)	576	1528
AI		
Nb family in training corpus	359'000	234'000
80/20 MacroAvg Precision	80%	78.4%
80/20 MicroAvg Precision	75%	71.1%
Worst Project	F003 - 73%	A050 – 59%

Summary WL – precision and recall

Working List		
Total families to be reclassified in 2020	118'000	357'000
Total families with EN documents	97'000 (82%)	314'000 (88%)
Total Symbols to be reclassified	191'000	663'000
Total Symbols predicted by AI with precision>80%	143'000	472'000
Total symbols needs a kind of “default” Symbols	48'000	191'000
MacroAvg Recall at 80% precision	83%	81%
MicroAvg Recall at 80% precision	75%	71%

Automatic Translation Impact

Translation and Other on 2012 and 2013		
Total families with documents in a WipoTranslate language	689	948
Total families with documents in another language	8	6

5,949,167 families still to be reclassified (1 family may be involved in several projects, etc.) - 100%

Of which **87.9%** have at least 1 member with English text

Of which **1.22%** (among the remaining 12.1% not processed by EN) have at least one member with a WIPO translate language (other than English)

Of which **0.23%** (among the remaining 10.88% not processed by WIPO Translate including EN) have at least one member with a language other than those of WIPO translate

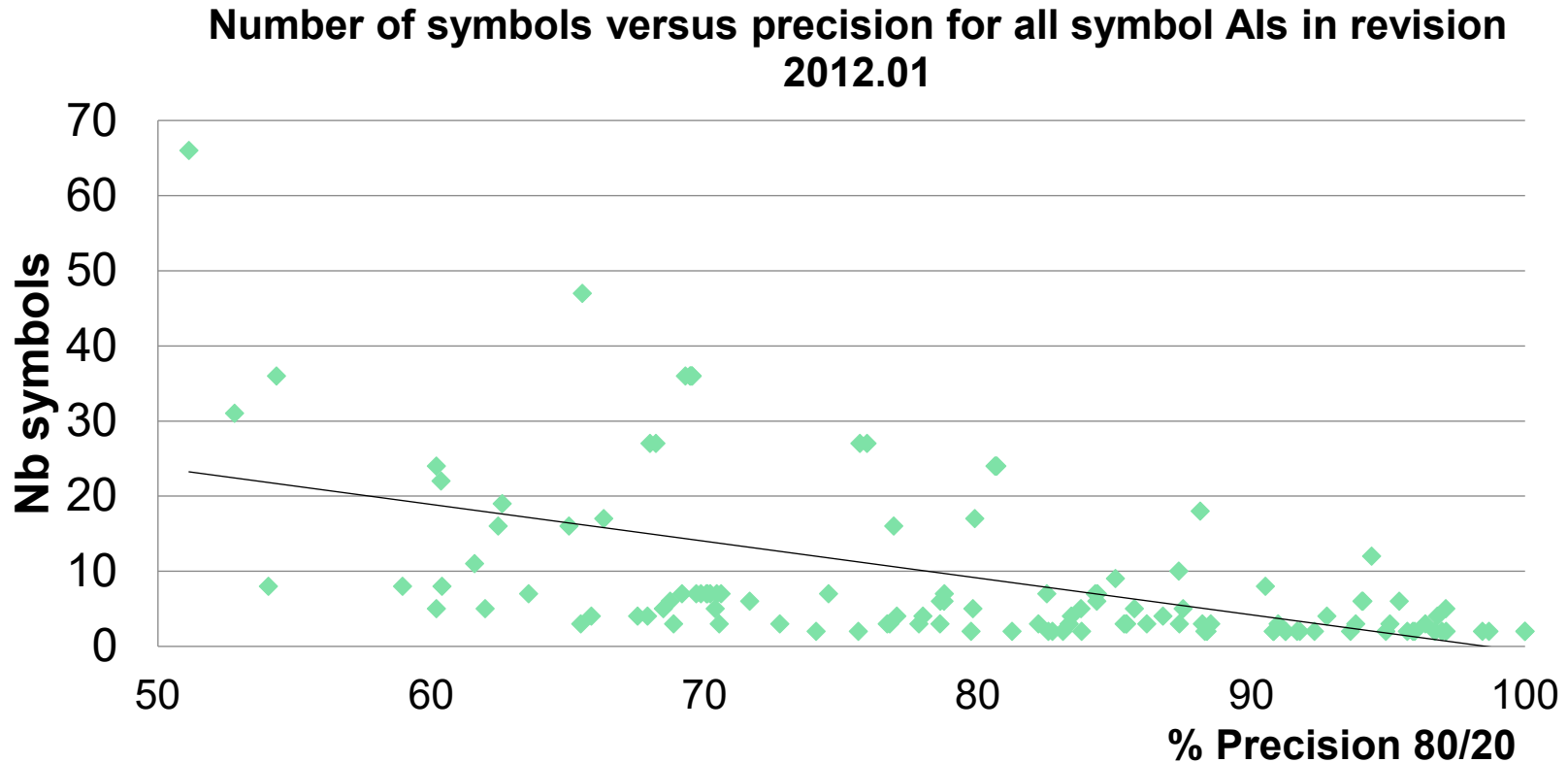
And of which **10.6%** have no member with text (therefore requires a default transfer)

Improve the coverage with automatic translation

- The existing language statistics for a document in DOCDB can be used as an indicator to quantify these different cases .
- 75% of families have an abstract. And the distribution of languages is:
 - 95.1% have an English abstract
 - 4.6% without abstract in English but having an abstract in one of the languages translated with WipoTranslate.
 - And 0.4% have an abstract in another language.
- This shows that the use of machine translation can increase the number of cases that can be handled by AI.

Number of symbols versus precision

More symbols less precision



Impact of precision limit on 2013

By choosing 80% precision, we have 191'886 symbols which will require a default transfer.

By dropping to 70%, we reduce that number to 97'207.

Going down to 65%, we reduce that number to 65'419.

By dropping to 60%, we reduce that number to 15'748.

Only those skilled in the art can decide on the choice between an AI allocation which may be affected by an error or a systematic allocation (default transfer)

project	Tot Symbols 2be reclassified	nb prediction with 100-80% precision	nb prediction with 80-70% precision	nb prediction with 70-65% precision	nb prediction with 65-60% precision	need a "default" transfert
A041	1 377	842	182	73	165	115
A042	36 015	35714	301	-	-	-
A043	4 017	4017	-	-	-	-
A044	101 629	72881	10 262	3 231	7 483	7 772
A045	144	144	-	-	-	-
A046	1 624	1548	73	3	-	-
A047	50 718	35357	9 463	1 779	2 093	2 026
A049	4 231	3491	740	-	-	-
A050	117 799	54659	28 406	11 029	23 489	216
A051	93 030	66273	13 823	6 831	6 103	-
A053	2 208	1486	128	95	170	329
A054	5 630	5582	48	-	-	-
A055	6 641	5848	793	-	-	-
C458	1 981	1981	-	-	-	-
F001	61 738	41938	8 792	5 195	5 813	-
F004	110 364	90268	6 899	3 552	4 355	5 290
F006	64 595	50826	13 769	-	-	-
Total/avg	663 741	472855	93 679	31 788	49 671	15 748
recall		71%	85%	90%	98%	

Adding Corpus B to the Training

- The addition of “corpus B” to the training part of “corpus R” has proven to be little use
- We could not find a rational explanation for this as the number of examples has been shown to have a positive impact on accuracy
- One explanation is to assume that over time, identity and definition become more clearly established and that symbols classified 9 years ago would be classified differently today.
- The use of the revision symbols differs somewhat and does not reinforce the vision of the reclassifications to be made today.
- **Verification of this would require validation on other revisions.**

Project	% Macro Average precision for first prediction at 80/20 (R)	% Macro Average precision for first prediction at 80/20 (B)
A023	79.6	78.7
A030	79.5	80.0
A031	86.0	86.2
A032	75.2	72.6
A033	83.5	82.5
A034	76.5	73.7
A035	80.6	79.1
A036	82.7	82.5
A037	91.2	91.6
A038	84.6	81.3
A040	75.8	72.7
F002	74.6	73.0
F003	73.3	73.4
F005	83.1	82.9
Total or %	80.4	79.3

Impact of the Size on training

Waiting improves precision,
but already a small % of documents brings good performance

Project	10/90	20/80	30/70	40/60	50/50	60/40	70/30	80/20	90/10
A023	76	78	79	79	79	80	79	80	80
A030	63	71	73	75	77	76	75	79	85
A031	81	83	83	84	83	83	83	86	88
A032	72	73	74	74	75	75	75	75	75
A033	76	79	81	82	82	83	83	84	85
A034	70	74	74	75	74	74	73	77	77
A035	77	78	79	80	80	80	80	80	80
A036	77	78	80	81	82	82	82	82	82
A037	75	80	80	80	79	80	80	83	83
A038	85	85	85	85	85	85	85	85	84
A040	55	57	59	59	60	59	59	59	66
F002	70	73	73	74	74	74	74	75	76
F003	69	72	73	74	74	75	75	75	75
F005	71	73	76	77	77	78	80	81	82
Average	73	76	77	78	78	78	78	80	81

Assisted Manual Reclassification

AI provides lists of the most relevant documents to classify (those which provide the most information)

→ **minimize the number of documents to be classified manually to achieve a certain accuracy**

random Reclassification Versus assisted Reclassification

