

A Simple and Uniform Way to Introduce Complimentary Asynchronous Interaction Models in an Existing Document Analysis System

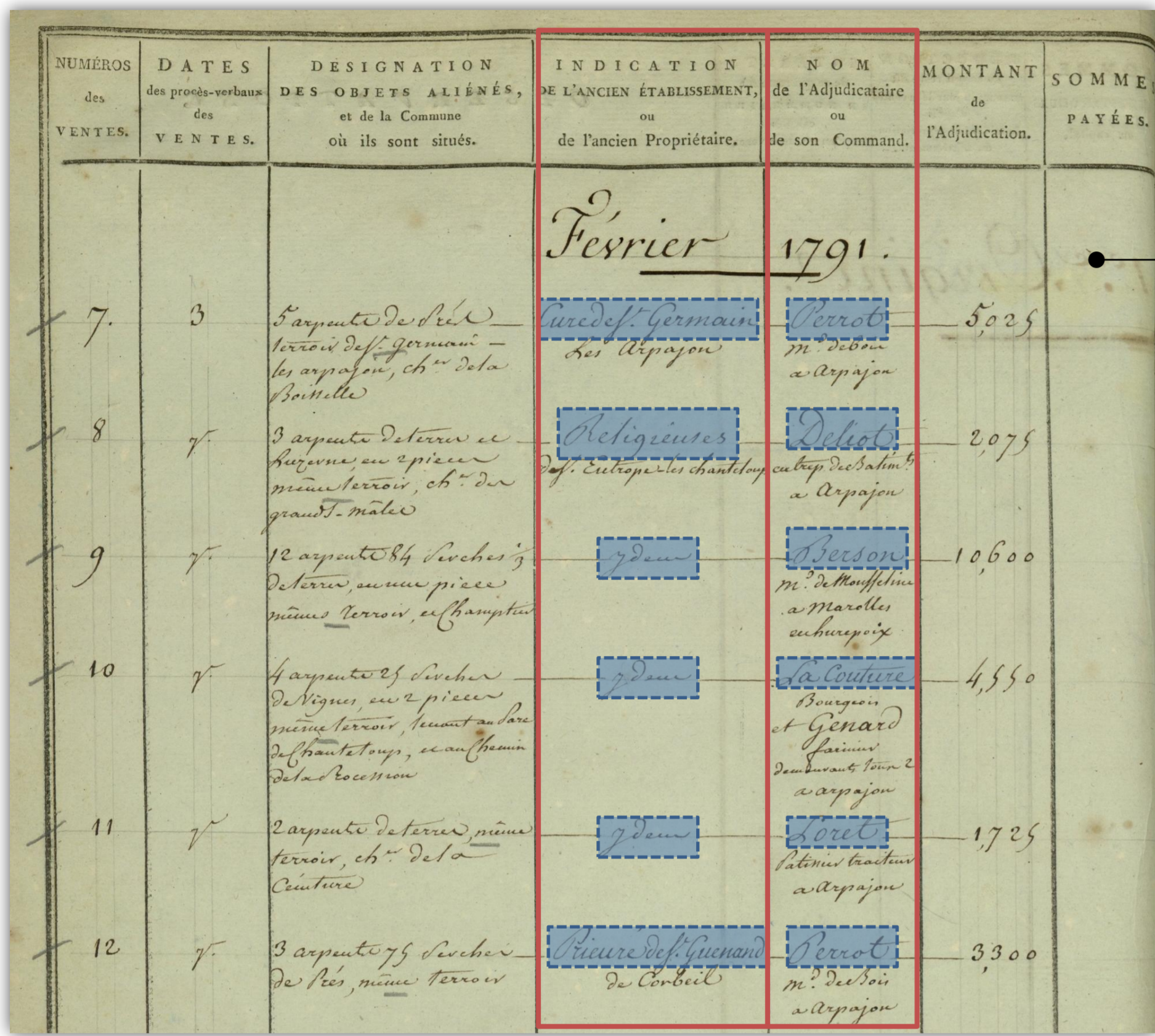
JOSEPH CHAZALON*, BERTRAND COÛASNON*, AURÉLIE LEMAITRE#

*INSA Rennes | #Univ. Rennes 2 – UMR IRISA, Université Européenne de Bretagne, France

Keywords: human interaction; asynchronous interaction; interaction models; degraded documents; historical documents

ABOUT THIS WORK

Motivations	Claims	Contributions
<ul style="list-style-type: none"> Document alteration and variability prevent DAR systems from producing reliable results when processing quantities of historical documents. No matter how fine you tune a system, lots of manual corrections will be required. 	<ul style="list-style-type: none"> The classical develop—process—correct scheme is inefficient for old or degraded documents. Interacting during the analysis stage, and not after, may permit an early correction of errors and a reduction of manual correction cost. 	<ul style="list-style-type: none"> An asynchronous interaction framework suitable for massive document processing 2 efficient interaction models for error handling Guidelines for adapting existing page analyzers using standard exception mechanism

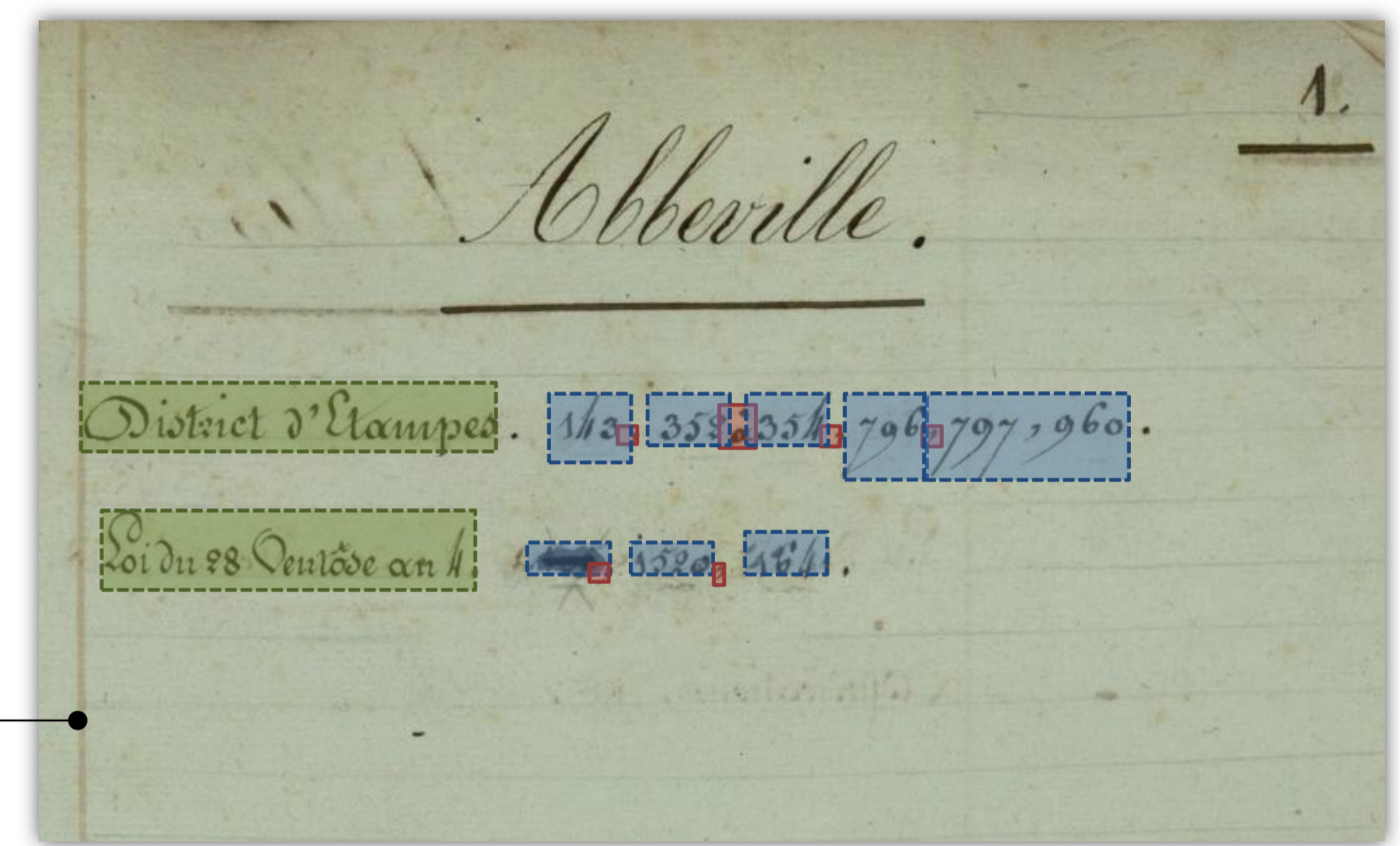


Example task 1 : Name recognition

Documents: Sale registers from the 18th century
Goal: Detect and index previous owner/buyer names
Challenge: Difficult handwritten recognition without lexicon (at the beginning) + semantics ("IDEM" = ditto)
Human interaction: Correct recognition and enrich a lexicon progressively (avoid keying same name twice)

Example task 2 : Number segmentation

Documents: Sales index (city, sale kind, sales numbers)
Goal: Export structured content to a database
Challenge: Under-segmentation of numbers (some separators are not detected, and numbers are merged)
Human interaction: Correct segmentation, but not recognition (has to be done automatically)



HOW TO ADAPT AN EXISTING PAGE ANALYZER?

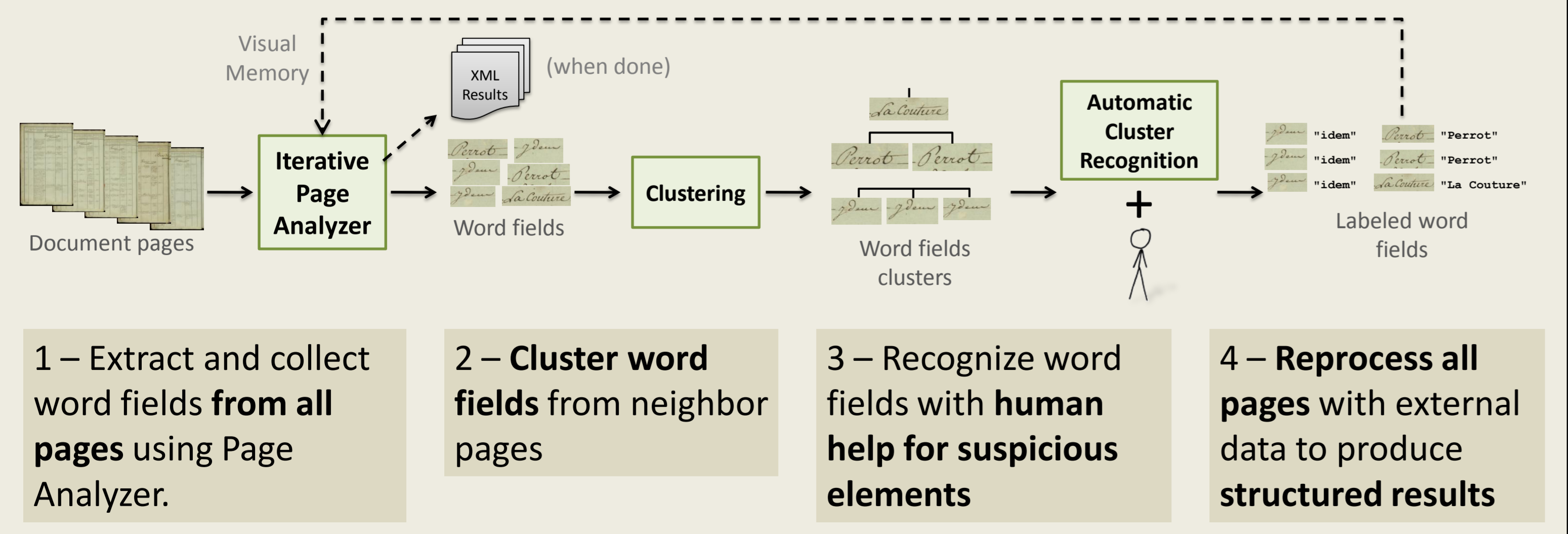
1 – Integrate the Page Analyzer in a collection processing system	2 – Enable an iterative analysis of pages using a Visual Memory	3 – Enrich page description to handle interactive exchanges automatically
<ul style="list-style-type: none"> "Collection" = set of documents sharing common properties: same writer, similar structure, etc. Using a global Strategy Module, it is possible to: <ul style="list-style-type: none"> ✓ orchestrate the analysis of subset of the documents using a Page Analyzer ✓ while interacting asynchronously about another subset with a Human Interface 	<ul style="list-style-type: none"> Reprocessing pages with external information permits to correct previous analysis. The Visual Memory is a dataset where each element has a shape and a position in its image referential <ul style="list-style-type: none"> ✓ it is sent to the Page Analyzer with each image ✓ it allows the fusion between image and external contents during the processing 	<ul style="list-style-type: none"> Page description is a document specific part of the Page Analyzer. New operators for this description indicate <ul style="list-style-type: none"> ✓ illegal or suspicious constructs (<i>error detection</i>) ✓ independence of subparts of the image (<i>error recovery</i>) ✓ less reliable or critical part which can be analyzed externally (<i>error correction</i>) The interaction framework is a generic part of the Page Analyzer which handles communication with external modules automatically Implementation is done with standard exception mechanisms which are easy to adapt

DIRECTED INTERACTION

Principle

- When error/lack of knowledge is automatically detected, the Page Analyzer **asks questions** to the Strategy Module which **collects and forward them**.
- External answers will be **used in a later invocation** of the Page Analyzer on the same page.

Experiment: Name recognition (Example task 1)
Dataset: 70 pages (1206 word fields) of sale registers from the 18th century



Results and Discussion

- Amount of manual work can be **adjusted** by tuning reject thresholds
- 28% relative reduction of manual work for 80% recognition (indexation task) when compared to post-processing correction**

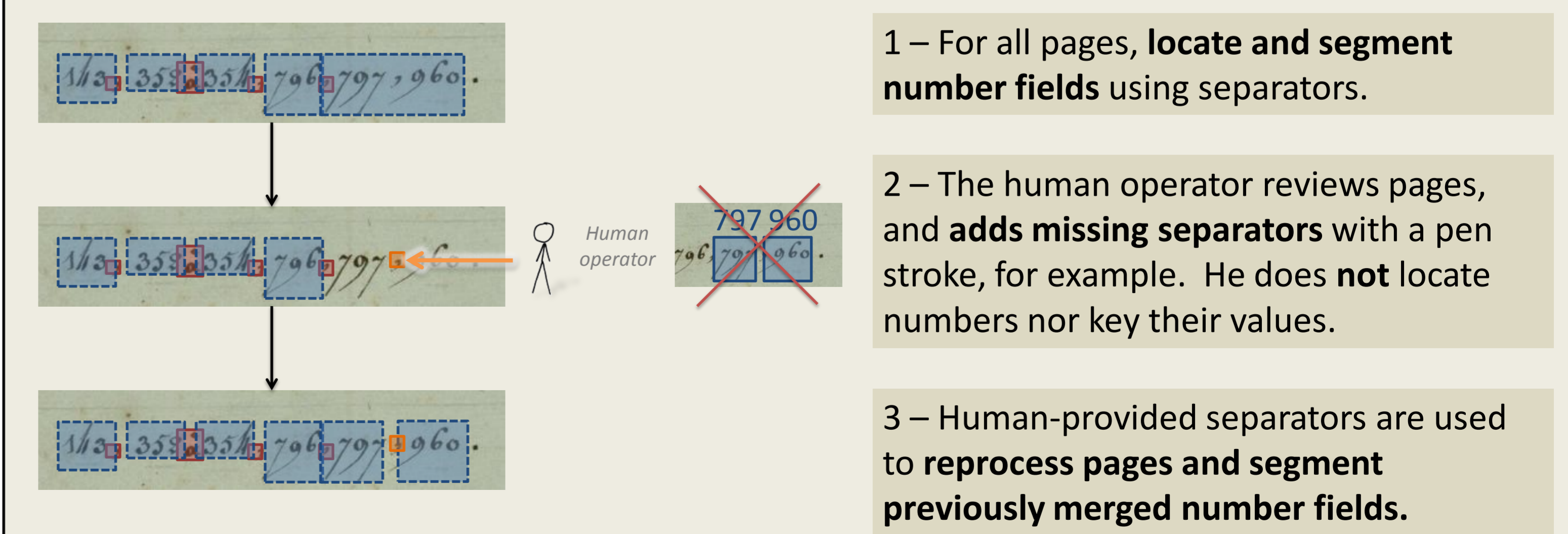
More details in L. Guichard, J. Chazalon, and B. Couasnon, "Exploiting Collection Level for Improving Assisted Handwritten Words Transcription of Historical Documents," in Proc. of ICDAR, 2011.

SPONTANEOUS INTERACTION

Principle

- Useful when error detection **cannot** be performed automatically.
- Page Analyzer **still can benefit** from **optional** human help/corrections.

Experiment: Number segmentation (Example task 2)
Dataset: 50 pages (1637 number fields) of sale indexes from the 18th century



Results and Discussion

- Manual edition is **simpler** than the one required during post processing
- 30% relative reduction of manual work (ground-truthing task) when compared to post-processing correction**

More details in J. Chazalon and B. Couasnon, "Iterative analysis of document collections enables efficient human-initiated interaction." in Document Recognition and Retrieval XIX, Proc. SPIE, 2012.