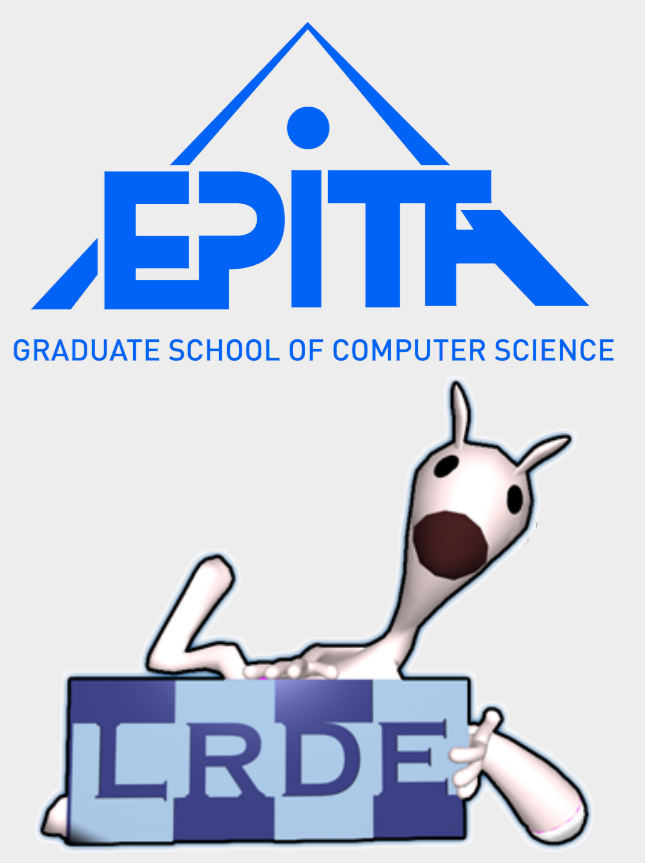




The SCRIBO Module of the OLENA Platform: a Free Software Framework for Document Image Analysis

Guillaume Lazzara, Roland Levillain, Thierry Géraud, Yann Jacquélet, Julien Marquegnies, Arthur Crépin-Leblond
EPITA Research and Development Laboratory (LRDE), France
olena@lrde.epita.fr



At a Glance

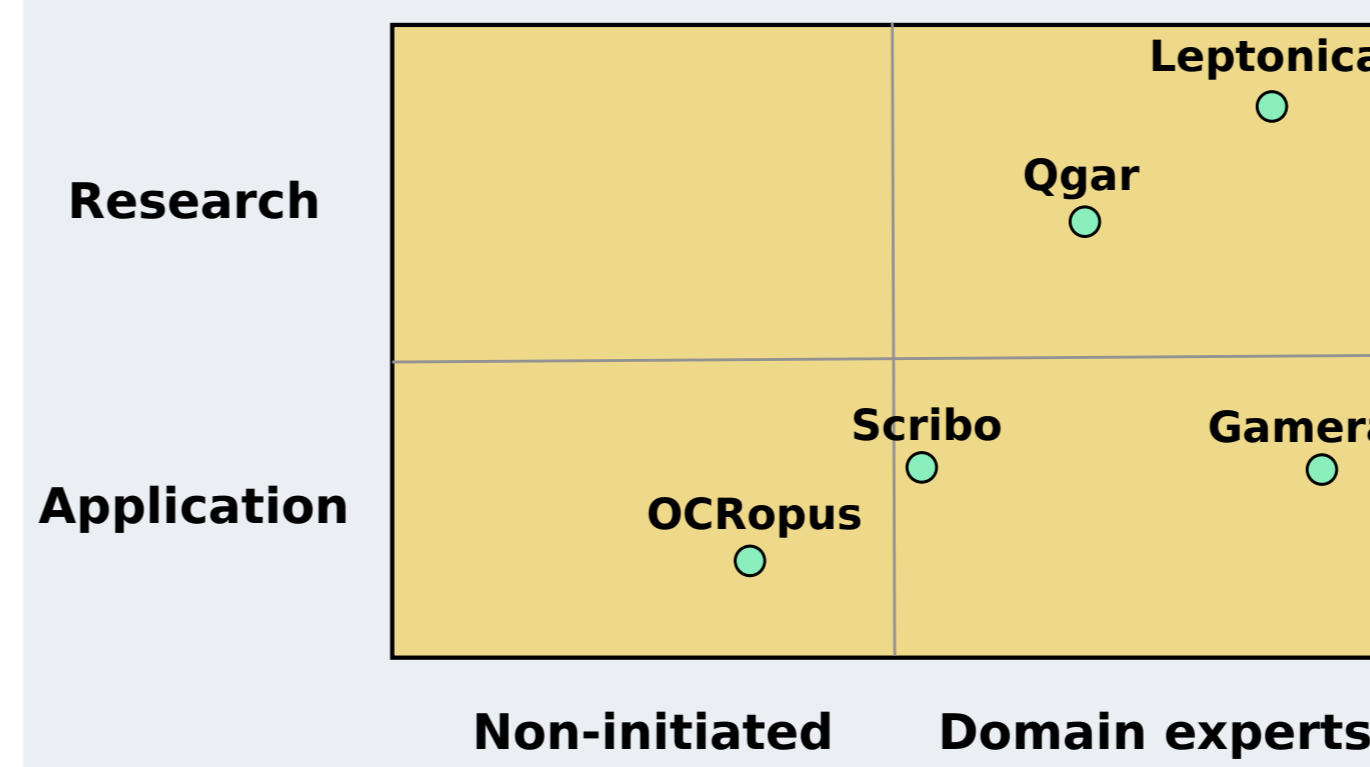
The Issue A Document Image Analysis (DIA) processing chain cannot handle all types of documents.
The Point It is necessary to provide specific treatments for each kind of documents.

Our Contribution A framework to design DIA software, preserving flexibility and efficiency.
The Outcome The implementation of our proposal, the SCRIBO module, illustrates the benefits of this approach.

Desired Properties of a Modern DIA Framework

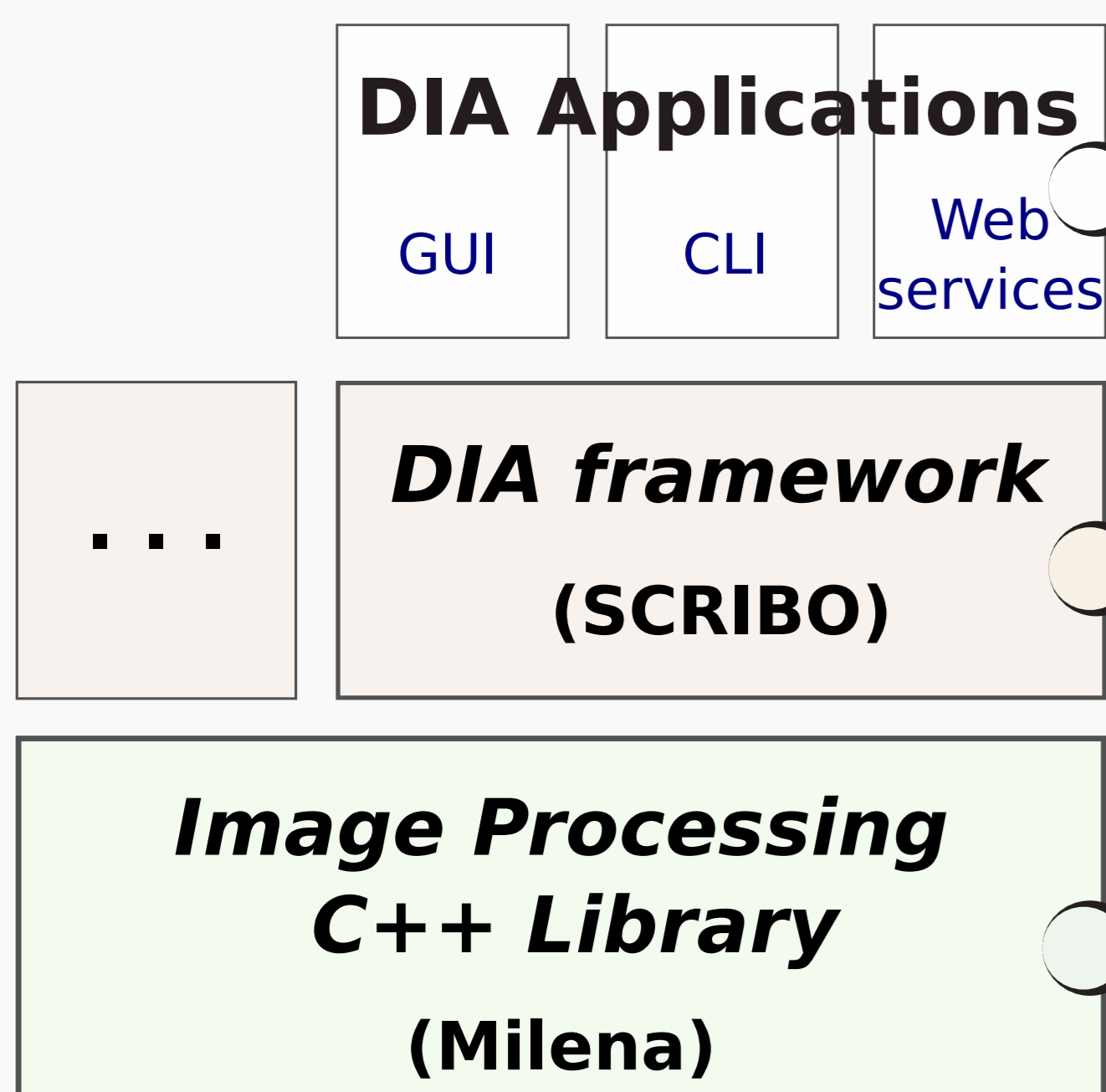
- Flexibility** Reusable building blocks to adapt processing chains to specific documents.
- Efficiency** Handle large amounts of documents.
- Multiple interfaces** Command line and Graphical Interfaces available.
- Easy to integrate** high-level Application Programming Interface (API) and support for various platforms.

Motivations



- Implement a framework with all our desired properties.
- Provide easy-to-use applications for DIA
- Make research progress in DIA accessible to end-user applications.
- Using our Image Processing (IP) library in concrete use cases.

The Olena platform



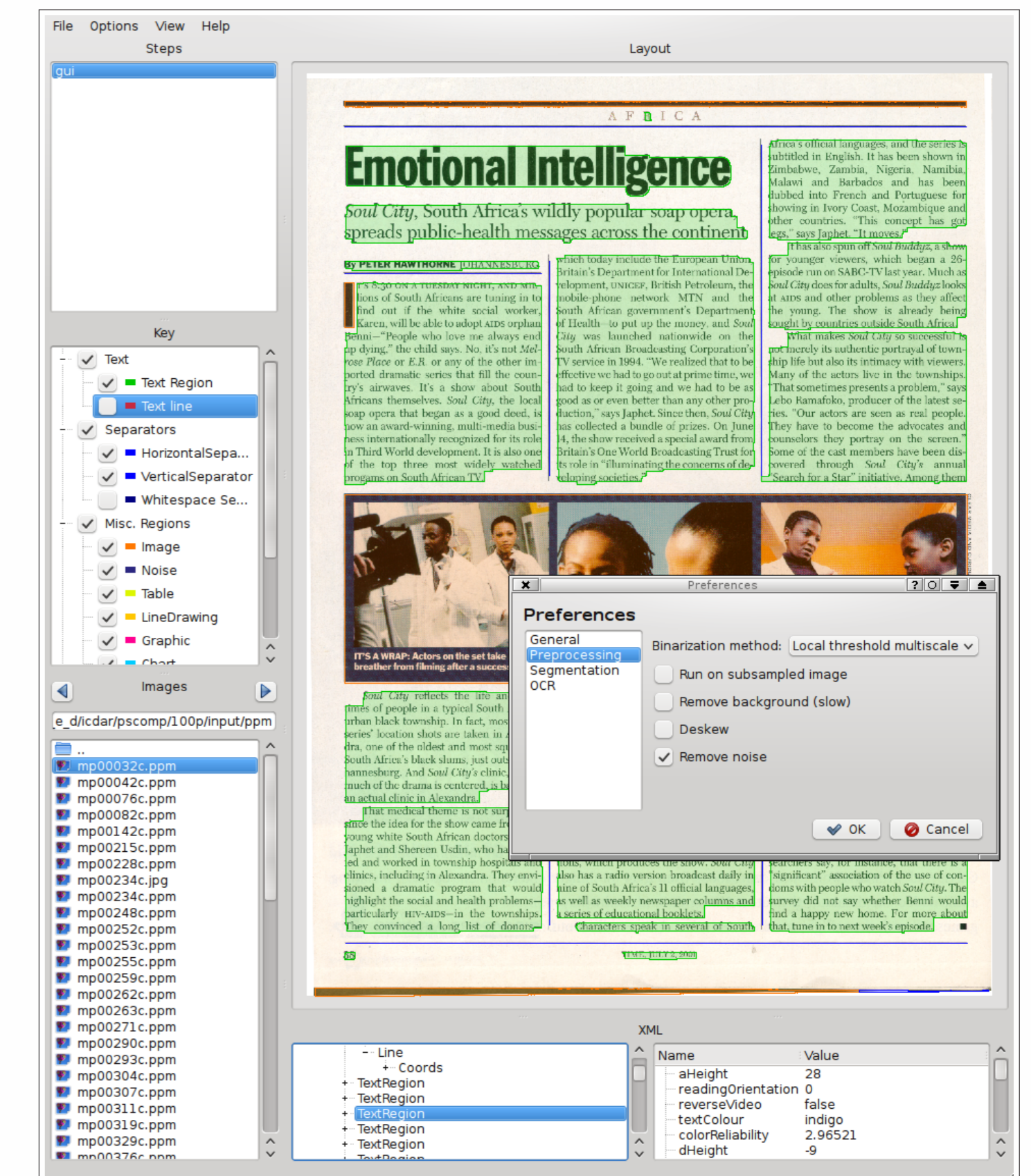
Applications and Uses Cases



Original document image.



Document reconstruction in PDF.



GUI for DIA and reconstruction.

More information

Online demos
<http://olena.lrde.epita.fr/Demos>
Website
<http://olena.lrde.epita.fr/>
Contact
olena@lrde.epita.fr

The SCRIBO module: a DIA Framework

Provides

- Basic routines
- Basic DIA toolchains
- Text in document
- Document layout analysis
- Text in picture
- High-level data structures
- Novel algorithms and techniques
- Standard I/O
- GUI and Command Line Interface (CLI)

Facts

- 3 years of development
- 40K lines of C++
- Open Source GPL v2
- Used in Nepomuk/KDE

Assets

- End-to-end tools → From digital document to HTML and PDF reconstruction.
- Based on a well established IP library.

The SCRIBO Project [1]



- Project conducted in the context of the "System@tic Paris-Région" Cluster (France).
- 9 Partners : AFP, CEA-List, EPITA, INRIA-Alpage, Mandriva, Nuxeo, Proxem, Tagmatica, XWiki.
- 3 years of development.
- Budget of 3,5M€

Milena: a Generic Image Processing Library [2]

Provides

- Data structures
- Safe data types
- More than 70 algorithms
- Memory management

Facts

- 10 years of development
- Version 1.0 released on July 2009
- 120K lines of C++
- Open Source GPL v2

References

[1] SCRIBO, Semi-automatic and Collaborative Retrieval of Information Based on Ontologies.
<http://www.scribo.ws>.

[2] Roland Levillain, Thierry Géraud, and Laurent Najman.
Why and how to design a generic and efficient image processing framework: The case of the Milena library.
In Proc. of the IEEE Intl. Conference on Image Processing (ICIP), 2010.