HITS
Heidelberg Institute for Theoretical Studies

www.h-its.org

Robust Extraction of Marked-Up Text Sections from Scientific Document Printouts

Mark-Christoph Müller

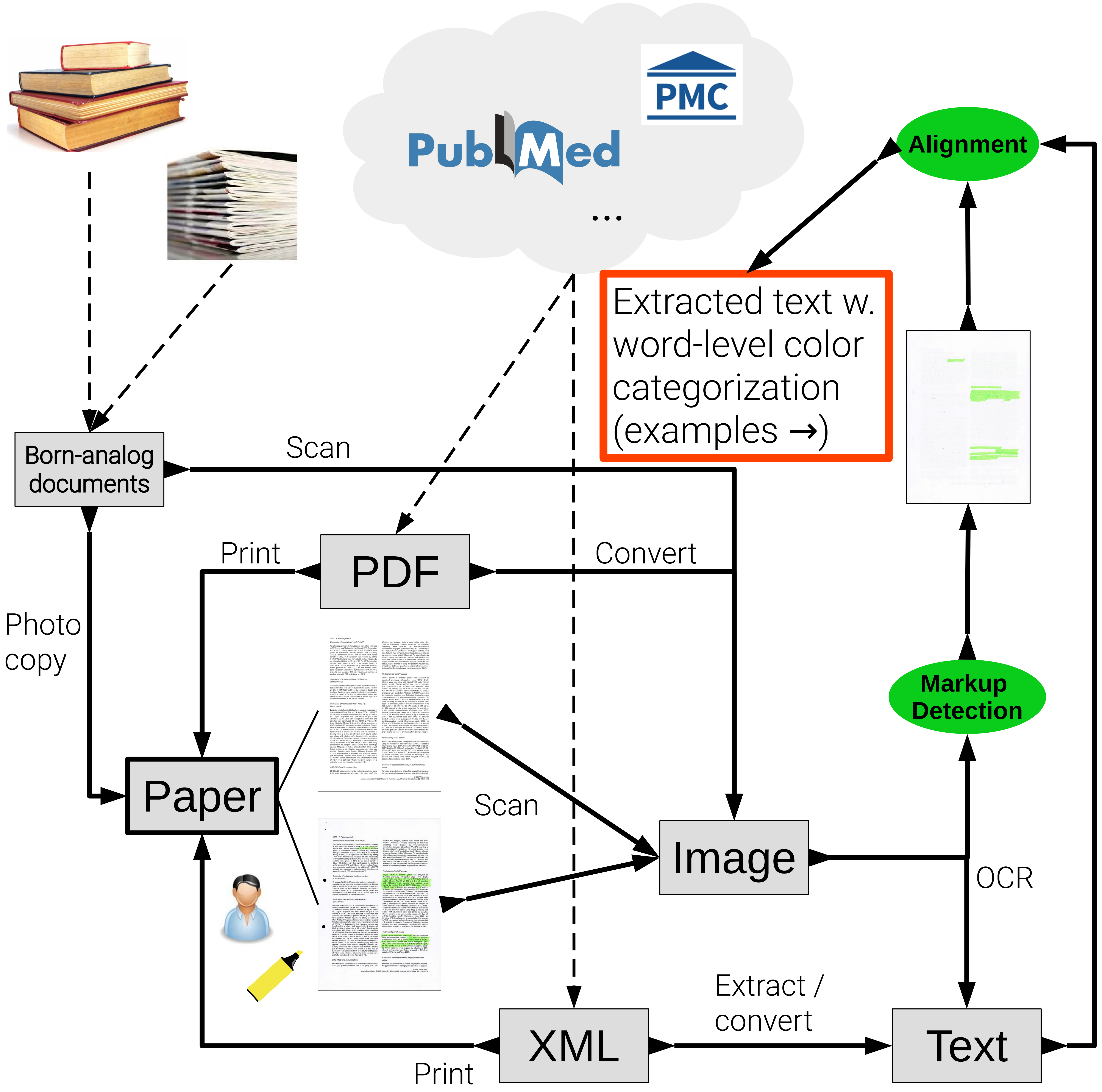
Heidelberg Institute for Theoretical Studies gGmbH, Heidelberg, Germany

Abstract

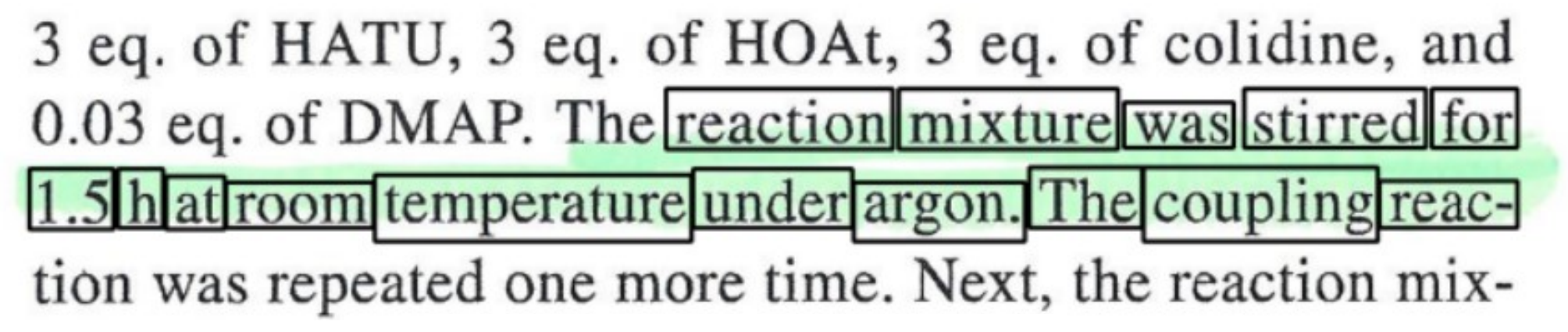
We present a simple tool for extracting text and markup information from **printouts** of (not only) scientific documents. While the heavy-lifting OCR is done by off-the-shelf **tesseract**, our focus is on detection, extraction, and basic **categorization** of color-highlighted text sections, as well as on providing a framework for downstream processing of extraction results.
The tool can be useful for document analysis tasks that must, or benefit from being able to, use printed paper. The code is available at **https://github.com/nlpAThits/docimg2mmax**

Background & Motivation

Despite the shift towards PDF and XML, **printed paper** is still crucial for scientific document use.
It is the medium of choice for **active reading**, supporting straightforward markup with highlighter pens, which is commonly done during manual excerption from scientific literature.
However, as soon as the highlighted text is supposed to undergo further **computational processing**, paper ceases to be practical.
Biomedical database curation is a case in point: Here, human domain experts often use paper printouts to mark up relevant sections in scientific documents, but for the subsequent database insertion (often done by other people), the data has to be re-keyed manually, which is both ineffi-cient and error-prone.



Example Results



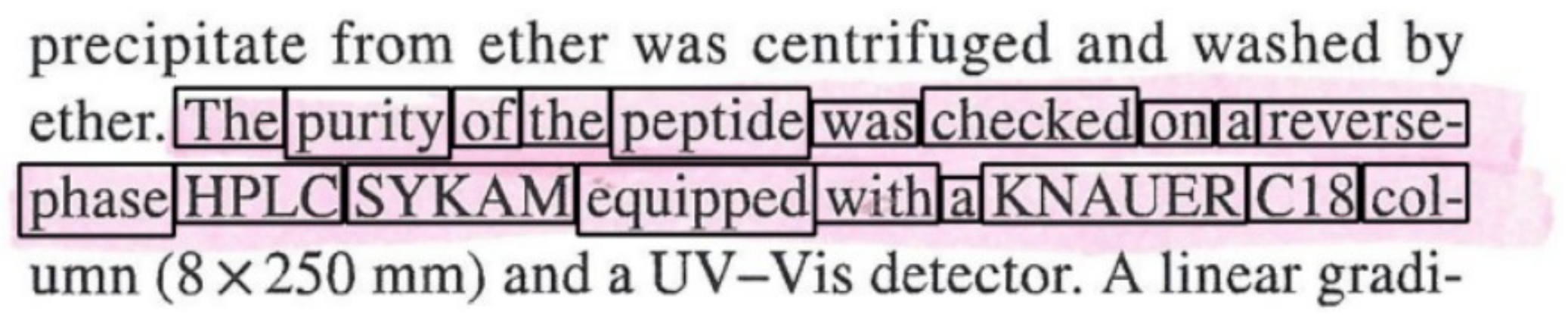
3 eq. of HATU, 3 eq. of HOAt, 3 eq. of colidine, and 0.03 eq. of DMAP. The reaction mixture was stirred for 1.5 h at room temperature under argon. The coupling reaction was repeated one more time. Next, the reaction mix-

"reaction"

Percentage colored 36%

Dominant color 245:255:244

3 eq. of HATU, 3 eq. of HOAt, 3 eq. of colidine, and 0.03 eq. of DMAP. The reaction mixture was stirred for 1.5 h at room temperature under argon. The coupling reaction was repeated one more time. Next, the reaction mix-



precipitate from ether was centrifuged and washed by ether. The purity of the peptide was checked on a reverse-phase HPLC SYKAM equipped with a KNAUER C18 column (8×250 mm) and a UV–Vis detector. A linear gradi-

"peptide"

Percentage colored 69%

Dominant color 253:224:246

precipitate from ether was centrifuged and washed by ether. The purity of the peptide was checked on a reverse-phase HPLC SYKAM equipped with a KNAUER C18 column (8 X 250 mm) and a UV-Vis detector. A linear gradi-

Acknowledgements

This work was done as part of the project **DeepCurate**, which is funded by the German Federal Ministry of Education and Research (BMBF) (No. 031L0204) and the Klaus Tschira Foundation, Heidelberg, Germany.