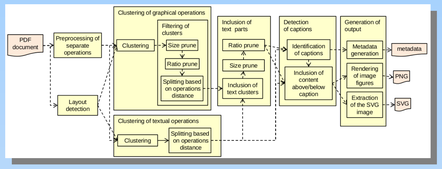
Capturing, Mining, Storing and Retrieving Scientific Images in INSPIRE

Piotr Praczyk (CERN, Universidad de Zaragoza)

Extraction of figures and tables from PDF documents

PDF has become a de facto standard in scholarly publishing. Being able to extract figures from PDF files is crucial when aiming at having a complete set of figures from HEP publications.

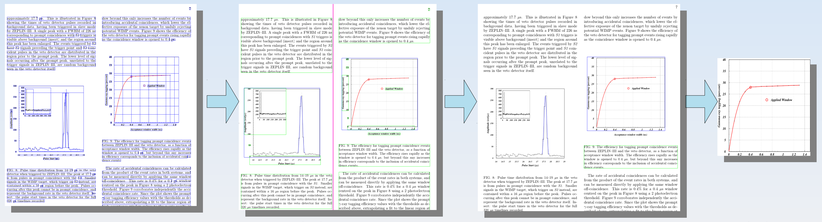


(1) PDF consists of a series of graphical and textual primitives. At first we identify these primitives.

(2) We cluster outcomes based on their position in PDF instructions stream, geometrical position and type.

(3) We use properties of figures to reject incorrect candidates, we merge parts of texts belonging to figures and identify captions. We also distinguish between figures and tables.

(4) We generate additional meta-data and export figures/tables in different formats (PNG, SVG).



(1) (2) (3) (4)

Extraction from LaTeX source and merging

• Most of the INSPIRE content comes from arXiv.org and the LaTeX source is available.

• Documents are parsed searching for standard markup used to construct figures.

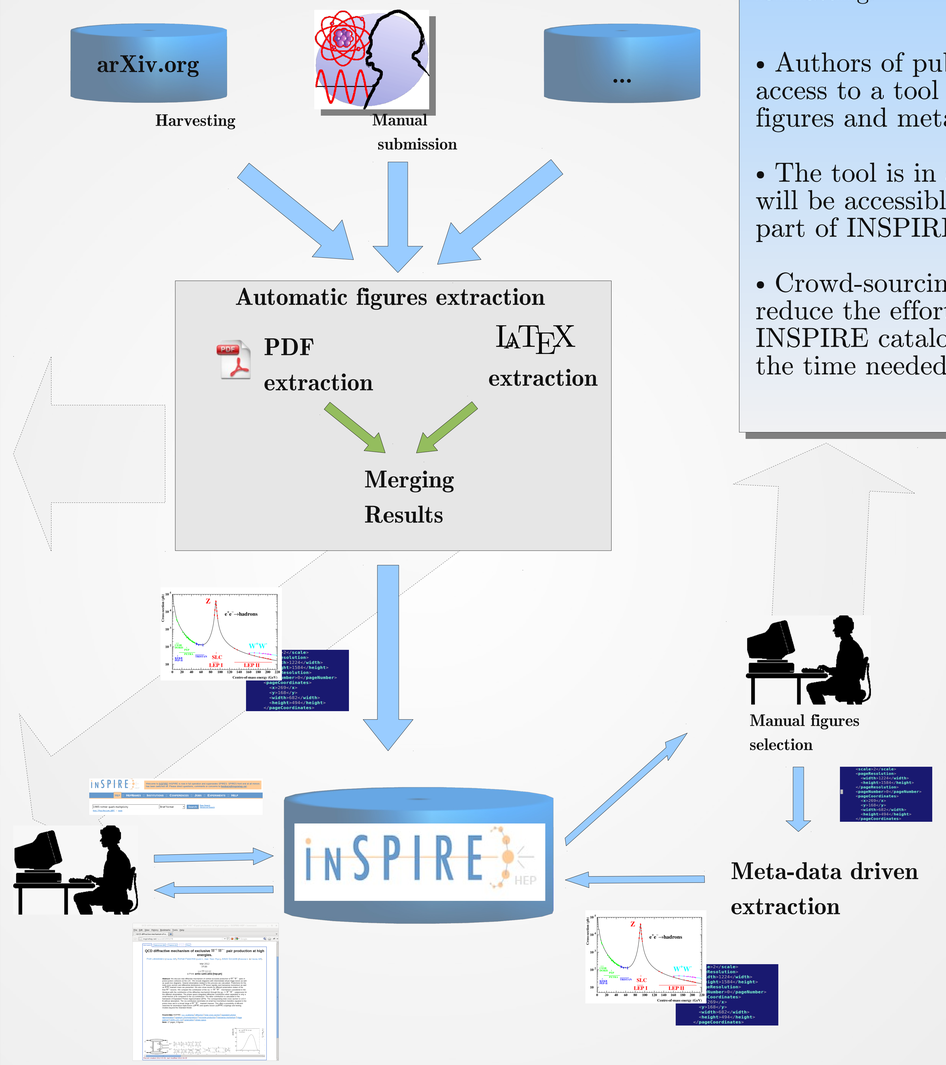
• LaTeX extraction allows more precise extraction of content that is described in a standard manner.

• Content generated by special LaTeX modules, graphics modified from within markup and so on, can not be extracted

• Some types of meta-data can not be extracted

• When both LaTeX and PDF are available, we execute both procedures and select partial results that are more accurate

INSPIRE, the successor of SPIRES, is a digital library allowing access to publications from High Energy Physics. We present a project aiming at extension of INSPIRE to include and take advantage of figures extracted from scientific publications.



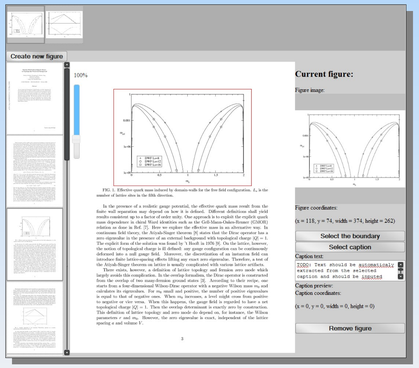
Crowd-sourcing[1] in difficult cases

In some cases, the automatic extraction generates incorrect results. The extraction has been designed to deal with documents born digital. Some older entries were created using more traditional techniques and later scanned and encoded as PDF. Also some newer content may significantly vary from usual practices of document formatting.

• Authors of publications are allowed access to a tool for manual selection of figures and meta-data

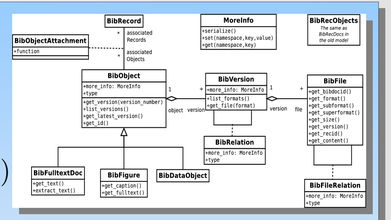
• The tool is in a prototype phase and will be accessible via web browser as part of INSPIRE.

• Crowd-sourcing allows to considerably reduce the effort required on the side of INSPIRE cataloguers and to decrease the time needed to correct mistakes



Storage in INSPIRE

• INSPIRE storage model has been extended to document resources (fulltext files, figures, data) independent from meta-data MARC records.



•We can model relations between documents (i.e. figure being extracted from a source document)

• Resources (also relations and other entities) can have data attached to them in a structure resembling key-value store (MoreInfo)

Future applications of figures

Search for figures based on their meta-data

• Clustering of related publications through similar figures

• Clustering of figures describing the same data

• Search for figures similar to a given one / describing the same phenomenon measured at different time (Evolution of knowledge about a phenomenon)

• Automatic extraction of data present in a figure

References

•A. Holtkamp, S. Mele, T. Simko, and T. Smith. INSPIRE: Realizing the dream of a global digital library in High-Energy Physics. In 3rd Workshop Conference: Towards a digital mathematics library, pages 83–92, Paris, France, 07 - 08 Jul 2010.

• P. Praczyk, J. Nogueras-Iso:Automatic extraction of figures from scientific publications in High-Energy Physics. (to be published)

• P. Praczyk, J. Nogueras-Iso, S. Kaplun, T. Simko : A storage model for supporting figures and other artefacts in scientific libraries: the case study of Invenio. In: Proceedings of the Fourth Workshop on Very Large Digital Libraries (VLDL 2011). Berlin, Germany (September 29th 2011)

• P. Praczyk, J. Nogueras-Iso, S. Dallmeier-Tiessen, M. Whalley. A bibliographic database model providing support for research data in scholar publications (submitted to TPDL 2012)

•[1] Brooks, T. C., Carli, S., Dallmeier-Tiessen, S., Mele, S., & Weiler, H. (2011). Authormagic in INSPIRE -- Author Disambiguation in Scholarly Communication. Proceedings of the ACM Web Science Conference - WebSci 11. Koblenz: ACM Press