PDFJailbreak - a communal architecture for making biomedical PDFs semantic

Alexander Garcia[1,*], Peter Murray-Rust[2], Gully APC Burns[3], Robert Stevens[4], Dominika Tkaczyk[5], Casey McLaughlin[1], Amaury Belin[6], Angelo Di Iorio[11], Leyla García[7], Célya Gruson-Daniel[8], Ross Mounce[9], Andrea Giovanni Nuzzolese[10,11], Silvio Peroni[10,11], Jeremy Spinks[1], Boris Villazon-Terrazas[12], Oscar Corcho[13], Olga Giraldo[13], Mike Wabiszewski[1]

[1]Institute for Digital Information and Scientific Communication, Florida State University, Tallahassee, Florida, USA. [2]Department of Chemistry, Lensfield Road, Cambridge, CB2 1EW,UK [3] Information Sciences Institute, Marina del Rey CA 90292. [4]*The University of Manchester, Oxford Road, Manchester, UK, M13 9PL*. [5]Interdisciplinary Centre for Mathematical and Computational Modeling, University of Warsaw, Poland. [6] Human-Computer Interaction, LIRIS, University of Lyon - France. [7] Temporal Knowledge Bases Group, Universitat Jaumé I, Castello de la Plana, Valencia, Spain. [8]http://hackyourphd.wordpress.com/. [9] Dept of Biology & Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY. [10]Department of Computer Science and Engineering, University of Bologna, Bologna, Italy. [11]STLab-ISTC, National Research Council, Rome, Italy. [12] iSOCO, Intelligent Software Components, Campo de las Naciones, 28042 Madrid. [13]Ontology Engineering Group. Facultad de Informática, Universidad Politécnica de Madrid, Spain.

Jail breaking the PDF

Most biomedical literature is only available in PDF form and this contains a wealth of unused data when liberated. PDFJailbreak is a communal project to create a formal flexible infrastructure to extract semantic information from such documents (papers, grants, theses, reports, guidelines etc.) PDFJailbreak covers the workflow from reading raw PDFs (both interactively and high throughput processing) to domain-specific annotations, argumentation, and data extraction.

PDFJailbreak incorporates a multiplicity of approaches to foster collaboration and optimize the process of settling on a practical single solution. These include:AMI2 (http://www.bitbucket.com/pdf2svg, LA-PDFText (http://code.google.com/p/lapdftext/), CERMINE [1] (http://services.ceon.pl/cermine/index.html), CiTalO [2], Citagora (https://github.com/jam31/citagoraweb/), PDFx (http://pdfx.cs.man.ac.uk/), xPDF (http://www.foolabs.com/xpdf/), Poppler PDF (http://poppler.freedesktop.org/), PDFMiner (http://www.unixuser.org/~euske/python/pdfminer/), PDFBox (http://pdfbox.apache.org/), PDF2SVG, PDFExtract (http://pdfextract.com/), PDF-extract (https://npmjs.org/package/pdf-extract), ParsCit+SectLabel, PDF2XML (http://sourceforge.net/projects/pdf2xml/), JPEDAL (http://www.idrsolutions.com/), and PDF2HTMLEX (http://coolwanglu.github.io/pdf2htmlEX/).

The functionality of these systems include: (1) extraction and normalization of PDF primitives (characters, paths, and images), (2) reconstructions of blocks (whitespace-separated chunks), (3) zoning and general annotation of blocks, (4) extraction of data from tables and figures, (5) extraction of citations, (6) scholarly-specific annotation of components (e.g. citation typing, bibliographic metadata, indexing of materials-and-methods), (7) linking to DBPedia (http://dbpedia.org/) and other scientific semantic services, and (8) domain-specific analysis (e.g., phylogenetic trees, chemistry, sequences).

In many of these there are alternative approaches, often complementary. For example, in blocking we use: machine learning, iterative parameter optimization, heuristics and crowdsourcing. In contrast, traditional XML offerings (e.g., JATS – http://jats.nlm.nih.gov/) only provides a subset of this and only about 10% of the literature is openly available as JATS. PDFJailbreak makes the information in most modern PDF documents fully accessible, including the detailed interpretation of tables and figures.

The architecture is designed for open community development and favours APIs that can be used at various stages in the workflow. Among the initial uses are citation typing, extraction of phylogenetic trees from diagrams, and bio/chemical structures and reactions.

ACKNOWLEDGMENTS

Special thanks to Pablo Mendes, Alexander Constantin, Steve Pettifer and Greg Riccardi.

REFERENCES

[1] D. Tkaczyk, L. Bolikowski, A. Czeczko, and K. Rusek. A modular metadata extraction system for born-digital articles. In 10th IAPR International Workshop on Document Analysis Systems. pp 11–16. 2012.

[2] Di Iorio, A., Nuzzolese, A. G., Peroni, S. Towards the Automatic Identification of the Nature of Citations. In: Proceedings of the 3rd Workshop on Semantic Publishing. pp 63-74. 2013. Montpellier, France