



SOFTWARE TOOL ARTICLE

REVISED MSL: Facilitating automatic and physical analysis of published scientific literature in PDF format [version 3; referees: 2 approved, 2 approved with reservations]

Zeeshan Ahmed^{1,2}, Thomas Dandekar³

¹Genetics and Genome Sciences, School of Medicine, University of Connecticut Health Center, Farmington, CT, 06032, USA

²Institute for Systems Genomics, University of Connecticut Health Center, Farmington, CT, 06032, USA

³Department of Bioinformatics, Biocenter, University of Wuerzburg, Wuerzburg, 97074, Germany

v3 **First published:** 16 Dec 2015, 4:1453 (doi: [10.12688/f1000research.7329.1](https://doi.org/10.12688/f1000research.7329.1))
Second version: 12 Apr 2017, 4:1453 (doi: [10.12688/f1000research.7329.2](https://doi.org/10.12688/f1000research.7329.2))
Latest published: 04 Apr 2018, 4:1453 (doi: [10.12688/f1000research.7329.3](https://doi.org/10.12688/f1000research.7329.3))

Abstract

Published scientific literature contains millions of figures, including information about the results obtained from different scientific experiments e.g. PCR-ELISA data, microarray analysis, gel electrophoresis, mass spectrometry data, DNA/RNA sequencing, diagnostic imaging (CT/MRI and ultrasound scans), and medicinal imaging like electroencephalography (EEG), magnetoencephalography (MEG), echocardiography (ECG), positron-emission tomography (PET) images. The importance of biomedical figures has been widely recognized in scientific and medicine communities, as they play a vital role in providing major original data, experimental and computational results in concise form. One major challenge for implementing a system for scientific literature analysis is extracting and analyzing text and figures from published PDF files by physical and logical document analysis. Here we present a product line architecture based bioinformatics tool 'Mining Scientific Literature (MSL)', which supports the extraction of text and images by interpreting all kinds of published PDF files using advanced data mining and image processing techniques. It provides modules for the marginalization of extracted text based on different coordinates and keywords, visualization of extracted figures and extraction of embedded text from all kinds of biological and biomedical figures using applied Optimal Character Recognition (OCR). Moreover, for further analysis and usage, it generates the system's output in different formats including text, PDF, XML and images files. Hence, MSL is an easy to install and use analysis tool to interpret published scientific literature in PDF format.

Keywords


Bioinformatics, Data mining, Images, Scientific literature, Text, OCR, PDF, Biomedical

Open Peer Review


Referee Status: ✓ ✓ ? ?

	Invited Referees			
	1	2	3	4
REVISED version 3 published 04 Apr 2018	✓ report	✓ report	?	?
REVISED version 2 published 12 Apr 2017	↑	↑	?	?
version 1 published 16 Dec 2015	?	?		

1 **Juilee Thakar**, University of Rochester Medical Center, USA

2 **M. Julius Hossain** , European Molecular Biology Laboratory (EMBL), Germany

3 **Florencio Pazos**, National Center for Biotechnology (CNB-CSIC) (Spanish National Research Council), Spain

4 **Karin Verspoor** , The University of Melbourne, Australia

Discuss this article

Comments (0)

Corresponding authors: Zeeshan Ahmed (zahmed@uchc.edu), Thomas Dandekar (dandekar@biozentrum.uni-wuerzburg.de)

Author roles: **Ahmed Z:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Dandekar T:** Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Ahmed Z and Dandekar T. **MSL: Facilitating automatic and physical analysis of published scientific literature in PDF format [version 3; referees: 2 approved, 2 approved with reservations]** *F1000Research* 2018, 4:1453 (doi: [10.12688/f1000research.7329.3](https://doi.org/10.12688/f1000research.7329.3))

Copyright: © 2018 Ahmed Z and Dandekar T. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

Grant information: This work was supported by a German Research Foundation grant (DFG-TR34/Z1) to TD.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 16 Dec 2015, 4:1453 (doi: [10.12688/f1000research.7329.1](https://doi.org/10.12688/f1000research.7329.1))

REVISED Amendments from Version 2

Following reviewers recommendations, here we present the newly revised manuscript striving for more clarity and better presentation of the results including language and style. Manuscript contains the updated affiliations of authors, updates in Introduction section, updates in Table 2 with the addition of cited manuscripts IDs, addition of new Table 4 about comparative analysis of MSL with different and earlier discussed related applications, addition of Layout-Aware software in discussion and comparison, comparative analysis of software applications, and updates in discussion, conclusion, acknowledgement, and reference sections.

See referee reports

Introduction

There has been an enormous increase in the amount of the scientific literature in the last decades¹. The importance of information retrieval in the scientific community is well known; it plays a vital role in analyzing published data. Most published scientific literature is available in Portable Document Format (PDF), a very common way for exchanging printable documents. This makes it all-important to extract text and figures from the PDF files to implement an efficient Natural Language Processing (NLP) based search application. Unfortunately, PDF is only rich in displaying and printing but requires explicit efforts in the extraction of information, which significantly impacts the search and retrieval capabilities². Due to this reason several document analysis based tools have been developed for physical and logical document structure analysis of this file type.

PubMed and some other publishing platforms (e.g. DOAJ, Google Scholar) provide search options to locate relevant published manuscripts but do not claims to search over the full-text literature and images. The recently, provided basic information retrieval (IR) system by PubMed is efficient in extracting literature based on published text (titles, authors, abstracts, introduction etc.), with the application of automatic term mapping and Boolean operators³. The normal outcome of a successful NLP and text based query brings a maximum of 20 relevant results per page; however, user can improve the search by customizing the query using the provided advanced options. So far, the current PubMed system, as well many other related system are unable to completely implement an efficient information retrieval system, capable of extracting both text and figures from published PDF files. One of the major and technical challenges is the availability of structured text and figures. To our limited knowledge, there still is no single tool available which can efficiently perform both physical and logical structure analysis of all kinds of PDF files and can extract and classify all kinds of information (embedded text from all kinds of biological and scientific published figures). Different commercial and free downloadable software applications provide support in extracting the text and images from PDF files:

A-PDF (<http://www.a-pdf.com/image-extractor/>),

PDF Merge Split Extract (<http://www.pdf-technologies.com/pdf-library-merge-split.aspx>),

BePDF (<http://haikuarchives.github.io/BePDF/>), KPDF (<https://kpdf.kde.org>),

MuPDF (<http://mupdf.com>), Xpdf tool (<http://www.foolabs.com/xpdf/>),

Power PDF (<http://www.nuance.com/for-business/imaging-solutions/document-conversion/power-pdf-converter/index.htm>)

However, these software applications do not provide text and images in a form where they could be considered for further logical analysis e.g. mining text in reading order from double or multiple columns documents (the text of first column followed by the text of second column, and so on), searching marginal text using keywords, removing irrelevant graphics and extracting embedded text inside single and multi-panel complex biological images.

So far, the current PubMed system as well many other related orthodox NLP approaches e.g. 4–13, are unable to completely implement an efficient information retrieval system, capable of extracting both text and figures from published PDF files.

To meet the technological objectives of this challenge, we took a step forward in the development of a new user friendly, modular and client based system (MSL) for the extraction of full and marginal text from PDF files based on the keywords and coordinates (Figure 1). Since MSL provides a module for the extraction of figures from PDF files and applies Optical Character Recognizer (OCR) to extract text from all kinds of biomedical and biological Images. MSL comprises three modules working in product-line architecture: Text, Image and OCR (Figure 2). Each module performs its task independently and its output is used as an input for the next module. It can be configured on Microsoft Windows platforms following a simple six-step installation process.

Methods

MSL extracts text and figures from the published scientific literature and helps in analyzing embedded text inside figures. The overall methodological implementation and workflow of the MSL is divided into two processes: (I) Text mining and (II) Image analysis. MSL is a desktop application, designed and developed following the scientific software engineering principles of three-layered Butterfly¹⁴ software development model.

Text mining

Physical and logical document analysis is one of the living challenges. To the best of the authors' knowledge, there is no solution available which can perform efficient physical and logical structural analysis of PDF files, implement completely correct rendering order and classify text in all possible categories e.g. Tile, Abstract, Headings, Figure Captions, Table Captions, Equations, References, Headers, Footers etc.

However, there are some tools available which are helping in this regard e.g. PDF2HTML towards contextual modeling of logical labelling¹⁵, PDF-Analyzer for object level document analysis¹⁶, XED for hidden structure analysis², Dolores for the logical structure analysis and recovery¹⁷ automatic conversation from PDF to XML¹⁸ and PDF to HTML¹⁹, Layout-aware²⁰ etc.

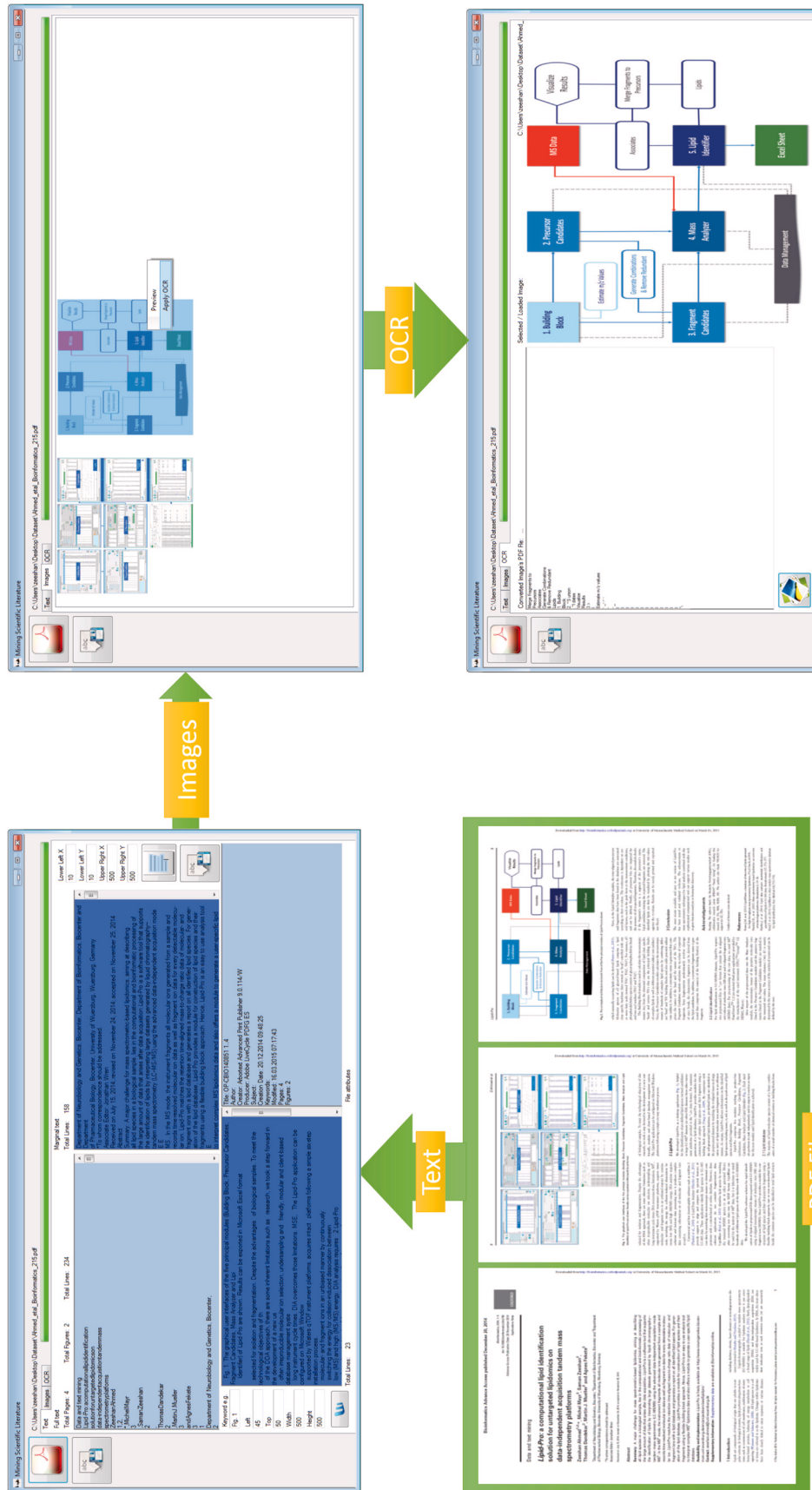


Figure 1. Graphical user interfaces of MSL and modular workflow. This figure shows the graphical user interface and modular workflow of three main components: Text, image and OCR. A PDF document⁴¹ is input and processed by MSL. Text module provides extracted, searched and marginalized text in reading order, and file attributes. Image component provides the preview of extracted images from the document. OCR component provides extracted text from selected and processed image.

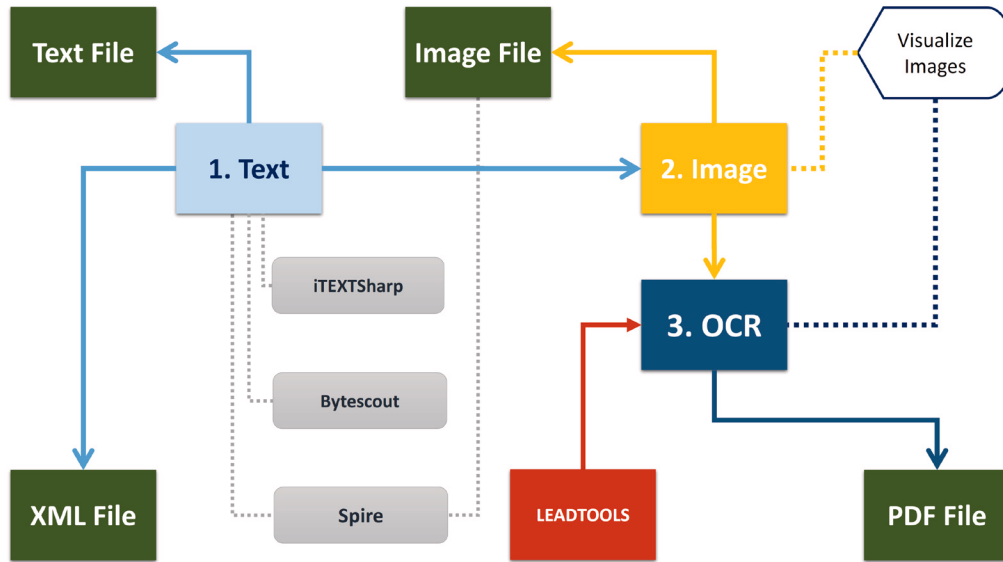


Figure 2. Conceptual architecture of MSL and component's workflow. This figure shows the conceptual architecture of the MSL application, which consists of three main components: Text, Image and OCR, and nine sub-components: Text File, Image File, Visualize Image, PDF File, LEADTOOLS, XML File, iTextSharp, Bytescout, Spire. As figure shows, Text component applies iTextSharp, Bytescout, Spire to extract the text from PDF document and write output in XML file. Image components applies Spire to extract images from the PDF document and visualize that using Visualize Image. OCR component applied LEADTOOLS to extract text from images and export that to PDF format.

We developed MSL's Text module, which is capable of processing PDF files with single, double or multiple columns. It divides the system's text based output in four sub-modules: full text, marginal text, keyword based extracted text and file attributes. Full text gives the complete text from PDF file, marginal allows user to give the coordinates (Lower Left X, Lower Left Y, Upper Right X and Upper Right Y) and extract the desired portion of the text from the PDF file. The keyword based text allows user to extract the information from PDF file based on keywords and respective coordinates (Left, Top, Width, Height) e.g. if a user is only interested in getting the figure caption or references, this kind of search will be helpful. The last sub module, File attributes gives the information about input file including title, author, creator, producer, subject, creation date, keywords, modified, number of pages and number of figures.

While implementing Text module, we researched and tried different available commercial and freely downloadable libraries with a focus on full text extraction, marginal text extraction, keyword based text extraction and text extraction from embedded images from PDF files. We tried different implemented systems and libraries (Table 1) e.g. iTextSharp Bytescout, Spire PDF Sautinsoft PDF Focus Dynamic PDF, PDFBox, iText PDF, QPDF, PoDoFo, Haru PDF Library, JPedal, SVG Imprint, Glance PDF Tool Kit, BCL SharpPDF etc.

One of the common problems in almost all libraries is merging and mixing of text, using double or multiple columns. Our developed system is the combination of different libraries, useful for different purposes. We have used Spire PDF to remove the

Table 1. Systems and Libraries tested for MSL. The table gives the list of different systems and libraries, which have been used for the extraction of text from PDF files.

Library Name	Weblink
iTextSharp	(http://sourceforge.net/projects/itextsharp/),
Bytescout	(https://bytescout.com)
Spire PDF	(http://www.e-iceblue.com/Introduce/pdf-for-net-introduce.html)
Sautinsoft PDF Focus	(http://www.sautinsoft.com/products/pdf-focus/)
Dynamic PDF	(https://www.dynamicpdf.com)
PDFBox	(https://pdfbox.apache.org)
iText PDF	(http://itextpdf.com)
QPDF	(http://qpdf.sourceforge.net)
PoDoFo	(http://podofo.sourceforge.net)
Haru PDF Library	(http://libharu.sourceforge.net)
JPedal	(https://www.idrsolutions.com/jpedal/)
SVG Imprint	(http://svgimprint-windows.software.informer.com)
Glance PDF Tool Kit	(http://www.planetpdf.com/forumarchive/53545.asp)
BCL	(http://www.pdfonline.com/corporate/)
SharpPDF	(http://sharp.pdf.sourceforge.net)

Book-marks, iTextSharp for the extraction of full and marginal text, Bytescout for the keyword based marginalized text search and producing output in the form of XML file (Figure 2). The generated XML file contains structured (tagged) text along with the information about its coordinates (placement in the file), font (Bold, Italic etc.) and size, which can be used for mapping and pattern recognition tasks.

Image processing

Image-based analysis is a versatile and inherently multiplexed approach as it can quantitatively measure biological images to detect those features, which are not easily detectable by a human eye. Millions of figures have been published in scientific literature that includes information about results obtained from different biological and medicinal experiments. Several data and image mining solutions have been already implemented, published and are in use in the last 15 years²². Some of the mainstream approaches are towards the analysis of all kinds of images (flow charts, experimental images, models, geometrical shapes, graphs, images of thing or objects, mixed etc.). There are not many approaches proposed for specific kinds of image-analysis e.g. towards the identification and quantification of cell phenotypes²³, prediction of subcellular localization of proteins in various organism²⁴, analysis of gel diagrams²⁵, mining and integration of pathway diagrams²⁶.

While implementing a new data-mining tool, one of our goals was to extract images from published scientific literature and try to extract embedded text as well. We analyzed different freely available and commercial OCR systems and libraries including Aspose, PUMA, Microsoft OCR, Tesseract, LEADTOOLS, Nicomsoft OCR, MeOCR OCR, OmniPage, ABBYY, Bytescout claiming to be able to extract embedded text from figures. During our research we found LEADTOOLS (Figure 2) as one of the

best available solutions for this purpose. MSL is capable of automatically extracting images from the PDF files and allowing the user to apply OCR to any extracted image by clicking and enlarging it for a better view (using Windows default image viewer).

Results and discussion

We tested MSL with similar parameters on randomly selected scientific manuscripts (ten PDF files) from different open access (*F1000Research*, *Frontiers*, *PLOS*, *Hindawi*, *PeerJ*, *BMC*) and restricted access (*Oxford University Press*, *Springers*, *Emerald*, *Bentham Science*, *ACM*) publishers, including some of the authors' published papers, details are given in Table 2. While testing MSL on the selected manuscripts, we observed best overall performance for the manuscripts²⁷⁻³², with satisfactory results from almost all publishers (including *Oxford University Press*, *BMC*, *Frontiers*, *PeerJ*, *Bentham Science*, *ACM*) in terms of both extracting text in reading order and extracting images. An observed poor performance involved manuscripts from *PLOS*³³, *Hindawi*³⁴, *F1000Research*³⁵ and *IEEE*³⁶ publishers. Here, in the case of text extraction we observed that the text was in reading order when using manuscripts from *F1000Research* and *IEEE* but text was without spaces in the manuscript from *PLOS* and with additional lines and extra spaces in the manuscript from *Hindawi*. In the case of figure extraction we observed one common problem among the four manuscripts from these publishers; along with the manuscript images (Figures), embedded journal or publishers' logos and images were also extracted. Additionally, while analyzing the manuscript from *F1000Research*, we observed that the images were broken into many pieces and it was not possible to find one single complete image. As we did not test all manuscripts from the mentioned publishers, we cannot claim that the results will be the same for all papers from a publisher, as the output may vary in

Table 2. Papers (PDF files) tested using MSL. The table gives the list of 10 of those manuscripts from different publishers, which have been used for testing and validating the MSL application.

Publishers	DOI	Manuscript Title
<i>F1000-Research</i>	10.12688/f1000research.5931.3	Ant-App-DB: a smart solution for monitoring arthropods activities, experimental data management and solar calculations without GPS in behavioral field studies ³⁵ .
<i>PLOS</i>	10.1371/journal.pgen.1006202	The Genomic Aftermath of Hybridization in the Opportunistic Pathogen <i>Candida metapsilosis</i> ³³ .
<i>Hindawi</i>	10.1155/2015/723451	Mathematical Properties of the Hyperbolicity of Circulant Networks ³⁴ .
<i>IEEE</i>	10.1109/IC4.2009.4909215	Design implementation of I-SOAS IPM for advanced product data management ³⁶ .
<i>BMC</i>	10.1186/1471-2105-14-218	Software LS-MIDA for efficient mass isotopomer distribution analysis in metabolic modeling ²⁸ .
<i>PeerJ</i>	10.7717/peerj.1319	Anvi'o: an advanced analysis and visualization platform for 'omics data ²⁹ .
<i>Frontiers</i>	10.3389/fninf.2015.00009	Ontology-based approach for <i>in vivo</i> human connectomics: the medial Brodmann area 6 case study ³⁰ .
<i>ACM</i>	10.1145/1838002.1838065	Intelligent semantic oriented agent based search (I-SOAS) ³¹ .
<i>Bentham Science</i>	10.2174/2213275906666131108211241	DroLIGHT-2: Real Time Embedded and Data Management System for Synchronizing Circadian Clock to the Light-Dark Cycles ³² .
<i>Oxford University Press</i>	10.1093/bioinformatics/btu772	Bioimaging-based detection of mislocalized proteins in human cancers by semi-supervised learning ²⁷ .

different papers. Our observed results using MSL are given in attached supplementary material ([Supplementary Table S1](#) and [Dataset 1](#)).

Dataset 1. Extracted images and text from papers tested using MSL

<http://dx.doi.org/10.5256/f1000research.7329.d108739>

Raw dataset is attached to this manuscript, which categorically provides all images and text in XML format, extracted from manuscripts (from different publishers (included in file names)) using MSL³⁷.

To apply MSL, published scientific literature has first to be downloaded in the form of a PDF file, from any published source. The validation process using MSL consists of three major steps: 1) Text mining, 2) Image extraction, and 3) Application of OCR to extract text from selected images as shown in [Figure 1](#), following the implemented workflow as shown in [Figure 2](#). Example results and graphics are shown in [Figure 1](#), [Figure 3](#) and [Figure 4](#). Representation includes the extraction of text and images from one of the randomly selected papers²¹, and application of OCR to one of the extracted images from another randomly picked publication²⁷.

[Figure 1](#) shows that one randomly selected published article's PDF file²¹ is inputted to the MSL's text, the extracted text is divided into three categories (i) complete text in excellent rendering order (ii) marginalized text and (iii) keyword based searched text. Two figures ([Figure 1](#) and [Figure 2](#)) are extracted and displayed in the image section, and one of those is selected to apply OCR. The applied OCR extracts textual information, which is displayed in and can be exported in a PDF file.

To further validate the application of OCR and discuss different results, [Figure 3](#) show another example of embedded text extraction from a complex figure³⁸, which includes three panels of images (i) colorful pie and circle charts, (ii) biological images and (iii) tabular information. Similar to our prior application of OCR, results are displayed in textual form as well as generated PDF file of extracted text. A noticeable difference between both outputs is that the textual information is presented in line-by-line order whereas in the PDF file the information is displayed in margins with respect to the original image.

The last resultant example is based on the validation of MSL by extracting the textual information from image based PDF files. We produced an image form of one of the randomly selected article²⁷ and then processed one of pages. As [Figure 4](#) shows, the obtained results were comprehensive in both textual as well as the PDF form. This kind of textual extraction can be very helpful, especially when the literature is available in only images e.g. in the case of old published literature in print only format but electronically available in scanned form. MSL produces several files as system output in the parent folder of the files. These files are: XML files (which include structured or tagged information), an Images File (extracted from the PDF file) and PDF files for all analyzed images using OCR. Additionally, extracted text can be split into different IMRAD parts and results can be searched and

categorized e.g. based on abstract, introduction, methods, discussion, conclusion and references.

We mentioned earlier that we have tried and implemented different libraries for text and image extraction and analysis. The best text based outcome was observed using iTextSharp, better image extraction was observed using Spire and OCR from LEADTOOLS was the most promising. While validating the implemented solution, other than the expected results (text and images), we observed some limitations in the used libraries: unexpected and irrelevant images were also extracted e.g. journal, publisher's logos and header-footer images embedded inside document (e.g. images added by the publishers, to provide publishing details), text was not always in good rendering order, especially when there were text-based mathematical equations with super and subscripts; and in case of double or multi-column PDF files, most of the libraries' rendering order is not correct. During extracting text, we found that some important symbols were missed and spaces were generated for some paragraphs. We found that it was not possible to extract particular images that are created as a combination of different sub-images and text objects in the manuscript. In these cases, text is found in extracted text area and all extracted sub-images are image sections, with the possibility of missing some sub-images as well. Moreover, when we applied OCR to different images (extracted or loaded), we found that its performance does vary with respect to the complexity of inputted images. In case of special characters (e.g. Greek delta, alpha, beta etc.), it does not perform well unless these are hard wired in the software.

In comparison to earlier mentioned tools; MSL possess some advantages as well as limitations. For instance, Dolores help user in adding custom tags to the PDF document and create semantic model associated to the processed class of documents, PDF2HTML implements conditional random fields (CRF) based model to learn semantics from processed PDF page's content, PDF-Analyzer devised a model based on rectangular objects for the analysis on PDF documents, XED applies method to combine PDF symbol analysis with traditional document image processing technique. MSL does not apply any of these methods and support such features. However, MSL does support segmentation of text, provides text in correct reading order, enable users with keywords based search and provide extraction of embedded text from figures (using OCR), which none of these tools does. To enhance the functionality of the MSL program (e.g. our standard version available here for download), we give a table of the most often used special symbols in biomedical literature ([Table 3](#)). Depending on your application in mind, you thus simply extend the MSL parser by considering also these special characters occurring often in your texts.

One in-house example is the DrumPID database³⁹, where different types of data and images are warehoused by us and an improved separation and retrieval of text versus figure legends, image descriptions etc. is highly useful and currently applied. The latest version of DrumPID allows understanding and screening of compounds for their effects in protein interaction networks. It is helpful in exploring potential antibiotic lead structures, studying individual pathways and potential targets in various organisms.

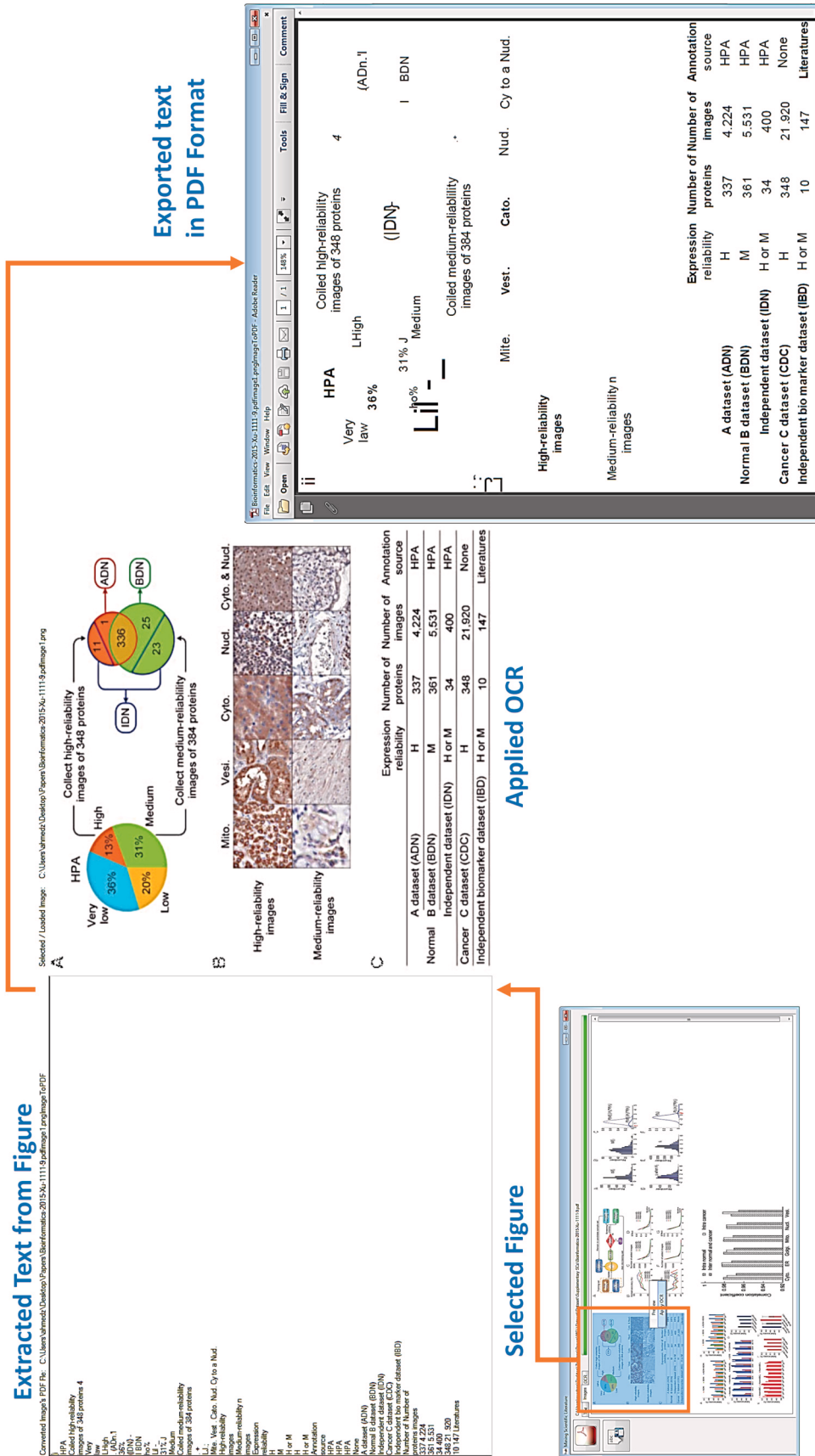
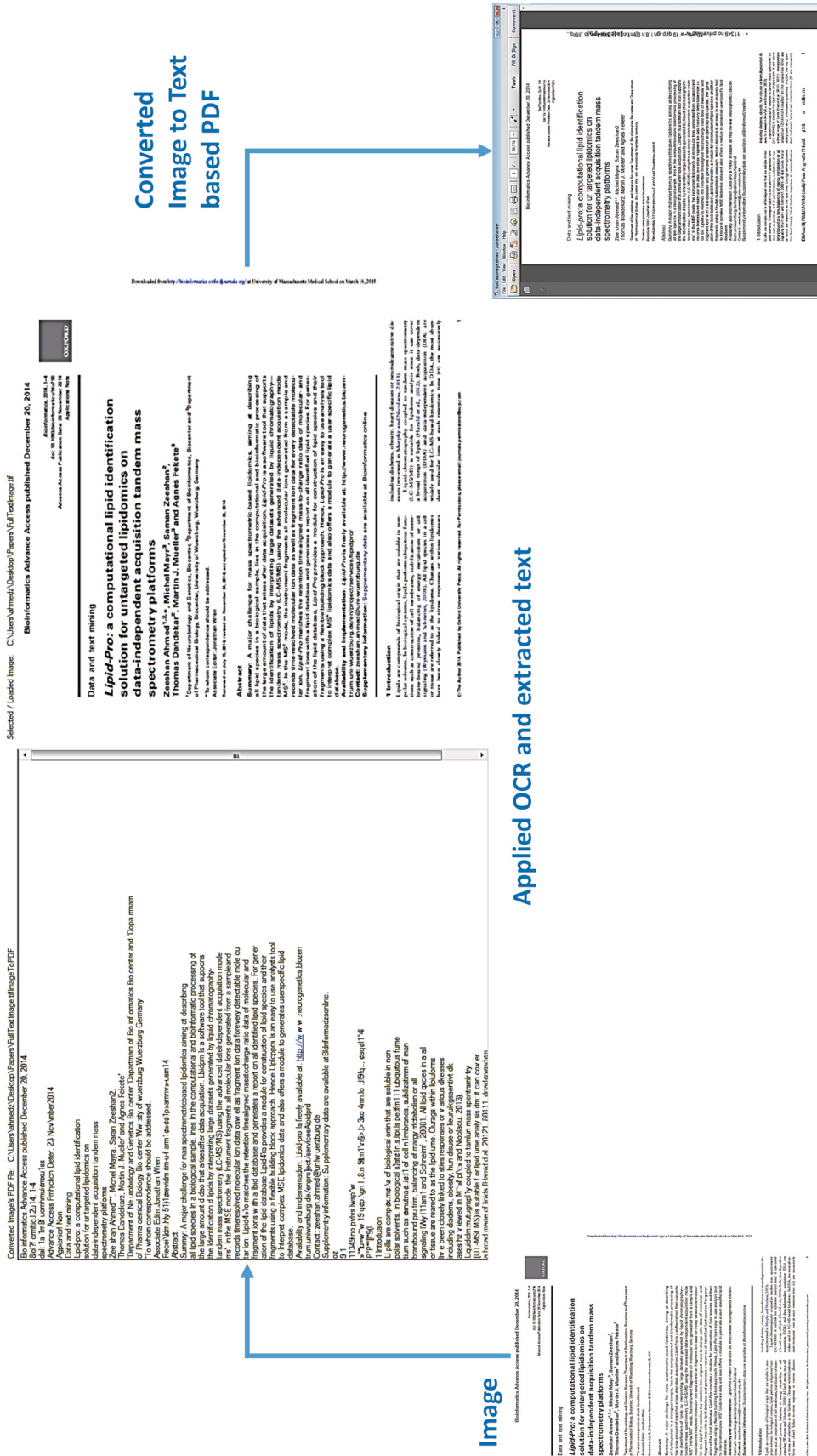


Figure 3. Example: Publication, Figure 1 of (YY et al., 2015). This figure shows document image analysis, text extraction and PDF conversion. A figure (based on three panels; including two charts, one image and a table) is selected from one of the randomly selected papers 2⁸. OCR (LEADTOOLS) is applied to extract and report the text from the figure in simple text form (section: Extracted Text from Figure) and in PDF file with similar margins to the original figure (section: Exported text in PDF format).



We compared (Table 4) MSL with with different, earlier discussed related software applications: A-PDF, PDF Merge Split Extract, BePDF, KPDF, MuPDF, Xpdf tool, Power PDF, PDF2HTML, PDF-Analyzer, XED, Dolores and Layout-aware. The chosen and compared tools were picked based on the following criteria: 1) able to extract text from single, double and multiple column PDF files, 2) able to extract images from from single, double and multiple column PDF files, 3) provide options to search text and image based elements, 4) help in segmentation of text and images, 4) open source, 5) commercial, 6) freely available, 7) easily customizable, and 8) additionally, capable of analyzing embedded images with the application of OCR. There is not mathematical or

statistical ranking of these tools were performed as most of these shared common features and obtained results were close. The reason to draw comparison is not at all to undermine the importance of any other available software applications but tried to justify the importance of MSL, as it's the only one among which is fully open source, capable of extracting text and images from PDF files, applies OCR to get text from extracted chosen images, allow users to search, easily customizable (users can replace libraries) and freely available.

Implementation & operation

MSL architecture is based on the Product Line Architecture (PLA) and Multi-Document Interface (MDI) developmental principles, and it is designed and developed (using C-Sharp programming language, Microsoft Dot NET Framework) following the key principles of *Butterfly* paradigm^{14,38}. The work-flow of MSL is divided into two processes: (I) extraction and marginalization of text with respect to the division and placement of text in PDF file and keyword based search by using the iTextSharp, Bytescout, Spire PDF libraries, and (II) extraction and analysis of figures by using the Spire PDF library and LEADTOOLS OCR.

It takes Portable Document Format (PDF) based literature files as input, performs partial physical structure analysis, and exports output in different formats e.g. text, images and XML files. It allows user to extract keywords and marginal (X and Y coordinates) information based text, have PDF file's metadata information (title, author, creator, producer, subject, creation date, keywords, modified, number of pages and number of figures) and save extracted full and marginal text in text files.

Table 3. Special symbols found in biomedical literature¹.

Number	Special Symbols	Name
1	Δ	Delta
2	α	Alpha
3	β	Beta
4	ϕ	Phi

The table illustrates that special characters occurring most often in the texts of choice enhance further MSL capabilities if incorporated in addition in the parser. This is, however, a text-dependent additional modification of the MSL program.

Table 4. Comparative analysis of MSL with different, earlier discussed related applications. This table compare different related applications (A-PDF, PDF Merge Split Extract, BePDF, KPDF, MuPDF, Xpdf tool, Power PDF, PDF2HTML, PDF-Analyzer, XED, Dolores, Layout-aware) to MSL. Comparison is drawn based on some major elements (Open source, PDF text extraction, PDF image extraction, OCR, Searching, IMRAD, Customizable and FREE).

Feature/Software	Open source	PDF text extraction	PDF image extraction	OCR	Searching	IMRAD	Customizable	FREE
A-PDF	No	Yes	Yes	No	No	No	No	No
PDF Merge Split Extract	No	Yes	Yes	No	No	No	No	No
BePDF	Yes	Yes	No	No	Yes	No	No	No
KPDF	Yes	Yes	No	No	Yes	No	No	Yes
MuPDF	Yes	Yes	Yes	No	Yes	No	No	Yes
Xpdf tool	Yes	Yes	Yes	No	No	No	No	Yes
Power PDF	No	Yes	Yes	No	Yes	No	No	No
PDF2HTML	No	Yes	No	No	No	No	No	No
PDF-Analyzer	No	Yes	No	No	No	No	No	No
XED	No	Yes	Yes	No	Yes	No	No	No
Dolores	No	Yes	Yes	No	Yes	No	No	No
Layout-aware	Yes	Yes	No	No	Yes	No	No	Yes
MSL	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Biomedical image extraction and analysis is one of the most complex tasks from the field of computer sciences and image analysis. Some of the mainstream approaches⁴⁰⁻⁴⁵ have been proposed towards the analysis of all kinds of images (e.g. flow charts, experimental images, models, geometrical shapes, graphs, image-of-thing, mix etc.). MSL allows user to automatically extracting images from the PDF files, let any selected image viewed via Windows default image viewer and apply implemented OCR. Other than extract images from PDF file, MSL allow user to load any image, apply OCR and export output in readable PDF file.

MSL produces several out files in the parent folder including XML files (which include structured or tagged information), Images File (extracted from PDF file) and PDF files for all analyzed images using OCR (Figure 5).

MSL application is very simple to install and use. It was tested and can be well configured on a Microsoft Windows platform (preferred OS version: 7). MSL follows a simple six steps installation process (Figure 6). After installation, it can be run by either clicking on the installed application's icon at the desktop or execute application following sequence of steps: Start → All Programs → MSL 1.0.0 → MSL.

Regarding using the MSL application, one important point to remember is that it is based on different PDF text extraction, marginalization and figure extraction libraries, which are automatically configured during installation but used OCR by the LEAD-TOOLS is not a freely available library, which we have used upon academic research (free) license. The OCR library is also automatically configured during installation but its performance at

different (non-licensed) machines is not confirmed. Moreover, the recommended display screen resolution size is 1680×1050 with landscape orientation.

Conclusions

As most of the publishers are publishing in simple HTML and PDF formats, its not possible to segment and analyze raw form literature using available commercial and open-source software applications, as those are helpful in mainly text and image extraction. It will be helpful to have literature available in semantically tagged formats (e.g. XML, OWL etc.), so literature can be efficiently parsed, categorized and searched.

The development of a virtual research environment to store and link molecular data, can be well achieved and established if first the mixture of text, protocols and omics data is properly separated from images, figures and figure legends – a task for which our tool can be well suited. There are a number of databases (e.g. *Alzheimer's Disease Neuroimaging Initiative (ADNI)*; *Breast Cancer Digital Repository (BCDR)*; *BiMed*; *Public Image Databases*; *Cancer Image Database (caIMAGE)*; *Collaborative Informatics and Neuroimaging Suite (COINS)*; *DrumPID*; *Digital Database for Screening Mammography (DDSM)*; *Electron Microscopy Data Bank (EMDB)*; *LONI image data archive*; *Mammography Image Databases (MID)*; *New Database Provides Millions of Biomedical Images*; *Open Access Series of Imaging Studies (OASIS)*; *Stanford Tissue Microarray Database (TMA)*; *STRING*; *The Cancer Imaging Archive (TCIA)*; *Whitney Imaging Center etc.*) which can directly profit from MSL by fast, automatic and rapid separation of text and text description from images and figure legends describing

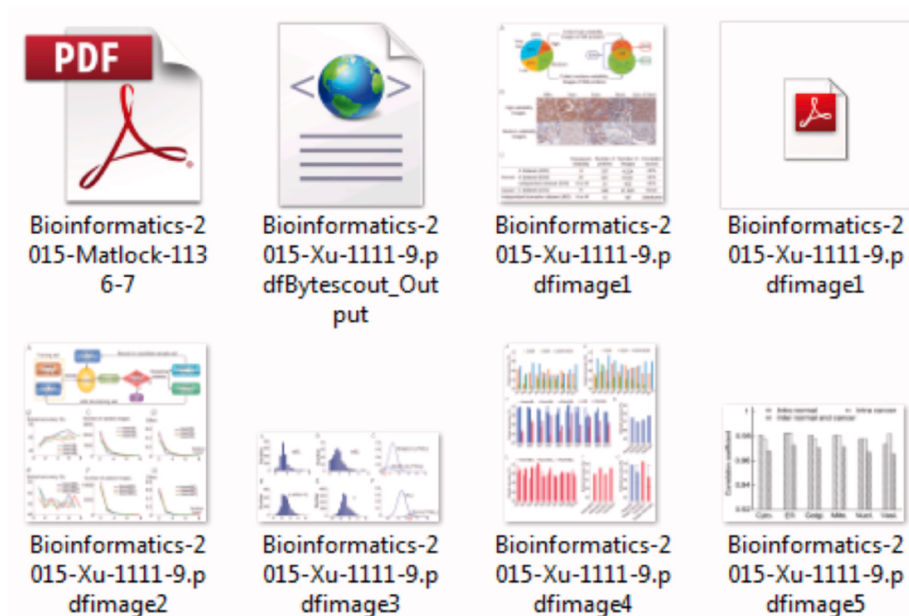
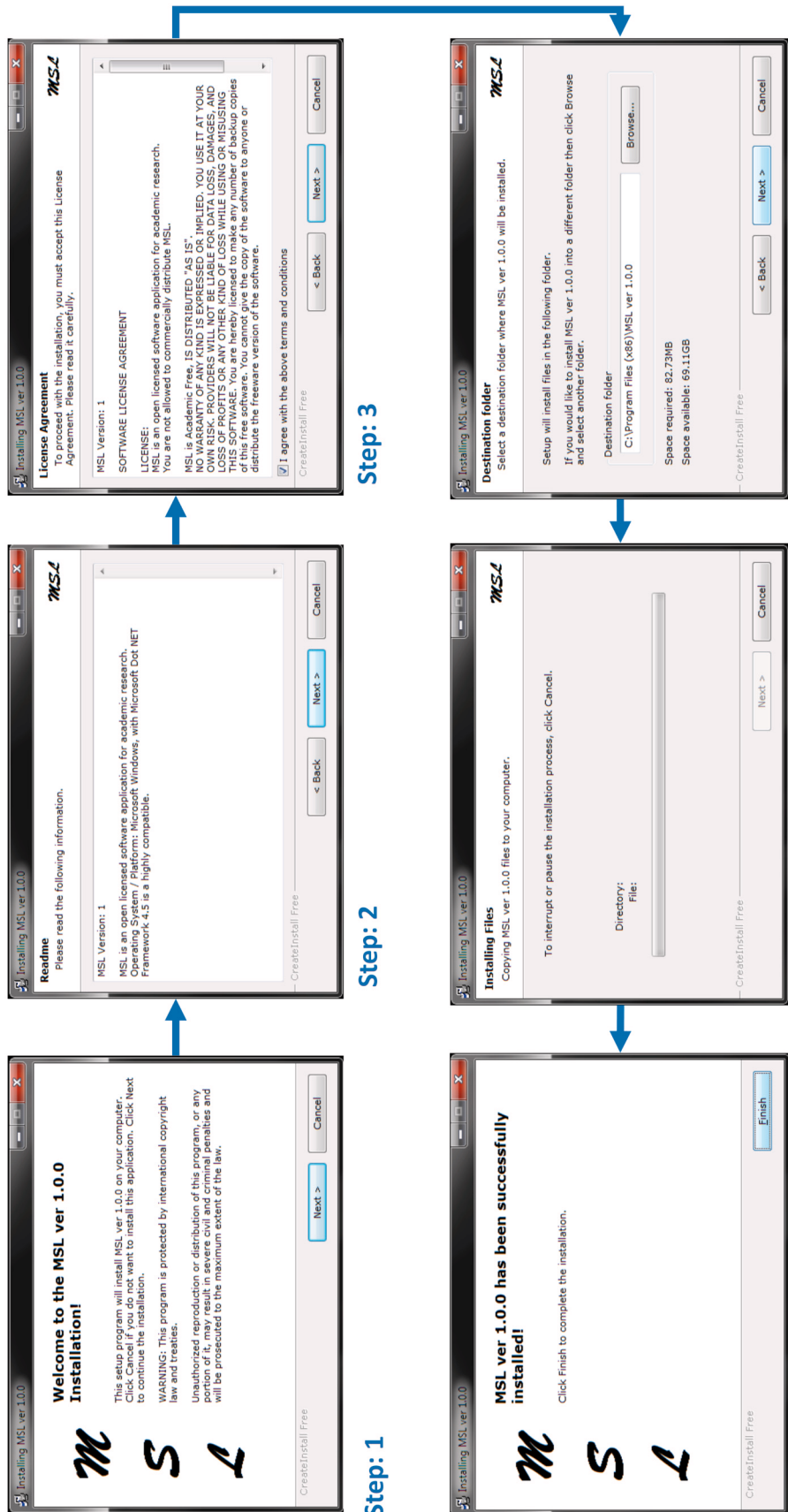


Figure 5. Screenshot of the all extracted images and generated files (XML and PDF). This figure shows different files generated during analysis of PDF document. PDF file (top, left) is the actual document, XML file is the structured (tagged) form of extracted text, second PDF file (top, right) is the extracted text from image (see Figure 3) and all other files are extracted image from PDF document.



Step: 4

Step: 5

Step: 6

Step: 3

Step: 2

Step: 1

Figure 6. MSL six steps installation process.

the images is important for further improvement of the database and its content.

The latest available and easy to use version of MSL has been tested and validated in-house. The advancements in information retrieval techniques for text and figure analysis combined with this sophisticated computational tool can support various studies.

Data availability

F1000Research: Dataset 1. Extracted Images and Text from Papers tested using MSL, [10.5256/f1000research.7329.d108739](https://doi.org/10.5256/f1000research.7329.d108739)³⁷

Software availability

Software access

The software executable is freely available at the following web link: <https://zenodo.org/record/30941#.Vi0PtmC5LHM>

The software download section provides one executable: MSL, setup to be installed on the Microsoft Windows platform.

MSL has been NOT been developed for any commercial purposes but as a non-commercial prototype application for academic research, analysis and development purposes.

Archived software files as at the time of publication

Mining Scientific Literature (MSL) Ver 1.0.0 (DOI: [10.5281/zenodo.30941](https://doi.org/10.5281/zenodo.30941)).

License

All associated files are licensed under the [Academic Free License 3.0 \(AFL 3.0\)](https://creativecommons.org/licenses/by/3.0/).

Supplementary material

Supplementary Table S1. List of Papers (PDF files) tested using MSL.

Supplementary which gives the list of some of those manuscripts from different publishers (*F1000Research*, *PLOS*, *Hindawi*, *IEEE*, *BMC*, *PeerJ*, *Frontiers*, *ACM*, *Bentham Science* and *Oxford University Press*), which have been used for testing and validating the MSL application. The attached table provides the information about some of the extracted images and observed full and marginal text.

[Click here to access the data.](#)

References

- Hunter L, Cohen KB: **Biomedical language processing: what's beyond PubMed?** *Mol Cell*. 2006; **21**(5): 589–594.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hadjar K, Rigamonti M, Lalanne D, et al.: **Xed: A New Tool for Extracting Hidden Structures from Electronic Documents.** In *International Workshop on Document Image Analysis for Libraries*. 2004; 221–224.
[Publisher Full Text](#)
- Sayers EW, Barrett T, Benson DA, et al.: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res*. 2010; **38**(Database issue): D5–16.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- States DJ, Ade AS, Wright ZC, et al.: **MiSearch adaptive PubMed search tool.** *Bioinformatics*. 2009; **25**(7): 974–76.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Poulter GL, Rubin DL, Altman RB, et al.: **MScanner: a classifier for retrieving Medline citations.** *BMC Bioinformatics*. 2008; **9**(1): 108.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Plikus MV, Zhang Z, Chuong CM: **PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm.** *BMC Bioinformatics*. 2006; **7**: 424.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Smalheiser NR, Zhou W, Torvik VI, et al.: **Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results.** *J Biomed Discov Collab*. 2008; **3**: 2.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Doms A, Schroeder M: **GoPubMed: exploring PubMed with the Gene Ontology.** *Nucleic Acids Res*. 2005; **33**(Web Server issue): W783–86.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Author contributions

ZA: developed the complete solution (including research, software designing, programming, testing, deployment and technical documentation). TD guided the study. All authors participated in writing of the manuscript and approved the final manuscript for publication.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by a German Research Foundation grant (DFG-TR34/Z1) to TD.

I confirm that the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We thank the German Research Foundation (DFG-TR34/Z1) for support. We would like to thank Ahmed lab, Genetics and Genome Sciences, Institute for Systems Genomics, School of Medicine, University of Connecticut Health Center USA, and Department of Bioinformatics, Biocenter, University of Wuerzburg Germany.

We would like to thank Dr. Chunguang Liang (University of Wuerzburg, Germany) for his help in testing MSL and all interested colleagues for critical community input on the approach and anonymous reviewers for their helpful comments.

We would like to thank all the open source, licensed and commercial library providers, for their help in this non-commercial and academic research and software development.

9. Kim JJ, Pezik P, Rebholz-Schuhmann D: **MedEvi: retrieving textual evidence of relations between biomedical concepts from Medline.** *Bioinformatics.* 2008; **24**(11): 1410–12.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Rebholz-Schuhmann D, Kirsch H, Arregui M, et al.: **EBIMed—text crunching to gather facts for proteins from Medline.** *Bioinformatics.* 2007; **23**(2): e237–44.
[PubMed Abstract](#) | [Publisher Full Text](#)
11. Douglas SM, Montelione GT, Gerstein M, et al.: **PubNet: a flexible system for visualizing literature derived networks.** *Genome Biol.* 2005; **6**(9): R80.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Eaton AD: **HubMed: a web-based biomedical literature search interface.** *Nucleic Acids Res.* 2006; **34**(Web Server issue): W745–47.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. Hearst MA, Divoli A, Guturu H, et al.: **BioText Search Engine: beyond abstract search.** *Bioinformatics.* 2007; **23**(16): 2196–97.
[PubMed Abstract](#) | [Publisher Full Text](#)
14. Ahmed Z, Zeeshan S, Dandekar T: **Developing sustainable software solutions for bioinformatics by the “Butterfly” paradigm [version 1; referees: 2 approved with reservations].** *F1000Res.* 2014; **3**: 71.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Tao X, Tang Z, Xu C: **Contextual Modeling for Logical Labeling of PDF Documents.** *Comput Electr Eng.* 2014; **40**(4): 1363–75.
[Publisher Full Text](#)
16. Hassan T: **Object-Level Document Analysis of PDF Files.** In *Proceedings of the 9th ACM symposium on Document engineering.* 2009; 47–55.
[Publisher Full Text](#)
17. Bloechle JL, Rigamonti M, Ingold R: **OCD Dolores - Recovering Logical Structures for Dummies.** In *10th IAPR International Workshop on Document Analysis Systems (DAS).* 2012; 245–249.
[Publisher Full Text](#)
18. Déjean H, Meunier JL: **A System for Converting PDF Documents into Structured XML Format.** In *Proceedings of the 7th international conference on Document Analysis Systems.* 2006; 129–140.
[Publisher Full Text](#)
19. Rahman F, Alam H: **Conversion of PDF Documents into HTML: A Case Study of Document Image Analysis.** In *Proceedings of Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers.* 2003; **1**: 87–91.
[Publisher Full Text](#)
20. Ramakrishnan C, Patnia A, Hovy E, et al.: **Layout-aware text extraction from full-text PDF of scientific articles.** *Source Code Biol Med.* 2012; **7**(1): 7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
21. Ahmed Z, Mayr M, Zeeshan S, et al.: **Lipid-Pro: a computational lipid identification solution for untargeted lipidomics on data-independent acquisition tandem mass spectrometry platforms.** *Bioinformatics.* 2015; **31**(7): 1150–1153.
[PubMed Abstract](#) | [Publisher Full Text](#)
22. Zweigenbaum P, Demner-Fushman D, Yu H, et al.: **Frontiers of biomedical text mining: current progress.** *Brief Bioinform.* 2007; **8**(5): 358–375.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
23. Carpenter AE, Jones TR, Lamprecht MR, et al.: **CellProfiler: image analysis software for identifying and quantifying cell phenotypes.** *Genome Biol.* 2006; **7**(10): R100.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. Chou KC, Shen HB: **Cell-PLOC: a package of Web servers for predicting subcellular localization of proteins in various organisms.** *Nat Protoc.* 2008; **3**(2): 153–162.
[PubMed Abstract](#) | [Publisher Full Text](#)
25. Kuhn T, Nagy ML, Luong T, et al.: **Mining images in biomedical publications: Detection and analysis of gel diagrams.** *J Biomed Semantics.* 2014; **5**(1): 10.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
26. Kozhenkov S, Baitaluk M: **Mining and integration of pathway diagrams from imaging data.** *Bioinformatics.* 2012; **28**(5): 739–742.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
27. Xu YY, Yang F, Zhang Y, et al.: **Bioimaging-based detection of mislocalized proteins in human cancers by semi-supervised learning.** *Bioinformatics.* 2015; **31**(7): 1111–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
28. Ahmed Z, Zeeshan S, Huber C, et al.: **Software LS-MIDA for efficient mass isotopomer distribution analysis in metabolic modelling.** *BMC Bioinformatics.* 2013; **14**: 218.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. Eren AM, Esen ÖC, Quince C, et al.: **Anvi’o: an advanced analysis and visualization platform for ‘omics data.** *PeerJ.* 2015; **3**: e1319.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
30. Moreau T, Gibaud B: **Ontology-based approach for in vivo human connectomics: the medial Brodmann area 6 case study.** *Front Neuroinform.* 2015; **9**: 9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
31. Ahmed Z: **Intelligent semantic oriented agent based search (I-SOAS).** In *Proceedings of the 7th International Conference on Frontiers of Information Technology.* 2009.
[Publisher Full Text](#)
32. Ahmed Z, Helfrich-Förster C: **DroLIGHT-2: Real Time Embedded and Data Management System for Synchronizing Circadian Clock to the Light-Dark Cycles.** *Recent Patents on Computer Sci.* 2013; **6**(3): 191–205.
[Publisher Full Text](#)
33. Prysycz LP, Németh T, Saus E, et al.: **The Genomic Aftermath of Hybridization in the Opportunistic Pathogen *Candida metapsilosis*.** *PLoS Genet.* 2015; **11**(10): e1005626.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
34. Hernández JC, Rodríguez JM, Sigarreta JM: **Mathematical Properties of the Hyperbolicity of Circulant Networks.** *Adv Math Phys.* 2015; **2015**: 723451.
[Publisher Full Text](#)
35. Ahmed Z, Zeeshan S, Fleischmann P, et al.: **Ant-App-DB: a smart solution for monitoring arthropods activities, experimental data management and solar calculations without GPS in behavioral field studies [version 2; referees: 1 approved, 2 approved with reservations].** *F1000Res.* 2015; **3**: 311.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
36. Zeeshan A, Detlef G: **Design implementation of I-SOAS IPM for advanced product data management.** In *IEEE 2nd International Conference on Computer, Control and Communication.* 2009; 1–5.
[Publisher Full Text](#)
37. Ahmed Z, Dandekar T: **Dataset 1 in: MSL: Facilitating automatic and physical analysis of published scientific literature in PDF format.** *F1000Research.* 2015.
[Data Source](#)
38. Ahmed Z, Zeeshan S: **Cultivating Software Solutions Development in the Scientific Academia.** *Recent Patents on Computer Sci.* 2014; **7**(1): 54–66.
[Publisher Full Text](#)
39. Kunz M, Liang C, Nilla S, et al.: **The drug-minded protein interaction database (DrumPID) for efficient target analysis and drug development.** *Database (Oxford).* 2016; **2016**: pii: baw041.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
40. Schindelin J, Arganda-Carreras I, Frise E, et al.: **Fiji: an open-source platform for biological-image analysis.** *Nat Methods.* 2012; **9**(7): 676–82.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
41. Schmid B, Schindelin J, Cardona A, et al.: **A high-level 3D visualization API for Java and ImageJ.** *BMC Bioinformatics.* 2010; **11**: 274.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
42. Schneider CA, Rasband WS, Eliceiri KW: **NIH Image to ImageJ: 25 years of image analysis.** *Nat Methods.* 2012; **9**(7): 671–75.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
43. Peng H, Ruan Z, Long F, et al.: **V3D enables real-time 3D visualization and quantitative analysis of large-scale biological image data sets.** *Nat Biotechnol.* 2010; **28**(4): 348–53.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
44. Lopez LD, Yu J, Arighi C, et al.: **A framework for biomedical figure segmentation towards image-based document retrieval.** *BMC Syst Biol.* 2013; **7**(Suppl 4): S8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
45. Sheng J, Xu S, Deng W, et al.: **Novel Image Features for Categorizing Biomedical Images.** In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* 2012.
[Publisher Full Text](#)

Open Peer Review

Current Referee Status:



Version 3

Referee Report 22 May 2018

doi:10.5256/f1000research.15700.r32753



M. Julius Hossain 

Cell Biology and Biophysics Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

The manuscript has been moderately improved by last two revisions. My main concern was that the work was not significantly new although it was claimed so. Authors tried to tone down their claim of novelty during the revisions. Authors added more comparisons and improved the readability of figures. While I am not fully convinced on the current version of the manuscript, I believe that the tool still could be useful for the community as it combines several features together.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 23 April 2018

doi:10.5256/f1000research.15700.r32754



Karin Verspoor  1,2

¹ School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC, Australia

² Health and Biomedical Informatics Centre, The University of Melbourne, Melbourne, VIC, Australia

Unfortunately I remain unconvinced about the contribution of the work presented in this manuscript. The results are presented through case studies rather than formal evaluation.

Claims are made relating to various capabilities of the system that are not justified; for instance, the statement that the tool can split the text into "various IMRAD parts" is neither explained (in terms of Methodology -- does this simply involve searching for specific key words such as "Abstract" or "Introduction"?) nor directly evaluated. There are no quantitative evaluations of performance that support the claims that are made, only statements that the tool has been "validated in-house". This does not constitute a scientific evaluation of the tool.

Closer inspection of the Text box of Figure 1 suggests that the text is not in natural reading order; I see several lines from separate columns juxtaposed. The XML file produced for this text retains the columnar structure and shows

The Lipid-Pro application can be configured on Microsoft Windows

immediately followed by

supported by Waters qTOF instrument platforms, acquires intact

in the next column field without any straightforward way to extract the linear order of the text; the second sentence here clearly does not follow on from the first. I don't see how this format would be useful at all to anyone wanting to do natural language processing on this text, as the authors state. (NB: information retrieval and natural language processing are not the same and the textual representation assumed for NLP is very different than for IR; it is very important to retain the natural text flow and sentence structure for NLP tasks.)

The premise of the work is still not adequately justified; the journals examined no longer (if they ever did; I doubt PLoS or PeerJ, for instance, ever did) publish work as images requiring OCR of the whole document. Lumping HTML and PDF together (as "simple formats") is clearly inappropriate; they have very different characteristics and indeed in HTML the text is nearly always in straightforward linear structure. However, not all PDFs are created equal and most PDFs have perfectly extractable (non-image/non-scanned) text these days through tools such as PDFbox. The authors have not responded to my prior comment that many publishers are already making documents available as nicely structured XML, in a format far preferable to what the authors are producing, due to richer provided semantic structure.

The extraction and analysis of images themselves may be useful; doing OCR of text in an included image could be valuable. The authors claim that this is unique to their tool. It may be, and this may be the most useful part of their tool. However it is unclear how well the tool does this; a close look at Figure 1 identifies some clear errors, e.g. "Coiled" rather than "Collect"; "Nud" rather than "Nucl"; etc. What level of error is tolerable?

Competing Interests: No competing interests were disclosed.

Referee Expertise: Biomedical natural language processing, text mining

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 11 April 2018

doi:10.5256/f1000research.15700.r32751



Juilee Thakar

Department of Microbiology and Immunology, University of Rochester Medical Center, Rochester, NY, USA

I have no further comments.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Referee Report 22 September 2017

doi:10.5256/f1000research.12318.r24964



Karin Verspoor  ^{1,2}

¹ School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC, Australia

² Health and Biomedical Informatics Centre, The University of Melbourne, Melbourne, VIC, Australia

This manuscript addresses the analysis of scientific articles published in PDF form, and introduces a software tool that essentially wraps a number of other tools to build a single tool that integrates a number of features.

There is no novel methodological contribution in the tool itself that I could determine; the value is primarily in integrating the other tools into a user-facing tool. However, the value of a user-facing tool (as opposed to an automated tool that proceeds without human input) is not made clear or evaluated (are users happy to do the work expected of them?).

As a very high level point on motivation, many scientific articles -- particularly in the open access literature which form the majority of the studied papers -- are already available as raw XML and it does not make sense to me to try to parse the PDF to infer structure (in a way that first requires user interaction, and second may introduce errors) that can be read directly from the publisher-produced XML. The PubMed Central repository is a repository of full-text, available online as structured HTML, and the open access collection which is downloadable in its entirety uses XML (<https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>). So in my opinion several of the key objectives of this tool are not needed for a large and growing proportion of the scientific literature. The argument could be made that older articles and articles from certain publishers are not available as "raw" XML, but the author neither make this argument nor specifically test this question. In addition, this XML resource provides a fantastic opportunity to automatically test (some aspects of) the performance of the authors tool. This is not done. The authors should consider the contribution of their tool in the context of such resources. The Introduction focuses on PubMed, but PubMed does not claim to search over the full-text literature (it is limited to abstracts by design), and PubMed Central is not mentioned here.

There is one more tool that I am aware of -- and I suspect there are others -- that the authors do not mention; the Layout-Aware PDF tool ¹. In general, I'd be interested to understand more deeply how the authors' tool differs from the various tools they mention, and why they selected the tools they did for their final system.

The Results presented by the authors are not framed in any well-defined evaluation framework. While they mention "best overall performance" no indication of either specific quantitative or qualitative criteria used to assess the tool have been provided. What were the criteria that were used? Were there guidelines for how to "score" systems on various criteria? What is presented is largely ad hoc qualitative observations. On what basis was performance of the tool ranked?

Comments on presentation:

It is poor practice to cite papers that are purely studied as artifacts -- a citation implies that you are referencing scientific content which you are not. Please remove citations to the papers that were used to test the system; they should be listed (preferably with DOIs) in Table 2, with a paper ID, and the paper IDs should be referenced in the article.

There are a number of phrases in the manuscript that are awkward. "marginalization" does not mean "to find margins". Queries in PubMed are not "NLP quer[ies]" but rather *user* queries processed by an IR system (NB: arguably, IR systems don't even use NLP). The authors talk about "orthodox NLP approaches"; I have no idea what makes an NLP approach "orthodox". I suspect the authors mean "pipeline" rather than "product line". What is an "inherently multiplexed approach"?

The long list of databases in the Conclusions doesn't add much to the manuscript; perhaps the example of the DrumPID should be pulled into a discussion section, and elaborated to clearly demonstrate the practical application and value of the tool.

References

1. Ramakrishnan C, Patnia A, Hovy E, Burns GA: Layout-aware text extraction from full-text PDF of scientific articles. *Source Code Biol Med.* 2012; 7 (1): 7 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

No

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 29 Mar 2018

Zeeshan Ahmed, University of Connecticut Health Center, USA

Reply: Thank you so much for your recommendations.

This manuscript addresses the analysis of scientific articles published in PDF form, and introduces a software tool that essentially wraps a number of other tools to build a single tool that integrates a number of features. There is no novel methodological contribution in the tool itself that I could determine; the value is primarily in integrating the other tools into a user-facing tool. However, the value of a user-facing tool (as opposed to an automated tool that proceeds without human input) is not made clear or evaluated (are users happy to do the work expected of them?). As a very high level point on motivation, many scientific articles -- particularly in the open access literature which form the majority of the studied papers -- are already available as raw XML and it does not make sense to me to try to parse the PDF to infer structure (in a way that first requires user interaction, and second may introduce errors) that can be read directly from the publisher-produced XML. The PubMed Central repository is a repository of full-text, available online as structured HTML, and the open access collection which is downloadable in its entirety uses XML (<https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>).

Reply: Thank you so much for your views.

So in my opinion several of the key objectives of this tool are not needed for a large and growing proportion of the scientific literature. The argument could be made that older articles and articles from certain publishers are not available as "raw" XML, but the author neither make this argument nor specifically test this question. In addition, this XML resource provides a fantastic opportunity to automatically test (some aspects of) the performance of the authors tool. This is not done. The authors should consider the contribution of their tool in the context of such resources. The Introduction focuses on PubMed, but PubMed does not claim to search over the full-text literature (it is limited to abstracts by design), and PubMed Central is not mentioned here.

Reply: We thank you so much for nice suggestion and have tried to revise manuscript accordingly.

There is one more tool that I am aware of -- and I suspect there are others -- that the authors do not mention; the Layout-Aware PDF tool 1. In general, I'd be interested to understand more deeply how the authors' tool differs from the various tools they mention, and why they selected the tools they did for their final system.

Reply: We thank you so much for nice suggestion and have tried to revise manuscript accordingly.

The Results presented by the authors are not framed in any well-defined evaluation framework. While they mention "best overall performance" no indication of either specific quantitative or qualitative criteria used to assess the tool have been provided. What were the criteria that were used? Were there guidelines for how to "score" systems on various criteria? What is presented is largely ad hoc qualitative observations. On what basis was performance of the tool ranked?

Reply: We thank you so much for nice suggestion and have tried to answer question in revision.

It is poor practice to cite papers that are purely studied as artifacts -- a citation implies that you are referencing scientific content which you are not. Please remove citations to the papers that were used to test the system; they should be listed (preferably with DOIs) in Table 2, with a paper ID, and the paper IDs should be referenced in the article.

Reply: We thank you so much for nice suggestion and have added DOIs.

There are a number of phrases in the manuscript that are awkward. "marginalization" does not mean "to find margins". Queries in PubMed are not "NLP quer[ies]" but rather user queries processed by an IR system (NB: arguably, IR systems don't even use NLP). The authors talk about "orthodox NLP approaches"; I have no idea what makes an NLP approach "orthodox". I suspect the authors mean "pipeline" rather than "product line". What is an "inherently multiplexed approach"?

Reply: We thank you so much for nice suggestion and have tried to revise manuscript accordingly.

The long list of databases in the Conclusions doesn't add much to the manuscript; perhaps the example of the DrumPID should be pulled into a discussion section, and elaborated to clearly demonstrate the practical application and value of the tool.

Reply: We thank you so much for nice suggestion and have added more details.

We thank you so much for your time and excellent suggestions, which have helped us in improving the manuscript.

With best wishes,
Authors

Competing Interests: No Competing Interests

Referee Report 06 September 2017

doi:[10.5256/f1000research.12318.r24952](https://doi.org/10.5256/f1000research.12318.r24952)



Florencio Pazos

Computational Systems Biology Group, National Center for Biotechnology (CNB-CSIC) (Spanish National Research Council), Madrid, Spain

The authors present a system for extracting the main text, the images, and the text within those (labels, etc), from scientific papers in PDF format. The system is implemented in an interactive desktop application for Windows and tested in 10 papers.

Extracting the large amounts of scientific and medical information stored in an unstructured way is an important challenge. Any effort in that direction, such as that presented in this work, is of potential interest.

My main concern with this work is that most of the features described are already available in existing software, even those described as "new". For example, tools like "pdftotext" can parse multi-column PDFs, "pdfimages" extract the images within a PDF file, "OCRFeeder" detects the image elements and extract the texts, tables, etc, ... On the contrary, some potentially newer features of MSL, such as recognizing the article parts (Abstract, References, ... -page 3-) mentioned in Methods are not described

later (Results). Problems of other tools mentioned in the Introduction, such as “removing irrelevant graphics” are not solved by MSL either.

I think the comparison with other tools should be presented in terms of what these fail to detect while MSL does not, and vice versa. E.g. For a particular article “pdftotext” was not able to recognize the columns while MSL does, etc.

In summary, better putting this system into the context of existing ones.

The system is implemented as an interactive tool intended for analyzing a single article at a time, even presenting some of the final results (e.g. text extracted from images) back as PDF. For a single article (or a small number) human inspection will be better than any automated system. I guess the real potential of this system is in the automated parsing of large collections of articles. In my opinion the authors should focus the manuscript more on that.

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 29 Mar 2018

Zeeshan Ahmed, University of Connecticut Health Center, USA

Reply: Thank you so much for your recommendations.

The authors present a system for extracting the main text, the images, and the text within those (labels, etc), from scientific papers in PDF format. The system is implemented in an interactive desktop application for Windows and tested in 10 papers. Extracting the large amounts of scientific and medical information stored in an unstructured way is an important challenge. Any effort in that

direction, such as that presented in this work, is of potential interest.

Reply: Thank you so much for your views and we agree with you.

My main concern with this work is that most of the features described are already available in existing software, even those described as “new”. For example, tools like “pdftotext” can parse multi-column PDFs, “pdfimages” extract the images within a PDF file, “OCRFeeder” detects the image elements and extract the texts, tables, etc, ...

Reply: Thank you so much and we respect your points that many competing applications do exist but that still doesn't not negatively impact our contributions.

On the contrary, some potentially newer features of MSL, such as recognizing the article parts (Abstract, References, ... -page 3-) mentioned in Methods are not described later (Results).

Reply: We thank for nice suggestion and have tried to revise manuscript accordingly.

Problems of other tools mentioned in the Introduction, such as “removing irrelevant graphics” are not solved by MSL either.

Reply: Thank you so much and we respect your point that it's not a complete solution, as future research and development is also recommended but still helps at some good levels.

I think the comparison with other tools should be presented in terms of what these fail to detect while MSL does not, and vice versa. E.g. For a particular article “pdftotext” was not able to recognize the columns while MSL does, etc.

Reply: We thank for nice suggestion and have added comparison.

In summary, better putting this system into the context of existing ones.

Reply: We thank for nice suggestion and have tried to revise manuscript accordingly.

The system is implemented as an interactive tool intended for analyzing a single article at a time, even presenting some of the final results (e.g. text extracted from images) back as PDF. For a single article (or a small number) human inspection will be better than any automated system. I guess the real potential of this system is in the automated parsing of large collections of articles. In my opinion the authors should focus the manuscript more on that.

Reply: We agree with you and thank so much for nice suggestion. We mainly tried to focus on that part but as one of the examples to show the strength of system, we have included that.

We thank you so much for your time and excellent suggestions, which have helped us in improving the manuscript.

With best wishes,
Authors

Competing Interests: No Competing Interests

Version 1

Referee Report 09 August 2016

doi:10.5256/f1000research.7898.r14625

**M. Julius Hossain**

Cell Biology and Biophysics Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

In this manuscript authors presented a computational tool that extracts text and images from PDF files. In general the manuscript is interesting considering that it can analyze various types of PDF files from different scientific areas based on the keywords and coordinates. However, it lacks technical novelty over the published literatures and needs additional input on the image analysis section before indexing.

Extraction of texts and images from scientific publications has been presented in various domains: computer science¹, biomedical²⁻⁴, chemistry⁵, proteomics⁶ and so on. The manuscript by Zeeshan Ahmed and Thomas Dandekar presents an incremental innovation without providing clear technological advancement in the field. The objective of performing both physical and logical structure analysis of all kinds of PDF files as mentioned in the manuscript has not been sufficiently supported by technological contribution described in Methods section.

The image processing section the manuscript has been very brief. It does not provide any advanced image analysis technique as mentioned in the abstract. Authors should mention how exactly segmentation of figures and labels are performed and how they are represented to make logical connection between different entities in order to perform further analysis and customized visualization.

The framework has been tested with a very small set of PDF files and no qualitative/quantitative result reporting the accuracy with respect to manually annotated files was presented. It would be good to increase the number test files and include the results of qualitative/quantitative analysis.

Some of the figures (Figures 1, 3, 4 and 6) in the manuscript are hard to see the details in both online and print format. These figures could be reformatted.

References

1. Clark C, Divvala S: Looking beyond text: Extracting figures, tables and captions from computer science papers. *AAAI 2015 Workshop on Scholarly Big Data*. 2015. [Reference Source](#)
2. Lopez LD, Yu J, Arighi CN, Huang H, Shatkay H, Wu C: An Automatic System for Extracting Figures and Captions in Biomedical PDF Documents. *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*. 2011. 578-581 [Publisher Full Text](#) | [Reference Source](#)
3. Lopez LD, Yu J, Tudor CO, Arighi CN, Hongzhan H, Vijay-Shanker K, Wu K: Robust segmentation of biomedical figures for image-based document retrieval. *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*. 2012. 1-6 [Publisher Full Text](#) | [Reference Source](#)
4. Lopez LD, Yu J, Arighi C, Tudor CO, Torii M, Huang H, Vijay-Shanker K, Wu C: A framework for biomedical figure segmentation towards image-based document retrieval. *BMC Syst Biol*. 2013; **7 Suppl 4**: S8 [PubMed Abstract](#) | [Publisher Full Text](#)
5. Choudhury SR, Mitra P, Kirk A, Szep S, Pellegrino D, Jones S, Giles CL: Figure Metadata Extraction

from Digital Documents. *2013 12th International Conference on Document Analysis and Recognition*.

2013. 135-139 [Publisher Full Text](#) | [Reference Source](#)

6. Kou Z, Cohen WW, Murphy RF: Extracting information from text and images for location proteomics. *Proceedings of the 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD 2003)*. 2003.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reader Comment 11 Apr 2017

Zeeshan Ahmed, University of Connecticut Health Center, USA

Reply: Thank you so much for your recommendations.

In this manuscript authors presented a computational tool that extracts text and images from PDF files. In general the manuscript is interesting considering that it can analyze various types of PDF files from different scientific areas based on the keywords and coordinates.

Reply: Thanks.

However, it lacks technical novelty over the published literatures and needs additional input on the image analysis section before indexing.

Reply: Thanks for raising this point, we have revised and tried to make it more clearer.

Extraction of texts and images from scientific publications has been presented in various domains: computer science¹, biomedical²⁻⁴, chemistry⁵, proteomics⁶ and so on. The manuscript by Zeeshan Ahmed and Thomas Dandekar presents an incremental innovation without providing clear technological advancement in the field. The objective of performing both physical and logical structure analysis of all kinds of PDF files as mentioned in the manuscript has not been sufficiently supported by technological contribution described in Methods section.

Reply: Thanks for raising this point and please accept apologies for the confusion. Its true that there have been many efforts from the past towards the same problem, and from different disciplines. We have taken a facile step by combining some of available technologies and tried to present a method, which can be adapted and further enhanced. To make our point more clearer, we have revised manuscript.

The image processing section the manuscript has been very brief. It does not provide any advanced image analysis technique as mentioned in the abstract. Authors should mention how exactly segmentation of figures and labels are performed and how they are represented to make logical connection between different entities in order to perform further analysis and customized visualization.

Reply: Thanks for raising this point and please accept apologies for the confusion. We agree with your point. We didn't discuss and go in to details of algorithmic image presenting because we didn't implement any algorithm for this work but tested and adopted some pre-existing OCR based libraries. We choose to go for commercial and licensed libraries because open source libraries and methods (we found) were unable to meet the developmental objectives this software. Without such details it is not possible to draw comparative (algorithmic, metrics based) conclusions. However, we give now a feature-based comparison in the results and discussion section of the manuscript.

The framework has been tested with a very small set of PDF files and no qualitative/quantitative result reporting the accuracy with respect to manually annotated files was presented. It would be

good to increase the number test files and include the results of qualitative/quantitative analysis.

Reply: Thanks for raising this point and we have revised with additional details. We tested our system with similar parameters on randomly selected scientific manuscripts (ten PDF files) from different open access (F1000Research, Frontiers, PLOS, Hindawi, PeerJ, BMC) and restricted access (Oxford University Press, Springer, Emerald, Bentham Science, ACM) publishers.

Some of the figures (Figures 1, 3, 4 and 6) in the manuscript are hard to see the details in both online and print format. These figures could be reformatted.

Reply: Thanks for raising this point and we have revised Figures.

Competing Interests: No competing interests.

Referee Report 13 January 2016

doi:10.5256/f1000research.7898.r11637



Juilee Thakar

Department of Microbiology and Immunology, University of Rochester Medical Center, Rochester, NY, USA

The manuscript titled “MSL: Facilitating automatic and physical analysis of published scientific literature in PDF format” addresses an important issue of extracting information from published manuscripts. However, the following issues must be clarified before indexing.

In the text mining section authors say that there is no tool to perform physical and logical structural analysis of PDF files. However, in the next paragraph they describe “Dolores” for logical structure analysis. Authors should describe how their method is different than Dolores.

Legends of all the figures should be more descriptive so that figures are understandable on their own. Each component of the figure should be described in the legend.

The results section is missing. Is it integrated in the discussion section? It is unclear what exactly the results were.

The article will be much clear if all the libraries (described on page 4 second paragraph) are described in the form of a table.

Authors should include a clear metric to estimate performance of the algorithm. This can be achieved by comparison with existing tools or through comparative analysis. A clear example showing the information extracted from several PDF files to address a biologically relevant example will be useful.

It is not clear whether the text extracted from the PDF files is actually coming from figure legends or related to the main body of the manuscript. Also, how is this text organized?

The authors mention that unexpected and irrelevant images were extracted. It is not clear how authors address that. It is absolutely essential to address that.

Minor corrections:

Page 2 second column: The definition of MSL is not the same as described in the abstract

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reader Comment 11 Apr 2017

Zeeshan Ahmed, University of Connecticut Health Center, USA

Reply: Thank you so much for your recommendations.

The manuscript titled “MSL: Facilitating automatic and physical analysis of published scientific literature in PDF format” addresses an important issue of extracting information from published manuscripts.

Reply: Thanks.

However, the following issues must be clarified before indexing.

Reply: Sure.

In the text mining section authors say that there is no tool to perform physical and logical structural analysis of PDF files. However, in the next paragraph they describe “Dolores” for logical structure analysis. Authors should describe how their method is different than Dolores.

Reply: Thanks for the nice suggestion. We have provided a brief comparison with Dolores, as well as some other mentioned tools in the paper. This includes Dolores, PDF2HTML, XED, and PDF-Analyzer, too and is now mentioned first time when mentioning Dolores.

Legends of all the figures should be more descriptive so that figures are understandable on their own. Each component of the figure should be described in the legend.

Reply: Thanks for pointing this out, we have revised the manuscript and added more details to the figure legends explaining the symbols used.

The results section is missing. Is it integrated in the discussion section? It is unclear what exactly the results were.

Reply: Yes, we have an integrated results and discussion section, please accept apologies for this confusion. To further clarify it, we have revised the heading titles and stressed the results achieved and subsequently discussed by subtitles.

The article will be much clear if all the libraries (described on page 4 second paragraph) are described in the form of a table.

Reply: Thanks for this important suggestion. We have added a table to the manuscript showing all libraries we tested for MSL as well as whether they are partly or completely integrated in MSL.

(i) Authors should include a clear metric to estimate performance of the algorithm. This can be achieved by comparison with existing tools or through comparative analysis.

(ii) A clear example showing the information extracted from several PDF files to address a biologically relevant example will be useful.

Reply: Thanks for these valuable suggestions.

(i). Initially, we also aimed to perform such comparative analysis, but most of the used and tested libraries are from different commercial and licensed sources, and algorithmic details were not given. We choose to go for commercial and licensed libraries because open source libraries and methods (we found) were unable to meet the developmental objectives this software. Without such details it is not possible to draw comparative (algorithmic, metrics based) conclusions. However, we give now a feature-based comparison in the results and discussion section of the manuscript. This also clearly shows the advantages of the MSL software. Furthermore, we discuss also the limitations and possible extensions of our MSL software, again referring to other existing software.

(ii): We have added a detailed example in the manuscript and some further examples of text and

image extraction to the supplementary material. Furthermore, we have added performance details, these highlight the biological importance as well. It is not clear whether the text extracted from the PDF files is actually coming from figure legends or related to the main body of the manuscript. Also, how is this text organized?

Reply: Please accept apologies for this confusion. The extracted text from PDF files is coming from the main text as well as the figure legends. The text is organized in reading order e.g. in the case of a two or multiple columns document; the text of the first column is followed by the text of the second column, and so on.

The authors mention that unexpected and irrelevant images were extracted. It is not clear how authors address that. It is absolutely essential to address that.

Reply: We apologize to have caused here some confusion. The “irrelevant images” are e.g. journal, publisher’s logos and header-footer images embedded inside document. These images are often added by the publishers of journal or conference, which have nothing to do with the actual manuscript’s content but to clear to the reader about publication details. We explain this in the manuscript including how these are removed.

Minor corrections:

Page 2 second column: The definition of MSL is not the same as described in the abstract.

Reply: Thanks for pointing this out; we have corrected this and are now more accurate when introducing MSL on page 2.

Competing Interests: No competing interests.

Referee Response 11 Apr 2018

Juilee Thakar, University of Rochester, USA

Thanks for responding to my suggestions.

Competing Interests: I have no competing interests to declare

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research