SOFTWARE TOOL ARTICLE

MSL: Facilitating automatic and physical analysis of published scientific literature in PDF format [version 1; peer review: 2 approved with reservations]

Zeeshan Ahmed¹, Thomas Dandekar²

¹The Jackson Laboratory for Genomic Medicine, Farmington, CT, 06032, USA
²Department of Bioinformatics, Biocenter, University of Wuerzburg, Wuerzburg, 97074, USA

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 figures from 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
Approval Status  ☑ ☑ ☑ ☑

version 3 (revision)
04 Apr 2018
view view ? view

version 2 (revision)
12 Apr 2017
? ? ? view

version 1
16 Dec 2015
? ? view view

1. Juilee Thakar, University of Rochester Medical Center, Rochester, USA
2. M. Julius Hossain, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
3. Florencio Pazos, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
4. Karin Verspoor, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
This article is included in the Bioinformatics gateway.

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

**Competing interests:** No competing interests were disclosed.

**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.

**Copyright:** © 2015 Ahmed Z and Dandekar T. This is an open access article distributed under the terms of the Creative Commons Attribution License, 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).

**How to cite this article:** Ahmed Z and Dandekar T. MSL: Facilitating automatic and physical analysis of published scientific literature in PDF format [version 1; peer review: 2 approved with reservations] F1000Research 2015, 4:1453 https://doi.org/10.12688/f1000research.7329.1

**First published:** 16 Dec 2015, 4:1453 https://doi.org/10.12688/f1000research.7329.1
**Introduction**

There has been an enormous increase in the amount of the scientific literature in the last decades. 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 capabilities. Due to this reason several document analysis based tools have been developed for physical and logical document structure analysis of this file type.

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 operators. The normal outcome of a successful NLP 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 orthodox NLP approaches 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, searching marginal text using key-words, 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.6-11, 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 analysis17, Dolores for the logical structure analysis and recovery18 automatic conversation from PDF to XML19 and PDF to HTML etc.

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
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.
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 e.g. iTekSharp (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), iTek 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://sharppdf.sourceforge.net) 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 Book-marks, iTekSharp 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 years\(^2\). 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 phenotypes\(^2\), prediction of subcellular localization of proteins in various organism\(^2\), analysis of gel diagrams\(^2\), mining and integration of pathway diagrams\(^2\).

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

---

**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.

---
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).

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, Springer, Emerald, Bentham Science, ACM) publishers, including some of the authors’ published papers, details are given in Table 1. While testing MSL on the selected manuscripts, we observed best overall performance for the manuscripts 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, Hindawi, F1000Research and IEEE 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 different papers. Our observed results using MSL are given in attached supplementary material (Supplementary Table S1 and Dataset 1).

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 papers, and application of OCR to one of the extracted images from another randomly picked publication.

Figure 1 shows that one randomly selected published article’s PDF file 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

| 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 MSL.

Table 1. 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 | Manuscript |
|------------|------------|
| F1000-Research | Ant-App-DB: a smart solution for monitoring arthropods activities, experimental data management and solar calculations without GPS in behavioral field studies. |
| PLOS | The Genomic Aftermath of Hybridization in the Opportunistic Pathogen Candida metapsilosis. |
| Hindawi | Mathematical Properties of the Hyperbolicity of Circulant Networks. |
| IEEE | Design implementation of I-SOAS IPM for advanced product data management. |
| BMC | Software LS-MIDA for efficient mass isotopomer distribution analysis in metabolic modeling. |
| PeerJ | Anvi‘o: an advanced analysis and visualization platform for ‘omics data. |
| Frontiers | Ontology-based approach for in vivo human connectomics: the medial Brodmann area 6 case study. |
| ACM | Intelligent semantic oriented agent based search (I-SOAS). |
| Bentham Science | DroLIGHT-2: Real Time Embedded and Data Management System for Synchronizing Circadian Clock to the Light-Dark Cycles. |
| Oxford University Press | Bioimaging-based detection of mislocalized proteins in human cancers by semi-supervised learning. |
Figure 3. Example: Publication, Figure 1 of (YY et al., 2015). This figure shows document image analysis, text extraction and PDF conversion.
Figure 4. Example Publication Page 1 (Ahmed et al., 2015). This figure shows document image analysis, text extraction, and PDF conversion.
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 shows another example of embedded text extraction from a complex figure, 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 article 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.

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, 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 multicolumn 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.

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 2). Depending on your application in mind, you thus simply extend the MSL parser by considering also these special characters occurring often in your texts.

**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. 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, Bytescoute, 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.

| 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.*
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 approaches 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 LEADTOOLS 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 1680x1050 with landscape orientation.

Conclusions
The latest available and easy to use version of MSL has been tested and validated in-house. The advancements in information

**Figure 5.** Screenshot of the all extracted images and generated files (XML and PDF).
Figure 6. MSL six steps installation process.

Step 1: Welcome to the MSL v1.0.0 installation!

Step 2: Installing file.

Step 3: License agreement.

Step 4: Select installation folder.

Step 5: Destination folder.

Step 6: MSL v1.0.0 has been successfully installed!
retrieval techniques for text and figure analysis combined with this sophisticated computational tool can support various studies. As future work, we are looking forward to further implement machine learning and pattern recognition methods for the extracted text classification.

Data availability
F1000Research: Dataset 1. Extracted Images and Text from Papers tested using MSL, 10.5256/f1000research.7329.d108739

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).

License
All associated files are licensed under the Academic Free License 3.0 (AFL 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
1. Hunter L, Cohen KB: Biomedical language processing: what’s beyond PubMed? Mol Cell. 2006; 21(3): 589–594. PubMed Abstract | Publisher Full Text | Free Full Text
2. 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
3. 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
4. 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
5. 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
6. Plikus MV, Zhang Z, Chuong CM: PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithms. BMC Bioinformatics. 2006; 7: 424. PubMed Abstract | Publisher Full Text | Free Full Text
7. Smalheiser NR, Zhou W, Torvik VI, et al.: Anne O’Tate: A tool to support
Open Peer Review

Current Peer Review Status:  ?  ?

Version 1

Reviewer Report 09 August 2016

https://doi.org/10.5256/f1000research.7898.r14625

© 2016 Hossain M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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 science\(^1\), biomedical\(^2\)-\(^4\), chemistry\(^5\), proteomics\(^6\) 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, et al.: 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, et al.: 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, 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
5. Choudhury SR, Mitra P, Kirk A, Szep S, et al.: 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 3nd ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD 2003)*. 2003.

**Competing Interests:** No competing interests were disclosed.

**I confirm that I have read this submission and 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**

**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, Peer), BMC and restricted access (Oxford University Press, Springers, 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.
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 confirm that I have read this submission and 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.

---

**Zeeshan Ahmed**

**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 clearer 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.

Reviewer Response 11 Apr 2018

**Juilee Thakar**

Thanks for responding to my suggestions.

**Competing Interests:** I have no competing interests to declare