ARTICLE

Accessibility of Tables in PDF Documents

Issues, Challenges, and Future Directions

Nosheen Fayyaz, Shah Khusro, and Shakir Ullah

ABSTRACT

People access and share information over the web and in other digital environments, including digital libraries, in the form of documents such as books, articles, technical reports, etc. These documents are in a variety of formats, of which the Portable Document Format (PDF) is most widely used because of its emphasis on preserving the layout of the original material. The retrieval of relevant material from these derivative documents is challenging for information retrieval (IR) because the rich semantic structure of these documents is lost. The retrieval of important units such as images, figures, algorithms, mathematical formulas, and tables becomes a challenge. Among these elements, tables are particularly important because they can add value to the resource description, discovery, and accessibility of documents not only on the web but also in libraries if they are made retrievable and presentable to readers. Sighted users comprehend tables for sensemaking using visual cues, but blind and visually impaired users must rely on assistive technologies, including text-to-speech and screen readers, to comprehend tables. However, these technologies do not pay sufficient attention to tables in order to effectively present tables to visually impaired individuals. Therefore, ways must be found to make tables in PDF documents not only retrievable but also comprehensible. Before developing such solutions, it is necessary to review the available assistive technologies, tools, and frameworks for their capabilities, strengths, and limitations from the comprehension perspective of blind and visually impaired people, along with suitable environments like digital libraries. We found no such review article that critically and analytically presents and evaluates these technologies. To fill this gap in the literature, this review paper reports on the current state of the accessibility of PDF documents, digital libraries, assistive technologies, tools, and frameworks that make PDF tables comprehensible and accessible to blind and visually impaired people. The study findings have implications for libraries, information sciences, and information retrieval.

INTRODUCTION

The web has a huge collection of documents, including pages, books, blogs, articles, reports, etc., available in different formats. These formats include HTML (HyperText Markup Language), EPUB (Electronic PUBlication), AZW (AmaZon Word), and the ubiquitous PDF (Portable Document Format) format. PDF is layout oriented and unstructured, having elements such as text, images, tables, and metadata. All these elements carry specific information and have their relative importance. Tables can be part of a structured or unstructured document. A structured table, like in HTML, is relatively easy to extract and interpret, as it has a starting and ending tag pair for the table itself, its headings, each row, and discrete values. However, unstructured documents, which can include books, journals, audio, video, images, and documents, do not follow a specified format or structure for the organization of information.[1] A table has levels of abstraction; the higher levels

Nosheen Fayyaz (nosheenfayaz@uop.edu.pk) is doctoral candidate, University of Peshawar. **Shah Khusro** (khusro@uop.edu.pk) is Professor, University of Peshawar. **Shakir Ullah** (ullah@ulm.edu) is Instructor, University of Louisiana Monroe.

https://doi.org/10.6017/ital.v40i3.12325

of abstraction have fewer details whereas a lower level gives more information. The human has to understand and comprehend the underlying semantics of the table content for sensemaking. The content of a table has a strong bond with its context, as it has concrete information regarding the surrounding text; therefore, tables are hard to comprehend when taken out of the context. Poorly conceived information is more dangerous, as it can lead to misconceptions and poor decisions. Any system or component that interacts with humans must be capable of offering comprehensible explanations.[2] A reader understands a table in at least three cognitive processes: *comprehension*, *searching*, and *interpretation & comparison*.[3] In contrast, blind and visually impaired persons need assistance in comprehending the tabulated information, for example, understanding table structure and its content, searching for particular information in a table, and comparing and interpreting tabular data. Therefore, they need technical solutions for reading documents.[4] According to the World Health Organization (WHO), the number of blind and visually impaired people has increased significantly and has risen to 2.2 billion, so technical solutions or assistive technology are a must for their reading.[5]

Assistive technologies are supposed to handle the three main kinds of print disabilities: vision problems, motor skill problems, and cognitive problems.[6] For vision problems we have tools like text-to-speech and screen readers to help blind and visually impaired people to read text documents. However, these tools work on the upper level of abstraction and give limited information to users because they focus on text and ignore components related to presentation such as tables, graphs, images, etc. This limitation is not only found on the web but also affects other digital environments including digital libraries, where more reliable document collections are present but their retrieval and presentation to blind and visually impaired people is challenging.[7] For example, a study identified the limitations of digital libraries in meeting the specific needs of blind and visually impaired people and suggested including help features for a more user-friendly experience.[8] Michigan State University has taken an initiative to make digital content accessible by adopting WCAG 2.0 as official technical guidelines. They presented a five-year accessibility plan for making new, existing, and purchased content accessible along with resource allocations, training for their staff, and future requirements.[9] These efforts may motivate other libraries to adopt such measures and make it a convenient place for people with disabilities. Common accessibility problems include the lack of alternate descriptions, using visual cues to describe interactions in the user interface, fuzzy visuals, and audios.[10] Furthermore, the sight-centered nature of the digital library creates problems for blind users, such as the absence of meaningful descriptions for nontext content and instructions, along with information about the digital library's features due to missing textual or verbal instructions.[11] The traditional usage of a digital library makes a canned or routine utilization of its collections, which may be broadened by making computational ready collections.[12] The accessibility of these documents will help in knowledge dissemination to blind and visually impaired people.

Researchers have presented frameworks and algorithms for exploring and interpreting PDF elements like images, charts, tables, and graphs. These interpretations are the basis for both humans and machines to gain meaningful insights out of tabular data. This paper highlights the significance of the rich semantics of PDF tables and the challenges in their interpretation and their presentation to blind and visually impaired people. It is proposed to present the tables' explicit and implicit information in a progressive manner to reduce the cognitive overload on blind and visually impaired individuals. This might be achieved by providing some basic information (such as a table caption and the number of rows and columns in the table), which may be followed by

navigation and querying within the table. This stepwise approach of leveraging a table's semantics may help in its better comprehension. Table semantics will also be helpful in libraries, information science, and information retrieval, as it has the potential to improve library cataloging and classification. The next section of this paper discusses and reviews the efforts and limitations in the existing literature and presents a general model of table processing and interpretation. The prominent issues and challenges are identified regarding general table structure, format, interpretations, and evaluation; the presentation of tables to blind and visually impaired people; and specifically the accessibility issues in digital library are presented in the following section. The last section contains some future research directions that will unleash some new dynamics of this domain.

THE CURRENT STATE OF TABLE PROCESSING

A table presents summarized information in a particular arrangement, where the structure of the table reveals some implicit semantics. In 1996, Xinxin Wang defined the logical structure and style rules of tables and presents Wang's abstract model.[13] The model separated the logical structure from layout specification and is considered a generic and complete model in the literature.[14] Unstructured tables can be regular or irregular. A regular table has intersecting vertical and horizontal borders that develops a table of cell bounding boxes, while in an irregular table, there is no relationship between the number of rows and columns.[15] Tables can be n-dimensional, having spanning cells or multiline cells. Tables can be long and span multiple pages and they can be floating (can be placed to the left or right of the page, with text wrapped around them). Sometimes tables have no explicit boundaries and even worse, cell separators may not be visible. A table can have a variety of content that includes numerical data, text, symbols, images, and equations.[16] The location of table and identification of table structure in documents is evaluated in the International Conference on Document Analysis and Recognition (ICDAR) table competition.[17] The methods used for the identification of table structure are rule-based methods, [18] data-driven methods,[19] and the graphical neural network.[20]

Multiple frameworks and approaches are used for the extraction and processing of tables from structured documents like HTML,[21] and unstructured documents like images and PDF documents.[22] Keeping in view all the conducted studies and research, we present a general model of table processing and interpretation in figure 1, showing the prominent inputs (web, PDF, and images), processes, and some notable outputs. The model has a table extraction process that is followed by processing which yields multiple outputs including organized data, analyzed structure, and analyzed content. Processing can be followed by establishing relationships within the table, within table and context, or with other related tables. Moreover, tables presented in open formats such as CSV, XML, JSON, and RDF extend their potential by exploration, creating or extending ontologies and knowledge bases, and publishing tables on an LOD cloud to establish links with other open data sources. Below is a detailed discussion of table extraction and processing and the relationships of tables with content and context.

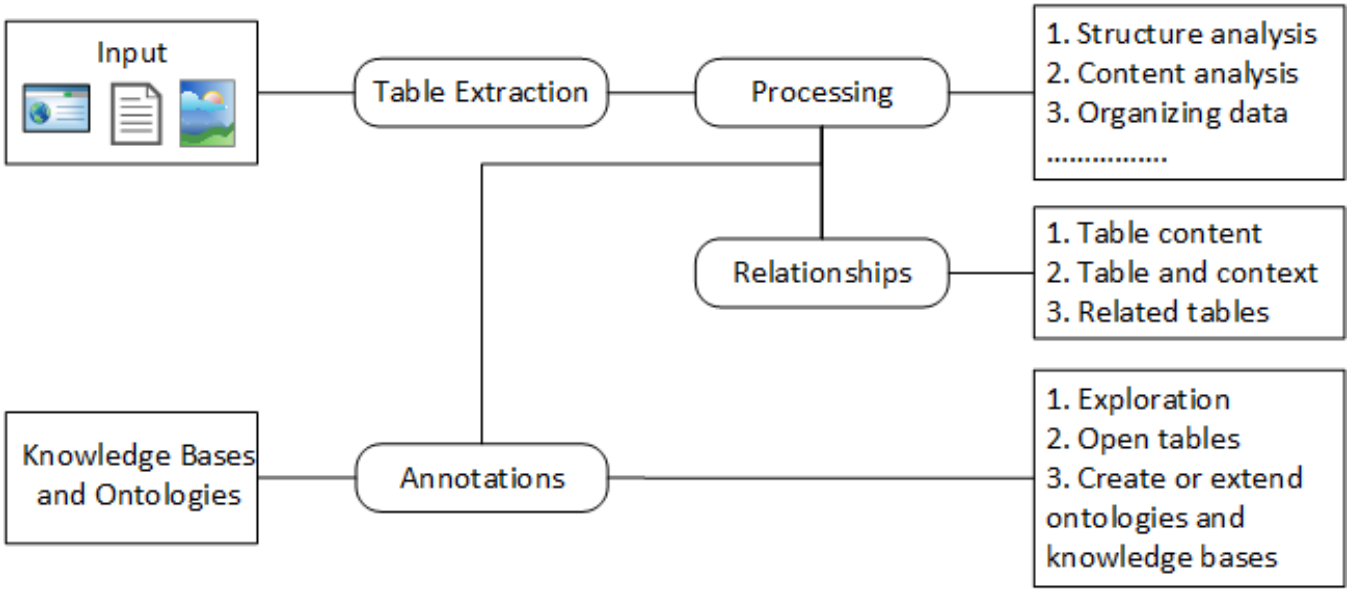

Figure 1. A general model for table interpretation.

Table Extraction and Processing

The recognition, extraction, and processing of tables, from a variety of documents, have used multiple approaches.[23] The hidden table semantics will not only help in understanding tables but can also contribute to digital library cataloging. These approaches are categorized in the following sections.

Using Heuristics

Different heuristic approaches are presented for extraction and processing of unstructured tables. For example, PDFTREX uses spatial features and follows a bottom up approach for the recognition and extraction of tables. It represents a table as a two-dimensional grid on a Cartesian plane and extracts the table as a set of cells along with their coordinates.[24] In another approach, natural language processing (NLP) features are used for deeper understanding of text. These tools use parts of speech and dependency paths for the extraction of tables and for finding relations among tables by using the NLP toolkit.[25] Milosevic et al. presented five steps for table processing, i.e., table detection, functional analysis, structural analysis, synthetic analysis, and semantic analysis,[26] while Roya Rastan endorses the first three steps in his PhD dissertation and proposed a framework for the processing of tables. This framework consists of four layers: input management, table processing, storage, and management.[27] To recognize and extract tables from documents, ad hoc heuristics are used with the existing methods and techniques, which includes three steps: (1) "preprocessing" to define and prepare text chunks from a source table by using the features of text like font, space, and bounding box; (2) "text block recovering" to identify the set of text chunks that could be treated as the content of a single cell; and (3) "cell recovering" to observe the arrangement of cells for identifying the rows and columns.[28] The authors exploited appearance features of text printing instructions and position of drawing cursor for table detection and structure recognition in their web-based solution. They claimed to attain an accuracy or F-score of 83.18% for table extraction and 93.64% for structure recognition.[29] Furthermore, an interactive document reader was presented by the researchers of Stanford University, in which structural analysis was combined with rule-based matching and natural language processing to associate a table's values with the related text to develop sentence-table

pairs in the document. They also tried to relate tables in two data sets but unfortunately obtained 48.8 % results.[30] Although the results were not satisfactory, this effort opens up new endeavors for practitioners and researchers.

Using Segmentation

The segmentation approach is used for identification of tables in unstructured and untagged PDF documents, along with its columns and rows.[31] Visual separators and geometric content layout information is used for the extraction of tables from multiple pages of documents, and the technique is tested on e-books and scientific documents.[32] Ali et al. adopted a segmentation approach to deal with incomplete, impure, and complex tables by extracting table schema, data, and reading paths of data to represent in a layout independent format.[33] The extraction of tables from images also used segmentation through a top-down pipeline approach. The text and tables in medical laboratory reports were identified, where the content of tables needs to be correctly captured and interpreted.[34] As the medical tables include text, numbers, characters, and symbols, therefore, their correct interpretations is critical in medical reports and a minor error can lead to very dangerous outcomes.

A system named TEXUS used segmentation to prepare text chunks for finding relations among cells. The system provided end-to-end processing of tables, which claimed to detect a variety of tables in layout-independent format from a data set of complex financial tables. The system interpreted the tables and produced an XML file about the structure of the tables showing the access paths of each data cell as an attribute.[35] Similarly, page segmentation is carried out by using deep learning methods to identify tables, text, and figures.[36]

Using Machine Learning and Deep Learning Approaches

Machine learning and deep learning techniques are also used for automatic detection and extraction of table data. Random forest classifiers are used to detect a table header.[37] Multitask fully convolutional neural networks (FCN) are used for page segmentation to identify the tables, text block, and figure elements.[38] The K-nearest neighbor method and layout heuristics are used in a system named TAO for the automatic detection, extraction, and organization of data in tables in order to generate an enriched representation of data.[39] Similarly, deep learning techniques like R-CNN are used to capture tables from a University of Las Vegas data set of document images. Based on the assumption that tabular data is mostly numeric, the researchers used color-coding/coloration to distinguish numeric and textual data and claim to have achieved improved performance.[40] Table detection and recognition in born-digital documents and images is carried out by using transfer learning for faster R-CNN in order to overcome the problem of labeled data sets and FCN semantic segmentation is used for table structure recognition. The method is evaluated using the ICDAR 2013 data set.[41] Another approach, named DeCNT, worked on images (in any format) for the extraction of tables by using a combination of deformable CNN with faster R-CNN/FPN. The method is evaluated using the ICDAR 2013 data set and the ICDAR 2017 POD data set, UNLV, Mormont.[42]

In other research, the authors pointed out the weakness in the existing methods and techniques for understanding tables and presented a “graph neural network” approach to analyze the structure of PDF tables and handle the spanning cells.[43] Along with that, multiple deep learning techniques are used for integrating and querying tables using word embedding, RNN, KNN, and LSTM for the classification of financial tables.[44] In the case of web tables, annotation of table columns is performed by using Convolutional Neural Network (CNN) along with transfer learning

in order to overcome the problem of a shortage of target data sets. The column semantics are embedded into vector space and are used for predicting the type of columns without using the metadata. The method is tested on two web table data sets, T2Dv2 and Limaye.[45]

Using Ontologies

Ontologies have also played a vital role in the detection, recognition, and annotation of tables from the web, images, and PDF documents. Ontologies consider the content and structure of a table for their conceptual representation. A system named TableMiner was used to interpret the tables semantically by identifying the semantic concepts for the columns and disambiguating the cell content by using RDFa and microdata for improved annotations of the table.[46] TableMiner considered relational tables and mapped the headers of table with the properties of ontology for linking the cell values with entities.[47] The relationships between a table and its context are also extracted and annotated, and to remove disambiguation, the provenance of relationships is preserved.[48] A framework named TABEL, developed by Varish Mulwad, has a module for converting a table to a graphical model to infer the semantics of table header, cells, and their relation to each other. These semantics are used to convert the graphical representation to RDF triples by using knowledge bases along with the author's own defined ontology or any other ontology.[49] The ontology is also used for finding the relevant tables in a domain of technical documents only.[50] For easy interpretation by ontologies and more usability of the government data, the unstructured tabular data is suggested to be published in open format like CSV (comma separated value).[51] The studies mentioned above have mostly used relational tables, technical document tables, government data, and medical data with a main objective of making tables open and interrelated. Along with that, another study argued that besides the metadata of a resource, user-generated content may also be considered and published as linked open data for improved consumption and would also contribute to better cataloging of digital libraries.[52] Unfortunately, it still has problems like disambiguation and correlation of complex tables, besides other issues involved in publishing and consuming data as open and linked data.[53]

Relationship of Tables with Content and Context

The content of a table is present in a particular arrangement in order to give some specific information. Therefore, the table content should be interpreted for the hidden semantics among the cell content, context, and with other related tables in a particular domain. In this regard, natural language techniques are used for the identification of relationships in the table and the related text using the NLP toolkit. The researchers claimed improvement in table schema identification and quality of relation.[54] Similarly, ontologies are used to identify the semantic relations among the text, table contents, and table structure.[55] Another research project used rule-based matching and structural analysis for finding the relationship in table cell and sentence text, by developing sentence-table pair in the document. This project tried to develop a relationship between tables of two data sets but achieved only 48.8% success rate.[56]

A system named TEXUS tried to find out the relations among the values of cells using cell entries, categories, and access paths. They used segmentation techniques for preparation of text chunks and produced an XML file about the structure of the table showing the access paths of each data cell as an attribute.[57] Narrowing the table-understanding domain to clinical literature, with a focus on just the numerical and textual data of tables from XML documents, Milosevic et al. extended their previous work and tried to identify the relationship between the table and the surrounding text. They added pragmatic analysis, cell selection, and syntactic analysis, defined five categories of cells, depending upon the data in the cells, and identified seven semantic categories for the

specification of table extraction process. The PubMedCentral data set is used to test the developed system with regard to task, variables and complexity. The authors claimed to achieve an accuracy measure F-score between 82% and 92%.[58] A "graph neural network" model was developed to build an undirected graph for the prediction of relations among adjacent cells. The model was tested on benchmark data sets, i.e., ICDAR 2013 and tableBank-2019, and claimed to outperform.[59] Another system, named Tablepedia, unified the tables of experimental results with regard to method, data set, metric, score, and source into tuples. The system extracted the related tables and identified the conflicted results by using the rule-based and learning-based methods with the help of SQL operations.[60] An SQL-like query was proposed for the financial tables in PDF formats by using deep learning approaches.[61] All the mentioned techniques for establishing relationships follow the rule-based, learning-based, segmentation, neural network, heuristics, and ontologies. Among these, the ontologies can establish inter domain relationships and explorations. However, it still has issues that will be discussed in the conclusion section.

Existing Accessibility-Driven Solutions for PDF Documents

Apart from the systems and frameworks for table understanding and processing, a mechanism or a solution is needed to present tables in a meaningful way to blind and visually impaired people. The accessibility of digital documents is based on the captured structural information and its availability for processing by other software and applications, such as tagged PDF, can help in summarizing, navigating, and providing structural information of the content.[62] Nazimi made an effort to present a framework for understanding the complex documents and its components, including images, charts, and tables, in a nonvisual representation to blind and visually impaired people.[63] The existing available solutions for reading PDF documents to blind and visually impaired people focus on text and give little attention to its elements such as tables, images, graphs, and charts. Particularly speaking for tables, these solutions either read the table caption and ignore the content, or read the table as if it were text, which renders it meaningless. These assistive technologies are divided into four main categories.

1. Text-to-speech tools
2. Screen readers
3. Voice assistant
4. Natural Language Generator (NLG)

Text to speech tools include products like WordTalk, Virtual Speaker, Audiobook Reader Voice, Voicepaper, Dream Reader, etc. These tools can read text from txt, PDF, or doc files aloud and have an interface for user interaction, where the user can copy-paste the text or mention the path of the file to read. Some of the tools are free with limited features while others are proprietary. They need human interaction, which might be difficult to use for a visually impaired or blind person. Screen Readers include JAWS, NVDA, COBRA, VoiceOver, and Talkback. They speak out every user activity that is taking place, like opening or closing a window, clicking on a button, reading text from a txt or PDF file. These tools are helpful for visually impaired people, as they do not need the user to open a specific software and then specify the path of files to read. The most popular tool, JAWS, is proprietary and used for Windows. NVDA is free software for Windows, while VoiceOver is a free tool provided with Apple's operating systems (including macOS and iOS). NVDA reads the text, taking note of punctuation; it reads the table row by row like text and then reads the caption of the table at the end. It can also read the alternate text of the table if it is included in Acrobat Pro. These tools need the user to be aware of what he or she is doing.

Similarly, there are voice assistants like Apple Siri, Microsoft Cortana, Google Assistant, and Amazon Alexa. All these tools take instructions from the user and then try to provide solutions. These tools may ask users several queries to clarify what they want and provide limited functionalities such as reading out GPS coordinates, playing music, etc. The Natural Language Generator (NLG) is used to convert raw text or data into narrations. Existing popular systems or tools include ARRIA NLG, Quill, IBM Watson, AX NLG Cloud, Amazon Polly, and Wordsmith. These systems are used to perform data analysis and convert the extracted analytics to narrations, which could be easily understand by the user. These tools are not for narrating tables from unstructured documents. However, a framework has been developed using a neural encoder-decoder to generate text from tables. It is claimed that the solution outperformed the existing solutions and achieved higher BLEU score and F-score using data sets WEATHERGOV, WIKIBIO, WIKITABLE, and a Chinese data set WIKIBIOCN.[64] This technique focuses on formal tables and ignores complex tables as well as unstructured PDF.

A new emerging category in this paradigm is the document-centered assistant, which tries to help users review documents by asking questions. The field is currently studied for the type of questions that a user may ask and the candidate machine learning models that can be used for answering them. The questions would be different from factoid questions and chitchat, because here the focus would be on relevant information from that specific document.[65] This category seems to have a big scope for understanding, reviewing, and inferring knowledge from documents.

Apart from the solutions mentioned above, there are some Java and Python tools and libraries that are used for table extraction from PDF documents and are shown in table 1. Some of the tools are commercial and claim to extract tables, table rows, and even table cells from documents and images, like PDFTables, DocParser, and PDFTron. Similarly, there are also open-source Java and Python libraries for table and metadata extraction from images and documents. The libraries that extract tables from images are Camelot, Excalibur, and PDFPlumber, whereas the libraries that extract tables from non-image-based documents are Tabula, PyPDF2, PDF Table Extractor, PDFPlumber, and PDFMiner. Among these five, PDF Table Extractor is browser-based and PDFMiner works with structured tables and digs out the semantic relations. For working with unstructured tables in PDF documents and developing a table extractor component for an integrated environment, Tabula, PDFPlumber, and PyPDF2 might be better choices.

The research and solutions mentioned above regarding table detection and understanding are carried out to make them meaningful to machines and humans who have no visual impairment or dyslexia. Therefore, the future documentation may consider the inclusion of translations and lay summaries (concise descriptions in simple words) of objects or elements within the document, as essential components, to make them accessible to blind and visually impaired individuals as well.[66] In this regard, the World Wide Web Consortium (W3C) developed the Web Accessibility Guidelines for developing web documents to make the nontext elements accessible. These guidelines include elements such as captions for tables and figures, description of figures, and summaries of the tables.[67] Similarly, HTML has tags to include summaries of a table, including <summary>, , <p id= “tblDEsc”>. Microsoft Word has an option “text alternative” to add a description of a table or figure for visually impaired people, who will use screen readers for reading the document. Adobe Acrobat Reader also has an accessibility pane to tag tables and add alternative text and descriptions of tables, which is used by the NVDA screen reader to read aloud. Moreover, CommonLook Office, whose motto is “build accessibility into documents early,” has add-ins for Microsoft Word or PowerPoint to add enough accessibility content to the documents to

Table 1. Solutions and libraries for table extraction and processing.

S no.	Tools	Open source	Image based	Comments
1	Tabula	Y	N	Extracts data tables from PDF and saves as CSV or Excel spreadsheet. It works on native PDF files and cannot extract scanned tables. It supports multiple platforms but does not support batch processing.
2	PDFTables	N	N	Extracts page, table, table row, and even table cell. It is a fully automated API. It supports multiple platforms and multiple programming languages.
3	DocParser	N	Y	Extracts information from images and forms. It is a cloud-based application and supports batch processing. It parses the documents and offers more features but needs human intervention. It shows poor accuracy in handwritten application forms.
4	PDFTron	N	N	Supports multiple platforms and multiple programming languages.
5	Camelot	Y	Y	A Python library that extracts table from images. It has built-in OCR.
6	Excalibur	Y	Y	A web-based solution which is powered by Camelot.
7	PyPDF2	Y	N	A Python library that can do batch processing with multiple files.
8	PDFPlumber	Y	Y	A Python library built on PDFMiner.
9	PDF Table Extractor	Y	N	A web-based tool built on Tabula. It supports scraping of multiple page tables and comparison of cell values.
10	PDFMiner	Y	N	A Python library that extracts information like location, fonts, and lines of the text. It focuses on analyzing text. It has a PDF parser. It figures out the semantic relationships among structured tables.

make the resulting PDF accessible. However, already-developed unstructured documents, without any accessibility features, still need some measures to make the documents understandable to visually impaired or blind users.

Keeping in mind the statistics of visually impaired people and the unstructured data of the future—the global data sphere will grow from 33ZB to 175ZB and 80% of this worldwide data will be unstructured—visually impaired individuals cannot be ignored for their access to knowledge.[68] Therefore, we would need mechanisms for making these unstructured documents understandable to as many people as possible by incorporating accessibility measures in the document readers. The following section highlights some of the key issues in this domain.

ISSUES AND CHALLENGES IN THE EXISTING SYSTEMS

Tables can be utilized in multiple scenarios including information extraction, table search, ontology engineering, conversion to DBMS, and document engineering.[69] The situation becomes difficult when a blind or visually impaired person needs to understand the tables. The issues and challenges in dealing with PDF tables are categorized in the following sections.

Table Structure

Tables in PDF documents need more focus on table structure detection because they do not follow a defined formal structure.[70] Several knowledge gaps are identified in literature regarding table structure, such as the identification of functional areas of tables, for which Silva argued the use of multiple heuristics and machine learning algorithms in parallel or in sequence.[71] The variety of structural layouts creates problems in their identification, which can be handled by defining more rules at the lexical and syntactic layer of table processing. This could also be fruitful for better semantic annotations.[72] In addition, the variety of cell content or inconsistent cell content, along with implicit header cells, creates problems in understanding the tables, especially by machines.[73] The vector representation of web tables may be applied to PDF tables for semantic annotations and identification of column types.[74] Along with that approach, graphical representation and a graphical neural network (GNN) can also be used for better structure identification in multiple domains.[75] New data sets need to be introduced for structure recognition in various domains, including business and finance, as they use a huge amount of tables in their documents.[76] From the discussion above, the table structure inconsistencies, cell content inconsistencies, functional and logical processing of tables needs more research effort to eliminate the stated problems. Along with that, the inclusion of more data sets will also help in handling the diversity in the field.

Table Formats

The existing format of tables in PDF lacks the metadata needed for further processing; therefore, the conversion of PDF tables to other formats, especially open formats, will open new endeavors. Some researchers have worked on converting tables to CSV format, which retains the basic structure but lacks some cell formatting. Researchers worked on the transformation of web tables to relational tables for easy manipulation.[77] In contrast, XML can handle complex data and is more easily read by humans. Therefore, a methodology is presented to work on tables in XML format, but it considers tables having text and numerical data only.[78] JSON, another format, can also be used as an alternative to XML; it is smaller in size than the XML and can handle complex and hierarchical data. The JSON format has less support than XML but is preferred for web application due to its interoperability and lightweight features.

Table Interpretation

The variable representation patterns of table values, dense content and natural language processing create problems in the correct interpretation of tables.[79] Anaphoric resolution techniques and documenting level discourse parsers are suggested to handle complex references among multiple domains.[80] Moreover, handling the locality features of a table and the annotation of its property feature can lead to better interpretation of tables.[81] The use of a knowledge base is suggested for understanding and annotating the relationships among tables and text to get more information about the extracted entities from tables and text.[82] Similarly, the extraction of data and its precision in medical and financial tables is an issue that needs the attention of researchers, as both fields have crucial and important data in its tables. [83] For easy interpretation of tables, machine learning classifiers, based on table headings and captions, can be used to classify them into their respective domain.[84] The relationship of tables in a specific domain and or among multiple domains can be achieved by developing ontologies.[85] This will enable the tables to be published on an LOD cloud that will establish more relationships and infer insights from multiple domains.

Table Evaluation

Most of the researchers working on PDF tables have tried to evaluate their work with popular data sets such as ICDAR 2013, ICDAR 2015, ICDAR 2017 POD, PubMed, UNLV, and Mormont. As we have PDF documents in multiple domains, therefore, new data sets should be introduced for structure recognition, especially in business and finance, as these domains use a large number of tables in their documents.[86] An evaluation methodology was proposed for table detection, structure recognition, and its functional and semantic analysis.[87] Unfortunately, there are no standard metrics, parameters, and formal methodology for table processing evaluation.[88] Therefore, standard evaluation metrics should be defined for PDF tables, in order to standardize the evaluation of algorithms and frameworks.

Table Presentation to Blind and Visually Impaired Users

The available tools and techniques for reading aloud documents to blind and visually impaired people either read the table caption only and ignore the content or treat the tables as text and read the rows line by line. This does not help these users to understand the semantics of the table and its content. Besides the content of the table, its layout shows grouping and connections among the content which is not presented to blind and visually impaired people by current solutions.[89] Therefore, tools and screen readers need to present tables in nonvisual format or give a summarized view of tables by following the guidelines of W3C, instead of reading the table like text.[90] The summarized view of tables can become part of bibliographic metadata and can contribute in cataloging in the perspective of linked and open data.[91] A study highlighted the accessibility of published PDF articles by four journal publishers and presented the findings in graphs to show the trend from 2009 to 2013, by taking parameters including meaningful title, alternate text for images, and logical reading order.[92] The author further applied the same methodology to analyze the articles published in next four years (2014 to 2018) and came to the conclusion that accessibility of PDF documents had improved. However, the journal publishers, who should be more aware of disability and accessibility, did not consistently follow the PDF/UA accessibility requirements and WCAG 2.0 when producing PDF versions of their articles.[93] Therefore, visually impaired individuals should be provided with a mechanism for understanding the digital content and underlying semantics at multiple levels of abstractions, like the general information about the document and its elements—including tables—its structure and content, navigation in the table, and querying the table to get more details and lessen cognitive overload.

Accessibility of Digital Library Collection

The accessibility of large-scale digital library collections can enhance content for sighted as well as visually impaired users. The traditional utilization of digital library collections needs to be broadened by making computation-ready collections meant to be used and consumed in multiple domains.[94] An effort was made by researchers to digitize and archive a digital repository of images and convert them to PDF/A documents but, unfortunately, the researchers came up with limited semantics as they did not consider the elements within the documents themselves.[95] The accessibility of these converted documents may be compromised with these limited semantics. The rich semantics of tables can be used in the bibliographic classification of a digital library's collection to increase the search width of the digital library.[96] Blind and visually impaired users can be assisted in using digital libraries, as they may need help at physical and cognitive levels. At the physical level, the blind may face difficulty in accessing information, identifying path and status, and efficiently evaluating information. At the cognitive level, they may face problems in understanding multiple structures, programs, information, features of the digital library, and the need to stick to some specific formats. Therefore, the inclusion of help features will make the

digital library friendly to blind and visually impaired people by incorporating meaningful descriptions for nontextual elements.[97] The sight-centered nature of the digital library creates problems for blind and visually impaired users due to missing textual or verbal instructions. Some researchers identified the inclusion of labels and meaningful descriptions for hyperlinks, instructions, structure, multimedia content and nontext content to make digital libraries friendly to blind and visually impaired people.[98] At the same time, others argue for improvement in usability by introducing help features in terms of usefulness, ease of use, and user satisfaction.[99] The accessibility of digital libraries in general and its content in specific may be improved by accommodating help features in the interface and meaningful descriptions for the contents' nontext elements including tables.

CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

This study discusses the accessibility of tables included in PDF documents in general as well as in the specific environment of digital libraries. Existing frameworks, algorithms, and solutions for the processing and interpretation of PDF tables, specifically their presentation to blind and visually impaired people, are thoroughly discussed. A general workflow of table processing is also presented in figure 1. The available solutions for reading out PDF documents to blind and visually impaired people are analyzed for their output, specifically for their attitude towards handling tables. Furthermore, a list of resources for table interpretation and presentation are discussed along with their different features. The issues and challenges in table structure, format, interpretation, evaluation, its presentation to blind, and accessibility of digital library collection are discussed. The researchers working in the domain of accessibility, digital library, and PDF tables can extend and modify the current solutions and algorithms by following the future research directions given below.

- The structure of a table has implicit semantic information which a sighted reader can infer but a blind reader needs assistance to understand. The structure of a PDF table is extracted using multiple approaches like heuristics, ontologies, machine learning and segmentation, whereas vectors are used for a web table.[100] Therefore, the combinations of multiple approaches and use of vectors for PDF tables may produce better results.
- The content of a table is usually numeric or very short text and needs proper interpretation. Therefore, a knowledge base can be used to get more information about the extracted entities from tables and text in order to understand and annotate the relationships among tables and text.[101] These knowledge bases can be predetermined or may be selected automatically according to the table content or domain.
- Table interpretation can become easy if tables are classified according to their domains by using machine learning classifiers. The classification can be based on table headings and captions, as well as the title and author of the document.[102]
- Ontologies are used to relate the tables in a specific domain and or among multiple domains, and publishing them on an LOD cloud will establish new relationships.[103] This will help in inferring new insights from complex, long, and numerical tables.
- Unstructured data and content can be made available for multiple usage and interpretations if it is converted to open formats like CSV, JSON and XML.[104] Among these, CSV comes with repeated content, XML needs special parsers, whereas JSON is lightweight and easy to write and read.[105] It has support from NoSQL databases like MongoDB and Apache CouchDB, and web Application APIs like Twitter, You Tube, and Facebook.

Therefore, JSON might be a better option for the conversion of PDF tables for its multiple interpretation and navigation within tables.
- The processes used for evaluation of tables have no defined matrices.[106] Therefore, the table evaluation processes should be defined with their respective matrices in order to standardize the research in this domain.
- The precision of extracted content of table is very crucial especially in medical, financial, and experimental tables that have numeric data. Therefore, the preprocessing of tables or conversion to other formats would need more attention to avoid any truncation or round off of the data.
- The presentation of tables to blind or visually impaired people can be in nonvisual or summarized form.[107] The summaries may be presented nonvisually, including the structural layout as well as a brief introduction of the table, to minimize the cognitive overload on these individuals.
- To evaluate the accessibility of digital library interfaces, 16 heuristics were proposed to make the digital libraries in reach of users, however, more heuristics are needed to make generalized interfaces for all individuals.[108]
- The nontext elements of digital library collections should have meaningful descriptions for better understandability of blind and visually impaired individuals. The user-generated content about these nontext elements could be used for cataloging.[109]
- The rich semantics of tables can be exploited for cataloging and classification that will be helpful in exploratory searching.
- As the Michigan State University Libraries has taken the initiative of assessing and improving the accessibility of digital library content by adopting the WCAG guidelines, other libraries can also adopt the model for providing accessible content to their users including blind and visually impaired individuals.
- The development of new data sets for tables in multiple domains can facilitate the researchers in interpreting tables and establishing relationships in cross-domains.

This review paper is an attempt to highlight the knowledge gap in processing the PDF tables and its accessibility for blind and visually impaired individuals. An efficient and open-source solution for making PDF documents accessible to blind and visually impaired people needs to exploit the heuristics, ontologies, machine learning, and deep learning by using open-source libraries and tools for understanding and interpreting the tabular content in order to reduce information overload.

ENDNOTES

[1] Roya Rastan, "Automatic Tabular Data Ex WCAG traction and Understanding" (PhD diss., University of New South Wales, 2017).

[2] Mark T. Maybury, "Communicative Acts for Explanation Generation," *International Journal of Man-Machine Studies* 37, no. 2 (1992): 135–72.

[3] Patricia Wright, "The Comprehension of Tabulated Information: Some Similarities between Reading Prose and Reading Tables," *NSPI Journal* 19, no. 8 (1980): 25–29, https://doi.org/10.1002/pfi.4180190810.

[4] Jean-Claude Guédon et al., *Future of Scholarly Publishing and Scholarly Communication: Report of the Expert Group to the European Commission* (Brussels: European Commission, Directorate-General for Research and Innovation, 2019), https://doi.org/10.2777/836532.

[5] World Health Organization, *World Report on Vision*, October 8, 2019, https://www.who.int/publications-detail/world-report-on-vision/.

[6] Mireia Ribera Turró, "Are PDF Documents Accessible?" *Information Technology and Libraries* 27, no. 3 (2008): 25–43, https://doi.org/10.6017/ital.v27i3.3246.

[7] Kyunghye Yoon, Laura Hulscher, and Rachel Dols, "Accessibility and Diversity in Library and Information Science: Inclusive Information Architecture for Library Websites," *Library Quarterly* 86, no. 2 (2016): 213–29, https://doi.org/10.1086/685399.

[8] Iris Xie et al., "Using Digital Libraries Non-Visually: Understanding the Help-Seeking Situations of Blind Users," *Information Research* 20, no. 2 (2015): 673.

[9] Heidi M. Schroeder, "Implementing Accessibility Initiatives at the Michigan State University Libraries," *Reference Services Review* 46, no. 3 (2018): 399–413, https://doi.org/10.1108/RSR-04-2018-0043.

[10] Joanne Oud, "Accessibility of Vendor-Created Database Tutorials for People with Disabilities," *Information Technology and Libraries* 35, no.4 (2016): 7–18, https://doi.org/10.6017/ital.v35i4.9469.

[11] Rakesh Babu and Iris Xie, "Haze in the Digital Library: Design Issues Hampering Accessibility for Blind Users," *Electronic Library* 35, no. 5 (2017): 1052–65, https://doi.org/10.1108/EL-10-2016-0209.

[12] Rachel Wittmann et al., "From Digital Library to Open Datasets," *Information Technology and Libraries* 38, no. 4 (2019): 49–61, https://doi.org/10.6017/ital.v38i4.11101.

[13] Xinxin Wang, "Tabular Abstraction, Editing, and Formatting" (PhD diss., University of Waterloo, 1996).

[14] Rastan, "Automatic Tabular Data Extraction," 25.

[15] Azadeh Nazemi, "Non-Visual Representation of Complex Documents for Use in Digital Talking Books" (PhD diss., Curtin University, 2015).

[16] Rastan, "Automatic Tabular Data Extraction," 14.

[17] Max Göbel et al., "ICDAR 2013 Table Competition," in *2013 12th International Conference on Document Analysis and Recognition* (2013): 1449–53, https://doi.org/10.1109/ICDAR.2013.292.

[18] Burcu Yildiz, Katharina Kaiser, and Silvia Miksch, "pdf2table: A Method to Extract Table Information from PDF Files," in *Proceedings of the 2nd Indian International Conference on Artificial Intelligence* (IICAI, 2005): 1773–85; Tamir Hassan and Robert Baumgartner, "Table Recognition and Understanding from PDF Files," in *Ninth International Conference on*

Document Analysis and Recognition (ICDAR 2007) (2007): 1143–47, https://doi.org/10.1109/ICDAR.2007.4377094; Alexey Shigarov et al., "Tabbypdf: Web-Based System for PDF Table Extraction," in *International Conference on Information and Software Technologies* (Springer International Publishing, 2018): 257–69, https://doi.org/10.1007/978-3-319-99972-2_20.

[19] Minghao Li et al., "TableBank: Table Benchmark for Image-Based Table Detection and Recognition," preprint, *arXiv:1903.01949*; Sebastian Schreiber et al., "Deepdesrt: Deep Learning for Detection and Structure Recognition of Tables in Document Images," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (2017): 1162–67, https://doi.org/10.1109/ICDAR.2017.192.

[20] Zewen Chi et al., "Complicated Table Structure Recognition," preprint, *arXiv:1908.04729*.

[21] Michael Cafarella et al., "Ten Years of Webtables," in *Proceedings of the VLDB Endowment* 11, no. 12 (August 2018): 2140–49, https://doi.org/10.14778/3229863.3240492.

[22] Shah Khusro, Asima Latif, and Irfan Ullah. "On Methods and Tools of Table Detection, Extraction and Annotation in PDF Documents," *Journal of Information Science* 41, no. 1 (2015): 41–57, https://doi.org/10.1177/0165551514551903.

[23] Hassan, "Table Recognition and Understanding"; Richard Zanibbi, Dorothea Blostein, and James R Cordy, "A Survey of Table Recognition," *Document Analysis and Recognition* 7, no. 1 (2004): 1–16, https://doi.org/10.1007/s10032-004-0120-9; Andreiwid Sheffer Corrêa and Pär-Ola Zander, "Unleashing Tabular Content to Open Data: A Survey on PDF Table Extraction Methods and Tools," in *Proceedings of the 18th Annual International Conference on Digital Government Research* (June 2017): 54–63, https://doi.org/10.1145/3085228.3085278; Christopher Clark and Santosh Divvala, "Looking beyond Text: Extracting Figures, Tables and Captions from Computer Science Papers" (paper, AAAI Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, January 25–26, 2015).,

[24] Ermelinda Oro and Massimo Ruffolo, "PDF–Trex: An Approach for Recognizing and Extracting Tables from PDF Documents," in *2009 10th International Conference on Document Analysis and Recognition (ICDAR)* (2009): 906–10, https://doi.org/10.1109/ICDAR.2009.12.

[25] Vidhya Govindaraju, Ce Zhang, and Christopher Ré, "Understanding Tables in Context Using Standard NLP Toolkits," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Sofia, Bulgaria: Association for Computational Linguistics, August 2013): 658–64.

[26] Nikola Milosevic et al., "Disentangling the Structure of Tables in Scientific Literature," in *Natural Language Processing and Information Systems, NLDB 2016, Lecture Notes in Computer Science* 9612 (Springer, Cham), https://doi.org/10.1007/978-3-319-41754-7_14.

[27] Rastan, "Automatic Tabular Data Extraction," 48.

[28] Alexey Shigarov, Andrey Mikhailov, and Andrey Altaev, "Configurable Table Structure Recognition in Untagged PDF Documents," in *Proceedings of the 2016 ACM Symposium on Document Engineering*, (2016): 119–22, https://doi.org/10.1145/2960811.2967152.

[29] Shigarov et al., "Tabbypdf," 262, 263, 265.

[30] Dae Hyun Kim et al., "Facilitating Document Reading by Linking Text and Tables," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (October 2018): 423–34, https://doi.org/10.1145/3242587.3242617.

[31] Hassan, "Table Recognition and Understanding," 1145.

[32] Jing Fang et al., "A Table Detection Method for Multipage PDF Documents via Visual Separators and Tabular Structures," in *2011 International Conference on Document Analysis and Recognition* (2011): 779–83, https://doi.org/10.1109/ICDAR.2011.304.

[33] Bahadar Ali and Shah Khusro, "A Divide-and-Merge Approach for Deep Segmentation of Document Tables," in *Proceedings of the 10th International Conference on Informatics and Systems* (May 2016): 43–49, https://doi.org/10.1145/2908446.2908473.

[34] Wenyuan Xue et al., "Table Analysis and Information Extraction for Medical Laboratory Reports," in *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)* (2018): 193–99, https://doi.org/10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00043.

[35] Roya Rastan, Hye-Young Paik, and John Shepherd, "TEXUS: A Unified Framework for Extracting and Understanding Tables in PDF Documents," *Information Processing & Management* 56, no. 3 (2019): 895–918, https://doi.org/10.1016/j.ipm.2019.01.008.

[36] Dafang He et al., "Multi-scale Multi-task FCM for Semantic Page Segmentation and Table Detection," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (2017): 254–61, https://doi.org/10.1109/ICDAR.2017.50.

[37] Jing Fang et al., "Table Header Detection and Classification," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (July 2012): 599–605.

[38] He et al., "Multi-scale Multi-task," 255.

[39] Martha O. Perez-Arriaga, Trilce Estrada, and Soraya Abad-Mota, "TAO: System for Table Detection and Extraction from PDF Documents," Florida Artificial Intelligence Research Society Conference, North America (2016).

[40] Saman Arif and Faisal Shafait, "Table Detection in Document Images using Foreground and Background Features," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*, (2018): 1–8, https://doi.org/10.1109/DICTA.2018.8615795.

[41] Schreiber et al., "Deepdesrt," 1163, 1164.

[42] Shoaib Ahmed Siddiqui et al., "Decnt: Deep Deformable CNN for Table Detection," *IEEE Access* 6 (2018): 74151–61, https://doi.org/10.1109/ACCESS.2018.2880211.

[43] Chi et al., "Complicated Table Structure Recognition."

[44] Rahul Anand, Hye-Young Paik, and Cheng Wang, "Integrating and Querying Similar Tables from PDF Documents Using Deep Learning," 2019, preprint, *arXiv:1901.04672*.

[45] Jiaoyan Chen et al., "Colnet: Embedding the Semantics of Web Tables for Column Type Prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence* 33, no. 1: 29–36, https://doi.org/10.1609/aaai.v33i01.330129.

[46] Ziqi Zhang, "Towards Efficient and Effective Semantic Table Interpretation," in *International Semantic Web Conference* (2014): 487–502, https://doi.org/10.1007/978-3-319-11964-9_31.

[47] Ivan Ermilov, Sören Auer, and Claus Stadler, "User-Driven Semantic Mapping of Tabular Data," in *Proceedings of the 9th International Conference on Semantic Systems* (September 2013): 105–12, https://doi.org/10.1145/2506182.2506196.

[48] Martha O Perez-Arriaga, Trilce Estrada, and Soraya Abad-Mota, "Table Interpretation and Extraction of Semantic Relationships to Synthesize Digital Documents," in *Proceedings of the 6th International Conference on Data Science, Technology and Application—DATA* (2017): 223–32, https://doi.org/10.5220/0006436902230232.

[49] Varish Mulwad, "TABEL—A Domain-Independent and Extensible Framework for Inferring the Semantics of Tables," (PhD diss., University of Maryland, 2015).

[50] Syed Tahseen Raza Rizvi et al., "Ontology-based Information Extraction from Technical Documents," in *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART)* (2018): 493–500, https://doi.org/10.5220/0006596604930500.

[51] Corrêa and Zander, "Unleashing Tabular Content to Open Data," 55.

[52] Irfan Ullah et al., "An Overview of the Current State of Linked and Open Data in Cataloging," *Information Technology and Libraries* 37, no. 4 (2018): 47–80, https://doi.org/10.6017/ital.v37i4.10432.

[53] Nosheen Fayyaz, Irfan Ullah, and Shah Khusro, "On the Current State of Linked Open Data: Issues, Challenges, and Future Directions," *International Journal on Semantic Web and Information Systems (IJSWIS)* 14, no. 4 (2018): 110–28, https://doi.org/10.4018/IJSWIS.2018100106.

[54] Govindaraju, Zhang, and Ré , "Understanding Tables in Context Using Standard NLP Toolkits," 660, 661.

[55] Perez-Arriaga, Estrada, and Abad-Mota, "Table Interpretation and Extraction," 227.

[56] Kim et al., "Facilitating Document Reading," 425, 426.

[57] Rastan, Pail, and Shepherd, "TEXUS," 906.

[58] Nikola Milosevic et al., "A Framework for Information Extraction from Tables in Biomedical Literature," *International Journal on Document Analysis and Recognition (IJDAR)* 22, no. 1 (2019): 55–78, https://doi.org/10.1007/s10032-019-00317-0.

[59] Chi et al., "Complicated Table Structure Recognition."

[60] Wenhao Yu et al., "Tablepedia: Automating PDF Table Reading in an Experimental Evidence Exploration and Analytic System," in *The World Wide Web Conference* (May 2019): 3615–19, https://doi.org/10.1145/3308558.3314118.

[61] Anand, Paik, and Wang, "Integrating and Querying Similar Tables."

[62] Turró, "Are PDF Documents Accessible?" 2, 4.

[63] Nazemi, "Non-Visual Representation of Complex Documents," 110, 111, 112, 118.

[64] Juan Cao, "Generating Natural Language Descriptions from Tables," *IEEE Access* 8 (2020): 46206–16, https://doi.org/10.1109/ACCESS.2020.2979115.

[65] Maartje ter Hoeve et al., "Conversations with Documents: An Exploration of Document-Centered Assistance," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (March 2020): 43–52, https://doi.org/10.1145/3343413.3377971.

[66] Guédon et al., "Future of Scholarly Publishing," 42.

[67] W3C, "WCAG 2.0."

[68] World Health Organization, "World Report on Vision"; David Reinsel, John Gantz, and John Rydning, "Data Age 2025: The Digitization of the World, From Edge to Core," IDC white paper, #US44413318 (Framingham, MA: IDC, November 2018), https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf/.

[69] Rastan, "Automatic Tabular Data Extraction," 18, 19.

[70] Arif and Shafait, "Table Detection in Document Images," 1.

[71] Ana Costa e Silva, "Parts that Add up to a Whole: A Framework for the Analysis of Tables," (PhD diss., Edinburgh University, UK, 2010).

[72] Milosevic et al., "A Framework for Information Extraction from Tables," 60.

[73] Rastan, "Automatic Tabular Data Extraction," 14.

[74] Chen et al., "Colnet," 31.

[75] Mulwad, "TABEL," 23; Zewen, "Complicated Table Structure Recognition."

[76] Siddiqui et al., "Decnt," 74160.

[77] David W Embley, Sharad Seth, and George Nagy, "Transforming Web Tables to a Relational Database," *2014 22nd International Conference on Pattern Recognition* (2014) 2781–86, https://doi.org/10.1109/ICPR.2014.479.

[78] Milosevic et al., "A Framework for Information Extraction from Tables," 56.

[79] Milosevic et al., "A Framework for Information Extraction from Tables," 55, 56.

[80] Kim et al., "Facilitating Document Reading," 432.

[81] Chen et al., "Colnet," 36.

[82] Asima Latif et al., "A Hybrid Technique for Annotating Book Tables," *Int. Arab J. Inf. Technol* 15, no. 4 (2018): 777–83.

[83] Rastan, Paik, and Shepherd, "TEXUS," 909.

[84] Milosevic et al., "A Framework for Information Extraction from Tables," 61, 62, 65, 66.

[85] Rizvi et al., "Ontology-based Information Extraction," 496.

[86] Siddiqui et al., "Decnt," 74160.

[87] Max Göbel et al., "A Methodology for Evaluating Algorithms for Table Understanding in PDF Documents," in *Proceedings of the 2012 ACM Symposium on Document Engineering* (September 2012): 45–48, https://doi.org/10.1145/2361354.2361365.

[88] Rastan, Paik, and Shepherd, "TEXUS," 917.

[89] David Pinto et al., "Table Extraction Using Conditional Random Fields," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (July 2003): 235–42, https://doi.org/10.1145/860435.860479.

[90] Nazemi, "Non-Visual Representation of Complex Documents," 118–44; W3C, "WCAG 2.0."

[91] Ullah et al., "Current State of Linked and Open Data in Cataloging," 47, 48.

[92] Julius T. Nganji, "The Portable Document Format (PDF) Accessibility Practice of Four Journal Publishers," *Library and Information Science Research* 37, no.3 (2015): 254–62, https://doi.org/10.1016/j.lisr.2015.02.002.

[93] Julius T. Nganji, "An Assessment of the Accessibility of PDF Versions of Selected Journal Articles Published in a WCAG 2.0 Era (2014–2018)," *Learned Publishing* 31, no. 4 (2018): 391–401, https://doi.org/10.1002/leap.1197.

[94] Wittmann et al., "From Digital Library to Open Datasets," 49, 50.

[95] Yan Han and Xueheng Wan, "Digitization of Text Documents Using PDF/A," *Information Technology and Libraries* 37, no. 1 (2018): 52–64, https://doi.org/10.6017/ital.v37i1.9878.

[96] Asim Ullah, Shah Khusro, and Irfan Ullah, "Bibliographic Classification in the Digital Age: Current Trends & Future Directions," *Information Technology and Libraries* 36, no. 3 (2017): 48–77, https://doi.org/10.6017/ital.v36i3.8930.

[97] Xie et al., "Using Digital Libraries Non-Visually," paper 673.

[98] Babu and Xie, "Haze in the Digital Library," 1057–59.

[99] Iris Xie et al., "Enhancing Usability of Digital Libraries: Designing Help Features to Support Blind and Visually Impaired Users," *Information Processing and Management* 57, no. 3 (2020): 102110, https://doi.org/10.1016/j.ipm.2019.102110.

[100] Chen et al., "Colnet," 31, 32.

[101] Kim et al., "Facilitating Document Reading," 432.

[102] Milosevic et al., "A Framework for Information Extraction from Tables," 61.

[103] Rizvi et al., "Ontology-based Information Extraction," 496.

[104] Embley, Seth, and Nagy, "Transforming Web Tables to a Relational Database," 2783; Milosevic et al., "A Framework for Information Extraction from Tables," 60.

[105] Nicholas J Tierney and Karthik Ram, "A Realistic Guide to Making Data Available Alongside Code to Improve Reproducibility," preprint, *arXiv:2002.11626*.

[106] Rastan, Paik, and Shepherd, "TEXUS," 917.

[107] Nazemi, "Non-Visual Representation of Complex Documents," 118–44; W3C, "WCAG 2.0."

[108] Mexhid Ferati and Wondwossen M. Beyene, "Developing Heuristics for Evaluating the Accessibility of Digital Library Interfaces," in *Universal Access in Human–Computer Interaction, Design and Development Approaches and Methods, UAHCI 2017, Lecture Notes in Computer Science* 10277 (Springer, Cham), https://doi.org/10.1007/978-3-319-58706-6_14.

[109] Ullah et al., "Current State of Linked and Open Data in Cataloging," 64.