JLSC

ISSN 2162-3309 | JLSC is published by the Iowa State University Digital Press | http://jlsc-pub.org

Volume 13, 1 (2025)

Making Your Repository (More) Accessible

Michelle Emanuel

Emanuel, M. (2025). Making Your Repository (More) Accessible. *Journal of Librarianship and Scholarly Communication*, 13(1), eP18308. https://doi.org/10.31274/jlsc.18308

This article underwent semi-anonymous peer review in accordance with JLSC's peer review policy.

ISSN 2162-3309 10.31274/jlsc.18308

PRACTICE ARTICLE

Making Your Repository (More) Accessible

Michelle Emanuel
University of Mississippi Libraries

ABSTRACT

Introduction: As colleges and universities make increasing and overdue efforts under the auspices of access, equity, and inclusion to make their resources accessible to all users, these efforts must extend to the institution's online presence, including its institutional repository. IR managers must first ask what "accessible" means for compliance with university policies as well as the Americans with Disability Act (ADA), immediately followed by plans for both remediating existing content and imposing best practices on new content, amid current workflows and budgetary restraints.
Literature Review: Literature on the topic of accessibility in IRs has mostly focused on the need to make collections accessible and the challenges for doing so. Advice on how to navigate the actual process is harder to come by.
Description of Service: The University of Mississippi established a goal that everything going into its IR would use OCR software to convert images of text into searchable text and create a process by which patrons could request remediation of older content from the IR, whether documents or recordings. A combination of shared tools (including Equidox and SensusAccess) and interdepartmental partnerships has made a significant difference in making these digital collections proactively accessible.
Next Steps: We continue to maintain partnerships with units around campus, made challenging by frequent turnover as in demand specialists take positions at other institutions. Despite our efforts to provide searchable text as a minimum level of service, OCR correction provides tags but not necessarily headings or alt-text. Hopefully future versions of OCR editors will include such features.

Keywords: accessibility, institutional repositories, OCR, captions

Received: 08/15/2024 Accepted: 12/02/2024

INTRODUCTION

The University of Mississippi (UM) launched its institutional repository (IR), eGrove, in October 2018 using Digital Commons (DC) from bepress/Elsevier. In addition to standard fares such as theses, dissertations, and open access faculty publications, the decision was made early in the proposal process to migrate the Libraries' digital collections from a local instance of CONTENTdm to the same platform as the new IR.[1] Not only would this decision remove the requirement for an additional contract negotiation, it would also keep the implementation team from being spread too thin if they only had to learn one new platform. Like many digital collections of archival materials, these collections include assets in multiple formats: text, images, and both audio and visual recordings. Printed archival materials, whether handwritten or professionally published, and images are scanned as high-resolution TIFF files which, when made available online in a digital collection, are duplicated into access PDFs that are easier to download. Any patrons needing a high-resolution image for publication can then be directed to an archivist to arrange licensing. But in either case – whether a TIFF or PDF – these are essentially photographs of text. Without correction of the optical character recognition (OCR), they are completely inaccessible to users with vision impairments. Multiple software solutions can be used to correct OCR, but they all require two very precious resources: time and money. While a document can be remediated using software, in addition to Artificial Intelligence (AI), in our current space and time it also requires a live person to verify that everything has been correctly updated – a process that is both time consuming and expensive. In addition to making printed materials from the IR accessible, we knew we would also need to make our non-commercial recorded items – both audio and video – accessible to users with hearing impairments, whether with captions in an SRT file and/or a transcript. The global pandemic of 2020 showed us the importance of placing our materials online, but anything we digitize for wide distribution must also be accessible to everyone. What does this process entail? What requires a live person, whether full-time staff or part-time student, vs. what can be done by AI? PDFs are among the most prevalent file format in IRs, since theses, dissertations, and articles are among the most commonly collected items. If created using born digital methods, such as Microsoft Word documents or Excel spreadsheets saved as a PDF, those files should at least have features such as headings and tagging[2], and OCR should not need correction. Within a text document, the use of headings and tags provides an

[1] At the end of that year, CONTENTdm's vendor, OCLC, would no longer support local instances of its platform.

[2] Headings are used to create a scaffolding for the document. The document's title is H1, the major sections are each H2, any subsections are H3, etc. Any heading can be repeated except H1; the document can only have one titles. Tags are used with coding to differentiate text, image, table, list, hyperlink, forms, and so on.

infrastructure that allows a screen reader to read any text aloud to a patron with visual impairments in the correct order. Microsoft provides tools within Office 365[3] that allow content creators to add descriptive or alternative (*alt-*) text for images before exporting to PDF. Though these tools have been available since Office 2010, and improved in Office 2016, they are not often used without a mandate from administration. For scanned materials, the workload can be considerably heavier, particularly in OCR correction. Although alt-text can often be found in the captions found underneath an image, mathematical notations, such as those found in many scientific articles or theses, can be extremely difficult to replicate or correct within OCR software. Typically, OCR software such as Abbyy FineReader[4] will highlight any questionable text in a document, depending on the font, or the age of the original print; perhaps a word like "learning" is reading as "leaming", or the number "1" is reading as a lowercase "l" or uppercase "I". The person cleaning the OCR would see this error and correct it. But with mathematical equations beyond basic algebra, this can be as challenging as a language with a non-Roman alphabet, such as Greek or Cyrillic. It is not uncommon for an equation to be scanned as an image instead of text, so that it doesn't have to be corrected, only described. This might help the employee in the short term, but not the user, thus raising the recurring issue of whose responsibility should it be to make the materials accessible? The content creator or the repository manager? In an ideal world, the author would submit a born digital document with appropriate headings, tags, and alt-text for publication in a journal or series. Or the publisher of the journal or series, after formatting and graphic design, would take the extra step and make sure that the work is screen readable. Or the video editor would add corrected captions before uploading to YouTube. But many IR managers know all too well that if faced with the opportunity to add interesting content to the IR in a timely manner – the recording of an important event on campus, or a collection of scanned letters donated to the Archives – the work of OCR and captioning is going to fall to them and to their staff to fit into their existing workflow.

In March 2020, accidentally coinciding with the pandemic, the IR team established a goal that everything going into the IR would use OCR software to convert images of text into searchable text and create a process by which patrons could request remediation of older content from the IR, whether documents or recordings. We were very nervous about how many requests there would be, for what kinds of materials, with what expectations for the turnaround time. But we found that by making a few key partnerships around campus, many of our initial fears were alleviated.

[3] Microsoft Office 365 includes Word, Excel, PowerPoint and other software.

[4] "ABBYY FineReader is an optical character recognition (OCR) system that converts scanned documents, PDF documents, and image files (including digital photos) into editable formats." https://help.abbyy.com/en-us/finereader/12/overview/

LITERATURE REVIEW

Literature on the topic of accessibility in IRs has mostly focused on the need to make collections accessible and the challenges for doing so. Advice on how to navigate the actual process is harder to come by. In articles from librarians in Sweden (Linde et al, 2011) and Croatia (Gudelj et al, 2023), the word "accessibility" is synonymous with "open access" – as in, is an article accessible without a subscription, anywhere in the world, without a paywall? But the other meaning of "accessiblity" – can something be *accessed* or used despite barriers such as physical disabilities or even limited internet access – has since gained importance. During the Pandemic, when campuses and public libraries were closed, the availability of high-speed internet and bandwidth in rural areas for streaming content like classes over Zoom, as well as the useability of web-based content on handheld devices became even more important than before. Walker and Keenan (2015) examine collections in both DC and CONTENTdm in terms of their accessibility with the observations of a visually impaired user. They end the article with a list of suggestions of ways to make DC more accessible, such as providing conference schedules in the form of an Excel spreadsheet. Darvishy et al (2023) looked at a sampling of PDFs available in Swiss IRs to check if they had at least "minimal" accessibility features, including both tags and hierarchical heading structures. These elements help a machine reader determine the section breaks of the document or differentiate between a paragraph and a table within the text. Less than 11% of their sample met this standard. Waugh et al (2020) examined accessibility practices in academic libraries and the content found in IRs to "understand the current landscape" and "identify the average level of content accessibility" (3) They found that in addition to the oft-cited needs of more time and more staff to implement best practices, their respondents also decried a lack of directive from their institution's administration. Somehow the IR administrators are supposed to figure it all out for everyone.

Rodriguez (2021) described using a grant-funded project at Florida State University (FSU) to partner with Georgia Tech's Center for Inclusive Design and Innovation (CIDI) to provide captions for audiovisual recordings. FSU was able to upload their files to the CIDI server and receive a time-stamped, plain text Web Video Text Tracks (WebVTT) file or a text file (.docx) that was "ingest-ready for the repository" (3). In the "Lessons Learned and Possible Futures" section (3), he does add that editing those files requires using other software. At UM, a grant-funded project used AI to generate captions for A/V recordings and hired student workers to then edit the outputs (Norris-Davidson & Emanuel, 2023). McLaughlin and Hoops (2021) shared their process to audit the accessibility of Indiana University's repository and the changes to their submission workflow. But without receiving additional resources to improve the IR's accessibility, "the majority of responsibility for the accessibility of the repositories' content remained on the users" (45). Hovious and Wang (2024) conducted an audit of PDFs of a

sample of articles found in EBSCO's Library and Information Source database to check if the aggregator provided HTML format, the optimal format for screen readers, and PDF accessibility conformance, or following the guidelines set by the Universal Accessibility standard (PDF/UA) or the Web Content Accessibility Guidelines (WCAG). In the end, they suggest that "libraries may wish to consider document remediation as part of their accessibility services" (9).

DESCRIPTION OF SERVICE

It can be overwhelming to consider what is required to make an entire repository accessible – where do we even start? – but before that process can even begin, there must be a shared understanding of what "accessible" means. It would be easy enough for the remediation of individual PDFs to take over an entire department's work week. There is no shortage of materials in need of the service, and for each one there is a range of granularity in terms of the level of detail needed. Do documents need to be machine readable, requiring that the document be tagged, with the order of paragraphs adjusted to allow a machine to read aloud in columns? Or is it enough that an item be searchable? Are there images that need alt-text added? Is the service provided by the repository an all-or-nothing proposition? We had to decide what our end goals should be for accessibility and be honest about what our team could consistently provide. A "blue sky" expectation, if financial and staffing resources were no object, would be to say that from a specific date going forward, everything uploaded to the IR would need to be accessible – however that is locally defined – and we would find ways to work through the backlog of collections either migrated from the previous platform or gathered during the first years of collecting for the IR. When we estimated that such an approach might never have an actual end date, the idea became immediately overwhelming and terrifying. But we also knew that the other end of the spectrum, to say, "We'll cross that bridge when we come to it or if someone asks for it," was no longer an acceptable game plan if we wanted to make our library's services more inclusive and equitable.

For our institution, we have found that somewhere between those two extremes works best for us. We compiled a best practices section (https://guides.lib.olemiss.edu/egrove/accessibility) collected from articles and websites from a number of aspirational and peer institutions to add to the Libguide (Springshare) that we had already created for the IR. We wanted to encourage content creators to be mindful about adding headings and tags to their text and providing alt-text with their images. We also wanted to convince our colleagues in Archives and Special Collections that the digitization of primary sources, some of which are handwritten, would also need to include transcriptions for text as well as audio/video before submitting them to add to the repository. But being realistic, we also knew we would need to have a system in place to accommodate requests for remediation of items when needed. But how would we do that in a sustainable manner?

Our first step was to determine what "accessible" needed to mean for our purposes. We looked to another unit across campus for advice: the specialists already hired by UM for digital accessibility services. A team spread across multiple departments, the specialists included one person in the Information Technology department ("Campus IT") and another in the Division of Outreach which administers online learning. They helped us understand the standards we should be looking for in making our content accessible to as many users as possible and showed us tools we could use to reach our goals. They clarified the meaning of "machine readable", that an automated voice could read a text document out loud for a user with visual impairments using the tags in each paragraph to read in a specific sequence whether in a column or from left to right. Furthermore, the machine will also read the alt-text to describe any image or decorative graphic to a user who cannot see it. In effect, all machine-readable content includes searchable text, but not all searchable text is machine-readable.

We soon decided that moving forward, on the advice of the campus accessibility specialists, we would need to commit to asking that all incoming content be searchable text, at an absolute minimum. Text that has been created using software such as Microsoft Word, even if exported as a PDF, automatically becomes searchable. We realized that in most cases, born-digital items would not pose as many problems as scanned materials. Any new scanned materials that we might need to add, including back issues of journals and any digitized archives, would require significant OCR correction as part of the collection preparation, and not as an afterthought. The impact of this decision would be nearly immediate; cleaning the OCR at the time of ingest would make everything in the IR that much more discoverable, therefore driving traffic to our site from Google. We would use Abbyy FineReader to correct the OCR because it was the most affordable option in providing licenses for multiple users, including our graduate assistant. But if the text runs in columns, as it does in many academic journals, or has images requiring alt-text, we would take the extra step and use Equidox[5] to correct the tags, headings, and any missing alt-text. As a PDF remediation platform, Equidox allows for the tags of a document – such as text, image, table – to be changed, as well as reordered. By adjusting the order of paragraphs (1, 2, 3, 4) to read from top to bottom instead of the default setting of left to right, the remediated document can now be read aloud by a screen reader. As many of the documents we receive are not printed using columns, for example, most theses and dissertations, we only need to use Equidox for a small percentage of our workflow. Our campus digital accessibility specialists allow us to share their license, since we are unlikely to need multiple seats at the same time. In the event our needs do increase over time, we are prepared to apply for our own license.

As the COVID-19 pandemic hit in 2020, UM's digital accessibility specialists were working to make university content online as accessible as possible, especially the materials in Blackboard,

[5] Equidox is "a Software as a Service (SaaS) PDF remediation solution" (https://equidox.co/).

UM's Learning Management System (LMS). They secured a campus-wide license for a service from SensusAccess[6] which they would brand as the "UM Document Converter" (Figure 1). This service, available to UM-affiliated students, faculty, and staff, is

> an online document conversion system supporting the transformation of text and image-based file types into different formats. Individuals upload files through a Web interface and select from a variety of output options, including MP3 audio, Braille, or e-book, and text formats including tagged PDF, Microsoft Word and Excel, Rich Text Format, CSV, TXT, and HTML (UM Document Converter webpage)

The affiliated user goes to this webpage (https://accessibility.olemiss.edu/home/document-accessibility-for-the-web/um-document-converter/), uploads a file to convert, chooses an output format, specifies options about that format, and enters an email for delivery of the converted file.

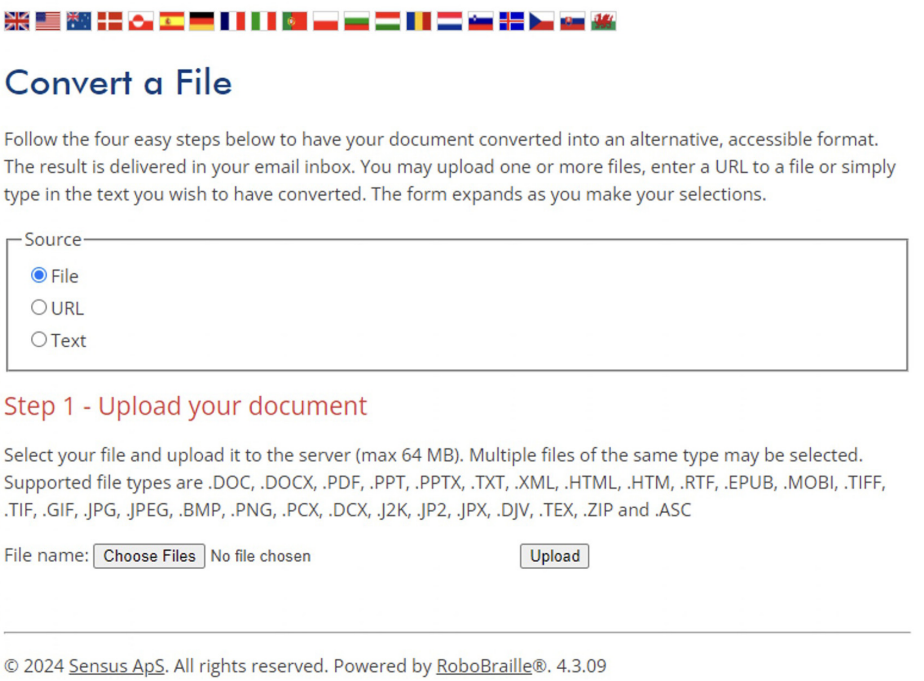

Figure 1. A screenshot from the UM Document Converter (SensusAccess), where a user would upload the file to be converted.

[6] "Depending on the document type of the source document, [SensusAccess offers] a number of conversion options are available for automated document remediation." (https://www.sensusaccess.com/accessibility/)

Unfortunately, SensusAccess has not had much success yet with handwritten archival items. Those still must be done by hand, without software. We soon realized that this would require an additional degree of partnership in-house with our own archivists. We had to convince them that for digital collections, making items accessible is now expected to be part of the collection processing workflow. Even though this would add significant time to processing newly digitized collections and accretions, it made the most sense that this work would come from them because they have more experience reading handwriting from the eras in which they collect.

For users not affiliated with UM, or for any user needing more than just searchable text, such as machine-readable access, we would need to have a request form in place. We created a request form in Qualtrics (Appendix 1) and worked with our DC consultant to find a way to place it in the sidebar as "Request an Accessible Copy" (Figures 2 and 3), in order to appear on all pages. We also asked the consultant to add an additional metadata field for accessibility (Figure 4) on all of our existing structures as well as any new structures moving forward, to indicate if remediation had already been done.

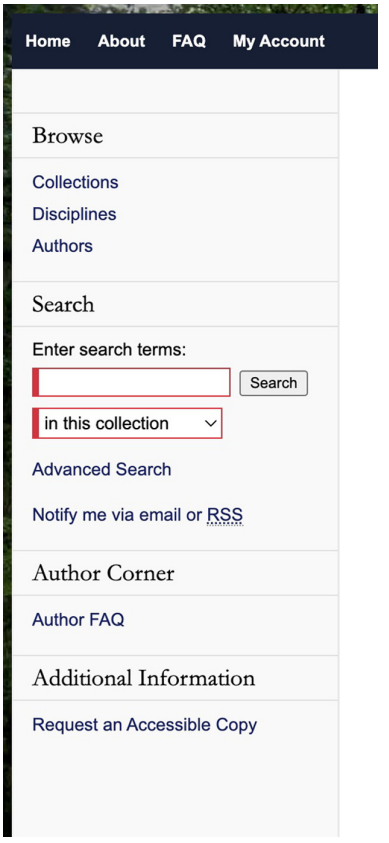

Figure 2. The IR's sidebar in Digital Commons, visible on each page, with a link to "Request an Accessible Copy".

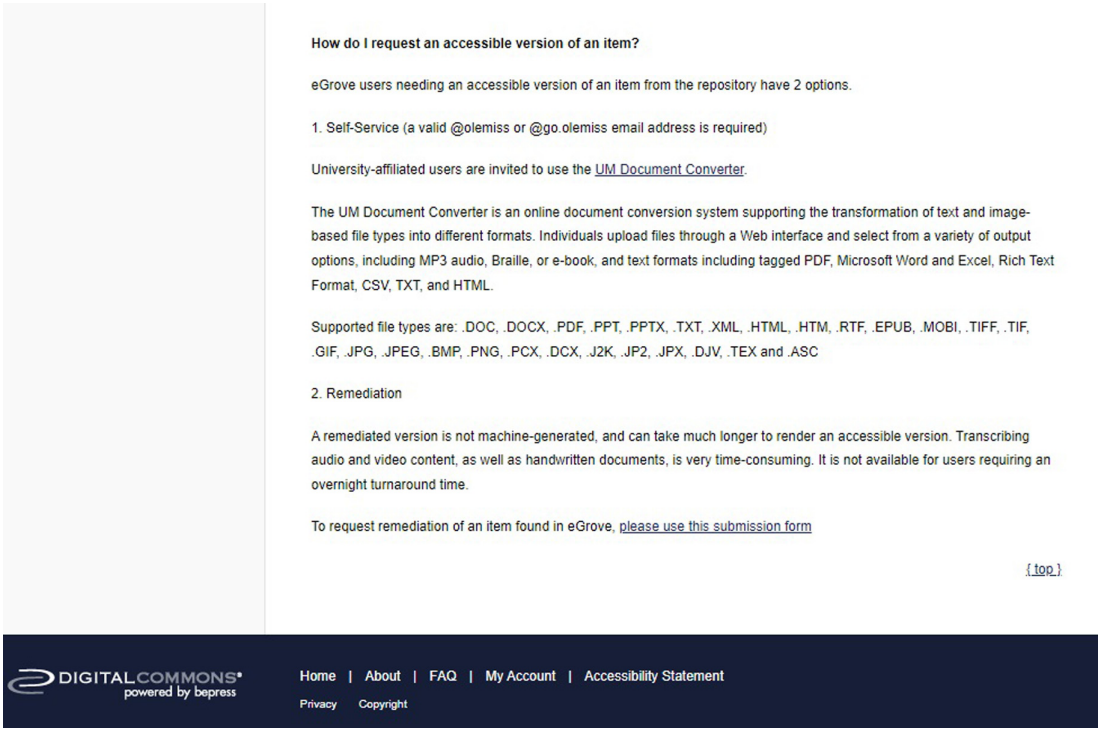

Figure 3. The link from Figure 2 goes to the last item of the IR's FAQ page.

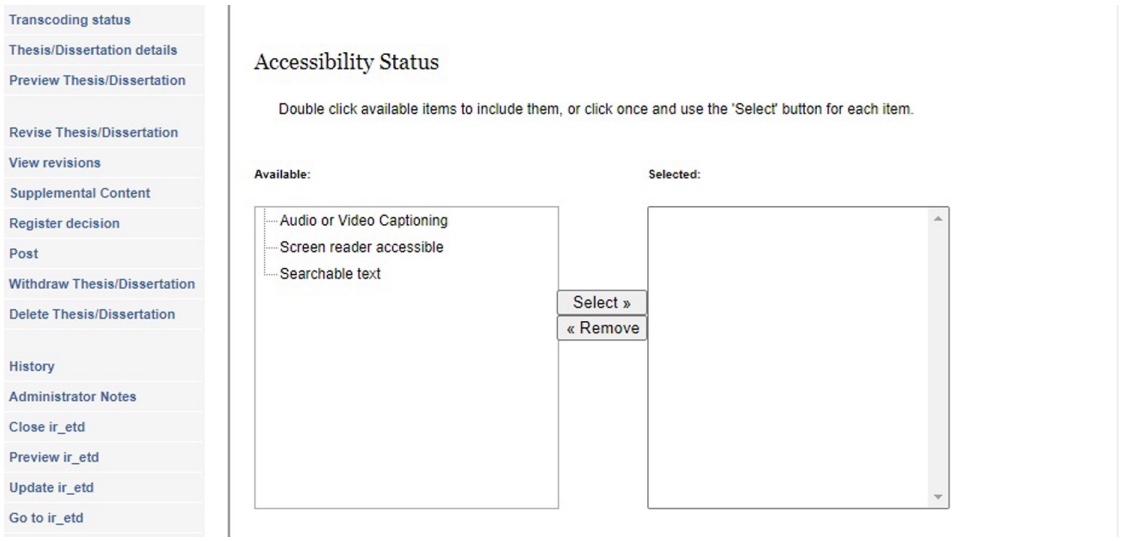

Figure 4. The additional metadata field in Digital Commons to indicate accessibility status.

After months of living in fear of being flooded with time-consuming requests for quickly remediated transcripts of handwritten letters from the nineteenth century, we realized that most of our users only wanted searchable PDFs of typewritten content. SensusAccess was acting as a triage agent. Unfortunately, we have not been able to see any usage statistics to see how many of the platform's requests have been IR-related, nor do we know if our patrons' needs accessibility needs are actually being met. We only know that they have not come back to ask for a remediated document. As helpful as SensusAccess has been to accommodate multiple formats including text and image files, it does not help with audio and video files, which require a different workflow.

Where our minimum standard for print materials is "searchable text", for non-commercial audio and video file accessibility, we aim for both captions in the form of a SubRip Subtitle (.SRT) file and a printable transcript. We used grant funding to experiment with low-cost AI tools, as documented in a toolkit we posted at the end of the funding cycle (Norris-Davidson and Emanuel, 2021). As this technology continues to advance quickly, we have already been able to adjust our workflow with newer tools that are even easier to use. In our first attempts at A/V remediation in 2020-2021, at the time of the first grant, we were using REV.ai as well as Panopto, because they were affordable as well as easy to use with minimal programming expertise. The collections we were trying to remediate were primarily oral history interviews, ranging from 30 minutes to over an hour each. At the time, REV.ai was charging $0.035/minute, meaning the price per video was around $2.10 or less. Meanwhile, during the pandemic, UM added Panopto as a plugin to Blackboard in an effort to help professors to "easily" add captions to the videos they were building for their classes, which had suddenly moved online. For our purposes, Panopto felt like a free license, only because we were not responsible for the cost. Both platforms generate an .SRT file, which is a plain text file with text and timing information for subtitles, as well as a transcript with timestamps, which then must be edited to make a printable document. Converting this text file to a more readable document is very time-consuming, mainly because it requires listening to the original recorded file in real time to differentiate speakers, and remove extra spacing, in addition to repeating any segments that need its text edited, but it is considerably less time-consuming than in the pre-AI era. At least there is a baseline to work from. In our more recent efforts, we have found that Otter.ai focuses on creating a readable transcript document that can then be exported as a PDF file, as well as an .SRT. Its AI can distinguish individual voices, which can be automatically labeled early in the document, considerably cutting down the required time for remediation. When the .SRT file is the primary product, like in Rev.ai or Panopto, conversion to a readable transcript requires removing the time stamp from each line and any extra spacing between lines. Whisper, a more recent transcription product of Open AI, is extremely robust, but it does require someone with a familiarity with GitHub and Python to do an initial set up for the workspace.

NEXT STEPS

By taking more time with new collections and incoming items as they are ingested to ensure that scanned documents include, at a bare minimum, searchable text, we save ourselves much of the anxiety of being slammed with multiple requests for remediated items within a short time span. We assume that born-digital documents, created using digital products such as Microsoft Word or Adobe Acrobat, will at least include searchable text even if alt-text is missing for some of the images. If a request were submitted to make a "screen reader accessible" document, we would then take the time to walk the item through Equidox, where we would adjust the tagging order for each page, add alt text where it is missing, and correct any other issues for each page of the document. Fortunately, of the few requests that we have received since implementing this best practice, they have mostly been requests for transcripts of legacy content, and requests for searchable text from users unaffiliated with the University who cannot upload a document to SensusAccess. We have maintained our partnerships with Student Disability Services, Digital Accessibility Solutions, and the Division of Outreach, which provided access to tools and skill sets we did not realize existed. The primary challenge to working with these departments is that the players change rather frequently. Specialists in this area are highly prized and have great mobility in the professional landscape and seem not to stay at our institution for very long. Fortunately, asking our colleagues in Archives and Special Collections to provide transcriptions for their digital collections has led to a new sense of collaboration to the management of the IR, and we have created new projects for several of their students needing internships. Finally, by adding OCR and transcription for our IR's holdings, we have increased the reach of our content by making it more discoverable, which can be measured through downloads. Despite our best efforts to provide searchable text as a minimum level of service, OCR correction provides tags but not necessarily headings or alt-text. We hope that future versions of OCR editors like Abbyy FineReader will include more features like Equidox.

BIBLIOGRAPHY

Darvishy, Alireza, Rolf Sethe, Ines Engler, Oriane Pierrès, and Juliet Manning. "The State of Scientific PDF Accessibility in Repositories: A Survey in Switzerland." *Learned Publishing 36*, no. 4 (2023): 577–84. https://doi.org/10.1002/leap.1581.

Fox, Elyse and Daina Dickman. "Tackling Accessibility in the IR: A Case Study in Leveraging Remote Student Workers." in *Rethinking Institutional Repositories: Innovations in Management, Collections, and Inclusion*. Edited by Josh C. Cromwell (ACRL, 2023), 209-217.

Gudelj, D., Matošić, M., Poljak, L., Tomić, V., Marušić, A., & Marušić, M. (2023). Accessibility of master's theses at SEA-EU Alliance universities in open access repositories. *International Journal on Digital Libraries*, *24*(4), 229–241. https://doi.org/10.1007/s00799-022-00340-3

Hovious, Amanda, and Congwen Wang. 2024. "Hidden Inequities of Access: Document Accessibility in an Aggregated Database." *Information Technology and Libraries 43* no. 1. https://doi.org/10.5860/ital.v43i1.16661.

Linde, P., Eriksson, J., Kullman, L., Fathli, M., Karlsson, K., Sikström, M., Sköld, Y., & Tång, I. (2011). Accessibility and self-archiving of conference articles: A study on a selection of Swedish institutional repositories. *Information Services & Use*, *31*(3/4), 259–269.

McLaughlin, M., & Hoops, J. (2021). Web Accessibility in the Institutional Repository: Crafting User-Centered Submission Policies. *Serials Librarian*, *80*(1–4), 40–45. https://doi.org/10.1080/0361526X.2021.1868217

Norris-Davidson, A. & Emanuel, M. (2021). Caption This: Creating Efficiency in Audiovisual Accessibility Using Automatic Speech Recognition Toolkit. https://egrove.olemiss.edu/libpubs/18

Norris-Davidson, A. & Emanuel, M. (2023). "Captions for All: Finding a Sustainable Captioning Workflow." in Rethinking Institutional Repositories: Innovations in Management, Collections, and Inclusion. Edited by Josh C. Cromwell. (ACRL, 2023). 187–196.

Rodriguez, D. (2021). Increasing accessibility of audiovisual materials in the institutional repository at Florida State University. *Journal of Academic Librarianship*, *47*(1), N.PAG. https://doi.org/10.1016/j.acalib.2020.102291

University of Mississippi, Digital Accessibility Serivces (2025). UM Document Converter. https://accessibility.olemiss.edu/home/document-accessibility-for-the-web/um-document-converter/

Vavrosky, Laura. "Accessibility of Institutional Repository Content: Current Landscape and Ideas for a Path Forward." Presentation at the Northeast Institutional Repository Day, December 3, 2020. https://repository.escholarship.umassmed.edu/handle/20.500.14038/37447?show=full

Walker, W., & Keenan, T. (2015). Do You Hear What I See? Assessing Accessibility of Digital Commons and CONTENTdm. *Journal of Electronic Resources Librarianship*, *27*(2), 69–87. https://doi.org/10.1080/1941126X.2015.1029395

Waugh, Laura, Colleen Lyon, Abigail Shelton, Kristi Park, William Hicks, and Nerissa Lindsey. "Accessibility in Institutional Repositories." (2020). *Copyright, Fair Use, Scholarly Communication, etc*. 165. https://digitalcommons.unl.edu/scholcom/165.

APPENDIX 1
Text of the Qualtrics form for requesting remediation of a document

(Also available here: https://uofmississippi.qualtrics.com/jfe/form/SV_8JuEVIBxV37mFX8)

This form is used to request a remediated transcription for items found in eGrove, the institutional repository of the University of Mississippi. If you are requesting more than one transcription, please fill out a separate form for each item. We are committed to making our collections as accessible as possible. Please keep in mind that the time required to complete a transcription will depend on the format, size and condition of the original item.

If you do not need a transcript but are looking for information about licensing or reproducing an archival image from one of our photo collections, or digitizing archival content found in a finding aid, please refer to these documents, and **do not submit the transcript request form.**

- Archives and Special Collections: Duplication Request Form
- Archives and Special Collections: Photograph Fee Schedule
- Archives and Special Collections: Other Forms and Policies

What is the name of the item needing transcription?

(URL preferred; title and/or author is also helpful)

Please select the desired format. *

- Searchable Text Document (PDF, Word.docx, text)
- Screen Reader Accessible (including tags and document structure for Screen Reader optimized PDF, Word.docx, HTML)
- Audio or Video Captioning (.srt, Word.docx, HTML)

Your Name

Your Email Address

What is your affiliation with the University of Mississippi? *

- Current student
- Employee (Faculty, Staff)
- Alumnus
- No affiliation

Powered by Qualtrics

** These bullet points represent radio buttons*