J. Willard Marriott Library
THE UNIVERSITY OF UTAH
295 South 1500 East Salt Lake City, Utah 84112-0860 (801) 581-8558

Toolkit to assess OCR'ed historical text in the era of big data

Harish Maringanti
Associate Dean for IT & Digital Library Services

Brain McBride
Head, Digital Infrastructure Development

Bohan Zhu
Software Developer

This project was made possible in part by a 2020 award from the Catalyst Fund at LYRASIS.

I. Goal

Primary goal of the project titled “Toolkit to assess OCR’ed historical text in the era of big data”, generously funded in part by the Catalyst Fund at Lyrasis, is to develop a set of workflows and a framework to enhance the quality of existing OCRed text available in digital repositories in the cultural heritage sector.

A central feature that allows cultural heritage institutions to open up their unique treasures, especially text-based documents, is by converting scanned images of each page into actual computer text – what is commonly called Optical Character Recognition (OCR). The accuracy of OCR technologies considerably impacts the way digital documents are indexed, consulted, and discovered. Accurate OCR data helps in better discoverability and accessibility of content. With the explosive growth in born-digital data that have high OCR accuracy rates, there is a higher risk of historical texts with poor OCR being not discovered in search results. With increase in demand for computational usage of available text, issues with poor quality of OCR in historical texts will prevent them from being used in projects involving computational tools. According to research done by Smith & Cordell, computational text analysis becomes hard because of the high number of misspelled words, and error rates are as high as 40% in nineteenth-century newspapers in English (Smith & Cordell, 2018). What might appear as minor OCR accuracy issues for single items will get amplified in the era of big data, where tools are created to work with content at scale. Digitized content with less accurate OCR data is not only likely to suffer in lower search rankings (poor discoverability) but users may also turn away from using such content to avoid frustrations. For example, if because of OCR errors, text that is supposed to be “Lincoln” is rendered as “Linco1n” in a collection (Torget, et, al, 2011), the OCR errors will disrupt search results significantly and users will not be able to locate the materials they are looking for.

Accurate OCR data can help with many downstream tasks such as text mining, topic modeling, NLP and impact of bad OCR on downstream tasks are being documented (Van Strein, et, al, 2020). If cultural heritage institutions want to make their collections available as datasets, and leverage current technologies such as machine learning, then improving the underlying OCR data is critical. Without accurate OCR data, institutions will struggle to improve discoverability, searchability, and accessibility of their collections, and are likely to be left behind in this age of big data. Though OCR technology has advanced significantly in the last decade, cultural heritage institutions are yet to take advantage of these developments, because of lack of available workflow tools that can assess whether investments in enhancing existing OCR text will yield results they may be looking for (Smith & Cordell, 2018). Our project aims to address this by creating a tool that can be used to assess the OCR accuracy of existing digital collections, and also test the feasibility of using off-the-shelf machine learning options to enhance the OCR accuracy. Note we are not assessing or comparing all the available solutions in this report as this depends on input quality, and current levels of OCR accuracy among other factors. Our goal is to provide a framework & a tool that institutions can use to run their own tests - our tool simplifies the process of the data preparation, management and integrates it with the existing

apis from a few cloud services such as Google & Microsoft, and also includes integration with Tesseract (open source).

II. Model/Process

a. **Technical Design**: We began the design process of the application with a goal of ensuring that we make it as easy for institutions to use as possible. There are many institutions with the resources necessary to implement and develop complex systems but one of our core goals was to ensure that the application can be used by the masses, with that in mind we used all open source libraries and dockerized the application to ensure that the application was easy to use and provided support for Windows, OS X, and Linux platforms.

   We originally scoped our development to include support for Google Vision, Microsoft Computer Vision, and Amazon Computer Vision. However, at the time of this writing, we have only implemented support for Tesseract 4.0, Microsoft Computer Vision, and Google Cloudvision. However, we do plan on implementing support for Amazon Computer Vision at a later day.

b. **Development and Prototyping**: The application is written in Python 3 and is built using the Django framework. There are additional python libraries that were used, these include:
   - requests
   - google-cloud-vision
   - pdf2image
   - Django
   - Pillow
   - Certifi
   - chardet
   - idna
   - setuptools
   - urllib3
   - matplotlib
   - pandas
   - Spacy
   - Spacy en_core_web_md
   - opencv-python
   - tesseract
   - uuid
   - pytesseract

   There are also additional system level dependencies which include:
   - Poppler-utils

- Libxext6
- Tesseract-ocr
- SQLlite

These libraries are then used to support the following functionality for the application:

- Evaluate existing OCR data accuracy (using template)
- Connect to remote resources and download and store objects
- Make calls to Tesseract 4.0, Google Vision, Microsoft Vision APIs for image processing
- Store OCR data from Tesseract, Google Vision, and Microsoft Vision
- Store all JSON response data from Google and Microsoft vision services. This data includes OCR information and other Machine Learning and Artificial Intelligence (ML/AI) information provided by these services
- Run OCR accuracy reports

c. **Identify test collections**: We set out to identify a diverse set of collection materials to be used to help evaluate our toolkit and help us determine if the cloud services provided for improved OCR accuracy. While we had plans to survey the community and gather datasets that could be part of this project, Covid-19 derailed these plans. Instead of a survey, we reached out to our collection partners (several organizations in state of Utah host their collections on Marriott Library's digital platform) for their input. We then identified collections which were scanned decades ago and lacked quality OCR data, both type and handwritten texts, and scans that had both images and text within the object. We reached out to our collection stakeholders, librarians, and other members in our organization to help curate the test collection sets.

d. **Evaluating original and Cloud based OCR results**: The application has the ability to compare the initial OCR word accuracy against the new OCR data collected from the three services the application supports. This functionality is critical to help stakeholders become better informed about the quality of their existing OCR data and to determine if using the cloud services or a newer version of tesseract will help improve the quality of their existing OCR data.

e. **Analysis and generate report**: The application has basic reporting function for users to generate reports that can be used to help inform stakeholders of potential improvements in the quality of OCR data

What did not work

We initially wanted to make the downloading of existing metadata as seamless as possible but that was a challenge. We implemented an OAI-PMH harvesting component to download metadata but there were limitations in capturing all of the necessary metadata necessary to run the tool due to inconsistencies across multiple repositories.. We then developed a template and some instructions on have to facilitate an export while also enabling a template to be

generated based on an OAI-PMH feed. The analysis functionality and the integration into external tools and cloud services went well and we were pleased with the end results.

What worked

Our results demonstrate that it is possible to extract high quality OCR from cloud services. When we tested available service options from Google, Microsoft, and Tesseract on our locally hosted digital collections, google services provided higher quality OCR consistently. Microsoft's Azure results were also better than the current quality of OCR. We could not compare the latest version of Tesseract results with current full-text in the repository because information about the version of Tesseract that was used to generate current OCR was not available.

While detailed instructions on using the tool are provided on the project's webpage (https://github.com/marriott-library), below is a basic overview of the tool along with an example.

Step 1

Institutions that may not have their current OCR metrics, can use the provided python script to analyze the word and character level accuracy information for any item, at a page by page level. For example, when an item Fig 1 is provided as input to the analyze python script, the word and character level accuracy results are shown in Fig 2.

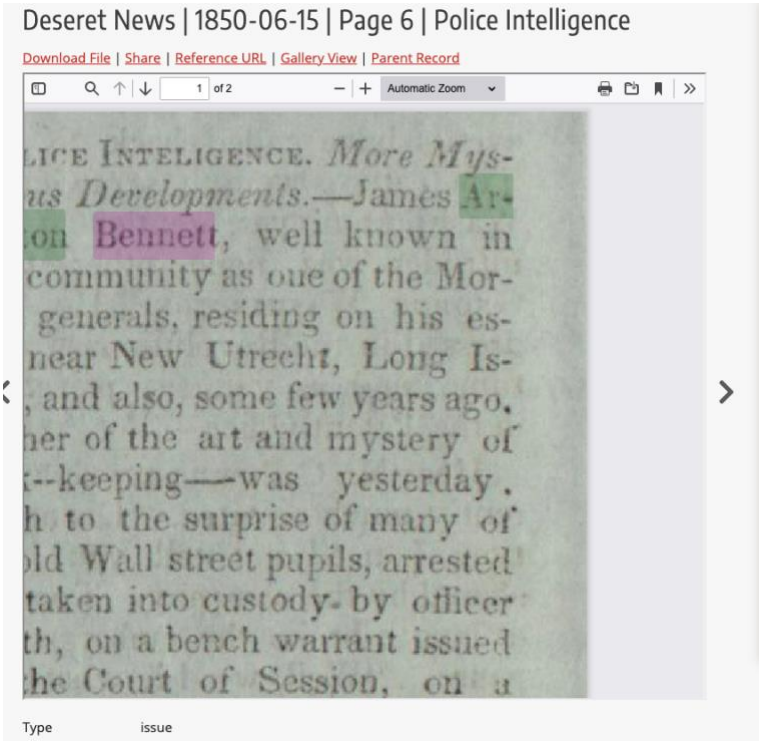


Fig 1: https://newspapers.lib.utah.edu/details?id=2568969&q=James+Arlington+bennett

```
E:\University_of_Utah\Projects\cloudvision-test\cloudvision>python plain_text_ocr_analyze.py
Word Accuracy
Max: 91.65
Min: 91.65
Avg: 91.65
Char Accuracy
Max: 87.24
Min: 87.24
Avg: 87.24
Total Words: 1114
```

Fig 2: word and character level accuracy information

These results are stored in an excel spreadsheet that will be described later in the section.

Step 2

The toolkit has options listed that will let institutions pick a service option (Azure, Google, Tesseract) to run a few tasks (download the OCR text, run accuracy results) as show in figures 3, and 4 below.

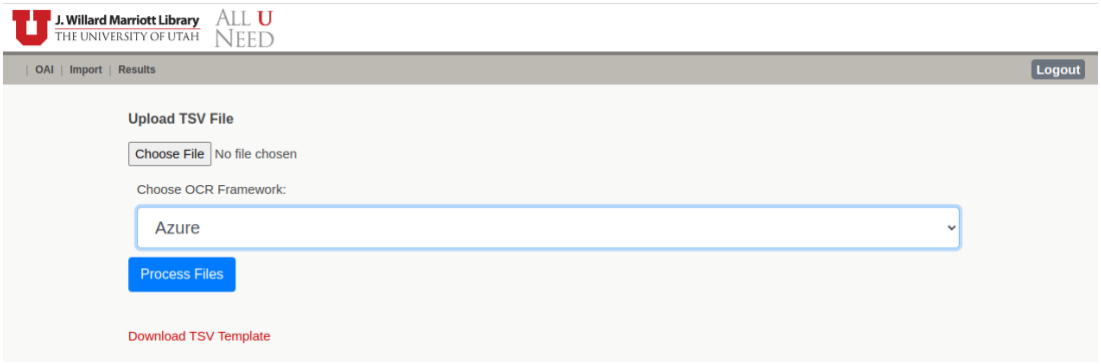


Fig 3: File upload screen

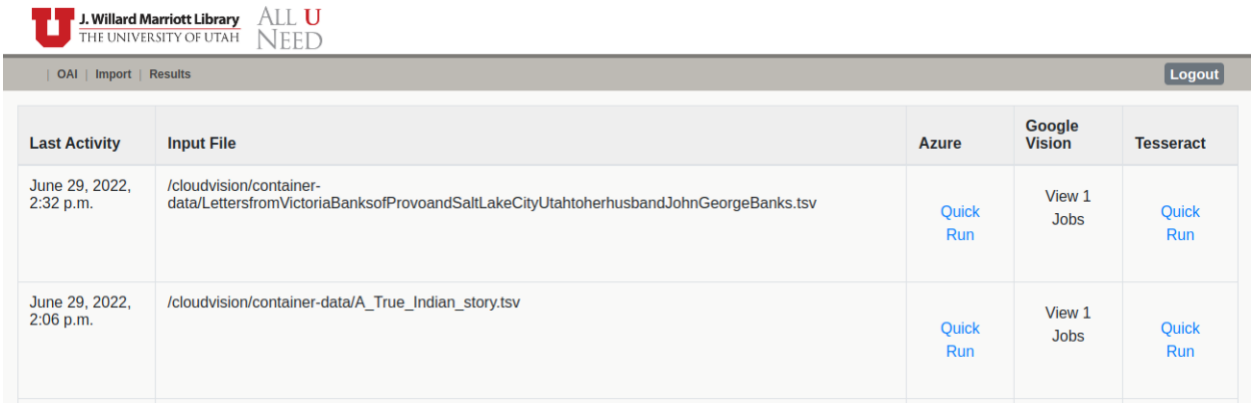


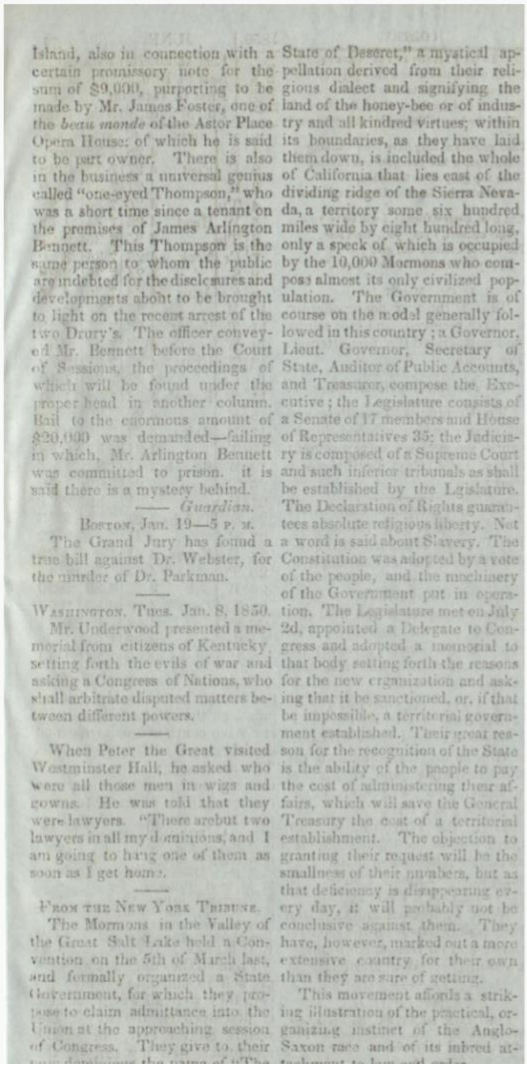
Fig 4: Tasks dashboard

Step 3

The tool lets the users see the results of the OCR analysis, by displaying the word accuracy, character accuracy, total words found, and not found – all displayed next to the item, so users can see where the OCR issues are, as shown in figures 5, and 6 below.

Total 1 file processed, Select an ocr file to view details:

File 1



Island, also in connection with a certain promissory note for the sum of $9,000, purporting to be made by Mr. James Foster, one of the *beau monde* of the Astor Place Opera House: of which he is said to be part owner. There is also in the business a universal genius called "one-eyed Thompson," who was a short time since a tenant on the premises of James Arlington Bennett. This Thompson is the same person to whom the public are indebted for the disclosures and developments about to be brought to light on the recent arrest of the two Drury's. The officer conveyed Mr. Bennett before the Court of Sessions, the proceedings of which will be found under the proper head in another column. Bail to the enormous amount of $20,000 was demanded—failing in which, Mr. Arlington Bennett was committed to prison. it is said there is a mystery behind.

—— *Guardian.*

Boston, Jan. 19—5 P. M.

The Grand Jury has found a true bill against Dr. Webster, for the murder of Dr. Parkman.

——

Washington, Tues. Jan. 8, 1850.

Mr. Underwood presented a memorial from citizens of Kentucky setting forth the evils of war and asking a Congress of Nations, who shall arbitrate disputed matters between different powers.

——

When Peter the Great visited Westminster Hall, he asked who were all those men in wigs and gowns. He was told that they were lawyers. "There arebut two lawyers in all my dominions, and I am going to hang one of them as soon as I get home.

——

From the New York Tribune.

The Mormons in the Valley of the Great Salt Lake held a Convention on the 5th of March last, and formally organized a State Government, for which they propose to claim admittance into the Union at the approaching session of Congress. They give to their State of Deseret," a mystical appellation derived from their religious dialect and signifying the land of the honey-bee or of industry and all kindred virtues; within its boundaries, as they have laid them down, is included the whole of California that lies east of the dividing ridge of the Sierra Nevada, a territory some six hundred miles wide by eight hundred long, only a speck of which is occupied by the 10,000 Mormons who compose almost its only civilized population. The Government is of course on the model generally followed in this country ; a Governor, Lieut. Governor, Secretary of State, Auditor of Public Accounts, and Treasurer, compose the Executive ; the Legislature consists of a Senate of 17 members and House of Representatives 35; the Judiciary is composed of a Supreme Court and such inferior tribunals as shall be established by the Legislature. The Declaration of Rights guarantees absolute religious liberty. Not a word is said about Slavery. The Constitution was adopted by a vote of the people, and the machinery of the Government put in operation. The Legislature met en July 2d, appointed a Delegate to Congress and adopted a memorial to that body setting forth the reasons for the new organization and asking that it be sanctioned, or, if that be impossible, a territorial government established. Their great reason for the recognition of the State is the ability of the people to pay the cost of administering their affairs, which will save the General Treasury the cost of a territorial establishment. The objection to granting their request will be the smallness of their numbers, but as that deficiency is disappearing every day, it will probably not be conclusive against them. They have, however, marked out a more extensive country for their own than they are sure of getting.

This movement affords a striking illustration of the practical, organizing instinct of the Anglo-Saxon race and of its inbred at-

Page	Word Accuracy	Char Accuracy	Total Words	Total Not Founds
1	96.09%	94.17%	741	29
2	93.83%	88.39%	162	10

Island, also in connection with a State of Deseret," a mystical ap- certain promissory note for the pellation derived from their reli- sum of $9,000, purporting to be gious dialect and signifying the made by Mr. James Foster, one of land of the honey-bee or of indus- the beau monde of the Astor Place try and all kindred virtues; within Opera House: of which he is said its boundaries, as they have laid to be part owner. There is them down, is included the whole also in the business a universal genius of California that lies east of the called "one-eyed Thompson," who dividing ridge of the Sierra was a short time since a tenant on da, a territory some six hundred the premises of James Arlington miles wide by eight hundred long, Bennett. This Thompson is the only a speck of which is occupied same person to whom the public by the 10,000 Mormons who com- are indebted for the disclosures and pose almost its only civilized pop- developments about to be brought ulation. The Government is of to light on the recent arrest of the course on the model generally fol- two Drury's. The officer convey- lowed in this country; a Governor, ed Mr. Bennett before the Court Lieut. Governor, Secretary of of Sessions, the proceedings of State, Auditor of Public Accounts, which will be found under the and Treasurer, compose the Exe- proper head in another column. cutive; the Legislature consists of Bail to the enormous amount of a Senate of 17 members and House $20,000 was demanded-failing of Representatives 35; the Judicia- in which, Mr. Arlington Bennett ry is composed of a Supreme Court was committed to prison. it is and such inferior tribunals as shall said there is a mystery behind. be established by the Legislature. The Declaration of Rights guaran- tees absolute religious liberty. Not a word is said about Slavery. The Constitution was adopted by a vote of the people, and the machinery of the Government put in opera- tion. The Legislature met en July 2d, appointed a Delegate to Con- gress and adopted a memorial to that body setting forth the reasons for the new organization and ask- ing that it be sanctioned, or, if that be impossible, a territorial govern- ment established. Their great rea- son for the recognition of the State is the ability of the people to pay the cost of administering their af- fairs, which will save the General Treasury the cost of a territorial establishment. The objection to granting their request will be the smallness of their numbers, but as that deficiency is disappearing ev- ery day, it will probably not be The Mormons in the Valley of conclusive against them. They the Great Salt Lake held a Con- have, however, marked out a more vention on the 5th of March extensive country for their own last, and formally organized a State, than they are sure of getting. Government, for which they pro- This movement affords a strik- pose to claim admittance into the ing illustration of the practical, or- Union at the approaching session ganizing instinct of the Anglo- of Congress. They give to their Saxon race and of its inbred at- new dominions the name of The tachment to law and order. Guardian. BOSTON, Jan. 19-5 P. M. The Grand Jury has found a true bill against Dr. Webster, for the murder of Dr. Parkman. FROM THE NEW YORK TRIBUNE. WASHINGTON, Tues. Jan. 8, 1850. Mr. Underwood presented a me- morial from citizens of Kentucky, setting forth the evils of war and asking a Congress of Nations, who shall arbitrate disputed matters be- tween different powers. When Peter the Great visited Westminster Hall, he

Fig 5: example 1, detailed results screen

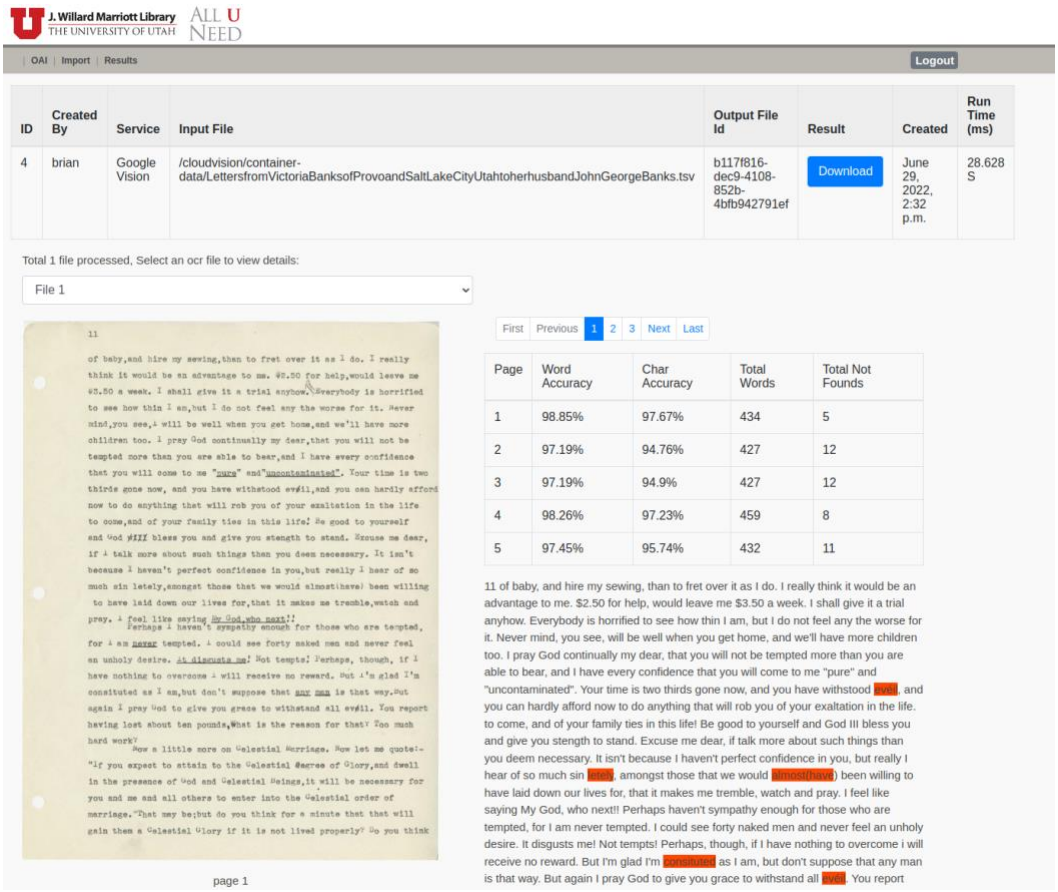


Fig 6: example 2, detailed results screen

Step 4

Finally, all the results from OCR analysis are stored in a spreadsheet for easy viewing as show in figure 7 below.

	Item_Name	Text_Type	Original_OCR	Original_OCR_Word_Accuracy	Original_OCR_Char_Accuracy	Original_OCR_Total_Wd_Cnt	Tesseract_OCR_Word_Accuracy	Tesseract_OCR_Char_Accuracy	Tesseract_OCR_Total_Wd_Cnt	Azure_OCR_Word_Accuracy	Azure_OCR_Char_Accuracy	Azure_OCR_Total_Wd_Cnt	Goolge_Vision_OCR_Word_Accuracy	Goolge_Vision_OCR_Char_Accuracy	Goolge_Vision_OCR_Total_Wd_Cnt	Imp_ORG_Azure_Word	Imp_ORG_Azure_Char	Imp_ORG_Google_Word	Imp_ORG_Google_Char	Imp_ORG_Tesseract_Word	Imp_ORG_Tesseract_Char
2	Letters from Victoria B	pdf/print	reed O~t 27/86 Provo,Uag.30,1886 M	92.7	88.32	5016	94.94	91.99		96.51	95		97.9	96.01	5574	3.81	6.68	5.2	7.69	2.24	3.67
3	Mighty Man Was Broth	pdf/print		96.3	94.66	10731	97.48	97.11		97.59	96.5		98.23	97.47		1.29	1.84	1.93	2.81	1.18	2.45
4	Transition - From Sage	pdf/print		93.65	89.45	76853	93.66	91.69		94.32	92.67		97.75	97.12		0.67	3.22	4.1	7.67	0.01	2.24
5	"A Brief Life Sketch of	pdf/print	, , U: • ' b • , • he , • Jacob • 1... aa 8 y	77.3	63.28		56.8	52.06		74.96	60.88		94.78	91.58		-2.34	-2.4	17.48	28.3	-20.5	-11.22
6	"A Sketch of William F	pdf/print	A sketch rf v-iJ.brn F'awcetts Life. cen	81.17	69.99		91.83	87.29		88.9	82.29		94.31	90.69		7.73	12.3	13.14	20.7	10.66	17.3
7	A True Indian story	pdf/handwritten		0	0		0	0	3	0	0		92.72	88.14	556	0	0	92.72	88.14	0	0
8	"Three Nephites" storie	pdf/handwritten		0	0		82.64	68.67	3727	62.63	61.42		91.49	83.81	4914	62.63	61.42	91.49	83.81	82.64	68.67
9	Deseret New	pdf/print		0	0	0	91.22	86.79	9792	87.04	79	6724	95.79	93.59	9699	87.04	79	95.79	93.59	95.79	93.59
10	William Fotheringham	pdf/handwritten		0	0		77.19	58.11	15181	22.97	20.67	3659	89.23	80.23	16111	22.97	20.67	89.23	80.23	77.19	58.11
11	Virginia Blair diary, 19(	pdf/handwritten		0	0		82.05	68.39	8608	23.7	22.95	675	92.55	87.06	8995	23.7	22.95	92.55	87.06	82.05	68.39
12	Atomic tests in Nevada	pdf/print		83.08	92.36	12502	90.18	88.18	11864	89.77	88.69	10530	97.19	96.35	12358	6.69	-3.67	14.11	3.99	7.1	-4.18
13	Elias Hicks Blackburn	pdf/print		83.1	71.25	24388	78.2	68.93	29644	73.99	67.5	11290	87.22	75.54	30565	-9.11	-3.75	4.12	4.29	-4.9	-2.32
14	A. David Every 1966	pdf/handwritten	(tia logue f mi. S. Green River j YO8O	80.78	69.14	1904	76.43	67.08	2164	74.76	63.06	848	93.97	87.48	2311	-8.02	-6.08	13.19	18.34	-4.35	-2.06
15	A Historical Overview (	pdf/print			0	0	96.54	96.04	56633	95.1	92.76	50904	98.56	97.51	56067	95.1	92.76	98.56	97.51	96.54	96.04
16	Diary of the Galloway/Stor	pdf/handwritten	Contains the diary documenting the 1(	100	100	57	60.32	51.59	2929	5.09	3.55	347	87.71	78.48	5744	-94.91	-96.45	-12.29	-21.52	-39.68	-48.41

Fig 7: Results stored in spreadsheet

Recommendations/Future directions

- Cloud services provide detailed responses to the OCR request via APIs, and includes such information as entity recognition which can be useful in providing additional services. For example, useful features such as browse by name or by location can be added to the repository by using the entity information from the APIs.

- Better reporting and visualization tools can be built to parse the information returned from the APIs.
- While basic cost estimation features are included, these features can be improved upon to provide overall cost analysis for a project involving multiple collections.

III. Accomplishments

The tool will allow users to process set of URIs (as input), download images to a docker container, provides options to analyze images against Google cloudvision API, Microsoft Azure API or tesseract (for OCR), will calculate accuracy at character & word level, and then generate a report that provides a summary of OCR improvements, if any.

We were able to develop almost all of the promised functionality of the tool. Because of the Covid-19 pandemic, we chose to get content from selected partners instead of doing a community survey. All the code and how to manual are published and available at the University of Utah, J. Willard Marriott Library's github page - https://github.com/marriott-library.

IV. Lessons learned

The following are the lessons that we learned as part of this project:

1. Toolkit would be useful to integrate with workflow management systems to get the best usage out of it. While the tool can be used as a stand-alone tool, institutions can get more usage from the tool by integrating it with their digital collections management tool.
2. While the cloud services have some level of support for non-English languages, not all languages were represented.

V. What's next?

Our roadmap includes adding more functionality to this tool, and working with partners to identify ways to integrate the tool with their digital repository workflow systems. We are also exploring available open source options to investigate support for multiple languages.

We are very thankful to the Lyraris organization for proving the seed funding to initiate this project, and supporting us throughout the development of this project in the initial phase.

VI. More information

For additional information on the project, and to follow updates, please visit https://github.com/marriott-library

References

Smith, David A., and Ryan Cordell. "A research agenda for historical and multilingual optical character recognition." *NUlab, Northeastern University.@ https://ocr. northeastern. edu/report* (2018): 36.

Torget, Andrew J., et al. "Mapping texts: Combining text-mining and geo-visualization to unlock the research potential of historical newspapers." *University of North Texas Digital Library* (2011)

Van Strien, Daniel, et al. "Assessing the impact of OCR quality on downstream NLP tasks." (2020).