Steps Towards Accessibility: Improving the User Experience through OCR

Melissa Eighmy Brown & Carol Nelson
University of Minnesota Libraries

Austin Smith & Hilary Thompson
University of Maryland Libraries
Melissa Eighmy Brown
Manager, Interlibrary Loan Borrowing & Digital Delivery Services
University of Minnesota Libraries

Hilary Thompson
Acting Director, User Services & Resource Sharing
University of Maryland Libraries
Carol Nelson
Minitex Resource Sharing Manager
University of Minnesota Libraries

Austin Smith
Resource Sharing Supervisor
University of Maryland Libraries
What is OCR?

Optical Character Recognition (OCR) is the recognition of printed or written text characters by a computer.

Benefits:
- Searchability
- Text to speech capabilities
- Screen reader compatibility
- Copy and paste
- Supports text mining
The ABO blood group system revisited: a review and update

L.I. Story and M.L. Drachen

Abstract

The ABO blood group system was the first to be recognized as a genetic trait and remains one of the most important in medical genetics. The ABO system is based on the presence or absence of antigenic substances on the surface of red blood cells, which are recognized by specific antibodies in the plasma. The three main blood groups are A, B, and O, with the A and B groups being further divided into subgroups based on additional antigens.

The ABO system has been extensively studied, and its importance in transfusion medicine, organ transplantation, and pregnancy management cannot be overstated. However, the diversity of blood group systems observed in different populations suggests that additional genetic factors may influence blood type.

Key Words: ABO, blood group, antigens, allele, genotyping

There have been many reviews of the ABO blood group system written throughout the years, covering different aspects of this fascinating topic. A limited discussion here is aimed for readers who want to focus more on the difference among ABO alleles and molecular genetics. Our intention is not to reproduce them but to follow the guidelines of this new, revised blood group.

Variation in A antigen expression was also recognized early in the twentieth century involving its role in disease. However, the A blood group is divided into A1 and A2, with A1 being the A antigen and A2 representing additional antigens. The A antigen is followed, and the A blood group was subdivided further based on characteristic reactions with human polysaccharides, i.e., strength of mixture and presence of various specific anti-A antibodies present in the sera of mixed subjects (Table 1). Weak forms of the antigen were also defined, but are typically more difficult to detect serologically.

The frequency of the common ABO phenotypes (A, B, AB, A-B, and O) varies greatly among different populations. The frequency of B was highest in Northern and Central Europe, and the frequency of A was highest in Southern Europe and the Middle East. O populations with a high frequency of B phenotypes are found mainly in Northern and Central Europe, and the frequency of O was highest in Native American populations.

Table 1: Subgroups of A antigens

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Reaction with Anti-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>+</td>
</tr>
<tr>
<td>A2</td>
<td>+</td>
</tr>
<tr>
<td>B1</td>
<td>-</td>
</tr>
<tr>
<td>B2</td>
<td>-</td>
</tr>
<tr>
<td>AB</td>
<td>+</td>
</tr>
</tbody>
</table>

https://youtu.be/MGE9VTmBtDU
Why OCR at Minnesota?

Digital Delivery Implementation
- 2016

OCR implemented for Lending/DD
- 2017

U of M Libraries Accessibility Committee
- 2019

Reserves Use of Digital Delivery Requires OCR
- 2016

OCR only for Digital Delivery
- 2018-
Why OCR at Maryland?

- **Spring 2015**: OCR #1 in ILL User Survey
- **Summer 2015**: Launched ALTR Service
- **Fall 2015**: ILL & Reserves Merger
- **Spring 2016**: OCR in USRS Strategic Plan
- **Fall 2016**: Explored OCR Options
- **Fall 2017**: New Web Accessibility Policy
- **Winter 2018**: OCR Integrated with ILL
2015 ILL UMD User Survey Results: Desired Service Enhancements

- Receive books faster: 211
- Receive recently published books sooner: 165
- Obtain items/copies from museum libraries, special collections, etc.: 329
- Keep books longer: 260
- Deliver items to my on-campus department or home address: 140
- Add books requested via ILL to the Libraries' collection: 183
- Pick up and return ILL items at additional UMD Libraries: 60
- Receive journal articles more quickly: 318
- Receive book chapters more quickly: 269
- Obtain EPUB articles ahead of print release: 106
- Receive articles that are text-searchable: 333
Fact or Fiction?

“When ILLiad receives PDFs, the OCR from the lender is not retained.”

Fiction (provided your cover sheet has OCR).
See ILLiad documentation.

“It’s not really important to users. I’ve never had a complaint.”

(Most Likely) Fiction. Have you asked your users?
→ 333 of 1,251 UMD respondents (27%) said it was.
Fact or Fiction?

“It would take too much staff time/effort.”

“It would slow delivery to our users.”

“It would cost too much.”

Whether fact or fiction depends on implementation.
University of Maryland: Borrowing & Document Delivery

Photo: UMD McKeldin Library by Bgervais / CC-BY-SA-3.0
Choosing a System

Automated delivery 24/7
- Server-based, not workstation-based
- No staff mediation required
- Minimal IT maintenance

Language & Script Coverage
In 2018 ~15K articles delivered in 32 languages & 10 scripts
Normal Workflow

Articles processed automatically when they arrive.
OCR complete within 10 minutes of delivery
(no staff intervention)

However:

Processing only handles languages written in Latin script.

*In 2018: 14,600 articles (97.5% of total) in 16 languages*
Exception Workflow

In some cases, default workflow cannot provide OCR.

In 2018 375 articles (2.5% of total) in 7 scripts

Patron can request OCR through their ILLiad account:

ILL staff process & re-deliver within 1-2 business days
Challenges & Considerations

- Page count must be estimated in advance
- Imperfect language differentiation
- Decreased need due to increased lender-side OCR?
University of Minnesota: Lending & Document Delivery
Software choices at MNU

- GdPicture + tesseract open source OCR
- pypdfocr + tesseract open source OCR
How do you decide what OCR software works for your library?

**ABBYY**
- Full featured
- Excellent quality OCR
- Used at U of MD
- Included with Scannx
- Cost

**Tesseract OCR**
- Works with scanning software
- Needs IT support to configure
- Open source

**Adobe Acrobat Pro**
- Your institution may already provide access
- High quality OCR
- Works best for low volume OCR needs
Audience Poll:

Is your ILL unit providing OCR?
If yes, how?
If not, do you plan to do so in 1-2 years?

Image: Raised Hands by Karen Arnold / CC0 Public Domain
BTAA ILL Coordinators Survey

Does your unit or library apply optical character recognition (OCR) to PDF files delivered electronically?

11 responses

- Yes: 36%
- No: 64%

For which service(s) does your library apply OCR to PDF files? Please check all that apply.

4 responses

- ILL Borrowing: 1 (25%)
- ILL Lending: 0 (0%)
- Document Delivery: 2 (50%)
- Electronic Reserves: 2 (50%)
- Some of our scanners have an OCR scanner: 1 (25%)
- Borrowing and Doc Del if patron indicator: 1 (25%)
How can we work together to provide OCR more programmatically?

Options include:

1. “OCR required” option + OCR provider group
2. ALA Interlibrary Loan code or consortial agreements
3. Feature of next generation resource sharing systems
OCR and Next Generation ILL Systems

Workflow A: Supplying Library Participates in Systematic OCR

1. Add language code from OCLC record to ILL request
2. Use language code to identify language script for OCR software
3. Identify if OCR is needed or already present
4. Apply OCR if needed
5. Deliver file via Article Exchange
6. If below threshold, flag for staff review
7. Check OCR confidence level
OCR and Next Generation ILL Systems

Workflow B: Vendor Provides Systematic OCR

1. Add language code from OCLC record to ILL request
2. Deliver file via Article Exchange
3. Use language code to identify language script for OCR software
4. Identify if OCR is needed or already present
5. If below threshold, flag for staff review
6. Check OCR confidence level
7. Apply OCR if needed
Learn More

Atlas Systems: **OCR Functionality in ILLiad**

UIU: **An Introduction to OCR LibGuide**

Natural Reader: **Text to Speech Application** -- try it out!

Convert scanned PDF to OCR using **pypdfocr**

Check out **GdPicture** SDK with add-on OCR
Melissa Eighmy Brown
University of Minnesota Libraries
eighm002@umn.edu

Austin Smith
University of Maryland Libraries
afsmith@umd.edu

Carol Nelson
University of Minnesota Libraries
c-nels@umn.edu

Hilary Thompson
University of Maryland Libraries
hthomps1@umd.edu