



LIBER 2026

Trondheim 1-3 July

Mapping Stakeholder Tension in Automatic Text Recognition (ATR) Practice

Dr Joseph Nockels, j.nockels@sheffield.ac.uk
University of Sheffield, Digital Humanities Institute;
University of Glasgow, Information Studies



[@libereurope.bsky.social](https://libereurope.bsky.social)



LIBER Europe



[@libervideo](https://libervideo)



[@libereurope](https://libereurope)



LIBER Europe

#LIBER2026



Research Background



Research Associate for *Digital Representations of Cultural Artefacts*,
University of Sheffield, Digital Humanities Institute, 2024 - 2026

Background as an accredited librarian, informing critical AI transcription workflows from a 'socio-technical' perspective. Bringing in information behaviour, critical library infrastructure, palaeography concerns.

'Making the Past Readable - The Impact of Handwritten Text Recognition on Libraries and Their Users', PhD - completed in 2024, supported by the Universities of Edinburgh and Glasgow, READ-COOP (who maintain the ATR tool Transkribus) and the National Library of Scotland.

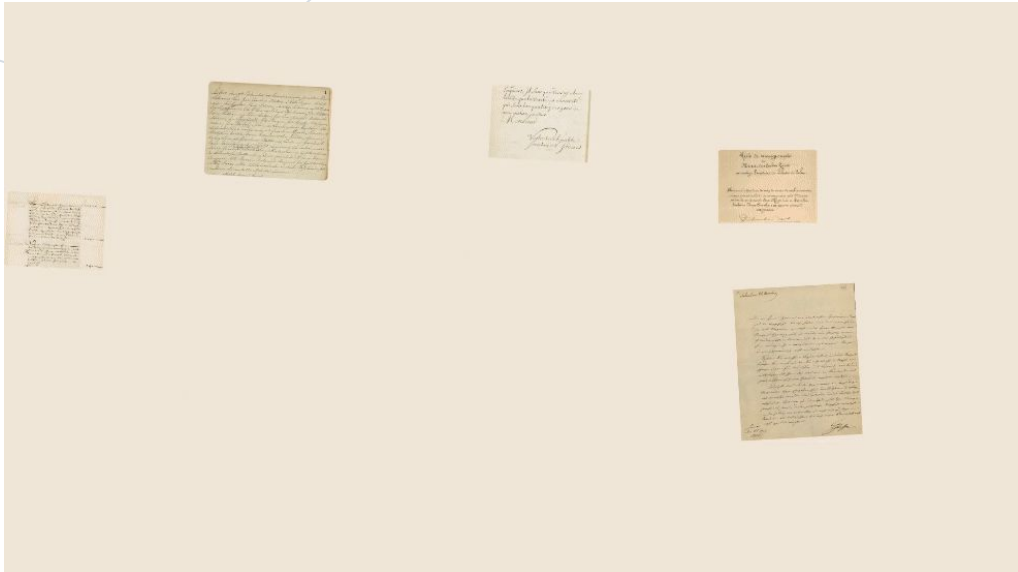
Also focus on broader digital sustainability in GLAMs, AI evaluation and benchmarking, as well as text and data mining for social and religious history.

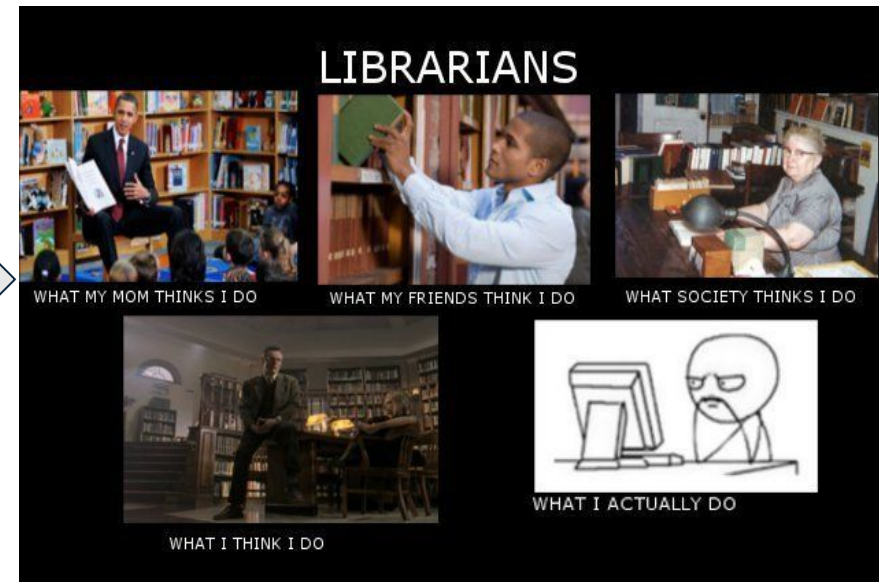
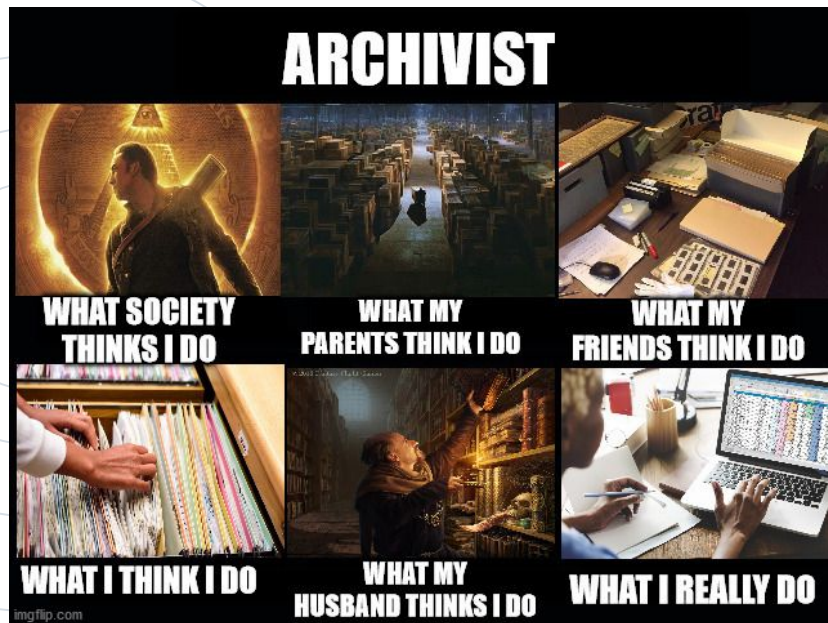
What is Automatic Text Recognition?



- AI-enabled method for converting images-of-manuscripts into computer-readable text (Pinche & Stokes, 2024).
- ATR can be used as a catch-all term for traditional Machine Learning algorithms (Optical Character Recognition) and newer Transformer-based AI (Transkribus, Large Language Models such as Claude).
- ATR, therefore, operates under a variety of business models - from open source to commercial closed products. This leads to a certain stakeholder tension.
- Libraries are actively exploring ATR affordances and current risks brought through unfamiliar patterns of use and a lack of availability evaluation data.

We require a '**socio-technical**' understanding of the technology, beyond purely technical considerations - what do users expect of ATR? How do they currently receive it? What issues is the user community (including libraries) actively trying to resolve?







... it *IS* complicated ...

Not all transcription is created equal ...



Methodology

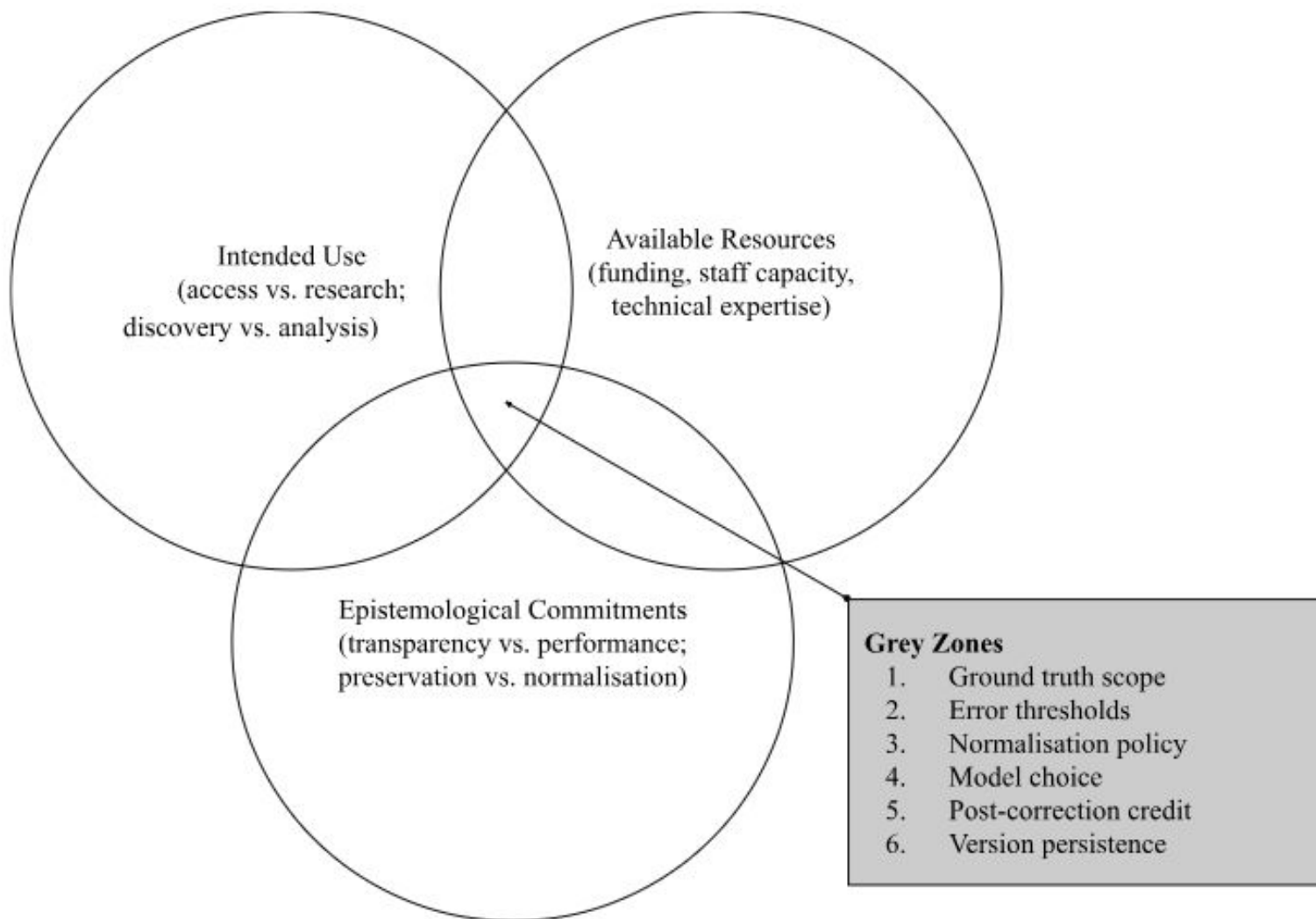


Priscilla Du Preez, Unsplash, 'Sharing'

Informal discussions with the Automatic Text Recognition community, Digital Humanities researchers, archivists, librarians, data scientists. Via Action Research processes.

Trading Zones Model (Galison, 2010)

Sharing local understanding of an entity without a full merging of meanings, symbols, and values, following Unsworth's (2000) Digital Humanities, 'scholarly primitives'. This enables points of difference to be considered, in a supportive and problem-solving environment.



Findable content: Discoverability and Contextualisation



- Accessibility to historical sources and insight into the processes undergone in generating ATR transcriptions, for all users wanting to find content.
- Prioritise the precise location of ATR resources, based on the context of the sources. Simple data provenance concerns, that libraries are specialists in considering.



Nacho Kamenov & Humans in the Loop /
<https://betterimagesofai.org/> /
<https://creativecommons.org/licenses/by/4.0/>



Accessible: How to View the ATR Outputs

- Sufficiently Descriptive Metadata, Minimal to High-Quality Enriched
- Citable persistent identifiers through DOIs, and via trusted online repositories (e.g. Zenodo, Hugging Face)
- Indicating versioning is important, since ATR transcriptions are continually edited.



Deborah Lupton / <https://betterimagesofai.org/> /
<https://creativecommons.org/licenses/by/4.0/>

Interoperable process overview/insight

- Indication of the best models / results, as well as the theorisation as to why they have been chosen (for the amount of time/money/expertise available)
- Which model?
- What settings... parameters e.g. revolution of model training, base model data used?
- Why ... ?
- How (come) ... environmental factors?



Yutong Liu & Digit / <https://betterimagesofai.org/> /
<https://creativecommons.org/licenses/by/4.0/>

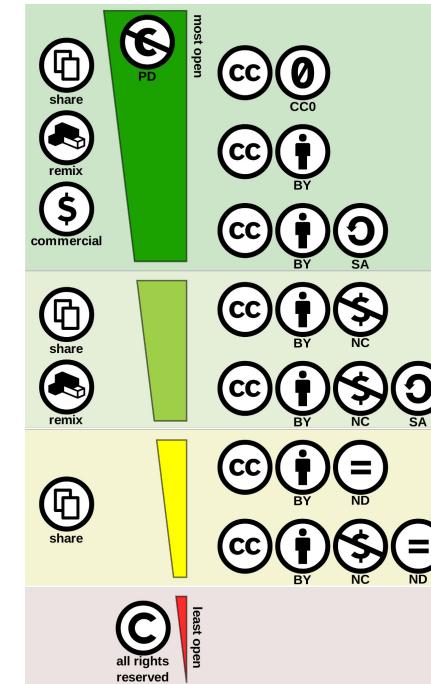


Reusable

- License C
- Transcript
- *De jure*: of



- License CC0
 - Transcription output as organisational metadata
 - *De jure*: often no set legal status for the transcription creator, especially when working as part of a large - crowdsourcing - effort
-
- A lot of work!
 - *De facto*: an important part of any ATR activity, if results and ground truth transcriptions are to be useful to the library sector and Digital Humanities research



Creative Commons (n.d.), Understanding Free Cultural Works,
<https://creativecommons.org/public-domain/freeworks/>

Reusable II – Renew(al) ATR-transcriptions

- Newer model(s) thus better quality
 - Storage capacity
 - Climate impact
-
- Citations/ accountability
 - Publishing can take a long time
 - Different results because of a “newer” dataset



Yutong Liu & Digit / <https://betterimagesofai.org/> /
<https://creativecommons.org/licenses/by/4.0/>

Collective benefit and/or responsibility



- Advocating for Open Source processes, in terms of ATR models, infrastructure development and results
- Everyone can add to the code, leveraging expertise from the Digital Humanities community, for instance eScriptorium, PaddleOCR, suryaOCR all make model weights and parameters easily accessible



- Sustainable infrastructure within a trusted responsible information repository
- Commitment to ATR resources being maintained
- Democratic access (not dependent on institute/employers - not available - servers)

Yutong Liu & Digit / <https://betterimagesofai.org/> /
<https://creativecommons.org/licenses/by/4.0/>



Ethics: Crediting

- Acknowledgement of people's work is important, as transparency, fairness and social justice. ATR projects can include varying levels of transcription contribution, from a volunteer transcribing a line of text to 'super transcribers' undergoing immense work. How should we ensure fair credit?
- How to balance with (updated version of) GDPR?
- Use of CreDiT models in published work
- Discuss with participants if and how they want to be represented





CRedit

CRedit is high-level taxonomy, including 14 roles, that can be used to represent the roles typically played by contributors to scientific scholarly output. The roles describe each contributor's specific contribution to the scholarly output.

Photo by [Jon Tyson](#) on [Unsplash](#)

