Documenting Workflows for HTR to TEI Conversions for Cultural Institutions: The Evolving Hands Project

Cummings, James
james.cummings@newcastle.ac.uk
Newcastle University, United Kingdom

Healey, Alexandra
Alexandra.Healey@newcastle.ac.uk
Newcastle University, United Kingdom

Jakacki, Diane
dkj004@bucknell.edu
Bucknell University, USA

Flex, Valentina
valentina.flex@newcastle.ac.uk
Newcastle University, United Kingdom

Jeffrey, Evie
e.jeffrey2@newcastle.ac.uk
Newcastle University, United Kingdom

Pirmann, Carrie
cmp016@bucknell.edu
Bucknell University, USA

Johnson, Ian
ian.johnson@newcastle.ac.uk
Newcastle University, United Kingdom

Background

This short paper will look at the work of the Evolving Hands project which is undertaking three case studies ranging across document forms to demonstrate how TEI-based Handwritten Text Recognition (HTR) workflows can be iteratively incorporated into curation by under-resourced cultural institutions. The case studies include materials that range from: 19th-20th century handwritten letters from the *UNESCO Gertrude Bell Archive*, 18th century German, 20th century French correspondence, and a range of printed materials from the 19th century onward in English and French. A joint case study converts legacy complex printed material of the *Records of Early English Drama*project. By covering a wide variety of periods and document forms the project has a real opportunity to foster responsible and responsive support for cultural institutions. Smaller cultural institutions often do not have the resource for the investigation of different methods in what remains a swiftly changing field of content enhancement. The short paper will introduce the project and report on the work done so far, but focus more intently on the Newcastle University Case Study on the Gertrude Bell Archive (given the limited time) and the methodologies used.

Newcastle University Case Study -- The Gertude Bell Archive

This case study uses Newcastle Special Collection's*UNESCO Gertrude Bell Archive*(http://gertrudebell.ncl.ac.uk/), which document the activities of the explorer, archaeologist, and political agent who was instrumental in establishing the Kingdom of Iraq in 1921. (Barr 2012, Bell 1927, Richmond 1937, and Dodge 2003) Bell is the subject of plays, documentaries, feature films, and recently was nominated as a BBC 20th Century Icon. (c.f. Barry 2017, Marozzi 2015, and Sluglett 2007) A separate centenary project is digitizing and cataloguing her archive of diaries, letters, and photographs. Piggybacking on that, we have trained an HTR base model of Bell's hand and applied it to carefully-selected content which has then been enriched in Transkribus to provide rich insights into political events in 1921. We are using the Transkribus platform's tagging feature to enhance the initial HTR transcriptions from which we generate (and then further up-convert) TEI P5 XML. The outputs from this will be displayed on the new Gertrude Bell research website, eventually alongside digital images presented using a IIIF viewer. (c.f. Kudella 2019)

Bucknell University Case Study -- Scholarly Production at Scale

The Bucknell case study centres on processes used across multiple discrete projects by staff with a range of digital experience. These projects represent different models for testing the HTR to TEI conversion process. Their sources are drawn from Bucknell's Special Collections and research of faculty working with archives in the US, UK, Europe, and Asia. They include scribal hands, life papers, correspondence, and semi-legible typed government files from 1700-1990 and are in English, French, German, and Maithili. This case study is directly benefiting multiple projects at the university, and this workflow is optimised for sharing with smaller cultural institutions around the world.

Joint Newcastle/Bucknell Case Study -- Transforming REED Print Collections

Cummings (AHRC PI) and Jakacki (NEH PI) have collaborated on a case study converting collections produced by the Records of Early English Drama (http://reed.utoronto.ca) project that has published since 1979 edited documentary records of pre-1642 performance in premodern England, Scotland and Wales. However, the semantic information provided in the print collections, through the use of special symbols and formatting, is lost in OCR. Earlier tests using HTR by Jakacki and Cummings demonstrated that

these distinctions can be preserved with HTR to TEI workflows. The project is documenting shared workflows for consistent up-conversion into viable materials ready to enter the REED project's digital publication workflow. (c.f Chagué et al 2022) This has the potential to be of use for all the other REED legacy print volumes (well over 20,000 pages of rich scholarly material).

Project Methodology

The project methodology, being based on individual case studies, takes a similar approach across different forms of work but also has intentional differences. In each case Transkribus is being used for HTR, though other providers of HTR were tested. (c.f. Stokes et al 2021) The generated text output is converted to TEI either through the basic Transkribus conversion or through standalone XSLT stylesheets. (c.f. Nockels et al 2022) However, it is the differences in methodology which are more interesting. In the Gertrude Bell case study the Transkribus Web/Lite version has been used to provide as much tagging as possible through that interface, in the Bucknell versions the desktop version has been used but additional tagging provided in Oxygen. The REED case study uses Transkribus Web/Lite and more detailed up-conversion stylesheets to protect formatting-based structures.

Overall the project is documenting these workflows in easy to understand how-to guides alongside other materials such as conversion scripts to make it easier for under-resourced cultural institutions to undertake such endeavours themselves.

Bibliography

Barr, J. A Line in the Sand. Simon and Schuster Ltd, 2012.

Berry, H. "Gertrude Bell: Pioneer, Anti-Suffragist, Feminist Icon?", In: Gertrude Bell and Iraq: A Life and Legacy, ed. by Paul Collins and Charles Tripp, Oxford University Press, 2017.

Bell, F. The Letters of Gertrude Bell (Vols 1 & 2). Ernest Benn, 1927.

Chagué, A., Scheithauer, H., Terriel, L., Chiffoleau, F., Tadjo-Takianpi Y. "Take a sip of TEI and relax: a proposition for an end-to-end workflow to enrich and publish data created with automatic text recognition". Digital Humanities 2022 : Responding to Asian Diversity, ADHO; University of Tokyo, Jul 2022, Tokyo, Japan. https://hal.inria.fr/hal-03739767

Gertrude Bell Website: https://research.ncl.ac.uk/gertrudebell/

Dodge, T. Inventing Iraq: The Failure of Nation Building and a History Denied. University of Columbia Press, 2003.

Kudella, C. & Göbel, M., Veentjer, U., Sikora, U. "Combining TEI and IIIF in a Virtual Research Environment", 2019 IIIF Conference - Göttingen, 2019. https://iiif.io/event/2019/goettingen/program/71/

Marozzi, J. Baghdad: City of Peace, City of Blood. Penguin, 2015.

Nockels, J., Gooding, P., Ames, S. et al. 'Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research'. Arch Sci 22, 367–392 (2022). https://doi.org/10.1007/s10502-022-09397-0

Richmond, E. The Earlier Letters of Gertrude Bell. Ernest Benn, 1937.

Sluglett, P. Britain in Iraq: Contriving King and Country. I.B. Tauris, 2007.

Stokes, P.A., Kiessling, B., Stökl Ben Ezra, D., Tissot, R., Gargem, H. 'The eScriptorium VRE for Manuscript Cultures. Ancient Manuscripts and Virtual Research Environments', ed. Claire Clivaz and Garrick V. Allen. Special issue of Classics@18 (2021).