Interview questionnaire about the selection criteria used by digital archives of historical newspapers

M.H. Beals, Julianne Nyhan, Tessa Hauswedell, Melissa Terras and Emily Bell

Corresponding publication: 2020. Hauswedell, T., Nyhan, J., Beals, M., Terras, M. Bell, E. Of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers. *Archival Science 20, 139-165.*

Appendix A: Interview Structure

Building the Corpus from Physical Collection

1. What forms of cultural, institutional or governmental policy were in place with regard to the collection and preservation activities that have affected the digitisation process? How did these affect the scope, shape and representativeness of the digitisation process? Was the aim to simply preserve and keep intact existing collections? Did you envision them fulfilling (only) the same functions as physical collections had?
2. How was the material obtained? Did the material have to be sourced to complete runs? Was only in-house material used? Was there an attempt to connect to material stored/digitised elsewhere electronically?
3. What was the physical condition of the original? Was digitisation conducted from the original newspapers or from microfilm/fiche? Were bound library editions of the newspaper runs used? Were bound editions left as single sheets when returned to storage? Was the status of bound / loose-leaf a factor in choosing digitisation?
4. How did the digitisers define "newspapers"? What were the defining characteristics for choosing if a publication qualified?
5. How were title changes handled; did they count as the same publication? How was chronology / length of run to be digitised decided?
6. Which were decisions regarding the image capture? (grayscale, bitonal) By whom?

Using the Corpus

7. What kind of search tools were implemented? Wild cards, Proximity, Boolean Operators, or others?
8. To what extent was the digitisation done with a specific end user in mind? How would you describe that end user? How did you envision access to the collection by that end user

(subscriptions, onsite, online)? Were these end users involved in decisions about the digitisation process? Were other possible end users involved in decisions?

9. Was any user testing conducted? At which stage, by whom? Which changes were implemented as a result?

Annotating Digitisation Choices in the Metadata

10. How were changes across a publication run annotated in the metadata, either the digital versions or the original collections? How were title changes demarcated? How were editors/owners/contributors demarcated? How were cut-off dates determined? Are missing editions clearly noted and highlighted (for example when there was a newspaper strike or there is simply a gap in the archive?

Categorization Metadata

11. Were newspaper-specific features represented in the metadata?
12. What level of semantic division (sentence, paragraph) are present in the metadata? What level of material division (page, issue) are present in the metadata? What level of informational division (article, insertion) are present in the metadata?
13. How are sub-genres (religious, trade newspapers) represented in the metadata?
14. How are multiple genres (magazine, pamphlet, newspaper) represented in the metadata?
15. Does the digital database pointed to named entities? How are persons (printer, owner, editor, compositor, contributor) entries integrated? How are physical locations (printing, sales, business office) integrated?
16. How are chronologies / run lengths for digitisation indicated in the metadata or database?

Metadata Population and Standards

17. Upon what were the current digital items metadata fields based upon? Which fields were populated from existing catalogue records of the physical items? Were these translated or standardised at the time of population or later on? Which fields were manually encoded for the digital items metadata and by whom? Which fields were automatically populated during digitisation?
18. Which fields were populated based on global standards (Dublin Core, etc)? Which fields were populated based on institutional standards? Which fields were developed/populated based on cataloguing standards for a particular item type/genre? Which fields were

developed/populated based on cataloguing standards for a particular department / sub-group of the institution?

19. Which fields allowed subjective or individual population?
20. Are the records for the specific physical item linked to the digital one?

Metadata Versionality

21. Which systems or persons decided on the metadata standards or methods for the current database? For previous versions of the database? Which systems or mechanisms exist for reviewing and changing the current methods / standards for populating metadata or for updating previous entries?
22. When was the current metadata standard adopted? If there were previous standards, when were they in effect? Do multiple standards co-exist in the current database or have they been updated / unified?
23. How homogeneous is the current database's metadata in fields and content? Is there a form of quality assurance in place for the consistency and accuracy of the XML?

Appendix B: Interview and Interviewee Details

The National Library of Australia

https://trove.nla.gov.au/

19 February 2018

Interviewed by M. H. Beals

Readex

https://www.readex.com

6 June 2018

Interviewed by Tessa Hauswedell

The National Library of Scotland

https://www.nls.uk/

20 June 2018

Interviewed by Tessa Hauswedell

Proquest

https://www.proquest.com/products-services/pq-hist-news.html

9 July 2018

Interviewed by Tessa Hauswedell

Gale, A Cengage Company

https://www.gale.com/intl/primary-sources/historical-newspapers

8 August 2018

Interviewed by Tessa Hauswedell

The British Library

https://www.bl.uk/

3 October 2018

Interviewed by M. H. Beals

Koninklijke Bibliotheek

https://www.delpher.nl/

4 December 2018

Interviewed by M. H. Beals

Funding

ES/R004110/1 (Oceanic Exchanges Project: Tracing Global Information Networks in Historical Newspaper Repositories, 1840-1914 funded by the 'Transatlantic Partnership for Social Sciences and Humanities 2016 Digging Into Data Challenge'). The research conducted by UK Institutions is funded by ESRC and AHRC.