[1]: Improving Computational Analysis with Humanities Narrative-Building Methodologies"/>

"EXALTING THE CULT OF GENTLEMANLY AMATEURISM"[1]: IMPROVING COMPUTATIONAL ANALYSIS WITH HUMANITIES NARRATIVE-BUILDING METHODOLOGIES

XML

1. ABSTRACT

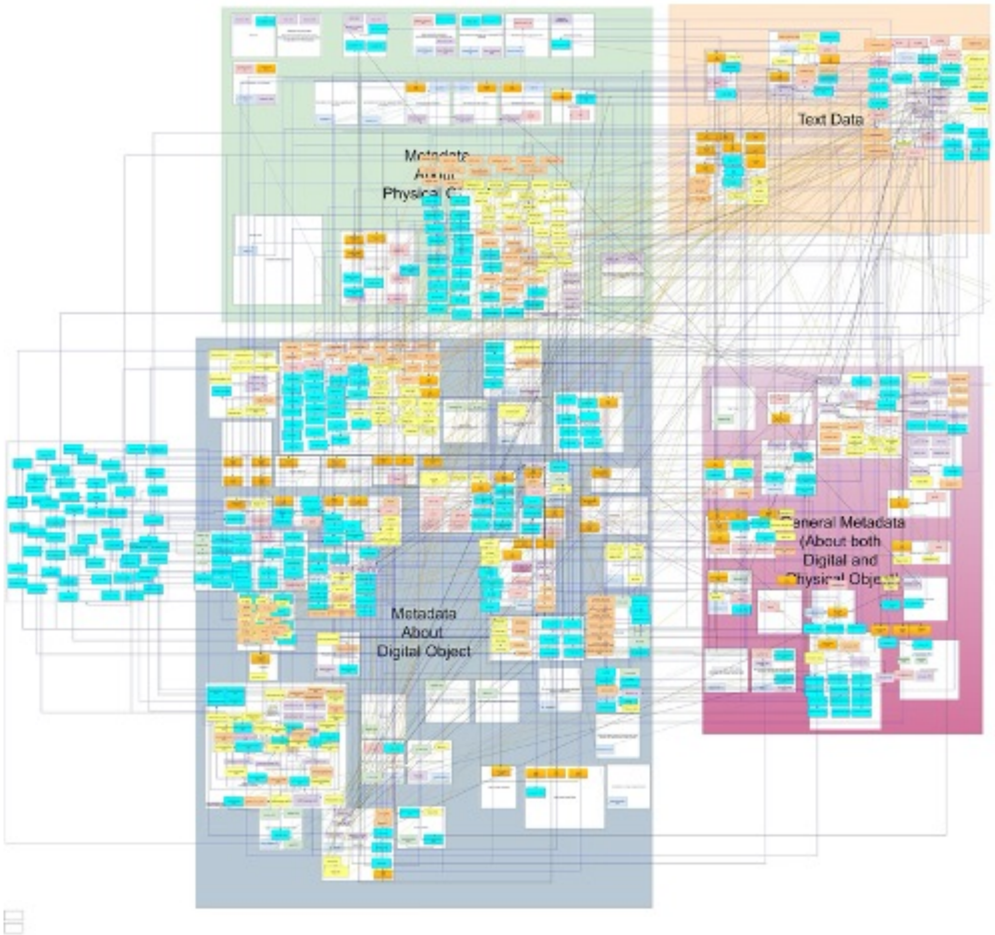
The nineteenth-century newspaper was a messy object, filled with an ever-changing mix of material—literary, factual and the suspiciously plausible—in an innumerable number of amorphous layouts. Working with digitised newspapers is no different. Each database contains a theoretically-standardised collection of data, metadata, and images, but the precise nature and nuance of this data is often occluded by the automatic processes that encoded it. Moreover, no true universal standard has been implemented to facilitate cross-database analysis, encouraging digital research to remain within existing institutional or commercial silos.[2] Where common standards have been asserted, such as the minimum standards for *Europeana* or *Chronicling America*, they have been standardised at only a very low resolution, with significant variance in the range and interpretation of the metadata within their direct collaborations as well as by independent programmes following their example.[3] These irregularities make the data highly vulnerable to misinterpretation by both end users and also those updating the collections in the future.

In order to better explore global exchanges (for example, scissors-and-paste journalism) in the nineteenth-century press, *Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840-1914* attempted to integrate and make interoperable the metadata used to store digitised newspapers in a variety of linguistic and institutional contexts. This paper will demonstrate how we excavated institutional decision-making from a variety of sources in order to understand the archaeology of digitised newspaper metadata, its vocabulary and structures, and how they related to the conceptions of the newspaper object by both modern end-users and the original nineteenth-century producers. It will explore how computational thinking and data management processes can be combined with discussions of the historical evolution of the newspaper to restore and integrate a narrative that is generally lost in the creation of digital archives: how the strategies and decision-making processes that shaped the composition and structure of the data have, and will continue to, impact user experience and the conclusions drawn from these materials.[4]

The *Ontologies* work package of *Oceanic Exchanges* had a simple remit: to catalogue and map the metadata terminology used by newspaper databases to one another and to an internal ontology, to support research into reprinting. Because complete documentation was not available for any of our collections, we retro-engineered the implementation of these vocabularies, beginning with document type definitions (DTDs) and schema specifications, complementing them with internal and public documentation on the cataloguing standards used. Some cases also required us to rely upon grey literature—discussions by users about how to manipulate the data— and direct examination of records. Finally, building upon previous research by team members and new interviews, we were able to develop a longitudinal understanding of how the data has been augmented or repackaged by institutions over the past twenty years.

Although most of the databases used variants of the METS/ALTO standard, these were not implemented in a way that would allow for simple equivalencies. The variance in terminology, and in the interpretation of the correct range of inputs for a given field, arose from the use of a hodgepodge of different vocabularies, including variants of Dublin Core, METS/ALTO, MPEG-21, PREMIS, as well as other bespoke or proprietary taxonomies. Overlapping and ambiguous vocabularies were also structured inconsistently, with some combining data at the article, page or issue level and others separating the metadata and content for these elements into multiple files. Our initial attempts to account for both internal structures and field equivalencies across these databases made the level of irregularity strikingly clear.

Figure 1: Map of all metadata fields from our samples (each one represented by a different colour), with connecting lines showing the internal hierarchy of each, broken down by metadata of physical object, digital object, metadata pertaining to both, and text data. Unmapped blue boxes represent an overflow of repetitive administrative technical metadata.



Moreover, the interpretation and implementation of these fields was inconsistent within collections owing to the turnover of staff during the digitisation process as well as the long history of metadata being drawn from existing library catalogues. Such layering is particularly evident in the metadata associated with *Trove*, the National Library of Australia's collections, which includes end-user annotations, categorisations and text corrections—layers which are valuable to humanities researchers but which remain in unintegrated grey literature and derived data for the other collections. The level of publically-available documentation about how to interpret both authorised and user-generated fields varied widely, and interviews and internal documents made it clear that consistent implementation of guidelines was unlikely across time. Working with these collections, therefore, requires a creative and flexible interpretation of these standards and an understanding of the history and character of the specific digital files.

After working with such disparate source materials, we concluded that the narrative of creation, archiving and digitisation might be most robustly and sustainably documented through a decentralised and layered medium, namely Linked Open Data. This decision is not without controversy. First, although the scale of periodical material makes it particularly tempting for large-scale analysis, the majority of newspaper metadata is in XML format, which presents specific challenges for semantic data modelling.[5] More philosophically, the possibilities and problematics of the semantic web have been theorised since the term was coined in 2001;[6] in particular, the importance of making and sustaining connections to humanistic forms of knowledge representation has been regularly emphasised.[7] Oldman, Doerr and Gradmann highlight the possibility of linked data combining "digital infrastructure, computer reasoning, interpretation, and digital collaboration", but warn of leaving a "mechanical meaningless shell"[8] if the endeavour is seen as an end in itself, or as a purely scientific exercise.[9] Indeed, linguistics and literary scholars have raised numerous concerns about tool-driven research questions and banal quantification for computation's sake.[10] Likewise, Berry and Fagerjord have claimed that linked data involves a fragmentation that "privileges knowledge divided into non-narrative shards of information",[11] seemingly putting it in direct opposition to the idea of reclaiming lost narratives of creation and use.

This paper will, therefore, explore the implications of competing claims surrounding the value of Linked Open Data within the specific domain of digitised periodicals, particularly when working with enriched metadata and data roundtripping (the process of integrating derived data back into the original collections), and demonstrate how combining institutional histories, interviews, metadata and historical narrative detail in a decentralised and layered structure can restore lost narratives and data provenance in a sustainable way that is intelligible across disciplines and at multiple resolutions–whether focusing on the textual content of the issue, the technical details surrounding digitisation, or the computational representation of the physical layout and materials.

[1] Noam Chomsky, *Language and Mind* (Cambridge: Cambridge University Press, 2006): 19.

[2] Developments in this area have been led by the *Europeana* Data Quality Committee, which has developed a *Metadata Quality Assurance Framework for Europeana* (see http://144.76.218.178/europeana-qa/), but engagement with this is limited by the project partners' existing metadata practices. Other cultural heritage metadata ontologies such as the CIDOC Conceptual Reference Model (http://www.cidoc-crm.org/) and FRBRoo (https://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf) do not include newspapers.

[3] See, for example, the guidelines provided by three key digisters, which generally conform to agreed METS/ALTO standards, but vary (and allow for variance) in metadata range and interpretation. "Technical Guidelines for 2018 Awards", *Library of Congress*, 18 December 2018, https://www.loc.gov/ndnp/guidelines/archive/guidelines1819.html. Accessed 10 October 2019; "Europeana Data Model", *Europeana Pro,* 2017, https://pro.europeana.eu/resources/standardization-tools/edm-documentation. Accessed 10 October 2019; Although internal documentation of METS/ALTO guidelines is not publicly available (it has been reviewed by the authors), the guidelines for the public API (with simplified field names) provide similar insights. "API version 1 technical guide", *National Library of Australia*, 2018, https://help.nla.gov.au/trove/building-with-trove/api-version-1-technical-guide. Accessed 10 October 2019

[4] See Bonnie Mak, "Archaeology of a Digitization", *Journal of the American Society for Information Science and Technology* 65.8 (2014): 1515–26 and Paul Fyfe, "An Archaeology of Victorian Newspapers," *Victorian Periodicals Review* 49.4 (2016): 546–77.

[5] See Fabio Ciotti and Francesca Tomasi, "Formal Ontologies, Linked Data and TEI", *Journal of the Text Encoding Initiative* 9 (2016-2017): http://journals.openedition.org/jtei/1480. Accessed 10 October 2019.

[6] See Tim Berners-Lee, James Hendler and Ora Lassila, "The Semantic Web", *Scientific American*, (2001).

[7] See Dominic Oldman, Martin Doerr and Stefan Gradmann, "Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge" in Schreibmann, Siemens and Unsworth, eds. *A New Companion to Digital Humanities* (Oxford: Wiley-Blackwell, 2016): 251–73.

[8] Ibid. 252.

[9] Sean Bechhofer et al, "Why linked data is not enough for scientists", *Future Generation Computer Systems* 29.2 (2013): 599–611.

[10] Chomsky, *Language and Mind*: 19; Nan Z. Da, "The Computational Case Against Computational Literary Studies," *Critical Inquiry* 45.3 (2019): 601–39.

[11] David M. Berry and Anders Fagerjord, *Digital Humanities: Knowledge and Critique in a Digital Age* (Cambridge: Polity, 2017): 77.

M. H. Beals (e.bell@lboro.ac.uk), Loughborough University, United Kingdom, Emily Bell , Loughborough University, United Kingdom, Julianne Nyhan , University College London, United Kingdom and Tessa Hauswedell , University College London, United Kingdom