Map of Digitised Newspaper Metadata

AN OCEANIC EXCHANGES DATASET

Document: DOI:10.6084/m9.figshare.11560059
Dataset: DOI:10.6084/m9.figshare.11560110

v.1.0.0
30 January 2020

Contents

Description

This dataset provides a comprehensive list of all the XML elements and attributes and JSON keys used within the fourteen (14) database instantiations utilised by the **Oceanic Exchanges Project** in the creation of their ***Atlas of Digitised Newspapers and Metadata***.

It was derived from a sampling of XML and JSON files from complete collections furnished by ten digitised newspaper providers: Hemeroteca Nacional Digital de México (National Digital Newspaper Archive of Mexico), Chronicling America (Library of Congress), the British Library, the Times Digital Archive (Gale, a Cengage Company), Delpher (Koninklijke Bibliotheek), Europeana, ZEFYS (Staatsbibliothek zu Berlin / Berlin State Library), Suomen kansalliskirjaston digitoidut sanomalehdet (Digital Newspapers at the National Library of Finland), Trove (National Library of Australia), and Papers Past (National Library of New Zealand). Information from these samples was supplemented by internal and publicly available documentation—document type definitions (DTDs), standardised metadata schema, and API technical guides—as well as interviews with library and digitisation staff.

It includes 3343 rows, each containing a unique element, attribute or key, and provides detailed information about their content, their placement within their separate hierarchies, and their equivalencies across the different instantiations and databases.

Background

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. 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. 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. It 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. Because exhaustive documentation was not available for any of the collections used, the project team retro-engineered the implementation of these vocabularies, beginning with document type definitions (DTDs) and schema specifications, and then complementing them with internal and public

documentation on the cataloguing standards used. Some cases also required the use of grey literature—discussions by users about how to manipulate the data—and direct examination of records.

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.

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 authoritative 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.

Although it currently represents only an initial snapshot of fourteen (14) digitised newspaper collections, this dataset acts as a framework to help bridge the interoperability gap between individuals who create the authoritative, standardised metadata for these collections and the end-users who attempt to create historical and other narratives through the use of these materials. It aims to enable researchers from a variety of disciplinary backgrounds to break through the barriers between collections as well as suggest historically informed principles for archivists and digitisers to consider when implementing their metadata standards and selecting which fields to make publicly available and searchable.

Methods and caveats

Essential background reading:

- Beals, M. H. and Emily Bell, with contributions by Ryan Cordell, Paul Fyfe, Isabel Galina Russell, Tessa Hauswedell, Clemens Neudecker, Julianne Nyhan, Mila Oiva, Sebastian Padó, Miriam Peña Pimentel, Lara Rose, Hannu Salmi, Melissa Terras, and Lorella Viola. *The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges.* Loughborough University, 2020. **doi: 10.6084/m9.figshare.11560059**.
- Hauswedell, Tessa, Julianne Nyhan, M. H. Beals, Melissa Terras and Emily Bell. '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: International Journal on Recorded Information, **forthcoming**.

The source data

In 2017-18, led by Paul Fyfe of North Carolina State University, Oceanic Exchanges gathered together fourteen instantiations of ten distinct digitised newspaper databases, alongside histories of their creation, composition and licensing. These collections were generously furnished by digitisation providers via a combination of text-mining harddrives, direct download packages and API retrieval systems. The collections were hosted on a secure server by Northeastern University, which could be consulted remotely by project partners around the world, and these datasets served as the primary source material for *The Map of Digitised Newspaper Metadata*. Information regarding the history and composition of these datasets was consolidated and made available via **the project website in 2018**.

Data processing

In 2018–19, a team led by M. H. Beals of Loughborough University worked to catalogue the data and metadata available across these collections, to undertake detailed interviews with data providers and libraries, and to develop a robust taxonomy for discussing the digitised newspaper not only as a facsimile but as a research object in its own right.

Collation of datasets

Project team members with the appropriate language knowledge worked through sample XML and JSON files, inputting into a shared spreadsheet the name of each element, key and attribute, its XPath/JSONPath, and an example of of its content. Each item was giving a unique identification number, which was used to describe internal hierarchies through lists of parents, children, attributes and attributing.

Provision of technical definitions

After a conflated catalogue was completed, team members used technical documentation, grey literature and a wider sampling of the collections in order to attached technical definitions, controlled vocabularies, data types and metadata standard information to each item. The language for these was largely standarised to aid mapping, but exceptions or unique characteristics were documented.

Attachment of ontological categories

As we finalised our master list of data and metadata fields, we attempted to visually group, or map, all possible elements across all collections, using the visualisation tool draw.io; we anticipated that the majority of fields would correspond directly to similar fields in other databases and thus a visual representation would be the clearest means of conveying the information overlap between collections. However, attempts to create a single map of all possible elements and attributes, and to provide provenance of internal structures while grouping object by type and subtype, raised significant ontological issues; the *ideal* relationship structure between elements varied widely depending on discpline and use case. As a result, we instead created an inclusive taxonomy of the metadata categories and sub-categories that was based upon the structure of the newspaper as both a physical and digital object and which considered the reality of the information that was available in each dataset. We hope this format will provide a deeper, more nuanced understanding of this ubiquitous and ambiguous

medium and allow for a generous mapping of similar fields while retaining sufficient detail to distinguish apples from oranges.

Caveats

- This dataset represents only a handful of digitised newspaper collections worldwide
- The collections it represents are predominately (50%) Anglophone
- The collections it represents were created by national libraries or large-scale commercial publishers in Europe, North America and Australasia.
- Where possible, technical definitions from metadata standards and DTDs have been modified to better reflect samples from individual collections but these changes have not been marked in the dataset
- Controlled vocabulary lists are derived from metadata standards, database-specific documentation and sampling. As a result, they may not reflect all possible values or may include values that were not instantiated in the final collection
- Format standard and data type are based on official documentation, where available, and sampling in all other cases.

We hope that the **dynamic version of this dataset** will lead to a more varied and globally representative selection of metadata connections. Those who work with the collections are encouraged to update and refine controlled vocabularies and definitions.

Data

General Notes

The current version is 1.0, released January 2020. This represents the static version of the dataset, a further description of which can be found in the *Atlas of Digitised Newspapers and Metadata* **DOI: 10.6084/m9.figshare.11560059**

The data table is supplied in TSV (tab separated values) format, which can be readily imported into spreadsheets and database software.

Table Columns

The columns of data are defined as follows:

Field	Description
UID	This column contains a unique numerical code used for cross-referencing between fields within the same database
Category	This column maps the field to a broad ontological category, used across collections
Sub-	This column maps the field a narrower ontological category, used across collections

Category	
Collection	This column contains the 4-character code for a particular collection and file type
XPATH	This column contains the full XPath or JSON Path to the field
Name	This column contains the name of the field (element or attribute) within the database
Format	This column contains the standard or schema to which the field belongs
Content Type	This column contains a 3-character code indicating the data type
Example Content	This column contains the contents of the field from an example XML or JSON file
MCHOICE Values	This column contains the controlled vocabulary of a multiple-choice field
Definitions	This column contains a definition or description of the field
Field Type	This column indicates is the field contains data that can mapped across collections, technical (usually automatically generated) data, or a container, which contains no data but rather only other elements or fields
Element Type	This column describes whether the field is an element/key (XML/JSON) or attribute
Parent	If this is an element, this column contains the unique numerical code of its parent element
Attributed	If this is an attribute, this column contains the unique numerical code of the element for which it is an attribute
Child(ren)	This column contains the numerical codes of all this element's child elements
Attributes	This column contains the numerical codes of all this element's attributes

Value Definitions

Categories

The top-level ontological categories used to map fields from one collection to another are as follows:

Category	Description
Abbreviated Newspaper Title	A standardised abbreviation of the newspaper title which may also appear in unique IDs for the newspaper and article
Alternate Newspaper Title	Provides an alternate title for the publication, where the title may have changed during its run. Occasionally this is a minor change, such as dropping the article, but this can also be a more radical restyling of the publication. Standardised title information can be found in newspaper title
Article Category	Specifies the genre of the article, such as Advertisements or News Article
Article	Smaller titles used to break up the article

Subheading	
Article Title or Headline	Provides the headline or title of the item. This may be hand-keyed or the result of OCR. Distinct from a section heading.
Attribution	The name of the author of the newspaper article, as printed
Comments and Social Tagging	Information about tags and comments added to an article by online users
Coordinates	Provides coordinates for a component of the image
Copyright	Specifies the copyright holder of the issue. Access conditions provides additional information about the status of a physical object, i.e. any restrictions on access
Database	Provides the name of the digital database in which the issue is stored. Distinct from the Holding Library information
Dimensions	Provides the dimensions of a component of the image. See measurement unit for the specific unit used; this is usually “mm10”
Document Type	Specifies the nature of the piece of writing, generally “article”
Edition	Provides edition information for the issue, including morning, afternoon, evening, day, special and supplemental. In SBMA and SBME, it also specifies that it is an electronic edition of the issue
Filename	Provides the filename of the image file attached to the XML text. This can take the form of file names, URLs or relative paths with filenames
Font Information	Provides information about the font of the text, as recognised by the OCR software. This includes font size, font style (whether bold, italics, underlined, small caps, etc.), font type (whether serif or sans serif), font width (whether proportional or fixed), and font family
Geographic Coverage	Classifies newspapers depending on their wider geographic area of publication and readership; it is listed as regional, local, or a specific territory. If not indicated, it can often be presumed to coincide with place of publication. It can be used to distinguish between different editions of the same paper aimed at different cities, towns or regions
Holding Library	Provides the name and details of the library or archive that contained the digitised material at the time of digitisation. For some databases, it is separated into library name and library location
Hyphenation	Provides information about words that have been typographically hyphenated.
ID	Provides a unique ID for the component of the image
Illustration Information	Provides information about any illustrations, including whether one is present, its type, and its colour information
Issue Date	Gives the date of the issue. May refer to the publication date, the date as printed on the issue, the ISO standard date or a part of the date, such as the day of the week, day, month or year. In some cases, this is normalised and in others it is the date as printed on the image

Issue Number	Gives the issue number for the item. This is sometimes a string, as printed, and other times a numerical value. It can also take the form of a unique identifier for the newspaper issue
Language	Specifies the language of the textual unit, often, but not always, using the ISO language code. This can refer to the language of the newspaper, the article text, or the specific block of text
Measurement Unit	Provides the measure unit for all values except the font size
Metadata Type	Defines the metadata type
Microfilm Reel	Most often provides a unique 4-digit ID to the microfilm reel used in the creation of the image associated with this XML. This does not translate to a MARC or library-based record number
Newspaper Subtitle	The subtitle, which is intended to provide clarification of the newspaper title, may be taken from the physical object, or an amalgamation chosen by the cataloguer
Newspaper Title	There are two kinds of newspaper title provided: first, the title as it appears on that particular issue; second, the title in a normalised format. This may or may not be a version of the title as printed but rather an amalgamation chosen by the cataloguer. It is usually derived from the earliest available issue from that newspaper, after which alternate newspaper titles will be recorded
OCR	Provides technical information on the OCR software used, including Description ID, Agency, Date and Time, Step Description, Step Settings, Creator, Name, Version, Relevance and Confidence Level
Page Count of Article	Provides the number of pages over which the particular article, as computationally zoned, is spread
Page Count of Issue	Total number of pages of the newspaper issue
Page Number	Provides an ID for the page. This is divided into: unique identifiers, page image numbers, identifiers across the database; URLs to web-accessible versions of the page; relative numerical identifiers, within the issue; and string descriptors
Page Position	Indicates the position of the page within the issue
Page Skew	Provides the degree to which the page image is skewed from the perpendicular
Paragraph	Information about paragraphs, including XML containers, text alignment, and UIDs
Place of Publication	Provides the geographical location associated with the printing or manufacture of the publication and generally listed as the city. They are determined by the imprint or cataloguer determination for the publication as a whole, except where specified as being the publication location of an alternate title for this newspaper
Publication Date Range	Provides the date range, in either years or full dates, of the publication. It does so without date or other restrictions and should be considered to refer to the newspaper as defined by ontological subcategory normalised title. There are three variants, the dates included in the collection (collection range), those that to our knowledge existed (full range), and all the individual days, months and years of

	publication. Newspaper start date and end date provide the full ISO date for the publication's first and last issue. Instantiations are divided into container elements Instantiations are divided into container elements, which hold no specific data, and attributes or specific elements, which hold the year, month and day separately.
Publication Frequency	Specifies the frequency of publication as a whole and should not be confused with the edition of a specific issue. Across these databases, it is usually listed as daily, weekly or quarterly
Publication Genre	Specifies the genre of the publication at the broadest level, i.e. whether a newspaper, periodical or magazine
Publisher	This category contains information about the publisher of the publication. It does so without date or other restrictions and should be considered to refer to the entire run of the newspaper as defined by category Normalised Title
Quality	The preservation status of the physical object
Section Heading	Specifies the printed title for a section; distinct from an article headline or title
Shelf Mark	This category contains information linking the publication, as a conceptual unit, to an item or record in an external database or catalogue. It does so without specific date or other restrictions and should be considered to refer to a specific physical volume rather than a volume as numbered by the original publisher
Starting Column for Article	Provides the column on the page (given as a letter) in which the article begins
Sub-Collection	Provides details of the sub-collection within which the item has been placed
Supplement Title	The physical newspaper or periodical section this article appeared in if not the issue itself; i.e. if it appeared in a supplement
Text	Article text content. For text content in article titles, See title of the article . For text content in subheadings, see article subheading
Volume Number	Provides the volume information, either a numerical volume number relative to the newspaper title or a unique identifier. One volume comprises many issues
Word Count of Article	The number of words in the article
Word Count of Page	The number of words on the page, as identified through Optical Character Recognition

Definitions of sub-categories can be found within the relevant category of the *Atlas*.

Data Types

The codes used to signify data types are as follows:

Code	Description
BOO	A Boolean char such as 0/1 or Y/N

COO	A set of numeric coordinates to delineate a segment of an image
DAR	A range of dates
DAT	A single date
FIN	A filename
MCH	Multiple pre-defined choices
NUL	Holds no content
NUM	Numerical value, may include the symbols . , -
STR	A string of alphanumeric content
UID	Any form of unique ID or acronym
URL	A URL

Databases

The codes used to distinguish different databases, and files within those databases, are as follows:

Code	Collection Name	Standard	Description
B1GI	British Library 19th Century Newspapers, Part I, Gale's Current Text-Mining Drives	GIFT	Issue Metadata XML File
B1GP	British Library 19th Century Newspapers, Part I, Gale's Current Text-Mining Drives	GIFT	Publication Metadata XML File
B1GT	British Library 19th Century Newspapers, Part I, Gale's Current Text-Mining Drives	GIFT	Text Content XML File
B2GI	British Library 19th Century Newspapers, Part II, Gale's Current Text-Mining Drives	GIFT	Issue Metadata XML File
B2GP	British Library 19th Century Newspapers, Part II, Gale's Current Text-Mining Drives	GIFT	Publication Metadata XML File
B2GT	British Library 19th Century Newspapers, Part II, Gale's Current Text-Mining Drives	GIFT	Text Content XML File
B1JI	British Library 19th Century Newspapers, Part I, British Library's Text-Mining Drives	Bespoke	Content and Metadata XML File
B1GL	British Library 19th Century Newspapers, Part I, Gale's Legacy Text-Mining Drives	GIFT	Content and Metadata XML File
B2GL	British Library 19th Century Newspapers, Part II, Gale's Legacy Text-Mining Drives	GIFT	Content and Metadata XML File
CAAL	Chronicling America	ALTO	Content and Layout XML File
CADI	Chronicling America		Directory Structure
CAME	Chronicling America	METS	Issue Metadata XML File

DEAL	Delpher	ALTO	Content and Layout XML File
DEMP	Delpher	MPEG	Issue Metadata XML File
DEOC	Delpher	Bespoke	OCR Text XML File
EUAL	Europeana	ALTO	Content and Layout XML File
EUME	Europeana	METS	Issue Metadata XML File
F1AL	Finnish National Library 1771–1910	ALTO	Content and Layout XML File
F2AL	Finnish National Library 1771–1910	ALTO+	Content, Layout and Metadata XML File
F1ME	Finnish National Library 1771–1910	METS	Issue Metadata XML File
HNME	Hemeroteca Nacional Digital de México	METS+	Content, Layout and Metadata XML File
HNDM	Hemeroteca Nacional Digital de México	Bespoke	Content and Metadata JSON File
PPAL	Papers Past	ALTO	Content and Layout XML File
PPDI	Papers Past		Directory Structure
PPME	Papers Past	METS	Issue Metadata XML File
SBAL	State Library of Berlin	ALTO	Content and Layout XML File
SBME	State Library of Berlin	METS	Issue Metadata XML File
SBMA	State Library of Berlin	METS	Publication Metadata XML File
SBMY	State Library of Berlin	METS	Publication-Issue Metadata XML File
TDAG	Times Digital Archive	GIFT	Content and Metadata XML File
TRAL	Trove	ALTO	Content and Layout XML File
TRAP	Trove	Bespoke	API XML Return
TRME	Trove	METS	Issue Metadata XML File

Suggested Citation

M. H. Beals and Emily Bell. (2020). Map of Digitised Newspaper Metadata v.1.0.0 [Data set]. Figshare.
DOI:10.6084/m9.figshare.11560110

Acknowledgments

This derived dataset was created using internal documentation, sample XML files, and interviews with the public and commercial providers of the digitised newspapers collections. We are grateful for the support and contributions made by members of the **Oceanic Exchanges** team—Ryan Cordell, Paul Fyfe, Isabel Galina Russell, Tessa Hauswedell, Clemens Neudecker, Julianne Nyhan, Mila Oiva, Sebastian Padó, Miriam Peña Pimentel, Lara Rose, Hannu Salmi, Melissa Terras, and Lorella Viola—as well as our friends and colleagues at libraries and publishing firms around the world—Seth Cayley (Gale), Steven Claeyssens (KB), Huibert Crijns (KB), Nicola Frean (NLNZ), Julia Hickie (NLA), Jussi-Pekka Hakkarainen (NLF), Chris Houghton (Gale), Melanie Lovell-Smith (NLNZ), Minna Kaukonen (NLF), Luke McKernan (BL), Chris McPartland (NLA), Maaike Napolitano (KB), Tim Sherratt (University of Canberra) and Emerson Vandy (NLNZ).

License

We ask that rather than create multiple independent variants of this dataset, that you create a pull request for (contribute) updates and additions to the dynamic dataset. This will allow us to maintain an ever more accurate and comprehensive map with full provenance tracking of your contributions.

Future plans

- Development of sample use cases
- Inclusion of additional datasets