THE ROUTLEDGE COMPANION TO LIBRARIES, ARCHIVES, AND THE DIGITAL HUMANITIES

Edited by Isabel Galina Russell and Glen Layne-Worthey

First published 2025

ISBN: 9781032356259 (hbk)
ISBN: 9781032356280 (pbk)
ISBN: 9781003327738 (ebk)

14

PUBLISHING LARGE COLLECTIONS OF DIGITISED PRINTED MATERIAL

The National Library of the Netherlands

Steven Claeyssens

DOI: 10.4324/9781003327738-17

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

14

PUBLISHING LARGE COLLECTIONS OF DIGITISED PRINTED MATERIAL

The National Library of the Netherlands

Steven Claeyssens

KB, the National Library of the Netherlands

The KB, National Library of the Netherlands, plans to have digitised a large part of its collection of printed material by 2030. Over 130 million pages have already been processed and are accessible via an online graphical search interface called Delpher, and as data, for computational research purposes. Like the Google Books project, HathiTrust and quite a few national digital libraries all over the world, this type of continuously expanding and evolving large collections of digital surrogates enables online searching, browsing and reading on an unprecedented scale, as well as scientific and scholarly analysis of vast amounts of digital text and images using machines, algorithms and software to "read" or mine the data. Based on the experiences at the KB, this chapter argues that (1) a better understanding of these massive collections of digitised material necessarily starts with the identification of the bibliographic objects that constitute the collections and (2) creating and opening up these large digital collections effectively turns libraries into publishers, urging librarians both to rethink and to reconfirm the traditions of curating library collections.

Large-scale digitisation at the national library of the Netherlands

Large-scale digitisation at the KB took off after the last turn of the century with a number of projects being initiated to digitise entire collections of printed material and subjecting the images to Optical Character Recognition (OCR) software to create machine-readable text.[1] In chronological order the first four projects were Staten-Generaal Digitaal in 2003 (2.1 million pages of Dutch parliamentary papers, originally published between 1814 and 1995), Databank Digitale Dagbladen in 2006 (8 million pages of newspapers, originally published between 1619 and 1995), Dutch Prints Online in 2007 (2.1 million pages of monographs, originally published between 1781 and 1800) and Digitalisering Tijdschriften in 2009 (1.5 million pages of periodicals, originally published between 1840 and 1950). They resulted in the publication of four dedicated online search interfaces: statengeneraaldigitaal.

DOI: 10.4324/9781003327738-17

nl (2007), kranten.kb.nl (2010), earlydutchbooksonline.nl (2011) and tijdschriften.kb.nl (2013).[2] While the digitised materials survived, the individual interfaces did not. In 2013 they were replaced by one national gateway. In a collaboration with several Dutch university libraries, the KB launched Delpher (delpher.nl), bringing together the collections created in several large digitisation programmes in the Netherlands in a full-text searchable and readable online graphical user interface.

On Delpher people can search the plain text (OCR) and some basic bibliographic metadata. Visually the scans of the original publications occupy a central position. Each book, newspaper or periodical is presented by means of an image viewer, based on the scans. Other versions of the publication (PDF and plain text) are not part of the interface but are offered as separate downloads. The KB has entered into agreements with right holder's organisations and publishers allowing for the in-copyright publications to be published on Delpher and to make copies available for research purposes.

Based on these agreements an additional service was launched in 2012 to give bulk access to all of the collections for computational research.[3] Data Services basically consists of a coordinator and a handful of manuals on how to search through a Java implementation of the Search/Retrieve via URL (SRU) protocol and harvest images, plain text and metadata through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Upon request a licence may be granted to get access to copyright-protected material for research purposes. The service has been successful in opening up the Delpher collections for a variety of national and international research projects (Polimedia, Translantis, Nederlab, Media Suite, Impresso, to name a few) as well as many individual researchers in the Netherlands, and beyond.[4] In conjunction with the KB Lab (lab.kb.nl), launched in 2014 at the first Digital Humanities in Belgium, Luxembourg and the Netherlands (DHBenelux) conference in The Hague, it has served as an inspiration for several European national libraries to give access to digital library collections for research.

The following years a few more options were added to give bulk access to the collections. The Delpher Open Newspaper Archive contains the plain texts and metadata of all Public Domain newspapers (1618–1876).[5] The archive is 111 GB in size and divided into 22 ZIP files. On the KB Lab a handful of derived datasets have been published: an example set, with a small selection of digitised publications from the years 1870 and 1871; the KBK-1M Dataset, a collection of 1,603,396 images and accompanying captions, extracted from digitised newspapers published between 1922 and 1994; 2,000 pages of newspaper "OCR ground truth," containing the complete and accurate record of every character and word on the pages, used to evaluate the output of OCR software, etc.[6] Over the last two years, a few instructional Jupyter Notebooks have been created to help users learn how to query the newspaper collection.

The library as a publisher

The KB's collections of digitised books, newspapers and periodicals have now grown into fully fledged large-scale national collections, actively maintained and well established. They are supplemented on a regular basis and access to the collections is facilitated according to the contemporary generally accepted primary and secondary access methods of digital cultural heritage: with an online graphical search interface and with a set of application programming interfaces (APIs), in line with the "Collections as Data imperative" first elaborated by Thomas Padilla and colleagues, making the collections "amenable to computational use."[7]

The global rise of digital scholarship and the Collections as Data paradigm in particular have spurred critical reflections on the production of large digital library collections in recent years. As Sarah Ames puts it: "With the shift towards digital, however, we are now seeing libraries increasingly becoming producers of their own collections. […] What, then, are libraries' responsibilities as data producers?"[8] The answers range broadly and include stressing the importance of documenting data provenance, institutional memory, transparency of decision-making processes and selection workflows[9]; support for Open Access policies, OpenGLAM principles and FAIR data practices[10]; urgent calls for collaboration with researchers, the development of an interdisciplinary hermeneutics[11]; and appeals for "responsible operations" particularly in managing bias.[12]

Although there are hardly any standardised solutions, for the most part libraries are making a reasonable effort to answer these calls. Surprisingly, one way of providing a framework for doing this so far remained conspicuously absent: acknowledge that libraries have become not so much the *producers* of their collections but first and foremost the *publishers* of their collections. Just like conventional publishers, libraries are not the creators of the works, nor do they necessarily print, i.e., manufacture or digitise the works. They do however make the decision and take the responsibility to make them *public*. They initiate, they select, they oversee the process, they deal with copyright etc. In other words, they do act very much like publishers, and if it looks like a duck, swims like a duck and quacks like a duck, then maybe we should start calling it a duck.

There are a few reasons why this shift in terminology is important and meaningful. (1) It fits better with the current practice of actual digitisation being largely outsourced. (2) It makes it clear that the library is in fact responsible for more than just the production of the collection, but also for making the material publicly accessible, via various channels to different audiences. (3) This, in turn, stresses that we should avoid equating a collection with a service. Whether it is a widely used graphical search interface, a command line interface, an app, an API or a notebook, it's a service running on top of an infrastructure containing the data.[13] (4) Lastly, it opens new windows on these collections. For instance, a better understanding of these massive collections of digital surrogates necessarily starts with the identification of the bibliographic objects that constitute a particular collection, and establishing their complex relationships. They are the result of a technical, scholarly, social and political process, as well as a publishing process. And publishing is an activity that has had significant scholarly attention that can bring to the table relevant perspectives on the question of how, what and in what way something is published. Applying this to collections of digitised material: what exactly is published, in what way and to what end?

A multi-layered publication of two different text editions

Both Data Services and Delpher basically publish the same data: they even run on the same technical infrastructure. However, they each offer a completely different view on the collections. The Delpher search interface is developed for the human eye and appeals to the historical sensation through digital mimicry, hiding the OCR. Data Services, on the other hand, puts a powerful emphasis on quantity and machine readability. This has significant implications for the status of the different components of the collections.

Take the plain text, the result of the OCRed scans. For Delpher, the plain text acts as a proxy. It's metadata, a source for the search index. For Data Services the plain text constitutes by far the most popular part of the collections. Expanding on the argument by Cordell that

"we might think of OCR as a compositor setting text in a language it does not comprehend," thus creating "a specific kind of edition for machine readers," the inevitable conclusion is that Delpher essentially publishes one edition of the text – in a bibliographical sense identical to the digitised paper copy – while Data Services puts the emphasis on a new, machine-generated edition of the text.[14] We can take this bibliographic dissection of the collections a step further by identifying the "discrete publishing efforts," bearing in mind that the "units of production are not necessarily the same as the units which make up the finished product."[15]

The largest and most queried online digital collection at the KB is the digitised newspapers collection. More than 17 million pages (and counting) of Dutch national, regional and colonial historical newspapers, originally published between 1618 and 1995, are being searched at least twice a month by more than 70,000 people. Apart from the various metadata, for each issue there is/are

- a searchable, bitonal PDF file at the issue level
- JPEG-2000 master images, mostly in greyscale, at the page level [archived, not published]
- JPEG-2000 access images, mostly in colour, at the page level
- Analysed Layout and Text Object (ALTO) files, with layout and text at the page level
- plain text in XML, segmented at the article level

We already observed that there are two type settings here, constituting two editions of the text: one primarily for the human eye (the images) and one primarily for machines (the ALTO and plain text files), with the PDF taking a hybrid position by essentially functioning both as an image and as a machine-readable file on its own.

The human versus machine divide leads to a crucial step in identifying the discrete publishing efforts. These collections have a layered composition, with some layers making a publishing effort from a human point of view and other layers making a publishing effort from a machine point of view. For people, the digital facsimile of the newspaper issue on Delpher (images and PDF) remains the central publishing effort. For machines there are at least two textual (plain text) publishing efforts: the article and the entire collection of articles; and two visual publishing efforts: the page (image) and entire collection of pages. These collections are more than just the sum of their parts. They embody "two seemingly opposed tendencies" of digital environments: "powerful dynamics of centralization and fragmentation."[16]

This multi-layered mode of publication of two different text editions accounts for much of the hidden complexity of these collections. It offers an explanation of why people working on the same collections and services may nevertheless have conflicting ideas about the importance of, for example, the comprehensiveness of the collection, the presence of multiple versions of the same text or publication, the status of the OCR or which files serve as mere "access copies" and which files as master images or "preservation copies" – a confusing and problematic notion in terms of publishing. It also clarifies why researchers who consult these collections might have incompatible expectations about its content. They all are in need of a shared understanding of what it means to publish a (national) collection of digitised material. That's where a more conceptual understanding of publishing can make a difference.

Publishing as modelling and framing

Publishing has had many exegetes. It is "the initial decision to multiply a text or image for distribution."[17] It is at once a "supply chain and a value chain."[18] Perhaps the most versatile

and sophisticated anchoring theory of publishing was formulated by Michael Bhaskar. He proposes a metaphoric vocabulary to understand publishing as a "content machine":

> The word content implies content of something. In other words, implied is an element not included in the word itself: that which content *fills*. This is the frame. Equally, content doesn't just appear. An interplay of causal factors, goals, motivations and ideological underpinnings shapes and provides the *raison d'être* for content: this is the model. Content always comes in specific frame-model couplings […]. We have here a context for publishing operations: to build a frame for content, according to a model.[19]

For example, a text is produced as a paperback and sold through mainstream bookstores (*framing*) with the intention of disseminating knowledge to a wider audience and making money (*modelling*). Both concepts are essential to understanding the publishing process, according to Bhaskar. They explain how (*framing*) and why (*modelling*) something is published in a certain way. Ultimately, however, they are of secondary importance to the activities that together form the actual core of publishing: *filtering* and *amplification*. Publishers decide which content to publish. They filter. Then they take the actions necessary to bring the content to the eyes of more than just its creator, thereby amplifying it. They make a whole series of technological and strategic decisions in order to increase the reach of the publication. Publishers are filters for content and constructors of amplificatory frames.

In line with this way of thinking, the Delpher interface is a *frame*, a specific way of making a large collection of digitised publications readable and searchable by people, free of charge on the web. Underlying the frame is a *model* related to the Open Access principles: to make as much content as possible available online for the widest possible audience – as the Delpher website announces, "It doesn't matter what kind of research you do and at what level."[20] What is crucial is that people can search the widest array of Dutch historical published texts. As a result, it's imperative to add as many Dutch published texts to the collection as possible. Whether the collection contains every edition of a text is of secondary importance, and if a single edition is digitised twice, that's not a big deal either. The OCR is simply metadata. If improved transcriptions are available, the old version can safely be overwritten.

Collections as published data

Publishing searchable surrogates of printed material online is one thing, making entire collections of digitised material publicly available as data is an entirely different publishing beast. The *model* might be pretty straightforward and self-explanatory – the intention to enable and foster computationally driven use of the collections – but the *framing* turns out to have many faces. Data Services at the KB, for instance, consists of documented APIs, a few ZIP files, some instructional notebooks and ad hoc assistance by a KB specialist. This approach reflects fairly accurately the conventional ways of facilitating data-level access at libraries nowadays: releasing datasets for downloading and giving access through one or more notebooks or an API.[21] The *ZIP file frame i*s easy to set up, demands little knowledge from the researcher to get started with the data, provided there is sufficient documentation available, but it scales poorly. The *API frame*, with notebooks as derivative, is a much more scalable approach, enabling advanced querying as well as bulk download through web services, and much more consistent with the FAIR principles of findability, accessibility, interoperability and reproducibility for both humans and machines alike, but it requires an

infrastructure that not every library has in place and necessitates (basic) coding skills from the researcher or library assistance. Based on interviews with researchers, developers and data providers, Edmond and Garnett are right to wonder whether the "lack of take up of web services among humanists perhaps shows a mismatch in what they want, and what developers think they want."[22]

There probably is no one-size-fits-all *frame* for publishing collections as data. But there might be a viable middle ground available where both researchers with programming skills and those lacking these skills can have effortless access to the data they need. This middle ground was central to a study, commissioned by the KB, into the desirability of developing a digital research environment for historical text collections to bridge the gap between simple search interfaces and advanced functionalities for computational analysis, usable for scholars (and other users) without programming skills.[23] The study employed a demand-driven approach, focusing on user requirements and based on interviews with employees of the KB, developers of comparable research environments and potential users. The resulting report observes that there are "insufficient commonly shared needs for developing an advanced analysis tool" and advises the KB to develop advanced search and selection functionalities. Its main recommendation is to

> Position a text suite as a corpus selection tool and support the discovery and selection research phases. A text suite hereby functions as a user-friendly front end to Data Services with more advanced features that do not fit within Delpher […]. Users can then make their own selection of sources and export them (possibly after approval by a KB employee).[24]

This new *frame i*s in line with the original *model*. It's still about enabling and fostering computationally driven use of the collections, but with a much sharper emphasis on usability, accessibility, flexibility and inclusiveness. It adds the intention to publish the collections in such a way that anyone can create and download any desired cross-section of a collection in a user-friendly manner, without having to programme or script.

Building a corpus (selection tool)

The KB plans to build a corpus selection tool. It should support easy corpus building and data gathering strategies for research purposes. What it will look like is still subject to debate. What we can do, however, is consider for a moment how scholars think about corpus building and reflect on the consequences for extracting corpora from large (national) collections of digitised material.

First of all, we have to bear in mind that different scholarly fields have different concepts of what constitutes an appropriate or balanced corpus. In linguistics, for instance, a corpus is a "collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research."[25] In Literary Studies, on the other hand, language variety is not the first concern, the complex notions of representativeness and sampling are. But what should be represented?[26]

> A text-genetic approach will tell us something about the developmental process of works with respect to authors and their intervening agents of editors, readers, and

> booksellers. A first edition approach will allow us to approximate new items that enter into the literary field while ignoring questions of penetration and circulation (how many were printed, how many were bought, how many were read).[27]

Underwood, Kimutis and Witte, for example, tackle this in a pragmatic way, by offering different ways to slice their corpus of English Language Fiction (1700–2009) drawn from HathiTrust Digital Library, with a volume list, containing all volumes; a record list, to "eliminate duplicate copies," a title list, "to identify one copy of each 'title' – by preference the earliest copy available in Hathi," and other subsets, like the gender-balanced subset or the frequently reprinted subset.[28]

There is also, almost by definition, an infinite list of individual pursuits of researchers, which underlie their potential selection criteria. Apart from the many requests for all articles in which a specific combination of words appears, in recent years academics, who have used the Delpher digitised newspapers collection as data, were interested in, for example, the advertisements,[29] the journalistic genres of articles,[30] the particular newspapers' "pillars" (groups in Dutch society divided by religion and associated political beliefs known as "zuilen", or "pillars" in English),[31] the illustrations published in the newspapers,[32] the place of publication and the places mentioned in the newspapers.[33]

Reflecting on what an online tool capable of building these disparate types of corpora could look like challenges and augments our understanding of digital collections. The standard online search interface as a *publishing frame* places few demands on the selection and contextualisation of the collection. A corpus selection tool on the other hand, in addition to the obvious requirements like strict versioning policies and granular persistent identifiers, requires deliberate, responsible selection and meticulous bibliographic curation. One can think of, for instance, the implementation of thesauri, and Work, Expression, Manifestation, Item (WEMI) entities, as part of the Functional Requirements for Bibliographic Records (FRBR) model or the IFLA Library Reference Model (LRM),[34] or the facility to compare and enrich the collection with metadata from the historical national bibliography to build substantiated arguments about the digitised versus historically published editions ratio. The possibility to select and filter based on extensive metadata and advanced search options is a crucial part of a corpus selection tool as a *publishing frame*. It makes an unexpected appeal to the publisher, i.e., the library: in this day and age, while the contents of an entire national bibliography are being fed to our digital reading machines, old-fashioned bibliographic analysis and curatorial expertise are still at the heart of collection building, especially when it comes to published digitised collections.

Conclusions

Digitisation of vast amounts of printed material is turning libraries into publishers of complex digital collections, for both humans to read and machines to mine. Through the lenses of publishing studies and digital bibliography, some of that complexity shows through. The Delpher collections at the KB, for example, contain two editions of each text and consist of several layers of publishing efforts, available via various interfaces. It's a powerful reminder that we should never equate the data with the service (or the *content* with the *frame*), and it raises a few fundamental questions for a (national) library, in terms of digital collection management, digitisation strategy and the increased significance of bibliographic control. On a very practical level, the recognition that digitisation makes a library a publisher may well,

and hopefully will, encourage these brand-new library-publishers to take facilitating computational research to the next level. The crucial first step to take: design easy-to-use services that support corpus building.

Notes

1 For the earliest history of digitisation at the KB, cf. Marieke van Delft, *Van wiegendruk tot word wide web. Bijzondere collecties en de vele geschiedenissen van het gedrukte boek* (Zutphen: Walburg Pers, 2015) and Astrid Verheusen, "Mass Digitisation by Libraries: Issues concerning Organisation, Quality and Efficiency," *LIBER Quarterly: The Journal of the Association of European Research Libraries* 18, no. 1 (2008): 28–38, https://doi.org/10.18352/lq.7902

2 Mass digitisation at the KB didn't stop just there. In collaboration with Metamorfoze (metamorfoze.nl), the Netherlands' national programme for the preservation of paper heritage, the KB keeps digitising Dutch printed material stored in archives and libraries in the Netherlands and elsewhere. The KB also became a partner in the Google Books programme in 2010 to digitise its books from the 18th and 19th centuries, and in the same year an agreement with ProQuest was signed to digitise the early printed books (1470–1700). The digital copies are now part of the Early European Books collection.

3 Data Services are not confined to the Delpher collections but support (bulk) access to all KB collections, catalogues and bibliographies, cf. kb.nl/dataservices (in Dutch).

4 lab.kb.nl/tool/polimedia, translantis.wp.hum.uu.nl, nederlab.nl, mediasuite.clariah.nl, impresso-project.ch

5 delpher.nl/data

6 lab.kb.nl/products/product_type/dataset

7 Thomas Padilla, Laurie Allen, Stewart Varner, Sarah Potvin, Elizabeth Russey Roke, and Hannah Frost, "Santa Barbara Statement on Collections as Data," Always Already Computational Collections as Data, 2018, https://collectionsasdata.github.io/statement

8 Sarah Ames, "Transparency, Provenance and Collections as Data: The National Library of Scotland's Data Foundry," *LIBER Quarterly: The Journal of the Association of European Research Libraries* 31, no. 1 (2021): 11, https://doi.org/10.18352/lq.10371

9 Henk Alkemade, Steven Claeyssens, Giovanni Colavizza, Nuno Freire, Jörg Lehmann, Clemens Neudecker, Giulia Osti, and Daniel van Strien. "Datasheets for Digital Cultural Heritage Datasets," *Journal of Open Humanities Data* 9, no. 1 (2023): 17, https://doi.org/10.5334/johd.124; Ames, "Transparency," 1–13; Paul Fyfe, "An Archaeology of Victorian Newspapers," *Victorian Periodicals Review* 49, no. 4 (2016): 551, https://doi:10.1353/vpr.2016.0039 and Tessa Hauswedell, 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 Newspaper," *Archival Science* 20 (2020)**:** 139–165, https://doi.org/10.1007/s10502-020-09332-1

10 Georgia Angelaki, Karolina Badzmierowska, David Brown, Vera Chiquet, Joris Colla, Judith Finlay-McAlester, Klaudia Grabowska et al, *How to Facilitate the Cooperation between Humanities Researchers and Cultural Heritage Institutions*, Warsaw: Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences, 2019: 12–13, https://doi.org/10.5281/zenodo.2587481 and Toma Tasovac, Sally Chambers, and Erzsébet Tóth-Czifra, *Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper*, 2020: 1–2, https://hal.archives-ouvertes.fr/hal-02961317

11 Angelaki, Badzmierowska, Brown, Chiquet, Colla, Finlay-McAlester, Grabowska et al, *Cooperation*, 8–10 and Sarah Oberbichler, Emanuela Boroç, Antoine Doucet, Jani Marjanen, Eva Pfanzelter, Juha Rautiainen, Hannu Toivonen, and Mikko Tolonen, "Integrated Interdisciplinary Workflows for Research on Historical Newspapers: Perspectives from Humanities Scholars, Computer Scientists, and Librarians," *Journal of the Association for Information Science and technology* 73, no. 2 (2022): 225–239, https://doi.org/10.1002/asi.24565

12 Thomas Padilla, *Responsible Operations: Data Science, Machine Learning, and AI in Libraries* (Dublin: OCLC Research, 2019), 1–36, https://doi.org/10.25333/xk7z-9g97 and Ryan Cordell,

Machine Learning + Libraries: A Report on the State of the Field, 2020: 42–50, https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf

13 Name recognition of one service, generally speaking the graphical search interface, can make a collection carry the name of that service in practice anyway. This should not detract from the importance of the adage not to equate the collection itself with the service.

14 Ryan Cordell, ""Q I-Jtb the Raven": Taking Dirty OCR Seriously," *Book History* 20 (2017): 199–200, https://doi:10.1353/bh.2017.0006

15 G. Thomas Tanselle, "The Bibliographical Concepts of 'Issue' and 'State'," *The Papers of the Bibliographical Society of America* 69, no. 1 (1975): 35, www.jstor.com/stable/24302244

16 Michael Bhaskar, *The Content Machine: Towards a Theory of Publishing from the Printing Press to the Digital Network* (London/New York/Delhi: Anthem Press, 2013), 60.

17 T.R. Adams, and N. Barker, "A New Model for the Study of the Book," in *A Potencie of Life: Books in Society: The Clark Lectures 1986–1987*, ed. Nicolas Barker (London/New Castle: Oak Knoll Press, 2001), 15.

18 J.B. Thompson, *Merchants of Culture: The Publishing Business in the Twenty-First Century* (Cambridge/Malden: Polity, 2010), 14.

19 Michael Bhaskar, *Content Machine*, 79–80.

20 "Het maakt niet uit wat voor onderzoek je doet en op welk niveau." (www.delpher.nl/over-delpher/wat-is-delpher/delpher-voor-iedereen)

21 See, for instance, the Data Foundry of the National Library of Scotland (data.nls.uk), the Open Data Platform of the National Library of Luxembourg (data.bnl.lu), the British Library Labs Digital Research Space (data.bl.uk), the BVMC Labs as part of the Miguel de Cervantes Virtual Library, Alicante (data.cervantesvirtual.com), in turn inspired by the GLAM workbench created by Tim Sherratt (glam-workbench.net), the data at the ÖNB labs of the Austrian National Library (labs.onb.ac.at), to name a few.

22 Jennifer Edmond and Vicky Garnett, "API's and Researchers: The Emperor's New Clothes," *International Journal of Digital Curation* 10, no. 1 (2016): 296, https://doi.org/10.2218/ijdc.v10i1.369

23 Max Kemman, Nick Jelicic, Guido de Moor, Marenne Massop, and Tommy van der Vorst, *User Needs for a Text Suite for Advanced Digital Research* (Utrecht, 2022), https://doi.org/10.5281/zenodo.6591572

24 Kemman, Jelicic, De Moor, Massop, and Van der vorst, *Text Suite*, 5.

25 John Sinclair, "Corpus and text: Basic principles," in *Developing Linguistic Corpora: A Guide to Good Practice*, ed. Martin Wynne (Oxford: Oxbow books, 2004), 16.

26 Andrew Piper, "Do We Know What We Are Doing?" *Journal of Cultural Analytics* 5, no. 1 (2020): 1–13, https://doi.org/10.22148/001c.11826

27 Andrew Piper, "Data, Data, Data: Why Katherine Bode's New Piece is so Important and Why it Gets so Much Wrong about the Field," TXTLAB, published June 23, 2017, https://txtlab.org/2017/06/data-data-data-why-catherine-bodes-new-piece-is-so-important-and-why-it-gets-so-much-wrong-about-the-field/

28 Ted Underwood, Patrick Kimutis, and Jessica Witte, "NovelTM Datasets for English-Language Fiction, 1700–2009," *Journal of Cultural Analytics* 5, no. 2 (2020): 7–11, https://doi.org/10.22148/001c.13147

29 Melvin Wevers, "Mining Historical Advertisements in Digitised Newspapers," in *Digitised Newspapers – A New Eldorado for Historians?: Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitization*, ed. Estelle Bunout, Maud Ehrmann, & Fréderic Clavert. (Boston/Berlin: De Gruyter Oldenbourg, 2022).

30 Marcel Broersma, and Frank Harbers, "Exploring Machine Learning to Study the Long-Term Transformation of News: Digital newspaper archives, journalism history, and algorithmic transparency," *Digital Journalism* 6, no. 9 (2018): 1150–1164, https://doi.org/10.1080/21670811.2018.1513337

31 Huub Wijfjes, "Digital Humanities and Media History: A Challenge for Historical Newspaper Research," *TMG Journal for Media History* 20, no.1 (2017): 4–24, http://doi.org/10.18146/2213-7653.2017.277

32 Melvin Wevers, and Thomas Smits, "The Visual Digital Turn: Using Neural Networks to Study Historical Images," *Digital Scholarship in the Humanities* 35, no. 1 (2020): 194–207, https://doi.org/10.1093/llc/fqy085

33 Antoine Peris, Evert Meijers, and Maarten van Ham, "Information Diffusion between Dutch Cities: Revisiting Zipf and Pred using a Computational Social Science Approach," *Computers Environment and Urban Systems*. 85, no. 101565 (2021), https://10.1016/j.compenvurbsys.2020.101565

34 International Federation of Library Associations and Institutions, *Functional Requirements for Bibliographic Records: Final Report* (München: Saur, 2009), https://repository.ifla.org/handle/123456789/811 and Pat Riva, Patrick Le Boeuf, and Maja Žumer, *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information* (Den Haag: International Federation of Library Associations and Institutions, 2018), https://repository.ifla.org/handle/123456789/40

References

Adams, Thomas R., and Nicolas Barker. "A New Model for the Study of the Book." In *A Potencie of Life: Books in Society: The Clark Lectures 1986–1987*, edited by Nicolas Barker, 5–43. London/New Castle: Oak Knoll Press, 2001.

Alkemade, Henk, Steven Claeyssens, Giovanni Colavizza, Nuno Freire, Jörg Lehmann, Clemens Neudecker, Giulia Osti, and Daniel van Strien. "Datasheets for Digital Cultural Heritage Datasets." *Journal of Open Humanities Data* 9, no. 1 (2023): 17. https://doi.org/10.5334/johd.124

Ames, Sarah. "Transparency, Provenance and Collections as Data: The National Library of Scotland's Data Foundry." *LIBER Quarterly: The Journal of the Association of European Research Libraries* 31, no. 1 (2021): 1–13. https://doi.org/10.18352/lq.10371

Ames, Sarah, and Stuart Lewis. "Disrupting the Library: Digital Scholarship and Big Data at the National Library of Scotland." *Big Data & Society* 7, no. 2 (2020): 1–7. https://doi.org/10.1177/2053951720970576

Angelaki, Georgia, Karolina Badzmierowska, David Brown, Vera Chiquet, Joris Colla, Judith Finlay-McAlester, Klaudia Grabowska et al. *How to Facilitate the Cooperation between Humanities Researchers and Cultural Heritage Institutions*. Warsaw: Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences, 2019. https://doi.org/10.5281/zenodo.2587481

Beals, Melodee H., and Emily Bell. *The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges*. Loughborough, 2020. https://doi.org/10.6084/m9.figshare.11560059

Bhaskar, Michael. *The Content Machine: Towards a Theory of Publishing from the Printing Press to the Digital Network*. London/New York/Delhi: Anthem Press, 2013.

Broersma, Marcel, and Frank Harbers. "Exploring Machine Learning to Study the Long-Term Transformation of News: Digital Newspaper Archives, Journalism History, and Algorithmic Transparency." *Digital Journalism* 6, no. 9 (2018): 1150–1164. https://doi.org/10.1080/21670811.2018.1513337

Candela, Gustavo, María Dolores Sáez, MPilar Escobar Esteban, and Manuel Marco-Such. "Reusing Digital Collections from GLAM Institutions." *Journal of Information Science* 48, no. 2 (2020): 251–267. https//doi.org/10.1177/0165551520950246

Cordell, Ryan. "'Q i-jtb the Raven': Taking Dirty OCR Seriously." *Book History* 20 (2017): 188–225. https://doi:10.1353/bh.2017.0006

Cordell, Ryan. *Machine Learning + Libraries: A Report on the State of the Field*, 2020. https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf

Edmond, Jennifer and Vicky Garnett. "API's and Researchers: The Emperor's New Clothes." *International Journal of Digital Curation* 10, no. 1 (2016): 287–297. https://doi.org/10.2218/ijdc.v10i1.369

Fyfe, Paul. "An Archaeology of Victorian Newspapers." *Victorian Periodicals Review* 49, no. 4 (2016): 546–577. https://doi:10.1353/vpr.2016.0039

Hauswedell, Tessa, Julianne Nyhan, Melodee 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* 20 (2020): 139–165. https://doi.org/10.1007/s10502-020-09332-1

International Federation of Library Associations and Institutions. *Functional Requirements for Bibliographic Records: Final Report*. München: Saur, February 2009. https://repository.ifla.org/handle/123456789/811

Kemman, Max, Nick Jelicic, Guido de Moor, Marenne Massop, and Tommy van der Vorst. *User Needs for a Text Suite for Advanced Digital Research*. Utrecht, 2022. https://doi.org/10.5281/zenodo.6591572

Mahey, Mahendra, Aisha Al-Abdulla, Sarah Ames, Paula Bray, Gustavo Candela, Sally Chambers, Caleb Derven et al. *Open a GLAM Lab*. Doha: Qatar: Digital Cultural Heritage Innovation Labs, 2019.

Mak, Bonnie. "Archeology of a Digitization." *Journal of the Association for Information Science and Technology* 65, no. 8 (2014): 1515–1526. https://doi.org/10.1002/asi.23061

McGillivray, Barbara, Thierry Poibeau, and Pablo Ruiz Fabo. "Digital Humanities and Natural Language Processing: "Je t'aime... Moi non plus"." *DHQ: Digital Humanities Quarterly* 14, no. 2 (2020). www.digitalhumanities.org/dhq/vol/14/2/000454/000454.html

Oberbichler, Sarah, Emanuela Boroş, Antoine Doucet, Jani Marjanen, Eva Pfanzelter, Juha Rautiainen, Hannu Toivonen, and Mikko Tolonen. "Integrated Interdisciplinary Workflows for Research on Historical Newspapers: Perspectives from Humanities Scholars, Computer Scientists, and Librarians." *Journal of the Association for Information Science and Technology* 73, no. 2 (2022): 225–239. https://doi.org/10.1002/asi.24565

Padilla, Thomas. *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*. OCLC Research Position Paper. Dublin: OCLC Research, December 9, 2019. https://doi.org/10.25333/xk7z-9g97

Padilla, Thomas, Laurie Allen, Stewart Varner, Sarah Potvin, Elizabeth Russey Roke, and Hannah Frost. "Santa Barbara Statement on Collections as Data." Always Already Computational Collections as Data, 2018. https://collectionsasdata.github.io/statement

Peris, Antoine, Evert Meijers, and Maarten van Ham. "Information Diffusion between Dutch Cities: Revisiting Zipf and Pred Using a Computational Social Science Approach." *Computers Environment and Urban Systems* 85, no. 101565 (2021). https://10.1016/j.compenvurbsys.2020.101565

Piper, Andrew. "Data, Data, Data: Why Katherine Bode's New Piece Is So Important and Why It Gets So Much Wrong about the Field." *TXTLAB*, June 23, 2017. https://txtlab.org/2017/06/data-data-data-why-catherine-bodes-new-piece-is-so-important-and-why-it-gets-so-much-wrong-about-the-field/

Piper, Andrew. "Do We Know What We Are Doing?." *Journal of Cultural Analytics* 5, no. 1 (2020): 1–13. https://doi.org/10.22148/001c.11826

Riva, Pat, Patrick Le Boeuf, and Maja Žumer. *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information*. Den Haag: International Federation of Library Associations and Institutions, January 2018. https://repository.ifla.org/handle/123456789/40

Sinclair, John. "Corpus and text: Basic principles." In *Developing Linguistic Corpora: A Guide to Good Practice*, edited by Martin Wynne, 1–16. Oxford: Oxbow books, 2004.

Tanselle, G. Thomas. "The Bibliographical Concepts of "Issue" and "State"." *The Papers of the Bibliographical Society of America* 69, no. 1 (1975): 17–66. www.jstor.com/stable/24302244

Tasovac, Toma, Sally Chambers, and Erzsébet Tóth-Czifra. *Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper*, 2020. https://hal.archives-ouvertes.fr/hal-02961317

Thompson, John B. *Merchants of Culture: The Publishing Business in the Twenty-First Century*. Cambridge/Malden: Polity, 2010.

Underwood, Ted, Patrick Kimutis, and Jessica Witte. "NovelTM Datasets for English-Language Fiction, 1700–2009." *Journal of Cultural Analytics* 5, no. 2 (2020): 1–30. https://doi.org/10.22148/001c.13147

van Delft, Marieke. *Van wiegendruk tot word wide web. Bijzondere collecties en de vele geschiedenissen van het gedrukte boek*. Zutphen: Walburg Pers, 2015.

Verheusen, Astrid. "Mass Digitisation by Libraries: Issues Concerning Organisation, Quality and Efficiency." *LIBER Quarterly: The Journal of the Association of European Research Libraries* 18, no. 1 (2008): 28–38. https://doi.org/10.18352/lq.7902

Wevers, Melvin. "Mining Historical Advertisements in Digitised Newspapers." In *Digitised Newspapers – A New Eldorado for Historians?: Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitization*, edited by

Estelle Bunout, Maud Ehrmann, & Frédéric Clavert. 227–252. Boston/Berlin: De Gruyter Oldenbourg, 2022. https://doi.org/10.1515/9783110729214-011

Wevers, Melvin, and Thomas Smits. "The Visual Digital Turn: Using Neural Networks to Study Historical Images." *Digital Scholarship in the Humanities* 35, no. 1 (2020): 194–207. https://doi.org/10.1093/llc/fqy085

Wijfjes, Huub. "Digital Humanities and Media History: A Challenge for Historical Newspaper Research." *TMG Journal for Media History* 20, no.1 (2017): 4–24. http://doi.org/10.18146/2213-7653.2017.277

Wilkinson, Mark D., Michel Dumontier, IJsbrand. Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3, no. 160018 (2016). https://doi.org/10.1038/sdata.2016.18