Beyond The Page-Break: Towards Better Tools for Remediation of Born-Digital Documents

Mark W. R. Anderson
Web and Internet Science Group
University of Southampton
Southampton, Hants, United Kingdom
mwra1r19@soton.ac.uk

ABSTRACT

A legacy of print is that much of our process and tooling is predicated on using text in paginated form, such as was required for (paper) printed media. Increasingly, digitally-created ('born-digital') documents will never be used non-digitally and yet their internal structure is still optimised for paginated presentation. As modern displays now offer the affordance of decomposition and remediation of complex documents, this article explores the limitations of this legacy design for post-paper-print era use and considers the design implications for writing tools intended for the new digital era.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; • **Applied computing** → **Hypertext / hypermedia creation**; • **Software and its engineering** → **Virtual worlds software**; *Software evolution*.

KEYWORDS

documents, hypertext, XR, VR, AR, data, remediation, exploration, narrative, provenance, metadata, linkbases, stand-off metadata, citation, addressing, AI, machine reading, viewspecs, visualisation

ACM Reference Format:
Mark W. R. Anderson. 2024. Beyond The Page-Break: Towards Better Tools for Remediation of Born-Digital Documents. In *35th ACM Conference on Hypertext and Social Media (HT '24), September 10–13, 2024, Poznan, Poland.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3648188.3678215

1 INTRODUCTION

January 2024 was a milestone for ACM-published papers as they will no longer be automatically available in print, i.e. non-digitally. Yet the long shadow of paginated print remains and in ways not necessarily helpful to consumption in emerging digital environments. A useful provocation is the XR (E**x**tended **R**eality) [1] environment that offers the possibilities of decomposing/deconstructing complex texts, but only if the source documents have suitable data to allow this. Academic papers, such as that which you are reading now, are generally constrained by the fixed nature of typeset print—at least within PDF format—thus limiting the reader's ability to explore the document's structure, linked data, and references.

This potential exploration involves a dynamic remediation, a reconstruction of the source in the virtual space. Beyond rendering a mere facsimile of a paper document, meaningful re-use of the source material needs to give accessible clues as to its internal structure and linkages—both internal and external. Today, these enhancements are generally lacking, or effected outside the document, e.g. on a host web page. Given the newness of XR, the tools for generating such environments are immature and still evolving. However, it is surprising how little our writing (i.e. text creation) tools have matured to service this new environment for consuming text.

Here, 'remediation' draws upon Bolter & Grusin's *Remediation* [22], extending the scope of the term. They were broadly considering transition between established, distinct, media—e.g. a work moving from book to film. Bolter's 2001 updated *Writing Space* [21, (Ed.2) Ch.3] also considered 'Hypertext And The Remediation of Print' as well as graphics (Ch.4) and the early eBook (Ch.5)[2]. He saw that (hyper)text, once freed from the shackles of fixed pagination, could be used more flexibly.

Earlier, Joyce—Bolter's co-creator of the hypertext authoring system *Storyspace*—noted [46] that a hypertext tool could be constructive even for non-literary creative use. The malleability of presentation in the XR environment is a natural provocation to on-the-fly remediation of sources previously only accessible in the fixity of paginated print. Sect. 3.1 discusses remediation further.

We are now quite used to the term 'document' and (computer) 'file' being used interchangeably, but that change is quite recent (c.1967) [64]. This semantic drift also plays into the notion that (primarily) textual works should be understood in terms of their printed form, i.e. as seen on paper—be it literal or digital.

Pertinent here to 'paginated-print-think', is McLuhan's notion of 'rear-view thinking' when faced with the new:

> The past went that-a-way. When faced with a totally new situation, we tend always to attach ourselves to the objects, to the flavor of the most recent past. We look at the present through a rear-view mirror. We march backwards into the future. [56, pp.74–5]

Indeed, now we need to look beyond the familiar conventions of (paginated) print. Happily, Hypertext's forward-looking community has always acknowledged a wider context to the understanding of text than as (paginated) print alone. From inception the Hypertext

[1] A collective term for immersive **V**irtual **R**eality (VR) and AR (**A**ugmented **R**eality). AR is reality overlaid with projected VR elements.

HT '24, September 10–13, 2024, Poznan, Poland

ACM ISBN 979-8-4007-0595-3/24/09
https://doi.org/10.1145/3648188.3678215

[2] Then, Web (hypertext) was far less capable and its use less widespread than nowadays.

Conference, eschewing Snow's 'Two Cultures' [74, pp.3–4], has not perceived a strict divide between formal writing and the looser creative and literary use of text: literary hypertext was never an 'other' but a peer to documentation and academic writing [5, Sect. 2.4], etc. As such, this community is well placed to consider how text creation tools may improve usefully, to move beyond the legacy constraints of print as their primary output.

Consideration has been given to given to 'digital reading', but focus tends to remain on the book metaphor [50, 67], and less on reading outside the page(-derived) form [51]. In exploring new interactions with text, we should not simplistically assume that they apply to *all* existing use of text. As will be seen, some contexts still need an immutable copy of record[3] (see Sect. 3.1.1).

To explore the implications of these new opportunities, we will start by reviewing some of the contributory background elements—documents, files, paginated (paper) print, remediation—before considering how hypertext's past work may contribute to better text creation tools and address (unintentional) limits of our current tools and interchange formats. Due to limited space, the following contingent issues are explicitly not addressed, including: annotation of documents, application coding and design, and application UI. These are relevant but require their own proper, discrete, exploration.

XR is also *just one example* of the challenge ahead. Herein lies the provocative idea: what of the print era is useful as opposed to merely familiar?

2 HOW WE ARRIVED HERE

2.1 What is a 'document'?

As traced by evolving definitions of 'document' in the the Oxford English Dictionary (OED) [63] the older meaning of an evidential artefact has been joined by a more diffuse meaning of a collection of digital data that is considered a single item. Whilst in the 1980s it might have implied a single 'file' [65], that too has blurred.

2.1.1 So is a 'document' one file—or more? Nelson noted "Computing is made up of files and directories and that's a tradition left behind from the 1940s that no one questions, … Another tradition is that one file equals one document." [10], indicating legacy behaviour from before the PC boom of the 1980s–90s which feeds into the blurring of what comprises a digital document.

If an early general notion of a one-to-one equivalency of a document and a digital file held true, the subsequent increased complexity of tools and formats has blurred that assumption. Commonly, files of major text-centric formats are actually a zip file masquerading as a single file: it holds the 'text', plus inserted assets like images, configuration, and metadata files: a DOCX file is XML+content, PDF is PostScript+content, and ePub is HTML+content, etc.

It may matter little to the author that accessing such a composite file that hides its wider structure, but it is less trivial for a process needing to extract the textual (or image, etc.) content for use in the context of a different medium of display (see Sect. 2.2.2).

2.1.2 Write vs. Read. It is also the case that many formats cannot be (screen) read as stored in their input form. They need either to be printed or first exported in some other form/file. Either the creating tool presents the user with a WYSIWYG[4] display or a discrete viewer app/mode is needed to 'read' the exported content. This leads to a notionally single document concurrently existing in several discrete forms, unlike print (ignoring deliberate print redesign, e.g in re-prints/re-issues).

2.1.3 How does a digitally native document differ from older notions of a document? Prior to digital text, a document would be written, or imprinted, onto a physical medium (paper, parchment, etc.). The physical artefact is the result of writing the document, yet it also constrains the text imposed on it. Form and function interact to give emergent norms such as today's A4 paper sheet. The source file(s) used to write a digital document is often not the format used to read or distribute it. Regardless of how many file(s) form the source, there will usually be a discrete output format (file). Normally, the latter is optimised for printing even if only ever read digitally. Thus, we tend treat the *output file's* form as the 'document' artefact.

2.1.4 Pagination and page numbering. Page numbering originated as a convenience when assembling a book [35, pp.124–125], and found use as an anti-tamper device [3, p.37], before taking on its current role in the 1500s [11, p.33][5]. Wider use of indexes and tables of content soon followed [35, pp.127 *et seq.*], though initially their purpose had to be explained to the general reader [31, Ch.1]. These ideas have lasted well through the print era[6], but re-flowable, re-mediated, digital texts pose a new challenge (see Sect. 5.0.3).

2.2 The long shadow of PDF and paginated print

2.2.1 PDF: the good part. Invented in the early 1990s, Adobe's Portable Document Format (PDF) [81] solved an erstwhile problem of the difficulty of achieving consistent *printed* output from different PC software and printers. PDF is thus an *output* file format (see Sect. 2.1.2). It was rapidly apparent that besides offering consistent layout and appearance a PDF document, with suitable 'reader' software, could stand proxy for its printed version as 'digital paper'. This both gave convenience, as well as saving on costs of paper, ink, and storage. So began the rapid ascent of PDF to its dominant role as the (output) digital document of choice. Indeed, it is most likely you are reading this article as a digitial PDF, albeit this *not* being PDF's original purpose.

2.2.2 PDF: the not so good part. In the rush to avoid still needing paper (cost!), overlooked was how a PDF works its magic. A key task is to take the desired typography and reproduce it faithfully, *as if on a (paper) page*. This is achieved via the PostScript language[7] that creates vector-based shapes representing small runs of text (often only a few characters). Surprisingly, there is no complete single copy of the input text in the PDF. To copy text, it has to be reversed out of the PostScript. As a result it is polluted with typographic artefacts added in 'printing' the source text such as ligatures, soft hyphenation, etc. Many users of such text will have little idea how to safely remove these artefacts without also damaging the text. Non-body-text items, such as footnotes, are also problematic as

[3] Immutable, even if only by controlled access to the original (digital document).

[4] "**W**hat **Y**ou **S**ee **I**s **W**hat **Y**ou **G**et'. The origin is disputed, but it appeared in the late 1970s. The PARC 'Bravo' system is generally held to be the first WYSIWYG text editor.

[5] Cites Saenger *The History of the Book in the West: 1455–1700, Vol.II* [72, p.403].

[6] For more detail on the evolution of the paginated print book see *The Book* [45].

[7] By Warnock and others (at Adobe) from 1982–84. Also: *Thinking in PostScript* [70].

their content is simply inserted into the recovered text where the page or column end occurs[8]. Table data and image captions are similarly poorly handled when retrieved from PDF.

Here, Nelson's dismissal of 'WYSIWYG', as being WYSIAYG—what you see is *all* you get—perhaps applies. Although his critique was actually of erstwhile word processor software, he rightly points up the limitations of over-concentration on the surface layer at the cost of what lies beneath and within. Extracting from PDF, you get what is 'on the page' and little more, including unwanted typographic elements never 'in' the source text, e.g. page numbers.

2.2.3 PDF: the legacy. Thus, even un-printed, the PDF cleaves strongly to a print representation, with text and typesetting commands freely intermixed in the stored (PostScript) data, in a fashion that is hard to de-interleave (see Sect. 4.0.1)[9]. However, we should not blame PDF unduly for not doing well things for which it was never designed.

In 2020 Adobe introduced a PDF 'Liquid Mode' [10], that re-flows text dynamically—via a means undocumented[11]. Whilst assisting small screen users, support is only for limited platforms, and impact on digital re-use of PDFs is limited. The limited success of the latter demonstrates the challenge of using AI to find—consistently and accurately—structure either never stored in the document or stored in a manner not amenable to simple recovery. Thus the 'problem' here is not use of PDF as such but the accidental over-dependence on being able to read the print-form render to understand fully the document's content. PDF has a rich potential internal structure but few general tools use that, even today.

The challenge for PDF is less, 'can the format be fixed?' than unthinking and inappropriate use of the format. Two-column print layouts, such as used by this Conference's Proceedings, offer efficient packing of text upon the page. Primarily, this lessens paper/ink costs, yet this saving is of no real value if a document is never printed. Conversely, multi-column use increases the occurrence of unwanted typographic artefacts (e.g. embedded soft-hyphenation[12]) in copied PDF text.

Post-hoc justification of PDF use abounds, often as an 'optimum format', but note that ACM is now among publishers offering papers in both PDF *and* HTML form[13]. This raises an interesting question of which form is considered the canonical copy of record; for instance, where there is later dispute over original content (q.v. Sect. 3.1.1).

3 ACKNOWLEDGING THE POST-PRINT ERA

Printing of books, magazines and newspapers continues, but digital consumption of text 'print' is declining [11, 24]. This change will be gradual, but for those moving away from printed media[14] there are already challenges in making best use of digitally native print, not least when remediated into short texts.

[8]This begs the question of what 'the text' comprises.
[9]Aspects of this issue were also reported in [4, Sect. 4.4].
[10]Note this mode has limited hardware and software support, and is tied to a single vendor as the PDF must be 'converted' for this type of use [1].
[11]Adobes credits a 'Sensei' AI. Liquid Mode does not alter the PDF but *tries* to detect structural elements to assist content re-flow for presumed small-screen use.
[12]Seen down the right edge of a column of text in PDF as unusual word breaks.
[13]The latter is currently only offered for papers generated using ACM's TAPS system.
[14]Either physical media or digital 'printed' media.

A further factor is the increase in variety of digital text display contexts. The replacement of early CRT monitors with lighter, flatter, devices has seen display embrace many sizes and formats: smart watches, mobile phones, tablets, TVS, desktop displays—and XR.

3.1 The future will be remediated

Apart from text imprisoned in paginated—fixed—typography, the new range of display sizes implies a need for dynamic remediation of the text. In simplest form, this means the ability to re-flow text within the display area rather than scaling fixed pages to fit the display. Whilst this works for a simple linear text-only work, such as a novel in HTML-based eBook form[15], more complex documents fare badly as these simple remediations give little thought or care as to inclusions in the body text such as images or footnotes.

XR opens a new aspect of display for digital text documents where there is no fixed display size and no particular style of display. In such an environment, is the traditional single linear presentation the most useful or informative presentation? Digital documents will benefit from separating the textual content (stream) from the data describing its use for a particular type of display. Yet the need for and value of remediations will vary with the type of document. It will also require additional internal metadata often not currently available.

3.1.1 Not all documents need nor want a mutable format. For some documents, their value is as a persistent record. Contracts, treaties, laws, bills of sale, and the like, all need a version of record: the 'document' version in the older sense of that word. It is important to be able later to refer to—and have faith in the veracity of—such a documentary record of note. Even so, other than by past convention, does such a record need to be in a fixed typographical form? How may the canonical form(at) be recognised?

3.1.2 Not all document structure is visible. Historically, to get at the structure of a document we have needed to break away from 'paginated-first' representation. The internal structure of the document and its internal (hypertextual) links normally have to be inferred, even in a document creation tool. This is a poor fit for the weakness of current 'AI', which finds patterns with ease but is less sure-footed where non-explicit structure or meaning must be inferred. for the author, being able to capture this implicit structure (in machine-understandable form) is useful both to our future selves and to other human and non-human users of the document, leaving less to guesswork in its remediation.

3.1.3 Addressing the correct text. When referring to known specific text it helps to be able to address a persistent anchor. Traditional outline headings do not necessarily offer the granularity desired for fine-grained reference (see Sect. 5.0.3).

3.1.4 Retaining context. Space-efficient typography results in contingent matter—text with attendant notes, tables, etc.—being placed away from the relevant text. With a single display, flipping from page X to page Y, as is needed to see both a table and the text that discusses it, is harder on screen than with paper.

[15]Confusingly the term 'eBook' is used to apply to both books set in fixed PDF typography and to those stored/rendered as (X)HTML.

Pertinent too is the issue of two-way links, lost in the Web's simplification of hypertext but long known to the Hypertext community[16]. It is as important to be able to come back to the original place as it is to be able to visit another place (see Sect. 5.0.5).

3.2 Writing for whom: author, reader or parser?

Traditionally, the author has control. In all but the most rigidly structured documents (e.g. a legal statute), the author is free to weave a narrative braid though the content that may result in differing strands of it being scattered in pieces through the overall work. Reconnecting these strands is a task the human reader learns to intuit: a parsing algorithm has less to go on. Print also sustains use of linearised presentation, which may work for, or against, the narrative. We should also be aware that the narrative is implicitly structural, even if not overtly so.

As more of our text lives entirely digitally, much that we 'write' may be read more often by software than by the human eye [43]. This makes it all the more sensible to give digital hand-holds to the 'reader' to expose the internal connections within a document.

3.2.1 Authorial control subverted. In Barthes' *Death of the Author* [12] post-structuralist literary criticism explored the notion that the meaning of the text resides not entirely in the author's exposition, or intent: the reader's perspective is considered as relevant. Historically this was an argument about interpretation. Now, a presentationally fluid medium like an XR increases the potential agency granted the reader in their (re-)interpretation of the work's order and narrative, whilst consuming its text.

In reading, exploring the internal and external relationships of the document are natural behaviours, though this is limited without appropriate access to the structure of the document. The reader's new ability for manipulation of text does not rob the author of agency, but it might suggest some changes as to how we write for consumption in a purely digital context (outside purely creative works). Rather than relying on the props (constraints?) of the past such as pagination and outlines, the author is free to offer additional metadata and structure[17] that allows a document to be explored in a remediated, non-linear manner. This is a style with which the Hypertext community is unusually experienced and thus is well placed to contribute to the exploration of such change.

3.3 Additional considerations for academic documents

Though documents of other types can share the issues below, academic 'papers' [66] make a good general example of some issues relating to internal structure[18]. Academic documents are atypically dense in terms of internal referencing (links) as well as in citation of external sources. There is now also the opportunity for data, either as source for or result of research, to be much more tightly bound to the documents, as digital links can be resolved in real time.

Some parts of a paper are explicitly labelled and set apart from the main narrative. The abstract and references are the most obvious examples of what would benefit from being easily accessed as a discrete unit. An abstract is only 'closed' by the presence of a next heading (and is not always only single paragraph), whilst a list of references (citations) normally comes at the end of all content—barring any appendices. These items' scope is obvious on the printed page, but may be less obvious internally, as in a PDF.

Footnotes (or end-notes) are challenging as these glosses are deliberate excursions from the primary linear course of the text. But in a digital, un-paginated setting they are of best use close at hand to the text they gloss.

Tables and figures (images) also often sit uneasily in the main text. They are often placed away from the text they explain due to now-legacy issues of print typesetting and paper costs. With suitable internal links or metadata, these collections of content—text and tables/figures—could become addressable units of content.

Data is less hidden than in the past. Published papers increasingly give explicit pointers to data underlying their narrative and often demand an *implicit* degree of contextual knowledge from the reader. A factual document's narrative is tightly bound to data, be it as provenance of sources or as the output of the research, particularly so in the sciences. In a digital environment, that separation of text and data is unnecessary. Thus tables and figures *could*, where pertinent, become less static in presentation. Note though, the need to have an immutable 'version of record' would likely still remain[19].

4 CHALLENGES FOR REAL-TIME REMEDIATION

Information overload, and the challenge of recording it effectively, is nothing new: consider Blair's *Too Much to Know* [19], Linnaeus' 'slips' [77], and Luhmann's *zettelkasten* [49]. Early in the use of personal computers, Nelson noted in *Computer Lib/Dream Machines*:

> Hierarchical and sequential structures, especially popular since Gutenberg, are usually forced and artificial. Intertwingularity is not generally acknowledged—people keep pretending they can make things hierarchical, categorisable and sequential when they can't. [61, p.'DM31']

and, expanding on this, he notes:

> EVERYTHING IS DEEPLY INTERTWINGLED. In an important sense there are no "subjects" at all; there is only all knowledge, since the cross-connections among the myriad topics of this world simply cannot be divided up neatly. [60, p.'DM45']

However, the ability to provide metadata to the narrative text of document, to indicate the internal links and invisible structures makes the latter now less challenging than in the past, and requires the author's attention to detail—something new tools *ought* to assist. Yet, this ability to honour a text's underlying intertwingularity has been widely overlooked in most popular text creation software.

4.0.1 Typeset text and remediation. Storing the source text of a document with hard-coded typesetting (e.g. ligatures[20], soft hyphens, etc.) privileges print-based viewing and passes, when extracted, 'dirty' text into any remediation.

[16] Ironically, PDF does support two-way intra-document link navigation.
[17] Text tools supporting such enhancement are missing or immature.
[18] Generally, these (peer-reviewed) papers are formally-published, long-form, essays or reports on a topic.
[19] For provenance, replication, and as a bulwark against corruption or malicious alteration of facts.
[20] Note: in some languages ligatures are structural and not merely stylistic artefacts.

Documents written with remediation in mind need—or gain significantly from— the ability to extract and expand the compressed narrative within. Whether this is to 'decompress' the document and see the main narrative in the context of its sources and peers, or to allow the information to be envisioned from differing perspectives, a task potentially aided by the potential malleability of an XR space.

4.0.2 Humans and computers read differently. Bush was early to contrast human associative thought, with the (then) strictures of erstwhile (computing) technology [26, Sect. 6, paras 1–2]. Despite the latest wave of interest in AI[21] it is not clear that algorithms can understand human thought patterns and how we infer meaning. This leaves work for the author, especially if the primary reader is non-human (the choice of reader is rarely as chosen by the author).

5 INTERNAL INTERWINGULARITY

Even with the Web's sub-set of wider hypertext, we have a fair appreciation of links between documents. Internally, this is less the case[22] and where our text creation tools hide this 'messiness' from the text creator. Yet, intra- and inter-document links are essentially the same, if differing in purpose and use. Hiding the internal structure of the document may simplify the writing experience, though not necessarily to the author's benefit.

Remediation, especially when effected as decomposition of a document in to discrete parts, benefits from rich metadata as to the document's structure (q.v. Sect. 4). For academic papers, such structure is normally not shown in output documents beyond outline headings and a few elements like a reference list. Much of the rest is inferred from the narrative and reader's prior knowledge of the domain. For enabling remediation, writing tools need to be able to assist authors to record the structure of their narrative even though it is not part of the visible text output.

5.0.1 Exposing content. Engelbart's *NLS/Augment* had the notion of 'viewspecs' [32] [33, Sect. 7b3]. These are easily-invoked, differing, visualisations of the *same source data.* Today, the notion of diverse viewspecs within a program can best be seen expressed in Eastgate's *Tinderbox* [16]. More widely, viewspecs are generally underused at present[23]. Sadly so, as a single UI (viewspec) fosters a singular view of the 'shape' of the document and its narrative/content.

Text need not be constrained to orderly print. Brath's *Visualizing with Text* [25] shows text to be tractable in presentation style without losing coherence—it can increase the meaning imparted. For those who balk at abstract presentation, the 'card' metaphor best known from Atkinson's *Hypercard*[24] [38] offers a less challenging but equally flexible approach to exploring the structure of a text[25].

5.0.2 Spatial Hypertext. An overlooked aspect of visualising (hyper)text is Spatial Hypertext (SH), which can help exploration of a document's structure and internal/external links. SH has seen parallel evolution as an algorithmically-mediated forms and as a manually generated map. The former, has lineage from *Aquanet*[26] [52], via *VIKI* [53] and *VKB*) [73]. Current work in this form is best seen in Atzenbeck's *Mother* system [7–9]. The non-algorithmically mediated SH started with *Storyspace* [15, 23] where it still remains in use; Eastgate's *Tinderbox* [16, 17] builds on this for more general writing use.

SH does not demand (visible) text, larger maps can be accommodating via abstractly styled notes (nodes) using simple shape, colour, etc. Link display can be (temporarily) suppressed if it adds visual 'noise' to the map. The spatial layout also lends itself to recognition of negative space: what is not mapped, linked, or simply unknown. Bernstein's *Can We Talk About Spatial Hypertext?* [18] provides a useful summary of the vocabulary and affordances of SH.

5.0.3 High Resolution Addressing. Effective remediation also needs good link addressing so that the disparate objects arising still know their relation to the whole and each other. Accurate addressing of content was a concern of the *NLS/Augment* project from outset [33, Sect. 6], and carried though to the early web by the notion of 'purple numbers' [47][27]. HTML allows targeting of headings and (unique within-page) ID anchors, though these affordances are not foregrounded by word processors, the most widely used text creation tools.

Within the paper/document-based metaphor, 'page', 'paragraph' and 'sentence' give some accuracy but are brittle if text is re-flowed, although close addressing should be robust to re-flow. The Kindle e-reader has the notion of a `#loc` every 150 bytes [28] [31, p.99]. However e-readers seem designed more for basic linear reading where the primary need is the last-read-to place in the piece, whilst still acknowledging the source pagination. The more complex structure of academic communication warrants better addressing tools. These are not unexplored—q.v. initiatives such as the long-lived TEI [29, 76] and JATS (Journal Article Tag Suite) [62, 80].

5.0.4 'Block-level' addressing. In the area of PKM[29] software there is the Web/HTML-derived notion of addressing paragraph 'blocks'[30] as the object unit. PKMs like *Roam* and *Obsidian*[31] use this approach. *Roam* also re-surfaces the pre-Web concept of links being two-way—see Sect. 5.0.5 below.

5.0.5 Links and linkbases. As well as visualising the structure within the document, and being able to accurate address content within the document, it is useful to be able to record (explicit) and explore (implicit) links within a document. Assistive (AI) processes can help with the former, but the latter benefits from being able to visualise the document and surface links not yet explicit. Before the Web set our general perception of hypertext and embedded links, the emergent Open Hypertext Systems (OHS) included the notion of a linkbases that stored links separately from the content. Normally this was via character offsets in the text stream. At the time text was generally simple unstyled 'plain' text, but if the text's style is stored discretely[32], this method can still be use for today's

[21] AI has been through a number cycles of (over-)promise and failure to deliver.
[22] A fair exemption being documents expressly written as hypertexts.
[23] Even MS Word has 4 viewspecs (View menu) but most only use the default 'Print'.
[24] Hypercard's origin is interesting [6], reflecting on the interconnectedness of things.
[25] Scrivener's 'corkboard' view uses a card metaphor for exploring narrative order [20].
[26] Aquanet itself drew on *Notecards* [41] and *gIBIS* [27]: also see *SEPIA* [75].
[27] Looking further back, in 1965 before displays were common or 'hypertext' was defined, Licklider was reflecting on issues of citations (and the implicit links) [48, p.56]
[28] N.B. This is not a standard. Also see: MobileRead wiki article'Page numbers' [57].
[29] PKM: **P**ersonal **K**nowledge **M**anagement.
[30] A line or paragraph being all the text between two literal line-breaks (or start/end).
[31] See: *Roam* (https://roamresearch.com) and *Obsidian* (https://obsidian.md).
[32] Thus separating pure textual content from semantic/stylistic data.

more styled text. An interesting new approach is 'standoff metadata' [58, 59] based on JSON use and addressing a need in Digital Humanities for overlapping link anchors and targets.

Unlike the Web's HTML-embedded links, earlier hypertext systems embraced discrete link storage: the linkbase. Linkbases could be multiple and contextually applied as in *Microcosm* [36, 42], or federated as seen in *Hyper-G* [55]. Grønbæk and Trigg's *From Web to Workplace* [40] also usefully summarises OHSs.

5.0.6 Link types. Important to the expressing a document's inner relationships is to capture the purpose(s) represented by a link. Early hypertext explored link types in close relationship to structuralised argumentation, most fully expressed by Trigg [78][33]. Sadly, distaste for the effort of setting link types led to disfavour, and link types were not actively adopted by the Web. However, they remain a strong potential affordance for remediation by categorising the relationships of content within a document.

5.0.7 Footnotes and other non-linear inclusions. The footnote arises mid-16c.[34] allowing branching from the primary linear narrative. A fully linear narrative can be a textual straitjacket, as glosses or insertions intrude. Thus in print, typesetting costs often shunted these to being 'endnotes'. A different presentational gloss, expensive typographically (and in paper paper use), is the marginal side note as is well displayed by Tufte's books (e.g. *The Visual Display of Quantitative Information* [79]). A large page margin might be wasteful of space in print form but can still elegantly allow space for text expansion[36], yet such a need is moot in a remediated environment like XR—marginal space being only as context requires.

The sidenote concept sits well with remediation: if a (foot/side/end) note knows its place in the text, in a remediated space its co-presentation as a sidenote seems an obvious provision. Similarly, figures, tables, and formulae are often inserted far from their associated text. Given metadata, representation of supporting content alongside text is an obvious effective re-use to aid presentation alongside its relevant text, even if there are multiple such pieces of text. Indeed, associated but discontinuous runs of text could be remediated as a continuous piece, alongside supporting material.

5.0.8 Structured abstracts. An innovation aiding granularity of addressing is the structured abstract [44]. This concept arose not in hypertext, but in the medical community in the late 1980s, and uses mandated labelled sections in abstracts. These guide both the human reader and offer tractable structure for computer readers. However, they could also, for remediation, use embedded metadata to indicate the purpose of sub-headings and link to relevant parts of the document.

In some areas of academic writing this structure could be extended further into document, though it is hard to generate generic heading/sections that are useful across all domains. But building on the structural notion it could inform document metadata and embedded anchors to aid both addressing and parsing of the whole.

5.0.9 Temporal aspects. Rosenberg's *Cartographies of Time* [71] shows visualising the temporal axis of information and text has a rich past. Others explore this through into the digital age [2, 25]. A remediation challenge is to capture dates meaningfully: not all dates have the same pertinence, so some authorial metadata hiting is useful. Anderson and Millard [4, p.103, Sect. 5 incl. fn. 21–23] explored the temporal axis of citation links within the hypertext Conference corpus[37] and such methods could be usefully expanded for documents that are collections or have large numbers of references. Time has a natural axis but temporal relationships can also be seen as a proxy for other exploitable sets of terms (code-accessed metadata?), such as people, organisations, or domains of study.

5.1 Text and data

Pressure for more effective re-use, such as expressed by *FAIR* [82][38] and Web/internet methods of close linking (q.v. Sect. 5.0.3), suggest academic papers and research documents could more gainfully link to their underlying data. By using SVG instead of bitmaps, charts could link (where pertinent) through to source data allowing for more expressive representation in remediation. Notebook tools [13, 69] could also leverage more expressive internal metadata.

6 CONCLUSIONS

Naturally-arising limitations of print place constraints on current tools for creating text. The advent of the likes of XR does not mean 500 years experience of printed text is suddenly of lesser value. However, print's conventions and constraints are not ideally suited these newer forms of textual presentation. This is the challenge for writing tools to address. A lesson from PDF is to not use a valid format outside its design intent without due consideration. It is too early to judge whether existing writing tools will evolve to embrace the needs of remediation by aiding the recording of significantly more structural metadata. Nor does it matter unduly, whether it is new or existing tools—or both—that take up the challenge.

The provocation of the likes of XR remediation is not simply a technical challenge. Rather it leads to (re-)consideration of our teaching and practice of writing. No longer are we only writing linear text. The latter can contain complex and multi-stranded narrative, but its understanding often (assumes) implicit communication direct from author to reader, disadvantaging computer-mediated exploration and consumption of text.

A brief essay leaves little room for depth but it is to be hoped that the challenges and opportunities here are clear. The Hypertext community has much to offer in this context. Some ideas may appear old but that shows the forward thinking nature of the original work, that often ran beyond the technological capabilities of the day: yet they still hold value.

ACKNOWLEDGMENTS

Dave Millard, Peter Wasilko, Dene Grigar, and Mark Bernstein kindly commented on drafts of this paper. Ideas here draw on the *Future of Text* group's [37] XR research (Sloan grant [39]), data re-use for UK's PSDI [68] and over two decades' use of Eastgate's *Tinderbox* [16] and *Storyspace* [15] tools and supporting their users.

[33] This work was part of the first PhD awarded for Hypertext.
[34] Zerby *The Devil's Details* [83, Ch.2] to give less intrusive glossing. Of course, footnotes can have footnotes[35].
[35] Hypertexually-informed narrative can support branching. See: Sect. 3.1 and [14].
[36] Body text expansion into margins in Even-Ezra's *Lines of Thought* [34].
[37] Partly inspired by *Citeology*'s [54] temporal visualisation, and by arc diagrams.
[38] Also see national data sharing initiatives in USA [28], UK [68], and Germany [30].

REFERENCES

[1] adobe.com. 2020. *What is Adobe Liquid Mode?* Adobe Inc. Retrieved January, 10 2024 from https://www.adobe.com/acrobat/hub/what-is-adobe-liquid-mode.html

[2] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. 2023. *Visualization of Time-Oriented Data.* Vol. 4. Springer Nature, New York, NY, USA. 441 pages. https://link.springer.com/book/10.1007/978-1-4471-7527-8

[3] Roland Allen. 2023. *The Notebook: A History of Thinking on Paper.* Profile Books, London, UK. 342 pages. https://profilebooks.com/work/the-notebook/

[4] Mark W. R. Anderson and David Millard. 2022. Hypertext's Meta-History: Documenting in-Conference Citations, Authors and Keyword Data, 1987-2021. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media* (Rome, Italy). Association for Computing Machinery, New York, NY, USA, 96–106. https://doi.org/10.1145/3511095.3531271

[5] Mark W. R. Anderson and David E. Millard. 2023. Seven Hypertexts. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media* (Barcelona, Spain). ACM, New York, NY, USA, 42: 1–15. https://doi.org/10.1145/3603163.3609048

[6] Bill Atkinson. 2018. *Triangulation 247: Bill Atkinson, part 2.* YouTube. Retrieved May 4, 2024 from https://youtu.be/INdByDjhClU?t=1341

[7] Claus Atzenbeck, Eelco Herder, and Daniel Roßner. 2023. Breaking The Routine: Spatial Hypertext Concepts for Active Decision Making in Recommender Systems. *New Review of Hypermedia and Multimedia* 29 (2023), 1–35. https://doi.org/10.1080/13614568.2023.2170474

[8] Claus Atzenbeck, Peter Nürnberg, and Daniel Roßner. 2021. Synthesising augmentation and automation. *New Review of Hypermedia and Multimedia* 27, 1–2 (2021), 177–203. https://doi.org/10.1080/13614568.2021.1942237

[9] Claus Atzenbeck, Daniel Roßner, and Manolis Tzagarakis. 2018. Mother: An Integrated Approach to Hypertext Domains. In *Proceedings of the 29th on Hypertext and Social Media* (Baltimore, MD, USA). ACM, New York, NY, USA, 145–149. https://doi.org/10.1145/3209542.3209570

[10] Lisa Banks. 2011. *Hypertext creator says structure of World Wide Web 'completely wrong'.* CIO Magazine. Retrieved May 4, 2024 from https://web.archive.org/web/20121103115842/http://www.cio.com.au/article/383395/hypertext_creator_says_structure_world_wide_web_completely_wrong_/

[11] Naomi S. Baron. 2015. *Words Onscreen.* Oxford University Press (USA), New York, NY, USA. 304 pages. https://global.oup.com/academic/product/words-onscreen-9780199315765

[12] Roland Barthes and S. Heath. 1977. Death of The Author. In *Image, Music, Text.* Fontana Press, London, UK, 142–148.

[13] Marijan Beg, Juliette Taka, Thomas Kluyver, Alexander Konovalov, Min Ragan-Kelley, Nicolas M. Thiery, and Hans Fangohr. 2021. Using Jupyter for Reproducible Scientific Workflows. *Computing in Science & Engineering* 23, 2 (2021). https://doi.org/10.1109/mcse.2021.3052101

[14] Mark Bernstein. 1998. Patterns of Hypertext. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia : Links, Objects, Time and Space—Structure in Hypermedia Systems* (Pittsburgh, PA, USA). ACM, New York, NY, USA, 21–29. https://doi.org/10.1145/276627.276630

[15] Mark Bernstein. 2002. Storyspace 1. In *Proceedings of the Thirteenth ACM Conference on Hypertext and Hypermedia* (College Park, MD, USA). ACM, New York, NY, USA, 172–181. https://doi.org/10.1145/513338.513383

[16] Mark Bernstein. 2002. *Tinderbox.* Eastgate Systems. Retrieved May 4, 2024 from http://www.eastgate.com/Tinderbox/

[17] Mark Bernstein. 2009. Shadows in The Cave: Hypertext Transformations. *Journal of Digital Information (JoDI)* 10, 3 (2009), 1–8. *JoDI* was archived without images, so see the Internet Archive.

[18] Mark Bernstein. 2011. Can We Talk About Spatial Hypertext?. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia* (Eindhoven, NL). ACM, New York, NY, USA, 103–112. https://doi.org/10.1145/1995966.1995983

[19] Ann M. Blair. 2010. *Too Much to Know.* Yale University Press, New Haven, CT, US. 416 pages. https://yalebooks.yale.edu/book/9780300165395/too-much-to-know/

[20] Keith Blount. 2007. *Scrivener.* Litterature & Latte Ltd. Retrieved May 4, 2024 from https://www.literatureandlatte.com/scrivener/overview

[21] Jay David Bolter. 2001. *Writing Space.* Vol. 2nd ed. Routledge, New York, NY, USA. 246 pages. https://www.routledge.com/Writing-Space-Computers-Hypertext-and-the-Remediation-of-Print-2nd/Bolter/p/book/9780805829198

[22] Jay David Bolter and Richard A. Grusin. 2000. *Remediation: Understanding New Media.* MIT Press, Cambridge, MA, USA. 295 pages. https://mitpress.mit.edu/9780262522793/remediation/

[23] Jay David Bolter and Michael Joyce. 1987. Hypertext and Creative Writing. In *Proceedings of the ACM Conference on Hypertext* (Chapel Hill, NC, USA). ACM, New York, NY, USA, 41–50. https://doi.org/10.1145/317426.317431

[24] Tom Bradshaw. 2004. *Reading At Risk: A Survey of Literary Reading in America.* Technical Report. National Endowment for the Arts. https://www.arts.gov/impact/research/publications/reading-risk-survey-literary-reading-america

[25] Richard Brath. 2020. *Visualizing with Text.* CRC Press, Boca Raton, Florida, USA. 268 pages. https://www.routledge.com/Visualizing-with-Text/Brath/p/book/9780367259266

[26] Vannevar Bush. 1945. As We May Think. *The Atlantic Monthly* 176, 1 (1945), 101–108. https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

[27] Jeff Conklin and Michael L. Begeman. 1988. gIBIS: A Hypertext Tool for Exploratory Policy Discussion. *ACM Transactions on Information Systems (TOIS)* 6, 4 (1988), 303–331. https://doi.org/10.1145/58566.59297

[28] Data.gov. 2009. *data.gov.* US Government. Retrieved May 4, 2024 from https://data.gov

[29] Steven J. DeRose and David G. Durand. 1995. The TEI Hypertext Guidelines. *Computers and the Humanities* 29, 3 (1995), 181–190. https://doi.org/10.1007/BF01830615

[30] DFG. 1951. *Deutsche Forschungsgemeinschaft.* Deutsche Forschungsgemeinschaft. Retrieved May 4, 2024 from https://www.dfg.de/en

[31] Dennis Duncan. 2021. *Index, A History of the.* Penguin UK, London, UK. 352 pages.

[32] Douglas Carl Engelbart. 1962. *About Viewspecs.* Stanford Research Institute. Retrieved May 4, 2024 from https://dougengelbart.org/content/view/218/ No known date, but date estimated on concept appearing in NLS before the 1968 demo.

[33] Douglas Carl Engelbart. 1984. Authorship Provisions in AUGMENT. In *Proceedings of the 1984 COMPCON Conference* (San Francisco, CA, USA). IEEE Computer Society, New York, NY, USA, 465–472.

[34] Ayelet Even-Ezra. 2021. *Lines of Thought.* University of Chicago Press, Chicago, IL, USA. 272 pages.

[35] Judith Flanders. 2020. *A Place For Everything.* Pan Macmillan, London, UK. 272 pages.

[36] Andrew M. Fountain, Wendy Hall, Ian Heath, and Hugh C. Davis. 1990. MICROCOSM: An Open Model for Hypermedia with Dynamic Linking. In *Proceedings of the First European Conference on Hypertext* (Versailles, France). Cambridge University Press, Cambridge, UK, 298–311.

[37] Future Text Lab. 2-24. *Future Text Lab.* Future of Text. Retrieved May 4, 2024 from https://futuretextlab.info

[38] Danny Goodman. 1988. *The Complete Hypercard Handbook.* Bantam Books, New York, NY, USA. 912 pages.

[39] Dene Grigar, Frode Hegland, Andrew Thompson, and Mark W. R. Anderson. 2024. Demonstration of The Future of Text in XR: Year 1. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media* (Poznan, Poland). ACM, New York, NY, USA. Submitted paper.

[40] Kaj Grønbæk. 1999. *From Web to Workplace: Designing Open Hypermedia Systems.* MIT Press, Cambridge, MA, USA. 424 pages. https://mitpress.mit.edu/9780262071918/from-web-to-workplace

[41] Frank G. Halasz, Thomas P. Moran, and Randall H. Trigg. 1986. NoteCards in a Nutshell. *ACM SIGCHI Bulletin* 17, SI (1986), 45–52. https://doi.org/10.1145/29933.30859

[42] Wendy Hall, Hugh C. Davis, and Gerard Hutchings. 1996. *Rethinking Hypermedia: The Microcosm Approach.* Kluwer Academic Publishers, Dordrecht, NL. https://doi.org/10.1007/978-1-4613-1335-9

[43] N. Katherine Hayles. 2010. How We Read: Close, Hyper, Machine. *ADE Bulletin* 150, 2010 (2010), 62—79. https://doi.org/10.1632/ade.150.62

[44] R. Bryan Haynes, Cynthia D. Mulrow, Edward J. Huth, Douglas G. Altman, and Martin J. Gardner. 1990. More Informative Abstracts Revisited. *Annals of Internal Medicine* 113, 1 (1990), 69–76. https://doi.org/10.7326/0003-4819-113-1-69

[45] Keith Houston. 2016. *The Book: A Cover-to-Cover Exploration of the Most Powerful Object of Our Time.* W. W. Norton & Company, New York, NY, USA. 448 pages. https://wwnorton.co.uk/books/9780393244793-the-book

[46] Michael Joyce. 1995. Siren Shapes. In *Of Two Minds: Hypertext Pedagogy And Poetics.* The University of Michigan Press, Ann Arbor, MI, USA, 38–59. https://doi.org/10.3998/mpub.10599

[47] Eugene Eric Kim. 2001. *An Introduction to Purple.* eekim.org. Retrieved May 4, 2024 from https://eekim.com/software/purple/purple.html

[48] Joseph C. R. Licklider. 1965. *Libraries of the Future.* MIT Press, Cambridge, MA, USA. 219 pages. https://mitpress.mit.edu/9780262120166/libraries-of-the-future/

[49] Niklas Luhmann. 2015. *Der Zettelkasten Niklas Luhmanns.* Niklas Luhmann-Archiv, Universität Bielefeld. Retrieved May 4, 2024 from https://niklas-luhmann-archiv.de/nachlass/zettelkasten

[50] Catherine C. Marshall. 2009. *Reading and Writing the Electronic Book.* Morgan & Claypool Publishers, San Rafael, CA, USA. 200 pages.

[51] Catherine C. Marshall. 2009. Reading Spatial Hypertext. In *Reading Hypertext*, Mark Bernstein and Diane Grecco (Eds.). Eastgate Systems, Inc., Watertown, MA, 211–220.

[52] Catherine C. Marshall, Frank G. Halasz, Russell A. Rogers, and William C. Janssen, Jr. 1991. Aquanet: A Hypertext Tool to Hold Your Knowledge in Place. In *Proceedings of the Third Annual ACM Conference on Hypertext* (San Antonio, TX, USA). ACM, New York, NY, USA, 261–275. https://doi.org/10.1145/122974.123000

[53] Catherine C. Marshall, Frank M. Shipman, III, and James H. Coombs. 1994. VIKI: Spatial Hypertext Supporting Emergent Structure. In *Proceedings of the 1994 ACM European Conference on Hypermedia technology* (Edinburgh, Scotland). ACM, New York, NY, USA, 13–23. https://doi.org/10.1145/192757.192759

[54] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2012. Citeology: Visualizing Paper Genealogy. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (Austin, Texas, USA). ACM, New York, NY, USA, 181–190. https://doi.org/10.1145/2212776.2212796

[55] Hermann A. Maurer. 1996. *The Hyper-G Now Hyperwave: The Next Generation Web Solution.* Addison-Wesley Longman, Harlow, UK. 672 pages.

[56] Marshall McLuhan and Quentin Fiore. 1967. *The Medium is The Massage.* Bantam Books, New York, NY, USA. 166 pages.

[57] MobileRead Wiki. 2024. *Page Numbering.* MobileRead Wiki. Retrieved May 4, 2024 from https://wiki.mobileread.com/wiki/Page_numbers

[58] Iian Neill and Andreas Kuczera. 2019. The Codex – an Atlas of Relations. *Zeitschrift für digitale Geisteswissenschaften* Sonderbände, 4 (2019). https://doi.org/10.17175/sb004_008

[59] Iian Neill and Desmond Allan Schmidt. 2021. SPEEDy. A Practical Editor for texts Annotated With Standoff Properties. In *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*, Elena Spadini, Francesca Tomasi, and Georg Vogeler (Eds.). Books on Demand, Norderstedt, DE, 45–55.

[60] Theodor Holm Nelson. 1974. *Computer Lib/Dream Machines.* Vol. 1978 Reprint of 1974 Edition. Self-published, n/a. 336 pages. https://computerlibbook.com

[61] Theodor Holm Nelson. 1987. *Computer Lib/Dream Machines.* Vol. 1987 Edition. Tempus Books of Microsoft Press, Redmond, WA, USA. 336 pages.

[62] US National Information Standards Organization. 2021. *ANSI/NISO Z39.96-2021, JATS: Journal Article Tag Suite, version 1.3.* US National Information Standards Organization (NISO). https://doi.org/10.3789/ansi.niso.z39.96-2021

[63] Oxford English Dictionary (online). 2023. *document, n. sense: 4.a.* Oxford University Press. Retrieved May 4, 2024 from https://doi.org/10.1093/OED/9116849622

[64] Oxford English Dictionary (online). 2023. *document, n. sense: additional sense.* Oxford University Press. Retrieved May 4, 2024 from https://doi.org/10.1093/OED/9116849622

[65] Oxford English Dictionary (online). 2024. *file, (n.2), sense I.4.b.* Oxford University Press. Retrieved May 4, 2024 from https://doi.org/10.1093/OED/4036492005

[66] Oxford English Dictionary (online). 2024. *paper, (n.), sense II.11.b.* Oxford University Press. Retrieved May 4, 2024 from https://doi.org/10.1093/OED/1184833433

[67] Jennifer Pearson, George Buchanan, and Harold Thimbleby. 2013. *Designing for Digital Reading.* Morgan & Claypool Publishers, San Rafael, CA, USA. 135 pages.

[68] PSDI. 2022. *Physical Sciences Data Infrastructure.* PSDI. Retrieved January, 10 2024 from https://www.psdi.ac.uk

[69] Quarto core team. 2022. *Quarto : An open-source scientific and technical publishing system.* quarto.org. Retrieved January, 10 2024 from https://quarto.org

[70] Glenn C. Reid. 1990. *Thinking in PostScript.* Addison Wesley Publishing Company, Reading, MA, USA. 221 pages.

[71] Daniel Rosenberg and Anthony Grafton. 2010. *Cartographies of Time.* Princeton Architectural Press, New York, NY, USA. 272 pages.

[72] Paul Henry Saenger. 2010. The Impact of the Early Printed Page on the History of Reading. In *The History of the Book in the West: 1455–1700, Vol.II*, Ian Gadd (Ed.). Routledge, London UK, 385–450.

[73] Frank M. Shipman, III, Haowei Hsieh, Preetam Maloor, and J. Michael Moore. 2001. The Visual Knowledge Builder: A Second Generation Spatial Hypertext. In *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia* (Århus, Denmark). ACM, New York, NY, USA, 113–122. https://doi.org/10.1145/504216.504245

[74] Charles Percy Snow. 2012. *The Two Cultures.* Cambridge University Press, Cambridge, UK. 194 pages. https://www.cambridge.org/core/books/two-cultures/506914BF3ADDD439EE8100FD2882D9DB

[75] Norbert A. Streitz, Jörg Haake, Jörg Hannemann, Andreas Lemke, Wolfgang Schuler, Helge Schütt, and Manfred Thüring. 1992. SEPIA: A Cooperative Hypermedia Authoring Environment. In *Proceedings of the ACM Conference on Hypertext* (Milan, Italy). ACM, New York, NY, USA, 11–22. https://doi.org/10.1145/168466.168479

[76] Text Encoding Initiative. 2024. *Text Encoding Initiative History.* Text Encoding Initiative. Retrieved January, 10 2024 from https://tei-c.org/about/history/

[77] The Linnean Society. 2023. *The Linnean Collections.* The Linnean Society. Retrieved May 4, 2024 from https://linnean-online.org/information.html

[78] Randall H. Trigg. 1983. *A Network-Based Approach to Text Handling for the Online Scientific Community.* Ph. D. Dissertation. University of Maryland, College Park, MD, USA. https://www.proquest.com/dissertations-theses/network-based-approach-text-handling-on-line/docview/303173380/se-2

[79] Edward R. Tufte. 1982. *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, CT, USA. 197 pages. https://www.edwardtufte.com/tufte/books_vdqi

[80] US National Institutes of Health. 2003. *Journal Article Tag Suite (JATS).* US National Institutes of Health (NIH). Retrieved May 4, 2024 from https://jats.nlm.nih.gov/faq.html

[81] John Warnock. 1991. *The Camelot Project.* Technical Report. Adobe Inc. https://pdfa.org/resource/the-camelot-project/

[82] Mark D. Wilkinson and 52 other authors. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (2016), 160018. https://doi.org/10.1038/sdata.2016.18

[83] Chuck Zerby. 2003. *The Devil's Details: A History of Footnotes.* Touchstone, New York, NY, USA. 162 pages. https://www.simonandschuster.com/books/The-Devils-Details/Chuck-Zerby/9780743241755