Hypertext's meta-history: Documenting in-conference citations, authors and keyword data, 1987-2021

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

David E. Millard
Web and Internet Science Group (WAIS), University of Southampton
Southampton, Hants, United Kingdom
dem@soton.ac.uk

ABSTRACT

Conferences such as ACM Hypertext have been running for many decades and the metadata on their collected publications represent a valuable scholarly meta-history on areas such as the community's health, diversity, and changing interests. But the metadata about these papers is not readily available for analysis, and the data collection and cleaning tasks appear substantial. In this paper we attempt to explore this challenge using the ACM Hypertext series as a case study. Taking the ACM Digital Library as a starting point, and using a combination of manual and automatic methods, we have constructed and released a 3-star Open Dataset representing over 1000 publications by almost 2,500 authors. An initial analysis reveals a modestly-sized but robust conference, with a changing pattern of in-citations that co-occurs with the arrival of social media, and a relatively consistent but imbalanced gender ratio of authors that shows some signs of recent improvements. The challenges encountered included identifying discrete author names, potential issues with text retrieval from PDF, and a disparate set of author keywords that reveals an absence of a common vocabulary. These insights are the results of a hard-fought process that is made complex by an incomplete digital record and a lack of consistency in naming. This Hypertext case study thus reveals a serious shortfall in the way that scholarly activity is captured and described, and questions PDF as the primary method of recording publications. Addressing these issues would make further analysis more straightforward and would allow larger events (with orders of magnitude more data) to be analysed in a similar way.

CCS CONCEPTS

• **General and reference** → **Reference works**; • **Social and professional topics** → *Gender*; History of computing; History of software; *Information science education*; • **Software and its engineering** → Maintaining software; Documentation; Software evolution.

KEYWORDS

hypertext, knowledge management, metadata, citation networks, links, linkbases, gender, Tinderbox, dataset, keywords, keywording, visualisation, analysis

ACM Reference Format:
Mark W. R. Anderson and David E. 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 (HT '22), June 28-July 1, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3511095.3531271

1 INTRODUCTION

It is the nature of technical conferences to look forward but in doing so it is easy to forget the past and, as a consequence, to lose the sense of research context that a rich history can provide. The meta-history of a conference (made up of the metadata around events, publications, people, themes, and topics) is one way to analyse and access this context. To varying degrees this data is digitally available in the online records of the conferences themselves, and thus through the collected set of publications and their metadata. But in practice, if the only digital offering other than standard bibliometrics is PDF documents, further information can be hard to extract, as the format is a wrapper for varying quality of content. For pre-Web and early Web era the situation is more challenging, as documents may simply be be scans of variable quality print and OCR will be of the lower quality of the time.

In this paper we investigate the challenges of reconstructing a conference's history from these digital resources, and explore some of the affordances and opportunities of that data. We use the ACM Hypertext Conference ('HT') as a case study, both for its relevance to the technologies at hand, and also because the size of its digital footprint is both small enough to be manageable as a test case, but also large and complex enough that the principles and challenges would scale to other events.

As the first part of this activity in 2021 we released an 33-year HT dataset[1] for public re-use as 3-star Open Data [21] with a CC BY-NC-SA 4.0 licence. Here, we present some initial analysis of what the data can tell us about the community and its changing character and behaviour. This type of approach has been applied in the context of Web Science, where the use of co-citation analysis, bibliographic coupling, natural language processing, topic modelling and network visualisation have led to a number of insights [27]. But Hypertext is older, so we also take the opportunity to reflect on the activity of deep diving this type of data, exploring the challenges of finding,

*Corresponding author & responsible for creation of primary dataset.

HT '22, June 28-July 1, 2022, Barcelona, Spain

ACM ISBN 978-1-4503-9233-4/22/06…$15.00
https://doi.org/10.1145/3511095.3531271

[1] See. http://dx.doi.org/10.5258/SOTON/D1870. The plain-text dataset tables cover HT 1987–2021 including the three ECHT Conferences.

cleaning, and annotating data that pre-dates the Web. Challenges that the Hypertext community is itself uniquely qualified to address.

2 BACKGROUND

HT is one of the longest-running events in the ACM family[2]. First held in 1987, the HT conference also predates the first of the Dexter workshops on hypertext in 1988 [20], but nonetheless came some twenty years after the first hypertext systems, NLS [12] and HES [34], were presented to the world. Separately, a European Hypertext conference ('ECHT') first started in 1990 [39], but has precedent workshops from 1988 [31] and 1989 [32].

Initially, both the US-based HT and European-based ECHT conferences were biennial, run in non-clashing years. From 1991,organisation of the conference came under the ACM Special Interest Group 'SIGLINK' (renamed 'SIGWEB' in 1998). After ECHT conferences in 1992 and 1994, from 1996 until today the conference has used the same short title HT, with locations alternating annually between the North America and other (mainly European) locations. Only two years—1988 and 1995–have seen no Hypertext conference. In early conferences there was a close link with the literary hypertext community, though the style and cost of computer conferences may have contributed to a weakening of that link over subsequent years[3]. HT is also unusual in that—for a while—it sought and accepted papers in hypertext form[4].

A number of other conferences also have their roots within HT, beginning life as workshops or conference themes. These include: The Semantic Web Conference ('SemWeb'); the workshop on Dynamic and Adaptive Hypertext ('DAH') which evolved into ACM's User Modeling, Adaptation, and Personalization ('UMAP'); and the Web Conference itself ('WWW')—the World Wide Web was first shown as a demo at Hypertext'91 in San Antonio, Texas (US). Other events that have grown independently also have an overlap with Hypertext, such as the Electronic Literature Organization ('ELO') conference, the International Conference on Interactive Digital Storytelling ('ICIDS') and the Wiki Symposium ('WikiSym'). This diaspora reflects the rapid evolution of the Internet and of the Web.

As a long-running event Hypertext has built up a sizeable presence in the ACM Digital Library ('DL.ACM') despite its relatively small size, with a docuverse of over 1,000 publications and almost 2,500 authors represented. What insight might an analysis of such records provide? In this paper we explore the possibility and the potential of a deep dive into ACM Hypertext docuverse. We hope to highlight the potential of this type of research community data, and to show the challenges of assembling and cleaning the datasets.

Our approach is inspired by work on Open Data [1] and efforts in Scholarly Publishing to look beyond conventional bibliometrics (activities like citations and downloads) to 'altmetrics' and 'scientometrics' [25, 28, 37, 38] [5] which suggest evaluating impact from a broader range of sources such as social media. Whilst altmetrics' original aim was to extend traditional bibliometric data by using a wider, post-print era, range of sources, inter-disciplinary approaches like Web Science [2, 24] seek to extend digital insight even further beyond pure bibliometric concerns. Such enriched data aids Digital Humanities studies: by creating an open dataset of a conference's history we want to explore an alternative, wider, notion of metrics concerned with the properties of the community itself, such as its health, interests, and diversity.

3 METHODOLOGY

3.1 The Dataset

The genesis of this dataset was an attempt to generate information to support exploration of in-conference citations, drawing upon HT's Proceedings in the DL.ACM. An exploratory survey of data was collected in 2016–17 and gradually expanded to its current form in 2018. since then the datasets has been updated to include HT conference papers up to HT'21. To address gaps in the record, some further information has has been obtained from other sources including conference websites—when they can be found. The earliest surviving HT conference website is that for HT'96[6].

Mindful of premature formalisation issues [40] and the uncertain nature and incompleteness of the records, the data was assembled and housed in the note-taking program *Tinderbox* [11] which is supportive of incremental formalism, and has strong, flexible, linking tools. It also has very configurable export allowing all data to be exported to a larger database system when the emergent HT metadata structure stabilises; XML-based data files also allow direct data access. It was also used to generate the dataset (see Section 5) provided with this paper. By interesting chance, *Tinderbox*'s design draws from the earlier hypertext system *Storyspace* [5] which debuted at the first HT conference in 1987.

Ironically, given the subject area is Hypertext, this dataset has been harder to assemble than envisaged due to errors and omissions in its data, and to the DL.ACM's website design. Data collection straddled both the old DL.ACM website—all `<table>`-based with no semantic labels, and the new—complex, deeply nested `<div>`-tags, but still lacking useful semantic labelling. Additionally, the DL.ACM has no API and prohibits screen-scraping forcing the researcher to use only the rendered page content[7]. Sadly, the new website design's many pop-up/roll-down decorative and marketing elements often obscure the page data about actual papers, hampering effective research using the website.

Once the structure of the data emerged it proved fruitful to look outside the DL.ACM for missing or additional information, conference cover artwork has been found for all ACM conferences except for 2001 and 2006. Conference websites have also been traced for 1996 onward, although most earlier sites are only in the Internet Archive and in some cases partially complete. In 2020, further research re-discovered over 700 further conference-related items, not published in the formal proceedings including: workshops, keynotes, tutorials, posters, and demos.

Initially, data collection was very time-consuming with automation being limited (by DL.ACM's T&Cs) to copy and paste into

[2]The oldest ACM Event is the ACM Southeast Regional Conference that began in 1967.
[3]More recent conferences have made efforts to reconnect with that part of the community (for example, through creative exhibitions).
[4]Sadly, such items were not archived as hypertext so are unavailable to later researchers.
[5]Also see *Beyond Bibliometrics*[8], Chapter 14. and *Bibliometrics and Research Evaluation* [14].

[6]Preserved via the Internet Archive: https://web.archive.org/web/19981202194738/http://info.acm.org/siglink/ht96/homepage.html.
[7]Alternative resources like rhe 'dblp' database proved to repeat errors in DL.ACM so were less useful than presumed.

Type	#	Type	#	Type	#
Full paper	701	Technical Briefing	20	Addendum	2
Short paper	377	Workshop	11	Case Study	1
Poster	169	Video	9	Exhibit	1
Panel	52	Late Abstract	5	Extended Abstract	1
Keynote	51	Doctoral Consortium Abstract	4	Invited Talk	1
Demo	38	Cultural Briefing	3	Tutorial	1

Table 1: Occurrence of each published type, 1987–2021 (1,447 items). ('Plenary' mapped to 'Keynote')

Tinderbox after which some automation could then be applied to leverage emergent patterns. But always-changing volunteer editorial teams meant conferences used inconsistent terminology and classification. Initial collection was a multi-pass process, both to error check manual re-keying and back-filling existing data as inconsistencies were revealed. A tool supporting incremental formalisation proved a good choice. As a broad indication of the time taken simply to generate the underlying dataset, assuming a 5 minute activity to record an item or conference, with a further 10 minutes for corrections and to identify and link citations, the dataset represents something of the order of 1,942 x 15min, i.e. 370 hours work or 46.25 8-hour workdays spread over a four year period. That estimate excludes all the later work on author name triage, gender metrics, keywords use and plain text quality. Author identification started in 2018, followed in 2019–20 by analysis of keywords, author gender, and plain text quality, all tasks taking further significant time to de-duplicate data and remove source errors. DL.ACM data was cross-checked with other data sources for validation, most notably 'dblp', HCI Bibliography, and Google Scholar[8].

3.1.1 Conferences. The current dataset comprises information from 33 conferences: ACM HT 1987–2021, including ECHT 1990[9] and ACM ECHT 1992 and 1994. Conferences since HT'17 have been added incrementally as the dataset became established. The primary facts stored are the conference's full name (and title, if any), its location, date, the website's URL, and the DL.ACM's DOI URL for the Proceedings. The per-conference notes also acted as a modular container for records of the conference's items and consolidating data from its papers.

3.1.2 Conference Papers. Each discrete entry in the DL.ACM was recorded as a *Tinderbox* note . In this analysis such a record is called an 'item'. Within this set additional data is collected for 'full' and 'short' paper items, this sub-set being referred to as 'papers'. From the 33 sets of Proceedings, per-item data was captured for 1,447 discrete items. The full range of item types—as described in the proceedings—are shown in Table 1. Of note is that early conferences often excluded keynotes as items so c.25% of these are not recorded in Proceedings (but note Section 3.1 above). Of these items, 1,079 were also identified as discrete 'Full' or 'Short' papers. For all items, counts were noted of both references used and of references which cite within the HT conference. Key facts are the item's title, (discrete) authors, item type, DOI/URL, and author keywords. Abstracts were stored, if available.

3.1.3 In-Conference Citations. Once all items were recorded, the references were reviewed in the DL.ACM for in-conference citation, validating them against the item PDFs if web page data was ambiguous (often so for early PDFs). These references were captured as links between per-item notes in the *Tinderbox* document, to give a potential viewable or exportable graph, reflecting the original aim of the dataset.

3.1.4 Author Triage. From the published item data, a note was generated for each unique author name and linked back to its associated item note(s). The overall list of published names gave an overall count of 2,682 discrete entries (to HT'21). All author notes link to their published items. Authors were then normalised to remove name variants, e.g. initials-only vs. full first name(s), typos, accents, name alternate shortenings, change of family name (e.g. marriage), change of gender, etc. All name variants of a single author were nested under the note of a canonical per-author record (the longest form of the name). This triage resulted in a count of 2,488 unique authors within the dataset.

3.1.5 Keyword Triage. As with each author, recorded per-item keywords were extracted into a list of 3,141 unique (lower-cased) per-term notes. These notes were then linked back to the per-item notes where they occur. Further triage, retaining provenance aggregated (for number, case and noun/adjective) to give 2,703 discrete terms.

3.2 Issues Encountered Creating The Dataset

3.2.1 Omissions from the DL.ACM. Possibly due to print costs, early Proceedings omitted some keynotes and other activities like workshops, tutorials, readings, exhibitions, etc. These occurred as part of the conference but were never recorded in proceedings. However, where they exist, conference websites may hold data. The subject of such conference events is important in establishing the contemporary state of the hypertext field. Occasionally, a single item may list all posters and demos with limited data and no abstracts or PDFs; in 2001 one such item was found to describe 30 otherwise discrete items. Alternate sources, revealed partial or full (e.g. paper/transcript) data of 708 'lost' items, many in turn describing many discrete items. Combined with the 1,447 published items this represents a loss to future research of almost *one third* of conference activity; this loss should be of concern for ongoing academic research.

[8]See, respectively, https://dblp.org/faq/index.html, https://hcibib.org, and https://scholar.google.co.uk.

[9]Though not strictly part of ACM HT, it falls in the overall sequence of Hypertext conferences. Data for ECHT'90 was taken from its Proceedings [39]. The proceedings are out-of-print and only available second-hand.

3.2.2 Incorrect PDF-derived data. PDFs for many early items yield poor quality OCR-ed text in some DL.ACM records (common with less common characters, e.g. if accented). Even cleaned, some items were insufficiently detailed to give an unambiguous reference. Thus a large number of item reference lists needed significant per-item effort before usable as a reliable record, including tracing possible (in-conference) references, which by intent DL.ACM should link; ligatures and accented characters appear problematic. PDFs are thus questionable as *reliable* machine readable resources.

3.2.3 Author family names, and gender. English-centric assumptions have resulted in family names being incorrect or partial. Authors can have multiple non-linked DL.ACM profiles. Having a full list of discrete authors of published items opened the possibility of assessing the degree of participation of women in the conference. However, this is a non-trivial task. Not only is there no formal gender recorded for authors, but some authors may have no preferred expressed gender, or a non-binary/gender-fluid identity; others may have changed gender identity since publishing. Nevertheless, with under-representation of women seen as a perennial problem in the computer sciences [4, 6, 7] we judged that it was a topic that was worth addressing. Our solution was to manually record an assumed gender (male/female/other/unknown) to each of the 2,488 individual authors in order to generate approximate overall data, but to exclude these annotations from the public dataset in order to avoid misgendering individuals.

International colleagues in the Southampton WAIS group helped with determination of Chinese, Indian and South American names. In addition, web search was used to find gendered references (e.g. in profiles or home pages) or to find official images of the author (and which had clear gender indication). Where any doubt remained an 'Unknown' value was recorded. The latter group mainly comprises East Asian names (China, Korea, Japan) and a few early initial-only first names. This work was an extremely time-consuming task but it was felt its sensitive nature warranted additional care so as to allow us to investigate gender representation based on clearer data; the task was by far the most effort-intensive single aspect of data collection[10]. It would be a useful approach for further work.

3.3 Plain text extraction

A variety of methods were tested for extracting plain text, such as is needed for automated Machine Learning or other analysis, but these gave inconsistent results. With no method showing particular accuracy, the plain text extraction built into macOS Finder was used to make a UTF-8 'TXT' file for available item PDFs (1,443 of 1,447 items). Font ligatures—a default LaTeX usage—resulted in characters losses in the output plain text: e.g. 'office' giving 'oce'. OCR errors resulted in corrupted sequences, seemingly at random, with either run-together words (i.e. thewordseranotspaced) or with letter-spacing errors (i.e. a s p a c e b e t w e e n e v e r y l e t t e r). Article plain text was also polluted with boilerplate text, headers, page numbers, line-break hyphenation and other irrelevant structural data.

[10] An automated method considered was using genderize.io (https://genderize.io) as used by Holman *et al.* [26, pp.9–11] https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2004956#sec008 but data had already been manually compiled form multi-source search

4 INITIAL FINDINGS

Despite the issues discussed above and the gaps in some of the records, much valid data has been extracted manually and a subset has been published—see Section 5. Full data, including sensitive aspects, such as author gender, are withheld but these may be requested, with justifying need, from the authors. While a full analysis of the dataset is beyond the scope of this paper we wish to show some representative findings in regard to the number and in-citation count of papers, the gender balance of authors, and keyword use.

4.1 Papers and In-Conference Citations

One of the most straightforward views of the Hypertext dataset is to look at the number of published papers, and the extent to which they have built on previous conference papers (measured as in-conference citations). In this analysis we used published items that were categorised as full or short papers (therefore excluding other types of submissions: posters, demos, workshop items, etc.—see Table 1).

Of the 1,079 papers, 567 have been cited by other HT conference papers at least once (52.54%). 651 papers have cited other conference papers at least once (60.33%). 789 papers have either cited or been cited once in conference (73.12%), meaning that almost a quarter of papers in conference history have no explicit relationship with other conference items. Thus 428 items (39.66%) have never cited or been cited, ever. Furthermore, 512 (47.45%) of all papers have never been cited in conference and 428 (36.67%) of all papers have never cited another conference item.

Figure 1 shows the number of papers published in every Hypertext event from 1987–2021, as well as the numbers of in-conference citations (1987 has no prior works to cite). After large initial interest[11] with hundreds of attendees, the conference has always been relatively modest in size[12]. This is reflected in the data, where the number of papers grows from 26 in 1987 to a peak of 50 in 2004. However, interest in Hypertext waned dramatically in the following years, to a low point in 2006 of only 22 papers and around 50 participants.

In the late 2000s, following the Web 2.0 revolution and the growth of popularity in early social media sites, Hypertext became one of the earlier venues for social media research. So much so that Millard [33] described the 2008 conference as a 'Great Safari', dominated by a new breed of in-the-wild systems that a few years earlier Walker [43] had identified as 'Feral Hypertext'. Despite the wane in interest in classic hypertext systems, with the inclusion of social media research HT recovered its numbers, and in 2012 changed its full name from 'Hypertext and Hypermedia' to 'Hypertext and Social Media' to reflect this new focus.

While the number of papers published has recovered to a healthy 30+ each year, the number of in-conference citations has dropped significantly, a change that co-occurs with the rise of social media research and these observations around the changing nature of hypertext research. Figure 2 shows this data as a percentage, and

[11] Attendance figures are not in DL.ACM. It does give acceptance rates, but only for 2000–19 [sic].

[12] Nielsen notes '87 being 100% overfull and HT'89 having 650 attendees [35, 36].

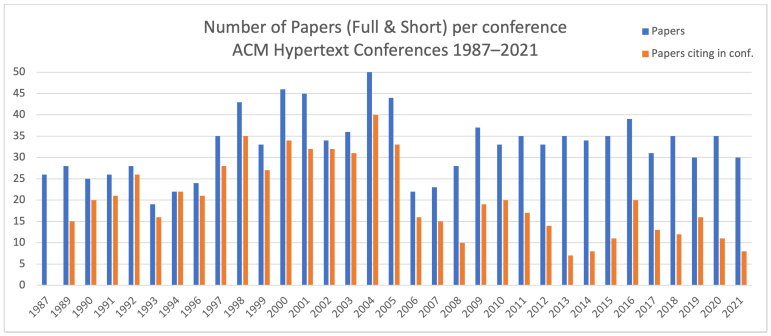

Figure 1: Number of Full & Short Papers per Conference (Hypertext Conferences 1987–2021)

the difference from 2008 onward is very clear. Manual inspection reveals that the majority of in-conference citations in this later period represent the classical hypertext-focused part of the conference community. On the one hand this reflects a healthy wider community for social media research—Hypertext is by no means the only, or even the most well known conference on social media—but it also reflects an aspect of direct scholarly conversation that has been lost, resulting in work that is more granular and less rooted than it might be.

The five paper most cited in-conference are:

- 58 cites. Bernstein 'Patterns of Hypertext', 1999 [3].
- 47 cites. Marshall, Shipman & Coombs, 'VIKI: Spatial Hypertext Supporting Emergent Structure', 1994 [30].
- 45 cites. Marshall *et al.*, 'Aquanet: A Hypertext Tool to Hold Your Knowledge in Place', 1991 [29].
- 37 cites. Zellweger, 'Scripted Documents: A Hypermedia Path Mechanism', 1989 [44].
- 37 cites. Streitz *et al.*, 'SEPIA: A Cooperative Hypermedia Authoring Environment', 1992 [41].

585 discrete authors have 5 or more published items, of which 26 have 10 or more items in the dataset. The author with the most papers in the dataset is Shipman with 27. The author with the most first-author papers is Bernstein with 17 and whom also has the largest overall item count of 34 published conference items.

An initial exploration of the health of a Conference is to look at whether authors of papers stay with a conference or publish once and move on. In Figure 3 we adapt[13] the 'author fluctuation' from Doerfel *et al.* [10, Fig. 2]. The top (grey) section is one-time only authors. Middle (blue) is previous authors who are still contributing albeit not necessarily every year of persistence; their 'life' is from first to last year of publication. Bottom (yellow) is new authors who will go on to contribute in subsequent years. The height of the bars is the annual count of discrete authors.

4.2 Gender Representation

As described in Section 3.2.3, we manually assigned an assumed gender to all 2,488 individuals in the dataset. While we are very aware of the sensitivities around gender identity (and cannot guarantee 100% accuracy) this assignment does allow us to paint a broad picture of gender diversity within the Conference, which tabulates as follows:

- Male: 1,881 (75.57%)
- Female: 524 (21.06%)
- Other (non-binary): 1 (0.04%)
- Unknown: 82 (3.29%), i.e. not unambiguously identifiable

Looking at this from the perspective of published *papers* across all conferences we observe:

- Paper includes a female author: 450 papers of 1,079 papers (41.70%)
- Paper has female first author: 222 of 1,079 papers (20.57%)
- The female author with most papers is Marshall: 21 (of which 14 as first author)

Figure 4 shows the percentage of papers with a female named first author, or which include a female named author. Despite some isolated spikes in the data, papers with female first authors as a percentage is relatively stable across the lifetime of the conference. However, the percentage of papers with female authors has risen slightly in the last 8 years (averaging 53.15% over this period, compared to 37.27% over the previous 25 years). Figure 5 shows this data alongside the same data for Male, Other and Unknown author genders. This rise runs counter to a perspective, often raised at Hypertext events, that there are fewer women participating than in the past (sometimes attributed to the loss of the literature and humanities parts of the community that were very active in the first ten years)[14]. Our analysis shows that such generalised assumptions are not true. But, rather than disproving the idea that representation is getting worse we should instead conclude that the situation regarding gender balance has never been good.

We also looked at gender across all published items. For the 21 conferences where the proceedings recorded keynotes, 15 of 46 (32.61%) were given by women[15]. Of the 45 panels in the record, 33 (73.33%) included women panelists, but only 6 of the total (13.33%) were chaired/introduced by women[16]. Hall is the female author with the most recorded items (27, including 16 papers), fourth highest

[13] 2021 data is omitted because until 2022 data is available continuing authors from 2021 cannot be assessed. As average 'life' of a continuing contributors is c.5 events (years), inherited authors from 2016 onwards *may* be an under-count as active authors may not have yet made their next contribution(s).

[14] Note too that not all contributing authors may *attend* the conference.

[15] Including keynote data recovered outside DL.ACM shows a slight increase: 28 of 75 (37.33%).

[16] Including panel data recovered outside DL.ACM shows a slight increase: panelists, 37 of 51 (72.55%); panel chairs 6 of 51 (17.65%).

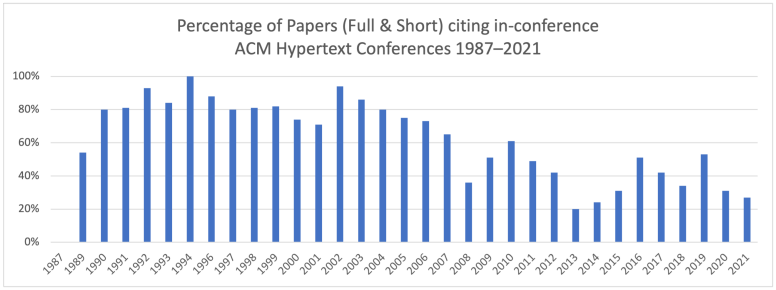

Figure 2: Percentage of Papers (Full & Short) citing in-conference (Hypertext Conferences 1987–2021)

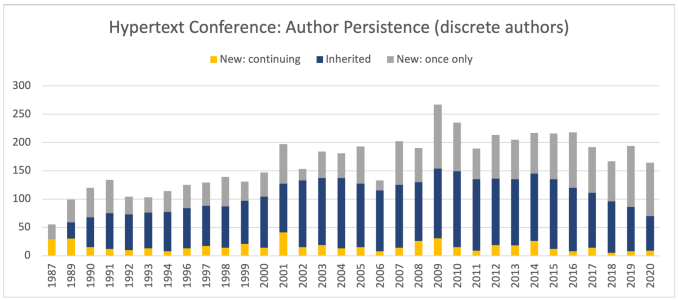

Figure 3: Hypertext Conference: Author Persistence (discrete authors), 1987–2020

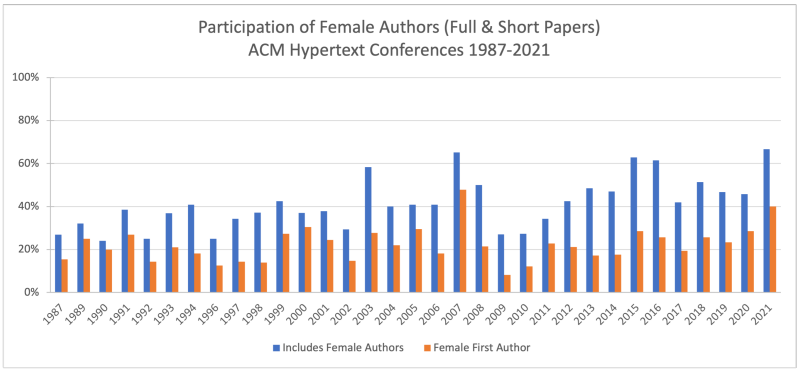

Figure 4: Participation of Female Authors, Full & Short Papers (ACM Hypertext Conferences 1987-2021)

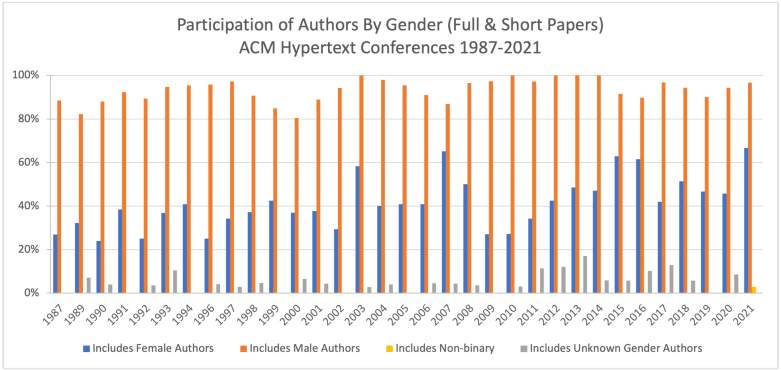

Figure 5: Participation of Authors By Gender, Full & Short Papers (ACM Hypertext Conferences 1987–2021)

Keyword	Years	#	Keyword	Years	#
Hypertext/Hypermedia	1989–2021	130	Information Retrieval	1990–2017	34
World Wide Web	1996–2021	66	Twitter	2010-2021	31
Open Hypermedia [system]	1993–2019	46	Social Networks	2006–2021	24
Spatial Hypertext	1993–2020	42	Adaptive Hypermedia	2000–2016	23
Navigation	1990–2015	35	Link	1990–2020	21
Social Media	2009–2021	35	Narrative	1998–2019	21

Table 2: All keywords with 20 or more occurrences, triaged 2,701 keyword listing

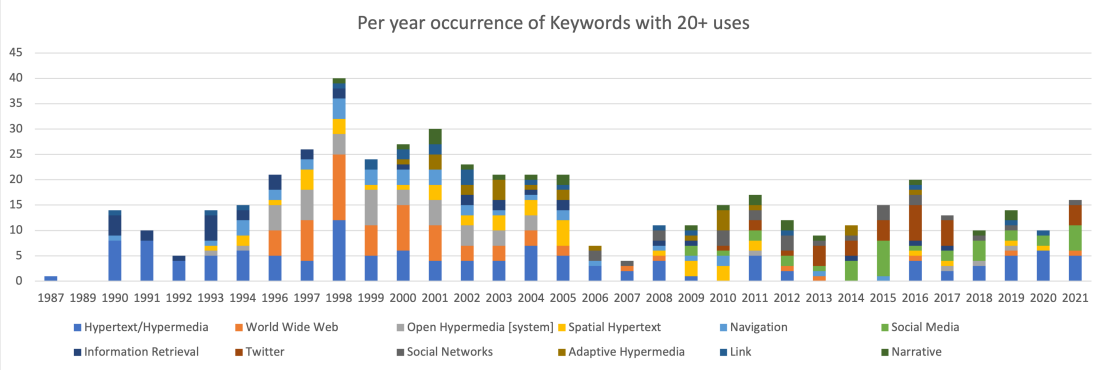

Figure 6: Per year occurrence of Keywords with 20+ uses (as per Table 2

in the whole dataset. These figures may be indicative of efforts by program chairs to create more balanced planned events, and shows that in some regards (especially panels) those efforts have been successful.

4.3 Author Keywords Over Time

Authors submitting to Hypertext assign keywords to their papers. This ought to be a rich source of information about the topics of interest to the community, and how these have changed over the years. Unfortunately the keywords in the dataset show little consistency, and provide only a limited picture. Across 1,079 short/long papers we found author's keywords as follows[17]:-

- Discrete keyword terms: 3,141. Triage for case/plural, etc., variations reduces discrete terms to 2,701
- Discrete uses of keywords in papers: 4,601
- Discrete keywords used 20 or more times: 13 (0.48%); highest use is *Hypertext* (88 times)—further merged with *Hypermedia* this rises to 130 times. See Table 2
- Only 569 of 2,701 (21.07%) aggregated keywords occur more than once
- 129 of papers (12.29%) have no author keywords.

Table 2 shows the 12 (aggregated) keywords occurred 20 or more times in HT items whose occurrence count is plotted by year in Figure 6. As might be expected the highest ranked keywords are broad generalisations: *Hypertext* and *Hypermedia*, *Spatial Hypertext*, *Hypermedia*, *Web*, etc. But it is possible to see how the emergence of some keywords match trends in the community. *Open Hypermedia Systems* were an early common topic in the late 1990s when the Open Hypermedia Systems Working Group was active [9], the *Semantic Web* [18] shows in the early 2000s, and of course *Social Networks*, *Social Media* and *Twitter* in particular are popular from 2006 onward. It is worth reflecting that these observations are made using less that 1% of all the keywords in the dataset. This limits their utility for organising the documents, and means that without some external knowledge base associating keywords together, mapping of topics over time is not viable.

In principle, a solution based on a rigid classification would solve this problem, e.g. the ACM Computing Classification System[19]. However, these are inflexible, and so difficult to evolve within the narrower scope of a single research community. Even where thematic terms emerge from author keywords, as has occurred with terms such as *Adaptive Hypermedia*, or *Narrative*, a term alone can hide semantic drift in use: early and more recent use of *Information Retrieval* show two periods of use with apparent differing meaning. Such insight is hidden without the analysed data retaining a link to source provenance: mere counts of total use hide much nuance. It would be preferential for a curated community vocabulary to foster a form of folksonomy[20] [15], and the fact that has not emerged is indicative of a lack of focus in the community (also reflected by those declining in-conference citations).

Compared to longer-lived subjects, social media and even hypertext are new and with no clear fixed taxonomy of keywords or, by proxy, recognised lines of theory or practice within the field

[17] Using all 1,447 published items yields a total of 3,167 discrete keywords.

[18] Semantic Web occurs 2001–2015. It is not in Table 2 as only 16 uses in total.

[19] ACM Computing Classification System: https://dl.acm.org/ccs.

[20] Term is credited to Thomas Vander Wal [42].

so most keywords are too broad in scope or overly narrow. This makes the keywords less useful for search and analysis. By adding, indeed enriching, author keywords with community terms post-publication would allow for more accurate identification of papers reflecting significant topics within the community. Being added *post hoc*, terms can also be adjusted or split should semantic drift occur in authors' use of keywords or in terminology within the community. In the moment, the author is not necessarily the best judge of what the paper reflects and for emergent topics, suitable terms my not yet exist. But, without them accurate search results are degraded. Search can only returns what it believes a match and so is critically dependent on appropriate target keyword terms being available.

4.4 Article Plain-text Recovery

Initial grep analysis indicates c.12% of the TXT files have run-together words (i.e. thewordseranotspaced) and c.1% have letter-spacing errors (i.e. a s p a c e b e t w e e n e v e r y l e t t e r). Error detection/correction tests using regular expressions showed a difficulty in fixing errors, such as run-on sequences original word boundaries not being unambiguous. In addition, the same technique proved inconsistent against lost ligature characters and line-break hyphenation as accurate detection requires understanding of both vocabulary and context. Thus the existing PDFs of record in DL.ACM served poorly as an accurate source of plain text for ML and other digital use without significant review and correction. Whilst the authors have had access to per-item PDFs as ACM members, the PDF articles themselves are not freely accessible to the public and so plain texts extracted from them can thus not be part of a public dataset. Original author-submitted source LaTeXand BibTeX files, which might help, are not accessible to researchers.

5 DISCUSSION

To give access to some of the data underpinning our initial findings an interim plain-text dataset is available [21] for re-use. This is a subset of the overall dataset and covers the conferences and conference items, including citation counts but not full author or keyword data (though these may be added in latter updates). For each item, 15 columns are supplied: record UID, published title, author abbreviated list, in-conference cites, in-conference cited by, number of references, parent conference/title/short title/year, DOI URL, full author names, first author, article type, abstract. A separate listing of UIDs records all in-conference links (using source/target UIDs). The Future of Text community [22] has already used the dataset in exploratory visualisations[22] of which an example[23] is shown at Figure 7, using *Tinderbox* with *Gephi*[24].

The visualization shows the sub-set of 789 papers cited, or citing, in-conference at least once. The plot organises the papers as nodes, coloured discretely per conference year, in a circle linking each cited and citing and with the node size reflecting the number of in-links; i.e. the largest notes are the most quoted papers. When a node is selected (the state illustrated), only nodes linked to the selection remain drawn and a side panel appears with information about the paper and linked papers and URLs for the DL.ACM source resource. All nodes remain active and the dataset can be explored by selecting a different node which then takes sidebar focus and the display of visible linked nodes is updated. The changing plot gives rapid feed back as to the degree, and spread in time of citing papers

5.1 Questions of Re-use

Part of the intent of this work is to see what can be done better in this task of interrogating (past) conference data. Although the work presented here draws only on ACM Hypertext, the choice of data and issues encountered offer a re-usable pattern for exploring creation of similar datasets for ACM Conferences and beyond. The 3 main sources—papers' references, authors and keywords—are publicly exposed in most conference or journal portals. Conferences such as ACM CHI or WWW would generate significantly larger amounts of data. Whilst tools like *Tinderbox* could still be valuable for initial investigation, where source data quality is still unclear. Strong support for incremental formalisation allows exploration of data structure before committing to a larger-scale database design. The resulting data, as shared with others, offers most benefit if presented in a form amenable to use with visualisations: i.e. aspects like node/edge tables, or as plain text tabular or JSON files.

In embracing such metadata collections, some questions arise:

- What are we not capturing that we ought to? Now that few read conference proceedings in print book form, are existing records too influenced by legacy print limitations (e.g. cost per printed page)? Is there a clear rubric for conference editors as to what needs to be recorded, other than accepted papers (and why)?
- What would improve extraction of data critical to creating a citation network? Practically, how usable—digitally—are the PDFs as the primary and *only* record, in an age of machine-based text analysis? Could techniques like 'Visual Meta' [23] help overcome the PDF format's textual inadequacies? Have the current ACM submission templates been reviewed with a mind to digital re-use of text? For example, does the DL.ACM's OCR configuration work well with current templates, e.g. to allow more rigorous avoidance of heading, page numbers, etc.?
- How might we improve the metadata/storage of metadata to aid analysis of the dataset? Could such measures allow for better (API?) quality querying of the corpus, or for supporting faceted browsing? By consideration of the metadata needs of such uses it may be possible to clarify additional (digital) data that can be captured as part of a conference or journal but which may be expensive or not possible to capture at a later date.

Creation of additional similar datasets for other conferences also opens the opportunity for cross-linking of data and even re-use of some existing information, such as author listings already reviewed for alternate names. In addition, data can also be visualised both for in-Conference analysis but also for exploration of wider patterns. It is important not to misread our findings as critique of the DL.ACM or the efforts of past conference chairs, or contributing authors.

[21] See. http://dx.doi.org/10.5258/SOTON/D1870.
[22] https://www.shoantel.com/proj/acm-ht/visualisations/index.html
[23] https://www.shoantel.com/proj/acm-ht/2020/index.html
[24] https://gephi.org

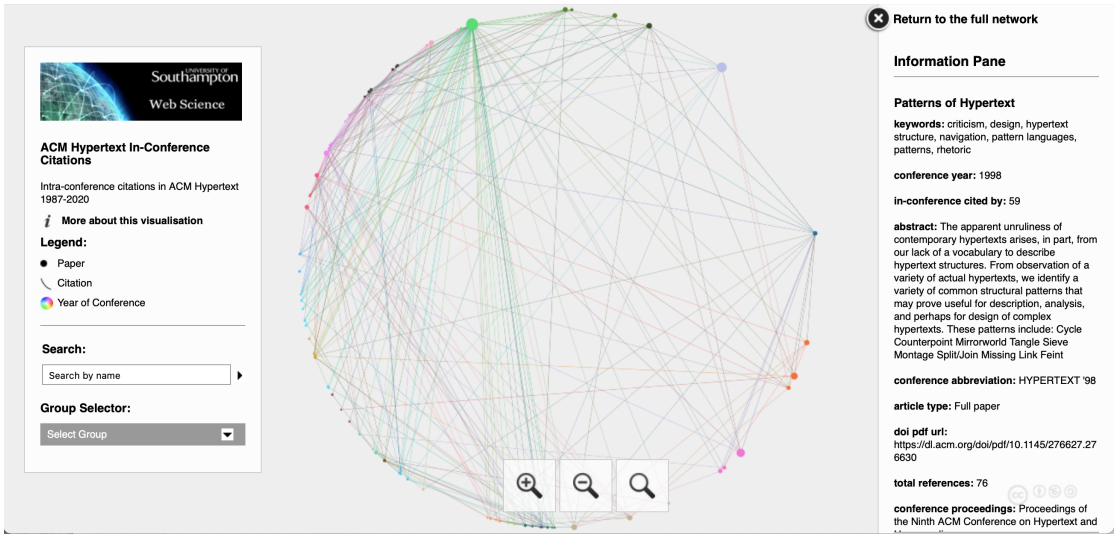

Figure 7: Visualisation of ACM HT Dataset.

The lesson here is that data not—or improperly—captured at the time is time-expensive to recover, not least as close inspection is needed for error detection.

5.1.1 What Is (Believed) Lost? Of concern for future scholarship is that the records of published proceedings only preserve a small part of the overall conference. Without a clear rationale, different proceedings include various secondary items like demos or side events (exhibitions, demos) were omitted from HT proceedings and thus may be already be lost to history and future study. Keynotes are often missing. This creates anomalies like Halasz's closing keynote at HT'91, omitted from the Proceedings [17] and this despite its importance in the context of his earlier 1987 treatise on this topic [16] and 2001 reprise [19]. The keynote has since been found in the Internet Archive [18], but aside from the content, the lack of formal record makes correct citation difficult. Whilst the papers provide the primary academic resource, other conference items and those outside the main programme—such as workshops—are useful to record as they give further insight as to the state of the subject at that time. Without paper print (cost) constraints, or as digital-only addenda, it is worth considering what is being discarded due to constraints that arguably no longer apply.

5.2 Future Work

Additional Conferences. Future work planned includes adding data for future HT conferences and assessing data from other hypertext conferences for inclusion. As there was no HT conference in 1995, IWHD'95 [13] was the only significant Hypertext-related conference that year. It is for consideration that hypertext-related journals of the period be considered for the dataset, though this needs consideration in terms of the scoping of 'in-context' citations. However, with careful application of link type values it is perfectly reasonable to be able to interrogate the dataset's link-base for links only to ACM conferences or into a wider source set. Such is the power of planned rather than harvested metadata.

Authors' Affiliations. We hope to record author affiliations, to explore geographical sources of conference input. This requires reviewing existing item sources to extract author affiliation, and as with the author data, de-duplicate and normalise it to give a canonical record for organisations. This is a large task, given there are c.26k references as of HT'21, albeit with some duplication. However, recording this could also enhance the citation trees and point to significant external influences and offer more geodata to enhance visualisations.

Keyword Coverage. As we know the conferences in which keywords were used. Do the keywords give us a usable lens through which judge the conference's interest over time?

Types of Citation. A further consideration is to look at the type of cited source, e.g. proceedings, journals, printed paper collections, vs. books/monographs, vs. websites (blogs, etc.) to reveal how this varies over time.

Curated tagging. Given the weakness of author tagging as a search resource, it would be useful to create some agreed thematic tags and apply them to the existing HT docuverse. As then-current terminology changes the keyworded terms, this would make it easier to find and follow, longitudinally, trends in the conferences (and in linked datasets).

Leverage Visual Meta. Given the poor portability of information to and from PDFs, attaching 'Visual Meta' [23] with enhanced metadata might improve the utility of existing papers in the HT docuverse.

6 CONCLUSIONS

In this paper we have argued that the meta-history of a conference—the metadata around its events, publications, and people—is an important resource that can help a community to reflect on its own health and development. Using ACM Hypertext as a case study we have explored how such a dataset might be constructed and analysed for insights, noting the challenges in the process, both legacy and ongoing.

ACM Hypertext is a conference with a long history, well represented in the ACM Digital Library. Through a combination of manual and automatic data extraction we have created a dataset for ACM Hypertext with over 1,000 publications, and more than 2,500 authors. An initial slice of this dataset has been released alongside this paper as 3-star open data (excluding gender information and full text representations for legal and ethical reasons).

An initial deep dive of this docuverse shows some of the potential. We have been able to see the health of the conference, its dip in publication volume around 2006, and subsequent recovery through the inclusion of social media research. We can also see the impact of this on the cohesion of the conference in terms of a reduction of incitations. We have been able to explore diversity, and have shown that the conference has consistently represented men better than women, although this may be improving (counter to the popular perception). Finally, we have looked at topics, revealing a broad and thinly distributed set of keywords that does not do a good job of reflecting the interests of the community, and may be a reflection of a lack of focus on key challenges or areas.

We have also reflected on the process of collecting the dataset, a challenge that extends to all academic conferences, some significantly larger than Hypertext. Restrictions around the ACM Digital Library meant that much of the primary data needed to be collected and input by hand (we used *Tinderbox* for collation), this included reviewing the PDFs of each paper separately, and applying OCR methods to create a plain-text representation for further analysis. Some key data (such as gender) was missing and needed to be reconstructed. These difficulties indicate that we are not curating our data in a way that is helpful to future generations of researchers. Our plan is to refine and extend the dataset (with information such as affiliation, citation types, and adding back missing data such as demos and workshop items). More complex automated analysis of the papers (through machine learning techniques such as clustering, topic analysis, anomaly detection, etc.) should also lead to new insights.

The research community dataset described in this paper requires a considerable amount of work to build, begging the question of whether that effort is worth it. The work shows that richer (alt)metrics are not 'free'—if not planned for, and that legacy print metadata his limited use beyond formal bibliometrics. However, it is also substantially easier to maintain a dataset than create one, and by releasing ours as open data we hope that others will be able to explore it, provide new views, and add to it through innovations of their own. Thus the effort over time is reduced, and the benefit maximised. After all 'Man will become better when you show him what he is like' (*Note-Book of Anton Chekhov*, 1921).

ACKNOWLEDGMENTS

Thanks to Charlie Hargood and Mark Bernstein for their historical perspectives on ACM HT and connected conferences, and to Amber Bu who assisted with the gender identification of Chinese names.

REFERENCES

[1] Tim Berners-Lee. 2006. *Linked Data.* W3C.org. https://www.w3.org/DesignIssues/LinkedData.html

[2] Tim Berners-Lee, Wendy Hall, James Hendler, Nigel Shadbolt, and Daniel J. Weitzner. 2006. Creating a Science of the Web. *Science* 313, 5788 (2006), 769–771. https://doi.org/10.1126/science.1126902 arXiv:https://www.science.org/doi/pdf/10.1126/science.1126902

[3] 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, Pennsylvania, USA). Association for Computing Machinery, New York, New York, USA, 21–29. https://doi.org/10.1145/276627.276630

[4] Sylvia Beyer. 2014. Why are women underrepresented in Computer Science? Gender differences in stereotypes, self-efficacy, values, and interests and predictors of future CS course-taking and grades. *Computer Science Education* 24, 2–3 (2014), 153–192. arXiv:https://doi.org/10.1080/08993408.2014.963363 https://doi.org/10.1080/08993408.2014.963363

[5] Jay David Bolter and Michael Joyce. 1987. Hypertext and Creative Writing. In *Proceedings of the ACM Conference on Hypertext* (Chapel Hill, North Carolina, USA). Association for Computing Machinery, New York, New York, USA, 41–50. https://doi.org/10.1145/317426.317431

[6] Joanne McGrath Cohoon and William Aspray. 2006. *Women and Information Technology: Research on Underrepresentation.* The MIT Press, Cambridge, MA, USA. 137–180 pages. https://doi.org/10.7551/mitpress/9780262033459.001.0001

[7] J. McGrath Cohoon, Sergey Nigai, and Joseph "Jofish" Kaye. 2011. Gender and Computing Conference Papers. *Commun. ACM* 54, 8 (Aug 2011), 72–80. https://doi.org/10.1145/1978542.1978561

[8] Blaise Cronin and Cassidy R. Sugimoto. 2014. *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact.* MIT Press, Cambridge, MA, USA. 432 pages. https://doi.org/10.7551/mitpress/9445.001.0001

[9] H. C. Davis, D. E. Millard, S. Reich, N. Bouvin, K. Grønbæk, P. J. Nürnberg, L. Sloth, U. K. Wiil, and K. Anderson. 1999. Interoperability between Hypermedia Systems: The Standardisation Work of the OHSWG. In *Proceedings of the Tenth ACM Conference on Hypertext and Hypermedia* (Darmstadt, Germany) *(HYPERTEXT '99).* Association for Computing Machinery, New York, NY, USA, 201–202. https://doi.org/10.1145/294469.294904

[10] Stephan Doerfel, Robert Jäschke, and Gerd Stumme. 2012. Publication Analysis of the Formal Concept Analysis Community. In *Formal Concept Analysis.* Springer Berlin Heidelberg, Berlin, Heidelberg, 77–95. https://doi.org/10.1007/978-3-642-29892-9_12

[11] Eastgate Systems, Inc. 2018. *Tinderbox: The Tool For Notes.* Eastgate.com. http://www.eastgate.com/Tinderbox/

[12] Douglas Carl Engelbart and William K. English. 1968. A Research Center for Augmenting Human Intellect. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I.* ACM, San Francisco, California New York, NY, USA, 395–410. https://doi.org/10.1145/1476589.1476645

[13] Sylvain Fraïssé, Franca Garzotto, Tomas Isakowitz, Jocelyne Nanard, and Marc Nanard (Eds.). 1996. *Hypermedia Design* (Montpelier, France). Vol. Proceedings of the International Workshop on Hypermedia Design. Springer Science & Business Media. https://doi.org/10.1007/978-1-4471-3082-6

[14] Yves Gingras. 2016. *Bibliometrics and Research Evaluation.* MIT Press, Cambridge, MA, USA. 136 pages. https://doi.org/10.7551/mitpress/10719.001.0001

[15] Thomas Gruber. 2007. Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web and Information Systems (IJSWIS)* 3, 1 (2007), 1–11. https://www.igi-global.com/article/ontology-folksonomy-mash-apples-oranges/2828

[16] Frank G. Halasz. 1987. Reflections on NoteCards: Seven Issues for the next Generation of Hypermedia Systems. In *Proceedings of the ACM Conference on Hypertext* (Chapel Hill, North Carolina, USA). Association for Computing Machinery, New York, NY, USA, 345–365. https://doi.org/10.1145/317426.317451

[17] Frank G. Halasz. 1991. "Seven Issues": Revisited. In *Proceedings of ACM Hypertext '91 Conference* (San Antonio, Texas, USA). Association for Computing Machinery, New York, NY, USA. Closing keynote, not in Proceedings.

[18] Frank G. Halasz. 1996. *"Seven Issues": Revisited.* PARC XEROX. https://web.archive.org/web/19970605120053/http://www.parc.xerox.com:80/spl/projects/halasz-keynote/ 1991 keynote, recovered transcript from audio; via archive.org.

[19] Frank G. Halasz. 2001. Reflections on "Seven Issues": Hypertext in the Era of the Web. *ACM Journal of Computer Documentation (JCD)* 25, 3 (2001), 109–114. https://doi.org/10.1145/507317.507328

[20] Frank G. Halasz, Mayer Schwartz, Kaj Grønbæk, and Randall H. Trigg. 1994. The Dexter Hypertext Reference Model. *Communications of the ACM (CACM)* 37, 2 (1994), 30–39. https://doi.org/10.1145/175235.175237

[21] Michael Hausenblas and James G. Kim. 2012. *5-star Open Data.* W3C.org. https://5stardata.info/en/

[22] Frode Hegland. 2018. *The Future of Text.* Future of Text. https://www.thefutureoftext.org

[23] Frode Hegland. 2019. Visual-Meta: An Approach to Surfacing Metadata. In *Proceedings of the 2nd International Workshop on Human Factors in Hypertext* (Hof, Germany). Association for Computing Machinery, New York, NY, USA, 31–33. https://doi.org/10.1145/3345509.3349281

[24] James Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, and Daniel Weitzner. 2008. Web Science: An Interdisciplinary Approach to Understanding the Web. *Commun. ACM* 51, 7 (jul 2008), 60–69. https://doi.org/10.1145/1364782.1364798

[25] Diana Hicks, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. 2015. Bibliometrics: the Leiden Manifesto for research metrics. *Nature News* 520, 7548 (2015), 429. https://doi.org/10.1038/520429a

[26] Luke Holman, Devi Stuart-Fox, and Cindy E. Hauser. 2018. The gender gap in science: How long until women are equally represented. *PLoS biology* 16, 4 (2018), e2004956. https://doi.org/10.1371/journal.pbio.2004956

[27] Clare Hooper, Isabella Peters, and Cécile Robin. 2018. Mapping the Topics and Intellectual Structure of Web Science. *The Journal of Web Science* 5 (2018), 1–15. https://doi.org/10.34962/xhmm-0425

[28] Nettie Lagace. 2016. NISO Releases Recommended Practice Covering Outputs of Its Multiyear Project in Alternative Assessment Metrics. *Serials Review* 42, 4 (2016), 337–338. https://doi.org/10.1080/00987913.2016.1246343

[29] 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, Texas, USA). Association for Computing Machinery, New York, NY, USA, 261–275. https://doi.org/10.1145/122974.123000

[30] 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, UK). Association for Computing Machinery, New York, NY, USA, 13–23. https://doi.org/10.1145/192757.192759

[31] Ray McAleese (Ed.). 1989. *Hypertext: Theory Into Practice*. Intellect Books, Oxford. https://www.intellectbooks.com/hypertext-1

[32] Ray McAleese and Catherine Green (Eds.). 1990. *Hypertext: State of the Art*. Intellect Books, Oxford. https://www.intellectbooks.com/hypertext

[33] David Millard. 2008. Hypertext 2008: A Great Safari (ACM SIGWEB Trip Report). *ACM SIGWEB Newsletter* October, 2008 (October 2008), 1–6. https://eprints.soton.ac.uk/266674/

[34] Theodor Holm Nelson. 1968. *Hypertext Implementation Notes*. Archive.org. https://archive.org/details/hin68

[35] Jakob Nielsen. 1988. Hypertext '87. *SIGCHI Bulletin* 19, 4 (1988), 27–35. https://doi.org/10.1145/43950.43953

[36] Jakob Nielsen. 1990. Trip Report: Hypertext '89. *SIGCHI Bulletin* 21, 4 (1990), 52–61. https://doi.org/10.1145/379106.379121

[37] NISO. 2016. *Outputs of the NISO Alternative Assessment Metrics Project*. Vol. NISO RP-25-2016. National Information Standards Organization (NISO), Baltimore, MD, USA. 77 pages. http://www.niso.org/publications/rp-25-2016-altmetrics

[38] Jason Priem, Dario Taraborelli, Paul Groth, and Cameron Neylon. 2011. *Altmetrics: A manifesto*. Altmetrics.org. http://altmetrics.org/manifesto/

[39] Antoine Rizk, Norbert A. Streitz, and Jacques André (Eds.). 1990. *Hypertext: Concepts, Systems and Applications*. Cambridge University Press. https://www.cambridge.org/gb/academic/subjects/computer-science/computing-and-society/hypertext-concepts-systems-and-applications-proceedings-first-european-conference-hypertext-inria-france-november-1990

[40] Frank M. Shipman, III and Catherine C. Marshall. 1999. Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems. *Computer Supported Cooperative Work (CSCW)* 8, 4 (1999), 333–352. https://doi.org/10.1023/A:1008716330212 based on their 1993 Technical Report ISTL-CSA-94-08-02, q.v.

[41] 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). Association for Computing Machinery, New York, NY, USA, 11–22. https://doi.org/10.1145/168466.168479

[42] Thomas Vander Wal. 2007. *Folksonomy*. Vvanderwal.net. https://www.vanderwal.net/essays/051130/folksonomy.pdf

[43] Jill Walker. 2005. Feral Hypertext: When Hypertext Literature Escapes Control. In *Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia* (Salzburg, Austria) *(HYPERTEXT '05)*. Association for Computing Machinery, New York, NY, USA, 46–53. https://doi.org/10.1145/1083356.1083366

[44] Polle T. Zellweger. 1989. Scripted Documents: A Hypermedia Path Mechanism. In *Proceedings of the Second Annual ACM Conference on Hypertext* (Pittsburgh, Pennsylvania, USA). Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/74224.74225

Visual-Meta Appendix

The data below is what we call Visual-Meta. It is an approach to add information about a document to the document itself, on the same level of the content (in style of BibTeX). It is very important to make clear that Visual-Meta is an approach more than a specific format and that it is based on wrappers. Anyone can make a custom wrapper for custom metadata and append it by specifying what it contains: for example @dublin-core or @rdfs.
The way we have encoded this data, and which we recommend you do for your own documents, is as follows:
When listing the names of the authors, they should be in the format 'last name', a comma, followed by 'first name' then 'middle name' whilst delimiting discrete authors with ('and') between author names, like this: Shakespeare, William and Engelbart, Douglas C.
Dates should be ISO 8601 compliant.
Every citable document will have an ID which we call 'vm-id'. It starts with the date and time the document's metadata/Visual-Meta was 'created' (in UTC), then max first 10 characters of document title.
To parse the Visual-Meta, reader software looks for Visual-Meta in the PDF by scanning the document from the end, for the tag @{visual-meta-end}. If this is found, the software then looks for @{visual-meta-start} and uses the data found between these tags. This was written September 2021. More information is available from https://visual-meta.info for as long as we can maintain the domain.

@{visual-meta-start}

@{visual-meta-header-start}
@visual-meta{version = {1.1},
generator = {ACM Hypertext 21},
organisation = {Association for Computing Machinery}, }

@{visual-meta-header-end}

@{visual-meta-bibtex-self-citation-start}

@inproceedings{10.1145/3511095.3531271,
author = {Anderson, Mark W. R. and Millard, David E.},
title = {Hypertext's meta-history: Documenting in-conference citations, authors and keyword data, 1987-2021},
year = {2022},
isbn = {978-1-4503-9233-4},
publisher = {Association for Computing Machinery},
address = {New York, NY, USA},
url = {https://doi.org/10.1145/3511095.3531271},
doi = {10.1145/3511095.3531271},
abstract = {Conferences such as ACM Hypertext have been running for many decades and the metadata on their collected publications represent a valuable scholarly meta-history on areas such as the community's health, diversity, and changing interests. But the metadata about these papers is not readily available for analysis, and the data collection and cleaning tasks appear substantial. In this paper we attempt to explore this challenge using the ACM Hypertext series as a case study. Taking the ACM Digital Library as a starting point, and using a combination of manual and automatic methods, we have constructed and released a 3-star Open Dataset representing over 1000 publications by almost 2,500 authors. An initial analysis reveals a modestly-sized but robust conference, with a changing pattern of in-citations that co-occurs with the arrival of social media, and a relatively consistent but imbalanced gender ratio of authors that shows some signs of recent improvements. The challenges encountered included identifying discrete author names, potential issues with text retrieval from PDF, and a disparate set of author keywords that reveals an absence of a common vocabulary. These insights are the results of a hard-fought process that is made complex by an incomplete digital record and a lack of consistency in naming. This Hypertext case study thus reveals a serious shortfall in the way that scholarly activity is captured and described, and questions PDF as the primary method of recording publications. Addressing these issues would make further analysis more straightforward and would allow larger events (with orders of magnitude more data) to be analysed in a similar way.},
numpages = {11},
keywords = {hypertext, knowledge management, metadata, citation networks, links, linkbases, gender, Tinderbox, dataset, keywords, keywording, visualisation, analysis}, location = {Barcelona, Spain},
series = {HT '22},
vm-id = {10.1145/3511095.3531271} }

@{visual-meta-bibtex-self-citation-end}

@{visual-meta-end