

# Digitizing the USPTO patent backfile

Simon Rowberry  <sup>1,\*</sup>

<sup>1</sup>Department of Information Studies, Faculty of Arts and Humanities, University College London, London, United Kingdom

\*Corresponding author. Department of Information Studies, Faculty of Arts and Humanities, University College London, London, United Kingdom E-mail: s.rowberry@ucl.ac.uk

## Abstract

The digitization of the US Patent and Trademark Office's (USPTO) backfile of six million patents undertaken between 1951 and 2001 was a five-decade struggle, featuring several media transitions from print and microfilm to CD-ROMs and, finally, the Web. This mass digitization project is on a similar scale to Google Books and the Internet Archive, but it is rarely discussed within critical digitization scholarship or for its significance as a tool for knowledge production. In this article, I focus on the USPTO's patent document's digital and physical material form and how the current paradigm of access and storage of the digital backfile emerged. Through this case study, I build upon Ian Milligan's distinction between the 'text' and 'platform' layers of a digitization project to demonstrate how historical decisions regarding format and metadata continue to influence how users retrieve and interpret documents, such as patents, online.

**Keywords:** digitization; formats; platforms; patent databases.

## 1. Introduction

In September 2022, the United States Patent and Trademark Office (PTO) shut down its long running Patent Full-Text and Image Database (henceforth 'PatFT') service, initially launched in 1999. The replacement, Patent Public Search (PPUBS), marked the PTO's first major interface overhaul in over 20 years. The front-end's perceived stability masked the more complex transformations that have occurred behind the scenes. Patent databases, especially those with a near complete backlist and long history of digitization such as the PTO's, present a useful test case for critical digitization studies that de-centres books to reflect upon how other types of documents have been altered by the transition to digital media. Our overfamiliarity with the format of books as physical objects and their digital representations masks the complexity of converting physical documents into digital media with embedded data that can be extracted for re-use.

In this article, I take a bibliographical and media historical informed approach to the development of the USPTO's publicly available online patent backfile. I propose a conceptual 'stack' model of digitization building on earlier work by Milligan (2022), Montfort and Bogost's (2009) model of platform studies, and

Bratton's (2015) media ecological stack. I emphasize the interplay between layers of the stack rather than considering them in isolation. Through tying this format-based analysis to the historical development of the USPTO's computerization projects, I demonstrate how the geological layers of long-term digital infrastructure projects continue to affect contemporary developments, even when they are multiple layers deep.<sup>1</sup> The USPTO's databases are a useful case study due to their lengthy history of digitization, detailed in the following section, as well as the complexity of the platform's infrastructure, and the Patent Office's emphasis on interoperability.

The 'illusionary order' of digitized archives (Milligan 2013) hides the more complex digital textuality underpinning platforms such as the USPTO's various databases. This 'illusory order' exists for many reasons: For example, the Patent Office conducted several digitization projects with competing aims that led to what Paul Edwards has termed 'data friction' (Edwards 2010: 100). This led to a two-tier system between more modern patents that are available in multiple forms, and patents granted before 1970, which are only available as facsimiles or in an arcane text-only format (Automated Patent System or APS). This inconsistent level of access renders the geological layers of the PTO patent

digitization process more visible, which in turn provides a fruitful case study to consider deeper questions about digitization and digital textuality more broadly.

## 2. The origins of digital patents

Interdisciplinary research has shed light on the development and impact of digitization projects ranging from Melissa Terras' (2010) documentation of amateur collections through to analyses of Google Books (Duguid 2007) and Early English Books Online (EEBO) (Mak 2014). Since the USPTO patent database contains millions of records, I focus here on mass digitization. Nanna Bonde Thylstrup describes mass digitization projects as assemblages 'consisting of humans, machines, objects, subjects, spaces and places, habits, norms, laws, politics and so on' rather than monolithic interfaces (Thylstrup 2019: 20).

There is a lack of temporality to Thylstrup's list, which is vital to consider in long-running digitization projects such as the USPTO's since the same digitized file can appear in a radically different context over the passage of time. We can account for this discrepancy through drawing upon what Helmond and van der Vlist (2019: 8) call 'platform historiography, [which] like historiography in general, foregrounds the methodological considerations and reflections associated with the use of multiple sources to interpret platforms' pasts'. Helmond and van der Vlist combine web archiving methods and insights to the exploration of social media, but this approach of considering the temporal dimensions of the platform itself, and what has been lost over time, is equally important as documenting the stages of the digitization process and the actors involved in that process.

Sociological approaches to digitization like Thylstrup's have been illuminating, but I build here primarily on the growing body of scholarship focusing on the bibliographic and material analysis of digitization (McKitterick 2013; Mak 2014; Gregg 2020) instead of policy or sociological dimensions.<sup>2</sup> As Adam Crymble reminds us, 'the history of digitization is part of a longer narrative of adaptive storage solutions, including early microfilming and microfiche efforts' (Crymble 2021: 46). Older digitization projects, such as work on the PTO patent backfile, which often involve an intermediary step of digitizing microfilm (Bellido 2023), require a different approach than initiatives such as Google Books that have shorter implementation periods with more stable infrastructural development.

Furthermore, critical digitization studies often focus on cultural heritage databases. PatFT and PPUBS sit awkwardly within that classification, as tools designed primarily for Patent Office examiners and customers to discover prior art. Nonetheless, secondary uses

include research and aesthetic repurposing through services such as printaparent.com which extend the cultural value of the dataset.<sup>3</sup> Despite this market, the USPTO does not encourage secondary uses beyond allowing access to the materials through their public databases or visiting NARA's physical holdings. The PTO is unique among Federal Offices as it is fully funded through user fees as codified by the 1990 Omnibus Budget Reconciliation Act and the 1999 American Inventors Protection Act that converted the agency to a 'performance-based' model (United States Patent and Trademark Office 2019b: 9). While this ensures the Office can withstand closure during moments of government shutdown, all costs must be justified to the core business users. Similarly, technological solutions prioritize 'pendency', or the length of the examination process. The financial imperative for the USPTO is therefore to orientate the design of their patent platforms towards improving the efficiency of the examination process rather than enriching the collection as part of cultural heritage. The platform design reflects this focus. Therefore, newer patents that are still actively protected are more readily available, and images are not integrated since they are not scrutinized in the examination process.

It is impossible to reconstruct a strict linear chronology of the PTO's digitization efforts as the Patent Office and its third-party contractors undertook several large-scale projects concurrently. This messiness is the focus of my argument rather than reconstructing a comprehensive and chronological history of the USPTO's adoption of computers in its workflow. To this end, this section briefly sketches a general history, summarized in Table 1.

The need to simplify the indexing of chemical equations led the Chemical Division to start working on automated patent searches as early as 1947 (Newman 1960: 734). The Dissemination of Technology, Scientific, and Engineering Information Act in September 1950 formed part of a broader push towards enabling private companies to maximize the use of public records (Bush 1950). The Patent Office

**Table 1.** Overview of stages of the USPTO's digitization programme

Date range	Digitization paradigm
1950–70s	Conversion from paper to microfilm
1970–90s	First round of digitization, extracting the text for information retrieval databases. Early attempts at standardization
1990s to early 2000s	Facsimiles distributed by optical media (CD-ROMs, laserdisc)
2000–22	Web-based search, separating PDFs from HTML
2022–	Hybrid HTML and PDF environment

picked up this initiative in 1954, when the US Senate Appropriations Committee directed the Department of Commerce to ‘make an aggressive and thorough investigation as to the possibility of mechanizing the search operations [...] to modernize, insofar as practical, the Patent Office operations’, a theme that would continue until the programme’s completion in the 2000s (Newman, 1960: 731).

In response to Congress’s demands, Robert C. Watson, then Commissioner of Patents, formed a cross-departmental committee to address these concerns. The Senate Appropriations Committee asked Vannevar Bush to chair the committee (Bush 1954: 3–4). Bush played a pivotal role in the World War II, leading the Office of Science Research and Development as well as being instrumental in the development of the atomic bomb (Zachary 2018). He had long been interested in information retrieval using analog computers, creating the Rapid Selector, a microfilm-based search device in the late 1930s (Zachary 2018: 408). The device was initially used for wartime codebreaking but was soon adopted by government agencies including the PTO (Bush 1954: 10; Zachary 2018: 276). Bush was a keen and persistent advocate for sweeping patent office reform including tackling overwhelmed courts and monopolies (Bush 1945, 1936: 227). Despite this advocacy, Watson recruited Bush for his expertise with the Rapid Selector and his scope for recommendations were limited to mechanization (Worthy 1954).

At a final meeting in November 1954, the Committee agreed that the Patent Office should proactively develop computers for its multifaceted needs (Advisory Committee on Application of Machines to Patent Office Operations. Minutes. Fourth Meeting 1954). The recommendations covered a range of areas from reclassification to research and development capacity building, but the PTO only adopted the recommendation to convert patents from print to microfilm. By 1962, the Patent Office focused on aperture cards with embedded microfilm to aid the search process (Bagg and Stevens 1962: 27). It took 10 years between 1962 and 1972 to fully microphotograph the backfile with help from Eastman Kodak and the microfilm was only made available to the public in February 1977 (USPTO 1981: 34–35). Early attempts to mechanize patents using microphotography paved the way for digital image-based patent storage, such as the laserdisc experiments by Pergamon and International Computaprint Corporation (ICC, later merged as Reed Technologies and now part of RELX) (White 1986: 178).

As part of a broader suite of reforms to the previous Patent Act of 1952, the Bayh–Dole Act (1980) consolidated previous computerization experiments. Section 9 requested a plan by the Commissioner of Patents and Trademarks on developing computerized systems for

‘all aspects of the operation of the Patent and Trademark Office’ (Bayh and Dole 1980: 9). In an early response to the Bayh–Dole Act, a 1983 report by Howard Bryant and Donald Stein emphasized the extent of the problem with analog holdings: ‘Maintenance of this massive paper file is costly and error prone. At any one time, an average of over 7 per cent of the documents may be misfiled or missing. The state of search file integrity jeopardizes the quality of patent examinations’ (Bryant and Stein 1983: 226).

Both internal and external factors drove the PTO’s push towards digitization. By the mid-20th century, national patent offices were investigating the potential for sharing data more efficiently using new distribution formats including microfilm and computers. The October 1961 International Patent Office Workshop on Information Retrieval in Washington led to the foundation of the International Cooperation in Information Retrieval Among Examining Patent Offices (ICIREPAT) and in 1978, the formation of the Patent Cooperation Treaty (USPTO 1981: 31). These projects focused on standardization and datafication of patents to ensure consistency between different national publications. This would future proof patent publication workflows and prepare the backfile for digitization and database entry. The pressure for easily spreadable data only became more important in the interim as more organizations including Google and IP5 (a consortium of the five largest national and supranational patent-granting institutions) sought to reuse national patent databases in new contexts.

The PTO invested heavily in digitization, ensuring that the complete backlist would be available to the public. Users can access all extant granted patents in the USA from Samuel Hopkin’s filing for the manufacture of potash in 1790 to weekly updates of new patents via Public Patent Search (see Table 2 for summary). Access is uneven, however, as only patents filed after 1975 are available as fully searchable text. Earlier patents are only available as facsimiles. The digital records available online are more complete than NARA’s physical records. At least 0.27 per cent (11,091) of patents published before 1979 are no longer available via NARA. Conversely, only withdrawn patents are unavailable as facsimiles via the PTO’s databases. The full-text database is less complete, with 187 missing patents, excluding a considerable number of withdrawn patents.<sup>4</sup> The USPTO acknowledges these absences on a separate webpage, removed from the context of the primary search engine (United States Patent and Trademark Office 2020). Users can only see skeletal metadata with little indication or reason for their absence. This lack of cohesion between different patent datasets has led to substantial variation in the presentation and contextualization of available patent records.

**Table 2.** Overview of patent records 1790–2001 and their digital availability

Dates	Last patent number	Total	Available as PDF	Available as HTML	Physical copies at NARA	Withdrawn
1790–1823 (X series)	10,280		2,625 [c.25.6%]	0		
1823–970	3,551,908	3,551,908	3,548,645 [99.9%]	0 [0%]	3,551,908 [100%]	3,262 [0.09%]
1971–5	3,930,271	378,363	375,549 [99.2%]	0 [0%]	378,363 [100%]	2,814 [0.7%]
1976–8	4,131,951	201,680	201,596 [99.9%]	201,029 [99.7%]	190,589 [94.5%]	84 [0.04%]
1979–2001	6,167,569	2,035,618	2,023,209 [99.3%]	2,010,800 [98.8%]	0 [0%]	12,409 [0.7%]

### 3. Critical digitization studies

Zach Lischer-Katz notes that we should pay attention to formats when considering digitization as it ‘is never merely a direct transmission of signals between formats but is perhaps better understood as a process of translation between two media formats constituted by fundamentally different representational systems’ (Lischer-Katz 2022: 1259). Lischer-Katz’s definition is broad enough that we can consider any conversion into a digital format as a digitization, even if it was previously available in a digital format. This allows us to compare analog to digital conversions with changes in digital formats that have been discussed within web archival contexts as ‘re-born digital’ (Brügger 2013: 758). We should account for ‘re-born digital’ content in longer term digitization projects where the underlying formats and standards change multiple times, which can increase the complexity of maintaining and updating records.

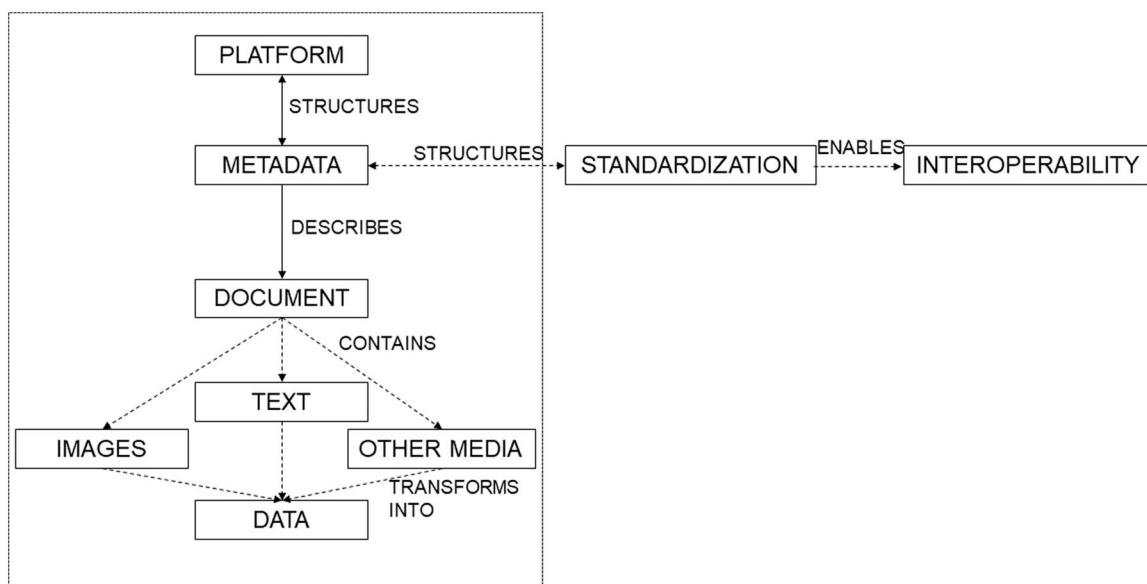
Through considering the longer history of patent computerization, I expand upon the work of scholars including Bonnie Mak and Stephen Gregg. Mak’s (2014: 1521) ‘Archaeology of a Digitization’ traces the development of EEBO from microfilm versions of books to a web service, concluding that ‘an archaeology of a digitization, then, should understand the digitally encoded entity as a cultural object, produced by human labor, and necessarily shaped by—and consequently embodying—historical circumstance’. Gregg’s analysis of Eighteenth-Century Collection Online (ECCO) argues that ‘the status of entities like ECCO and their digitized books actually challenges the notion of a linear, progressive history’ (Gregg 2020). As digitized and born-digital collections become an integral part of contemporary historical research, it is vital to understand how the interplay between platform and text affects the digital surrogate, and what this says more broadly about digital textuality.

Ian Milligan’s work on digitization provides the foundations of my theoretical intervention into critical

digitization studies. In *The Transformation of Historical Research in the Digital Age*, Milligan distinguishes between the ‘text’ and the ‘platform’ layers within the digitization process, with the platform mediating access to the ‘raw data’ (Milligan 2022). This is a useful model, especially with further adaption to account for the complexity of the digitization process. In this article, I expand this model to a five-layer stack (visualized in Fig. 1) as a generalizable model of digitization platforms: *platform*, *metadata*, *document*, *media/content*, and *data*.

Milligan avoids defining the term *platform* but, in this context, it refers to a website or database for accessing media (Montfort and Bogost 2009; Gillespie 2010). The second layer, *metadata* (data about data), may be visible either in or around the document, or rendered hidden in the source code. The *document* constitutes the representation of the digitized content on screen as a facsimile or a born-digital document such as a webpage. We can transform the second layer of Milligan’s model, *text*, into the more generalizable form of *media*, such as patent documents where images are significant parts of documents that might be lost through the digitization workflow. The final layer, *data*, can refer to two distinct yet interconnected phenomena: first, the underlying layers of alphanumeric code that represent the text and other media that flow through binary, hexadecimal and various other encoding systems that interact with a wide variety of formats and standards; and secondly, ancillary uses of the document as a data source from either an internal or external user.

The distinction between media and data is a particularly rich area for exploration in understanding how digital documents and files such as patents are re-circulated and used beyond their original purpose. Choosing which media qualify as important data will lead to different design decisions. In terms of the USPTO’s patent database, this can clearly be seen in the treatment of various media types in the process of conversion. The main text of the patent and its textual



**Figure 1.** The five-layer stack of digitization. The stack model is adapted from both Bratton (2015) and Montfort and Bogost (2009).

metadata have been prioritized, appearing in both the HTML and PDF versions of patents. Alongside facsimile copies of patents, this is the material that is most likely to be re-used by third-party datasets. Images are seen as less important from both a media and data perspective, but this then further ignores the other types of media and data that might be included in a patent filing such as computer code, DNA, or other ancillary documentation.

This internal model of a digitization project does not exist in a vacuum, so *standardization* and *interoperability* are two major external forces that shape the stack. While Edwards (2010: 268) optimistically notes, ‘Standards act as lubricants. They reduce friction by reducing variation, and hence complexity’, Pargman and Palme (2009: 186) argue that standards are an inertial force ‘restricted by decisions that were made long ago’ and any ‘alterations are expensive’. Nonetheless, typographic and format standards shape how users interact with the underlying media. Standardized metadata ensure that users can locate relevant information, while formats maintain consistent tags which makes it easier for third parties to create content using those standards. Equally, interoperability is vital in areas such as patent examination where multiple governing bodies need to cooperate, as well as other secondary uses.

#### 4. From page to screen

The first step of any digitization project is converting the analog material into digital form, which is often a

time and resource intensive process. Scanning and correcting paper copies of the patent records was an on-going project involving multiple third-party contractors. For example, Access Innovations won a contract to conduct a round of digitization. The company, led by Marjorie Hlava, conducted the digitization in a former limestone mine in the small town of Boyers, Pennsylvania, where the federal government had contracted Iron Mountain to facilitate storage of personnel records (Hlava 2014a).<sup>5</sup>

Given the fragility of some of the documents and the challenging working conditions, Access Information developed a unique workflow to ensure maximum readability and preservation of the original documents. There were trade-offs in the process: Access Information captured the images at the then high resolution of 300 bpi (bits per inch), but optical character recognition was not a priority and the PTO have not attempted to re-scan the documents at a higher resolution. Hlava (2014a: 53) acknowledges the limitation in this approach in a 2014 keynote: ‘we only scanned and delivered the OCRed text at 97 per cent [accuracy]. After much debate, I finally agreed that we could use the dirty OCR because statistically your term was likely to be spelled right at least once in the average thirty pages of a patent’. Smith and Cordell (2019) corroborate Hlava’s claims, suggesting that ‘dirty OCR’ is sufficient if researchers are aware of how it impacts interpretation. Since patent offices rely on other mechanisms such as classification to enable examiners to identify relevant patents, a higher level of accuracy was not desirable.

Focusing on text extraction alone ignores a large part of the patent backfile. William Rankin notes the drawings within patents are vital to the construction of the patent's originality, which is a problematic proposition since images are now largely uncoupled from the primary HTML patent reading interface. Rankin (2011: 58) argues 'intellectual property rights may not be granted to novelty which is claimed in writing but not shown in the drawing, and disclosure in a drawing can establish precedence even when not included in the text'. Conversely, Kang (2023: 6) notes: 'In contrast to public imagination of patent with geeky inventions depicted in technical drawings, the legal system is not particularly concerned with patent drawings'. This is reflected in the setup of the USPTO's database structure whereby it is more challenging to access a PDF of a patent than its full text, increasing the challenge for interested secondary users accessing this material.

As Grooms recalls, 'the PTO began capturing patents as ASCII [the American Standard Code for Information Interchange character set] data in 1970 as one of its first automation projects. This only included a small portion of the patents that issued but was increased each year until 1975 when all patents that issued, with some minor exceptions, were captured. (Drawings and other non-textual material were not captured.)' (Grooms 1988: 163). These dates represent the limits of full-text search still, indicating a historical divide between fully indexed documents and those only accessed through direct means. As a result, by a 1994 attempt to consolidate databases, there was 'a two-component system. The first component of the system contains all U.S. patents scanned to create digital image files with both patent text and drawings. [...] The second component of APS is a database containing U.S. patents, text only, issued since 1975' (Auyang 1994: 858).

There were clear reasons for separating these two projects: microfilm was a mature visual standard while computer screens were not set up for high-definition facsimiles, so it was more convenient to instead return just the text for users retrieving information. This split underlies one of the core tensions in the early history of digital publishing that reverberates to contemporary debates around the benefits of the PDF and EPUB formats: Should digital documents attempt to look like a facsimile of print or should they instead work in a new standard? These design choices within the document layer of the digitization stack in turn shape the platform layer. When the USPTO began to create an online patent search service, it had to reconcile these two different datasets, which led to the development of two separate servers: PatFT, the full-text search engine that renders the main body of a patent in HTML; and the

separate 'PIMG' server where single-page PDFs, based on older TIFF renderings from the digitization process, are available.

## 5. Are patents documents or data?

The opportunities of digitization included reconsidering what was the top priority for computerized access to patents. In one of the few published case studies of the digitization of patent records, Kang (2019: 58) argues that 'the referent of a patent is no longer the document or image, but digital data disconnected from its diagrammatic format and the physical media of paper or the .pdf'. In a digital-first patent office, every representation is just an assemblage (Thylstrup 2019: 20) of various different data points with no single source. While Kang correctly identifies a trend towards datafication of patent records, the documents remain rooted in print conventions rather than making best use of more accessible digital formats. Until the recent redesign of the search interface through the launch of PPUBS, the HTML version of a patent replicated the header structure of the print copy. This shift away from facsimiles shifted the skeuomorphism back a stage, as the new interface introduced more folder-based metaphors to replicate the desk environment of a patent examiner.

The long, difficult, history of digitization at the Patent Office left an indelible mark on current access to patents. Even when documents have been rescanned, a palimpsest of this earlier digitization remains in the USPTO's complex infrastructure, revealing a longer history in the metadata that the Patent Office often re-enforce through visual design. For example, Figs. 2–4 show how the USPTO designed the HTML version of the patent's header to reflect the aesthetic of the print-ready version. While the rest of the HTML design diverges from the PDF, this is a small reminder of the material history of patents as documents and their formats' contexts within the history of computing. As a result, the USPTO patent archives exist in a liminal space that highlights the awkward transition between print and digital databases. A digitized patent collection had the potential for full-text searching, which could increase opportunities for spotting emerging trends above print-based methods such as patent classification. While newer patents would be available in born-digital format and easily accessible, older patents required further work.

Previous research into the material history of patents emphasizes patents as text or their broader context, largely ignoring their history as bibliographical objects. Partially, this stems from the lack of a single 'published' form of a patent. Instead, new patents exist as data records distributed via various workflows

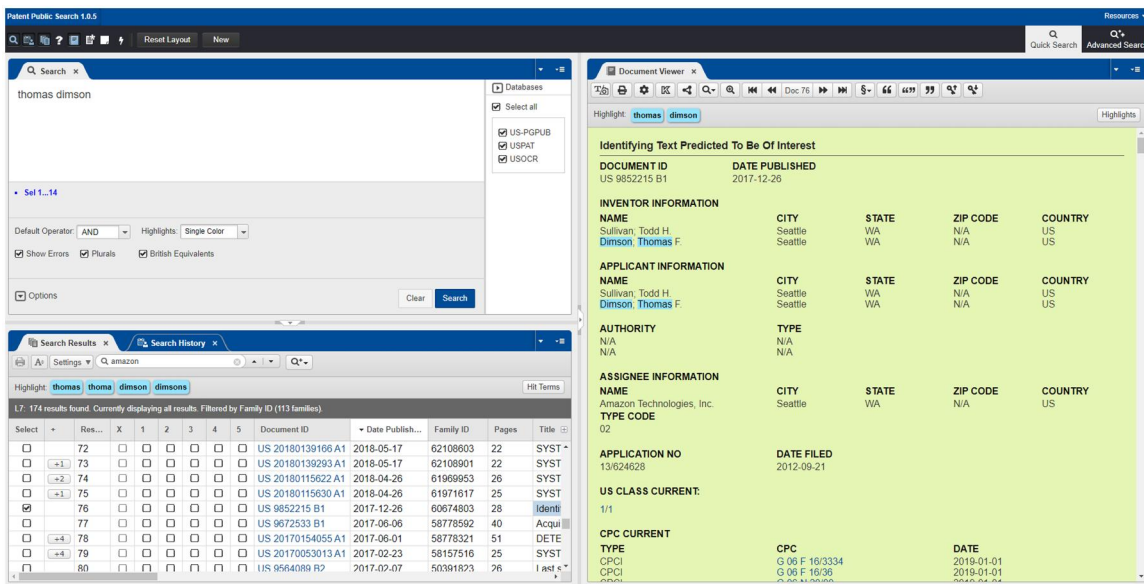


Figure 2. The PPUBS search interface with the full text on the right-hand side.

**USPTO PATENT FULL-TEXT AND IMAGE DATABASE**



( 1 of 1 )

United States Patent  
Sullivan , et al.

9,852,215  
December 26, 2017

Identifying text predicted to be of interest


**Abstract**

A body of text may be compared with one or more user-selected text portions to rank a plurality of text portions of the body of text, such as for predicting which of the text portions are likely to be annotated by users. As one example, the text of a content item may be compared with excerpts of other content items that have been highlighted or otherwise annotated by a plurality of users. Based at least in part on the comparison, some implementations identify one or more portions of text of the content item that are likely to be selected or highlighted by users that access the content item. In some examples, a classifier may be trained based on popular highlights determined for a plurality of content items. The classifier may be applied to a body of text to determine portions that users are likely to consider profound or interesting.

Figure 3. The HTML version of the patent as rendered by the now defunct PatFT interface.

including their destination as both PDFs and HTML on the USPTO’s website. Despite the perceived fluidity of digital patent records, the PTO still regards print as the format of record. In response to a discussion on ‘more stinky biblio’ on the Patent Information Users Group mailing list, Larry Larson, a PTO employee admitted that the Patent Office’s ‘electronic data, both in databases and in bulk data on magnetic media, is not intended to be a collection of absolutely correct

information; rather, it is intended to be an accurate rendering of PTO’s [print] legal publications’ (quoted by Calvin E VanSant in *United States Patent and Trademark Office 2002*).<sup>6</sup> Larson’s statement reveals a central constraint in the digitization process: it was seen primarily as a means of improving distribution rather than an opportunity to reconsider what a patent document might look like in the digital form. Interoperability thus shapes both the data and



United States Patent and Trademark Office

Home | Site Index | Search | FAQ | Glossary | Contacts | eBusiness | eBiz alerts | News

Patent #: US009852215 Section: 0 of 28 pages Help

US000009852215B120171226 1 / 28 220%

US009852215B1

**United States Patent**  
Sullivan et al.

(10) Patent No.: **US 9,852,215 F**  
(45) Date of Patent: **Dec. 26, 201**

(12) **United States Patent**  
Sullivan et al.

(54) IDENTIFYING TEXT PREDICTED TO BE OF INTEREST

(71) Applicants: Todd H. Sullivan, Seattle, WA (US);  
Thomas F. Dimson, Seattle, WA (US)

(72) Inventors: Todd H. Sullivan, Seattle, WA (US);  
Thomas F. Dimson, Seattle, WA (US)

(73) Assignee: Amazon Technologies, Inc., Seattle, WA (US)

2005/060643 A1*	3/2005	Glass et al.	715/50
2007/0073745 A1*	3/2007	Scott et al.	707/
2007/0150801 A1*	6/2007	Chidlovskii	G06F 172/715/
2007/0282824 A1*	12/2007	Ellingsworth	70
2008/0107338 A1*	5/2008	Furmaniak	G06K 9/00/382/
2008/0114756 A1*	5/2008	Konig	G06F 17/30/
2008/0228749 A1*	9/2008	Brown	G06F 17/30/
2009/0094178 A1*	4/2009	Aoki	706
2009/0157572 A1*	6/2009	Chidlovskii	G06N 99/1706
2010/0145927 A1*	6/2010	Kasbekar et al.	707/

Sections:  
 Front Page  
 Drawings  
 Specifications  
 Claims  
[Full Document:](#)

Figure 4. The PDF version of the same patent.

document layers of the digitization stack. Consequently, the USPTO has prioritized a sense of print-based ‘documentness’ over the potential for data to be rendered in different, more accessible, formats such as reflowable HTML with all media and data included.

## 6. The importance of format

The role of formats as the fulcrum between digitized media and platforms cannot be understated (Jancovic, Volmar and Schneider 2020). The choice of format can make an outsized impact on how audiences interact with the digitized materials and how they might be able to reuse it in novel ways. For example, Project Gutenberg’s insistence on plain text files has ensured the corpus is a foundational data source for digital humanities projects while digital library books that are only browsable as image facsimiles via a web browser are far more restrictive. The USPTO went through three major shifts in its format. The initial digitization push in the 1970s and 1980s worked with APS, a proprietary format optimized for information retrieval systems with little consideration for the print versions. This was followed up in the 1990s with a move towards Standardized General Mark-Up Language (SGML), the precursor to HTML used across the publishing industry. Finally, the USPTO adopted XML which was flexible enough to work for both print and screen optimization.

The PTO collaborated with ICC between 1970 and 1980 to develop digital workflows for printing and

distributing patents. ICC developed strategies for deploying formats for both *storage* and *output* starting with APS, the first computer-first approach to storing patent data for screen reading. APS encompassed the range of systems and the text/metadata format that end users would interact with through the Messenger system. APS was designed specifically for the USPTO and it was not interoperable, which caused friction in an era of greater international cooperation (White 1985). The PTO costed the system to run using the APS ASCII system for 20 years until 2004 (Sage 1984).

A text-first, informational retrieval-oriented systems came with limitations in terms of how to capture the rich data available in patent documentation. Members of the digitization project referred to ‘Complex Work Units’, or pages that did not just contain text (Nixon et al. 1984: 2–33). While images represented the most common challenge in terms of representation, the patent backfile also contained a range of material objects such as ‘samples consisting of chemically treated fabrics, and coated hardboard and wallboard’ (Glasgow, Passante and Meadows 1984: 2–21). DNA samples and source code often remain excluded from the published digital version, demonstrating the importance of the main body of the text over most ancillary evidence including the images.

Since APS was designed for information retrieval in a computational paradigm with lower storage and bandwidth, allowing users to assess if a patent was relevant to their search quickly was a top priority, and there were on-going debates as to whether the digitized patents should be stored in ‘composed’ (i.e. text) or



‘image’ format (Smith 1987). The rapid return of the first document within 30 seconds was an important consideration for the USPTO in 1986, leading to a focus on text-only rather than a ‘Centralized Image Data Base’. (Carpenter 1986). Beyond these considerations, the results had to be optimized to show the first ‘logical’ page rather than the complete document or the first page of the print publication (Smith 1988). The ‘logical first page’ was an established practice for patent examiners through the *Official Gazette for Patents*, a publication that summarized new inventions each week with a snapshot of the first logical page. Such decisions around what to display first in a digital reading environment continue to be important, as seen in Alan Galey’s discussion of where a Kindle book opens (Galey 2012).

Since APS was focused purely on digital distribution, the typesetting of print patent publications required a separate workflow. In 1984, Glasgow et al noted that ‘the current typesetting process for US patents uses markup and keying from modified original application and amendment pages. These pages contain handwritten additions, changes, and deletions. Entry of the bibliographic data for the first page is more complex due to the several information sources (e.g. file wrapper, application, correspondence)’ (Glasgow, Passante and Meadows 1984: 4–17). The typesetters would first work with a template containing all the markup and then manually key in the relevant data. The shift to SGML in the early 1990s enabled a more effective workflow for printing work as the SGML could be directly converted to PostScript (Klopfenstein 1992).

The conversion of the backfile from APS to SGML to XML is an example of re-born digitization. In April 2004, the Patent Office announced a new solicitation for ‘Patent Data Capture 2 (PaDaCap2)’, which was eventually awarded to Reed Technology and Information Services, the successors of ICC (United States Patent and Trademark Office 2004). An initial contract of a year had options through to the end of 2011 and included the full suite of publishing activities from digitizing incoming material to typesetting publications for digital and physical publication via XML and PostScript respectively. Users of the USPTO backfile do not directly encounter these formats in the presentation of the material, but their logics for structuring patent data, nonetheless, effects how users interact with the HTML or PDF publications and can provide further details for the composition and provenance of digitized materials (Misson and Singh 2022).

The burden of recurrent upgrade cycles and changing standards affects all patent records. Even the so-called born-digital records filed as late as 2000 have undergone at least two changes in data format. Initially, they were created as semi-structured text

records, enduring a brief transition to SGML in 2001, before settling on the emerging XML (Extensible Mark-up Language) standard in early 2002 (United States Patent and Trademark Office 2019a). Both formats act as ‘wrappers’ for metadata and content. Within this context, the wrapper acts as a core intermediary between the underlying document and the platform through providing relevant metadata. Even the PDF standard, which Adobe presents as a facsimile of print, includes extensive metadata in its headers and footers (Eve 2022). The USPTO’s implementations of SGML and XML were initially near identical, but subsequent revisions to the XML specification in 2002 required the patent wrapper metadata to be rewritten to comply with the new standard. While this did not have a dramatic effect on the visible final product, it increased the possibility for ‘data friction’ and conversion errors when moving from one standard to another. Mark-up language syntax—for example, the use of brackets or semi-colons—can be sensitive and a single incorrectly placed punctuation mark could cause cascading error messages.

The facsimile versions of the patent backfile have equally gone through a period of transition. The PTO originally stored the scanned documents as multi-page TIFF (Tagged Image File Format) documents, which was still in common use in federal agencies’ digitization projects in the late 1990s (see, e.g. Puglia and Roginiski 1998). While multi-page TIFF is an uncommon format for end-user consumption, it is an archival standard and was accepted as an ISO standard as early as 1998 compared to the PDF in 2008 (ISO 1998, 2021). There is little visual difference between a PDF and TIFF in this context: PDFs available via the PTO are even reformatted versions of a multipage TIFF generated through the digitization process. There is only a problem when a web browser interprets the two file formats as multipage documents. Most browsers do not natively support TIFFs, requiring a new document to be loaded for every page. By default, users are only able to access one page at a time because of this early format choice despite multipage PDFs rendering directly within modern web browsers. The limitations of a previous format have remained and shaped the platform layer as the USPTO has not updated the interface to meet the affordances of the updated archival presentation.

## 7. Platforms beyond the walled garden

Documents and formats cannot exist in isolation, but they need to be discoverable and often housed on a platform, as indicated by my generalized stack model. As noted earlier, Milligan eludes defining platform in establishing his model. ‘Platform’ is a loosely defined

word that has largely fallen into two overlapping yet distinct definitions. First, as ‘the abstraction level beneath code’ (Montfort and Bogost 2009: 147) that enables a level of platform-specific creativity, most commonly discussed in terms of a video game console. The second, more popular definition refers to social media as a platform as opposed to a publisher (Gillespie 2010). It is tricky to see where the PTO database sits within these definitions, as it lacks social media affordances and the PTO bears responsibility for the content it publishes. While its content has generated considerable creative outputs, this is not due to the interplay between the website and the documents. Nonetheless, PaFT and PPUBS are unique in their way of delivering content and reflect a certain aesthetic, so may benefit some classification as platform. Nonetheless, we need some further nuance to understand the PTO’s offerings as part of a larger patent platforms ecosystem rather than an isolated platform. This follows recent scholarship that attempts to pluralize the concept of platform studies (Apperley and Parikka 2018; Boellstorff and Soderman 2019).

There is an inherent tension between interoperability and the concept of a platform as a ‘walled garden’ (Zittrain 2009) in the case of PaFT, and several similar more open platforms such as Project Gutenberg. I have previously described Project Gutenberg as an ‘anti-platform’, which has the chief intention of creating and distributing public domain texts as ‘spreadable media’ rather than retaining control of the content within a more limited platform (Rowberry 2023). Wikipedia is another notable example of an unwalled garden, albeit without a history of digitization, which Jankowski (2023: 1) has framed as the ‘Wikipedia imaginaire’ due to how its ‘data is woven deep into the fabric of how we imagine the relationship between knowledge and digital culture’. Wikipedia’s content is reused as data for a wide range of projects across the Web, including other open source encyclopedias, Google’s search results, and self-published ebook content.

In a similar manner, due to posting public domain material, the USPTO prioritizes interoperability to introduce a degree of ‘platform spillage,’ where material digitized for one context can be transformed dramatically by a third party. This process is only exacerbated by the rise of Large Language Model (LLM) start-ups such as Open AI who can extract valuable data, often stripping any associated metadata.<sup>7</sup> The platform model is flattened as plain text circulates in new contexts. As authors become more litigious around their re-use in these datasets, it is likely that public domain sources such as Project Gutenberg and patent databases, as well as creative commons content including

open access academic work, will become more important to developing LLMs.

After we have accounted for the slippery notion of ‘platform’ in relation to the PTO, there still needs to be some infrastructure that enables users to find the documents they are interested in, either on the original platform or secondary locations such as Google Patents. This can be achieved through patent classifications. Unfortunately, the PTO’s implementation of classification updates alongside its commitment to accurately replicate the initial print version of the patent with the PDF creates a disconnect between the PDF and HTML version. While the Patent Office updates the classifications on the HTML version, the PDF remains a static, outdated representation. In turn, this generates value for the PTO platform since the HTML stays within the closed system while the PDF version is more spreadable. The accurate metadata are no longer inscribed in the text of the patent itself as per the principles of INID and it is not even associated with the metadata of the PDF files. In some cases, the PDF facsimile of the original patent has a completely different US Classification set to the HTML version. For example, the American Newspaper Publishers Association’s 1984 patent, ‘Method and Apparatus for Digital Serial Scanning with Hierarchical and Relational Access’ (Cichelli and Thompson 1984) initially had a primary classification of ‘370/92’, which has now been replaced with ‘705/30’ with no revision history publicly available. Conversely, the international CPC classifications are only visible on the HTML version but come with a version update.<sup>8</sup> Even though other patent databases display the same facsimile PDF, they also update the metadata to ensure discoverability.

In a commitment to interoperability and breaking down walled gardens within the patent platforms ecosystem, the PTO’s databases prioritize metadata over platform standards. Since there are still remnants of the print infrastructure, the metadata have to be transferrable between the two media, which reduces the possibility for more advanced digital searching tools with the PTO infrastructure. Other patent bodies, such as the European Patent Office, have experimented with more advanced search but this supplements the print-oriented options rather than replacing them. Patent records therefore exist in a liminal space where digitization has altered some aspects of the document while the facsimile format has retained its authority over the platform.

## 8. Conclusion

The PTO’s patent digitization programmes and platforms provide a unique perspective for critical digitization studies as an example of a decades-long project

that looks to convert a complex document format beyond books. Such projects cover multiple paradigms of computational material—from microfilm and early optical storage through to PDFs and web-based search—which allow us to assess the longer history of digital textuality. Even though platforms are seen as foundational within digital projects, they can be more transient when there is less interplay between platform, standard, and format, as is often the case in relatively straightforward text and media formats. The PTO's drive to reduce complexity by not digitizing or recording non-textual aspects of patent applications reduces the importance of the platform and increases the degree of interoperability. Instead, both format and metadata become more important in a long-term view of digital textuality. Good metadata ensure that digitized documents can flow between platforms and survive through platform updates. Formats become a greater stumbling block as they can ossify and become difficult to update without significant labor. The PTO's three main format standards—PDF, HTML, and TIFF—are all mature, open, and backwards-compatible standards, having been initially released between 1986 and 1993. There is no clear challenger to these formats such as the World Wide Web Consortium's failed attempt to create Portable Web Publications, a mixture of an ebook and webpage (Gylling, Herman and Siegman 2015). PDF and HTML will likely remain the default standards for both the PTO and other digitization projects, largely maintaining the current status quo of the platform, and likely digital textuality for the foreseeable future.

### Author contributions

Simon Rowberry (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration).

*Conflict of interest statement.* None declared.

### Funding

This project was funded by the Carnegie Trust for the Universities of Scotland (Grant Number: RIG008587).

### Notes

1. With digital media, even a short period can lead to dramatic change. I started this research in late 2018, and as of writing in mid-2023, many of my initial claims were no longer valid due to system updates.
2. For example, this might take the form of an analysis of the various agencies and third-parties involved in these digitization projects and how this was documented in correspondence available through NARA. There is also a rich body of test logs in the

DB Deliverable Files 1985–1989 collection of the USPTO's NARA archives that would provide valuable insight into the labour behind testing the digitization process.

3. Examples of history of technology and science research that draw heavily from patent archives include Schmookler (1950), Trajtenberg and Jaffe (2002), O'Reagan and Fleming (2018), and Kranakis (2019).
4. The PTO has withdrawn a total of 18,641 utility patents granted before 2001. At launch in 2001, there were 8,702 patents still to be digitised. All these numbers are far lower than the estimates of 100,000 missing patents from discussion in 2002, although that included an additional 380,000 patents dating back to 1971.
5. Images of both the state of the archives and the digitization process can be found in Hlava's slides from the speech her article is derived from Hlava (2014b).
6. This tends to be a core tension across digitization projects that can either choose to accurately render the original documents in digital form or focus on secondary uses.
7. Recent submissions to both Arxiv (Beliveau and Ma 2022; Pelaez et al. 2023; Yoo et al. 2023) and Hugging Face (Suzgun et al. 2022; Cariaggi, 2023) demonstrate that the PTO dataset is actively being used by Natural Language Processing and LLM research groups for training data.
8. The circulation of multiple versions of patent documents, of varying accuracy, is reminiscent of the various versions of records and manuscripts that have become prevalent in open access policies for journal articles.

### References

- Advisory Committee on Application of Machines to Patent Office Operations. *Minutes. Fourth Meeting* (1954) Patent Office Mechanization Study (Nov–Dec 1954). Box 90. Vannevar Bush Papers, Manuscript Division, Library of Congress, Washington, D.C.
- Apperley, T., and Parikka, J. (2018) 'Platform Studies' Epistemic Threshold', *Games and Culture*, 13: 327–48.
- Auyang, H. L. (1994) 'The Electronic Filing of Applications with the United States Patent & (and) Trademark Office', *Hastings Communication and Entertainment Law Journal*, 17: 853–66.
- Bagg, T. C., and Stevens, M. E. (1962) *Information Selection Systems Retrieving Replica Copies: A Start-of-the-Art Report*. NBS Technical Note 157. Washington, DC: U.S. Department of Commerce.
- Bayh, B., and Dole, B. (1980) *Patent and Trademark Law Amendments Act*. Public Law 96-517.
- Beliveau, S., and Ma, J. (2022) 'Recent Developments in AI and USPTO Open Data', <https://doi.org/10.48550/arXiv.2207.05239> accessed 2 Jan. 2024.
- Bellido, J. (2023) 'Patents in Miniature: The Effects of Microfilm as an Information Technology, 1938–68', *Technology and Culture*, 64: 407–33.
- Boellstorff, T., and Soderman, B. (2019) 'Transplatform: Culture, Context, and the Intellivision/Atari VCS Rivalry', *Games and Culture*, 14: 680–703.
- Bratton, B. H. (2015) *The Stack: On Software and Sovereignty*. Cambridge: MIT Press.
- Brügger, N. (2013) 'Web Historiography and Internet Studies: Challenges and Perspectives', *New Media & Society*, 15: 752–64.
- Bryant, J. H., and Stein, D. P. (1983) 'Automated Patent Searching: Preliminary Results of USPTO Studies', *World Patent Information*, 5: 226–9.

- Bush, V. (1936) 'Science in a Changing World', *Journal of the Patent Office Society*, 18: 227–36.
- Bush, V. (1945) 'Patents—A Proposal for Legislation', American Patent Law Association, Box 4, Vannevar Bush Papers, Manuscript Division, Library of Congress, Washington, D.C.
- Bush, V. (1950) 'An Act to Provide for the Dissemination of Technological, Scientific, and Engineering Information to American Business and Industry, and for other Purposes', Public Law 776—81st Congress, Chapter 936—2D session, S. 868.
- Bush, V. (1954) *Report to the Secretary of Commerce by the Advisory Committee on Application of Machines to Patent Office Operations*. Washington, DC: Department of Commerce.
- Cariaggi, F. (2023) 'BERT for Patents', <https://huggingface.co/anferico/bert-for-patents>, accessed 30 Aug. 2023.
- Carpenter, W. (1986) 'Architectural Alternatives to Centralized Image Data Bases', Box 3. Working Papers and Progress Reports relating to System Development and Acquisition 1983-1992 Records of the Patent and Trademark Office, Record Group 241, National Archives at College Park, MD.
- Cichelli, R. J., and Thompson, M. O. (1984) 'Method and Apparatus for Digital Serial Scanning with Hierarchical and Relational Access', United States Patent No. 4,429,385.
- Crymble, A. (2021) *Technology and the Historian: Transformations in the Digital Age*. Chicago: University of Illinois Press.
- Duguid, P. (2007) 'Inheritance and Loss? A Brief Survey of Google Books', *First Monday*, 12, <https://firstmonday.org/ojs/index.php/fm/article/download/1972/1847> accessed 2 Jan. 2024.
- Edwards, P. (2010) *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge: MIT Press.
- Eve, M. P. (2022) 'New Leaves: Riffing the History of Digital Pagination', *Book History*, 25: 479–502.
- Galey, A. (2012) 'The Enkindling Reciter: E-Books in the Bibliographical Imagination', *Book History*, 15: 210–47.
- Gillespie, T. (2010) 'The Politics of "Platforms"', *New Media & Society*, 12(3): 347–64.
- Glasgow, A. H., Passante, E. C., and Meadows, H. E. (1984) *Technical Assessment of U.S. Patent and Trademark Office Data Capture*. MTR-84W137, December. McLean, VA: The MITRE Corporation.
- Gregg, S. H. (2020) *Old Books and Digital Publishing: Eighteenth-Century Collections Online*. Cambridge: Cambridge University Press.
- Grooms, D. W. (1988) 'Quality Assurance of text and image databases at the U.S. Patent and Trademark Office', *Information Services & Use*, 8: 161–5.
- Gylling, M., Herman, I., and Siegman, T. (2015) 'Advancing Portable Documents for the Open Web Platform: EPUB+WEB [Unofficial Draft]', <https://w3c.github.io/epubweb/>, accessed 20 Apr. 2023.
- Helmond, A., and van der Vlist, F. N. (2019) 'Social Media and Platform Historiography: Challenges and Opportunities', *Sound & Vision*, 22: 6.
- Hlava, M. M. K. (2014a) 'Making Information a Business: The Voice Behind the Curtain', *Information Services & Use*, 34: 49–64.
- Hlava, M. M. K. (2014b) 'NFAIS 2014 Miles Conrad Award Lecture', [slideshare.net/accessinnovations/nfais-2014-miles-conrad-lecture-presented-by-marjorie-mk-hlava](https://slideshare.net/accessinnovations/nfais-2014-miles-conrad-lecture-presented-by-marjorie-mk-hlava) accessed 2 Jan. 2024.
- ISO. (1998) 'ISO 12639:1998 Graphic Technology—Prepress Digital Data Exchange—Tag Image File Format for Image Technology (TIFF/IT)', <https://www.iso.org/standard/2181.html>, accessed 17 Feb. 2023.
- ISO. (2021) 'ISO 32000-1:2008 Document Management—Portable Document Format—Part 1: PDF 1.7', <https://www.iso.org/standard/51502.html> accessed 17 Feb. 2023.
- Jancovic, M., Volmar, A., and Schneider, A. (eds) (2020) *Format Matters: Standards, Practices, and Politics in Media Cultures*. Lüneburg: Meson Press.
- Jankowski, S. (2023) 'The Wikipedia Imaginaire: A New Media History beyond Wikipedia.org (2001–2022)', *Internet Histories*, 7(4): 333–353. <https://doi.org/10.1080/24701475.2023.2246261>.
- Kang, H. Y. (2019) 'Ghosts of Inventions: Patent Law's Digital Mediations', *History of Science*, 57: 38–61.
- Kang, H. Y. (2023) 'Patents as Capitalist Aesthetic Forms', *Law and Critique*. <https://doi.org/10.1007/s10978-023-09349-2>.
- Klopfenstein, R. C. (1992) *Transmittal of MITRE Briefing Entitled Joint GPO/PTO/MITRE Meeting on SGML Printing*. Box 16. Working Papers and Progress Reports relating to System Development and Acquisition 1983-1992. Records of the Patent and Trademark Office, Record Group 241, National Archives at College Park, MD.
- Kranakis, E. (2019) 'A Tale of Two Inventions: Monsanto, Biotechnology, and the Geography of Postmodern Science', *Isis*, 110: 701–25.
- Lischer-Katz, Z. (2022) 'The Emergence of Digital Reformatting in the History of Preservation Knowledge: 1823–2015', *Journal of Documentation*, 78: 1249–77.
- 'Marconi Papers', MS Marconi. Oxford, Bodleian Libraries.
- Mak, B. (2014) 'Archaeology of a Digitization', *Journal of the Association for Information Science and Technology*, 65: 1515–26.
- McKitterick, D. (2013) *Old Books, New Technologies: The Representation, Conversation and Transformation of Books since 1700*. Cambridge: Cambridge University Press.
- Milligan, I. (2013) 'Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997-2010', *Canadian Historical Review*, 94: 540–69.
- Milligan, I. (2022) *The Transformation of Historical Research in the Digital Age*. Cambridge: Cambridge University Press.
- Misson, J., and Singh, D. (2022) 'Computing Book Parts with EEBO-TCP', *Book History*, 25: 503–29.
- Montfort, N., and Bogost, I. (2009) *Racing the Beam: The Atari Video Computer System*. Cambridge: MIT Press.
- National Microfilm Association records, University of Michigan Library (Special Collections Research Center).
- Newman, S. M. (1960) 'Information Retrieval Research in the U.S. Patent Office', *Journal of the Patent Office Society*, 42: 731–41.
- Nixon, M. H. et al. (1984) *Automated Patent System Function Requirements*. McLean, VA: The MITRE Corporation.
- O'Reagan, D., and Fleming, L. (2018) 'The FinFET Breakthrough and Networks of Innovation in the Semiconductor Industry, 1980-2005: Applying Digital Tools to the History of Technology', *Technology and Culture*, 59: 251–87.

- Patent Case Files, Record Group 241, National Archives at Kansas City, KS.
- Pargman, D., and Palme, J. (2009) 'ASCII Imperialism'. In: Lampland, M. and Star SL (eds) *Standards and Their Stories: How Quantifying, Classifying and Formalizing Practices Shape Everyday Life*, pp. 177–99. Ithaca: Cornell University Press.
- Pelaez, S. *et al.* (2023) 'Large-Scale Text Analysis Using Generative Language Models: A Case Study in Discovering Public Value Expressions in AI Patents', <https://doi.org/10.48550/arXiv.2305.10383> accessed 2 Jan. 2024.
- Puglia, S., and Roginiski, B. (1998) *NARA Guidelines for Digitizing Archival Materials for Electronic Access*. College Park, MD: National Archives and Records Administration.
- Rankin, W. J. (2011) 'The 'Person Skilled in the Art' is Really Quite Conventional: U.S. Patent Drawings and the Persona of the Inventor, 1870–2005'. In: Biagioli, M., Jaszi, P., and Woodmansee, M. (eds) *Making and Unmaking Intellectual Property: Creative Production in Legal and Cultural Perspective*, pp. 55–75. Chicago, IL: University of Chicago Press.
- Records of the Patent and Trademark Office, Record Group 241, National Archives at College Park, MD.
- Rowberry, S. (2023) *The Early Development of Project Gutenberg, c.1970–2000*. Cambridge: Cambridge Elements in Publishing and Book Culture.
- Sage, K. (1984) 'Automated Patent System Costing Analysis', Box 2, Working Papers and Progress Reports relating to System Development and Acquisition 1983–1992, Records of the Patent and Trademark Office, Record Group 241, National Archives at College Park, MD.
- Schmookler, J. (1950) 'The Interpretation of Patent Statistics', *Journal of the Patent Office Society* 32: 123–46.
- Smith, A. (1987) 'Evaluation of Storing and Retrieving APS Images at a Resolution of 150 Dots Per Inch', Box 5, Working Papers and Progress Reports relating to System Development and Acquisition 1983–1992 Records of the Patent and Trademark Office, Record Group 241, National Archives at College Park, MD.
- Smith, A. (1988) 'Logical First Page Image Retrieval Mode', Box 7, Working Papers and Progress Reports relating to System Development and Acquisition 1983–1992, Records of the Patent and Trademark Office, Record Group 241, National Archives at College Park, MD.
- Smith, D. A., and Cordell, R. (2019) *A Research Agenda for Historical and Multilingual Optical Character Recognition*. Boston: Northeastern University and the Andrew Mellon Foundation.
- Suzgun, M. *et al.* (2022) The Harvard USPTO Patent Dataset. Available at: <https://huggingface.co/datasets/HUPD/hupd> accessed 30 Aug. 2023.
- Terras, M. (2010) 'Digital Curiosities: Resource Creation via Amateur Digitization', *Literary and Linguistic Computing*, 25: 425–38.
- Thylstrup, N. B. (2019) *The Politics of Mass Digitization*. Cambridge: MIT Press.
- Trajtenberg, M., and Jaffe, A. B. (2002) *Patents, Citations, and Innovations: A Window on the Knowledge Economy*. Cambridge: MIT Press.
- United States Patent and Trademark Office. (2002) 'Public Comments Resulting From: Notice of Public Hearing and Request for Comments on the Proposed Plan for an Electronic Public Search Facility', <http://web.archive.org/web/20140228125035/http://www.uspto.gov/web/offices/com/sol/comments/epubsearch/index.html> accessed 2 Jan. 2024.
- United States Patent and Trademark Office. (2004) 'Solicitation DOC52PAPT0410001 Patent Data Capture 2 (PaDaCap2)', <http://web.archive.org/web/20071015224942/http://www.uspto.gov/web/offices/ac/comp/proc/padacap2/padahom.htm> accessed 2 Jan. 2024.
- United States Patent and Trademark Office. (2019a) 'Bulk Data Storage System (BDSS) Version 1.1.0', <https://bulkdata.uspto.gov/>, accessed 11 Feb. 2019.
- United States Patent and Trademark Office. (2019b) *FY 2018 Performance and Accountability Report*. Alexandria, VA: United States Patent and Trademark Office.
- United States Patent and Trademark Office. (2020) 'Database Contents, Patent Full-Text', <http://patft.uspto.gov/netathtml/PTO/help/contents.htm>, accessed 1 Jul. 2019.
- United States Patent and Trademark Office. (1981) *The Story of the United States Patent and Trademark Office*. Washington, DC: Patent and Trademark Office.
- Vannevar Bush Papers, Manuscript Division, Library of Congress, Washington D.C.
- White, M. (1985) *An Evaluation of Selected Alternative Media for International Data Exchange*. McLean, VA: The MITRE Corporation.
- White, M. S. (1986) 'Impact of Optical Disc Technologies on the Storage and Distribution of Patent and Trademark Information', *World Patent Information*, 8: 177–81.
- Worthy, J. (1954) 'Memo: Project to Explore Mechanization of Search in U.S. Patent Office', Patent Office Mechanization Study (Aug–Oct 1954), Box 90, Vannevar Bush Papers, Manuscript Division, Library of Congress, Washington, D.C.
- Yoo, Y. *et al.* (2023) 'Multi Label Classification of Artificial Intelligence Related Patents using Modified D2SBERT and Sentence Attention Mechanism', <https://doi.org/10.48550/arXiv.2303.03165> accessed 2 Jan. 2024.
- Zachary, G. P. (2018) *Endless Frontier: Vannevar Bush, Engineer of the American Century*. New York: Free Press.
- Zittrain, J. (2009) *The Future of the Internet*. London: Penguin.