Digital Textuality

John Lavagnino, King's College London, August 2020

To appear in *The Oxford Encyclopedia of Literary Theory*, edited by John Frow (Oxford University Press, forthcoming)

Summary

Digital textuality has its roots in the most familiar digital system, the alphabet. In defining rules for what aspects of an inscription contain information, the alphabet makes exact copying of writing possible; such exact copying is the fundamental digital characteristic, without which digital machinery could not work. But copyability can have practical limitations, when more complex forms are built up out of basic digital elements: documents, in particular, often assume particular concepts and systems. Digital document systems can be based on many different theories of documents, and typically combine incompatible theories in one document; they also hide considerable amounts of information from users. Very different digital approaches to texts are found in databases, which atomize texts and render all relationships explicit; this degree of formalization is not common in the humanities, but it enables the creation of widely-used research tools (such as library catalogues). The principal innovation in digital documents so far is the hypertextual link, which in connecting texts more closely together created new possibilities for expression and exploration. The creation of vast amounts of digital text led to the unexpected importance of searching, which was made more usable by exploitation of the information provided by links. Searching has overturned ancient hierarchies of importance and attention, by making forgotten texts as accessible as canonical ones.

Keywords

alphabet, Unicode, copyability, markup, standardization, formalization, databases, hypertext, searching

Alphabets

Alphabets are the original digital systems. In the Latin alphabet, every letter A is the same, even if it is written by a different person or with a variant shape or on another surface. It is an A because it is not a B, or any other letter in the alphabet. It is an A despite any variations, because there is a fixed gamut of letters in the alphabet and every letter must be one of them, or it isn't a letter at all. These letters may be used to represent words and sounds, or not: an alphabet (or syllabary, or set of ideographs) works whether the symbols mean a great deal or nothing, as long as there is a defined range of symbols and a way to categorize every mark as one symbol.

Of course, in the everyday use of writing these rules are broken all the time. It is easy to write something by hand that suddenly switches to a different alphabet, or is a picture rather than letters, or demands considerable effort in interpretation to work out what some of the symbols are supposed to be. But readers also normally expect these rules to be followed: the art historian or the typographer may focus on the look of letters, but readers normally do not. And in digital machinery, these rules are not just the usual practice: the hardware is designed to work that way.

One consequence of making alphabets the basis of writing is that exact copies can be made: that if a copy gets the letters and punctuation and spacing of a quotation from Charlotte Brontë right, there is scope for disagreement on the interpretation of her text, but not on what the letters and punctuation are. Readers do not all need to read the same copy of her book. The assumption is different among art historians, who do not consider it adequate to study only photographs of a painting or sculpture, even of a work significantly based on words as Jenny Holzer's are. Art historians assume that different copies of a print are different rather than identical. Some copies of Brontë's texts—her manuscripts, or first editions, or copies she presented to friends, or the copy that is a gift from one's own friend—may have special status as evidence or as talismans, but the convention is to regard those as special cases; they do not invalidate the widespread reading of her works in other copies.

The theory of the alphabet is also the reason why machines can read books, in a limited but useful way. Optical character recognition (OCR) is the source of much of the searchable text available online, particularly that of older printed books: it is successful because any particular book has a comparatively limited range of symbols that machinery can be fairly successful in identifying, without needing to understand anything that the text expresses.

Making copies of copies by photocopying or photography tend to degrade the information; but quoting words does not, as long as it is done with sufficient care. Digital machinery is designed to make copies very reliably, and must do it in order to work; copying happens not just when something is transferred to or from a disk drive, but as a constant part of the machine's internal functioning. Most explanations of digital machinery naturalize the data, but data integrity is a result of work by the machinery to keep it the same. Digital data is artificial. Its basis in a set of discrete symbols makes it possible, with some effort, to support exact copying and to ensure that the data does not change.

By the 1960s the fundamental unit of information in most digital machines was binary, with two possible states; but the essential feature is the existence of a set of distinct and limited set of basic symbols. Digital machines have been built that use decimal numerals and even Roman numerals rather than binary numerals as their basis, and the fundamental units stored in flash memory often have four or eight possible states rather than two; the case has also been made for a three-state rather than two-state system as the best choice.[1] Most operations are on larger groups of these basic units, so that to programmers and users alike it mostly seems as though letters or numbers are the basic units of digital text and not bits; at either level, digital machinery offers a closed set of discrete symbols. Though the apparent

similarity to binary oppositions in thinking has often prompted discussion, nobody has shown that binary representation, rather than discreteness, actually matters. It is the ability to be exactly copied that is the really fundamental digital characteristic.

The internal representation of each character is a pattern of bits, with no natural connection to what the character means or how it looks. There is nothing but the inertia of practice to stop you from using a different pattern for your own characters, and at one time a number of different systems of "character encoding" coexisted. This led to the most common form of error with digital data that is observed in everyday use: *mojibake*, when random-looking characters appear in text that was encoded using one encoding but displayed using a conflicting encoding.

Many early digital systems were monolingual, and expansion to support other languages often still meant a choice of using only one language and encoding at a time. By the 1980s this seemed too restrictive, and work began that led to the creation of the Unicode Consortium, with the aim of creating a single character encoding that would work for every language. The idea of universal support is taken seriously: as of 2020 Unicode included 154 different scripts, and a total of 143,859 different characters.[2] The standard is not just a list of characters; it includes a thousand-page reference on how these characters and scripts work. Unicode does not include every known script or character, but there is continuing research to develop the standard to include further alphabets from history and from the present day. Inertia, the persistence of old software, and some disagreements on approach mean that other systems live on, but its general success means that "all modern writing systems" (and "most historic writing systems") really are available on any system based on Unicode.[3]

This is the technical and practical program of Unicode; but it also had a political program behind its success. Although the Unicode initiative was sponsored early on by many of the biggest companies in the industry, other initiatives with strong backing have failed. The intention was that Unicode would replace the need for all other character sets: it was to be "a superset of all characters in widespread use today".[4] But it also chose to provide straightforward support for earlier character sets, to ensure that conversion was not burdensome. This support for legacy character sets means the new standard was much less tidy than it might have been if designed anew from scratch: there is repetition that would otherwise have been eliminated, and a proliferation of characters that might not otherwise have been included. For example, there are dozens of different asterisk characters, including many small variations in appearance: there is both an "eight teardrop-spoked propeller asterisk" and a "heavy eight teardrop-spoked propeller asterisk".[5] These differences are more like the differences between typefaces, which the standard doesn't deal with, than like distinct characters; but they were inherited from other character sets.

The world of characters for writing languages has an intrinsic limitation that the world of symbols does not: in the pre-digital age the creation of new symbols in printed text was constrained by the burden of creating the necessary materials for printing, but now it is far easier to devise and use new symbols. Pictorial material could easily outstrip the number of

Unicode characters, except that providing a universal set of pictorial images was not a Unicode goal. But in the twenty-first century Unicode began all the same to include pictures, because Japanese character sets expanded to include emoji on mobile phones. Unicode's legacy policy, its inclusion from an early date of symbols from European character sets, and an awareness of the popularity of emoji online, led to its decision to support these characters, though their potential range is essentially infinite.

The eventual result was that the Unicode standard now includes not only a theory of the alphabet, but also a theory of race and gender. It is not a complex theory. “Human-form emoji” may be specified as male, female, or “gender-neutral”. Race is not directly specified, but there is instead a way to select skin color: “people and body parts” may be specified as having one of five skin tones, or left as the default, which should be a “generic non-realistic” tone. This skin-tone scale was originally devised by the dermatologist Thomas B. Fitzpatrick to classify people on the basis of their skin's response to sunlight exposure; it talks about lightness and darkness and not about other variations in appearance. But you can combine it with a “hair component”: red, white, curly, or bald; “person with blond hair” already has a separate symbol, and dark hair is the usual default.[6]

At the same time, “the general recommendation is to be as neutral as possible regarding race, ethnicity, and gender”; and representation is assumed to be “in a colorful cartoon form” rather than highly realistic.[7] Unicode's approach emerges from conflicting goals: acknowledging diversity but also simplifying it. But all the same, you can choose these possible variations in the representation of human beings on systems that implement Unicode properly, whereas there is no way to select what kind of saxophone the saxophone emoji depicts; the consortium recognizes that race, ethnicity, and gender, now that they have a connection with the implementation of the character set, have to be taken account of. It is still a character set, though. The alphabetic approach fundamental to Unicode cannot escape absolute distinctions and enumerable characteristics, even when it is referring to phenomena that are open, overlapping, and changing.

Documents

There is always more to texts than just characters. Unicode has only limited support for specifying letters of different size or design, always stemming from its incorporation of legacy character sets; it does not offer to do anything at all about the separation of one text from another, spatial arrangement to indicate kinds of textual material, division of texts into sections, and many other such elements of significance. These are all regarded by the Unicode Consortium as issues for the document level, a level that necessarily embraces a wide range of expressive possibilities and approaches to text, rather than the restrictions and sharply-defined distinctions of character sets. Before the digital world, inscribed texts were called manuscripts or papers or books; in the digital world they are files or documents. The usual belief is that digital files or documents have no physical form, even though their physical form, or at least the way they look, still dominates thinking about them.

Digital documents have two features rarely found in paper documents: they can express an explicit analysis of the text into parts, and they can also hide material from readers, who may be able to detect its presence from the document's behavior in reading machinery but will typically find it hard to inspect directly. These features are the reason for the great multiplicity of ways to represent documents as opposed to characters, and the difficulties often encountered in conversion from one form to another. There is one theory of the alphabet, but many theories of documents.

Users of nearly every kind of word-processing software build up their documents from characters, though versions of this explicit and machine-processable way of doing it are not the only way to do it: images of text still work for readers even without any analysis of whatever looks like writing in those images into letters and words. Some texts consist exclusively of characters, as is the case with text messages and tweets, but the document level normally adds markup. In its most familiar form, markup specifies presentational features such as spatial layout and italics, features about the way you want it to look when it is presented for display and reading. Both layout and type style can get far more complex than these simple examples, and in practical use are closely bound to the capabilities and limitations of particular display systems.

But there are other ways to conceptualize a document and its associated markup. Markup can specify document features in a descriptive way, for example, with direct specification of what the parts of the text are—as in this article, with its title, summary, and subsections—rather than how they look: this is important for online publication, where one text may need to be displayed very differently on a phone and on a laptop. Descriptive markup involves a significantly different approach to the text, because one presentational feature can have many meanings. Italics, for example, are freely used for very diverse things, such as emphasized text, book titles, words in other languages, names of ships, mathematical variables, and the unspoken thoughts of serial killers. For readers, interpreting the text is in part interpreting the presentation, and not only the words. But descriptive markup can record a specific reason why some text is in italics, while still supporting the conventional display. Where presentational markup becomes complex because the technical capabilities of display systems are extensive, descriptive markup runs into the problem of the enormous range of expressive uses even of a simple feature like italics: it is an approach often most effective in a restricted domain rather than in creating something that will work for every kind of document. Both kinds of markup therefore raise problems of translation from system to system. And present-day systems do not usually force a choice of one theory or the other: a free mixture of both approaches is common.

The appearance of text in many cases does not diverge from the traditions of print: even highlighting to mark links is only a development of established methods for marking cross-references. But if writers are not typically doing much to expand the displayed analysis or articulation of the text, the digital document often includes much more that is hidden. Digital documents all have metadata associated with them, data that is about the document but not part of what is normally read: unlike paper documents, digital documents always have names,

and digital systems usually record dates of creation or modification. Digital documents can also contain their own history: version control systems make it possible to record extensive information about document evolution, in principle down to recording the date and time of every character added or changed. The informational content of digital documents is often vastly greater than what is visible in the version displayed for reading.

A particularly common kind of document is the digitized version of a pre-digital book that contains the text twice: as page images and as text mechanically transcribed by OCR. Copying a section of such a text can often produce something different from what the displayed page says, because the display is based on the image but the copied text is the OCR transcription, often imperfect. Most popular word processors since the late twentieth century present themselves as WYSIWYG systems, in which "what you see is what you get". But all such systems in fact make it possible to include extensive information in documents that readers do not normally see, and may have no way to examine. The document that contains its own revision history and identifies all the different people who worked on it over time represents much more than is apparent to readers of the completed text. The email message that silently reports back to its author when it is read, or the document whose content changes depending on the time or place of reading, usually do not explain to the reader their connections to the reading environment. The techniques and effects of tracking and advertising online to tailor what particular people see are not so distinct from methods available within individual documents. A familiar consequence of hidden data is the problem of deleted text that persists when document authors assume it was gone forever, and that comes back to embarrass them.

Printed texts can hide things too, but more through ciphering techniques such as acrostics, or through literary expression: texts without markup can still suggest more than they say directly. But the kind of extensive logging of work on a text that digital files easily include is much less common with paper, and also has to be much more visible; extra information has to have paper to support it, and more of it when there's more information. Hidden information in paper documents is more likely to be visible but incomprehensible, rather than simply invisible.

Although readers and writers seem little concerned about the specific perspectives represented by markup, they are a problem for moving documents from one system to another. A descriptively marked-up text that identifies elements as titles, or as anything else, has a different kind of information from a presentational one that simply specifies italics, and cannot be converted to that form without losing something. But it may also be hard to convert to a different descriptive scheme that incorporates a different set of concepts: some descriptive schemes include lines and stanzas of poetry, some do not. Limiting markup to the presentational level does not solve the problem, because not all display systems have the same capabilities. Many readers and writers never find this a problem, if they are consistently using one word processor and not trying to move documents between systems, or are not attempting to publish the same material both in print and on the web, or are relaxed about loss or change of details.

A basic reason why people wanted word processors in the first place was that they made documents dynamic: you would never need to retype a whole page because you revised part of it. The copyability of the alphabet made that part possible. But to support further features—document structure and everything else above the level of the alphabet—you needed a great deal that was hidden from everyday inspection. Moving from the world of the alphabet to the world of the document is moving from a world of deliberate restriction that enables digitality, to a world closer to the whole range of expressive possibility and ways of thinking about text—but also much more dependent on particular systems.

Databases

There is another tradition of digital text, with its own goals, theorists, practices, and software, descended from different intellectual and institutional spheres. Database systems developed from the world of account books and other financial records, from ledgers, catalogues, tallies, and lists: from all kinds of recordkeeping where the item rather than the sentence is the primary element. In popular discussion, it's common to hear any kind of digital collection called a database, but the specific expertise of database practitioners is in collections constructed and operated in a very particular way: as with the term “archive”, professionals work with a much narrower sense. Word processors have minimal capability to build a database; spreadsheets are optimized for building very rudimentary databases; real database systems are much more rarely used directly by the general public than either of these forms of software, though they work unseen behind a vast range of widely-used applications and web sites.

A scholar’s monograph about a film director might contain a list of films and their performers, writers, and other collaborators; and searching the text of such a book would be one way to find, for example, all that director’s films in which a particular actor appeared, though the task would require reading the text to extract the details and eliminate extraneous material, such as suggestions that a performer was considered for a role but rejected or comparisons with other performers. A good film database can answer that kind of question at once, without requiring any further work on the results: the Collaborations function of the online Internet Movie Database can list the films any pair of people were both involved in. Databases can provide exact answers to complex queries that work with their carefully-constructed data.

Databases still typically contain many words even if the sentence does not matter. But their approach to language, and to everything else, is based on atomization and explicitness. A sentence combines many references to things, ideas, actions, and their interactions; the database method is to separate nouns and verbs, break everything down into small parts, and represent the connections of these items explicitly. A monograph might say “Sally Potter directed *Orlando*”; a database would analyze that into two items, Sally Potter and *Orlando*, along with the relationship of authorship that identifies Potter as director of the film. Uncertainty is not an obstacle to the database approach, but it needs to be defined. The qualification in the sentence “*The Revenger's Tragedy* is most likely by Thomas Middleton”

would be captured in a database by identifying a different relationship: “is probably by”. The shades of meaning that different ways of putting it might suggest—“could be by”, “is probably by”, “is almost certainly by”—either vanish in database representation or must be explicitly distinguished. They don't have to be ranked: you don't have to define the relationship of “could be by” to “is probably by”, but usual database practice would be to define and rank those shades of meaning so that you can do more with the data.

Atomization is essential, but the degree of atomization depends on the database's intended use. Even the name “Sally Potter” is too large a unit for some purposes: if you wanted to provide an effective and accurate search by last name, you'd need to identify and separate the first and last name components of the name and store them as separate items. That analysis is still too simple for most real populations, which have names with more components (middle names, possibly several; various kinds of titles and suffixes), possibly in different orders, and possibly changing over time. Any practice that you want to take account of has to be designed into the database.

In this way, databases routinely involve more work to create and maintain than documents: they require a very disciplined approach to conceptualizing and organizing information, in the same way that an account book is very different from a stack of receipts and notes on expenses. The database is above all a very heavily processed form of information. This is poor for tentative and exploratory work, but powerful for analysis of a known domain. The simple task of finding the name of *Orlando*’s director does not really need database infrastructure; but analysis or filtering with more complex specifications, such as collaborations or limitations by date or language, is what the technology excels at. Databases are not typically designed for undirected browsing or extensive reading; the normal pattern of use involves asking a question and reading the report that is generated, a selection and display that nobody else may have ever seen.

Move beyond what looks like safely factual data, such as names, and databases are still valuable for humanities scholars but also unnerving: the formalization of ideas and concepts has to go farther than is otherwise customary. Art historians, for example, have created a system called Iconclass to support art databases; it defines an organized set of codes for “subjects represented in images”—something much broader than the set of Unicode emoji, but still not quite the same as a description of everything you could possibly look at, because the classification responds to what you actually find in art around the world and the way it is discussed by scholars.[8] Out of ten primary divisions, three are “Religion and Magic”, “Bible”, and “Classical Mythology and Ancient History”. But there are also “Abstract Ideas and Concepts”, including, for example, “the Universe”—a notation within the subdivision “Existence”—which can be qualified with keys to indicate its use as personification, allegory, symbol, or emblem.[9] This system unavoidably recalls the “Chinese encyclopedia” invented by Jorge Luis Borges, and discussed by Michel Foucault as illustrative of the contingency of classification.[10] But even when knowledge and analysis of the universe are incomplete, a classification that responds to features of the domain and ideas about it is useful—above all for enabling people simply to find things they can then proceed to reflect on. For practical

purposes what matters is having a way to group together everything that symbolically represents the universe, or any other artistic motif, and make it findable.

This world of digital practice seeks to eliminate connotation and implicit information as much as possible; one original motivation was to eliminate the problem of data that could not be moved to a new system without breaking. The copying problem that exists with documents was a problem for early digital databases, and going even farther with explicitness was the solution. From the document point of view, databases are inadequate or impossible most of the time: turning sentences into sets of explicitly-defined items and relationships can be a vast task even for simple material, and sentences typically do not specify everything completely enough. That the information in documents is partial, implicit, and contradictory is a valuable feature. A common method for handling documents in databases is to process the metadata, with its appropriate items such as titles, authors, dates, and times, and then to treat the actual text as one big chunk, because the database machinery can't really do much with it. This is the approach of many content-management systems, databases used to manage the content for web sites.

To the database practitioner, document systems are slow and limited in their features. To document experts, databases are missing the point: overformalized, not really suited to sentences. To non-experts, both kinds of system, and indeed their free combination, seem useful but full of puzzling constraints.

Networks

Much of the work in digital document systems was devoted to reproducing the world of conventional paper documents, while making some things easier to do: revision, in particular. But even before the digital age, there were proposals for new ways to organize and connect documents, and this line of thinking led to the first true prophet of hypertext—indeed, the creator of that word, Ted Nelson, who starting in the 1960s insisted on the revolutionary nature of properly hypertextual systems. He was an eccentric visionary, rarely assimilated for very long by academic or commercial organizations, who turned out to be right. What makes texts into hypertexts is linking: making references to other places in the text active, requiring the most minimal effort to go to one of those places. The effort to do this with a printed book might often be minimal, too, but the evidence of practice is that being even easier makes a difference. Links can be as various in their function and meaning as the kinds of reference in print are. But they can go still further: they do not have to be visible, or always refer to the same destination, or only point to other texts. Nelson particularly wanted to support the creation of "a body of written or pictorial material interconnected in such a complex way that it could not conveniently be presented or represented on paper".[11]

Nelson's fundamental interest was in the thinking and exploring individual, creating new thoughts and responding to those of others: he was not talking only about literary expression. The most familiar form of hypertext now lives in the kind of practical space that Nelson often imagined: focused on the usual aims of nonfictional writing but with further possibilities for

organization and connection. Its dominant form, the World Wide Web, was originally designed as a vehicle for “information management”, in the context of a large scientific laboratory. The rationale stated by Tim Berners-Lee in 1989 for his invention was not about expression, but about supporting the kind of decentralized, nonhierarchical working practices that were the norm in such a research organization. Hierarchical systems of organizing and retrieving documents were not really effective, just as hierarchical organization was inadequate to produce the work that actually needed doing. This repeated an idea found earlier in Nelson and other advocates of hypertext: that it actually fit how people thought and worked better than existing ways of organizing texts and knowledge. But Berners-Lee imagined a system with far less linking than Nelson did, one that was a way to organize and publish quite conventional documents.[12]

That it was the World Wide Web that got so big, rather than any of the other systems developed in the period, is due to its simplicity rather than its sophistication, and to its ready availability: it was free, and clients and servers alike could be run on many different systems. It appeared at a time when people outside the business and academic worlds were just starting to use networking extensively. Later, in the era of social media, it became a commonplace that users created all the content, but in the initial period of the web that was also what drove growth; commercial ideas about exploiting networking in the first years of the web had been mostly built on the model of broadcast TV rather than user contribution. As the set of web sites expanded, it quickly became the largest and most interactive resource available in the online world, leading to a new kind of experience: that of infinite connection. You could keep following links forever, it seemed: not just to accomplish the kind of practical tasks that were initially assumed as your goal, but also for free exploration and endless pursuit of things that just looked interesting. In this respect hypertext went beyond the initial, rational idea about representing associations and had a place for speculation and fantasy as well; it mirrored the mind better than it had intended.

In the same period as the rise of the web, diverse experiments with the literary possibilities of hypertext flourished, and with very different goals, with the mind's imaginative capacities as their center rather than as an accidental consequence. It was a world that tried to widen possibilities rather than to develop a comfortable norm, and instead of making information accessible it thought about what the link might express: what could inaccessibility or changing destinations mean? It was also very interested in nonlinearity: far more than print, digital media could offer a changing experience. Though there has long been an active digital-poetry world, and many digital works that had no narrative basis, it was hypertext fiction that attracted widespread attention in the 1990s—with somewhat unfortunate results, because the very traditionalist world of mainstream fiction did not welcome works designed to disrupt storytelling and character identity. Some electronic-literature writers moved into the visual-art world, where there was no problem at all if your story had no fixed start or end, and where many more possibilities for representing the active mind were welcomed.

The original idea of the web was that everyone would normally be both a reader and a writer. In the late 1990s it seemed to be in danger of domination by commercial content providers,

but the twenty-first century saw the rise of successive forms of social media that once again gave everyone a place to write and publish. As activity moved from blogs, to Facebook, to Twitter and Instagram, the tools became easier to use, and the emphasis on interaction and linking became stronger and stronger. This was not hypertext as representation of the mind, but as representation of conversation. You could certainly craft tweets that were perfect miniature poems, in this system that restricted utterances to a few dozen words; and commercial organizations vigorously pursued the possibilities for advertising their activities. But the dominant use was in conversations where the exchange was what was meaningful, not the individual tweet.

That interest in conversation was so strong that it led to an important example of innovation by users. It was they who popularized the hashtag, on Twitter and elsewhere, the practice first proposed by Chris Messina of identifying topics by a hash sign followed by a label: #digitaltextuality. Twitter's managers were not interested, until the practice became widespread and they found it expedient to support hashtags in their software; the users understood the service better than the service provider.

Nelson had imagined collaboration as one of the fundamental activities in hypertext systems, but in the social-media world collaboration worked by aggregation: responses accumulated, but were not often collected and revised by groups of writers into new or larger texts. The leading online work developed through collaborative writing was Wikipedia, which focused primarily on composition of encyclopedia articles and deemphasized individual voices and contributions—even while recording in great detail the development of articles, and providing spaces for discussion without promoting their existence to readers very much. Its scale, its openness to contributors from anywhere, and its prominence in the online world were all exceptional; but collaborative writing online became more widespread with the rise of cloud-based word-processing platforms such as Google Docs.

Machines

In principle, following links requires no machinery: scholars have long looked up references to other books. But in practice, minimizing the effort to move to some other place makes a difference. The same is true for searching massive amounts of text, the other fundamental operation transformed by the digital world. Even before the rise of print, there was a desire for indexes to all the words in texts, known as “concordances”, but the cost of producing these meant they were created for only a few texts of particular importance, most often religious texts. In the 1960s, there was a great increase in the production of concordances, using early digital technology, but they still only covered a few hundred chosen works. But as digital storage and interactive access became cheaper and more widespread, and more and more texts were created in digital form or converted to it, word searching became a routine rather than a rare possibility.

That this makes a difference was not widely recognized until it had happened. Although early hypertext systems had often included searching, many discussions of hypertext barely

mentioned it, and work on searching mostly happened in a different research area called "information retrieval". One hypertext system that lacks searching is the World Wide Web: Berners-Lee in his original proposal does not mention searching, and all web-searching systems are later additions, not incorporated into the web at a fundamental level but operating as unprivileged observers. During most of the first decade of the web, experts assumed that the normal starting point would be an online guide or directory; as with hashtags, users proved to have different preferences, particularly once the web became so large that manual indexing could not keep up.

Web searching had the limitation that it was a search for strings of characters—and this is still what it fundamentally is, despite the efforts by operators of major search engines to recognize common patterns of use (such as searches for film times or popular products), or to correct misspellings, and tailor results to the intended goal. But as information-retrieval research had found, it was difficult to make word-based searching select the most relevant documents out of a large set. Once the web became big, many searches matched so many documents that some sort of ranking and filtering was necessary to make the results useful: not every mention of Charlotte Brontë was of equal significance. But how was a machine to know which documents were more significant?

The vast Google empire began with its founders' recognition that the web already knew. But it wasn't the text that told you what mattered; it was the links. If you ranked web pages based on how many other web pages were linking to them, you had a measure of significance based on human activity. The original idea—the "PageRank" algorithm—has required endless development and adjustment, especially as further human activity took place in attempts to trick the system; but analyzing the link structure of the web remains one of the principal ways of judging significance.[13] It means, though, that search results widely regarded as reliable and objective are very far from being that. They can change dramatically over time, not just because the contents of the web change, but because the techniques used by search engines to filter them changes. Particularly as people came to ask about broad subjects, such as "*Jane Eyre* and colonialism", the results depended greatly on the system. Web searching is an oracle: it provides answers, but the relationship between question and answer is inscrutable.

The link structure is full of valuable information about what has seemed significant to the community of users, and about associations that those users made between different pages. But web searching still has many limitations, and is usually an approximate way to get at what is wanted; another reason it has worked better than what information-retrieval experts expected, though, is that it is fast and cheap. This means that iterative attempts to find something are practical, in a way they had not been in many earlier full-text systems, which often charged by the query. But many kinds of searching are poorly supported. Image searching systems online still work primarily by indexing text around images, rather than images themselves. Two other great limitations are in searching for concepts such as colonialism that might be expressed using many different forms of words, and in searching for material from particular places or times. In the end, searching depends above all on the alphabet: because texts online are mostly analyzed already into letters, they are searchable,

and it takes far more work to identify the elements of images, or to determine the date of any text when it is not expressly specified in metadata, to support these other forms of access. And beyond these limitations is the informational bias of the web's content, which, exactly like printed output, is far more extensive as a record of writing and creation in richer countries.

Specialized online collections exist to get around some of these problems: collections of specially indexed images, as in some art-historical sites, or of text such as newspapers with searching that knows the dates. Web searching is so dominant, though, that every other kind of searching has been made to resemble web searching in the first instance. Library catalogues are one kind of database that can be searched exactly and reliably for works by (for example) a particular author; but almost all such library catalogues now offer by default a web-style interface, because it is the more familiar thing, and it usually works well enough.

At the same time, one common kind of query, at least among literary scholars, is for a particular form of words—and in this case many limitations of web searching go away: what is needed is precisely that phrase and no other. Here the shaping effect of search-engine filtering is much less strong: this is an area where it is possible to get five results for many queries rather than millions. In this specific kind of use, web searching is not an oracle but a database: the effects of filtering mostly vanish and there is a reliable answer instead, limited only by the available contents of the web. Lexicographers got here first: long before the web, full-text searching on large collections became the preferred foundation for writing new dictionaries. But their methodology has spread throughout the humanities.

This represents a large but little-discussed change in our relationship to written texts. What was formerly hidden in the vastness of printed output is now often easily accessible; kinds of research that were once unthinkable are now routine. In the era of concordances, word indexing was reserved for the most revered texts; in the era of web searching it is very much the same for those texts, for long-forgotten books, and for many current online conversations. The long-standing tendency to privilege canonical texts, and to identify them as sources much more readily than little-known texts, is no longer so thoroughly embedded in tools for research. The obscure quotation and the forgotten author are greatly diminished phenomena. ("In an age of search engines, misquotation and misattribution are largely psychological matters. It has become very easy to check if you try, but not everyone tries—self-reliance hardening into an inability to consider the possibility of being wrong."[14])

This has transformed academic research in the humanities, particularly since the launch of Google Books in 2004, which added millions of digitized print books to the web: there is a boom in reference to works that nobody knows about and nobody has read, because now the machine has read them. In literary studies, these references usually come without any mention of how it was possible to find something so obscure; historians are much more likely to explain their use of digital resources. The same phenomenon now operates in life generally, where messages intended for a narrow audience, published in exclusive or obscure

ways, become widely visible and constitute part of a public figure's profile. At one time the world would have waited decades for a biographer to discover such things.

But if a name or phrase has even a slight change, a search may never turn it up; and a spot on the page may prevent the machine from reading a crucial word correctly when the OCR operates. The increasing reliance on web searching rather than manual indexing to guide readers to material is only a further step in our dependence on our research tools. The common assumption of the universal accessibility of everything online is not justified; but the process of reading now centers less on the person and the page, and more on the person's interaction with a vast digital system.

Discussion of the Literature

The case for the alphabet's digital nature was most thoroughly made by Nelson Goodman; among many competing accounts of the fundamentals of digital representation, the most valuable is by John Haugeland.[15] The effort required to keep digital data from changing is rarely discussed in popular accounts: one of the few to do this is W. Daniel Hillis's.[16] The extensive documentation published by the Unicode Consortium is the best starting point on present-day approaches to digital character representation and specifics of the Unicode standard.[17] Deborah Anderson and other associates of the Script Encoding Initiative at Berkeley have written in more detail about the work of extending the standard to support more scripts.[18]

Arguments over document representation from the humanities point of view have turned partly on issues of practicality and utility, but more crucially on theories of texts.[19] The specific topic of remediation, and the nature of the relationship between the paper and digital worlds, has been well analyzed, but the consequences are not widely recognized.[20] Matthew G. Kirschenbaum's work has combined critical reflection with detailed study of the real workings of digital systems and their use by writers and readers.[21]

David C. Blair offered one of the best discussions of the differences between document and database approaches.[22] Lev Manovich made an influential case for databases as a significant expressive form.[23] The later discussion in *PMLA* turning on this issue usefully isolates a number of conflicting positions.[24]

The early discussion of something like a hypertext system by Vannevar Bush in 1945 is still cited and discussed, but it was Ted Nelson's contributions that made the case for digital hypertext systems and inspired many.[25] The topic of hypertext had a surge of activity in the 1990s, following on from the impetus of Nelson's advocacy, the inspiration of poststructuralist theory, and numerous practical implementations, with work along several lines: literary-critical approaches, writerly perspectives, and practical advice for utility and effectiveness.[26] With time the focus shifted to electronic literature more generally rather than hypertext specifically, and indeed much interesting work was more connected with the

visual-art world than with the literary world.[27] Historical accounts of the field are also now appearing.[28]

The library-science literature on text searching is extensive; Marcia J. Bates's work in the pre-web era remains valuable for its close analysis of behaviors and uses.[29] Evaluations of the possibilities and limitations of full-text collections and searching, and discussions of a theory of their use, remain too few.[30] Renewed work on the historical study of keywords, reviving Raymond Williams's approach, offers one way to think more deeply about these practices.[31]

Links to Digital Materials

Emojipedia, an independent reference on Unicode emoji.

FileFormat.info, a reference website on the vast range of formats for documents and other kinds of digital data.

Iconclass, a classification system for subject matter in visual art: main page for the project; browsing and searching interface; RKDImages, an Iconclass-indexed online collection at RKD–Nederlands Instituut voor Kunstgeschiedenis.

IMDb, a commercial film database, with some advanced searching features that demonstrate what databases can do.

Script Encoding Initiative, for the development of Unicode encoding for writing systems not currently supported.

The Unicode Consortium.

Further Reading

Barnet, Belinda. *Memory Machines: The Evolution of Hypertext*. London: Anthem Press, 2013.

Bolter, Jay David, and Richard Grusin. *Remediation: Understanding New Media*. Cambridge: MIT Press, 1998.

Caton, Paul. “On the term ‘text’ in digital humanities”. *Literary and Linguistic Computing* 28, no. 2 (June 2013): 209–220.

Dechow, Douglas R., and Daniele C. Struppa. *Intertwingled: The Work and Influence of Ted Nelson*. Heidelberg: Springer, 2015.

Funkhouser, Christopher T. *Prehistoric Digital Poetry: An Archaeology of Forms, 1959–1995*. Tuscaloosa: University of Alabama Press, 2007.

Gottlieb, Nanette. "Technology and the Writing System in Japan". In *Language Life in Japan: Transformations and Prospects*, edited by Patrick Heinrich and Christian Galan, 140–153. Abingdon: Routledge, 2011.

Gray, Jonathan. "Text". In *Keywords for Media Studies*, edited by Laurie Ouellette and Jonathan Gray, 196–200. New York: New York University Press, 2017.

Hayles, N. Katherine, and Jessica Pressman, eds. *Comparative Textual Media: Transforming the Humanities in the Postprint Era*. Minneapolis: University of Minnesota Press, 2013.

Hayles, N. Katherine. *Electronic Literature: New Horizons for the Literary*. Notre Dame: University of Notre Dame Press, 2008.

Higdon, David Leon. "The Concordance: Mere Index or Needful Census?" *TEXT* 15 (2002): 51–68.

Kirschenbaum, Matthew G. *Mechanisms: New Media and the Forensic Imagination*. Cambridge, MA: MIT Press, 2008.

Kirschenbaum, Matthew G. *Track Changes: A Literary History of Word Processing*. Cambridge, MA: Harvard University Press, 2016.

Lang, Anouk, ed. *From Codex to Hypertext: Reading at the Turn of the Twenty-first Century* Amherst: University of Massachusetts Press, 2012.

Lavagnino, John. "Digital and Analogue Texts". In *A Companion to Digital Literary Studies*, edited by Ray Siemens and Susan Schreibman, 402–414. Oxford: Blackwell, 2007.

McGann, Jerome. *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*. Cambridge, MA: Harvard University Press, 2014.

Putnam, Lara. "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast". *American Historical Review* 121, no. 2 (April 2016): 376–402.

Unicode Consortium. *The Unicode Standard: Version 13.0.0—Core Specification*. Mountain View, CA: Unicode Consortium, 2020. https://www.unicode.org/versions/Unicode13.0.0/.

Notes

[1] Brian Hayes, "Third base", *American Scientist* 89, no. 6 (November–December 2001): 490–4.
[2] Unicode Consortium, "Unicode 13.0.0", last modified March 10, 2020, http://www.unicode.org/versions/Unicode13.0.0/.

[3] Unicode Consortium, *The Unicode Standard: Version 13.0.0—Core Specification* (Mountain View, CA: Unicode Consortium, 2020), 14, https://www.unicode.org/versions/Unicode13.0.0/.
[4] *Unicode Standard*, 3.
[5] Unicode Consortium, *The Unicode Standard, Version 13.0: Archived Code Charts*, last modified March 10, 2020, https://www.unicode.org/Public/13.0.0/charts/CodeCharts.pdf.
[6] Mark Davis and Peter Edberg, "Unicode Emoji", Unicode Technical Standard #51 (Mountain View, CA: Unicode Consortium, 2018), http://www.unicode.org/reports/tr51/tr51-14.html.
[7] "Unicode Emoji".
[8] "Iconclass", accessed August 15, 2020, http://www.iconclass.nl/home.
[9] "Iconclass Browser", accessed August 15, 2020, http://www.iconclass.org/rkd/51A11/.
[10] Jorge Luis Borges, "The Analytical Language of John Wilkins", in *Other Inquisitions 1937–1952*, trans. Ruth L. C. Simms (Austin: University of Texas Press, 1964), 103; Michel Foucault, *Les mots et les choses: une archéologie des sciences humaines* (Paris: Gallimard, 1966), 7.
[11] T. H. Nelson, "A File Structure for The Complex, The Changing and the Indeterminate", in *Proceedings of the 20th National Conference* (New York: Association for Computing Machinery, 1965), 84.
[12] Tim Berners-Lee, "Information Management: A Proposal", March 1989, https://www.w3.org/History/1989/proposal.html.
[13] Larry Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Stanford InfoLab Technical Report, 29 January 1998, http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.
[14] Adam Mars-Jones, "Mrs Winterson's Daughter", *London Review of Books*, 26 January 2012, 8.
[15] Nelson Goodman, *Languages of Art: An Approach to a Theory of Symbols* (Indianapolis: Bobbs-Merrill, 1968); John Haugeland, "Analog and analog", *Philosophical Topics*, 12, no. 1 (Spring 1981): 213–25.
[16] W. Daniel Hillis, *The Pattern on the Stone: The Simple Ideas that Make Computers Work* (New York: Basic, 1998).
[17] Unicode Consortium, *The Unicode Standard: Version 13.0.0—Core Specification* (Mountain View, CA: Unicode Consortium, 2018), https://www.unicode.org/versions/Unicode13.0.0/; Mark Davis and Peter Edberg, "Unicode Emoji", Unicode Technical Standard #51 (Mountain View, CA: Unicode Consortium, 2018), http://www.unicode.org/reports/tr51/tr51-14.html.
[18] Deborah Anderson, "Unicode and Historic Scripts", *Ariadne* no. 37 (30 October 2003), http://www.ariadne.ac.uk/issue37/anderson; Script Encoding Initiative, Department of Linguistics, University of California at Berkeley, home page, http://linguistics.berkeley.edu/sei/.
[19] James H. Coombs, Allen H. Renear, and Steven J. DeRose, "Markup Systems and the Future of Scholarly Text Processing", *Communications of the ACM* 30, no. 11 (November 1987): 933–947; Allen H. Renear, Elli Mylonas, and David G. Durand, "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies", *Research in Humanities Computing* (1996), 263–280; Jerome McGann, "Marking Texts of Many Dimensions", in *A Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, and John Unsworth, 198–217 (Oxford: Blackwell, 2004); Desmond Schmidt, "The inadequacy of embedded markup for cultural heritage texts", *Literary and Linguistic Computing* 25, no. 3 (September 2010): 337–356.
[20] Jay David Bolter and Richard Grusin, *Remediation: Understanding New Media* (Cambridge: MIT Press, 1998); Ian Gadd, "The Use and Misuse of *Early English Books Online*", *Literature Compass* 6, no. 3 (May 2009): 680–692; Bonnie Mak, "Archaeology of a Digitization", *Journal of the Association for Information Science and Technology* 65, no. 8 (August 2014): 1515–1526.
[21] *Mechanisms: New Media and the Forensic Imagination* (Cambridge, MA: MIT Press, 2008); *Track Changes: A Literary History of Word Processing* (Cambridge, MA: Harvard University Press, 2016).
[22] David C. Blair, "The Data–Document Distinction in Information Retrieval", *Communications of the ACM* 27, no. 4 (April 1984), 369–374.
[23] Lev Manovich, "Database as Symbolic Form", *Convergence* 5, no. 2 (June 1999): 80–99.
[24] Ed Folsom, "Database as Genre: The Epic Transformation of Archives", *PMLA* 122, no. 5 (October 2007): 1571–1579; Jonathan Freedman, "Whitman, Database, Information Culture", *PMLA* 122, no. 5 (October 2007): 1596–1602; N. Katherine Hayles, "Narrative and Database: Natural Symbionts", *PMLA* 122, no. 5 (October 2007): 1603–1608; Jerome McGann, "Database, Interface, and Archival Fever", *PMLA* 122, no. 5 (October 2007): 1588–1592; Meredith L. McGill, "Remediating Whitman", *PMLA* 122, no. 5 (October 2007): 1592–1596; Peter Stallybrass, "Against Thinking", *PMLA* 122, no. 5 (October 2007): 1580–1587; and Ed Folsom, "Reply", *PMLA* 122, no. 5 (October 2007): 1608–1612.

[25] Vannevar Bush, "As We May Think", *Atlantic Monthly* 176, no. 1 (July 1945): 101–108; Theodor Holm Nelson, *Literary Machines* (Sausalito, CA: Mindful Press, 1981).
[26] George P. Landow, *Hypertext: The Convergence of Contemporary Critical Theory and Technology* (Baltimore: Johns Hopkins University Press, 1992); Nancy Kaplan, "Politexts, Hypertexts, and Other Cultural Formations in the Late Age of Print", *Computer-Mediated Communication Magazine* 2, no. 3 (March 1995), http://www.ibiblio.org/cmc/mag/1995/mar/kaplan.html; Michael Joyce, *Of Two Minds: Hypertext, Pedagogy and Politics* (Ann Arbor: University of Michigan Press, 1995); Robert Coover, "Literary Hypertext: The Passing of the Golden Age", *FEED*, no. 10 (February 2000), https://web.archive.org/web/20000303231210/http://www.feedmag.com/document/do291_master.html; Jakob Nielsen, *Hypertext and Hypermedia* (San Diego: Academic Press, 1990).
[27] Dene Grigar, "Electronic Literature: *Where* Is It?", *electronic book review*, December 28, 2008, http://www.electronicbookreview.com/thread/technocapitalism/invigorating; N. Katherine Hayles, *Electronic Literature: New Horizons for the Literary* (Notre Dame: University of Notre Dame Press, 2008).
[28] Christopher T. Funkhouser, *Prehistoric Digital Poetry: An Archaeology of Forms, 1959–1995* (Tuscaloosa: University of Alabama Press, 2007); Alice Bell, *The Possible Worlds of Hypertext Fiction* (Basingstoke: Palgrave Macmillan, 2010); Belinda Barnet, *Memory Machines: The Evolution of Hypertext* (London: Anthem Press, 2013).
[29] Marcia J. Bates, "Information search tactics", *Journal of the American Society for Information Science* 30, no. 4 (July 1979): 205–214; "The Design of Browsing and Berrypicking Techniques for the Online Search Interface", *Online Review* 13, no. 5 (October 1989): 407–424; and "The Getty End-User Online Searching Project in the Humanities: Report No. 6: Overview and Conclusions", *College and Research Libraries* 57, no. 6 (November 1996): 514–523.
[30] Charles Upchurch, "Full-Text Databases and Historical Research: Cautionary Results from a Ten-Year Study", *Journal of Social History* 46, no. 1 (Fall 2012): 89–105.
[31] The Keywords Project, *Keywords for Today: A 21st Century Vocabulary*, ed. Colin MacCabe and Holly Yanacek (Oxford: Oxford University Press, 2018).