[White, A. (2014) Digital Media and Society: transforming economics, politics and social practices. Basingstoke and New York: Palgrave Macmillan. Chapter One, pages 3-25].

1
From the public to the private: the digitization of scholarship

Introduction

Depending on your perspective, digital media threatens either to destroy or revolutionize millennia-old scholarly practices. The way in which we seek information online as a means of helping us to construct knowledge differs significantly from tried and trusted academic methods. This would not be significant if we did not use the Internet so much for this specific purpose, but as will be outlined in more detail throughout this chapter, even those of us who carry out academic research are increasingly reliant on the Internet. For this reason, we need to understand both how we search for information online and the extent to which our online practices help or hinder us in our quest for knowledge.

Traditional forms of knowledge creation are based on structures and processes designed to lead the reader carefully along the path from the gathering of raw information to the development of understanding. These structures and processes have involved the categorizing of information (*classification*), testing its authenticity (*provenance*) and exposing us (*access*) to a plurality of it (*universality*). These core principles have underpinned scholarship and its attendant institutions like universities, libraries and archives. As this chapter will go on to illustrate, the rapid growth of digital media to its present pre-eminence as platform *par excellence* for the dissemination of information has called into question the validity of these long-held principles.

To some that it is to be celebrated. After all, in many ways, scholarship has become easier as we do not have to physically situate ourselves in the archive or the library. It also easier to search for sources without using cumbersome catalogues. But as more educational resources have moved from public institutions to online platforms, largely private, sector how does this impact on the ways in which we develop our knowledge about the world around us? This chapter will attempt to provide answers to these questions, beginning at the place where modern institutional forms of knowledge building began.

The Library at Alexandria

Private libraries had been in existence for centuries, even millennia, before Greek antiquity. Battles (2003: 25) reports that the first libraries of clay tablets appeared in Mesopotamia around 5,000 years ago and by the seventh century BC, a library at Nineveh not only contained an impressive 25,000 tablets, but also was organised in a semi-systematic way. But it was the Greeks who, in Alexander the Great, built in Alexandria just over two millennia ago what is commonly believed to be the first attempt to construct an institution that could properly be regarded as a research library. Alexandria represented the first serious institutional attempt to put into practice the principles that govern academic research to this day. While, for various reasons ranging from elitism to low levels of literacy, one of those principles, access, was not at the forefront of the librarian's priorities, Alexandria did develop rudimentary forms of classification, provenance and universality. As Simon Goldhill elaborates:

> Without the practices of the library, we wouldn't have the university in the form we have it today, we wouldn't have the organization of knowledge we have today, we

> wouldn't have the whole institutions of scholarship that we recognize. And that seems to me to be the sort of legacy that is really profound (Bragg *et al.*: 2009).

For these reasons, Alexandria is as good a place as any to begin a discussion on traditional forms of scholarship.

The Library at Alexandria and universality

In its attempt to secure as much Greek literature as it possibly could, as well as substantial collections in other languages, Alexandria can be regarded as the first viable attempt to establish a 'universal' library (Manguel 2008: 22, 24; Bragg *et al.*: 2009; Cavallo and Chartier 1991: 10).[1] There has long been a strong association between the concept of universality and knowledge which has provided the inspiration for the construction of libraries, archives and museums that attempt to collect 'everything' (White 2008). In the case of Alexandria, this was manifest in the library's housing of a community of scholars in its *Museon*, thus making a concrete link between universality of collection and knowledge creation. Despite – or perhaps because of – being destroyed after a few centuries, the idea that it encapsulated lived on in projects like the eighteenth century *Encyclopedie*, Dewey's nineteenth century decimal library classification system and, of course, contemporary national libraries and archives; all these projects and institutions are predicated on the Alexandrian concept of an inter-relationship between knowledge creation and universal access to documents. This inter-relationship will be explored further in a later discussion on digital media, before which there will be a discussion on other scholarly values which Alexandria popularized.

Classification/cataloguing

Gathering together under one roof as much information as possible is futile if it cannot be made accessible in a way that is convenient for those who seek to access it. This requires that the incoherent mass of information that users can access is given a structure or, to use terms germane to libraries and archives, classified or catalogued. It is believed that one of the librarians at Alexandria, Callimachus, created the world's first alphabetical catalogue, a 120-volume set which provided references of the library's most important Greek authors (Manguel 2008: 50; Polastron 2007: 15). As well as utilizing the alphabetic system which had been invented by Alexandria's first librarian, Callimachus also created tables based on different categories of knowledge (McNeely and Wolverton 2008: 20). This created not only the classification of different forms of knowledge but also a canon of important authors and texts. While the lack of historical sources makes any judgement necessarily provisional in nature, the very fact that the Museon (museum) attached to the library attracted the region's finest scholars suggests that there must have been some way for them easily to retrieve written material.[2] This theory is bolstered by our knowledge that the catalogue was alphabetical within the categories, a system which McNeely and Wolverton (2008: 21) argue was successful in making 'books readily and rapidly accessible to roaming encyclopedic intellects [the scholars *in situ* at Alexandria]'.

Provenance

Another important element of the Alexandrian library which retains its significance to this day is provenance, the need both to identify the creator/author of an individual record and to establish that it is the original version. The seeming desire for the Ptolemies to possess original texts rather than copies appears to support Derrida's (1996: 91) contention that there has always been an

obsession in western culture with 'origin', best illustrated in the archive. Both Manguel (2008: 24-5) and Battles (2003: 31) report that the Ptolemies often did not return the scrolls that they 'borrowed' for copying. This might have been unintentional, but it is certainly plausible to suggest that their desperation to retain the original documents was precisely because of the importance which they attached to provenance. Supporting evidence for this stance comes from Simon Goldhill's (Bragg *et al*.: 2009) reporting of the lengths to which the Ptolemies would go to secure original texts, sometimes to the extent of paying huge sums of money for 'borrowing' them – as above, he states that once secured these texts often were not returned to their original owners; this in addition to their practice of impounding books from ships that docked at Alexandria. And there was certainly an important practical reason for this: copies almost always contained many textual inaccuracies (McNeely and Wolverton 2008: 17).

The core elements of information-gathering at Alexandria

What emerges from this short discussion on the Alexandrian library is a general formula for information-gathering, much of which survives to the present day. These are:

- 1. Order/classification - cataloguing
- 2. Provenance – collected original documents
- 3. Access – limited to a small number of scholars
- 4. Universal – tried to collect everything
- 5. Public institutions hold and control most information
- 6. Nation-building – Alexander the Great was able to spread Greek culture throughout the Middle East

The last of these elements, nation-building, constitutes the modern imposition of a term on a historical epoch that did not have the same political construct, or at least not in the same form that we understand it today. Nonetheless, its inclusion is useful in illustrating the extent to which libraries have also played a wider political role. In relation to this chapter, that political role is associated with the relationship between memory institutions and the state, particularly during the emergence of the European nation-state in the eighteenth century to which we will now turn.

Francis Bacon, evidential scholarship and the emerging nation-state

In the modern area, the idea that there was an inter-relationship between scholarship, universality, classification and provenance received intellectual ballast from Enlightenment philosophies on knowledge construction. An acceptance of the supposition that there was a connection between universality and the attainment of knowledge led to great Enlightenment projects like Diderot's eighteenth century *Encyclopedie* and the increasing proliferation in that same century of *bibliotheques*, or catalogues (White 2008: 114-15). This demonstrates not only the enduring legacy of the Alexandrian library, but the influence of one of the most prominent philosophers of this or any other age, Francis Bacon.

Bacon is famous primarily for his invention of 'induction', the idea that scientific theories should be based on the observation of large amounts of data or of experiments. A fundamental requirement of Bacon's philosophy is that a 'great storehouse of facts should be accumulated' (Sargent 1999: xx). Thus can be discerned a link between Bacon's scientific method and the universal library:

> Once gathered, this experience had to be compiled into organized national histories, that could be printed and distributed throughout the learned world and thus could foster communication and the free exchange of ideas and information. As early as his advice to Elizabeth I in the 1590s, he had been urging the establishment of institutions that would advance this goal, such as *"a most perfect and general library," containing all "books of worth" whether ancient or modern, printed or manuscript, European or of other parts"; a botanical and zoological garden for the collection of all plants as well as rare beasts and birds; a museum collection of all things that had been produced "by exquisite art or engine"; and a laboratory "furnished with mills, instruments, furnaces and vessels" (vol. 8., pp. 334-35)* [emphasis in the original] (Taken from Bacon's *Book One, aphorisms*, and cited in Sargent 1999: xx).

Universalist projects became *de rigueur* for the emerging European powers as demonstrated by the archives which were constructed during this period: the House of Savoy archive in Turin in the early eighteenth century; Peter the Great's 1720 St. Petersburg archive; Maria Theresa of Vienna's 1749 archive; the establishment of princely and civic archives in Warsaw, Venice and Florence in the 1760s and 1770s; the creation of the French national archives in 1790; and the establishment of the UK Public Record Office (PRO) in 1838 (Steedman 2001: 68). And the methods of the historians working within these institutions were remarkably similar to Bacon's notion of induction, where the evidence was believed to speak for itself. We see this in the figure of influential twentieth century British archivist Hilary Jenkinson who, like Bacon, believed that the hypothesis should follow rather than precede the evidence (Gilliland-Swetland 2000: 12).

This type of evidential scholarship was based on principles that have altered little since Alexandria: namely universality, provenance and classification. The same could be said for librarianship as well, which, in the nineteenth century became more systematic in its acquisition and storing of books, especially after Melville Dewey's invention of a standardised decimal form of classification. Universality was promoted, like it was at Alexandria, through trying to collect virtually everything – in the UK and Ireland, for instance, there are six legal deposit libraries to which all publishers and/or authors in those territories must send two copies of their books.

As people became more literate throughout the nineteenth century and libraries and archives became, in theory at least, more accessible, these institutions had growing political influence. This role has been identified by McNeely and Wolverton (2008: 165) in the using by nineteenth century nationalists of public education in an attempt to unify European societies which were riven with ethnic, religious and class tensions. Similarly, the UK's Public Library Bill of 1850 was underpinned by a utilitarian philosophy which supposed that giving people greater access to information would make them more disposed to 'reason' (Battles 2003: 137).

The core elements of information-gathering in the modern, democratic nation-state

Let us remind ourselves of the earlier general formula for information-gathering at Alexandria and compare it to that in the modern, democratic nation-state.

Alexandria

- 1. Order/classification - cataloguing
- 2. Provenance – collected original documents

- 3. Access – *limited to a small number of scholars*
- 4. Universal – tried to collect everything
- 5. Public institutions hold and control most information
- 6. Nation-building – Alexander the Great was able to spread Greek culture throughout the Middle East

Modern, democratic nation-state

- 1. Order/classification – cataloguing
- 2. Provenance – collected original documents
- *3. Access – public libraries, museums and archives widened access*
- 4. Universal – tried to collect everything
- 5. Public institutions hold and control most information
- 6. Nation-building – information gathering institutions linked to power and the nation-state

The only major difference between the two is that of access, which, as a result of greater levels of literacy, better modes of transport and a democratic impulse to share knowledge as widely as possible, gave citizens of nineteenth century Europe much better opportunities than those in Egypt two millennia ago. That this is the only key difference illustrates the enduring legacy of Alexandria's values. Similarly, the nineteenth century typology above also accurately represents the contemporary situation today … or at least until the recent exponential growth of digital media. This last point alludes to the argument of many that these information-gathering

principles are no longer relevant to our modern informational environment, a debate to which this chapter will now turn.

The existential threat to the library and the archive

Those who first visit the impressive-looking building in downtown Washington DC which houses the USA's national archives and records might be surprised when they discover that NARA (National Archives and Records Administration) is dwarfed by its over-spill building in Maryland. Similar tales can be told about national archives and libraries in other countries, and are an enduring reminder of the capacity of the universal library to confound those who try to build it. The lack of affordable space in many of the world's capital cities, a problem made more acute by the strain on the public purse engendered by the global financial crisis, imperils the continued, seemingly unlimited, growth of national libraries and archives.

There is another threat to the national library and archive which is more existential in nature. The role of these national institutions has been so successful in developing civic consciousness among their citizens that they are seen to embody the values of their nation. The logic of this is that their destruction will not only result in physical loss but will also threaten those very values that they are seen to embody. And such is the identification of the nation-state with these values – national archives and libraries help to shape public consciousness which both reflect and propagate their own nation's values – that in war-time the destruction of archives and libraries can cause considerable loss of morale among the citizens of the nations to which they belong. While the wilful destruction of archives and libraries can be traced all the way back to ancient Alexandria and perhaps beyond, the sophisticated technologies that modern armies have at their disposal means that this can be carried out in a much more efficient and systematic

manner. During the Bosnian conflict in the 1990s, for instance, the Serbs targeted a number of cultural institutions, reaching an apotheosis with the destruction of most of the 1.5 million volumes in the National and University Library of Bosnia (Battles 2003: 188). According to András Riedlmayer, there was a rationale, albeit twisted, to this destruction:

> Throughout Bosnia, libraries, archives, museums and cultural institutions have been targeted for destruction, in an attempt to eliminate the material evidence – books, documents and works of art - that could remind future generations that people of different ethnic and religious traditions once shared a common heritage. … (Riedlmayer, A.; [Full reference not given by Battles] cited in Battles 2003: 188).

And, in a further twist, the person who signed the directive ordering General Ratko Mladic to shell the Vijecnica neighbourhood within which the National and University Library of Bosnia stood was Nikola Koljecvic, a former Shakespearian scholar who had often patronized the institution during the cosmopolitan tranquillity of pre-war Sarajevo (Battles 2003: 186-7). As in other such incidents throughout history, this was not merely a by-product of war: perhaps better than anyone else in Bosnia, Koljevic realised the importance of the role of culture in conflict. As Battles (2003: 156) points out, it is likely that the actualization of the fear of loss has had a profound influence:

> It may not be too much to say that the sudden disembodiment of the book in the late twentieth century – as text disappeared first into the grainy obfuscations of microfilm

> and eventually into the pixelated [sic] ether of the Internet – began with crude renewals of violence against the book in the First and Second World Wars.

Alluded to above is one of methods employed by archivists and librarians in response to the potential threats posed by diminishing space or destruction in war-time, namely micro-filming. Novelist Nicholson Baker's (1992) polemic on microfilming practices in British and American libraries illustrates the unsuitability of this method of preservation. Baker's details a litany of destruction of original newspapers, periodicals and books in some of the world's most prestigious public and university libraries in the process of their micro-filming. The Library of Congress's chief of photo-duplication in the 1970s when it was the midst of destroying thousands of newspapers, Charles La Hood wrote that: 'Microfilming came at a propitious time, as the Library of Congress was experiencing an acute space problem in its newspaper collection' (cited in Baker 1992: 35). This lack of space is, of course, not absolute, but an acknowledgement that the institution was either unwilling or unable to buy additional storage space, as Baker (1992: 36) would have liked it and other institutions to do.

Baker's polemic, though, has been challenged most prominently by Richard Cox (1992), whose critique of it touches on some of the issues that animate discussion of the purported impact of digital media on knowledge and society. This includes associating Baker's stance with the existential fear of losing information as articulated by many western writers – perhaps most eloquently expressed in Ray Bradbury's novel *Fahrenheit 451* – and with the long-held dream to construct the universal library (Cox, 1992: 11-14). He uses a quotation from a review by Julian Dibbell in the *Village Voice Literary Supplement* to emboss his point:

> Bush's [Vannevar] fantasy is no crazier than Baker's. In essence, both of them dream of having access to all the information ever published, and it drives them nuts to see a single scrap of it fall through the cracks. But the world and all the order in it are always slipping through the cracks, and the failure to reconcile oneself to that is, among other things, as good a definition of obsessive compulsion as any (Dibbell 2001; cited in Cox 1992: 124).

While not commenting on the cost of Baker's proposal that public libraries and archives build warehouses to store material rather than micro-film it, Cox (1992: 43) argues that this would make collections less accessible to the public.

But the debate about micro-filming is now largely academic, as digitization has become the primary method for the reproduction of original scholarly sources. The advances in computing technology and storage capacity over the two decades since Cox and Baker's books were published would appear to supersede concerns about a trade-off between storage and access. In the world of paper archives and libraries, the increase in amount of material, entailing as it does the building of storage space on remote sites where property is cheaper, decreases the quality of access. In addition to providing an elusive target for those with baleful intentions, the virtualization of the archive and library offers a platform for a seemingly unlimited amount of scholarly material as well as the capacity, through the operation of highly sophisticated search engines, to deliver any digitized document to the user's PC within seconds. This is an obvious lure for those of us who spend inordinate amounts of time both visiting different research institutions and waiting for material to be delivered from over-spill stores *within* institutions (in the British Library, an increasing number, possibly a majority, of books take two days to arrive

at a reader's desk after a long journey from Boston Spa to London). Electronic access is purportedly egalitarian in its exposure of these collections to a much wider audience than those who have the time, money and inclination to travel to the best archives and research libraries (White 2011: 317-18). Accordingly, a number of great public institutions like the USA's Library of Congress and the British Library have long-running digitization projects, but their efforts are dwarfed by the search engine giant, Google, especially since the onset of the global financial crisis (White 2011).

The Google Books Project

Since its December 2004 announcement of its agreements with some of the world's leading research libraries, Google has made steady progress towards its goal of digitizing every single one of the 32 million books in the WorldCat (Stross 2008: 98, 107–8; Vise 2006: 238). The project is an extension of the corporation's belief that gathering together as much information as possible and making it as easily retrievable will enhance humans' capacity for knowledge construction:

> It [Google] seek to develop "the *perfect* search engine", which it defines as something that "understands *exactly* what you mean and gives you back *exactly* what you want". ….
>
> In a 2004 interview with *Newsweek*, Brin [Sergey] said: "Certainly if you had all the world's information directly attached to your brain, or an artificial brain that was smarter than your brain, you'd be better off." [emphasis in the original] (Carr 2008: 4).

Explicit in Google's mission is creation of an information gathering regime that will not only supersede the author's schema for information-gathering in the modern, democratic nation-state, but will continue to evolve until perfection in information retrieval is attained. Like much that pervades digital media, this is a revolutionary approach in its abandoning of the careful construction over two thousand years of an information-gathering regime that differed little from that during the time of the Ptolemies for a model which is perpetually re-calibrated. Given the number of books that Google has already digitized – Auletta (2010: 257-8) claims that this was 7 million by October 2008 - it seems that the universal library (of books) is within reach. While this is potentially a boon for researchers and the general public, this will profoundly alter existing information-gathering models.

But we should never lose sight of the metaphysical rationale for the building of these huge electronic libraries. Manguel's (2008) view that the desire for the universal library appeals to our need to establish a sense of order in a complicated world, echoes Google's proclamation that its massive digitization project is not merely to make these books 'universally accessible and useful' but a more ambitious plan to '*organize* [my emphasis] the world's information' (cited in Appleyard 2007). These large-scale digitization projects should not, then, be viewed only as a means of improving access to existing archives and library collections, but also as a re-organization of the information contained within them. The question that needs to be asked in relation to this is what implications does this have for the type of evidential scholarship based on classification and provenance that has been the basis of knowledge construction for centuries? The answer to this question will be answered though a consideration of the extent to which the scholarly methods outlined earlier are applicable to our digital media ecosystem.

Information-gathering in Googleverse

Before considering the way in which the virtualization of so much of our scholarly heritage will affect how we search for information, let us remind ourselves of the existing typology:

Modern, democratic nation-state

- 1. Order/classification – cataloguing
- 2. Provenance – collected original documents
- *3. Access – public libraries, museums and archives widened access*
- 4. Universal – tried to collect everything
- 5. Public institutions hold and control most information
- 6. Nation-building – information gathering institutions linked to power and the nation-state

The sixth element, nation-building, can be set aside for now and more fully considered in chapter three on the theorization of politics in our digital age. How are the other five elements to be conceptualized in societies where the search engine is often the first port-of-call when we look for information?

It is easy to exaggerate the potency of the search engine, especially in the light of the findings of Head and Eisenberg's (2009) study of 2,318 US college students that most students refer first to course readings when they are writing assignments. However, those same figures show that around 96 per cent also use Google and 85 per cent Wikipedia for help with their

course assignments; the figure for course readings is 97 per cent (Head and Eisenberg 2009: 18). For information-seeking that is not related to their courses, students are most reliant on Google (around 98.5 per cent) and Wikipedia (around 90 per cent). A previous study by Head (2007) demonstrated that even for research for course assignments students were most likely (47 per cent) to go to the World Wide Web first. Other studies by the OCLC in 2006 (cited in Rowlands *et al.* (2008: 292)) and Van Scoyoc and Cason (2006) report similar findings, with the former observing that 89 per cent of college students use commercial search engines when they begin research; only 2 per cent start with a library website. In all these studies, the assistance of both library websites and librarians themselves is low on the list of students' priorities. Thus the search engine's epistemological break with traditional models of information-gathering cannot be ignored when we are discussing knowledge construction in a digital age.

1. Classification and the mathematical algorithm

As stated throughout, even though electronic catalogues have been in use for decades in libraries and archives, the concept of classification has retained its value within those institutions. This is partly because it helps us to navigate our way around the increasing proliferation of information that digital media has ushered in. The need to manage large surges in information is not new, Postman reporting how there was a significant increase in the number of schools in England from the late fifteenth to early seventeenth centuries as a response to Gutenberg's print revolution (1993: 62-3). But earlier forms of information management have taken place within the broad information-gathering framework referenced throughout this chapter. Thus the search engine is so radical not only because of its technical capacity but for its rupture of existing information-gathering protocols.

And how is the search for information facilitated by search engines if there is no discernible method of classification? The short answer is that search engines use sophisticated mathematical algorithms to deliver to the reader the information most relevant to his/her search. Unlike traditional classification systems, these algorithms are secret and being constantly re-calibrated. Google is thus the most efficient search engine because its engineers have designed a better algorithm than its competitors. The lack of obvious structure runs contrary to the type of evidential scholarship that has been the mainstay of the best models of education for centuries. There are, though, those who believe that we should celebrate the supersession of formal classification.

David Weinberger's book *Everything is Miscellaneous* is a paean to the wonders of a classification-free digital world (2007). Weinberger's basic argument is almost technologically-determinist in its assertion that the rules of the so-called world of atoms do not apply to digital information. In relation to scholarly information, the tagging of each digital file with as much metadata as possible supersedes, according to Weinberger (D. 2007: 17-23), the need for classification. This will enable researchers to access the most relevant information without the intercession of a scholarly guide. This belief is partly based on the idea that digital media has caused the death of distance, not only bringing those far-flung archives and libraries nearer to us through the act of digitization, but information generally, which is now at our 'fingertips' and hence does not need require a mediator (in the form of classification) in the same way that it did in the past (Friedman 2006: 176-85). To paraphrase Sergey Brin's words above, who needs mediation when 'all the world's information [can be] directly attached to your brain' (Carr 2008: 4)? And in an era where many of us are bypassing traditional library and archival classification systems and finding rich resources online, then there is some merit in this argument. But there

are two major flaws in it too: the argument that there is no structure in the search engine is problematic, as is the belief that structure no longer matters.

Classification in academia has always been a problematic concept, as seen in Petrucci's (1991) argument that creating a canon of the best authors in any given discipline is a project designed to sustain governing ideologies. In this sense, the creation of reading lists and the classification of material in the library and the archive could be read in the same way; after all national memory institutions are designed to sustain *national* memory perhaps more so even than general knowledge. These types of post-modernist critiques of academic classification are not new, but have been given greater potency by the advent of technologies that provide a viable practical alternative to formal classification schemes. But the belief that the mathematical algorithms that drive search engines are random or neutral is fundamentally flawed. Google's PageRank technology bears similarities with academic peer-reviewing in its ranking of websites. The algorithm elevates those websites that are physically linked to larger numbers of other websites, with links to the higher-ranking websites giving additional privileges. Gleick (2011) contrasts this with methodologies of earlier search engines which ranked websites purely quantitatively, emphasised with an anecdote about how the Oregon Center for Optics appeared as the first result of an *Altavista* search for 'university' simply because that word appeared many times in a headline about the Center.

One of the no doubt unintended effects of the PageRank methodology is that it reinforces rather than challenges existing hierarchies of information, a phenomenon O'Neil (2009: 58) likens to the 'Matthew Effect' in academia where established researchers are much more likely to gain citations than their less experienced colleagues. Also, though their ideology is global, search engines are biased towards certain regions and languages. The most commercially

successful search engines are American and there is a bias towards English language sources generally and American websites specifically in their rankings, partly as a result of their longer vintage (Halavais 2009: 89-90). In a widely publicized critique of Google, the then French national librarian, Jean-Noel Jeanneney, was concerned, among other things, about the search engine's bias towards English language sources even on subjects where sources in another language would be more appropriate. From his own country's perspective, there was a danger that people would be referred primarily to English translations of the work of France's greatest novelists or English language versions of significant events in its history (Jeanneney 2007: 42-3). The French government tried to counteract these tendencies through the development of a European search engine during the middle of the first decade of the millennium, but this – Quaero – petered out when Germany withdrew its support in 2007 (Doueihi 2011: 166-7; Vaidhyanathan 2011: 25). Despite this, Vaidhyanathan (2011: 138-9) reports that Google is increasingly tailoring search based on location of the user. While this might be an efficient strategy for locating your nearest pizza parlour, searches that pander to the user's particularities are probably not the best way of developing broad-based intellectual knowledge.

2. Reputation-building rather than provenance

The tendency of techno-utopians to argue against traditional forms of classification while ignoring the way in which the Internet classifies information is repeated in relation to provenance. The students who, in the empirical studies earlier in this section, spurned the authority of the librarian for the supposed efficiency of the commercial engine can justify their actions by invoking a whole class of Internet theorists, social commentators and commercial corporations who believe that challenging the experts is the intrinsic duty of Internet users

(Weinberger, D. 2007; Friedman 2006; Brin, cited in Carr 2008: 4). But those of us who teach in universities know that encouraging students, or anyone else for that matter, to launch critiques of established theories and concepts without reading authoritative texts is not good pedagogical practice. My impression is that by and large students are aware of this too. At the same time, the convenience of Internet searching makes it an attractive option that they are not going to forgo in their academic studies anytime soon.

There are some, especially the commentators mentioned in the last paragraph, who believe that the Internet represents a democratisation of information and therefore is a welcome departure from authority-based knowledge development (Friedman, 2006: 176-85). But the need to sift credible sources of information from the not-so-credible is accepted even by some of the most enthusiastic advocates of user generated content (Gillmor 2010). Provenance in the library and archive is partly based on the book, document or record's location within a wider structure, be it a series of records or canon of literature. The Internet is not structured in that way, so how does provenance operate within it?

The first approach is through the establishment of reputation. Online reputation is determined mainly quantitatively. This is the core of Google's PageRank which, while giving additional credence to websites linked to their highly-ranked peers, ranks websites by the number of times they are appear as links in others. The success of this method is illustrated by Vaidhyanathan's (2011: 59) review of some empirical studies which demonstrate that users exhibit a 'trust bias' in relation to Google's ranking. There are other websites, like Reddit, Digg and del.ic.ious, devoted to ranking reputation through the use of 'folksonomies', a method which allows users to tag those websites or sources of which they most approve (O'Neil, 2009: 49-50). Some of these methods are taken from commercial websites, most notably Amazon's ratings of

books and the reliability of second-hand booksellers (O'Neil 2009: 50). This has led to the championing of the 'wisdom of crowds' or the 'hive mind', a form of collective intelligence that the Internet can easily facilitate (Surowiecki 2005; Leadbetter 2008; Shirky 2009). The belief that the hive mind is superior to individual experts is manifest in projects like Wikipedia. But this approach dangerously conflates popularity with authority, risking in its extreme form, to use the delicious phrase of Jaron Lanier (2011: 79), 'digital Maoism'.

Like the dismissal of academic classification, the casual suppression of the principle of provenance in the online world does not actually push it to the margins but encourages it to mutate into a more virulent form. Thus, rather than exposing us to a greater plurality of information, all too often those with the greatest reputation online are interchangeable with popular figures offline. In the London *Independent* newspapers 100 most influential 'Twitters' in 2012, included in the top ten were a footballer, celebrity chef, DJ and famous illusionist, as well as four comedians/actors (Burrell 2012). This illustrates the populism rather than pluralism of the so-called 'hive mind' and much of what passes for information on the Internet today.

3. Greater access?

The Internet continues to give ever greater access to its content. While many websites are still censored in China, the world's most populous nation now has an estimated 590.6 million Internet users, representing 44.1% of the population (Pew Research Center 2013a). On the African continent, where access to technology has traditionally lagged behind other regions, the recent rapid take up of mobile phones and wifi technology has greatly improved access (see chapter nine). In the developed world, the migration through digitization of scholarly materials to online platforms continues seemingly unabated. While some people continue to have much greater

access to the Internet than others, the global diffusion of digital media technologies seems to be closing this gap rapidly (Friedman 2006). But even this rosy scenario contains some caveats.

Let me explain this by reflecting on my own access to Internet content as I type these sentences on my university PC in China. In many ways I am information-poverished, as my access to websites like Google is severely restricted or blocked. I could, of course, buy a VPN to deal with this problem, but this involves a fee for what ideally should be free, could make me more vulnerable to viruses and advertising, and does not always work properly. Sadly, this tale of patchy coverage and increasing potential financial cost of surfing is all too familiar in an Internet where government restrictions are increasing in many places and where commercialisation is taking a firmer hold. I do, though, through my university have greater electronic access to articles from some of the world's leading educational journals and databases than the vast majority of my fellow citizens. I can also access the *New York Review of Books* on my e-reader (which requires a small subscription each month) but not the London *Times* (which also requires a small subscription but that I do not want to pay). My anecdote has highlighted the main difference between access in China and the Anglophone world.

In the Anglophone world, intellectual property serves as the most important gate-keeper to online content. Were I not a member of faculty at a prestigious university which is prepared to pay the expensive institutional subscriptions for many of the leading journals in my own discipline, it would be difficult for me to have anything more than limited access to the journal articles of the main educational publishers. As my example of the London *Times* shows, an increasing number of the world's most prestigious newspapers are also retreating behind online paywalls. Access to copyrighted material is not so much of a problem (again, depending on your perspective) in China, where Montgomery (2010: 108) estimates that up to 90 per cent of film

and music watched and listened to in China is pirated, but random blocking of access to content that does not have copyright restrictions clearly is. What we can conclude from all this is that while the Internet is giving ever greater access it is not clear what access models will dominate in the future. The Google Books project illustrates the folly of trying to predict which model will prevail. The project has slowed down after a number of legal challenges by publishers and at the time of writing it is not clear whether all these books the corporation has digitized will be accessible to the general public in the future. There is no guarantee that Google will even exist in a few decades time; if it does, it is not beyond the realm of possibility that it might decide that providing free access to its digitized books is no longer a priority or economically viable (Vaidhyanathan 2011: 165, 202). With some European politicians, including the British Prime Minister David Cameron, thinking aloud that suspected rioters maybe should be banned from using social media, we should not assume that the Internet is on an ever upward curve of greater access (Halliday 2011). What can be stated with more conviction, though, is that our present generation lives in a much richer information environment that at any point in history.

4. OCR'ing technology, transcription and universality

It is surely safe to assume that greater access has also been mirrored by a serious move toward the once seemingly impossible goal of the universal library. Notwithstanding its legal problems, Google's plan to digitize every single book in the WorldCat could be completed in this decade. This project is supplemented by initiatives to digitize special collections in libraries, archives and universities throughout the world (White 2011: 321). Does this mean that the universal (digital) library is attainable? The short answer is not in the near to medium future; a longer explanation will follow below.

What is good scholarly practice, namely the digital copy's fidelity to the original, is good for the universal library too. This is because inaccuracy in copying is not only detrimental to the quality of individual scholarly resources, but also reduces a digital library's coverage, in other words makes it fall short of universalism. Every time that text is omitted or obfuscated in the journey from the printed page to the electronic file, or metadata not recorded, there is a reduction in the amount of information in the universal library. When these omissions and inaccuracies reach a certain level (Library of Congress (2013) guidelines for its digitization programmes stipulate that these should occur in no more than 0.05 per cent of total characters, or one in 2,000), then the universality of the library is called into question. Only a very few of the accounts or critiques of Google Books discuss the actual quality of its digital copies. Of those that do, Jones (2010) and Duguid (2007) identified an alarmingly high number of images that it would not be possible to convert into machine-readable text, while a more positive account by James (2010) identified errors in less than one per cent of the pages he sampled. It might be possible to 'clean up' the text by human hand either individually or by crowd sourcing.[3] But cleaning up millions of words is time-consuming and expensive, and it is not clear that crowd sourcing can quickly reach the level of accuracy recommended by the Library of Congress (White 2011: 322). In many respects this discussion is futile because the secrecy of Google's work practices means that not only do we not know how it is applying quality assurance procedures, but more importantly we have no idea whether Google is even working towards Library of Congress levels of accuracy.

This concern extends to metadata generation too. Metadata, which is literally 'data about data', is automatically generated when digital files are created. Of most importance to scholars is the writing of bibliographic metadata, of the type that librarians and archivists have appended to

records for centuries. This can only be added manually, which is a daunting task when one considers the vast number of individual and series of digital files that are produced every year. Early studies of the Google Books project suggest that its metadata is not as accurate as similar projects in the public sector and contains more omissions (JISC 2007: 3; Townsend 2007). Even publicly-funded projects have been lax in creating bibliographic metadata and generally it is difficult to agree on or enforce common standards (White 2011). All this highlights the naivety of Weinberger's (D. 2007: 17-23) earlier comment that the mass generation of metadata can replace classification. But this is not a problem that relates only to classification. Every inaccurate or omitted piece of data in a database limits the amount of information that it can provide. The most disturbing element of this is that while errors in traditional libraries and archives can be easily identified, it is much more difficult to identify errors in hidden databases, especially when they are run by corporations who might see no commercial gain in admitting error.

The long-term sustainability of digital texts cannot be guaranteed and is another reason why we should temper the hubris of the most zealous techno-utopians. There is as yet no answer either to the problem of rapid obsolescence and replacement of digital storage formats and platforms (Doueihi 2011: 119-21) or to the danger posed by changes in digital storage policies as a result of alterations in the financial or organisational structures of corporations (Vaidhyanathan 2011: 165, 202). One of the reasons why micro-filming was pursued so zealously by librarians in the 1960s and 1970s USA was that newspapers were deemed to be acidic and hence at risk of destruction. The result of that campaign was not only the destruction of huge runs of newspapers but also the loss of the information within them as large amounts of micro-film are now unusable. It would be perverse if the pursuit of universality was used an excuse to destroy

original books and other documents, leaving their content solely in digital files whose deterioration or loss would not only imperil the universal digital library but our great existing scholarly paper heritage too.

5. Private *search engines have superseded* public *institutions as the main source of scholarly information*

The point in the previous paragraph about the potential dangers involved in the holding of scholarly information by corporations that could change drastically or even fold within a short period of time will continue to be a salient one in this period of financial instability. There is nothing inherently wrong with private corporations holding this type of information, as the important intellectual role of long-established academic publishers, newspapers, private libraries, archives and educational institutions testifies. But in most countries, the state has acted as a guarantor of educational information through licensing, legal deposit and the public funding of knowledge and educational systems. Furthermore, what makes search engines different from these other private institutions is that their revenue is advertising-based and therefore not directly derived from the information that they disseminate. This makes them less sensitive to the quality of the information that their search engines uncover, unless of course it threatens their advertising revenue. Academic publishers, newspapers, private libraries, archives and educational institutions cannot be so blasé about the quality of their content and hence have more in common with the great public institutions than they do with commercial search engines. For these reasons, commercial search engines should not be the main custodians of scholarly information (see also Vaidhyanathan 2011: 202). Furthermore, despite the national bias of most

search engines, they are not as concerned about developing the public sphere as traditional knowledge institutions, the political implications of which will be discussed in chapter three.

Conclusion

Let us remind ourselves of the earlier general formula for information-gathering in the modern, democratic nation-state and compare it to that in the age of digital media:

Modern, democratic nation-state

- 1. Order/classification – cataloguing
- 2. Provenance – collected original documents
- *3. Access – public libraries, museums and archives widened access*
- 4. Universal – tried to collect everything
- 5. Public institutions hold and control most information
- 6. Nation-building – information gathering institutions linked to power and the nation-state

Digital Media

- 1. Order/classification – *seen as unimportant*
- 2. Provenance – *hard to establish*
- 3. Access – universal in theory
- 4. Universal – attempt to collect everything

- 5. *Increasingly, it is private institutions which hold and control most information*
- 6. Nation-building – *not important*

A number of studies of hypertext reading have shown that students find the lack of narrative structure disorienting (White 2007). There is no reason to believe that this does not continue to be the case. This would suggest that digital media's discouragement of formal classification and imperfect methods of determining provenance has a similar disorienting effect on the user. As was outlined earlier, there is a hidden form of classification online, but it is both intellectually problematic and perverse in that by privileging what is already popular it can often expose users to an ever narrower range of information than more academic forms of classification. The attempt to establish some sort of provenance through reputation threatens to degenerate into digital Maoism or mob rule. The lesson, then, is that the Internet needs to learn from academia, the library and the archive. One of the ways in which Google has done this is through the creation of Google Scholar, a search engine for identifying academic papers and their citations. Other initiatives like academia.edu, a 'Facebook for academics', which is more structured than Google Scholar, is also a promising development. But it is easy to bash the Internet and the powerful global search engines and neglect the role of the user. The onus is on us too to be less dependent on the search engine for sourcing academic work, which should involve consulting experts, particularly librarians, more often (Halavais 2009: 113).

Focusing on the user leads to a broader point about media literacy. While scholars have been concerned with this for many years (Kubey 1997; Livingstone 2004), the growth of powerful search engines and the advent of the Google Books project in the past decade makes media literacy all the more important. This would begin to address the general lack of intellectual

curiosity about the structure and biases of search engines, as illustrated by Deborah Fallows's research in 2005 which found that 68 per cent of users in her sample thought that the search engine was unbiased, while 62 per cent were unaware of the distinction between paid results and those where no money had changed hands (Fallows 2005: i-ii; cited in Van Dijck 2010: 581-2). While one would hope that Internet users have become more savvy since then, initiatives need to be developed to help people make a clearer distinction between authority and popularity (Gillmor 2010; Halavais 2009 110-11). Ideally this type of media literacy education should not be solely technical, namely merely teaching students how best to play the system (Van Dijck 2010: 575). If it does not engage with wider epistemologies about the construction of knowledge, there is a danger that it will merely reinforce the view that the search engine is the best vehicle for delivering the most useful information and it is our task merely to improve the efficiency of our searching, rather than consider alternative strategies (Halavais 2009: 94).

We should not, though, lose sight of the potential of the new research methods that digital media has encouraged. While, as argued earlier, crowd sourcing is not a viable alternative to provenance, it does have a role in quickly and efficiently correcting factual errors. The digitization of archives and library collections has enabled us to find information almost instantaneously rather than spend days wading through original documents for one reference (White 2011: 318). (This presupposes that institutions are digitizing to Library of Congress standards of accuracy). And this searching need not only take place within one database but across a huge number. This can enable the identification of patterns that individuals or even teams of researchers would not be capable of detecting on their own. While there are legal implications in relation to data mining, it does have the potential to facilitate inductive research for ground-breaking research in diseases and other socially beneficial areas beyond even the

wildest dreams of Francis Bacon and the evidential scholarship of the nineteenth century (Vaidhyanathan 2011: 178-9; Van Dijck 2010: 585). The research potential of the search engine makes it all the more important for it to be more transparent about its structure and methodologies (Halavais 2009:115; Gillmor: 2010). While there are many reasons for this opaqueness, primarily commercial confidentiality and their status as private not public organizations, search engines will struggle to retain their credibility as serious facilitators of knowledge construction if they do not make some effort to employ typologies of information-gathering that have elements roughly similar to that of traditional libraries and archives.

While this chapter has focused on knowledge construction, many of the issues raised here have wider political ramifications. One of those issues, the role of social media in altering how we think about individual identity, will be explored in the next chapter. The relationship between knowledge and identity in the age of digital media is profoundly political, the implications of which will be theorised in greater depth in chapter three, the concluding section of Part I.