The Sacred Unreadable Artefact: Digital Preservation, Computational Abundance, and Scarce Access

Martin Paul Eve

Birkbeck, University of London

There is a short story by the famous Argentine author, Jorge Luis Borges, of a civilization possessed of a holy book. The book must, at all costs, be protected and preserved for the future. It is encased within a dark and mighty sarcophagus to ensure its safety from *quote* "humidity, heat, damp, cold, ice, fire, wind, rain, snow, sleet, prying fingers, hard stares, the gnawing of rats, sonic disintegration, the dribbling of infants, and the population at large" *end quote*. The special caste of custodians in the story – a kind of priesthood of knowledge – are confident that they can protect the book; especially from this last and most damaging group, the population at large. Indeed, as time goes by and greater swathes of this growingly democratic population request access to the book, the priesthood formulate ever-more contrived rationales for the protection of the artefact. The intrinsic value of the book, to use a term from the report that forms the basis of today's symposium, seems, in the story, to be increased by its scarcity of access, even as its instrumental value to society grows

lesser by the day. For even the priesthood do not really know or understand the contents of the book that they guard. They have only the peripheral metadata context within which to work: the sacredness of the artefact, but also the sacredness of the notion of preservation. As preservation becomes an end in itself for the priesthood, the barbarian populace eventually overwhelm the fortification and prise open the sacred sarcophagus. The story draws to a close as the lay tribes examine the holy book, over the corpses of the priesthood, to find that it is written in a language and script that is completely indecipherable and that has been lost to time; as meaning has eroded over the span of artefactual preservation.

Borges, of course, never actually wrote such a story. But he could have and it did sound vaguely plausible as a transparent allegory of the phenomenon under discussion today. Namely: what is the tension between, and the resolution of, preservation and access for non-print legal deposit? How is it that we have come to a situation where the path-dependence of print has so thoroughly conditioned the access possibilities for the digital that its most salient property – that of non-rivalrous dissemination – must be once more made rivalrous and discarded? And what of the structures of meaning that themselves naturally erode over time, like an entropic process, in the digital space? How, without some form of continuous access, can we ensure that we can still read our digitally preserved heritage over even a decadal timespan?

[SLIDE]

Perhaps a more basic question that many of you are asking, but are too polite to voice, is: why am I (Martin Eve) here? The simple answer to this is that I have been actively involved in attempting to reshape academic publication cultures in the humanities disciplines for the past decade in order to take advantage of digital affordances. I should also note that, as a result, my observations in this keynote are inflected towards the academic publishing end of the spectrum, whereas digital legal deposit covers artefacts from diverse sources. Most notably, I have worked on eradicating paywalls to humanities scholarship in the belief that it will be educationally advantageous to all societies to

allow the general public to read about cultures, artforms, and history for free. Along the road that I have travelled, I have met much resistance to this opening of access, often from those who believe, I think wrongly, that they are thereby "defending" the humanities subjects. But I have also met resistance from publishers in the form of partial moves to greater access that do not fully harness the digital. Let me give an example. In the UK, as an attempt (in my view) to stymie the totally free and open consumption of research material by anyone online, The Publishers Association and several other bodies established an "Access to Research" pilot, in which anyone could access the content of research articles – for free – so long as they did so *on site* and *in person* in a library. In the first 19-month pilot period of "Access to Research", the UK national total for access was 89,869 searches from 34,276 users.

That doesn't sound so bad, you may think. But the 909 articles published or supported solely by the platform that I run, the Open Library of Humanities, in its first year accumulated 118,686 unique views. That is, this tiny number of open access articles were viewed by more people than a UK national-level pilot giving on-site access to vast quantities of subscription material across all disciplines over almost double the same time period. This kind of study is most often used to show that "very few people want to read this material, so why should an industry reconfigure its economics to accommodate such changes?" I think that our platform shows exactly the opposite, though. For this is where my interests in open access coincide with issues of user-centric thinking about non-print legal deposit. In a world where we can demonstrate by example that there is an audience for even the most abstruse types of humanities scholarship, it is becoming increasingly problematic to separate preservation from any kind of distributed networked access.

And an important part of this is thinking about disability; that is, opening access across many bounds. It is a notable feature of the white paper that existing provisions for disability in non-print legal deposit justifiably come in for a rough ride. As a disabled academic myself who has vasculitis, rheumatoid arthritis, neurological hearing loss, as someone who has had a stroke and

pneumonia and sepsis and acute bowel dysmotility, I am very much fed up with people telling me that we do not need open access to monographs and articles in the humanities – or perhaps even elsewhere – because people can "go to a library" to access these. There are many people for whom "going to a library" is an extremely difficult thing to do. From those with bowel problems to those with Parkinson's and a whole host in between, the reality on the ground of difficulty of physical access gives the lie to the intrinsic assumptions of ability by those who make such claims. I could not "go to the library" to read academic books when I spent most of March 2019 in hospital, despite the fact that I was still mentally alert enough to conduct academic work. So while it is heartening to hear of voluntary arrangements to get around the poorly worded exceptions only for those with visual impairments in non-print legal deposit, it is clear that nobody's rights were ever really respected by voluntary arrangements. A user-centric model of access, to legal deposit artefacts, or indeed to any kind of contemporary artefact in a library, must consider what it means to insist on embodied user presence in an era of potentially digital access. For it is this insistence that causes the disability, by disabling entire groups from being able to conduct research. Users is not a homogeneous group, but a set of intersecting identity formations that require accommodation; of which digital dissemination is merely one aspect.

[SLIDE]

Fundamentally, though, what we are dealing with when we deal with non-print legal deposit and the difficulties put up in order to access this archive is, in my view, the history of keeping ideas scarce. We do so in order to bestow an economic quality (scarcity and value) onto a fundamentally non-economic concept (an idea). How are ideas non-economic? Namely: that they are not rivalrous. Like stories, music, and other forms that we have come to call "intellectual property", ideas can be transmitted from one person to another without the original owner having to lose anything. This stands in stark contrast to material objects, which must be rescinded upon their transfer. Material

objects, as such, have a scarcity function embedded within them; each can exist only at one place and time. Ideas, on the other hand, do not.

Digital objects are a lot like ideas, although not identical. Although they rest upon the materiality – and therefore scarcity – of material infrastructures (hard drives, computer monitors, processors etc.) they give the appearance of unlimited worldwide replication in an instant, without the original owner losing his or her version. That is, the digital appears nonrivalrous. To encode an idea within a digital form gives the first moment in human history where the form could do justice to the content.

The problem always was, and still is: how does one make ideas and digital objects economic, in the sense of scarcity? One much-touted recent solution – although it's rarely billed as such – is the blockchain. The blockchain can work as a "currency" because it puts a requirement of cryptographic work – a scarce resource – on the verification of transactions. This is a technical solution to a technical problem – although it comes with its own technical problems, most notably the question of who will be left to spend bitcoin after its astronomical power consumption has accelerated global warming to the point of human annihilation.

The social "solution" to this problem of non-economic forms is copyright and patents, the former of which protects expression and the latter of which protect ideas. Copyright as it exists today in most jurisdictions offers the twofold protections of economic and moral rights. This is usually now recognised as a time-limited monopoly on an expression in order to reward original creativity. In the contemporary world of academic publishing, the common rationale that publishers give for demanding copyright transfer or exclusive assignment is that they have invested labour in the production of the article and should, therefore, be able uniquely to benefit from selling that article during the term of copyright.

[SLIDE]

But John Willinsky draws attention, I think brilliantly, to a most curious aspect of the origins of modern copyright. The common legal touchstone to which most histories of copyright point is the 1710 Statute of Anne. This enshrined the first copyright term – a mere 28 years, trivial by today's standards of the author's life plus seventy years – in law. It was also implemented in order to incentivize the production of works. BUT, of critical importance, I think, following Willinsky, is that the Statute of Anne is titled, in brief, "An Act for the Encouragement of Learning". In the history of implementing copyright, learning holds a special place and is to be encouraged. Over time, the centrality of learning and its importance has been lost from our understanding of copyright – as, I often fear, has the notion that copyright is time limited and designed for the eventual decommissioning of exclusive right of expression to the public domain. Yet learning was central to copyright's initial establishment.

Legal deposit is a mechanism that recognises the importance of learning, of cultural heritage, of archival records and of building that archive, even when it has to operate under limiting conditions of nationalism in a world where learning and culture are now global. The question is: does it recognise the *encouragement* of learning, or merely the fetishisation of an abstract learning, conducted by nobody, upon material inside a black sarcophagus to which no-one really has meaningful access?

Furthermore, the assumptions behind some of the protections of legal deposit do not make great sense in an open-access world. The first is the question of what it means *to publish* something. The second is the fact that there is no provision for more liberal access and licensing built in to non-print legal deposit. I want to turn briefly to each of these.

The usual, overly simplistic definition that is given to "publication" is "the act of making public". Michael Bhaskar sufficiently problematises this, I think, though, when he asks whether it is really "publishing" a work to leave a single copy of it, say, **[SLIDE]** on a park bench. This would, in all senses fulfil the definition of "making public", but it is a pretty questionable basis on which to

say that a work was "published". Bhaskar instead re-defines publishing as the combined threefold functions of filtering, framing and amplification. In this way, publishing becomes not an act that pertains to the availability of the work, but instead is about the labour functions that are undertaken.

Does this context cross over into the digital domain? Certainly, Clay Shirky has told us that, in the digital world, the labour of publishing is, instead, a button. Yet what does it mean in a fire-and-forget culture to return to this notion of "making public" as being akin to "publishing"? For the functions of filtering, framing, and amplification are now widely dispersed among many different entities. For instance, consider the act of filtering. What or who is the gatekeeper in a world of digital self-publishing? In our digital world, *discoverability* is provided by third-party black boxes, like Google, who help us to sift through vast quantities of material (obviously with a degree of self-interest in so doing). This is not strictly what Bhaskar means by "filtering" – which he thinks of as a gatekeeping function on the publishing side. But if one is able simply to make available one's own material, then the idea of filtering becomes much more closely tied to discoverability. Filtering is needed by *users* – but *after the fact*. Indeed, these sifting (filtering) services are now also indistinguishable from various amplifcatory functions; with paid placements and so forth, it is difficult to amplify except by separation from the otherwise indistinguishable haystack. For a fee, Google will give honed magnets to one's readers, with which they might extract your needle.

Where this perhaps gets most interesting, though, is in the notion of framing. Non-print legal deposit brings a new slant to the framing of born-digital and digitally-published objects. For it confers a legitimacy upon such objects that gives them a parity, at least of sorts, with their previous print counterparts. Indeed, the frame reads: this work is significant enough to be preserved for posterity. In a world in which many remain sceptical of the digital – believing it to be ephemeral (although have *you* ever tried getting something you want removed *off* the web?) – this legitimation function should not be overlooked as a significant publication feature; albeit, again, devolved away from an active publication agent.

To return to my earlier point, though, it also confers upon the digital object an economic quality. Most acts of free online web publishing – including, say, the formal versions of open-access academic publishing, which include peer review etc. – have the de-economizing functions of open indefinite and free dissemination. Yet, as today's report notes, in the non-print legal deposit system, articles that we publish at the Open Library of Humanities are restricted to on-site access by a single patron. We have no way to signal an exception that we want broader dissemination of this work than the assumed economic default of rivalry. Indeed, one of the instrumental values that the current system of non-print legal deposit *forces* upon works is an economic scarcity. It presumes that one wishes to sell the work and that this is the publication and authorship rights that must be protected. Yet, I query whether this is actually the case in the digital world. It seems more likely, to me, that the publishing industry will move to become a supply side and distributed / unbundled / disaggreagated set of services – as it already is in the world of academic publishing. Certainly, one might still wish to protect certain author rights – say those who write fiction. But there are also categories of authors for whom the prime motivation is to be read, rather than to sell – and these two phenomena are in tension with one another.

Or, at least, that's the theory. Of course, just because the digital world seems to effect a de-economization for dissemination doesn't mean that society has kept pace. We still operate in a world run on the exchange of material artefacts (which are scarce), mediated by human structures of finance. That said, some economic models are better suited to the digital world than others. Namely: systems of patronage make unlimited online dissemination compatible with the digital world. There are some good examples of this. Patreon, for instance, is a system whereby "fans" can purchase recurring monthly "memberships" in order to support an artist or writer or any other type of "content provider", as the comedian Stewart Lee might humorously term them. The *Guardian* newspaper made a similar move in recent years where, despite all its content being free online, it

has managed to solicit enough voluntary contributions from those who wish to support it to break even, from a previously dire financial situation.

Such models are prone to free riders; those who will not pay because they realise they can gain access without so doing. Sometimes these models provide exclusive perks to members in order to entice them to continue to pay. Sometimes, they simply point out that, if all consumers were free-riders, the producer would not be able to exist and they can thereby solicit contributions from more ethical audiences.

[SLIDE]

Academics, though, are in an interesting category here. In the UK, academic research is funded via a number of streams, one of the most significant of which is the quality-related (QR) funding dispensed by Research England through the Research Excellence Framework (REF). This comes from the public purse ("taxpayer money"), but academics are paid their salaries in order to produce research work to which they own the copyright and that they are free to give away to whichever publisher they choose. This is one of the most compelling and frequently cited arguments for open access; academics are paid by the public, so why should the public not be able to read the work that they fund? Of course, many academics do not have full-time, permanent jobs and are precariously placed. But academics of any ilk are very unlikely ever to be able to raise enough money through the sale of their work to earn a living – after all, what is the market for "T cell receptor antagonist peptides are highly effective inhibitors of experimental allergic encephalomyelitis"? This is an authorship group who write to be *read* rather than writing to *sell*. It is, in some ways, a *freedom from the market* that allows the academic *freedom to investigate niche topics*. This remains, I believe, a good rationale for open access.

This category of patronage-driven authors might be willing to pay Google, or other "new publisher-like" services, for amplification and filtering services that tip in their favour, so that

readers can access and discover their material for free. But the underlying assumption of how digital legal deposit should work for such authors is badly skewed. These authors do not want the re-economization of their works. They want information to be free, as the now-old adage goes. There is, beneath non-print legal deposit, as it currently stands, a specific set of economic assumptions about authorship that held mostly in the non-digital realm. It is the economic scarcity correlation.

Yet, if policymakers have seemingly misunderstood the new digital context – or have rather simply been constrained by path-dependency of print – authors often misunderstand the nature of digital preservation. As above, many believe digital preservation to be a technical problem. They say: "look at format degradation and look at link rot – digital resources do not have the stability of print". Yet, as Kathleen Fitzpatrick reminds us in her excellent book *Planned Obsolescence*, the reason that print endures is not due to any inherent quality of permanence, but rather due to the fact that we have invested in institutions called libraries and people called librarians whose role it is to *ensure that permanence*. **[SLIDE]** This echoes the thoughts of the well-known digital preservation expert David S. H. Rosenthal (DSHR), who has noted that the problems of digital preservation are all socio-economic problems. That is, with infinite resources, one could preserve any and all kinds of digital object forever. Because you'd have someone whose job it was, every day, to ensure that you could access the digital object. And if it didn't work, they would fix the underlying problem. (And, by the way, the most common argument levelled against digital preservation by those who are new to it is: "what about an apocalyptic failure of the electricity grid?" I hate to break it to you, but if that fails, there are bigger problems for the entirety of the world's knowledge, which relies on temperature and humidity controlled storage, for instance.)

More specifically, though, the problem is the economic correlation between scarcity of provision and a belief, in the digital realm, that we can and should collect, store, and preserve *everything*. Part of this attitude comes from a Silicon Valley technical solutionist mentality. Google's corporate mission to "organize the world's information and make it universally accessible

and useful" has become the *de facto* universal standard against which serious digital information initiatives are meant to compete. But Google is a *terrible* advocate for digital preservation in many ways. **[SLIDE]** Not only does it perpetuate the myth that preservation is a purely technical problem, but it is also guilty, continually, of shuttering no-longer profitable services that it cannot continue to maintain, but upon which userbases depend. "The world's information" is discarded by Google and made neither accessible nor useful when they will no longer maintain various applications. Somehow, though, this apparently doesn't count because, they will say, it is not a technical problem – "just" one of resourcing. But that problem of resourcing *is* the central problem of digital preservation.

In other words, even while we often base our presuppositions around digital preservation on a technical mentality promoted by Google, even this organization understands that it is a *principle of economic selectivity* or, as I have returned to multiple times already, *scarcity*, that drives what we can preserve. Again, this can be considered in a ratio. Objects that are abundant are likely to remain accessible. PDF files, for instance, are so widely used that the effort of employing someone whose job it remains to ensure the readability of this openly specified format is a very high objects to people ratio. By contrast, the custom format that I have invented for storing my own data and that is used only by two people worldwide, for a fictional example, has a much higher ratio of investment for preservation.

And this is where we can come back to Bhaskar's notion of the importance of filtering in publication. When there is some form of material constraint on publication – say, the overhead and unit costs of printing, distribution fees, warehousing, author advances, legal costs, and so forth – there is an economic filter built in to the publication process. This doesn't mean that nonsense wasn't published, just rather that only people with money they were willing to spend could publish nonsense. In the current digital environment, though, many of these unit costs can be eradicated – the costs inhere in cost to first copy. Different types of publishers will invest different levels of

effort in creating that first copy. Academic publishers who organize peer review, copyedit, typeset, proofread, and digitally preserve their titles actually have most of the same costs as they would otherwise. Less reputable publishers who do not undertake such tasks, though, can still claim to have "published" material, but they encounter fewer economic checks on the volume that they publish.

Legal deposit then picks up the pieces of this twofold attitudinal approach: that we must preserve everything and that more and more can be published. Yet, delving deeper into this attitude of what we should preserve – and whether it is "everything" – is worth some of our time today.

I want to consider here the case of retractions from the scholarly record. When a piece of scientific work has been shown to be incorrect, unethical, fraudulent, or otherwise inadmissible, the standard procedure at an academic journal is to issue a retraction. That is, the piece is clearly marked as "struck from the record" in order to show that it should no longer be cited or considered valid. Usually, it is still possible to *read* such work, it's just that it is marked as retracted.

Is this work that we want to preserve? Do we want our national deposit libraries to ingest work that is incorrect, unethical, or fraudulent? I think it actually depends upon whom you ask. Consider the example of Andrew Wakefield's notorious paper on MMR. The overwhelming consensus in the medical profession is that this paper has done immeasurable harm and that the anti-vaxx movement is now single-handedly responsible for the return of preventable and extremely dangerous diseases in nonetheless economically developed countries. I haven't run a survey, but I would anticipate that many in the medical sector would suggest that this paper should not be preserved for posterity. However, if you ask a *historian of science* whether this work should be preserved, the answer will absolutely be a "yes". For how else are we to understand the societal contexts and histories of the anti-vaccination movement without access to the original source materials? There are very different instrumental use cases for the same material here, one based on

appraising the truth of the work (at which it fails) and the other based on a historicizing and contextual approach that values its falsehood.

Perhaps another good example of this is to be found in the field of software and digital document preservation. Certain types of document – such as Microsoft Word files – can be vulnerable to worms, viruses, and other forms of malware, usually encoded within their macro procedures. Should we preserve these malicious artefacts along with the document itself? Again, it depends upon whom you ask. The person who wants to read the original document, without the threat of damage to the computer system on which they are viewing the object, would probably prefer a cleaned-up version of the file (although who is going to do this cleaning up? This is likely, again, a lot of work in the context of rare or long-forgotten malware). On the other hand, the historian of computing might precisely want to study those viruses and malware and how they functioned within the document-object's historical context. Value for preservation is in the eye of the beholder and I would argue that it is very difficult for us to predict, in advance, which phenomena will be wanted by future unknown audiences.

[SLIDE]

Thus, despite my scepticism, I come to a "preserve everything" conclusion, even while noting the absolute economic untenability of that stance. We *must* select what to preserve in the present – and that includes the twofold functions of storage and making accessible – and we *must* consciously select what to discard. History will likely judge us poorly for our decisions to discard, but the contemporary economic principles do not offer us the option of keeping everything accessible, even if they do give us the possibility of *storing* everything. But, as with those who opt to be frozen upon their death here in the present, there is no guaranteed possibility of future resurrection for our stored artefacts, without continuous accessibility checking.

There are also questions of retroactive subject permissions that circle around this mentality of preserving everything. Istvan Rev tells of the problematic situation he faced in his archive of material on the Rwanadan Genocide. A previously broadcast clip from a BBC documentary featured a woman who incriminated herself on camera – thereby putting herself at great personal risk – who then retrospectively requested that she be removed from the footage. The material had, again, already been broadcast (one could call this analogous to publication). However, there is a substantial difference between a one-off broadcast and open, unlimited, online dissemination of the same footage. How does one balance these matters of delicate personal consent, data controller issues in the era of GDPR and the "right to be forgotten", and even the potential harm of these subjects against the need to create an archive of material for posterity, and an archive that must, in some way or another, be accessible.

This brings us back, full circle, to these matters of open access to national non-print legal deposit. Is it a difference of type, or degree, to have only on-site access to preserved materials. Let us say, in the above case, that the archive refused to remove the video for the sake of history, but permitted only on-site, rather than on-line, access. How would one feel, in terms of personal safety, being that woman? Could one imagine selective restriction of access to certain materials? What about embargoes? Of course, this runs the risk of the Streisand effect, calling attention to the very thing that one wishes to bury, once more, in that haystack.

This is a little bit of a straw-person argument in some ways, though. The challenge with high-volume archives such as the UK Web Archive and the non-print legal deposit system is not hiding material, it's finding it. Google does a good job of retrieving the information that most users want through a combination of weighted linking algorithms and assumptions of social homogeneity among users. The first of these is the famed PageRank method, in which sites confer "votes" for each other through hyperlinks. Sites that have already been ranked more highly carry more weight when they link to other sites. This then creates a chain of value among documents and creators. It

also led to the humorous GoogleBombing phenomenon whereby one can attempt to influence a search term by generating a horde of sites that all link to a particular artefact using the same term. This is partially but not wholly why a search for "idiot" returned a picture of Donald Trump. The second phenomenon that Google exploits is an assumption that users look for the same things. This can be helpful. If most people click a specific link in relation to a particular search term, then it is likely that this is a high-quality response that should be prioritised in future searches. The problem, as Safiya Noble has expertly shown in her *Algorithms of Oppression: How Search Engines Reinforce Racism*, is that this can lead to racist, sexist, ableist, homophobic, transphobic, and other discriminatory categories being condoned as worthy of higher placement, merely because a subset of users search using these terms.

This aside, how do we imagine that this might work in an academic-archival setting? Citation parsing and inter-reference calculation are extremely difficult in a free-text environment. In XHTML, with which Google primarily deals, you have a very well-formed structure for linking that we know how to parse. References, academic or otherwise, in a massive free-text corpus are far harder to parse computationally. Let us say that I have referenced a Borges short story in this paper. Except I didn't, I referenced a made-up Borges short story. Our computational natural language parsing is not good enough to work out that this is a reference to a fictional, non-existent text. This causes us substantial difficulties if we want to weight return results by documents that reference other documents, in a totally free-text fashion.

Secondly, at present, we don't have a huge userbase for the non-print legal deposit archive; because it is on-site only. This means that we have a very specific subset of users who are working with this archive and they are unlikely to represent a broader population. In other words, we have a situation in which, if we prioritise discoverability by the current userbase, we are hardly reflecting a broader potential grouping. We also need to ask how helpful this is in the context of a historical archive. People are often, in such contexts, searching for the rare and exceptional, not the same

thing as everyone else. All of this poses huge problems for discoverability within a system of non-print legal deposit.

[SLIDE]

Ultimately, though, and here is where I want to begin to conclude: it is easy to feel pessimistic about the prospects for the collection, preservation, and ongoing access to our digital cultural heritage. We have mindsets conditioned by print path dependency and a simultaneous culture of universal preservation coupled to place-restricted access. We are not sure what we need to preserve, so we opt to think that the answer should be "everything", without thinking through the resource implications of this in the present and the future. We have heterogeneous user categories who want different things from an archive; some, seeking truth, others seeking contextual falsehood for historical understanding. We have needles and haystacks, with very few effective magnets.

Yet these unanswered queries, these unresolved questions are why I feel that today's report – on *user-centric* evaluation of the UK Non-Print Legal Deposit – is a valuable frame. For we do not want, to return to my non-Borges short story, to end up with The Sacred Unreadable Artefact. We must think beyond the theology of the gatekeeping priesthood, beyond the economics of print scarcity, and beyond the false idealism of digital abundance, to consider the fundamental question: for whom is legal deposit designed? If the answer is "nobody" – or if we cannot identify a substantial audience – then, I would suggest, we have a far bigger problem.