**Submitted on:** 14/08/2014

# The paradox of selection in the digital age

**Titia van der Werf**
Netherlands Memory of the World National Committee, The Netherlands.
E-mail address: titia.vanderwerf@oclc.org

**Bram van der Werf**
Van der Werf Technologieadvies, The Nederlands.
E-mail address: vanderwerfbram@gmail.com



**Abstract:**
*With this essay, the authors aim to contribute to the digital preservation thinking by highlighting some of the aspects of the digital environment that seem pertinent to digital heritage. It gives a high-level perspective of changes in the way digital content is produced, consumed and perceived. With the massive shift towards open data and open access, the changes in the logic of economics and the thriving of digital consumerism, it asserts that digital waste production and the corresponding carbon footprint are becoming a huge challenge in the digital age. In particular, the authors state that since digital waste and assets cannot be distinguished from each other, this poses a greater preservation challenge than technology obsolescence or other digital preservation issues. According to them, the traditional models for selecting and preserving content as heritage no longer apply. Being digital flips the selection model from production filtering to consumption filtering. The authors propose the trias hereditaria – a collaboration framework between the information industry, the public authorities and the cultural heritage institutions – to balance the forces at play in the digital environment and to stimulate digital behaviour in positive ways. Raising awareness, cleaning-up waste and personal archiving are examples of measures that can help promote digital heritage. This essay is meant to be thought-provoking and to advance the collective thinking looking for answers to the challenges of digital preservation. To that effect, it is written in a style inspired by Nicholas Negroponte's Being Digital and draws heavily from his work and terminology.*

**Keywords:** Digital Heritage, Digital Preservation, Selection, UNESCO

## Preface

In 2003 UNESCO issued a Charter on the Preservation of Digital Heritage[1], which states that this heritage is at risk of being lost due to its digital nature and urges the responsible stakeholders to take appropriate preservation measures. More than ten years later, the sense of urgency has not diminished. Nor have the concerns of securing an authentic record of the digital heritage been reduced.

It is telling that the UNESCO/UBC Vancouver Declaration on Digitization and Preservation, released in 2012, speaks of the need for a better understanding of the digital environment in order to establish

*"digital preservation models that respect fundamental legal principles enshrined in institutional regulatory frameworks, and balance access with privacy, right to knowledge with economic rights, and respect ownership and control of indigenous cultural heritage and traditional knowledge in digital format*"[2].

This statement shows how digital preservation, initially perceived as essentially a technological challenge, is now much more conceived of as a practice that needs to take into account the norms and values of society. We would argue even more specifically: the norms and values of digital society – a space that has become a major part of the world we live in today and that we are shaping by the way we live our lives as digital consumers, citizens, workers and individuals. In the digital environment, societal norms and values around access to information, privacy and ownership are shifting; the logic of economics is changing; the digital divide between North and South becomes more evasive; government administrations and legislation lag behind developments – issuing ad-hoc and not very long-standing regulatory laws, directives and incentives. Balancing the forces at play in the digital environment is an evolving and continuous process and we are only beginning to understand digital society and how it interacts with our analogue world.

With this essay we wish to contribute to the digital preservation thinking by highlighting some of the aspects of the digital environment that seem pertinent to digital heritage and in doing so, understanding better what digital preservation might mean in our emerging digital society.

We propose to do this by following the footsteps of Nicholas Negroponte, one of the first and most influential thinkers on digital futures. In his introduction to "*being digital*", he writes:

"*The methodical movement of recorded music as pieces of plastic, like the slow human handling of most information in the form of books, magazines, newspapers, and videocassettes, is about to become the instantaneous and inexpensive transfer of electronic data that move at the speed of light. In this form, the information can become universally accessible. Thomas Jefferson advanced the concept of libraries and the right to check out a book free of charge. But this great forefather never considered the likelihood that 20 million people might access a digital library electronically and withdraw its contents at no cost."*[3]

[1] The Charter on the Preservation of the Digital Heritage can be found here.
[2] The UNESCO/UBC Vancouver Declaration (September 2012) can be found here.
[3] Nicholas Negroponte *Being Digital*, New York, 1995.

We will draw from his work and use his terminology, where appropriate. His contrast of atoms versus bits continues to prove useful, as do his evocative concepts of "*the bits-about-the-bits*", "*digital butlers*" and many others.

**From technological quicksand to bit flooding**

Negroponte did not discuss what "*being digital*" means for our capacity to remember our digital past nor explain how digital technology impacts our digital heritage. Published in the same year as Being Digital, Jeff Rothenberg's Scientific American article[4] brought the problem of digital longevity to the attention of a larger audience. His message was: digital formats and the programs that render them become obsolete very quickly, as do the computers on which they run, making it unlikely that our grandchildren will be able to use the digital documents, records and art-works we are currently creating, unless we take appropriate action. His call for action was picked up by a number of large national archives and libraries in the States and in Europe and was the start of a new research area: digital preservation. With Rothenberg's paper and the ensuing polemic around viable technological solutions, the tone was set. Technology has been in the foreground and in sharp focus, but the background is a blur. On hindsight, developments happening in the background have had more impact on digital heritage in the past twenty years, than technology obsolescence. Some of these developments, like the unlimited drive of individuals to self-publish on the web and the conscious policy of public authorities to make public sector information freely available on the web, are contributing to the information overflow that we all experience today and to another digital preservation challenge: selection.

**More is never enough**

Beginners don't understand 'less is more'. Negroponte's account of the first use of home video cameras – with endless pans and zooms - is very recognizable. The reaction was similar – at an even larger scale - when the camera cell phone appeared on the market. Since then – some 15 years ago – more has not become less. Everywhere we go, we record, document, and upload the minutiae of our lives. When we visit the Eiffel Tower, we do not look at the iron construction - instead, we immediately lift our phone to capture an image that everybody in the world has already seen a million times. More is never enough.

As consumers, we are getting used to taking pictures of everything and anything, all the time. This is a golden opportunity for businesses to dream up services that help us make the best of the instant pics (or audio, or video; there's little that modern cell phones can't capture) that we share on social media sites: think fashion, think food, think DIY, etc. Take for instance the "Photo a Day" apps – they let us capture our everyday moments and make journaling our life easy and simple. Or apps that let us share with friends what we bought in web shops, or what we didn't buy but liked, while perusing online showrooms.

Consumer behaviour is of all times. It is not a generation xyz thing. Generation x does it on Twitter, generation y with the mobile phone. Over time, consumerism shifts. As the way we engage with the world moves from atoms to bits, our consumer behaviour shifts from things we own (cars, houses, TVs, etc.) to information or experiences we share (music, games, pictures, etc.). And, with this shift come other, related trends. One is that information costs

[4] "Ensuring the Longevity of Digital Documents" Jeff Rothenberg, Scientific American, January 1995 (Vol. 272, Number 1), pp. 24-29.

nothing to consume. From images on the web to open source software and tools, from free apps to streamed digital content, never before have we had access to so much without having to pay a cent for it, and this is impacting our expectations as consumers. While we are expecting to upload and download more bits and for free, we do not seem to be too much bothered about the quality of the content – as long as it moves easily up and downstream, to and from the far corners of the web. Ever higher resolutions and more bandwidth on the web are taken for granted; owning the devices – the atoms – that support these higher resolutions is still seen as a status symbol. We don't mind viewing a low-quality recording video on YouTube, but we take pride in having the latest and most performing quality recording device. It is like driving faster cars. Even if we are not allowed to drive faster than 120 km/hour, we are obsessed with having a car that can drive faster. With bandwidth, it's a similar story. To quote Negroponte:

"*There really are some natural laws of bandwidth that suggest that squirting more bits at somebody is no more sensible or logical than turning up a radio's volume to get more information."*

He doubted the dogma that says "*we should use high bandwidth just because we have it*". Likewise, we could ask ourselves: Should we be recording and sharing every minute of our lives just because we can? Probably not. But consumerism follows its own logic, and it is not sensitive to moral motivation.

**The biggest recording machine of them all**

Not only do consumers record endless amounts of bits, the whole public sector does so as well. It registers every bit of information about us - as civilians and tax-payers - and files every aspect of our life: health, education, employment, contacts with law and order, etc. It lists and matriculates all assets and properties, be it cars, houses, land or pets; indexes companies and their stock prices; encodes with minute precision the characteristics of all natural resources that fall under its jurisdictional control: climate, sea life, forests, wildlife, soil composition and minerals, etc.; collects data for research and registers all scientific outputs of universities; catalogues all cultural heritage artefacts kept in memory institutions. A growing number of cameras mounted in traffic intensive areas continuously monitor the highways and roads. With aerial photography every square centimeter of a state's geographical territory is recorded. We could go on and on and on … and let's not forget all the intelligence information that is gathered for national security reasons.

Authorities have always collected data and have done so more or less systematically over the past centuries. Obviously this is more the case in the world's highly regulated societies and, since the advent of the digital age, in the world's most digital public sectors. Digital data collection can be much more granular and voluminous, thanks to the capacity of computers. This allows for fine-grained modeling and simulation of real-world situations – which in turn allows policy makers and decision-takers to better address real world questions. And the potential is huge. Climate modeling is probably the biggest success story of big data modeling. We have all witnessed the significant improvements in weather forecasting over the past twenty years.

Is "*being digital*" for public authorities the license to become data farms? They certainly have a compelling story to justify it.

## Pump up the digital economy

The reuse of public sector information has become the cornerstone in the "Open data" and "Open government" strategies of public authorities. Whereas the "right of access" to information held by public authorities has always been restricted by the pen-and-paper bureaucracy and commercial exploitation of public information an unfulfilled dream of enlightened policy makers – open data is now seen as a pillar for democratic transparency and a lever for economic growth. In making the data digitally available for re-use at marginal costs, authorities hope to stimulate the markets which can generate new innovative businesses and jobs and provide consumers with more choice and value for money. Think of GPS, weather forecasts, financial and insurance services.
At the same time we see in countries with a well-developed public service, that governments are retreating and privatizing public services for which they collected the data in the first place. The running of public transportation, hospitals, schools, cadastres, weather forecasting and other services is handed over to the private sector together with tightened regulations. In countries with a history of lesser centralization, governments are tightening their control over the private sector as well. Both trends converge to support the transition of public authorities into digital governments that control the execution of the ever expanding public service provision by private parties and the corresponding collection of data. Taking over control of the banks during the financial crisis in 2008 is just another such example.
The movement towards more openness of public sector information and privatization of public services is extending its reach to cultural, educational and research establishments and public service broadcasters – the core of the traditional heritage sector. This is an area where, if we take a closer look at it, we see how the public authorities are having trouble in imagining the right incentives for realizing their transition to being digital.

## The gold open access fiasco

Making all the country's publicly funded scientific research free for anyone to read, is one of those government policies meant to benefit the content industry and thereby the country's economy. Scientific articles used to be accessible only through very expensive journal subscriptions. They are now becoming available as "open access" articles – which means accessible to anyone online and for free. This is, however, not such a radical change as it seems, because the scientific publishing industry is being allowed to keep doing business as usual. Their business is one of filtering papers and packaging them into higher or lower ranking journals. Theoretically, one would have expected the public authorities to make research papers available as "open access without filters and without ranking stamp"- in other words "equal from the onset". Sure, not all research bits are equally worthy of the space they take, so filtering and ranking stays very important. In line with expectations, the content market would develop innovative web-based filtering and rating services that are more fit for purpose than the traditional paper-journal model. However, an unexpected flaw in public authority reasoning occurred, no doubt under the lobbying pressure of the publishing industry. The current regulations being put in place in the UK and in the Netherlands favour "gold open access", meaning that the scientific publishers get paid up front to publish public-funded research articles in their journals. The filtering-packaging-and-rank-stamping system stays in place, the publishers' revenues are secured and the authorities agree to pay market prices for publishing their own public sector information as open access. Now that public-funded research can pay its way into scientific journals, peer-review - the quality filter that scrutinizes the scientific value of a paper – seems to be by-passed altogether.

The gold open access example shows the hiccups on the path to open data and open access. Regardless of the price to make it happen, public authorities seem pretty determined to make public sector data and information available for free on the Internet. The unsuspecting civilian would almost believe it didn't cost anything – the tax-payer in him not being aware he is indirectly covering all the public sector information expenses. To the public at large it seems as if all these bits are public domain because they are available to everyone at no cost. And nobody owns them – or so it seems.

**When we share, do they care?**

Although the digital commons are still somewhat nebulous, it is striking to observe how different groups are trying to appropriate the commons for their own cause. In the digital commons people share their software code (SourceForge and GitHub), their knowledge (Wikipedia), their pictures (Flickr: The Commons), etc. The commons favor use and reuse, rather than ownership and exchange as a commodity. The digital commons movement comes in many flavours: from anarchism to collective activism and they all have some form of idealism in them. As the commons are branching out in different directions, one new young shoot is growing rapidly: the commons of digital heritage. And we are seeing cultural heritage institutions moving into that space as contributors of digitized content, advocates for removing copyright barriers, guardians of the public domain and more recently, stewards of the commons.

The digital heritage space, that asserts it belongs to the digital commons, is populated with the digitized content of analogue collections held in libraries, archives and museums. This space is very much fragmented, with numerous digital library websites, digital archives, museum of the future and cultural discovery initiatives such as Europeana and the Digital Public Library of America.

The main drivers for providing these free content services are best expressed in their own words:

"*We, as memory organisations, have the wealth of human knowledge and experience within our collections and it is our responsibility to share that with the world – we should seek to educate, to enlighten and to entertain. And increasingly, our ability to share is becoming ever more feasible because, just like a candle's flame, when we share digitally we enable lots of other flames to be lit at little cost other than our initial willingness to share.*"[5]

These so-called "digital library" initiatives are the result of projects initially funded by public monies and they are largely unsustainable on the long run. Indeed the spectre of having to justify the underlying economic model of such approaches looms large over the digital libraries. On the one hand there is the realization that cultural content should be made freely available because digital content has become a "ubiquitous utility" – not something anyone would consider paying for. On the other hand, governments and other funders are "obsessed" with the economic consequences of maintaining such digital libraries into the future.

The root of the problem is that they are set-up as a public service, not as a proper commons initiative or as a commercial service. And, public authorities do not want to run such services.

[5] This is a citation form Simon Tanners's Keynote address at *Sharing is Caring* (Copenhagen, April 2014) posted on his blog *When the Data Hits the fan!*

Their purpose for investing in digital library projects is to make the content of libraries, archives and museums available as free, online digital content to boost the creative industries. So they have been requiring for some time now, that the projects and initiatives sustain themselves on the long-run by developing business-models and by entering into public-private partnerships – all of which is inherently not what the commons are about. The commons are based on voluntary effort and donations. Membership-based financing of such initiatives, drawing from public-sector institutional budgets, provides only short-term relief – as there is very little stretch in these budgets. As yet we have seen very little uptake by the commons themselves. So who is going to adopt this new content and take care of it in the long run?
The reason why this digital library adventure is worrisome, from a digital preservation perspective, is that it is distracting memory institutions from tackling the preservation of "real" digital heritage: bits that do not have atoms as counterparts in the analogue world. The content shared in the digital commons consists mostly of digitized books, newspapers and photographs – copies of analogue originals that have been selected in the past and are preserved in physical stacks. Calling all attention on the preservation of digitized content under the guise of digital heritage honestly seems bizarre – when we know that an abundance of newly born-digital heritage content is produced on a daily basis which needs to be catered to.

**Follow the golden clicks**

Using and sharing content on the internet is altogether largely free – not just in the digital commons, also on the platforms owned by large private companies. You can search Google for information on the Internet, Skype with your relatives at the other end of the world, chat with friends via Facebook, link up with colleagues via Linkedin, play music from Spotify or share files via Dropbox – all for free – or so it seems. By now, we know that we are paying a price: that of "sharing" our personal data. We are all implicitly and explicitly allowing the services we use to collect information about ourselves and to track our online behaviour. The large sites, those where we reside day and night, are collecting enough data for the compilation of personal profiles with detailed information about our health, sport preferences, socioeconomic status, family connections, emotional concerns, etc. This information feeds the backbone of digital marketing and advertising, which in turn finances free content and services on the web. So despite appearances, the information and social media services are not about the content, they are about our digital footprint as internet residents.

IBM's CEO Ginni Rometty, predicted in 2013:

"*What you will see with rapid data and social sharing is the death of the average and the era of you. Business will be able to truly serve the individual."*[6]

She is bringing to life Negroponte's interface agents or "digital butlers" – whom he described in 1995 as living both in the network and by our side, acting as our personal filters and delivering the information we want at the time we need it. She is echoing what he predicted: "In the near future, the filtering process will happen by using headers, those bits about the bits". The bits about the bits are not only metadata about the content and the products; they are also metadata about us, the individuals.

[6] Published on the website of Forbes on 03.08.2013.

According to the 2011 Digital Universe study :

"The amount of information individuals create themselves — writing documents, taking pictures, downloading music, etc. — is far less than the amount of information being created about them in the digital universe".[7]

In other words the bit trail we leave behind on social media, search engines and websites is far more than the bits of information we generate ourselves on such platforms. Ponder that!

**It's getting hot in here**

When you were listening to your favourite playlist on YouTube, with Google Chrome or Apple Safari, did the noise of your laptop cooling fan predominate? Well, that's probably because you hadn't disabled the ad blocker in your browser and what you heard was an overheated computer trying to process Google or Apple's tracking application on the background.

In August 2013 science.time.com's headline read: "*The Surprisingly Large Energy Footprint of the Digital Economy*"[8]. It revealed that the digital economy used a tenth of the world's electricity – a share that is only increasing as we keep on pumping more and more bits through the ICT-system that connects us with the Internet via smartphones, laptops and digital TVs. The electrons follow the bits. They are consumed by the electron-thirsty computer-server farms that make up the backbone of "the cloud" and keep running 24/7, together with the cooling systems that prevent the computers from overheating. We hardly realize the consequences for the environment.

Following the logic of more is never enough, the devices we use to interface with the digital world - which by the way, use more electricity than our refrigerator – are replaced within 1-3 years of purchase because they become unwanted or obsolete. This phenomenon – which stimulates demand by encouraging purchasers to buy sooner if they still want a functioning product or new features – not only fills the for-profit pockets, it also grows the global mountain of waste.

Ultimately, technology innovation is consumer-driven. As consumers, we buy new gadgets, new features and new design. We continuously replace old stuff with new stuff and we know that we are disrupting the environment and destroying ecosystems while doing so.

Waste is not the prerogative of atoms. Our global waste mountain of bits is becoming just as intimidating. Because each time bits are no longer used or relevant, their duplicate-bits and back-up bits also become unused and irrelevant. We can only guess how many percent of the total digital information consists of duplication, through formatting and reformatting of the same content. Those numbers are not part of the digital economy metrics. But even without metrics, we know there is a lot of it going on. Think of discussion forums that can generate both regular and stripped-down pages targeted at mobile devices; store items shown or linked via multiple distinct URLs; printer-only versions of web pages.

[7] IDC iView "Extracting Value from Chaos," June 2011, sponsored by EMC.
[8] The article can be retrieved following this link.

How many deposited bits are lying unused, like layers of sediment, on the bottom of the web? How much of it is waste? How much digital heritage? And how can we distinguish the one from the other?

## The expanding heritage record

The digital heritage record includes all things digital that a society deems important enough to keep and preserve into the future – be it cultural heritage, scientific knowledge, government information, business information, personal information. More precisely, it is the totality of the evidence of digital societal activity that has survived the past. It represents the bits that are preserved in archives, libraries, museums and a whole range of other digital archives that have a legal or self-assigned mandate to preserve the bits. Some examples of digital archives are: research data archives, broadcasters' archives, the Internet Archive, business archives, archives documenting specific historical events, like the USC Shoah Foundation, etc. In addition to these institutional repositories, personal archiving initiatives abound.

In taking decisions about what is worthy to be kept into the future, both individuals and institutions are converting societal data into historical data.

The OECD observed in 2011 that "*more data was created in 2011 than in the whole of human history, or at least, since the invention of the alphabet*"[9]. Considering the growth rate of data creation, it's no surprise that the amount of historical data collected by memory institutes is growing at a fast rate as well. The backlogs of unprocessed materials piling up in the archive stacks are daunting. They can reach several tens of kilometers. There is now also the growing backlog of computer files, databases and email-archives sitting in data storage facilities waiting to be processed – the volume of which is not yet part of their metrics, a sign that digital preservation practice is still in its infancy. It seems reasonable, from a tax-payer perspective, to expect memory institutions to tighten their selection criteria in order to keep the accretion in control. This is not altogether what is happening. If we take a closer look at what memory institutions are trying to keep, we see an ever expanding heritage record.

## Less is more: no need for selection?

Over the past centuries, public archives have evolved from keepers of "left-over" artefacts into systematic collectors of public authority records as evidence. By law their task is to keep such evidence in order to assure the transparency of government operations. They need to preserve the record of past governments that have, through time, steadily become the biggest recording machines of them all and have expanded their regulatory system into all corners of society. They also need to deal with requests for information from civilians, journalists and researchers who are eager to inspect the most recently disclosed records. In 2014, the National Archives of Australia reported a backlog of "*12,453 applications from the public dating as far back as 2009 and all of these should have been dealt with within 90 days.*"[10] With the promotion of open government principles and open data, expectations of the public are mounting and are changing in disillusion because archives cannot deliver.

[9] Quoted from Martine Durand, 'Can big data deliver on its promise?' OECD Observer, November 2012.

[10] Quoted from the Phillip Thomson 'National Archives' crackdown on massive backlog' in the Canberra Times, May 1,2014.

There are a lot of external factors that impact on the ability of archives to deliver, and there are also internal factors. The profession is very much aware of its responsibility to select records of enduring value and of the fact that "*all other archival activities hinge on the ability to select wisely.*"[11] Even though 80% or more of government administration is systematically de-selected and destroyed, it just doesn't do the job. For many years now archivists have been unsuccessful in reducing the flood of atoms entering their building and have had to extend the limits of their physical storage areas in every possible direction. In the physical world, where space is scarce, the urgency to act is much stronger. In the digital space, this feeling tends to slack off, as the limiting variables for archival records - that is, in spatial terms, storable and, in human terms, usable – suddenly no longer seem to apply. Being digital means storage is limitless and, to paraphrase Negroponte, instead of reading what archivists justify as worthy of the space it takes, being digital flips the selection model and puts you in the driver's seat – with your interface agent acting as your personal filter and delivering the information you want at the time you need it.

This thinking misses two very important factors. Firstly, digital storage is not limitless because government budgets are limited and because it affects our physical ecosystem of atoms. We cannot just simply keep storing all the bits being produced over time. Secondly, not all the bits are worthy of the space they take. There is a lot of trash out there, a lot of duplicate bits and a lot of bits nobody will ever care about anymore. We need filters to massively de-select the irrelevant, redundant, meaningless bits and bit hoovers to clean them up. We need archivists in all the corners of the digital realm to de-select and dispose of 80% or more of the bits we produce and to keep the worthy bits, from which Negroponte's digital butlers will select and compile *Daily Me's*.

**Collecting the web as a document**

De-selecting is not typically what libraries do. Their practices and workflows have developed in the past centuries around building collections. They think in categories of published information: academic publications, national imprint, literature, fiction, cookbooks, etc. National libraries have a long-standing cultural heritage task to record and preserve a country's published output. Their instrument is twofold: the national bibliography, a registration system that lists "all" the documents published in a country and the legal deposit that secures and preserves "at least one copy" of every publication published in a country.

The example given by the Bibliothèque nationale de France[12] is illustrative of the expansion of the legal deposit's collection remit in the past 500 years:

| **Type of material** | **Year** |
|---|---|
| Printed material | 1537 |
| Prints, maps and plans | 1648 |
| Sheet music | 1793 |
| Photographs and sound recordings | 1925 |
| Posters | 1941 |
| Videos and multimedia documents | 1975 |

[11] 'Planning for the Archival Profession: A Report of the SAA Task Force on Goals and Priorities (Chicago: Society of American Archivists, 1986), 8.

[12] Data from: Peter Stirling and Gildas Illien 'The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future', a paper presented at the IFLA 2011 Puerto Rico Conference, retrievable here.

| | |
|---|---|
| Cinema | 1977 |
| Multimedia, software and databases | 1992 |
| Internet | 2006 |

From 1995 onwards, most national libraries have steadily extended their remit from atoms to bits and in doing so their task became very much more complex. In 2008 electronic resources were still seen as a "material type", distinct from other material types such as books, journals, newspapers, audio-visual recordings, etc. Online resources were characterized as "remote access" and resources stored on off-line carriers (CD-ROMs) as "direct access"[13]. Just to show how difficult it was in those days – six years ago - to understand the nature of digital information.

When all published media are bits and distributed online via the internet or broadcast via TV and radio, the collection space becomes somewhat less complex and more homogeneous – one would think. Well, it doesn't. It turns out that on the internet publishers are no longer the only players in the field and there is a lot of self-publishing going on. In theory, publishing is an act of making something publicly known – therefore, all information resources accessible on the web can be considered to have been "published". This way of thinking explains why national libraries have been harvesting resources from the web for some years now, as part of their legal deposit task. The British Library is legally allowed or obliged (as you prefer to see it) to crawl the UK-web since 2013. As a result they are now harvesting 1TB a day. Like their colleagues in other countries, they are archiving mostly static web pages and providing mainly document-centric access. They select web resources, sometimes at the "item level", like analogue publications, and stop following the hyperlinks when they reach the borders of their national web domain. The web archive is like a huge historical document, in which the web has been deconstructed, flattened, parceled out by country of origin and no longer interactive.[14] This does not feel like the way forward.

**Documenting the web**

Even if information on the web is generally available as recorded bits, most of these bits are not meant to be treated as publications. Most information on the web is for interaction purposes via website shops, e-government sites or social platforms and is comparable to verbal and non-verbal information exchanged during interactions in the physical world. Libraries have never recorded and preserved conversations of people in streets, shops, information booths, restaurants, concerts … why would they want to do it on the web? In 2010, the Library of Congress proudly announced that they had signed an agreement with Twitter to receive all its public tweets, from day one of the company's inception. Many national libraries take periodic snapshots of their countries' entire web domains – like taking a picture of the DNA of digital society. They do it, because it's possible: the bits are out there, available for harvesting and that is enough to raise a collector's blood pressure.

Whilst the onus used to be on publishers to push their paper-based publications into the legal deposit, being digital flips the responsibility to the national library to pull publications into its deposit. We see the filtering mechanisms being inverted and it is clear that librarians, being collectors at heart, have difficulty in operating the lock gates to digital heritage. And it is getting hot in there.

[13] For example in Maja Žumer (ed.) *Guidelines for National Bibliographies in the Electronic* (Draft), 2008 that are accessible here.

[14] from Helen Hockx-Yu's presentation at the WebArt CATCH meeting – 19 April, 2013.

The British Library gave an impression of its carbon footprint in 2010 (before they started to systematically archive the web), as follows:

- The BL's digital collections amount to nearly 100Tb now, expected to grow to 300Tb in 2013 and 2Pb in 2018;
- The files are stored in 3 large server banks, at St Pancras, Boston Spa and the National Library of Wales;
- The servers run 24/7 – they use energy to run and energy for cooling;
- It is estimated that for every £1 spent on running servers, £1.20 is spent on cooling.[15]

We not only need digital archivists to clean-up our digital information society, we also need dikes and locks to contain bit flooding of the digital heritage space. Don't ask librarians to be the gatekeepers. Their strengths lie elsewhere: metadata.

**The bits-about-the-bits**

In Being Digital Negroponte explains the power of metadata, as the best way to deal with a massive amount of data:

"*In the next few decades, bits that describe the other bits, table of contents, indexes, and summaries will proliferate (…). These will be inserted by humans aided by machines, at the time of release (like closed captions today) or later (by viewers and commentators). The result will be a bit stream with so much header information that your computer really can help you deal with the massive amount of content.*"

We are seeing that the big Search Engines that index the web are becoming increasingly "intelligent" in recognizing what interests you and what the content is about. They are becoming much smarter at uniquely identifying the entities which people search for: persons, locations, events, products, books, scientific articles, etc. They do this by collecting metadata from different sources. For book information they collect information from Wikipedia and from libraries. If you search for "Being Digital", Google will present you with a knowledge card about Negroponte's book. This card looks like a library catalogue card, with data identifying the author, date of publication, and subject. The more these data are accurate and interlinked with other related information on the web, the better they will serve you or your digital butler to find what you are looking for at the time you need it.

Digital heritage will also need to develop intelligence about itself and carry signals that future agents will be capable of understanding. Its historical features in particular will be important to capture: When was it created? by Whom? How? Why? and increasingly interesting will be the information about its contemporary usage and users. Digital data by itself is very little relevant. It only becomes relevant when it is actually used. It's about the way people use the data. This information transforms data into societal data and that's when it also starts becoming interesting as historical data.

[15] Data from the website of the British Library.

## The paradox of selection in the digital age

The boundaries of our digital heritage are stretching to the far corners of the web and expanding at the same pace as the web itself, into infinity. Common sense tells us that it is not realistic to keep the bits that we all produce, as historical bits. Memory institutions are trying to come to grips with the exponential nature of digital information. In the analogue world, they know exactly what to keep and what to dispose of; in the digital world they seem to be overwhelmed and to have trouble in selecting what to keep. *Why is that?*

The major sources of heritage - culture, science and governments - have not changed with the advent of the digital age. However, they produce MUCH more information in digital form than they ever did in analogue form and there is a perception that EVERYTHING they produce is worth keeping. In the paper society producing ink on paper and shipping books from A to B is very cumbersome compared to producing bits and distributing them across the world. There are physical constraints and also heavy filter mechanisms in place to make sure only the information worth the effort of publishing, printing, distributing makes it to the bookshelves. A publisher will set a limit to the number of pages an author is allowed to write. In the world of bits, all these constraints and filters just don't exist. As Negroponte observes, being digital gives new degrees of freedom and it takes time to use these "more adroitly and sparingly". We are not there yet. We are newbies in a new world that promises to bring the endless possibilities of the future with it. The digital commons propagate much of this new idealism and we see a hundred flowers bloom but no digital heritage garden yet being cultivated. We see abandoned projects and unattended content.

Meanwhile consumers and the industry thrive exuberantly on the web. Consumers connect 24/7 to communicate, take pictures, listen to music, play games, share jokes, like, date, rant, expose themselves, shock others, etc. The most successful internet companies are now big data processors, pulling data from the full Twitter fire hose, Tumblr, YouTube, Facebook, Instagram, and others. Their secret is the way they filter social data into something meaningful to their customers and provide brands real-time feedback on their campaigns and products. Consumers and industry are also producing the most pollution on the web and their carbon footprint is growing exponentially with their digital activity.

Public authorities are taking up the role of big data producer, joining the ranks of the large internet companies in pulling data from citizens and igniting the fire of the digital economy – in many ways aggravating the bit flooding and the carbon footprint. It took centuries for governments to take appropriate measures to clean up our material world – with sewage systems, dump sites, recycling initiatives - which have grown into an industry in itself -, regulating CO2 emissions, investing in alternative energy, etc. Being digital for public authorities does not mean they have no role in regulating digital society. No one else – not the consumers, the industry, nor the commons – will take measures to regulate the negative effects of the digital economy. Only governments can do this. Currently, they seem to see only the positive effects, endless opportunities and limitless growth.

As a result, public archives are not really encouraged to do the job they are supposed to do: namely appraisal and disposal of electronic records. Instead, they are distracted into digitizing historical archives to make these available as free, online content. For libraries and museums with a mission to collect digital heritage, the same holds true.

## The filters are us

Digital content is massive, free and open. On the web, the content bits are all treated the same. The garbage and the gems: there is no difference. There are no filters in place. Not at the waste level to distinguish waste from content. Not at the quality level to distinguish the relevant from the irrelevant. The filters that commingle the bits-about-the-bits of the content with those of our online behavior are the new gold. Obviously, they are proprietary - not open and available for free like the content is. They are the key to successful internet entrepreneurship. They match personal preferences, context and content on the fly. Pushing content to mass-audiences is no longer the game; neither is narrowcasting or niche marketing. The filters have become so granular that they can target at the individual level. Ultimately these filters are only successful if the individual does consume the proposed content – in other words, the filters are us.

What does this mean for the digital heritage? The published heritage used to be collected by piggybacking on the content filters of publishers and the media. It reflected the different market segments and audiences for which the filters were designed: from popular best-sellers to special interest magazines for car lovers, golf players or culinary food enthusiasts. There are no equivalent filters on the web for collecting the published digital heritage. National libraries are forced to re-think radically their national bibliography remit in an age in which it is not about the national production of content but the national consumption of content. They will need new filters based on use. After all, heritage is about what people have read and listened to and how that is reflected in their own writings or works of art. It is not about ideas that were never shared, words that were never read and speech that was never listened to. All that information – which bits have been used and which ones have not - is recorded in the digital space we live in. It's a matter of building smart filters that distinguish waste and trash from the digital heritage.

## Barking up the wrong tree

It is important for the cultural heritage institutions to think about the filters they want to use and about the bits-about-the-bits they need in order to be able to make their selections of digital heritage. The analogue filtering mechanisms don't work in the digital world – even if ultimately, the same selection criteria apply. What is suitable for retention and what is not hasn't radically changed in being digital. What is different is the method of filtering, selecting, de-selecting, and disposing. There are currently too many initiatives that are trying to re-define the selection criteria and that are expanding the boundaries of the digital heritage record. The focus is on the WHAT, whilst it should really be on the HOW.

Likewise, the mass digitization spurt of the past years has seduced heritage institutions into thinking that all these digitized atoms are part of the digital heritage. It is diverting their focus from the born-digital bits, which they urgently need to start taking care of in systematic ways.

And it gets worse. Some technologists, like Jeff Rothenberg, have warned against the dangers of digital obsolescence. This warning has been taken up by the community of archivists and librarians and the ensuing discourse has distracted them from thinking about their core task: to select and collect digital heritage. As a result they have spent the past two decades discussing hardware and software standards that could provide a reasonable level of confidence that digital heritage will still be accessible in the coming decades. They have talked about archival formats, format registries, software repositories, preservation strategies,

preservation metadata, etc. Even the preservation metadata focuses on technical characteristics instead of those that give machine understanding of the content and why it was selected as digital heritage.

Obviously it is important to ensure the long term accessibility of our digital heritage – otherwise all the effort put into selection will go down the drain of digital obsolescence. But it's more urgent for cultural heritage institutions to think about selection than to think about long-term accessibility. There are no signs that digital longevity is a technological problem that cannot be solved. There are more signs that digital longevity is endangered by bits that get lost because they are not taken care of.

**Scapegoating the industry**

Many digital preservation guidelines have appeared in the past years that recommend limiting the diversity of file formats selected for retention and migrating to open formats in order to manage the problem of technological obsolescence. Memory institutions have been lobbying that the tech industry should continue supporting old versions of software and be constrained to use standard open formats that better withstand the test of time. This has fuelled the widespread belief in the heritage sector that open formats are more stable than proprietary formats and that private companies are generally evil because they are the cause of accelerated obsolescence and impose licensing conditions on the use of proprietary formats. What they fail to see is that the "shelf life" of most tech-companies is shorter than ever before and that those businesses are themselves hit hardest by accelerated obsolescence. Take Nokia or Blackberry's dramatic fall from market grace. It's the marketplace that is rapidly changing and it's us, the consumers, who are driving obsolescence. We decide what will become a commodity and what will not. We always want more and different. We want to publicly say and write anything on the web, and when our digital footprint embarrasses us, we demand that it be removed. What is the default reaction of regulators? It is to impose measures on the industry. In 2014 the European Court of Justice ruled that citizens could ask search engines to remove particular links from results for a search made under their name, if the material was deemed to be out of date, no longer relevant or excessive. Why isn't it our duty to remove such materials that we have (caused to be) published in the first place? Why aren't we treated as responsible citizens? Why aren't we accountable for our online behavior as we are for our offline behavior?

**The beneficial role of government intervention**

Being digital for public authorities means they have a duty to regulate digital society. They have a role to influence our digital behavior in positive ways.

In relation to our cultural heritage, it is first and foremost our behavior - consumptive ("more is never enough") and polluting (leaving our trash behind) - that impedes the selection of digital information to be preserved. Governments can influence this behavior in positive ways, like they do in the physical world. Think of the measures taken to reduce carbon emissions from cars. Consumers are incentivized to drive clean cars by levying higher taxes on polluting cars. The growing demand for clean cars in turn instigates car manufacturers to achieve green compliance by optimizing engine performance. Public authorities can make citizens aware of their digital footprint and its effect on the environment. They can stimulate consumers to produce less bit pollution by setting targets in consultation with the industry. When consumers are sensitized to their duty to clean-up their unused bits, the new demand

for clean-up facilities will stimulate the industry to develop more user-friendly trash cans, filter applications, bit-hoovers that clean-up expired bits left behind by self-destructing links. Online games will appear in which players are challenged to seek and destroy waste bits. As the awareness of distinguishing between trash-bits and meaningful bits grows, consumers will want personal archiving facilities to keep the bits that are dear to them. The industry will invent preservation agents and additives that attach themselves to the bits and links that need to be kept. And with awareness of personal archiving, comes appreciation for cultural heritage and memory institutions. Once the importance of digital heritage is understood by the general public, there will be support for the cause of heritage institutions – not only because of their wonderful atom collections and digitized atom collections, also for their born-digital bit collections.

**The lobbying role of cultural heritage institutions**

Usually public authorities take measures when there is a sense of urgency. Too often, it is only when refuse dumpsites on the outskirts of cities start to create health problems that public authorities close them down and look for alternative measures. This is where the memory institutions come into play. They have a role to alert public authorities about the urgency of barriers that impede them from carrying out their core tasks. These barriers are best identified by focusing on the digital information landscape, not on technology. The issues that keep practitioners awake at night are those of selection. They ask themselves: "Does it make sense to keep the web as a document?"; "Is it scalable to continue selecting online resources at the item-level?"; "Do I need to keep all the blogs, tags, likes, tweets that have been published about an online resource as well?"; "How do I get rid of the trash that pollutes my daily web-harvests?"; "They say appraisal and disposal are no longer necessary in times of declining storage costs and improved discovery: what is my job's future?". As professionals, they are keen to deliver good work and they have become very uncertain about the way they can fulfill their tasks in the digital environment. Not only because they feel paralyzed by the risks of technological obsolescence, also because they have no answers to fundamental selection questions – which lie at the core of their competencies. Developing a well-informed understanding of the digital information landscape is the way forward. Understanding why the traditional filters do not work in the digital environment and why trying to control the growth of content via bibliographic control no longer works on the web. Understanding that free, unlimited online storage, even in the open cloud, costs money and pollutes our physical world – and that we therefore need to clean-up the bits and distinguish between waste and assets. With this understanding, memory institutions can start thinking in new ways to carry out their task of selection in the digital age.

**Trias hereditaria**

Improving the interactions between the information industry, the public authorities and the cultural heritage institutions – we call this framework the trias hereditaria – can potentially resolve the current digital preservation impasse. A constructive collaboration between the three parties can achieve much, like tackling the waste that ends up silting the web and triggering the organization of content in meaningful ways for consumers, citizens, workers and individuals alike. A note of caution is in place here. In no way should these interactions lead to measures of control or censorship of the web. As Negroponte warns in his after words on the growth of the Net:

"*The only hazard is government in the form of politicians who want to control it. Usually under the banner of sanitizing the Net for children, people all over the world are trying to censor its contents. Worse, some countries, including the Unites States, want to make sure that there is some means for them to listen into messages, like wiretapping. If that does not give you the willies, it should."*

In the same way as he supports "*having safe text*", we support "*having clean text*". And none of both will happen overnight. We have to want it to be so. That said, the trias hereditaria can help make it happen. It is a construct that can potentially make Cyberspace a clean space, with clear signage for web visitors and memorable heritage sites. This can be achieved through positive incentives, no need for draconian measures. Consumer behavior can be influenced for the better, like the anti-smoking campaigns have demonstrated. Consumers will be just as happy cleaning up their trash as they are producing it – as long as you use the right incentives.

If I clean up my trash and you don't, we have a problem. This is the danger of trying to regulate web consumer behavior at the nation-state level. Individual nation states are not able to impact on the wider Cyberspace, which is not about geography and national borders. It is about communities. Like the digital commons, the academic or cultural heritage communities. This is why both memory-institutions-as-a-community and UNESCO - which can act as the glue of public authorities at the international level – have a significant role to play in the trias hereditaria. We believe in their mission and hope this think piece about the paradox of selection in the digital age will guide them forward in the right direction.

**About the authors**

**Titia van der Werf** graduated in History and Informatics from Groningen University, the Netherlands. She is a well-known manager in the library and archive sector with extensive experience in the strategic application of information technology. She has worked for the Koninklijke Bibliotheek, the African Studies Centre and the International Institute of Social History. She is an expert and has published and spoken extensively on digital preservation issues, metadata and the design of repository/e-deposit systems. She was on the team that drafted the UNESCO Charter on the preservation of digital heritage (2003) and is a member of the Netherlands Memory of the World Committee. Currently, she is Senior Program Officer at OCLC Research.

**Bram van der Werf** has an MSc in Total Quality Management. He has extensive experience in improving operations, customer services and software development practices as senior manager in tech companies such as CANON, Rational and IBM. In the past seven years he has mostly worked as consultant for public services to help manage and improve the relations with industry partners, to professionalize open software development and also as crisis manager. More specifically in the cultural heritage sector, Bram has worked for Europeana (access to culture heritage) as technical director, was executive director of the Open Planets Foundation (digital preservation) and most recently, advisor for PrestoCentre (preservation of audio-visual content).