

The future of saving our past

As letters and diaries give way to e-mails and laptops, fresh challenges and opportunities have emerged for archivists. **Jeremy Leighton John** explores the digital wilderness for the British Library.

In 2000, the tenaciously original evolutionary biologist and Royal Society professor Bill Hamilton died after an expedition to the rainforests of Africa, and within months his archive was delivered to the British Library in London. Unlike archives received in the past, this one did not consist only of boxes of papers — handwritten letters, typed draft essays and the like. It also included a hoard of computers and storage media from the early 1960s onwards — from 5-hole paper tapes and 80-column punched cards through to optical disks — which occupied 26 of the 200 boxes.

Bill was my mentor, and I grew to know him well, living with his family for four years while completing my PhD in evolutionary biology. At the library, his archive came to me, along with the question of how best to deal with its digital elements. Over the years, Bill's digital archive has been joined by those of other eminent researchers, including evolutionary biologist John Maynard Smith, developmental biologist Anne McLaren and computer scientist Donald Michie.

Digital revolution

The digital revolution is transforming the nature of personal archiving: from curation techniques to the kinds of lives being preserved for posterity — not just the rich and famous, but now everyone participating in the digital age. Most surviving ancient and medieval documents relate to legal, ecclesiastical and formal secular knowledge. Few bear direct witness to the details of everyday lives. Since the sixteenth and seventeenth centuries, paper has helped to open the way for widespread personal writing, printing has facilitated greater literacy and a more robust postal system has promoted distant communication. Even so, unless an archive has gone to a repository or

collector, personal papers have been kept in numbers by only the wealthiest families.

With the emergence of personal computing in the 1970s, more and more people are passing on details of their lives to future generations as digital files. It has been estimated by the International Data Corporation that by 2010

computers and devices that can make exact copies of disks — an approach that is now being adopted by archival institutions worldwide.

The original files residing on a scientist's disks and tapes will eventually become unreadable. It is therefore imperative to make exact copies of these files on fresh media in a demonstrably rigorous way. Merely turning a computer on risks altering files, so devices called write-blockers are used to prevent this from happening.

At the heart of the process is the forensic 'imaging' of an original disk. A single 'image' file, or bitstream, is made of the entire disk. It incorporates a 'map' that allows exact digital replicates of the original files to be recreated from it. Hash values are calculated for every file: short sequences of alpha-numeric characters akin to 'digital fingerprints', unique to each file. If, decades later, the same hash value is obtained for a file using the same hashing algorithm, one can be confident that the file has not been altered.

So as not to rely on software that is subject to obsolescence or is esoteric, we also create (with varying degrees of fidelity to the look and feel of the original) digital facsimiles that, through conversion to suitable file formats such as PDF or XML, are readily usable on modern computers and (let us hope) easily transferred to new computer systems.

However, the digital replicates must be preserved to retain — to the fullest extent possible — the information represented in the original files. This information will be needed should, for example, a future scholar wish to interpret precisely and accurately the original styles, layout and dynamic behaviour of a scientist's files — including home-made computer programs. These replicates can be presented with high-fidelity emulators of ancestral software and hardware.

The British Library houses a small range of classic computers with tape and disk



ROSEMARY WOODS

nearly 70% of the digital universe will be created by individuals rather than organizations. This offers an immense source of fascinating and useful personal data. But it also presents technical, ethical and legal challenges.

In 2003, I was formally appointed as the first curator of what we call eMANUSCRIPTS at the British Library, responsible for these personal digital objects and for the digital manuscripts project, which aims to research and develop the necessary procedures to curate such archives.

How can the authenticity of files be demonstrated years and centuries later? How can we ensure that a file's dates and other embedded metadata are not inadvertently changed, and are properly extracted? These kinds of considerations motivated the use of computer forensic techniques and the establishment of a secure digital scriptorium in the British Library. The facility is equipped with forensically sound

drives, which are used not only in digital capture, but also to help us understand ancestral computer systems and to appraise the fidelity of emulators.

The propensity for digital media to become unreadable, or to be protected by passwords and encryption, has led to other important changes in personal archiving. Curators are now approaching scientists directly, proactively, on their retirement or even earlier. In doing so, opportunities arise for gathering even more information.

Enhanced curation activities include audio interviews that seek a scientist's memories and life story, panoramic photography and three-dimensional graphical imagery of home and study environments, conversational video walks through the local landscape, and recordings of recollections as the person looks through photos and manuscripts. The British Library is producing interactive panoramas for the research and writing environments of eminent scientists and writers: from Fellows of the Royal Society to Poets Laureate.

Personal progress

The archived materials of influential scientists are of immense interest to historians. But life information also has the potential to be of incalculable benefit to scientific and human advancement. Diaries of travellers, letters of diplomats, logbooks of ships' officers and local family archives have long yielded geological, meteorological and sociological data. Records of measurements of magnetic north found in ship logs, for example, have been used to reconstruct the history of Earth's magnetic field.

Some important surveys have actively collected personal information: the UK National Child Development Study has over the years sought the participation of 17,000 individuals born in one week in 1958, and more recently has begun collecting DNA to look for relationships between genetic, medical and lifestyle factors. But this premeditated gathering of structured information is typically expensive or limited. In the digital era, vastly greater quantities of 'freestyle' personal information can be beneficially garnered.

A key challenge is to find a way to give scientists and bona fide researchers access without impinging on privacy. Ideally, some trusted repositories (public, non-governmental or even commercial) would ask users to what extent they would like their information, perhaps anonymized, to be available for

research. While mediating appropriate access, the repositories would help to ensure authenticity. For the most influential and creatively productive individuals, including scientists, the eMANUSCRIPTS themselves might be held for safekeeping; for others, the files — existing, as they would, in enormous quantities — might remain in the care of individuals (as 'archives in the wild', so to speak), with a repository holding only their hash values. If and when an eMANUSCRIPT becomes of interest to a scholar or scientist in the future, the hash value would vouch for its long-standing existence and integrity. Various models for future access can be imagined; the most suitable will ultimately be determined by the nature of emerging technologies and the legal and ethical contexts in which they operate.

The digital lives project — the British Library's first research project under its newly gained status as an Independent Research Organisation, funded by the Arts and Humanities Research

Council — aims to further understand personal digital archives and their research potential in the twenty-first century. Through this project we are seeking to ascertain how files are generated, obtained, organized, used and stored by the public and by academics, and to understand the legal and ethical environment in which they exist.

One of the project's components is to look at the effect of the proliferation of cloud computing. People are relying increasingly on online service providers such as Google, Flickr, Facebook and Carbonite to create their files, to make them available to others or to store them remotely. Of particular relevance to the project are the web locations that are not publicly available but are restricted to the individual or friends and family — sites that typically would not be harvested by web archiving programmes. Is this information being archived at all or simply lost? If kept, who maintains legal ownership of it or could make it available for research? We are only beginning to answer these questions.

Many other important issues are arising and will do so with ever greater frequency. New forms of personal information continue to emerge. Future generations may well have comprehensive video and global-positioning-system recordings of their lives, along with records of neurological and physiological parameters such as heart rate, not to mention personal DNA sequences. Bionic devices that partially restore, enhance or extend an individual's physical or sensory capabilities will

be digitally tuned for each individual, just as digital hearing aids are today. Personal fabricators will allow individuals to create for themselves useful and ornamental physical artefacts. People already interact online through virtual versions of themselves in online gaming environments, and in future this will be advanced with immersive visualization, complemented by touch, taste and scent. What will happen to these digital representations in the long run?

Evolving archives

Perhaps the most fundamental change arising from the digital nature of personal archives lies in their passing from one generation to the next. With a paper letter or diary only one sibling could inherit the original from a parent, but today siblings are receiving identical eMANUSCRIPTs (from texts to videos).

After many generations, there will be highly diverse personal digital archives distributed throughout the population, containing many replicates of identical eMANUSCRIPTs (as well as versions that have been modified, deliberately or inadvertently). It will almost certainly be possible to create phylogenetic networks or trees from these extant personal digital archives, and even to surmise from them the composition of ancestral archives. In the digitally networked era it may be possible for the first time to capture, map and analyse in powerfully quantitative detail the movement and influences of eMANUSCRIPTs with the ideas, observations of nature and accounts of events borne by them.

It will not escape readers with an interest in evolutionary computing that such changing archives, like artificial life, evolving software and computer viruses, may provide another complex system for research — a natural history of evolving digital lives manifested in sets of eMANUSCRIPTs. Bill Hamilton's own archive makes evident his strong interest in evolutionary computing. I believe he would have been intrigued, and pleased, to find that it has been stimulating this line of thinking. ■

Jeremy Leighton John is curator of eMANUSCRIPTS at the British Library, and is principal investigator of the digital lives research project.
e-mail: jeremy.john@bl.uk

FURTHER READING

Gantz, J. F. *The Expanding Digital Universe. A Forecast of Worldwide Information Growth Through 2010* available at www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf (2007).
Garfinkel, S. L. *Int. J. Dig. Crime Forens.* 1, 1–28 (2009).
John, J. L. *Adapting Existing Technologies for Digitally Archiving Personal Lives* available at www.bl.uk/ipres2008/programme.html (2008).
Williams, P. et al. *Digital Lives: Report of Interviews with the Creators of Personal Digital Collections* available at www.ariadne.ac.uk/issue55/williams-et-al (2008).

"Archives in the wild have the potential to be of incalculable benefit to scientific and human advancement."