British Library PhD research placement scheme – project profile:

Virus checking in long-term digital collection management and digital preservation

Reference: 2018-7-PEN

Supervisor	Maureen Pennock, Head of Digital Preservation
Department, Location	Digital Preservation. This placement can be hosted either at the Library's St Pancras site in London or at the Library's Boston Spa site in Yorkshire (or split across both sites)
Start date/timeframe	The placement can take place between May-July 2018 or alternatively anytime between September 2018 and April 2019
Duration	**3 months** (or part-time equivalent – see below) *For a part-time placement, the student would be required to spend a minimum of **3 days/week** on site at the Library.*
Remote-working	If required, the placement could be undertaken working from home for **1 day/week**.

Context for placement

The British Library is responsible for the custodianship of ever-growing collections of digital content. These collections are extremely heterogeneous, and include content such as e-books, e-journals and the UK Web Archive, digitised newspapers, books and manuscripts, digital maps and digital sheet music, electoral registers, patents, personal digital archives, sound and audio-visual content. We also have an extensive collection of legacy digital collection content acquired on hand-held media and dating back to the early 1980s. The Digital Preservation Team, within Collection Management, is responsible for ensuring that these digital collections remain accessible and authentic over the very long term, despite inevitable changes in the underlying technology used to deliver them.

Digital content acquired by the Library undergoes significant processing before delivery to readers, including testing for computer viruses. This is done before content is added to the repository system, both to protect the integrity of the Library's networks and to ensure that viruses cannot be transmitted to readers accessing the Library's digital content. Some of the Library's collections are at particularly high risk of containing viruses and require special treatment. This is particularly true for harvests of the UK web domain collected for the UK Web Archive. Similarly, we know that cover disks from computer magazines (and similar) were often used as a means to spread computer viruses, so the Library's legacy content stored on hand-held media is extremely likely to contain some examples. The Library is currently conducting a programme of work to produce disk images from legacy media (including magazine cover disks), which may be able to shed additional light on the likely incidence of and risks from malicious content stored on floppy disks and CDs.

Computer viruses are currently an under-explored aspect of digital preservation. While there is a widespread assumption that virus checking is an essential part of any practical digital preservation implementation, particularly at ingest, there are a range of interrelated questions that have not been explored in that much detail.

In hosting this PhD research placement, the British Library is keen to gain a greater insight into the types of viruses in our digital collections, a better understanding of the real issues and risks that we might need to deal with as a result, and an increased awareness of how virus checking software deals with legacy viruses and its role in the Library's processing of digital content.

Expected tasks and outcomes

This PhD research placement will explore preconceptions around virus checking in long-term digital collection management and digital preservation. The Library's growing corpus of disk-images from legacy media would be available for research, alongside an emulation rendering environment. The placement student will be expected to:

- Undertake some analysis of our collections to identify the types of viruses that they contain and the risks that they pose

- Review the available literature on the effectiveness of viruses over time and across different computing systems
- Review the available literature on the effectiveness of virus-checking software, particularly as it relates to identification of legacy viruses and "false positives"
- Choose one or more of the following questions (or any others that are relevant) to focus on and explore in more detail, using our collections and preservation/access environments as a test bed:

► **When should virus checking take place?** The current assumption seems to be that all content being preserved needs, as a minimum, to be virus tested prior to submission to long-term storage (sometimes this is combined with an additional "quarantine" step, intended to provide the opportunity for virus checking tools to be updated with the latest risk information). Yet malicious content does not typically present a risk until it is accessed, and this typically happens not within the storage infrastructure but on an access machine. Alternatively, given that virus-checking software updates and improves over time, does consideration need to be given to conducting checks at other stages of content lifecycles? Should virus checking better be seen as one of the checks carried out when content is being accessed? What impact might 'delayed' virus checking have on the overheads associated with collection processing, particularly if only a small percentage of the collection is ever accessed? And what new risks might emerge as a result?

► **How do changes to virus-checking tools interact with risks?** Virus checking tools evolve over time to respond to the ever-changing methods used by those that wish to spread computer viruses and other types of malicious content. It is less clear, however, how the current generation of virus-checking tools would deal with older legacy viruses, e.g. those spread through physical media. It would be interesting to understand more about how virus-checking tools deal with older viruses, and whether current virus checking techniques might be prone to identifying false positives.

► **What are the risks of allowing viruses into a typical emulation environment?** There is a lot of interest now in exploring how emulation might work as a means to provide access to certain types of collections, e.g. the content stored on legacy media. Emulation environments are often now provided via emulation-as-a-service platforms, which generate temporary virtual machines within web browsers. How might different viruses, old and new, work within a virtual machine environment, and how might this change the risks? It would be interesting to consider the risks of running virus-laden content within these virtual machines, and whether malicious content could "escape" into the wider environment or network, particularly over time and across different generations of technology.

As part of their application, students should identify which of the research questions they intend to tackle or, if relevant, propose an alternative. ***Please outline the research questions that you intend to tackle in the section of the application form that asks for "any information not covered above that you feel supports your application for the research placement".*** Please note however that this is not a technical placement.

Outputs from this research would include:

- A literature review report (for internal circulation)
- A separate report on the unique research undertaken during the placement (for internal circulation)
- One or more blog posts on the placement and the research being undertaken. This could make use of a relevant BL blog or other channels suggested by the student.

There would also the possibility of working on a peer-reviewed conference or journal paper on the subject.

Training and experience expected to be gained by student through the placement

The placement will be supervised by the British Library's digital preservation team, which is part of Collection Management. The student will be supervised either in London by Michael Day or in Boston Spa by Maureen Pennock. Both have backgrounds in academic research. There will also be an opportunity to engage with the British Library's digital preservation and web archiving teams during the course of the placement.

Induction will take place at both sites in order to meet all members of the digital preservation team face-to-face. A full introduction to the British Library's *Flashback* legacy conversion project will be provided. The induction will also cover digital preservation main concepts and challenges.

Weekly meetings will be held with the student to monitor progress and to provide direction and feedback, using an informal performance management template. The student will be asked to identify training needs during the first two weeks of the placement. If appropriate, the student may have the opportunity to attend an introductory 'Digital Preservation 101' training course run by an external provider.

Travel expenses incurred travelling between London and Boston Spa for placement-related meetings and other

activities during the course of the placement will be covered by the British Library.

This placement opportunity provides:

- Practical experience working with historic computer viruses and virus-checking software in a secure environment
- Wider knowledge of the British Library's incredibly diverse digital collections
- Training in digital preservation and digital imaging
- Experience in using emulators for access to collection content

Opportunities for professional and personal development include:

- Experience in writing reports for business consumption
- Experience in writing content for circulation on the web
- Experience in engaging with mixed-skill, multi-level staff, both technical and non-technical, and from entry level to senior management.

The placement will also provide a good platform for the student to develop a network of contacts in other institutions interested in the same subject matter.

Required knowledge and skills

This is not a technical placement. However, the student will be expected to be computer literate and comfortable in extending their computing skills to analyse files and script-based output as required.

The placement could appeal to computer literate students in a range of academic disciplines and subject areas, including the History of Science and Technology, Digital Humanities, Museum Studies, Archival or Library Science, Information Management, Computer Science. However, we are open to applications from students from any discipline who have an interest in this area.

This is a training and development opportunity open to current PhD students only. It is not intended to lead to a permanent post at the Library. Please note that the Library is unable to provide a stipend for PhD research placements. Applicants must obtain the support of their PhD supervisor and Graduate Tutor (or someone in an equivalent senior academic management role) in advance and, as part of their process, consult their HEI to ascertain what funding is available to support them.