British Library PhD Placement Scheme 2021/22 – Project Profile

Investigating Digital Preservation Recommendations for File Format Validation Errors

Reference: 2021-DP-FFV

Supervisor (name & role)	Dr Peter May (Digital Preservation Technical Architect)
Department & Site	Digital Preservation Team, London (St Pancras; preferred) or Boston Spa
Start date/Timeframe	Anytime from May 2021 onwards, but to be completed by March 2022 (exact start date to be agreed with the supervisor)
Duration	**3 months** (or part-time equivalent) For a part-time placement, the student would be required to spend a minimum of **3 days/week** on the project.
Suitability for remote working	This placement could be undertaken remotely. The placement student would need a good broadband connection. To be discussed with the supervisor.

Context for placement

Digital content plays an increasingly important role in the British Library, with publishers delivering more content digitally and existing print collections being digitised. Assuring readers of accuracy and authenticity is vitally important. Digital preservation is a key endeavour in this pursuit, providing business and technical guidance to the issues affecting acquisition, long term storage, and access to digital content.

The Library is at the forefront of digital preservation research, with a dedicated digital preservation team who undertake research and engage with the international community – particularly through the Open Preservation Foundation and the Digital Preservation Coalition – to improve our understanding of practical approaches to digital preservation and guide their application.

The Digital Preservation Team's work includes technical research and development, collection-level assessments, risk analysis and policy setting. Specifically, part of their role involves assessing the validity of digital content with respect to format specifications, understanding validation failures, and deriving technical guidance and policies on mitigating such failures. They use specialist format validation software tools (e.g. JHOVE) to check for compliance issues and inform production implementations in the Library's long-term repository. These tools can identify a wide range of validation errors, though the varying impact of these errors on the long term renderability of digital content is not always clear.

The placement student will work with the Digital Preservation Team on a deep-dive investigation of format validation errors. The project will help the Library and the broader community to develop a better understanding of real-world format validation failures, and influence the preservation of digital collections in the Library's new repository system.

Expected tasks and outcomes

The student will undertake research into file format validation errors, using digital collection sample files and validation tools (i.e. JHOVE), to identify and document (on the Open Preservation Foundation's JHOVE wiki page) the following:

- Root cause(s) of validation problems;
- Impact those problems have on the long term preservation and rendering of those files;
- Remediation actions that may be undertaken.

From this research the student will develop initial format validation "profiles" detailing significant errors that may need to be remedied for specific formats, whilst also identifying errors which are likely to have less impact. To do this the student will need to:

- Develop detailed technical knowledge about file formats, e.g. through reading specifications;
- Develop an understanding about preservation risks and their application;
- Understand the validation tool code base (to understand implemented validation tests);
- Collate sample files, or create exemplar files, that illustrate validation problems;
- Run validation tools over files (possibly in debugging mode), gather outputs, and analyse results.

The student will be expected to collaborate and share findings with the preservation community, particularly through blog posts and, potentially, webinars.

A final report/paper documenting the work undertaken, key results, lessons learned, challenges, and future direction would summarise the project and provide future benefit to the Library and broader community who may undertake similar work.

Training and development opportunities

All placement students are welcome to access the Library's offer of workshops, talks and training. For further details please refer to the Application Guidelines on the British Library website.

In addition, this placement project will offer the following:

- Core Digital Preservation training, delivered either through our internally developed course, through 1-2-1 sessions with team experts, or through an external online course, depending on placement/course timings. Additional externally-provided digital preservation webinars may also be available (e.g. through Open Preservation Foundation or Digital Preservation Coalition).
- Specific training and guidance around file formats and validation tools will be provided throughout the placement by appropriate team members, and potentially through the Open Preservation Foundation (timing dependent). The exact nature will be dependent on the student's needs, but could cover topics such as: running/debugging validation tools, command line tool invocation, source code version control (Git), and Github wiki editing.
- Sharing of knowledge and results across the digital preservation community is important, and there will be opportunity to develop presentation skills through blog writing and giving webinars.
- Subject to availability at the time of the placement, internal Library courses, which may cover topics such as digitisation, metadata, content mining and cleaning data.

Required knowledge and skills

Please ensure that you are aware of the general expectations for all applicants, as detailed in the Application Guidelines on the British Library website.

For this specific project, the following additional criteria apply:

- Computer programming experience, with a deep awareness of how computers operate, digital file storage and computation;
- Able to read and understand Java code (Python advantageous, but Java essential);
- Able to present and explain technical terms to a lay audience;
- Able to quickly read, understand and apply learnings from detailed technical documents;
- Able to focus and be detail oriented, but sufficiently grounded to see the bigger picture and explain findings in an accessible way;
- Knowledge of or interest in digital preservation.
- A detailed knowledge of format specifications is advantageous, but not essential.

Application deadline: 5pm on Friday **18 December 2020**

Further information on eligibility, conditions and how to apply is available on the British Library website: https://www.bl.uk/news/2020/october/phd-placement-adverts-2020