Monitoring Bodleian Libraries’ Repositories with Micro Services

|  |  |  |
| --- | --- | --- |
| **Edith Halvarsson** | **James Mooney** | **Sebastian Lange** |
| *Bodleian Libraries, University of Oxford*  *United Kingdom*  *edith.halvarsson@bodleian.ox.ac.uk* | *Bodleian Libraries, University of Oxford*  *United Kingdom*  *james.mooney@bodleian.ox.ac.uk* | *Bodleian Libraries, University of Oxford*  *United Kingdom*  *sebastian.lange@bodleian.ox.ac.uk* |
|  |  |  |

**The Digital Preservation Micro Services and Reporting (DPMS) service is Bodleian Libraries’ monitoring system for digital collections. DPMS interacts with the storage layer of the Libraries’ existing repository systems. This method is an alternative to the monolithic systems model for digital preservation [1]. The decoupled nature of DPMS has meant that preservation tools could be incorporated into the Libraries’ existing repository services, without needing to migrate assets to a separate digital preservation platform.**

**This paper outlines how the DPMS service was developed. It describes the components of DPMS’s technical framework and the process of implementing it in Bodleian Libraries’ digital repositories**

**Keywords – Micro Services, Open Source, Reporting**

**Conference Topics – Innovation**

# Background

Bodleian Libraries is a group of 28 libraries that serve the University of Oxford (England). It is the largest academic library service in the United Kingdom and one of the largest library services in Europe [2]. In addition to its sizable physical collections, the Libraries have collected born-digital content for over 15 years. During this time the Libraries have developed specialist repositories for digitized content, born-digital archives, research data and research publications. The Libraries’ digital collections are now primarily managed in three core services: 1) Digital.Bodleian, 2) the Oxford University Research Archive (ORA/ORA-Data), and 3) Bodleian Electronic Archives and Manuscript (BEAM).

In 2017 Bodleian Libraries and Cambridge University Library (CUL) undertook a joint market review of digital preservation systems to assess their suitability for the Libraries’ existing repositories. The review was completed as part of a research project called the Digital Preservation at Oxford and Cambridge (DPOC) project [3]. At the time of the review, none of the assessed systems met all the different repositories’ essential requirements. The option of migrating all digital collections to one joint monolithic digital preservation system was therefore deemed unsuitable, as it would have resulted in the repositories being unable to undertake some of their core administrative activities. A micro services architecture provided the Libraries with an alternative approach for monitoring and reporting on digital collections, where digital files could still be retained in their respective heterogenous repositories.

In collaboration with Dave Thompson, Digital Curator at the Wellcome Library, a digital preservation solution which could work with existing digital repositories was scoped. This resulted in a paper for Library Hi Tech in 2018. The paper introduced a design proposal for a new architecture to work with Big Data volumes of preserved digital resources [6]. This paper came to form the basis for a business case to realise a micro services based digital preservation approach at Bodleian Libraires. In 2018 funding was secured for an initial proof of concept from the University of Oxford’s IT Capital fund. The funding enabled the creation of a micro services reporting platform which has been in usage in the Libraries for the past two years, with new micro services being added to the platform on a regular basis.

# Micro services architecture overview

The novel features of micro services architecture are a focus on small size and interdependency. Micro services are typically created to address a single function/capability and operate independently of each other, only communicating with other micro services via their published interfaces [5]. While micro services can introduce additional administrative overhead compared to a monolithic system, they can also bring great benefits. Among these are scalability (each micro service can be scaled individually to meet service needs) and portability. Micro services can be deployed across different heterogenous platforms [5]. As illustrated by the micro services approach proposed by California Digital Library in 2010, micro services can bring great benefits to diverse digital curation infrastructures as they “can be deployed in the context in which it makes most sense, both technically and administratively” [1]. In Bodleian Libraries, the usage of micro services has for example enabled the Libraries to scale individual micro services to meet collection growth needs over the past two years.

# Micro services uptake

While micro services have been discussed in the digital preservation field for over 10 years, it is not yet a commonly adopted institutional approach. Based on the experience of the Libraries, it is possible that the more modular costing model involved with a digital preservation micro services approach (potentially covering multiple individual support contracts and inhouse staff costs) is prohibitive for organizations needing to cover all digital preservation activities within a single support contract. This could however shift in the future if a micro services reporting framework was provided as a commercial solution.

# The DPMS project

The Digital Preservation Micro Services and Reporting (DPMS) project begun in 2018 and is scheduled to complete in early 2023. During the DPMS proof of concept (2018-2019), the service’s technical framework was developed. The platform, based on Elastic Stack (ELK), is described further in the technical overview section below. The technical framework from the proof of concept formed the basis for all micro services which were subsequently added to DPMS.

Development of the DPMS service is running in sequential stages. Each new stage looks at a particular category of digital preservation tools/micro services. These categories are as follows:

1. File integrity
2. Virus checking
3. Backup and restore analysis
4. Characterisation
5. Validation
6. Digital preservation copy monitoring

Once a stage is completed, the new micro services are deployed in production so that they can be actively used by the Libraries core repositories. So far, micro services relating to categories 1-4 have been developed. Micro services relating to the final categories (5-6) will be developed in 2022.

# The DPMS framework components

The DPMS framework is built on open-source tools. The Libraries’ preference is for using supported open-source projects, rather than writing custom tools [4]. DPMS has incorporated the outputs from several open-source tools and utilities (such as Siegfried, MediaConch, ExifTool, Zabbix, rsync and others). These tools and utilities make up the individual micro services offering. Each was assessed during the project to ensure that it was well supported and (where relevant) could provide metadata as JSON output.

![Graphical user interface

Description automatically generated]()

Figure 1: Example of a Grafana dashboard

The DPMS platform itself is built on Elastic Stack (ELK). It was chosen as the preferred framework, as an instance of ELK was already actively maintained by the University of Oxford’s IT Services on behalf of the wider University. Using the existing stack took some of the overhead of running the service away from the Libraries. The ELK stack comprises of Elasticsearch (a search engine), Kibana (a data visualisation tool), Beats (for centralizing log data) and Logstash (a processing pipeline). In addition to the Kibana software, Grafana (another open-source visualization tool) is also used for creating graphs which provide DPMS users with a more high-level overview of their data.

# The DPMS workflow

DPMS interacts with the Libraries’ existing repository systems via their back-end storage. Files held in the repositories’ storage areas are scanned by the micro services. Each repository can mix and choose which micro services are most relevant to their collection profile. Metadata output from the micro services scans are aggregated into JSON log files via a log merging tool.

A copy of the JSON log file is sent to preservation storage and another copy is sent to Oxford University IT Services for indexing in ELK. Elasticsearch provides the search engine for interrogating metadata gathered by the micro services. This metadata can then be visualised in Kibana and Grafana for the end user.

Chart

Description automatically generated

Figure 2: Example of a Kibana visualization (breakdown of file formats)

From the repository user’s perspective, DPMS consists of dashboards which aggregates statistics about digital files and the storage they are held on. Grafana provides the main high-level overview of the repositories’ digital collections and is generally how users choose to first access their statistics. The Grafana dashboards then link out to Kibana for more detailed statistics about the collections. Users have the option to create more advanced queries and searches in Kibana, or to drill down into the detail about particular files. Where issues have been highlighted by DPMS (such as a virus or deleted files), these are highlighted in the high-level Grafana dashboards for further investigation. A ticket for the issue is also raised in GitLab, which is the Libraries’ central location for technical documentation.

Micro services scans are repeated every 1-3 months, with fixity monitoring always undertaken on a monthly basis. DPMS is currently tracking approximately 58 million files (300TB of storage) across the Libraries’ repositories. As ELK retains historic logs, DPMS can also illustrate changes to digital collections over time.

# Supporting the DPMS service

The DPMS service is run by the Libraries’ digital preservation team, with support from the University of Oxford’s ELK service. The digital preservation team are responsible for scheduling scans, aggregating JSON logs, onboarding new collections, providing training, and assisting with user queries. In total, the Oxford ELK team and the Libraries’ digital preservation team dedicate 0.8 FTE to running the DPMS service.

However, the Libraries’ repository staff are by necessity also actively involved in the preservation of their digital collections. Repository service owners are responsible for reviewing and (where possible) investigating preservation alerts from DPMS. As experts on their content and workflows service owners are often best placed to interpret ‘unusual’ activity in their repository, such as large-scale deletion of files or unexpected colour profiles in images. The time spent by service owners on investigating preservation alerts vary from each repository (with more active repositories requiring more staff engagement) and on how long the repository has been scanned by DPMS. As service owners become more familiar with the DPMS reports and findings each month, the time spent on investigating issues generally decreases.

# Onboarding new repositories to the DPMS service

Four of the Libraries’ repositories currently use the DPMS service. When onboarding new repositories, the digital preservation team follows a standardised onboarding workflow. It can take several scans to fine-tune the Grafana and Kibana dashboards, to ensure that they meet the needs of the new repository. The onboarding workflow consists of the following steps:

* The repository owners provide initial set-up information for configuring the DPMS dashboards. Including:
  + Overview of the repository directory structure
  + How/when digital collections are updated
  + User credentials
* The repository owners are provided with training on using Grafana, Kibana, and GitLab
* DPMS completes an initial scan of the files in the repository
* The digital preservation team reviews the scanning results together with the repository service owner
* DPMS scanning patterns are updated to exclude certain files and directories if appropriate
* The digital preservation team creates a generic reporting dashboard for feedback
* The repository service owner provides feedback, and the dashboard is updated accordingly
* A configured JSON metadata log is sent to ELK for indexing

Once a repository has been onboarded it is added to DPMS’s regular monthly scans.

# Next steps

Bodleian Digital Libraries Systems and Services department will take on the full system support costs following the completion of the initial DPMS project, covering software support contracts and inhouse staff (the Libraries’ digital preservation team). The creation of the DPMS service is only an initial step towards enhancing the Libraries’ understanding of its digital collections. Going forward, the Libraries will look at using the DPMS generated metadata to address areas of preservation concern. Such activities will include decreasing the percentage of unknown file types in the collections (by submitting new file signatures to PRONOM) and assessing gaps in available viewers and file conversion tools.

# Conclusion and summary

DPMS is an example of an alternative model for digital preservation monitoring using a micro services approach. This paper has illustrated how Bodleian Libraries could utilise existing digital preservation tools and open-source frameworks to improve its monitoring capacity. The DPMS service model may also have applicability to other organizations who do not want to migrate from their current repository systems, or who have chosen to not implement a DAMS like system. As illustrated by the DPMS service, the knowledge which can be gained from implementing only a few micro services can greatly improve an organization’s understanding of its collections.

# Acknowledgements

Many thanks to Kristian Kocher (University of Oxford) for his continued guidance and expertise on ELK. Thank you also to Bram Hauer and Matt Thorpe (Bodleian Digital Library Systems and Services) for their work supporting the DPMS project. Last but not least, we would like to acknowledge the contributions of the late Dave Thompson and thank him for his support and forward-thinking spirit.

# References

1. Abrams, S., Kunze, J., Loy, D. (2010). ‘An Emergent Micro-Services Approach to Digital Curation Infrastructure’, *The International Journal of Digital Curation*, vol. 5, no.1, pp.172-186
2. Bodleian Libraries (2022). ‘About the Libraries’. Available at: https://www.bodleian.ox.ac.uk/about/libraries (Accessed: 26 January 2021)
3. Bodleian Libraries (2020). ‘Holding Page: Digital Preservation at Oxford and Cambridge’. Available at: <http://www.dpoc.ac.uk/> [Accessed: 26 January 2021]
4. Bodleian Libraries (2018). Digital Preservation Policy. External policy. Available at: <https://web.archive.org/web/20211121141920/https://www.bodleian.ox.ac.uk/sites/default/files/bodreader/documents/media/digital-preservation-policy.pdf> (Accessed: 26 January 2021)
5. Dragoni N., Lanese I., Larsen S., Mazzara M., Mustafin R., et al. ‘Microservices: How To Make Your Application Scale’. *A.P. Ershov Informatics Conference* (the PSI Conference Series, 11th edition), Jun 2017, Moscow, Russia. ffhal-01636132f
6. Gerrard D.M., Mooney J.E, Thompson D. (2018). 'Digital Preservation at Big Data Scales: Proposing a Step-change in Preservation System Architectures', Library Hi Tech, vol. 36, no. 3, pp.524-538, <https://doi.org/10.1108/LHT-06-2017-0122>