# DigiSpec

## Scoping Future Born-Digital Data Services for the Arts and Humanities

# CASE REPORTS

Edited by:
David De Roure
John Moore
Kevin Page
Toby Burrows

Oxford e-Research Centre, University of Oxford
The National Archives

2022

The "DigiSpec: Scoping future born-digital data services for the arts and humanities" project was funded by the Arts and Humanities Research Council (AHRC) under research grant number AH/W007592/1.

**DigiSpec: Scoping Future Born-Digital Data Services for the Arts and Humanities**

# CASE REPORTS

## Table of Contents

**DigiSpec: Scoping Future Born-Digital Data Services for the Arts and Humanities**

# Case Reports

## Introduction

The DigiSpec Project was funded by the UK Arts and Humanities Research Council in 2022 under its "Scoping Future Data Services for the Arts and Humanities" programme.

As part of the evidence base for its report, the project collected twenty short Case Reports from experts in the field of born-digital data services. Each of these describes the current and future activities and requirements of a significant UK project or service.

These Case Reports are published in this collection. The DigiSpec Project also commissioned three longer Case Studies, which are being published separately.



## DigiSpec Project Team

Oxford e-Research Centre, University of Oxford:
Professor David De Roure, Dr Kevin Page, Dr Toby Burrows, John Pybus

The National Archives:
John Moore, Mark Bell, Ashleigh Hawkins, Santilata Venkata, Nicola Welch

July 2022

# Living with Machines

**Title of project: Living with Machines**

Revision: 1.0; date: 2022-03-25; licence: CC BY 4.0

Authors: David Beavan (Acting PI [with Emma Griffin], Acting Principal Research Software Engineer, The Alan Turing Institute) <dbeavan@turing.ac.uk>; Timothy Hobson (Co-I, Senior Research Software Engineer, The Alan Turing Institute) <thobson@turing.ac.uk>

## Description of resource or infrastructure

Living with Machines (LwM)[1] is a ground-breaking five-year £9m+ research partnership between The Alan Turing Institute, the British Library, and the Universities of Cambridge, East Anglia, Exeter, London (QMUL) and King's College London, in which historians, data scientists, geographers, computational linguists, and curators have been brought together to examine the human impact of industrial revolution. Funded by the UK Research and Innovation (UKRI) Strategic Priority Fund, it is a multidisciplinary collaboration delivered by the Arts and Humanities Research Council (AHRC) (project reference: AH/S01179X/1). Focussing on the long nineteenth century (c.1780-1918), the Living with Machines project aims to harness, on a hitherto unprecedented scale, the combined power of massive digitised historical collections and computational analytical tools to examine the ways in which technology altered the very fabric of human existence.

Much of the data the project uses is not born-digital, but is digital at the point of entry to the project. Our 63 datasets are extremely varied, featuring textual, tabular, geospatial and image content, including Ordnance Survey maps (National Library of Scotland, British Library), historical census records (CamPop, UK Data Service), historical newspaper content (British Newspaper Archive, Findmypast, and targeted digitisation), natural language processing (NLP) language models, historical gazetteers and newspaper press directories. LwM has a dedicated Rights Assurance Manager and Data Wrangling work strand, in addition to Turing Research Engineering and BL staff of Research Software Engineers and Research Data Scientists who work in a 'radical collaboration'[2] with their research and professional colleagues in a number of work strands with overlapping staffing. Our digital infrastructure is non-intrusive and designed to facilitate new questions in a nimble manner. It includes staff skills and knowledge exchange on a number of levels from reading groups to formal training in software tools and methods. Centralised infrastructure has focussed on data storage (100s of TB), indexing of datasets for quick exploitation, linking data and data exploration through literate programming (e.g. Jupyter Notebooks). We are now beginning to see the scale of this project shine, as we are in the phase of maximally exploiting the knowledge we can get from our datasets. The breakthroughs are being made where we link and move between modalities: from maps showing change in the built environment, to the census of those who lived there, to the newspapers that represented different social groups, to the attitudes in them.

As we are two thirds through the funded phase, our attention is turning to fully exploiting the potential of the data we have ingested, with two main threads: increased linking across our datasets, using richer contextual information to assert improved confidence in these links, such as the Meta Gazetteer work; and scaling-up our processing, by using the power of cloud compute to take deeper dives into our datasets for a richer, more nuanced analysis. We will invest heavily in opening up our code and data as well described, documented and tested reproducible outputs to further build communities and broaden the reach of the project and secure our (digital) legacy. Capturing the impact of these outputs is important for us, our institutions and our funder, and we hope to shape what a good case study looks like in this environment, embedding the knowledge and experience of this project with the wider A&H community, and aiming to share experience with Towards a National Collection and the UKRI Digital Research Infrastructure scene and beyond.

[1] https://livingwithmachines.ac.uk

[2] https://livingwithmachines.ac.uk/how-we-collaborate

**Current state**

- Hardware
    - Microsoft Azure Cloud infrastructure
- Software
    - Azure Services:
        - Blob storage
        - Linux virtual machines
        - HDInsight managed Spark cluster
        - Azure managed PostgreSQL database
        - Virtual network & access rights services
        - Azure Cloud Shell
    - Processing and analysis
        - Python libraries for ML, NLP, image processing, geospatial data analysis
        - Pytorch & TensorFlow
        - Apache Spark
        - ElasticSearch
        - Django framework
        - XSLT
    - User interface
        - Jupyter notebooks, JupyterLab, JupyterHub, BinderHub
        - Zooniverse crowdsourcing platform
        - D3 Javascript visualisation
        - Kepler geospatial analysis tool
    - Project, code and content management
        - GitHub
        - Google Docs
        - HackMD
- Data and content
    - Historical newspaper collections: <58 TB of text in XML format (27 TB) and selected images in JP2 format (31 TB)
        - Four collections: British Newspaper Archive (46 TB) on loan from Findmypast, JISC 1 & 2 (2.5 TB), Heritage Made Digital (0.5 TB), LwM project digitised newspapers (6.5 TB)
    - Mitchell's Newspaper Press Directories (95 GB) XML, plain text and TIFF images
    - UK census archive for years 1851, 1861, 1871, 1891, 1901, 1911 in CSV format (<10 GB)
    - Historical Ordnance Survey maps
        - NLS collection (4.5 TB) GeoTIFF, Shapefiles, CSV
        - LwM project digitised maps (5.5 TB)
    - Geospatial datasets (<10 GB):
        - Map boundary metadata in GeoJSON format
        - OS Open Roads vector line data in Shapefile format
        - Quick's rail station directory
        - Lewis's Topographical dictionary of England
        - GB1900 Historical Gazetteer
        - Wikidata mirror
    - Acquired data is static and data volumes fixed
    - Rate of growth of processed/derived datasets is small (<10%) relative to the size of the static (source) data
    - Derived data includes:
        - Newspaper metadata in PostgreSQL relational database
        - Annotated images (maps & newspapers) and text
        - Artefacts generated by/for Machine Learning models & pipelines

        - Mapping tables for data linkage
- Audiences, users, disciplines, subjects
    - The principal users are the project team of Digital Humanities researchers and data scientists
    - Disciplines represented in the team include historians, computational linguists and NLP specialists, computational modellers and statistical analysts, curators and content managers
    - Wider audiences include external researchers, exhibition visitors, crowdsourcing participants
    - Source and derived datasets produced by the project are, where possible, made publicly available via the British Library and are of potential interest across a wide range of disciplines and levels of scholarship
- Access methods
    - Discovery methods:
        - Relational database queries
        - Defoe[3], a bespoke tool for querying historical textual datasets at scale based on Apache Spark, enables queries targeting both source metadata and full text (e.g. keyword/ngram search)
        - Dataset-specific access tools built into Jupyter notebooks (e.g. DeezyMatch[4])
        - Subsampling tooling for creating representative samples, having prescribed distributions along arbitrary metadata dimensions
    - Analysis typically involves attaching data, residing in cloud storage, to a virtual machine, or cluster, running on the same cloud platform
- Place in research lifecycle
    - Data collection has been completed
    - Processing is an ongoing task
    - Analysis began shortly after the project commenced, running in parallel with data collection and processing on the most immediately available data. Large-scale analysis on the main newspaper dataset is now beginning to be feasible
    - Sharing of code and data has started but is lagging the analysis, and will be a focus for the final 18 months of the project
- Staffing: number and type
    - RSE/RDS: 6
    - Academics: 5
    - Professional staff: 6
    - Researchers: 5
    - Leadership (PIs and Co-Is) are a mix of the above

**Issues and problems**

Our challenges revolve around bias and representativeness of our datasets; when and where can they be trusted, particularly when linked; quality (e.g. optical character recognition (OCR)), and ambiguity (e.g. toponymical disambiguation). The diachronic nature of datasets and semantic changes in the language used can be challenging with software not built for historical content. We wish to build better computational models by using crowdsourcing and human-in-the-loop methods. There are often no established persistent identifiers (PIDs), or community of use around some of our resources, as they are not openly available to researchers and are sufficiently commercially sensitive they must be stored in a Trusted Research Environment (TRE). That brings challenges of legacy, and how we share derivative datasets or enhancements to datasets that are not open, and the impact this has on the reproducibility of our research and its engagement. The experience of the project and both Arts and Humanities (A&H) and Data Science communities is that there is a shortage of research software engineer and research

[3] https://github.com/alan-turing-institute/defoe
[4] https://pypi.org/project/DeezyMatch/

data science skills, particularly in the humanities. This project has, through the DH RSE Summer School and other initiatives, made some progress to addressing this pipeline of talent, but it will require a much larger investment to make the impact that is required for the ongoing success of digital projects.

**Future requirements (taking a view of follow-on projects too)**

- Hardware
    - Scalable (cloud) compute for intensive HPC or GPU tasks that can operate on sensitive data (e.g. the commercially sensitive British Newspaper Archive)
    - A focus on software-defined infrastructure and communities of practice, so that compute and storage research infrastructure can be maximally reused
- Storage
    - Centrally negotiated intellectual property rights for access to common or large datasets
    - Shared (cloud) storage that contains these datasets, if the data provider cannot provide a query endpoint
- Software
    - Supported (staff and compute) tools and environments to run workflows and pipelines of connected processes and algorithms
    - Framework(s) for linking inhomogeneous datasets (e.g. tabular, semi-structured, plain text, images) by shared features such as temporal and spatial attributes in a way that is systematic, configurable and reproducible could greatly accelerate the research workflow and promote cumulative enrichment of data
    - Improved open source tools for annotating datasets providing support for a range of content types and research domains. Current tools are clunky and do not provide fully interoperable annotations suitable for data linkage applications
- Access methods
    - TREs that have policies tailored to historical and/or commercially sensitive content rather than the focus on personal data, to minimise excessive barriers to research
    - More open data from the cultural heritage sector, especially when funded by public funds
    - Modernised licence conditions for data providers (e.g. UKDS) that are tailored to cloud compute environments
    - More endpoints and opportunities for running analysis next to the data at the data provider (e.g. non-consumptive research/analytics)
- Skills
    - We desperately need skilled Research Technical Professionals (RTPs), in particular Research Software Engineers and Research Data Scientists with A&H backgrounds
    - A&H researchers experienced in digital methods and technologies
    - Cross-disciplinary centres that lower barriers to radical collaborations
- Preservation
    - Preparation for the preservation of the research-ready, data-wrangled large datasets takes months, time that would be duplicated by each project using the dataset
    - Standards and guidelines for data archiving and preservation that facilitate future data science endeavours would be helpful both at the input phase, improving accessibility and reducing data wrangling effort, and at the output phase when preparing derived datasets for dissemination

**Case Reports**

**NLS Chapbooks**

Authors: Abhishek Dutta, Giles Bergel, Andrew Zisserman (University of Oxford)

Contact: Giles Bergel (giles.bergel@eng.ox.ac.uk)

**Description of resource**

- Its nature and purpose

This resource presents visually searchable images of the illustrated pages of the National Library of Scotland's Chapbooks Printed in Scotland dataset. It is an outcome of a collaboration between the NLS and the University of Oxford's Visual Geometry Group, arising from the lead author's term as the National Librarian of Scotland's Fellow in Digital Scholarship in 2020-1. The resource is designed to showcase the application of computer vision techniques to open, curated data, specifically for research in the history of cheap printed 'chapbooks' printed in Scotland during the seventeenth, eighteenth and nineteenth centuries. Many chapbooks have been lost and little is known about their origins, but it is clear from the number of editions printed in various locations that a substantial trade existed to supply a large and broad readership, pointing to growing levels of literacy and to some extent indicating popular tastes. By combining visual search technology with library metadata, the resource makes it possible to browse all the illustrated pages within the dataset by standardised metadata; perform visual searches on any one illustration to find matching illustrations; use the search results to bridge incomplete metadata and then infer the likely origins of a publication (its date or place of printing, or its printer); and to visualise clusters of matching illustrations across the dataset.

While a standalone resource, the demo can be accessed through a project page entitled Visual Analysis of Chapbooks Printed in Scotland, hosted by VGG, which provides context and background. At the time of writing the page included a short narrative about the project; links to Git repositories hosting project software and data; a link to a peer-reviewed article on the project; links to the NLS Data Foundry repository and to an information page about the original collections; links to several workshop presentations (slides) and to a public lecture (YouTube video); and credits. Such project pages, which present a stack of assets such as a paper, demo, data and code, are increasingly common in computer science and the digital humanities.

- Technological approach

The resource employs a variety of computer vision and data science techniques, described in detail in this article - https://doi.org/10.1145/3476887.3476893. The first task was to download image files (jpgs) and metadata (METS XML) from the NLS Data Foundry, along with further metadata (in JSON) giving standardised names of people, places and dates that was supplied on application to the library. A bespoke image annotator tool, VGG List Annotator (LISA) was created for the project, allowing the display of lists of chapbook pages (numbers) in a web browser. Annotations consisted of simple rectangular bounding boxes, the coordinates of which were stored offline in a JSON format. The resulting image regions were used as ground truth for a repurposed object detector, EfficientDet, which is a general-purpose neural network-based object detector trained on the Common Objects in Context (CoCo) training and evaluation dataset. After several rounds of training, inference and verification, over 3600 pages containing regions of interest (printed illustrations) were obtained, including the coordinates of those regions.

Images and metadata are presented in an instance of VGG Visual Search Engine (VISE) hosted by VGG. Images can be browsed by way of the NLS metadata, which is stored in an SQL database, and searched using VISE's visual search interface. Search is based on SIFT features within a 'Bag of Visual Words' model, which has proven effective for searching printed book illustrations along other instances of objects. Searches are made using a pre-computed index, which is amenable both to internal and external (uploaded) queries, presenting ranked results. A further step was to compare each indexed image with every other image in order to obtain clusters of matched illustrations, which can be browsed by population size (from 2 to 22 instances).

- Future direction and goals

While a demonstrator, the resource is already a functioning research, teaching and cataloguing tool for studying the history of the Scottish printed chapbook and for teaching computer vision in higher education and library and cultural heritage settings. Future goals include adding more images from the NLS and other institutions that have digitised their chapbooks; porting the resource to various cloud services to improve scalability, storage, sustainability and integration with other tools (such as image classifiers); and to provide the ability to directly edit and enhance the stored metadata. An ingest stage that included the ability to obtain images and metadata from IIIF and other image repositories is a longer-term goal, as is integrating the illustration detector into a single pipeline, the goal being to make the image-processing services more widely available for teaching and research.

**Current state**

- Hardware
  The resource uses a server shared with a number of other VGG demos
- Software
  VGG Image Search Engine (VISE) with an Nginx server proxying external requests to the VISE server
- Storage
  Project data – 240MB.
  Software and dependencies – 40MB (VISE only)
- Data and content
  Image (jpeg) files – 3600 images; 200MB
  Metadata – SQL tables; 40MB
- Audiences, users, disciplines, subjects
  Book historians, especially of popular printed Scottish literature, balladry, folklore and poetry.
  Digital humanists and digital scholarship professionals, especially based in the library and GLAM sectors
- Access methods
  Open-access web publication
- Place in research lifecycle
  The resource has been through the research lifecycle, but offline analysis continues on the data, the results of which will feed back into it.
- Preservation
  Local (Oxford University) disk-based storage based on IBM Spectrum Protect software, mirrored to a remote location.
- Staffing
  Technical support is provided by RSEs (Dutta); user support by the project PI (Bergel).

**Issues and problems**

Sustainability is not currently scoped beyond the length of the Visual AI EPSRC programme grant (until 2025). The VISE index that facilitates visual search cannot at present be incrementally updated, necessitating batching images for updates.

**Future requirements**

If the resource was to grow to integrate images of all extant Scottish and related chapbooks, the resource might require up to 10GB of storage space, and a doubling of server RAM to 16GB. There is also a requirement for an infrastructure capable of sustainably hosting and preserving complex, full-stack scholarly outputs such as the Visual Analysis web page and its linked resources, which includes but is not limited to NLS Chapbooks itself. Such an infrastructure has yet to be fully scoped: collaboration between researchers, digital preservation professionals, reproducibility advocates and suppliers of research data management systems is necessary to understand the complexity and usability of these hybrid scholarly outputs.

**References**

Visual Analysis of Chapbooks Printed in Scotland:
https://www.robots.ox.ac.uk/~vgg/research/chapbooks/
Visual AI project:
https://www.robots.ox.ac.uk/~vgg/projects/visualai/
The National Library of Scotland's Fellowship in Digital Scholarship 2020-1:
https://data.nls.uk/projects/the-national-librarians-research-fellowship-in-digital-scholarship/

**Oxford University Museums Digital Collections Service**

***Dr Aruna Bhaugeerutty***

Oxford University's Gardens, Libraries and Museums (GLAM) contain some of the world's most significant collections. This division within the University comprises the Ashmolean Museum of Art and Archaeology, History of Science Museum, Museum of Natural History, Pitt Rivers Museum, Bodleian Libraries, and Botanic Garden & Harcourt Arboretum. These departments are integral to the delivery of the University's strategic aims of teaching, research and widening participation. They also embody the public face of the University, representing the front door to the wealth of knowledge and research curated by and generated at Oxford. Collectively they hold over 20 million collection items and welcome approximately 3.2 million visitors each year.

**Description**

The University has recently invested in the development of an IT infrastructure and digital service(s) to manage digital collections across the GLAM museums. Key systems within this service include Collections Management Systems (CMS), Digital Asset Management Systems (DAMS), Collections Online and middleware applications that support data preservation and validation. The service infrastructure allows the museums to effectively manage their digital data and surrogates of their collections, thereby supporting the management and care of the physical collections as well as improving access to and engagement with the collections both for staff and the diverse audiences of the GLAM museums. Collectively the systems are designed to enhance productivity, improve the quality and accuracy of collections data, and provide a framework for the ongoing management, development and provision of digital materials.

The GLAM service predominantly comprises cloud-based applications hosted by external suppliers containing collections metadata and multimedia (e.g. zetcom MuseumPlus CMS, Axiell EMu CMS and Montala Resourcespace DAMS). The CMS stores data in XML format and the DAMS stores a variety of file formats although the repository currently largely comprises JPG, TIFF, MP3, MP4, PDF, and OBJ files. Groovy (Java-based) scripts generate CMS data exports to templates written in Word, Excel, Freemaker and PHP scripting provides additional DAMS configuration. A series of modular integration scripts via plugins pull and push data between CMS and DAMS systems via their APIs. Local python-based middleware for data validation parses systems content using their APIs to assess data quality and status for online publication. Each system includes plugins for MS Azure AD authentication connecting to University Nexus365 user profiles. Cloud-based backups are provided via supplier SaaS agreements and additional 3rd party suppliers (e.g. Bytes, Cirrus HQ) manage second-line backups of system data and configuration in AWS for DAMS and VEEAM for CMS. The Collections Online application uses ElasticSearch and S3 hosted on AWS and uses APIs and scripting to draw and present content from the CMS and DAMS on a localised Drupal-based platform (Mosaic).

GLAM wish to enhance the Museums Digital Collections Service through the development of preservation, digitisation, search and discovery, and additional online platform offerings.

For example, integration with preservation micro-services that monitor digital collections and alert owners where issues that may impact long-term accessibility are detected will improve understanding of the nature and health of as well as risks to museum digital collection repositories through the application of file integrity, virus scanning, backup and restore analysis, characterisation and validation checks. Continued digitisation of the museum's collections is required in order to complete and enrich the content currently managed and delivered by the service's systems. This will provide the content infrastructure required for collections search and discovery to enable both academic research and public engagement. A digitisation service infrastructure requires ongoing investment in both physical and digital estate as well as complex, specialist resourcing and workflows. Technical infrastructure development is also required to enable implementation of interoperability initiatives such as IIIF, Linked Art, Wikidata and authority/ontology creation that will greatly facilitate sharing and aggregation of museum digital collections data across GLAM and University networks and beyond for research discovery. Integration between Collections Online and additional online platforms/technologies that can package and present digital collections content aligned to specific teaching, learning, research, engagement needs must also be considered following a Research Data Management model to ensure future sustainability and development of the service.

**Current state**

- Hardware
    - CMS – virtual application server and database (4GB RAM, 1 GigE, RAID 1, HTTPS port 443; Linux OS, PostgreSQL, Payara server, Java Oracle JDK 8) externally hosted and managed
    - DAMS – virtual application server and database (6 x 2.4GHz Intel Xeon with 32GB RAM) externally hosted and managed
    - COL – AWS cloud infrastructure
- Software
    - CMS – zetcom MuseumPlus, Groovy (Java-based language), API; VEEAM backup; Axiell EmU, Perl
    - DAMS – Montala ResourceSpace, PHP scripting, API; full AWS backup
    - COL – Oxford Mosaic (Drupal-based web platform), ElasticSearch, S3 bucket
    - Python middleware
    - Plugins for integration (between MuseumPlus and ResourceSpace) and authentication (with Azure AD)
- Storage
    - DAMS – 45TB x 2
    - CMS – 750KB x 2
    - COL – as above
- Data and content
    - DAMS – JPG, TIFF, MP3, MP4, PDF, and OBJ files
    - CMS – XML
    - COL – as above
- Audiences, users, disciplines, subjects
    - Researchers (e.g. university students, postgraduates, academics)
    - Educators & students (e.g. primary and secondary schools, informal learning classes, families)

    - General public (e.g. interested adults, hobbyists)
    - Fine art, Archaeology, Anthropology, Natural history, Scientific history; Art, History, Culture, Humanities; Digital humanities/Data science
- Number and nature of resources/front-ends supported
    - CMS – 4 instances
    - DAMS – 4 instances
    - COL – 4 sites, and 50+ legacy sites (under review)
- Access methods: discovery, analysis
    - Internally via API
    - Externally via COL front-end, using search and browse functionalities (or internal exports on request)
- Place in research lifecycle
    - Live collection, i.e., data and assets updated/created/enhanced on a daily basis
- Staffing: number and type
    - 9 members of the team – 2 technical data/scripting (akin to RSE)

**Issues and problems**

- Lack of core resource to maintain and develop service (often reliant on project funding); and challenge finding people with requisite combination of subject-specific knowledge and technical skills
- Limitations of off-the-shelf systems within the sector resulting in service complications and/or additional layers (e.g. not sufficiently cloud-based, or lacking functionality to enable adequate preservation or interoperability).
- Lack of minimum infrastructure requirements relating to data integrity/interoperability for external suppliers leads to major resource from GLAM to improve these systems and invest in their development roadmaps
- Backlog of legacy databases and microsites that need to be brought into the central systems through an ongoing programme of work to improve access to the collections; resource intensive and in the meantime poses digital preservation risk of data loss
- Need for more computational expertise and resource to help de-duplicate and link data using machine learning etc.
- Lack of unified domain-specific standards/authorities to assist linking up data across the sector and internationally
- Copyright law makes the provision of GLAM collections content for research difficult without major resource investment (e.g. legal advice, risk management, content licensing). Embedding copyright policy into automated, technical workflows to determine whether an asset can be shared internally/externally currently results in a large amount of withheld content.

**Future requirements**

- Hardware - Scalable and secure cloud-based infrastructure
- Storage - Estimated increase of 5% per year (e.g. 5-year projection 60TB DAMS and 1TB CMS x 2 including backups; COL likely to scale at same rate)
- Software - additional web applications to feature collections for different audiences and purposes,machine learning or image recognition to provide users with content

suggestions, use of linked data and IIIF to provide additional collections visualisation/ browse tools, crowdsourcing tool for users to suggest data/content improvements

- Access methods – APIs to enable researchers or content aggregators to search, query and extract data and images via COL
- Preservation – expansion of microservices to monitor and report on digital collection, application of RDM practices to the development and maintenance of collection resources or microsites

**References**

University of Oxford, *GLAM Digital Programme* <https://www.glam.ox.ac.uk/digital-strategy> [accessed 08.04.2022]

**ALICE: The Aggregate Line Inspector & Collaborative Editor**

Dr. Samantha Blickhan, Zooniverse & Adler Planetarium, Chicago IL
samantha@zooniverse.org

**Description**

ALICE is a browser-based web application which includes an editorial interface for working with data generated by text transcription projects on the Zooniverse crowdsourcing platform. It was built in response to a direct need from teams running crowdsourcing projects with the aim of transcribing historical and archival records from digitized images. As a quality control method, Zooniverse projects use multi-key transcription, meaning that multiple users transcribe a given unit of text multiple times and the results are then aggregated together to determine the consensus result. Long strings of text require extremely complex aggregation methods, and teams have struggled to work with the data outputs from their projects, particularly those from smaller or under-resourced institutions with fewer opportunities to partner with specialists.

ALICE works with two main data types: 1) digital images of archival materials containing text; and 2) crowd-generated transcription data and metadata generated through Zooniverse projects. First, a subject (usually an image of a manuscript or other archival document page) is transcribed by volunteers via a Zooniverse crowdsourcing project. Fully-transcribed images and their transcriptions are automatically sent to ALICE and the line-by-line data is aggregated as part of that process. Teams use their Zooniverse credentials to log into ALICE, where they can review and edit their project results. ALICE also includes expanded data export options, including .txt files alongside raw, unparsed .json and .csv files, thereby increasing usability of Zooniverse data and expanding access to crowdsourcing as an option for creating digital versions of archival text.

Initial support for ALICE was provided by a Level III Digital Humanities Advancement Grant from the U.S. National Endowment for the Humanities. Our next steps include implementing feature requests submitted by our first round of users and expanding the tool types supported in ALICE (e.g. including support for structured transcription data like tables or forms).

**Current state**

ALICE consists of an Application Programming Interface (API) called TOVE (Transcription Object Viewer/Editor), as well as the ALICE user interface (UI), which allows project team members to view, edit, approve, and export transcription data. ALICE is free to use, and the underlying code is open source and publicly available through GitHub, with one repository for TOVE, and one for ALICE. Both code repositories include User-facing documentation is available via the ALICE About page. The aggregation scripts used in ALICE can be found in the Zooniverse aggregation documentation page.

ALICE viewing features include: side-by-side comparison of aggregated transcription results with the original manuscript page; viewing all submitted individual transcriptions for any given line; automatic flagging of low confidence aggregate results (i.e. high disagreement among

transcribers); access within the viewer to all metadata for a given manuscript page (author, title, date, catalog URL, etc.) as well as for a given transcription (user ID, timestamp, etc.).

ALICE editing features include: ability to designate a correct transcription from individual submissions or input a novel transcription for any given line; adding additional metadata for in-app interactions like editing and saving (e.g. editor ID and timestamp); ability to export both raw and aggregated transcription data as CSV files; ability to adjust aggregation parameters on a per-subject basis, or choose between available aggregation algorithms depending on the data format and transcription methods used.

The initial development and implementation phase for ALICE is complete. Since development finished in August 2020, sixteen teams have used the app so far, with half a dozen more currently developing crowdsourcing projects that they intend to use with ALICE. The ALICE user base is international, and represents public libraries, state archives, public charity archives, and universities. At least two of the teams identify as being small or understaffed, and one team is working with documents from the archives of a Historically Black College or University.

The main effort creating ALICE involved six team members: the Project Director, a Designer, Front- and Back-End Developers, and a Data Scientist, as well as the broader Zooniverse team.

**Issues and problems**

So far, the main issues faced by ALICE users include the need for batch actions, like approving or deleting data. In its current state, all images must be reviewed and approved manually in order to be exported from the app. There have been a few minor bugs identified and fixed, mostly due to unanticipated user needs, e.g. needing to paginate more than 100 pages for a given index.

**Future requirements**

To continue supporting our growing community of ALICE users, the primary need is continued funding to incorporate feature requests and expand the remit of ALICE to include other tool types and data formats. Access to aggregated data is a frequent request for Zooniverse users from many research areas, and ALICE has already inspired a number of efforts around viewing and accessing aggregated data from Zooniverse projects, including a prototyping effort with the British Library to incorporate methods for working with IIIF subjects and data exports.

**References**

Blickhan, Samantha. New developments in crowdsourced text transcription. Poster presented at Association for Computers and the Humanities 2021, July 21-23.

Ridge, Mia, Blickhan, Samantha, Meghan Ferriter, et al. Case study: making Zooniverse data more easily re-usable with ALICE. In *The Collective Wisdom Handbook: Perspectives on Crowdsourcing in Cultural Heritage*, Chapter 10: Working with crowdsourced data.

**Knowledge graphs – Mapping Manuscript Migrations**

Toby Burrows, University of Oxford

**Description of resource**

The primary goals of the **Mapping Manuscript Migrations (MMM)** project, funded by the Digging into Data Challenge of the Trans-Atlantic Platform between 2017 and 2020, were to bring together data relating to the history and provenance of medieval and Renaissance manuscripts and to explore the research potential of the aggregated dataset. Three different data sources were incorporated into MMM: the Bodleian Library's online catalogue *Medieval Manuscripts in Oxford Libraries*, and two manuscript provenance databases – the Schoenberg Database of Manuscripts (US) and Bibale (France). The main aim was to support large-scale research questions across metadata documenting the characteristics and history of more than 220,000 manuscripts. These questions include: production and ownership patterns, movements of manuscripts across time and space, and historical trends in manuscript collecting and sales.

MMM used the Linked Open Data publishing model and a range of W3C Semantic Web standards and technologies, including Universal Resource Identifiers (URIs), the Resource Description Framework (RDF) data model, ontologies like CIDOC-CRM and $FRBR_{OO}$, and the SPARQL query language for querying RDF data. The RDF data were published for direct exploration and downloading, via Linked Data Finland and Zenodo, and a public SPARQL endpoint was also implemented. A public portal (built with the Sampo-UI software) provides browsing, filtering, and searching functionality, as well as some map-based visualizations. Transformation pipelines were built to create RDF data from relational databases and TEI-XML descriptions of manuscripts, and to map the RDF data to the MMM Data Model. Vocabularies were reconciled automatically using VIAF, TGN, and GeoNames identifiers, and also semi-automatically using the "Recon" tool developed at Aalto University.

Future goals include expanding the coverage of the MMM data to include additional manuscripts from other European and British sources; linking the MMM data to IIIF manifests for digitized copies of manuscripts; implementing an identifier framework for medieval and Renaissance manuscripts (since none currently exists); and extending the Linked Open Data cloud for medieval studies to include more specialized vocabularies for people, organizations, and works (rather than the more generic vocabularies used in MMM). There is also the potential to incorporate the MMM data into a broader "Medieval Britain" Linked Data service, for example, enabling manuscripts to be studied alongside other resources for art, architecture, archaeology, musicology, history, and other disciplines, as well as into the Linked Open Data cloud more generally.

**Current state**

- Hardware
    - Virtual server on the Linked Data Finland platform, maintained by Aalto University with funding from Tekes – the Finnish Funding Agency for Technology and Innovation.
- Software:
    - Apache Jena Fuseki triple store – for housing the RDF data and providing the public SPARQL endpoint.
    - Sampo-UI (Open Source) for user interface with browsing, searching, and some visualization capability. This consists of a NodeJS backend built with the Express framework, and a client based on React and Redux.
    - 3M tool for transforming data to RDF.
    - Recon tool for semi-automatic reconciliation of vocabularies.
    - ResearchSpace (Open Source) plus BlazeGraph triple store was deployed as an alternative user interface and delivery platform, housed at the University of Oxford.
- Storage
    - 1.4GB in a triple store for the RDF data (a copy is also stored on Zenodo)
- Data and content
    - 24 million RDF triples (Turtle serialization) + schemas
    - Source data which have been converted to RDF triples but are unreconciled and not yet transformed to MMM Data Model are housed on GitHub
    - Documentation on GitHub
    - Static content (seeking future project funding to extend)
- Audiences, users, disciplines, subjects
    - Academic researchers, curators, and librarians interested in the history and provenance of medieval and Renaissance manuscripts
    - Relevant to an extensive range of humanities disciplines, including history, art history, musicology, history of science, languages and literatures
- Access methods: discovery, analysis
    - Browsing, filtering, and searching through Sampo-UI interface
    - Querying with SPARQL at public SPARQL endpoint
    - Some visualizations through Sampo-UI interface and Yasgui SPARQL query interface
    - Data export via CSV and as complete RDF files for analysis in external systems
- Place in research lifecycle
    - The MMM data have been extracted from a variety of printed, handwritten, and online catalogues; processed into three major source datasets managed by cultural heritage institutions; and transformed into a Linked Data knowledge graph for sharing through user interfaces and by data downloads. Some sample analyses using the MMM data have been published separately.
- Staffing: number and type
    - Technical support
        - No specific or formal support; ad-hoc support from former project team members (Linked Data researchers in Aalto University's Semantic Computing Research Group).

        - Linked Data Finland is supported by Aalto University's Semantic Computing Research Group.
    - User support
        - No assigned staff; ad-hoc support from members of former project team (librarians, manuscript researchers, Linked Data specialists)

**Issues and problems**

- The MMM knowledge graph uses the Linked Open Data set of standards, including URIs, RDF triples, ontologies like CIDOC-CRM and $FRBR_{OO}$, and the SPARQL query language. It would also be possible to use graph database technology like Neo4j to achieve most of the same aims, but without the potential interoperability of the MMM standards framework. The MMM data can be reused in any software environment which supports these Linked Open Data standards.
- The MMM data are made available for reuse under a CC-BY-NC 4.0 licence. The source data were provided to the MMM project under the same licence.
- Future development of this knowledge graph is dependent on further funding for additional projects. It is currently unclear how best to make the transition for MMM from a project output to ongoing digital infrastructure.
- The MMM Project encountered limitations and ambiguities in its source data, arising from incomplete data modelling and loosely structured narrative text fields. Using Named Entity Recognition technologies to identify and extract persons, organizations, and places as components of historical provenance events could overcome some of these limitations.
- The expertise to develop and maintain Linked Open Data knowledge graphs is highly specialized, and currently tends to be available mainly in the context of academic research projects. A more systematic national approach to ensuring and sharing ongoing skills and expertise in this field is highly desirable.

**Future requirements**

- Hardware
    - An environment capable of running graph databases and the Linked Open Data software stack, such as AWS Neptune. (Microsoft's CosmosDB supports graph databases but not RDF triple stores.)
- Storage
    - Additional storage for expanding the scope of the RDF data relating to medieval manuscript studies would amount to perhaps 5GB in total. Extending this to medieval studies more broadly could increase this by an additional 5-10GB.
- Software
    - For data produced in conformity with the various components of the Linked Open Data stack, a range of different software could be used. The main types of software would be: triple store; identifier minting and management service; software for vocabulary reconciliation; software for data modelling and ontology development; NER tools; visualization tools; exploration and discovery tools.

    - There are various Open Source options for these components. Commercial solutions like Metaphactory or GraphDB might provide more consistent support, but at a significant cost.
- Access methods
    - Data download; direct data querying with SPARQL or similar; exploration, discovery, and visualization through user interfaces and portals.
- Preservation
    - Institutional and national ownership and support for Linked Open Data services and data, provided in the medium to longer term by IROs and universities, rather than relying on individual projects, would do a great deal to preserve this kind of data and make it interoperable across subject domains.

**References**

Eero Hyvönen et al., “Mapping Manuscript Migrations on the Semantic Web: A Semantic Portal and Linked Open Data Service for Premodern Manuscript Research,” in: *The Semantic Web: Proceedings of the 20th International Semantic Web Conference (ISWC 2021)* (Lecture Notes in Computer Science, vol. 12,922) (Cham: Springer, 2021), pp 615–630.

Mikko Koho et al., “Harmonizing and Publishing Heterogeneous Pre-Modern Manuscript Metadata as Linked Open Data,” *Journal of the Association for Information Science and Technology (JASIST),* vol. 73, no. 2 (May 2021), 240-257.

Toby Burrows et al., “Medieval manuscripts and their migrations: Using SPARQL to investigate the research potential of an aggregated Knowledge Graph,” *Digital Medievalist* 15 (1) 2022. https://journal.digitalmedievalist.org/article/id/8064/

## Architecture diagrams

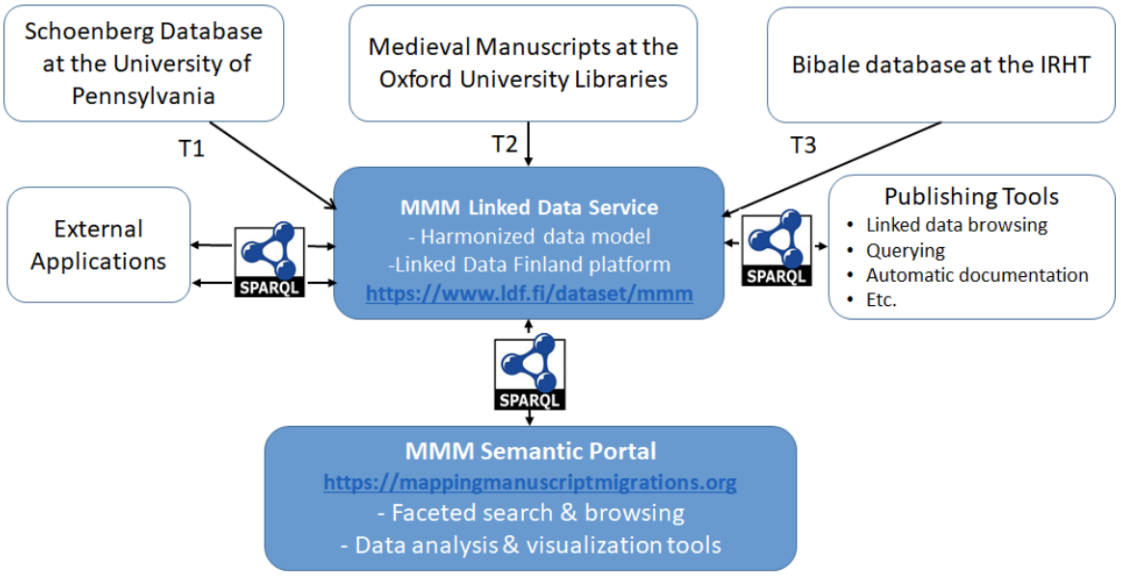



**Figure 1: MMM Publishing Model**

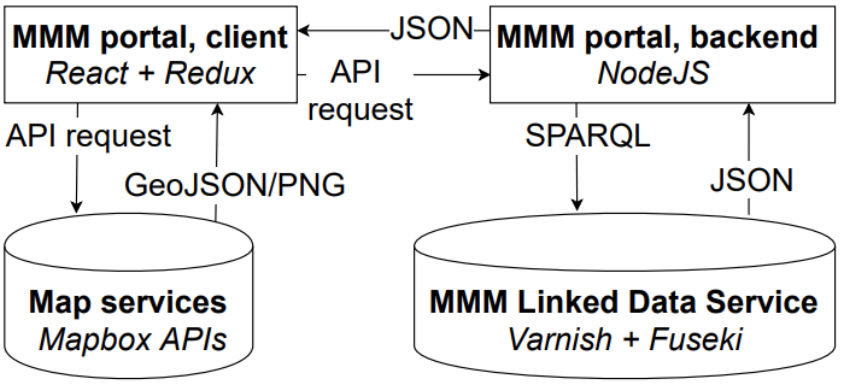



**Figure 2: MMM Portal Architecture**

**Rider Spoke – Riders Have Spoken (Archive)**

Alan Chamberlain Alan.Chamberlain@Nottingham.ac.uk

**Description of resource or infrastructure**

- Its nature and purpose

Rider Spoke was a joint project between the Mixed Reality Lab and Blast Theory exploring the notions of gaming, narrative, location, movement and the ways in which people engage and understand digital infrastructure when it is framed in narrative. In this case the cellular network. Riders Have Spoken was an initial investigation in how to start to archive and re-use the content in a scholarly way.

From Blast Theory "Rider Spoke continues our fascination with how games and new communication technologies create novel social spaces where the private and the public intertwine."

"Rider Spoke invites you to cycle through the streets of the city alone with a smartphone on your handlebars and a voice in your ear. You're asked questions about your life, searching for a place to hide your answer to each one. As you cycle, you get to choose - answer another question, or look for the hiding places of others and hear what strangers have to say?"

- Technological approach

The initial technological approach of the experience Rider Spoke was based around the use of Cell IDs to locate digital content. This works by identifying a Cell ID and attaching content to that Cell ID which is then triggered by the software on the mobile device. The person taking part in the experience believes that they have discovered the content (wherever they think it has been placed – under a bench or behind a wall or so on, but the content is available across the cell). Participants and the artists were also able to leave content in a given Cell using an authoring tool. The system is now run by Blast Theory and has been updated to use apps and geolocate content. The original Riders Have Spoken system was desktop computer based.

- Future direction and goals

The experience Rider Spoke is now hosted by the artist collective Blast Theory. It has been deployed recently, more information can be found on their web page, listed above. The

archival system *Riders Have Spoken* was moth-balled and would need updating to work with the current system, although the content would effectively be the same – Audio Files and Textual notation. This could be extended to incorporate to add script notations and sketches if linked to an iPad or something with those capabilities if this was going to be used in a long-term research context. Having the ability to notate on the fly and see other people's notations that go beyond typed text would present users with a flexible way to develop and dig into an archive, adding map layers and transparencies which relate to the temporal aspects of the experience could also enhance the archive. Public facing annotations would also be an interesting feature as this would enable people to collaborate. Using an archival system would also allow people to re-experience the content from the audience's perspective – imagine being able to take part in a geo-located experience from 50 years ago! There could be a whole plethora of content that is useable in a range of different formats, which go beyond the digital.

**References**


Riders Have Spoken – Top Level Video for a look at the graphics
https://vimeo.com/54374073

Chamberlain, A. et al. "Locating Experience: touring a pervasive performance", *Personal Ubiquitous Computing Journal*, Volume 15 Number 7. Springer Verlag. ACM Library. DOI: 10.1007/ s00779-010-0351-3. pp. 717-730 – PDF

Jonathan Foster, Steve Benford, Alan Chamberlain, Duncan Rowland & Gabriella Giannachi (2010) "Riders Have Spoken: A practice-based approach to developing an information architecture for the archiving and replay of a mixed reality performance", *International Journal of Performance Arts and Digital Media*, 6:2, 209-223.
https://www.tandfonline.com/doi/abs/10.1386/padm.6.2.209_1

Rider Spoke Video (in re-development)

From the Blast Theory Website:
Rider Spoke 2022
https://www.blasttheory.co.uk/projects/rider-spoke/

Riders Have Spoken
https://www.blasttheory.co.uk/projects/riders-have-spoken/

**King's Digital Lab Infrastructure at King's College London**

Author: Arianna Ciula (with contributions by James Smithies as ex KDL Director, Brian Maher, Tim Watts, Miguel Vieira, Pam Mellen, Tiffany Ong; reviewed by Matt Penn in eResearch)
Arianna.ciula@kcl.ac.uk (lab contact address: kdl-info@kcl.ac.uk)

**Description of infrastructure**

- Its nature and purpose

King's College London (**KCL**) has been at the forefront of Research Software Engineering (RSE) for the arts and humanities (A&H) for over fifty years. Several incarnations of what is currently the renowned Department of Digital Humanities have developed a considerable technical infrastructure, currently hosted at the University of London Computing Centre (ULCC). King's Digital Lab (**KDL**) was established in 2015 to rationalise and manage this infrastructure, with a particular emphasis on RSE best practice, security, and sustainability. Primary KDL partners include the UK HE and cultural heritage sectors with secondary partnerships in the creative industries.

KDL inherited 113 projects (50 AHRC) when it was established, worth a total of £39,886,078 research funding (£25,705,694 AHRC). After implementing an internationally recognised archiving and sustainability programme to manage projects beyond their funding period, the Lab cares for 55 projects under Service Level Agreements (30 AHRC), has remediated a further 28 to low-cost hosting solutions (10 AHRC) and migrated 22 to other institutions (9 AHRC). The team manages 22 active projects (6 AHRC). Projects hosted on KDL infrastructure have been submitted to the Research Excellence Framework (REF) as research outputs or part of Impact Case Studies. The Lab's infrastructure is a significant node in the UK's A&H e-infrastructure, providing long-term support for dozens of high-impact research projects.

The Lab's holistic conception of infrastructure, which aligns to UKRI definitions that include human and procedural elements alongside technical components, is viewed as a best-practice model within KCL as well as externally. Its blend of scholarly RSE and IT industry methods for infrastructure design, development, and maintenance position it as a key interface between the institution's eResearch and IT functions. This central alignment to institutional strategy is enhanced by the integral relationship between KDL and King's eResearch strategy, which has recently gained significant institutional investment for High Performance Computing (HPC), storage, and permanent FTE staff.

- Technological approach

KDL infrastructure is based on VMware virtualisation technology and is regularly upgraded to ensure sustained levels of performance and reliability. Daily backup procedures are in place, with each backup being retained for 30 days. Backups can be used not only for data recovery but also to fully recover a virtual server into an operational state.

In December 2021 KDL completed the remediation and upgrades of its infrastructure to increase its long-term sustainability and ensure alignment with KCL eResearch and IT. As part of this effort: 44 servers were upgraded and 28 decommissioned; 38 applications were upgraded or rebuilt; 25 websites were converted to static; 6 websites were migrated (3 to the cloud); 17 applications were decommissioned. It should be

noted that some KDL-hosted sites have multiple applications and some applications are distributed across or access multiple servers as per the diagram below.

- Future direction and goals

The significant attack surface presented by KDL's public-facing estate and the range and diversity of serious security exploits (contained and remediated in conjunction with KCL Cybersecurity) having to be fended off have created an urgent need to **migrate** from on-premises infrastructure to the more secure central **eResearch** operated CREATE **platform** (https://doi.org/10.18742/rnvf-m076). In addition, the KDL infrastructure is approaching end of life and a replacement infrastructure would have been needed by December 2022 at the latest. Integration with eResearch services will provide KDL with a **flexible long-term infrastructure roadmap**, for the first time in its history. This will begin with an immediate migration to eResearch (testing ongoing at the time of writing) to mitigate urgent security and end of life issues.

KDL technical strategy is also in evolution, informed by RSE best practices and the specialisms of the areas of knowledge and production the Lab operates in (mainly Arts and Humanities, cultural heritage and digital creativity). The technical stack will therefore evolve to support an archiving-first approach and produce modular software, re-usable components and generous interfaces.

## Current state

- Hardware

Hardware includes 5 Dell R640 servers (Intel Xeon Gold 6154 3GHz, 18 cores 25MB Cache, 512GB RAM per host, upgradeable to 768GB RAM, for a total of ~2.5TB), a largely solid-state SAN (44 x 960GB SSD disks, 6 x 6TB 7.2k spinning disks) comprising 33TB RAID6 and an 18TB slow archive. Lab infrastructure is connected to the university network via a 10GB network connection.

A powerful desktop machine (Threadripper 1950x 16 Cores, 128GB, 2 X NVidia 1080Ti GPU) currently housed at KDL is also accessible by team members and other staff or research students on an on-demand basis.

- Software

KDL's current infrastructure comprises ca. 110 virtual machines, running Ubuntu Linux and a software stack that includes a range of software (Java, PHP) but primarily Django/Python, JavaScript, and associated management tools (Solr/Elastic Search, Vagrant, Docker). 11 additional servers are used for centralised services such as image and other data storage, mail and user authentication. A range of network, security, and website analytics tools are used for monitoring and updates. See architectural diagrams below.

- Storage

55TB of disk space provided by a local iSCSI SAN. The entire system is backed up via KCL's eResearch ArcaStream off-site storage daily.

- Data and content – type and size; rate of growth

The KDL team builds solutions ranging from archives and repositories to digital scholarly editions, data visualizations, software modules, notebooks and extended reality (XR) products. Projects content range across the following non-exclusive categories:

1. access to metadata, texts, and image sources of multiple types of cultural objects;
2. analytical resources of selected sources, artefacts and works;
3. experimental projects using immersive technology, augmented reality and machine learning methods.

Category 1 includes longstanding and pioneering projects in the digital humanities that continue to have substantial reach and impact in the academic community and wider public domain. Examples include the *Prosopography of Anglo-Saxon England*, the AHRC *Research Centre for the History and Analysis of Recorded Music,* and the *Clergy of the Church of England Database*; during the past REF cycle (2013-2020) these together reached almost 64 million page hits (excluding requests for images, css, or javascript files, local server access requests, crawlers and bots). One project alone in this category, *Nineteenth-Century Serials Edition*, provides over a million digital assets.

Very different projects comprise category 2, from critical scholarly editions related to a set of works or authors such as *The Values of French*, *The community of the realm in Scotland (1249–1424)* and *Jane Austen Fiction Manuscripts* to interpretative analysis supported by data collection and curation such as the popular *The Redress of the Past: Historical Pageants in Britain (1905-2016)* and *The Art of Making in Antiquity*.

Category 3 represents a smaller proportion of the KDL estate; high profile projects such as *The Digital Ghost Hunt* are driving innovation at the intersection between research organisations, creative industries, and cultural heritage. Our current infrastructure is not designed to host creative products emerging from these projects but requirements are expected to evolve in the future. Notebooks and generous interfaces are also part of this category.

- Audience(s), users, discipline, subject

KDL's infrastructure is used to host multiple legacy projects that remain in active use by the research community as well as native projects (i.e. started after its set-up in 2015). Core **audiences** include researchers and professionals affiliated to HEIs and other research organisations, students enrolled in HEI programmes, independent researchers, government employees, secondary school students, and the public. KDL's portfolio of projects covers a plethora of **disciplines** in the A&H, from history, archaeology, historiography, art history, epigraphy, palaeography and musicology to information studies, literary and modern languages studies, theatre studies, performance and practice-oriented research in the cultural and media sectors.

133 websites (out of the total 154 for which logs could not be aggregated) hosted by KDL during the past REF eligible period up to September 2020 gained a total of 242,795,194 hits (a conservative estimate due to the semi-automatic methodologies used to aggregate web logs). Of those sites, 14 only contained 1,485,229 digital research assets (any file in the following format: png|jpg|jpeg|jp2|mpeg|avi|mp3|mp4|wav|tif|tiff|avi).

- *For infrastructure cases: number and nature of resources / front-ends supported*

KDL manages ~60 high profile DH projects, supported by a Systems administration team responsible for management of ~106 running virtual servers (plus ~140 Virtual Machines planned to be decommissioned in 2022-23) hosted on local infrastructure. It also hosts 116 static sites, many of which are legacy static sites.

- Access methods: discovery, analysis

Access to KDL services is primarily via public (and occasionally private) websites. A small subset of users (internal to KDL SDT team plus a very small number of contractors and visiting researchers) access KDL servers through Secure Shell (SSH). Access control is provided either by a KDL operated LDAP and Kerberos servers or Shibboleth, depending on project requirements.

- Place in research lifecycle, e.g., collected, processed, analysed, shared

Projects in the KDL portfolio defined or are defining new **methodologies** for data collection, processing, analysis and presentation typically involving more than one discipline and cutting across academic and cultural heritage sectors, and increasingly the cultural industries. Development of projects is undertaken using a mature Software Development Lifecycle (SDLC) that accounts for the full lifecycle from design – inclusive of the definition of estimated infrastructural requirements and associated costs – to sun-setting. It is guided by quality guidelines (including REF outputs checklist) and a robust archiving and sustainability policy (overseen by a Service Level Agreement (SLA) Committee to determine web hosting and infrastructure SLA renewals balancing scholarly value with contractual commitments, cost, security, and reputation). Project data are usually released gradually over the course of the project period under a license agreed with partners. KDL makes any project application source code freely accessible at KDL's GitHub repository (https://github.com/kingsdigitallab/). Unless otherwise specified, users can download this material and reuse it under the terms of the MIT License (https://opensource.org/licenses/MIT).

KCL has an official Research Data Management (RDM) Policy and a comprehensive RDM service including a research data repository facility for depositing and sharing project-generated data and metadata. This is currently an institutional instance of Figshare which uses Amazon web storage infrastructure and has been certified for the ISO/IEC 27001 information security management standard. Backups are provided by both the Amazon web storage service and Figshare's own daily backups and weekly snapshots of the entire repository. The repository services are free at the point of use with a commitment to preserving research data for a minimum of ten years after the end of the project. All datasets are issued a DOI. If agreed as part of archiving and sustainability planning for a project, selected datasets can also be archived on KDL CKAN catalogue https://data.kdl.kcl.ac.uk/ (with standard open licenses and DOIs).

- Staffing: number and type (e.g., RSEs)

KDL team includes 13 Research Software Engineers (RSE) equivalent to 12.5 FTE now. The roles (descriptions at https://zenodo.org/record/2559235) include RSE analysts, developers, UI/UX designers, project and systems managers.

    - Technical support

KDL Director delegates technical infrastructure to a Senior Systems Manager (1.0 FTE) and a Senior RSE (1.0 FTE) with responsibility for security hardening and migration. Additional technical support is provided by KDL's Principal Research Software Engineer, with input from KCL IT, eResearch, and KCL Cybersecurity where required.

- User support

KDL project analysts manage the scope of requirements for active projects (under development) and work is performed by the KDL Solution Development Team (SDT) in a series of targeted, dedicated timeboxes. Any bugs or other issues can be reported to those analysts for active projects or using the contact web form for completed projects with frontend or admin interfaces. Capacity assessment and scheduling of tasks for the SDT are then coordinated by KDL lab and project managers.

### Issues and problems

As summarised across this document:

- Security risks
- End of life of technical infrastructure
- Support for archiving and sustainability effort (inclusive of funding to sustain cyclical sustainability assessment and decommissioning process – see more details on costs below)
- Challenges around monitoring methodologies across servers (e.g. aggregation of web logs)
- Future requirements to be scoped (e.g. with respect to digital creativity, use of AI/Machine Learning methods and heritage science)

KDL's legacy project estate costs approximately £119k per year to maintain (projected to be ca. 116k by early 2023); part of these costs is covered from contribution of external partners. The balance is covered by KCL Faculty of A&H. About £32k of the Faculty total covers legacy projects which are still within the funder-mandated post-project support period (forecasted to be ca. £20k by early 2023), leaving a balance of about £45k for projects which FAH has chosen to support for scholarly reasons (forecasted to amount to ca. £63k by early 2023).

### Future requirements

- Hardware

KDL system is currently in the process of being migrated to central KCL eResearch OpenStack infrastructure. KDL will host servers on the KCL CREATE OpenStack instance with 2 dedicated hypervisors. This equates to 128 CPU cores, 3TB of RAM and 8TB of local storage. The institutional target is to be net zero carbon by 2025. Upgrading from on-premise infrastructure to eResearch will eliminate e-waste from KDL's core production infrastructure. KDL will retain a smaller on-premise footprint for projects with specialist needs such as use to test Machine Learning methods and Digital Creativity applications. This flexibility and scalability will be particularly important for delivery of future solutions related to digital creativity and data science.

- Storage

As part of the migration to eResearch, KDL's storage requirements will be fulfilled by a CephFS SSD backed volume (40TB) and an additional CephFS HDD backed archival-tier volume (25TB). In addition to this, 120TB of off-site storage, provided by ArcaStream, will be used to hold backups.

- Software

No major changes from current status are foreseen but future requirements are to be defined based on the evolution of our technical stack as discussed above.

- Access methods

Whilst Shibboleth access will be maintained for projects that need it, in the future KDL plans to migrate websites currently using LDAP to use SSO provided by the KCL Active Directory service. This will negate the need for colleagues to have a second account purely for accessing KDL services. External users will be provided with affiliate accounts where required. Finally, KDL are in the process of enabling the use of SSH from behind a gateway provided by KCL eResearch, which features both SSO provided by the Active Directory and two-factor authentication.

- Preservation

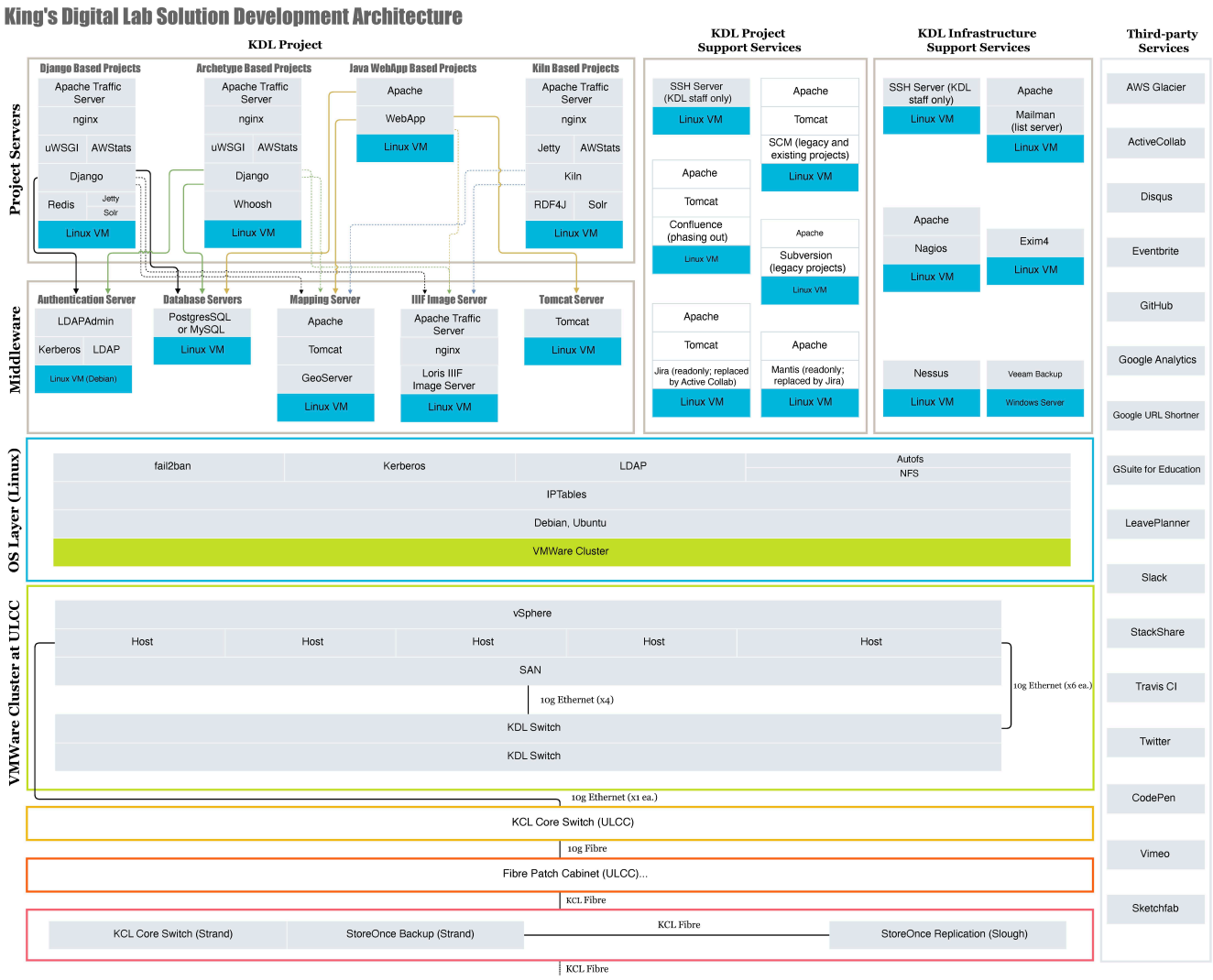
KDL projects sit in a crucial space between national and regional-scale outputs and small locally hosted initiatives and have been underserved for infrastructure funding despite being retained in many cases for years beyond their initial funding period. In many cases the cost of their infrastructure has been supported by King's College London gratis, in some cases for a decade or more: without **funding support** this generosity will be difficult to sustain. In this sense, KDL not only enables the future strategic priorities of AHRC and UKRI, but bridges sustainability gaps that the emergent AHRC Digital Research Infrastructure investment and the ongoing Scoping Data Services scheme appear designed to fill.

## References


- KDL, *Archiving and Sustainability: KDL's pragmatic approach to managing 100 Digital Humanities projects, and more.*. https://kdl.kcl.ac.uk/our-work/archiving-sustainability/ (updated 2022)
- Jakeman, N. *Safeguarding an inheritance and ensuring a legacy: Software Development Lifecycle for Research Software Engineering* https://kdl.kcl.ac.uk/blog/sdlc-for-rse/ (2020).
- Smithies, J., Westling, C., Sichani, A.M., Mellen, P., & Ciula, A. Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King'sDigital Lab. *Digital Humanities Quarterly* 13.1 (2019).

## Architecture diagrams

### Present Infrastructure Architecture




### Present Development Stack

| Dependencies | Languages | Frameworks | Building | Testing | Continuous Integration | (Continuous) Deployment |
|---|---|---|---|---|---|---|
| pip | Python | Django | Django | Pytest | | |
| | | Django REST | | Tox | Fabric \| GitHub Actions \| GitLab | |
| npm | JavaScript | Vue | Vue Cli | Jest | | |
| | SCSS | Bootstrap | Gulp | Siteimprove | | |
| PostgreSQL | | | | | | |
| Elasticsearch / Kibana | | | | | | |
| Pa11y | | | | | | |
| | | | | | | Redis / nginx / |
| Docker / Docker Compose | | | | | | |

# The UK Web Archive

Ian Cooke <Ian.Cooke@bl.uk>

## Description

The UK Web Archive captures and preserves UK web sites, making them available for access and research. It is a collaborative effort funded across the UK Legal Deposit libraries, led by an interdisciplinary team at the British Library. The archive is formed from an annual Domain Crawl which creates an archived copy of all openly accessible web resources that can be identified and crawled. In addition, web pages and web sites that are identified as being of particular value are crawled more frequently throughout the year. This includes the creation of event based and thematic collections. The UK Web Archive contains archive copies of millions of web sites and billions of resources.

The team consists of 4.5 FTE curatorial roles and 4 FTE technical roles.  This is a small team, which means our tactics are:

- Automate all routine processes:
    - Requires higher-level technical skills to maintain, but scales better.
- Use commodity/*de facto* standard infrastructure software and services where possible:
    - For example, we use Apache Hadoop as it is a very widely used storage and compute platform, and build tooling around those standard interfaces.
- Where custom tools cannot be avoided, share the burden through open source projects, collaborating with international partners operating in the same niche:
    - We work with other web archives to maintain and improve our tools, like crawlers, playback systems, search platforms and researcher tools/documentation.
- Decoupled Architecture:
    - Ingest and Access processes run against the same Storage platform. Strong data standards (WARC format) and stable interfaces (e.g. Apache Hadoop) allow Ingest and Access services to run and evolve independently.

The terms of the legal regulations we operate under mean we cannot make the full archived pages available on the open internet without explicit permission from the original publisher.Having said that, as full corpus is now tens of billions of resources from millions of websites, totally around a Petabyte of content, it is not really practical for researchers to download the whole thing. Therefore, we offer a range of access routes:

- If you know the URL, you can search for that directly.
- We have manually catalogued over ten thousand websites, and browsing these collections can help if they match your area of interest.
- We are working towards complete coverage for our full-text search service, which allows you to find pages of interest from the whole corpus (i.e. this does not rely on sites having been curated by hand).

- We publish non-consumptive APIs and datasets generated from the documents, and try to help researchers work directly from these surrogates. (A good example here is the Web Archives section of the GLAM Workbench: https://glam-workbench.net/web-archives/).
- Where time allows, we work directly with researchers, analysing data on their behalf to produce the datasets they need. (A recent example was generating word statistics for a computational linguistics research from the Alan Turing Institute, who could use them to study how words have changed meaning over time).

In terms of future goals, we're are mostly focussed on addressing some aspects of the current approach that are lacking:

- Making the curation process more accessible and scalable, so more people can help us catalogue the archive web, so more people can find themselves in it.
- Expanding the full-text search service to full coverage of our holdings, while maintaining acceptable performance and through a reasonably usable interface.
- Keeping datasets up to date automatically (they are currently rather out of date).
- Empowering researchers to work directly on the data where possible (rather than us being a bottleneck).

### Current state

- Hardware
    - Commodity servers, commodity drives, in an institutional server room supplying power, cooling, networks.
- Software
    - Commodity platforms and services, like Apache Hadoop, Apache Solr.
    - Niche tools, like crawlers and playback, maintained through open source collaboration with other web archives.
- Storage
    - Around 1PB, with three-way replication in each Hadoop cluster.
    - Replica cluster being constructed for the National Library of Scotland.
- Data and content – type and size; rate of growth
    - WARC files (containing data in thousands of formats) and provenance data.
    - Growth rate c. 120TB/year.
- Audiences, users, disciplines, subjects
    - General public
    - Researchers in eg history, electronic literature, geography, public policy, design and other visual arts, human-computer interaction, computational linguistics
    - Journalists
- For infrastructure cases: number and nature of resources / front-ends supported
    - The UK Web Archive Website & APIs
    - Datasets
    - GLAM Workbench
- Access methods: discovery, analysis
    - Find by URL

    - Find by curated topic/theme
    - Find by full-text search and facet-based filtering
    - Analysis of trends in full-text search terms and facets
    - Analysis of metadata and extracted data via datasets and APIs
- Place in research lifecycle
    - A data source
    - An analysis platform
- Staffing: number and type
    - Technical support:
        - 1 FTE hardware/systems/monitoring
        - 1 FTE testing/analysis/reporting
        - 1 FTE development
        - 1 FTE technical lead (not unlike a senior RSE)
    - User support

### Issues and problems

- Fundamentally, not enough technical roles paying competitive wages. This means we do not have time to support research.
- Legal constraints of non-print Legal Deposit regulations mean most items cannot be made available publicly, and also that we cannot take advantage of cloud services.
- The UK Web Archive is a very large and complex collection. This presents challenges for discovery, and for understanding the web archive as a concept, to end users. Different approaches to discovery are used, such as full-text indexing for parts of the archive, and the creation of thematic and event-based collections to identify parts of the archive. However, there remains a challenge to provide entry points to the archive for research that meet user skills and resources.
- The size and complexity of the archive also presents challenges for collection management and preservation (for example, providing information on file formats within the archive, and supporting quality assurance at scale).
- Related to this, there are needs to support skills and methods training for researchers and for curatorial roles.
- Not meeting expectations around social media collecting:  It is often assumed that we are archiving social media in a comprehensive manner, but harvesting social media presents significant legal and technical challenges. Social media is highly influential but it is not being adequately archived at present.

### Future requirements

There are common challenges for support for collection management (e.g., file format identification, preservation), curatorial management (e.g., the creation of representative and inclusive collections; interpretation and  presentation of the archive), and research (supporting search in a sustainable way, creating services that meet researcher skills and tools). These include the ability to create tools that support analytical access to the archive at scale, and the embedding of research skills needed to use these tools effectively.

The UK Web Archive has experience in working with researchers to understand and develop tools and methods to support research using the archived web (e.g., Big UK Domain Data for Arts and Humanities, and The Archive of Tomorrow). This approach could be taken to shape training and tools for researchers, curators and collection managers.

- Hardware
    - The most flexible option would be to move to cloud-based storage and services. This would allow e.g. giving researchers direct but carefully-controlled access to data without impinging on any other operational processes.
    - If cloud services cannot be used, the approach will be the same as now but evolving to higher storage density. If the BL adopts a hybrid cloud approach, this may allow more flexibility.
- Storage
    - S3 (either cloud or S3 compatible on-site, but must be horizontally scalable)
- Software
    - Some switch to standard cloud APIs rather than relying on local software.
    - Focusing on tools and partnerships that help us grow the number of people using the shared custom tools.
- Access methods
    - Improve APIs and datasets (send data to the researcher's tools/code).
    - Allow researcher access to a compute platform (send researcher's code to where the data is).
- Preservation
    - Regularly scan for the formats within the WARCs.
    - Run automated browsers to look for changes in rendering.
    - Adopt JavaScript renderers for more difficult/obsolete formats (as has been done for Flash).

## References


The UK Web Archive user interface
https://www.webarchive.org.uk/

Ian Milligan, 2020, You shouldn't Need to be a Web Historian to Use Web Archives: Lowering the Barriers to Access Through Community and Infrastructure. WARCNet papers, Aarhus, Denmark
https://cc.au.dk/fileadmin/user_upload/WARCnet/Milligan_You_shouldn_t_Need_to_be__2_.pdf

Big UK Domain Data for the Arts and Humanities
https://buddah.projects.history.ac.uk/

The Archive of Tomorrow, National Library of Scotland
https://www.nls.uk/about-us/working-with-others/archive-of-tomorrow/

## Emerging Formats

**Ian Cooke and Stella Wisdom, British Library; Graeme Hawley, National Library of Scotland**

## Description

With the introduction of new digital media in the publishing landscape, cultural heritage institutions including Legal Deposit Libraries are investigating how new formats and their technical dependencies are shaping their collection policies, and how they can ensure meaningful collection and preservation of current complex digital publications over the long term. Since 2013, UK Legal Deposit Libraries (LDLs) have actively collected, preserved and provided access to e-Journals, e-Books and archived websites acquired under the UK's Non-Print Legal Deposit Regulations. Furthermore, in recent years the LDLs have been researching the collection management needs of acquiring and preserving more complex digital content types, referred to as emerging formats. These are defined as born-digital publications with no print counterpart that have strong software and hardware dependencies, and often consist of more than one media type, such as e-Books created as mobile apps and web-based interactive narratives. The marketplace creating these works is continuously changing, and most of these new formats are already at risk of rapid obsolescence.

To date the most practical progress in collecting emerging formats at the libraries has been with web-based interactive narratives, as the collection of these works has been supported by workflows and tools already employed by the UK Web Archive (UKWA). An Interactive Narratives collection (https://www.webarchive.org.uk/en/ukwa/collection/1836) was established by Dr Lynda Clark as part of a post-doctoral placement in 2019 entitled 'Emerging Formats: Discovering and Collecting Contemporary British Interactive Fiction'. This has been followed by the New Media Writing Prize collection (https://www.webarchive.org.uk/en/ukwa/collection/2912), which has collected shortlisted and winning works from this prize organised by Bournemouth University, which showcases and rewards innovative examples of digital interactive storytelling. In addition to archiving interactive websites, there have also been experiments in collecting eBook apps, for example the Android app and PC version of *80 Days* by inkle studios, also apps such as one about T.S. Eliot's *The Waste Land*, which were made by Touch Press and Faber & Faber.

The LDLs meet regularly to discuss emerging formats, and have recently undertaken to do further investigatory and experimental work between them. For example, the National Library of Scotland (NLS) will be looking in closer detail at a piece of digital literature called *All This Rotting*, by Alan Trotter, and *Goldilocks and Little Bear* published by Nosy Crow, to understand not only the long term digital preservation issues, but also what good access looks like in reading rooms. NLS will also attempt to preserve Minecraft St Kilda, and will be talking to another LDL, Cambridge University Library, about their experience of preserving Minecraft St Catherine's. In working together like this, the LDLs can collectively cover more territory, share different approaches, and also, crucially, test interoperability between the six different institutions.

It is not always possible to acquire and preserve partial or complete captures of these complex digital publications, so the collection and creation of contextual information is an approach currently being explored for filling in the gaps to enhance the curation for these items. Examples of contextual information relating to emerging formats include webpages, author interviews, reviews, blog posts and screenshots/screencasts of usage of a work. This work is influenced by The Electronic Literature Organization, which was established in 1999 to promote and facilitate the writing, publishing and reading of electronic literature. They host and curate The NEXT, a virtual space, a multimedial museum, library, and preservation environment, which contains a wide variety of media related to

born-digital interactive writing, including digital files of the works; artist's notes; ephemera including material associated with performances, exhibitions, publications; and research.

## Its nature and purpose

The meaningful collection, preservation and access to complex digital publications over the long term, including relevant contextual information.

## Technological approach

The main tool used to build the UKWA Interactive Narratives and New Media Writing Prize collections was W3ACT, or ACT, the Annotation and Curation Tool designed by the British Library to help curators and subject specialists curate specific parts of the Web. It interfaces with the Heritrix crawl engine built by the Internet Archive. Whenever multi-media content proved to be an obstacle to achieving a complete capture, Webrecorder's ArchiveWeb.page and Rhizome's Conifer were used to patch the original Heritrix capture.

Apps, and similar material, have been collected under legal deposit regulations, either as file transfers from the publishers or downloaded from an app store or repository. In some cases, files have been transferred as uncompiled content, in others as packaged applications. These are stored on a secure network (for preservation) and are installed on a small collection of mobile tablet devices (for investigation and access).

## Future direction and goals

UK Legal Deposit Libraries will trial a 'contextual collecting' approach for a selection of material collected so far. This will provide rich documentation of each work in its intended context, to describe the creator's intent and the user experience. Documentation will include text, images and videos that describe the work (e.g. press releases, reviews); recordings or screen shots of the work; and interviews with the creators. This methodology is similar to the use of 'traversals' to describe archived works of e-literature. Examples of this can be seen on The NEXT and the Cartografia Digital Archive of Digital Literature in Latin America.

The goal is to establish a methodology to create collections of electronic literature and documentary. This will create a collection, shared by and accessible at the UK legal deposit libraries, that will preserve examples of innovative digital publication for use by researchers and creative industries, to understand past practice, support research and enable teaching.

NLS is going to attempt to build a small "test" collection of emerging formats during 2022/2023, and will also set up an internal community of interest on a broader topic of "difficult formats" which will help more recent emerging formats challenges to be seen in the round, especially in terms of gaps in collection, under-represented audiences (some formats might attach especially strongly to certain audiences), and user experience in reading rooms.

## Current state

### Hardware

Collection using web archiving tools relies on the UK Web Archive's technical infrastructure (see Software and Storage below)

2 x iPads and 2 x Galaxy tablets for investigation of significant properties of apps

One objective of the work at NLS will be to investigate hardware requirements

### Software

Collection by web archiving uses W3ACT (the UK Web Archive Annotation and Curation Tool) to manage administrative and descriptive metadata, schedule crawls and manage rights. Web crawling uses Heritrix, supplemented by Webrecorder.

Access to web archive content uses the UK Web Archive user interface (www.webarchive.org.uk) and PyWB. Webrecorder captures can be accessed through the ReplayWeb.page browser-based viewer, or the standalone desktop app. For viewing web archives that include Flash, the ReplayWeb.page desktop app is preferrable, as it includes the Ruffle plug-in, which supports Flash content written in ActionScript 1 and 2. There is currently no access to Webrecorder captures through the UK Web Archive user interface.

### Storage

Content in the UK Web Archive is stored on a Hadoop Distributed File System.

Apps and other material is stored using the British Library's Minimum Preservation Tool (MPT), which provides for replication of content across two geographically separated storage "nodes" and regular fixity checking (checksum validation) of all files on both nodes to help ensure content remains authentic and unchanged.

### Data and content – type and size; rate of growth

Relevant collections in the UK Web Archive stand at over 400 entries.

35GB data is held for offline content in the MPT

Growth is limited by the resources currently in place, including staff resources, and represents a small percentage of the total amount of material that exists and is at risk of loss. The New Media Writing Prize web archive collection grows by around 20 new entries per year.

### Audiences, users, disciplines, subjects

Research uses are cross-disciplinary, including literature, visual arts, media studies, archive and library science, human-computer interaction design, and software development.

The collection is highly relevant for creative industries, to see examples of work created previously, and for gaining inspiration. The collection in particular represents the works of individual creators and small studios.

### Access methods: discovery, analysis

The collections in the UK Web Archive can be discovered and accessed using the user interface (www.webarchive.org.uk) Access is restricted to legal deposit library premises, under legal deposit regulations, with a proportion of the collection permissions-cleared for wider access.

Offline content is currently inaccessible as the infrastructure to support this has yet to be developed.

### Place in research lifecycle

Work is currently experimental, with small collections made so far.

### Staffing: number and type

There is 1 dedicated curatorial post (based at the British Library).

Additionally, there is a proportion of time from across the UK legal deposit libraries for curation, engagement, web archiving and preservation research. Staffing is spread across a variety of skills and functions, including curation, digital preservation, cataloguing and metadata, public access provision, and acquisition and ingest.

## Issues and problems

Work at present is largely restricted to research into issues and collection building using web archive infrastructure. This is due to the very small number of staff funded to support this work (which in turn creates problems for sustainability). This is partially mitigated by a small but supportive network of GLAM institutions with similar collection management challenges.

Similarly, resources are lacking for technological development, in particular to support access.

Cultural and legal barriers, in particular a highly complex rights environment and lack of engagement from large corporations (who are owners of rights in software and firmware).

## Future requirements

Resources to trial and develop a methodology and access system to support archiving of digital literature by collecting and creating information to record the context and use of a work, alongside the preservation at bit level of the digital work and its components. This would follow a traversal method similar to that used by The NEXT archive.

This would require additional staffing in curatorial, rights management, and technology infrastructure management (in particular related to web archiving) to increase the volume of digital works that could be collected, mitigating the risk of obsolescence and loss.

### Hardware

Examples of mobile devices and other technology, such as VR headsets, used to access emerging formats, to allow experimentation, collection of contextual material and as documentation of context.

### Storage

Distributed server system, with geographic replication, for preservation.

### Software

Continued development of high fidelity (e.g., browser based) web archiving and playback tools, e.g., Browsertrix, Webrecorder.

Emulators to support access to publications built using technology and software that has become obsolete (e.g., Ruffle for Flash emulation).

### Access methods

Development of a browser-based platform to support discovery and access.

### Preservation

Development of preservation infrastructure across legal deposit libraries to ensure long-term preservation of digital works and related assets. This would build on the infrastructure already in development for the legal deposit libraries.

## References

Smith, C. and Cooke, I. “Emerging Formats: Complex Digital Media and its Impact on the UK Legal Deposit Libraries.” Alexandria: The Journal of National and International Library and Information Issues. August 2018.

Clark, L., Rossi, G.C., Wisdom, S. (2020). Archiving Interactive Narratives at the British Library. In: Bosser, AG., Millard, D.E., Hargood, C. (eds) Interactive Storytelling. ICIDS 2020. Lecture Notes in Computer Science(), vol 12497. Springer, Cham. https://doi.org/10.1007/978-3-030-62516-0_27

**F-Tempo (Full-Text search of Early Music Prints Online)**

**Tim Crawford (Goldsmiths University of London): t.crawford@gold.ac.uk**
Alastair Porter (Universitat Pompeu Fabra, Barcelona, Spain)
Laurent Pugin & Rodolfo Zitellini (RISM Digital, Bern, Switzerland)
Golnaz Badkobeh & David Lewis (Goldsmiths University of London)

**Description of resource**

*Nature and purpose*

F-TEMPO is a publicly-accessible online resource for content-based, full-text searching of renaissance and early baroque music distributed across music libraries. This musical period (roughly 1500-1700) is full of intertextual reference and quotation, reuse and re-texting, so a tool to draw attention to passage-level relationships across millions of pages of music is particularly valuable. It is accessible both via its own web-based front end and via an API allowing developers to make their own custom applications. Queries can use a variety of methods: entry of a simple alphabetic code representing a pitch-sequence to be located; choice of a particular page within the database as query in order to locate other instances of similar music; upload of a digital image of a single page from a similar early printed music collection (intended as a means to identify unknown music). With tools such as this we hope to be able to offer an unrivalled degree of enhanced access to the printed repertory of music from the early-modern era.

*Technological approach*

The tool builds on Optical Music Recognition (OMR), which transcribes the music within the page-images (optionally using IIIF) into encoded editions, but with inevitable errors. The storage format is MEI (Music Encoding Initiative), which can handle the special features of the historical music notation. From the transcriptions we extract features which represent the sequences of notes in the music; these are encoded as strings in an alphanumeric format. Robust and fast matching of entire pages is possible using indexes built with Minimal Absent Words (MAWs) based on the extracted strings, and more fine-grained retrieval (e.g. searching for all occurrences of a short theme or motif) can be achieved by standard approximate string-matching methods (e.g. ngrams). Indexes are stored and retrieved using a SOLR index accessed from a node.js server and exposed through a URL-based API. All code is public and open source, and since the web-based client GUI and the server interact via the API there is a great deal of flexibility for future custom development.

*Future direction and goals*

We anticipate significant expansion of our database as other leading music libraries are incorporated. Our use of state-of-the-art search technologies means that we can accommodate this without substantial effects on performance. As well as extending the list of libraries involved, we intend to improve our ingestion (OMR, encoding, indexing) process as we do so.

We are also exploring other indexing and matching strategies. Further work will assess the range of musical research questions that can be tackled using this technology. An enhancement will be the incorporation of metadata searching to support and refine our content-based methods. In the opposite direction, we intend to explore how a content-based system such as F-TEMPO can act as a very useful tool in cataloguing digital music resources, offering the means to automatically align multiple instances of very similar pages of music, distributed across multiple collections.

**Current state**

- *Hardware*
    - Virtual Machine, 4 CPU cores, 32GB RAM
- *Software*
    - Aruspix OMR library[1], Solr for search, custom search application (nodejs/express) and frontend (react)[2], docker for development and for production deployment
- *Storage*
    - 2TB
- *Data and content:*

The system currently works with around 400,000 pages from five libraries (British Library, London; Bavarian State Library, Munich; Berlin Staatsbibliothek; Bibliothèque nationale, Paris; Polish National Library, Warsaw). Although the basic input data is high-resolution graphics (usually scanned tiffs, but often derived from intermediate formats such as PDF), we do not in principle need to store these once the features needed for indexing have been extracted, but we use lower-resolution grayscale jpeg images for the purposes of result display and visual verification. However, the OMR system we use (*Aruspix*) saves about 1 MB of data per page in a compressed format, which we currently store in complete form.

The note-sequence features we extract are typically very small (typically 300 bytes per page), and from these we derive MAWs and ngrams of varying length. The indexes generated by SOLR occupy rather more space. Our total current storage requirement is about 7 GB for 500,000 pages. We estimate this is likely to double over the next two years.

- *Audiences, users, disciplines, subjects*

Although F-TEMPO was conceived as a tool for specialised discovery by musicologists, it has a wider range of possible uses. In particular, it has potential as a learning resource for music students who will be able to engage directly with the graphical content of sources used for modern editions; this will lead to a greater understanding of the way the printed music needs to be interpreted before it can be performed. As the resource draws nearer to complete coverage of the repertory, there will be the possibility of incremental nuanced analyses of various

[1] https://github.com/DDMAL/aruspix/
[2] https://github.com/TransformingMusicology/f-tempo

sub-repertories and their special natures (e.g. large-scale stylistic comparisons of sacred vs. secular music, or changes over time and place) which have hitherto not been possible. Via the API it should also be possible to provide content-based musical data, and statistics based on this, to external searches which could be used to support various kinds of musicological discourse. In this way, the resource could be of use to an indefinite range of humanities disciplines.

*Access methods:*

Online search via web GUI

Programmable search via API

*Place in research lifecycle:*

Working prototype using static resource

*Staffing:*

Technical support

Informal and unfunded at present

User support

Informal and unfunded at present

***Issues and problems***

Smooth incorporation of library holdings benefits from consistent metadata standards. Although this would most easily be taken directly from IIIF manifests, these lack sufficient bibliographic control. MARC records could be an acceptable substitute, provided these can be linked reliably to the images, and that they record musical works in sources at an appropriate granularity (the results of a search should be at the piece level, but this is not possible if piece metadata and individual images cannot be related).

IIIF permits control of the images themselves to remain with the publishing institutions, so search results can simply be links to the relevant library's holdings.

***Future requirements***

- Hardware
    - A long-term hosting solution
    - As the project scales up, indexes can be distributed across servers (for example, hosted by libraries), whilst images can remain on institutional IIIF servers. This means that the hardware requirements can remain small. Processing requirements are comparable with SOLR-based text searching elsewhere.
- Preservation
    - Hosting solution should include future-proofing of code, backup, etc.
- A longer-term funding model

- Promotion activities to reach our target audience

**References**


- F-TEMPO website (provisional): https://f-tempo.org
- Aruspix website: https://www.aruspix.net
- T. Crawford, G. Badkobeh and D. Lewis, ‘Searching Page-images of Early Music Scanned with OMR: A Scalable Solution Using Minimal Absent Words’, ISMIR 2018: https://archives.ismir.net/ismir2018/paper/000210.pdf

**Challenges of Digital Scholarly Editions**
**(Voltaire Foundation, University of Oxford)**

*Digital Voltaire* and *Digital d'Holbach*.

Contact: Professor Nicholas Cronk (Director), Dr Birgit Mikus (Administrator),
Voltaire Foundation, University of Oxford

**Description of resource**

- Its nature and purpose

The aim is to produce high-quality scholarly editions of 18th-century French writers in digital form that will have the same citation authority as the definitive print editions of the past, all to be grouped under the umbrella title *Digital Enlightenment*. This presents a number of challenges:

• In some cases, such as *Digital d'Holbach*, we are building a born-digital resource.

• In other cases, such as *Digital Voltaire*, we are building the digital resource from pre-existing print editions. But even in this case of such an archival edition, there will be a number of born-digital additions:

corrections and additions (that must be peer-reviewed and signed)

IIIF images of manuscripts, to be put with text (TEI data and metadata models)

Other IIIF images, notably book illustrations, to be put with text (TEI data and metadata models)

The addition of the catalogue of Voltaire's library, more complex because it needs to link to the library list of titles, to the text, to footnotes, and to the marginalia.

The addition in due course of a time-line biography of Voltaire that will link to the biography itself, to the correspondence, and to the relevant works.

The addition in due course of a peer-reviewed online journal that will publish research created with the help of the resource (and that will in its turn enhance it).

The addition in due course of an editing tool, that would allow, for example, a born-digital edition of a newly discovered letter to be seamlessly integrated into the larger resource, once it had been peer-reviewed.

etc.

- Future direction and goals

* We need to meld the born-digital and the archival editions, to ensure maximum cohesion and cross-searchability of the resource.

* We need to ensure that the born-digital additions to the archival edition are fully integrated and function well.

* We need to ensure a business model that will allow maximum diffusion while retaining maximum complexity of the resource.

**Current state**

- Hardware
    - Staff computers (Macs), VF internal network drive, cloud hosting
- Software
    - Currently used: Oxygen XML Editor for TEI files, freizo, Invenio, and Github for IIIF files and editing platform/basic viewer
- Storage
    - Files stored on staff computers, network drive, cloud hosting
- Data and content – type and size; rate of growth
    - TEI files, IIIF files; size approaching 500GB for both together; pilot IIIF files now completed, no further growth without more funding; TEI files growing at slow rate, ca 5GB a month
- Audience(s), users, discipline, subject
    - Currently staff and developers audience; funders; Humanities, 18th-century/French/History studies
- Access methods: discovery, analysis
    - To be developed
- Place in research lifecycle, e.g., collected, processed, analysed, shared
    - Both TEI and IIIF in collection and processing stages
- Staffing: number and type (e.g., RSEs)
    - Technical support
        - 2-3 (data engineer for TEI/edition data, 1-2 IIIF engineers)
    - User support
        - None (currently no external/non-VF staff users)

**Issues and problems**

- We had issues finding a data engineer to develop the data and metadata models when starting on the digitisation of the archival and creation of the born-digital material; we found an external partner on consultancy basis.
- Same issues with finding support for IIIF processing, storage and hosting; we found an external company and support for the pilot IIIF project.
- Not able to set up hosted/administered and maintained Linux VM for data experiments; internal services unable to provide administered Linux VM, external options too expensive

- Most issues encountered so far are on the most basic infrastructure and personnel basis, resulting in time- and funds-intensive searches for competency, hosting, and support of access interfaces (IIIF).

**Future requirements**

- Hardware
    - Mostly in form of server hosting and storage, sustainable long-term solutions; no specialist equipment
- Storage
    - Sustainable long-term storage and maintenance, repository, working environment, and interface, in TB region
- Software
    - Current editing software probably fine
- Access methods
    - Specifically tailored interface required for TEI editions; link up with IIIF viewer/TEI editing platform – RSE required
- Preservation
    - Long-term hosting required, plus RSE time for maintenance

## anəstor

Contact: Neil Jefferies, Bodleian Libraries and Data Futures GmbH,
Peter Cornwell, Data Futures GmbH and ENS-Lyon, France

Data Futures is a Leipzig-based not-for-profit that works on sustainable scholarly resources, and rescuing and redelivering legacy resources.

### Description

- Nature and purpose

  anəstor is a IIIF-based cloud service which can be licensed on an annual basis, supporting multiple projects, operated by Data Futures on behalf of client organizations, or integrated with existing infrastructure such as institutional repositories. It provides workflow management for both manual and automated creation, as well as ongoing enrichment of annotation collections relating to digitized heritage imagery as well as scientific literature. anəstor provides the following functions, which are missing in existing IIIF infrastructure, enabling complete end-to-end annotation tasks to be operated securely:

1. automated accession of image files to create an annotatable IIIF service where this is not already available
2. secure storage of annotation data for later creation of annotation collections or primary research records via serialization into potentially multiple output representations
3. transformation of existing annotations (often made using legacy tools and pre-IIIF image services) into specific OADM dialects, or WADM for preservation (dynamic serialization between OADM variants and WADM is also supported—for example to enable repositories storing WADM to present annotations via particular instruments such as specific Mirador3)
4. authentication of human contributors via ORCID to enable management of permissions granted to specific users; prevent unwanted modification of annotations, and time- and ID-stamp annotations for traceability
5. creation and operation of manual annotation work-packages: automated chunking of large annotation tasks so that contributors can accomplish annotation work within a limited period and commit it; potentially proceeding with multiple subtasks over an extended period or moving to another research activity—facilities for embedded task-specific training and reminder-style instructions are available, as are multiple-visit workflows in which annotation tasks are presented to multiple contributors for checking
6. connection of automated annotation infrastructures external to anəstor, so that training datasets, for example for neural-net instruments can be develop; subsequently automatically generated annotations can checked for exceptions and merged with manual correction workflows

- Technological approach

  anəstor is a Data Futures *freizo*-based application—using distributed MongoDB infrastructure to provide high-reliability annotation services to client user communities. Developed during 2016, and subsequently employed by a growing European and U.S.

community, anəstor employs *freizo's* infrastructure of multiple communicating database members to replicate research data on geographically-distributed computing instances. This strategy supports scalability but also imparts flexibility to employ multiple providers of leased computing and storage services as well as bare-metal IT equipment. anəstor has been used for annotation tasks on large sources of greater than 10k images with contributor communities of more than two hundred concurrent users over extended periods, and also with machine annotation agents producing hundreds of thousands of annotations.

- Future direction and goals

  While many annotation tasks can be addressed with stock workflow functionality, anəstor requires configuration for a particular task by Data Futures personnel. For example Mirador dialogs, which define annotation metadata, are necessarily tailored for training and use by human contributors and machine annotation requires training datasets and/or APIs in order to uses specific source imagery efficiently. Development goals for anəstor therefore include creation of tools to enable client organization administrator personnel to carry out such workflow tailoring, as well as contributor permission management and task operation, independent of Data Futures, for a growing class of annotation tasks.

  Another future direction is to integrate textual materials into image and annotation workflows, leveraging the standards developed by the AHRC/NEH-funded Unlocking Digital Texts project.

- Data and content – type and size; rate of growth

  typical client data size is between 200GB and 50TB per project (though one project has >1PB of moving imagery)

  typical number of annotations created per project is around 5000 though multiple projects have exceeded 800,000

  content is mix of large imagery and digitized scientific literature, publishing and numerical data, with average individual image file size of 15MB

  growth has been relatively linear over a decade, but rate of new clients has been restricted
- Audiences, users, disciplines, subjects

  primarily biodiversity, medical and social science and humanities; some heritage
- Access methods: discovery, analysis

  website, annotation collections exported as JSON, XML, XLS, CVS; annotated corpora comprising primary research records currently exported as InvenioRDM repositories or OCFL archive files

  annotated InvenioRDM corpus repositories support Elasticsearch and invenio-oaiserver for external discovery
- Place in research lifecycle

  approximately 60 anəstor client annotation projects are either: completed data resources; data resources supporting ongoing enrichment; annotation workflows in operation (prior to completing repository or project outcomes)
- Staffing: number and type
  - Engineering
    - three FTEs
  - Technical support
    - five FTEs - of which three provided by client institutions

- User support
    - none - provided by client  institutions

**Issues and problems**

**Future requirements**

- Hardware
  leased VM as well as leased bare metal and real Hewlett-Packard Enterprise
- Storage
  leased VM as well as leased block storage; long-term storage via CERN-Data Futures *hasdai* Partnership
- Software
  JavaScript, Python 3, MongoDB
- Access methods
  website, ORCID authentication
- Preservation
  export in proprietary and standards-based formats; InvenioRDM corpus repository as part of *hasdai* Partnership with CERN; OCFL archive file

**References**

https://ejt.biodiversity.hasdai.org/
https://avisblatt.dg-basel.hasdai.org/
https://voltaire-pilot.ox.hasdai.org/

## *freizo*

Contact: Peter Cornwell, Data Futures GmbH and ENS-Lyon, France,
Neil Jefferies, Bodleian Libraries and Data Futures GmbH

Data Futures is a Leipzig-based not-for-profit that works on sustainable scholarly resources, and rescuing and redelivering legacy resources.

**Description**

- Nature and purpose

1. Redelivery of existing research data—especially digitized print and image sources annotated using legacy technologies; making corpora portable so that they can be organized using standards-based data, search and citation models (WADM for annotation) and subsequently output to new research and repository platforms for more effective reuse and preservation
2. Replacement of existing workflow software by making research data accessible to current standards-based tools via APIs
3. Creation of new research data corpora and data resources through aggregation at scale from heterogeneous sources—documents, metadata, data files—generating services such as IIIF to support annotation with mainstream tools; and output, in particular of annotation to a range of research and repository platforms

- Technological approach

*freizo* is a MongoDB application first developed during 2010, which employs multiple communicating database members to replicate research data on geographically-distributed computing instances. This strategy supports scalability but also imparts flexibility to employ multiple providers of leased computing and storage services as well as bare-metal IT equipment, in order to reduce outage and supplier-related vulnerabilities. Necessity for this underlying approach was demonstrated graphically during one of the OVH catastrophes (10 March 2021) in which both virtual and bare metal equipment which operated a major *freizo* node in Strasbourg was destroyed by fire. Several hundred active client users of that particular *freizo* infrastructure were able to continue working while the fire was being extinguished, using other *freizo* nodes including Helsinki and Atlanta operated by other suppliers. *freizo* DB members are employed primarily for metadata (administrative and traceability, as well as client workflow metadata) with a range of storage services for bulk data, including cloud object storage offerings such as Amazon's S3.

Specific *freizo* support for mainstream IIIF annotation tools was developed during 2016 and subsequent collaboration with programs such as Transkribus has led to workflows creating annotations at scale using programmatic as well as machine-learning applications. Multiple projects based on *freizo* have each produced >800 thousand annotations. Anəstor, a *freizo* application launched during 2020, is a stand-alone annotation store for research communities which supports ORCID-based authentication to determine contributors' access rights to create/edit annotations. Anəstor can serialize between WADM and multiple OADM representations, maintain versions of annotations and timestamp and mark them using ORCID.

Data Futures GmbH, which operates *freizo*, originated the *hasdai* Partnership with CERN during 2018 to more fully develop long-term preservation of corpora processed using freizo. This partnership, described in attachments, has led to bulk accession tools for the Invenio platform

(CERN's repository architecture), enabling the automated generation of DataCite-oriented corpus repositories. In 2019 CERN and Data Futures launched the InvenioRDM consortium, gaining European and U.S. institutional partners which contributed funds and developers to create packed open source repository software. Data Futures is the community lead for IIIF and OCFL for the InvenioRDM consortium. Release v9.0 of InvenioRDM supports automated IIIF manifest generation upon publishing image and literature records; Data Futures biodiversity and humanities corpus repositories (see examples) present IIIF annotations, and Data Futures and Plazi are extending Zenodo to support display and community enrichment of annotations.

During December 2021 a CERN-Data Futures InvenioRDM sprint led to launch of the ocflcore.py library (https://pypi.org/project/ocflcore/) which enables *freizo* and InvenioRDM records to be aggregated automatically to produce OCFL archive files—corpus snap-shots which support versioning and eliminate external dependencies—making them effective for platform-agnostic and 'dark' archive use. Such OCFL archive files are intended for long-term preservation using off-line media, including LTO tape and fused silica, and also allow the future creation of new repositories and research workflows, including platforms undefined at the time of generation.

- Future direction and goals

Approximately 100 research activities and redeliveries of vulnerable legacy data resources currently use *freizo.* Increasing numbers of these are being output as Invenio Framework3 and InvenioRDM repositories when research projects terminate, so that *freizo* services can be shut down. Long-term support of such Invenio corpus repositories is managed under the *hasdai* Partnership agreement. Significant numbers of projects, which support long-term research, retain both *freizo* and Invenio instances. While heterogeneous leased data center resources remain an important component of Data Futures strategy,

Institutional Guarantee (IG)—in which participating organizations host equipment managed by the *hasdai* Partnership—is expected to lead to an increase in the proportion of real machinery operated.

The future direction of the freizo infrastructure has several trajectories:

- Fully integrating OCFL archive creation and maintenance functionality with *freizo* and InvenioRDM—making preservation and creation of long-term data resources from annotation into non-engineering functions that can be accomplished by system administrators
- Adding a new IG equipment layer, based on the *hasdai* Distributed Scientific Data Library, which is a network of LTO-based server nodes already having a pilot tier of early adopters, which incorporates robotic libraries and is hosted by participating organizations
- Extending anəstor functionality to provide annotation and enrichment of scientific literature as a service
- Integrate textual materials into image and annotation systems, leveraging the standards developed by the AHRC/NEH-funded Unlocking Digital Texts project.

- Hardware

  Heterogeneous leased virtual machinery and bare metal, plus increasing Hewlett-Packard Enterprise (HPE) DL380 servers and MSL family LTO libraries (located a various client sites)
- Data and content – type and size; rate of growth

  typical client data size is between 200GB and 50TB per project (though one project has >1PB of moving imagery)

content is mix of large imagery and digitized scientific literature, publishing and numerical data, with average individual image file size of 15MB
growth has been relatively linear over a decade, but rate of new clients has been restricted

- Audiences, users, disciplines, subjects
  primarily biodiversity, medical and social science and humanities; some heritage
- Access methods: discovery, analysis
  website, InvenioRDM corpus repositories support Elasticsearch and invenio-oaiserver for external discovery, InvenioAPI's (including repository API's such OAI-PMH and SWORDV3)
- Place in research lifecycle
  All stages, evenly distributed
- Staffing: number and type
  - Engineering
    - three FTEs
  - Technical support
    - five FTEs - of which three provided by client institutions
  - User support
    - none - provided by client institutions

**Issues and problems**

**Future requirements**

- Hardware
  leased VM as well as leased bare metal and real Hewlett-Packard Enterprise
- Storage
  leased VM as well as leased block storage; long-term storage via CERN-Data Futures *hasdai* Partnership
- Software
  JavaScript, Python 3, MongoDB
- Access methods
  website, ORCID authentication
- Preservation
  export in proprietary and standards-based formats; InvenioRDM corpus repository as part of *hasdai* Partnership with CERN; OCFL archive file

**References**

https://ejt.biodiversity.hasdai.org/
https://avisblatt.dg-basel.hasdai.org/
https://voltaire-pilot.ox.hasdai.org/

**Cambridge Digital Library – TEI metadata**

Huw Jones, Head of Digital Library, Cambridge University Library

**Description of resource**

Cambridge Digital Library https://cudl.lib.cam.ac.uk/ is the main platform for the surfacing of content-driven digital humanities research in Cambridge. It consists of c.500,000 digital images, c.37,000 descriptive records, many thousands of pages of transcription, and a range of other resources such as 3D, video, multispectral etc. Its main funding stream has been through partnerships with research, and the nature of both the digitised material and the accompanying data reflect this, with the metadata which underpins the items on the site containing what we would consider to be the outputs of research in addition to the kinds of information usually held by libraries about their collections.

TEI is the base metadata format behind all content held on the Digital Library. TEI was selected as the preferred format for its ability to hold both descriptive data and transcription, and to combine the outputs of research with more "formal" library and archive data. Its wide adoption in the digital humanities (and specifically by some of our major partner projects) was also an important factor. As a unified platform, the Digital Library is an aggregator of data from a multitude of sources (and often from multiple sources for the same record). Much of our metadata has been converted to TEI from other formats – text documents, spreadsheets, databases etc. – or extracted from library or archival systems. A common workflow is to derive source metadata from an existing system or resource into a simple TEI record which is then enriched with the outputs of research – though we increasingly deal with large amounts of born-digital TEI records, particularly from large partner projects such as Fihrist, the Newton Project and the Darwin Correspondence Project. An inevitable result of this aggregation is that the data been created in different contexts with different outcomes and research questions in. TEI records are often surfaced through project-specific platforms in addition to the Digital Library.

The purposes of this data can be broadly divided into publication and analysis. The Digital Library itself offers a deliberately simple view onto the data, with an image viewer, metadata viewer, and basic search and faceting. More complex modes of publication are often handled through project-specific interfaces. Analysis and reuse of data is supported through open licensing and API access to the source TEI data (along with IIIF access to images and other resources). We have had considerable success in promoting reuse of data in the digital humanities (for example, we have two live projects which reuse data generated by earlier projects, and three bids currently submitted which make significant use of datasets available through the Digital Library).

**Current state**

The TEI data is stored and managed in a private repository on BitBucket and surfaced through the Digital Library platform, which has now been made available as open source (technical details here: https://github.com/cambridge-collection/cambridge-collection.github.io). The data is also available through APIs which are part of the platform software, which allow access to full TEI

records and also HTML fragments of transcription (where it exists). The open source platform has been adopted by Manchester University https://www.digitalcollections.manchester.ac.uk/ and by Lancaster University https://digitalcollections.lancaster.ac.uk/ which effectively ensures a basic level of consistency across the data model used by the three institutions. The TEI also has a close relationship (with some minor differences owing to platform requirements) to that underlying the suite of TEI catalogues hosted and maintained by the Bodleian (e.g. Fihrist https://www.fihrist.org.uk/, Medieval Manuscripts in Oxford Libraries https://medieval.bodleian.ox.ac.uk/). Data consistency across these platforms is theoretically based on a shared TEI schema for manuscript description https://github.com/msDesc/consolidated-tei-schema though in practice the heterogenous nature of the TEI coming into the platforms, and the demands of publishing schedules means that some TEI data has diverged considerably from this model. The range of materials now being published (including modern material and objects) may mitigate against the use of a single shared schema for all TEI records in the future.

The primary generators of the data are researchers, who also form the primary audience, though we also get considerable usage in teaching, from news stories and social media, and from general public interest. The data is interesting in that it is a product of research one of whose main purposes is reuse in further research, with an emphasis on the application of digital humanities methodologies such as network analysis and natural language processing. One challenge inherent in this process is that the data inevitably alters considerably over the course of time which creates problems both for citation and for the relation of changes in the data to particular actions or methodologies. It is envisaged that a more sophisticated use of the versioning available through Git and methods internal to TEI such as revisionDesc would go some way towards addressing these problems – particularly for data enrichment or alteration as the result of automated processes.

While the Digital Library platform is developed and maintained by a team of developers at Cambridge (now with contributions from Manchester and Lancaster), overall responsibility for the creation, processing and maintenance of the TEI data, along with associated training and documentation, is handled by a minimal team of two staff, each of which have other responsibilities. In practice TEI creation is mainly handled by project teams, which are either trained and supported by Digital Library staff, or, in the case of large and long-running projects such as Fihrist, Darwin and Newton, trained and supported by the projects themselves. Data conversions from other formats and systems are almost exclusively handled by the core team, as are transformations to formats for publication and dissemination (JSON, IIIF manifests, HTML etc.). Pressure of work tends to mean that collaborative work to consolidate practice (and therefore technical approaches) comes as an afterthought rather than as part of a planned or structured programme of work, which means that many of the internal workflows and transforms have developed a rather ad hoc and responsive feel.

**Issues and problems**

Consistency of practice in TEI is a key factor both in data reuse and in promoting efficiencies around the development of platforms and tools for working with the data. As an aggregator of data from a number of projects and sources, the Digital Library acts as a microcosm of a

more general tension between data reflecting the specific needs and research questions of projects and the kind of data consistency which facilitates reuse and data-driven approaches. We have attempted to mitigate this problem through the adoption of a consolidated schema, but initial experience indicates that a single schema or practice for all projects and all types of material is not a realistic solution. A possible approach here would be a modular set of schemas where basic structures and elements are handled in the same way, with material or project specific sub-schemas to add or override certain aspects of the data. A schema or set of schemas which was closely tied to tools/platforms would certainly be very helpful – e.g. if your data validates against schema A then tools B, C and D and platforms E and G will work without the need for further adjustment.

A related and perhaps more serious problem is the consistent use of standard vocabularies and identifiers across the data. As has been pointed out, data crosswalks are far easier to handle than identity resolution. For certain core elements we promote the use of standard identifier schemes – e.g. VIAF for names, Getty for places, LCSH for subjects – but a number of projects deal with material where a large number of entities do not appear in standard schemes (names being a particular issue). Here projects tend to create their own authority files and identifier schemes which work well within particular datasets but do not allow for the kinds of interdisciplinary data work which might generate new research pathways. A particular problem is the lack of a standard identifier scheme for manuscript material, though there have been some initial efforts on this front through the International Standard for Manuscript Identification project. A similar problem exists for vocabularies and ontologies, especially for the descriptive elements of the TEI (materials, scripts, bindings, seals etc.). Here, vocabularies tend to be developed either on a project basis or within a subject area, leading to similar problems when working across datasets.

As discussed above, one major issue is the fluidity of data over time. This creates problems with citation, which might be partly resolved through a more structured and granular approach to versioning, but perhaps a more serious problem is the relation of changes in the data to particular actions or methodologies. This is a particular issue with automated approaches, where it is important to be able to relate a particular element of the data to the methodology which generated it, especially where there are chains of dependency in the data – e.g. method 1 says A is true, from which method 2 creates a hypothesis that B is true, from which method 3 says that C is true – which causes problems if method 1 revises its conclusions.

**Future requirements**

- A central (modular?) set of TEI schemas and documentation which are closely aligned with platforms/tools/methods
- A more organised and centralised approach to the creation/use/maintenance of authority schemes and vocabularies, including a more streamlined way for projects to contribute to existing schemes such as VIAF
- A structured approach to citation/versioning and to the granular recording of the actions and methods which result in changes to datasets.

- Potential for a centralised data repository (modelled for instance on the NLS' Data Foundry https://data.nls.uk/) to promote reuse of data across institutional boundaries

**PRiSM Sample RNN**

*Contact*: Christopher Melen <christopher.melen@rncm.ac.uk>

**Description of resource or infrastructure**

- *Its nature and purpose*

PRiSM SampleRNN is a computer-assisted compositional tool released on GitHub in June 2020. It generates new audio outputs by 'learning' the characteristics of an existing corpus of sound or music. Changing parameters of the algorithm and how the dataset is organised significantly changes the output, making these choices part of the creative process. The audio generated can be used directly in a composition or to inform notated work to be played by an instrumentalist. The software was developed by RSE Chris Melen, who has a dedicated multi-GPU hardware resource for development and to assist people in using the tool. Development of the software is funded by Research England, Expanding Excellence in England (E3) under the PRiSM award to RNCM.

- *Technological approach*

The original SampleRNN architecture was described in the paper SampleRNN: An Unconditional End-to-End Neural Audio Generation Model (arXiv:1612.07837 [cs.SD]). Available implementations used deprecated technologies so it was decided that PRiSM would offer its own implementation, based on Google's popular Machine Learning library TensorFlow 2 which is actively maintained. Although remote access to GPU clusters (such as JADE) is available, the code was developed and deployed on the same sort of hardware that the users in the community would be using, i.e. linux PCs with GPU cards. Clusters have also been used alongside this resource. The code is released on github and a colab notebook is available.

- *Future direction and goals*

  - The software has been used extensively and continues to be supported, with experience gained from its use feeding into evolving best practice.

  - Recent developments have improved efficiency and enable higher resolution audio.

  - SampleRNN was used in the event Future Music #3 (16-17 June 2021), which featured work by the Machine Learning for Music (ML4M) Working Group (a collaboration with University of Manchester), and will be used in forthcoming performances including Future Music #4.

**Current state**

- *Hardware*

PRiSM currently possesses two dedicated 3XS Workstations, built according to PRiSM's specification by SCAN UK. The machines have nearly identical specification, apart from their Graphics Processing Units, vital for any Machine Learning task. The first machine, obtained in early 2020, was equipped with two NVIDIA Titan RTX cards, each with 24GB of onboard RAM. The second, purchased around a year later, swapped these for a pair of GeForce RTX 3090 cards, utilising NVIDIA's next-gen Ampere GPU architecture, each again with 24GB of RAM. In our Machine Learning experiments we have found this second card to be around 50% faster than the Titan RTX.

The following table outlines the additional hardware specification for both machines:

| | |
|---|---|
| RAM | 128GB |
| Primary Storage | 1TB SSD |
| Secondary Storage | 4TB SATA3 HDD |
| CPU | Intel i9 (10 cores, 20 threads) |
| Cooling (GPU) | Liquid |

- *Software*

Both machines came with the Ubuntu 18.04 LTS operating system pre-installed, as per PRiSM's specification. Each also has the latest CUDA drivers and toolkits installed. Although we initially made use of the Anaconda data science platform, most of PRiSM's Machine Learning tasks are currently run using Docker containers based on NVIDIA's own Docker base image. This has the advantage of obviating the need for the incremental upgrade of the CUDA installation (which can often be problematic on Linux), since it has these baked in. Docker has also proved a very convenient platform for running Machine Learning tasks.

- *Storage*

Training data, models and generated outputs for PRiSM's Machine Learning projects are stored on the secondary storage (4TB HDD) attached to each machine, and backed up to one of a pair of Synology DS220j NAS devices (16GB RAID). For convenience data generated for specific clients will also be uploaded to PRiSM's own cloud storage, on a case-by-case basis.

- *Data and content – type and size; rate of growth*

Current data collections include datasets, models, and generated outputs (both test material and final outputs). For PRiSM SampleRNN datasets are invariably derived from an initial audio file or

set of such files, supplied by the client. This file is then split into numerous smaller chunks, of a uniform duration of about 6-8 seconds. A typical dataset will consist of around 2000-5000 such chunks, for optimal training. Smaller datasets may be enhanced by overlapping of chunks, thus increasing the number.

All PRiSM SampleRNN training currently requires mono input files in wav format, and generates mono wav output. For sake of speed much of the initial experiments and training for SampleRNN was done on audio with a sample rate of 16kHz, but recent optimisations have made it possible to work at much higher sample rates such a 44.1kHz, with marked improvement in the quality of the generated output.

- *Audiences, users, disciplines, subjects*

PRiSM SampleRNN currently has a small, but steadily increasing, community of active users centred around the project's GitHub repository. PRiSM is committed to maintaining this and all its software repositories into the foreseeable future and beyond.

- *Access methods: discovery, analysis*

SampleRNN is available on GitHub and as a colab notebook. Information is available by the RNCM PRiSM website https://www.rncm.ac.uk/research/research-centres-rncm/prism/prism-collaborations/prism-samplernn/

- *Place in research lifecycle, e.g., collected, processed, analysed, shared*

SampleRNN is used in practice-based research.

- *Staffing: number and type (e.g., RSEs)*

SampleRNN is supported by one RSE and is only part of their role.

**Issues and problems**

**Future requirements**

- *Hardware*

Local machines for development and test are essential and GPU upgrades would make sense to replicate what others are using. We are looking to make further use of GPUs provided by universities and as national facilities.

- *Storage*

Increased storage on the server may become necessary to accommodate future work.

- *Software*

The area of machine learning tools is notorious for code ceasing to run as libraries are further developed and deprecation occurs. We have done what we can to use supported platforms and to make the code available to be sustained by the community.

- *Access methods*

- *Preservation*

The PRiSM centre has a pro-active approach to creating a “PRiSM Archive”. We have also discussed with IP and copyright experts the status of future models trained on source materials with rights restrictions – this is an unresolved issue in the community at this time.

**High Speed 2 (HS2) Historic Environment Digital Archive**

Kieron Niven, Archaeology Data Service

**Description**

The construction of the new HighSpeed2 (HS2) rail network will create "the largest historic environment digital archive ever compiled in the UK". The programme of work is being undertaken in three phases (1, 2a, 2b) with a predicted completion date of 2045. The project has, and will continue to generate, large amounts of born-digital data at a number of levels: advance survey work such as geophysical and structural surveys; standard excavation and evaluation datasets; 3D survey data from laser scanning, lidar, and photogrammetry; programme-level GIS; and specialist and programme-wide synthesised reports.

Much of this is born-digital by definition of the survey equipment used, which logs the data directly in digital format. But digital methods of site recording have been increasingly adopted by 21st century archaeologists, with traditional paper-based pro forma replaced by the use of digital tools. This change has been accelerated by the scale of the HS2 project, and the importance of avoiding delays to the construction. Thus, over 40,000 human burials recorded in advance of the construction of the new HS2 station at Euston were logged on hand-held tablet computers, using digital pro forma. Such excavation archives provide the primary record of the archaeology which has been destroyed during the development process, and whereas paper records could be boxed up and consulted in museums decades, if not centuries, after the excavation, the digital record is much more fragile and requires active curation. However, it also offers an opportunity for widespread Open Access and online dissemination for researchers, and the wider general public.

The Archaeology Data Service (ADS) has entered into an agreement with HS2 Ltd which will ensure it can safeguard "their digital legacy for future generations and deliver HS2 Ltd's commitment to making data easily available under the Government's Transparency agenda". The long term commitment from HS2 also allows ADS to implement infrastructure changes essential to the management of the digital archive.

At the time of publication, over ninety HS2 digital archives have been deposited with ADS. Although generated by various specialist sub-contractors, the majority of these archives fit within a 'standard structure' of archaeological archives as outlined in the Historic Environment Research and Delivery Strategy (HS2 Ltd 2017) and more specifically in the accompanying Historic Environment Digital Data Management and Archiving Technical Standard (2020). However, as components of a larger ongoing programme, additional complexity is introduced

into the HS2 digital archive through scale. In one aspect scale refers to the sheer number of HS2 digital archives being deposited with ADS (i.e. multiple datasets from different contractors, potentially from the same site and at different stages of investigation) and the problems in ensuring that datasets are clearly related to one another. Secondly, 'scale' can also be seen in regard to the physical extent of individual sites and the corresponding size of datasets generated, particularly where data-intensive surveys such as laser scanning, lidar, or photogrammetry have been undertaken.

Overall, the technological infrastructure at ADS underlying the HS2 digital archive will stay consistent with that supporting other fieldwork archives with HS2 datasets fitting within an expanded version of the existing ADS storage and archiving infrastructure. On a wider level however, the HS2 digital archive has required significant preparation of specific guidance and the development of bespoke workflows for HS2 Ltd and their contractors. These developments ensure that digital data are captured and documented at the point of creation and that deposited datasets can be ingested, processed, and disseminated within a timely manner. The agreement between HS2 Ltd and ADS also ensures that ADS can support and continue to develop this infrastructure for the duration of the HS2 project, whilst the data will be preserved into perpetuity. At the technical end, in order to address issues of 'joined up' datasets and scale, HS2 digital archive interfaces will see a greater use of mapping interfaces, both within archives and at various overview levels, along with more widespread use of appropriate preview mechanisms for large datasets such as 3D data.

It is highly likely, considering the timeframe of the HS2 project, that both the archaeological methods generating data, and the archival and technological approaches adopted by the ADS, will continue to develop and evolve. The long-term agreement with HS2 Ltd will allow ADS to continue to develop existing processes for ingesting, preserving, and disseminating datasets within a core framework. This underlying consistency between HS2 digital archives and other ADS collections ensures that any developments are mutually beneficial. Key areas for development will likely focus on increased access to data through more efficient search mechanisms and text mining, preview mechanisms, and through specialist services for data types such as spatial data.

**Current state**

- Hardware
    - University-managed systems consisting of 32 virtual machines hosting 35 servers.
- Software
    - Bespoke ADS Collections Management System and related tools to manage ingest and archiving process.

    - ORACLE databases.
    - DROID (file identification, checksums).
    - Lucee (web development).
    - Leaflet (web maps).
    - 3DHOP (web 3D previews).
- Storage
    - University-managed Network File System (90TB), backed up for 3 months (instantly retrievable), and on tape drive longer term.
    - 15Tb storage allocated for HS2 within the first five years of the project, to be reassessed at the end of this period.
    - AWS S3 and Glacier storage (preservation copies)
- Data and content – type and size; rate of growth
    - Wide range of raw and processed data from an array of techniques; tabular data; spatial data; images; publications.
    - There are currently over 300 archaeological interventions expected from phases one and two of the HS2 project. These will range in size and may generate multiple datasets.
- Audience(s), users, discipline, subject
    - General Public
    - Academic Researchers
    - Students (from primary school level to PhD)
    - Commercial Archaeologists
    - Local and National Heritage Bodies
    - Community Groups
- Access methods: discovery, analysis
    - Access is primarily through the web interface to individual collections although no HS2 archives are currently released. Data is primarily made available to download although certain data types (images, 3D content) can be previewed without download.
- Place in research lifecycle, e.g., collected, processed, analysed, shared
    - Data exists at all stages, from raw collected datasets through to processed, analysed data, to synthesis and summary reports.

**Issues and problems**

- Current issues mainly focus on ensuring data is deposited in a timely, consistent, and complete manner in line with specifications, and that datasets are sufficiently documented and related to one another where relevant.

- Issues with scale at dataset level have predominantly focussed on dissemination mechanisms for large data sets, specifically those resulting from laser scanning or photogrammetric survey (i.e. single multi-terabyte datasets in which the size presents barriers to both access and reuse).
- Variation/consistency in deposited data over the long term may present issues e.g. where raw data exists for some of the other datasets but only synthesised data (summary reports) for others.
- Working within larger (external) project workflows, business structures, and systems can be potentially problematic, for example where data flows differ based on data type. Within the HS2 project certain data types such as GIS and geophysical survey data go from data creator to HS2 Ltd for sign-off before coming to ADS. In other cases, such as with digital reports, these first go into OASIS (an online reporting system for archaeological investigations) from which they are archived directly by ADS. All other data types are deposited directly with ADS by the archaeological contractors.
- Successful integration of data into singular search interfaces for a programme-wide front end.

**Future requirements**

- Hardware
    - More dedicated systems and processing power for various aspects of the repository, particularly focussed on dealing with large data: ingest (upload and processing); storage (see below); archiving (data/file characterisation, assessment, normalisation); dissemination and access.
- Storage
    - Storage requirements for HS2 archive beyond the initial 15Tb allocation to be reassessed at the end of the first five year period.
- Software
    - Solutions for allowing easy access to large datasets e.g. streaming or multi-resolution solutions to access 3D data and point clouds.
    - Possible automated online solutions for data deposit and ingest.
- Access methods
    - As above, better options for users to assess and access large datasets such as point clouds and 3D models.
    - Potential access to geospatial datasets via services such as WMS and WFS.
    - Integration of frameworks such as IIIF to allow more standardised reuse of data.
- Preservation
    - As above, increased access to high-performance systems for assessing and processing large datasets.

**References**

High Speed Two (HS2) Limited, 2017. Historic Environment Research and Delivery Strategy. Phase One. https://www.gov.uk/government/publications/hs2-phase-one-historic-environment-research-and-delivery-strategy

High Speed Two (HS2) Limited, 2020. Technical Standard - Historic Environment Digital Data Management and Archiving. Version 4.

## White Rose Etheses Online: theses as complex digital objects

Authors: Rachel Proudfoot, Nicola Barnet, Masud Khokar, John Salter, University of Leeds, UK

### Description

White Rose Etheses Online (WREO) is a joint repository infrastructure that holds research theses awarded by the Universities of Leeds, Sheffield and York. WREO is based on the EPrints platform and has been hosted at the University of Leeds since 2008.

Theses submissions with complex multiple parts is not a new concept. Many physical theses had complex components in the form of 'back of the thesis' content such as CDs, memory sticks, 3D objects such as textile samples, and additional elements bound into theses such as maps and diagrams. The development of electronic versions of theses and the move towards e-only submission have opened new possibilities and challenges.

While most theses still consist of a single PDF file, an increasing number of them have additional examined or supplementary components. This growth of complex, interconnected theses objects is related to our support of practice researchers who are permitted to submit a portfolio of practice material to support their work, and researchers in other fields producing non-textual materials central to their research. Capturing practice research can be challenging, in terms of definition, selection of materials, formats and finding suitable infrastructure to host outputs [1]. Investigation at University of Leeds found interest in submitting more complex theses across a variety of academic fields. Files may be uploaded to WREO, but they link to associated materials often stored in external repositories systems (such as the Research Data Leeds repository) and to externally hosted websites. WREO also holds theses digitized from print by University of Leeds or the British Library, and digital materials from physical media submitted with paper theses (CDs, USBs, hard-copy supplementary material).

In addition to our own analysis and research, the British Library has also previously highlighted the possibilities of more complex theses in a series of case studies [2]. However, recent discussion with the British Library's EThOS (Electronic Theses Online Service) suggests most research theses harvested into ETHOS continue to comprise a single PDF. This case study illustrates the digital complexity of theses at one institution, including the growth of and increasing appetite for theses as complex digital objects. The identified issues and possibilities will be evident in many other institutions, including specialist arts institutions. There is huge potential to improve the advice and infrastructure supporting complex digital theses.

## Current state

- Platform: VMs: 1 x Application (RHEL, 8CPU); 2 x Database (RHEL, 8CPU) and network storage for files.
- EPrints software: EPrints 3.3.16 (Gelato Blizzard) => ulcc-core 1.2 (Year of the Whopper)
- EPrints plugins: Multiple including REF, coversheet generator, arkivum, IRUS, RIOXX, etc.
- Data and content: 20,200+ thesis records (7,000 for Leeds). Growth is approximately 1,500 theses per year across the three institutions. Most theses are uploaded as a single PDF file, but associated files are often uploaded as part of the formally examined content, or as supplementary material. EPrints accepts any file type: over 80 formats are recorded in WREO including zip files which could contain any file formats (see Appendix A for all formats).
- Network storage: approx. 500GB
- Audience / users: the material is available for use by anyone, including other researchers and the general public. Etheses are some of the most frequently downloaded materials in institutional repositories. Metadata is also available for harvesting by third party services, such as the DART Europe Ethesis Portal. Many Universities in the UK partner with the British Library EThOS service which harvests both metadata and associated files for research theses.
- Access methods: most material is openly available. Associated files are exposed for full text indexing by Google and metadata exposed via OAI-PMH. Some file types rely on the end user having an appropriate viewer or software to open them.
- Place in research lifecycle: Theses are uploaded after examination, viva and corrections. However, it is vital that thesis content and structure for multi-part / complex theses is considered much earlier in the research process (see Issues section)
- Staffing number and type (University of Leeds): 0.3 FTE repository management and administration. 0.2 FTE repository development and technical support

## Issues and problems

Thesis infrastructure is optimized for single file upload, assuming a ‘textual’ thesis as this is what most theses, historically have been. Most repository platforms are designed for multiple, non-connected files to be uploaded.

The options for structuring and submitting multi-file theses and theses with non-textual elements are not always clear. The thesis structure and choice of materials is sometimes interwoven with the academic practice itself.

Diverse digital formats already in evidence and likely to proliferate, with the associated risks for digital preservation and of obsolescence.

Training and awareness is needed for supervisors, PGRs, professional services staff involved in thesis advice and administration and for library staff involved in advice, training, upload and thesis management.

Where researchers are choosing from a complex portfolio of materials, it can be challenging to assess what to keep, where, in what form and for how long. Views vary on the extent to which indiscriminate ‘data dumping’ is, or could be, a problem where PGRs are invited to deposit a body of work as part of, or

in addition to, the examined thesis. A key requirement is to make sure the theses remain ‘examinable’ - so that the structure is clear and volume not overwhelming for an examiner.

Work submitted for examination is not always the same as work for public dissemination: rights issues can lead to redacted theses or ambiguities about what may, or may not, be made available and what license(s) should be applied. Thus, a coherent examined work can become more fragmented when it enters the public sphere. This problem further extends with multipart, complex digital thesis.

Component parts of the examined work or supplementary material may be available outside the WREO infrastructure, with links to external content provided (often not using PIDs and pointing to unstable platforms). Parts of the ‘work’ are at risk of disappearing or becoming detached from the other materials that form the thesis.

There is a significant challenge in accessibility of content, particularly in creating accessible original copies and enhancing digitised content for web accessibility and general accessibility standards.

## Future requirements

### Infrastructure

- A cloud-based, scalable, repository infrastructure that handles interconnected, complex, multi-part and multi-media materials by design.
- Uses PIDs for the component parts of a complex thesis and a PID for the overarching work i.e., the work submitted for the award of the research degree.
- Infrastructure that is capable of generating Analysis-Ready Data for Machine Learning.
- Interoperable with digital preservation infrastructure that supports multi-format, multi-component digital etheses spread across multiple platforms and sites.
- Human infrastructure that understands practice-based research and how it is stored in digital repositories.

### Content

- Accessible by design to as many people as possible (e.g., people with disabilities).
- Accessible and preservable content, supporting the diverse, complex, inter-connected digital formats across same and multiple platforms.
- Easily understandable in its entirety, and easily linked individually.

### Operations and Support

- A well understood way to identify component parts of a thesis and how they relate to each other. It should be clear which parts were directly examined and which form part of the wider portfolio of work.
- Templates are used for theses but tend to be designed for textual forms. Not all researchers are keen on templates, finding them constraining – particularly in creative areas. Nonetheless, there is a potential role for accessibility and preservation ready structures for presenting material in different ways: for example, for presenting research as a website.

- A research showcase built gradually over the course of the degree so that issues of structure, format, selection of materials, licensing and rights can be tackled on an ongoing basis, supporting researchers to create a ‘showcase’ in addition to the examined thesis.
- Support for supervisors and PGRs to consider the structure and forms of complex theses, including file formats and any rights considerations and accessibility.
- Preview options and examples for PGRs to engage with what the research will / could look like well in advance of public release.

## References


1. Bulley, James and Şahin, Özden. Practice Research - Report 1: What is practice research? and Report 2: How can practice research be shared?. London: PRAG-UK, 2021. https://doi.org/10.23636/1347.
2. British Library. Case studies on Non-text doctoral theses. 2016. https://www.bl.uk/case-studies?service=ethos%20and%20theses

## Appendix A - List of File Formats in WREO (April 2021)

| Format | Leeds totals | Sheffield and York | WREO totals | % format in Leeds |
|---|---|---|---|---|
| pdf | 6569 | 13056 | 19625 | 33% |
| png | 823 | 2417 | 3240 | 25% |
| zip | 49 | 206 | 255 | 19% |
| wav | 43 | 149 | 192 | 22% |
| avi | 34 | 76 | 110 | 31% |
| mp4 | 30 | 122 | 152 | 20% |
| xlsx | 19 | 68 | 87 | 22% |
| cif | 10 | 10 | 20 | 50% |
| docx | 9 | 845 | 854 | 1% |
| mp3 | 6 | 426 | 432 | 1% |
| xls | 6 | 46 | 52 | 12% |
| ppt | 5 | 9 | 14 | 36% |
| xlsm | 5 | 2 | 7 | 71% |
| mov | 4 | 68 | 72 | 6% |
| gif | 3 | 4 | 7 | 43% |
| 7z | 3 | 0 | 3 | 100% |
| pptx | 2 | 1 | 3 | 67% |
| eps | 2 | 0 | 2 | 100% |
| flv | 2 | 0 | 2 | 100% |
| xml | 2 | 0 | 2 | 100% |
| doc | 1 | 344 | 345 | 0% |
| txt | 1 | 71 | 72 | 1% |
| wmv | 1 | 13 | 14 | 7% |
| bak | 1 | 0 | 1 | 100% |
| mdl | 1 | 0 | 1 | 100% |
| jpg | 0 | 195 | 195 | 0% |
| cda | 0 | 33 | 33 | 0% |
| jpeg | 0 | 27 | 27 | 0% |
| m4v | 0 | 27 | 27 | 0% |
| mpg | 0 | 25 | 25 | 0% |
| mtp | 0 | 17 | 17 | 0% |
| rtf | 0 | 14 | 14 | 0% |
| wma | 0 | 14 | 14 | 0% |
| aif | 0 | 11 | 11 | 0% |
| m3u | 0 | 11 | 11 | 0% |
| dbf | 0 | 10 | 10 | 0% |
| fam | 0 | 10 | 10 | 0% |
| vob | 0 | 10 | 10 | 0% |
| rar | 0 | 9 | 9 | 0% |
| tv | 0 | 9 | 9 | 0% |
| m | 0 | 8 | 8 | 0% |
| tiff | 0 | 8 | 8 | 0% |

| | | | | |
|---|---|---|---|---|
| db | 0 | 7 | 7 | 0% |
| mdb | 0 | 7 | 7 | 0% |
| px | 0 | 7 | 7 | 0% |
| val | 0 | 7 | 7 | 0% |
| flac | 0 | 6 | 6 | 0% |
| bup | 0 | 5 | 5 | 0% |
| html | 0 | 5 | 5 | 0% |
| ifo | 0 | 5 | 5 | 0% |
| bmp | 0 | 4 | 4 | 0% |
| dna | 0 | 4 | 4 | 0% |
| emf | 0 | 4 | 4 | 0% |
| m4a | 0 | 4 | 4 | 0% |
| nvp | 0 | 4 | 4 | 0% |
| tif | 0 | 4 | 4 | 0% |
| xps | 0 | 4 | 4 | 0% |
| csv | 0 | 3 | 3 | 0% |
| exe | 0 | 3 | 3 | 0% |
| fsl | 0 | 3 | 3 | 0% |
| ini | 0 | 3 | 3 | 0% |
| java | 0 | 3 | 3 | 0% |
| sbn | 0 | 3 | 3 | 0% |
| sbx | 0 | 3 | 3 | 0% |
| shp | 0 | 3 | 3 | 0% |
| shx | 0 | 3 | 3 | 0% |
| wr1 | 0 | 3 | 3 | 0% |
| aiff | 0 | 2 | 2 | 0% |
| au | 0 | 2 | 2 | 0% |
| gb | 0 | 2 | 2 | 0% |
| mpeg | 0 | 2 | 2 | 0% |
| ppsx | 0 | 2 | 2 | 0% |
| accdb | 0 | 1 | 1 | 0% |
| asf | 0 | 1 | 1 | 0% |
| css | 0 | 1 | 1 | 0% |
| gz | 0 | 1 | 1 | 0% |
| inf | 0 | 1 | 1 | 0% |
| odt | 0 | 1 | 1 | 0% |
| pps | 0 | 1 | 1 | 0% |
| qbe | 0 | 1 | 1 | 0% |
| skp | 0 | 1 | 1 | 0% |
| wdp | 0 | 1 | 1 | 0% |
| wri | 0 | 1 | 1 | 0% |

## Open Geospatial Data Application and Services viewer (OGDAS)

Gethin Rees, Lead curator, digital mapping, The British Library

**Description of resource or infrastructure**

● Its nature and purpose

This report describes a web maps interface and viewer that would make open geospatial data collected by the Legal Deposit Libraries (LDL) openly available on the UK web. Our collection of geospatial data collected under legal deposit has been made available to readers via the GDAS viewer on-site since 2017. This user-friendly web-map platform is hosted and maintained by Idox providing functionality that facilitates visual examination and querying of the collection. The viewer provides access to the entire collection, both proprietary and open geospatial data collected under legal deposit and in most LDLs access is restricted to a single reading-room terminal.

● Technological approach

The proposed OGDAS resource would make those datasets that have been and will be collected under open licenses freely available to the UK public for remote, off-site access. The Geospatial Commission of the Cabinet Office are working to make more geospatial data available under open licenses to boost economy and society. However, there are significant barriers to discovering historical geospatial data and for their use by non-experts. OGDAS would offer benefits to the many members of the public who want to access geospatial data but do not have the motivation and skills to download and use open data in GIS software. In offering simple functionality focused on visual use, the OGDAS could appeal to a broad range of users albeit without the benefits of more complex analysis. The resource would open our collection to a far wider potential audience than readers visiting LDL Maps Reading Rooms. Furthermore, open data becomes more valuable over time, particularly if the collection of data at risk of loss is prioritised. Providing remote access to such data 5-10 years after collection can offer significant value to readers as the data may be otherwise unavailable. There are three key areas of activity required:

- Understand how we can retain open licenses for geospatial data collected under non-print legal deposit.
- Develop an infrastructure for the rights management of datasets.
- Develop a viewer for remote use, beginning by evaluating the suitability of the current GDAS viewer from both technical and user perspectives.

Each area could act as a pilot or learning exercise for a wider approach to the LDLs making open data available remotely.

● Future direction and goals

The development of the viewer could lead in two future directions:

The first is the ability to capture, archive and make available in the GDAS/OGDAS geospatial data from interactive maps on the UK web. The volume of web maps data that is available on the UK web is very large and in many cases there is a substantial risk of loss. Significant scoping work is required

in this area, first to determine whether this falls within legal-deposit legislation and second to scope and develop a bespoke tool that would work on a variety of websites.

A second future direction would be to provide analytical access to historical open data. Analysis performed through queries, statistical analysis and algorithms is the key use case in the UK economy and society. Access facilitated by a user interface, download or application programming interface (API) access could offer benefits to both research and business.

**Current state**

The OGDAS viewer could follow the same technical approach as the GDAS.

● Hardware

The GDAS viewer is made available on Legal Deposit Library terminals and served using Amazon Web Services. The collection is stored in a PostgreSQL/PostGIS database on Amazon Web Services and in the Minimum Preservation Tool on British Library servers.

● Software

The GDAS viewer is based on technologies that will be familiar to many if not most readers, namely, a web maps interface including 'slippy' maps with functionality akin to popular platforms such as Google, Bing or Apple maps. This interface offers benefits in terms of familiarity and ease-of-use enabling a broad range of users to access the collection for visual use. The GDAS viewer is written using an open-source stack of PostgreSQL/PostGIS, GeoServer and OpenLayers. Back-end code is written in Python, the front-end is written in the Angular.js Javascript framework and the platform is hosted on Amazon Web Services.

● Storage

Dissemination Information packages of open data are currently stored in Idox's PostgreSQL/PostGIS and GeoServer databases. Submission Information packages of open data are stored in the British Library's Minimum Preservation Tool.

● Data and content – type and size; rate of growth

As of March 2021, the collection of proprietary and open data consisted of 3.2 tb, 1087 datasets in the GDAS viewer (DIPs), 1500 SIPs consisting of 5,675,874 files in 9,099 folders. Of these 1087 datasets, 419 were collected under open licenses. The collection as a whole is growing at a rate of several hundred gb per year. It is likely that open data will make up a significant proportion of data collected each year.

● Audiences, users, disciplines, subjects; Access methods: discovery, analysis

Geospatial data has been collected under non-print legal deposit regulations since 2016. Collection is a collaborative exercise across all legal deposit libraries and managed by Legal Deposit Map Librarian's group (LDMLG). Readers use the GDAS viewer on-site for higher education study, recent local history and personal historical interest. These are generally specific, closed-ended tasks that the readers had in mind prior to travelling to the library. Users of the viewer include semi-experts, hobbyists and experienced lay users. The overwhelming focus of GDAS use is the Ordnance Survey Great Britain large-scale vector datasets, Mastermap and Landline.

Providing off-site access to open data would open up our collections beyond this core group of users and datasets to a much wider audience including casual users with an interest in learning more about

geospatial data but without the time to travel to the library, and professionals or experts. The latter might use visual functionality to discover data for formal research in environmental science, geography, civic planning or quantitative social sciences. Furthermore, there are opportunities to engage private sector users and support business.

● Place in research lifecycle

Although many data portals exist they rarely visualise geospatial data or account for historical datasets. The OGDAS might act as a discovery portal to facilitate visual comparisons of older data.

● Staffing: number and type

The GDAS viewer has been developed and maintained by a third-party vendor, Idox. We could follow a similar approach for the OGDAS, under which Idox would provide technical support. If research demonstrated that a different viewer was required, we could look at alternative models including in-house development. In terms of user support, the OGDAS would require documentation. However, the familiarity of our potential audience with web map interfaces means that user support might otherwise be minimal.

**Issues and problems**

The proposed OGDAS offers few potential problems from a technical perspective as the content is currently being stored, preserved and delivered for on-site use in the GDAS.

On the other hand, licensing presents two main challenges, the first being the preservation of open licenses for content collected under legal deposit. The second is ongoing rights management within a viewer that could potentially deliver content for both on-site and off-site access.

**Future requirements**

● Hardware

● Storage

● Software

● Preservation

Future requirements for the OGDAS could largely follow developments for the GDAS viewer in terms of hardware, storage, software and preservation.

● Access methods

However, delivering the analytical functionality that a subset of expert users might demand, encompassing complex querying, statistics or algorithms such as those available in GIS software, would be more intensive requiring bespoke software or a collaboration with a GIS software company.

**References**

Please see the process described in the journal article ‘Map Collecting in the Digital Age’ by Phil Hatfield and Chris Fleet (Fleet & Hatfield, 2017). Accessible here: https://journals.sagepub.com/doi/abs/10.1177/0955749017730712

# Environmental monitoring system and its integration with digital twins

Pedro Maximo Rocha, The National Archives, UK
Amy Sampson, The National Archives, UK

## Description of infrastructure

The preservation of, and access to, the physical collections stored by The National Archives (TNA) is central to our Public Task. We execute this responsibility through the monitoring and management of the document storage space environment, ensuring temperature and humidity are line with BS EN 4971:2017. As sector lead, TNA has a responsibility to demonstrate best practice in this area, engaging with up-to-date tools and technologies. In the Kew repositories, environmental conditions are monitored by a network of Hanwell sensors. This system is at the end of its lifecycle. Additionally, the platform used for this is outdated, restraining workflows and lacking interoperability.

TNA needs to establish an up-to-date approach to ensuring that the environment in the repositories is effectively monitored. The application of digital systems for management of heritage buildings and collections is already in use elsewhere: Building Information Modelling (BIM) is being deployed in the restoration of The Houses of Parliament; the Natural History Museum is producing a Digital Twin model of its building to create a cross-departmental system for data collection, storage and visualization; and the Museum of London is deploying Internet of Things (IoT) sensors to monitor environmental conditions, with an integrated system that allows critical analysis of performance.

The aim for TNA now is to identify and implement a new environmental monitoring system, including both sensors and a digital management platform that provides real-time data, along with the ability to access and use data more effectively. This would establish a system that could, in the future, be integrated with a digital twin model, creating an infrastructure of sensors which can evolve to a common digital management platform for multiple resources and services, and have applicability to a range of departments across the organisation. To initiate this stage would require software, hardware and storage capacity. An integrated digital management tool will improve critical cross-department communication, collaboration and management, and allow remote assessment of the environment on-site - crucial when staff cannot be on-site, as demonstrated during the recent Covid-19 lockdowns.

## Current state

- **Hardware**
    - 191 x ML4106 T/RH Transmitters
    - 2 x ML4106 lux transmitters
    - 5 x SR2 receivers
    - 7 x Pro repeaters

- **Software**
    - Data Capture and export
        - Hanwell EMS dashboard - Real-time data is accessed via a web-based browser (EMS), which provides a range of functions, including:
            - Real-time data graphing, with the ability to view up to one year's worth of data
            - Data overlay between different sensors
            - Graph notation
            - Alarm parameters
            - Pre-programmed reporting capability
            - Calibration
    - Data Processing
        - Microsoft Excel
        - Pre-programmed reporting function with a variety of report types available

- **Storage**
  All EMS data (including reports etc…) resides in Kew on a virtual server. This is managed by the IT Infrastructure team

- **Data and content**
    - Rate of growth: Each sensor captures 2 data points for T and RH every 5 minutes

- **Audiences, users, disciplines, subjects**
    - This system is used for internal monitoring in collection care department (CCD)
    - Reports generated by this system are used for communication with other departments on TNA

- **Staffing: number and type (e.g., RSEs)**
    - Technical support: 1 x dedicated member of Infrastructure team, who has developed specialist knowledge/relationships
    - Rolling Bronze-level annual service and support contract with Hanwell, commissioned and managed by IT Infrastructure

## Issues and problems

The current system for environmental monitoring does not comply with FAIR principles of data management. Tools used to display and export data from EMS are difficult to access and reviewing data can be resource-intensive. The software used is proprietary and is also outdated given current trends around computational capabilities. The current system can export data into Excel format, but it is not readily interoperable with other data processing tools or programming languages. As such, no automation of data processing outside of what EMS offers is currently possible.

## Future requirements

- **Hardware**
    - 200 LoRaWAN T/RH/VOC/lux transmitters
    - 1 MQTT broker

- **Software**
    - Data Capture, export, automated reporting features
        - IoT sensors dashboard - Real-time data to be easily accessed via a platform, providing a range of functions, including
            - Real-time data graphing, with the ability to visualise data from all the sensors from all monitoring life-cycle in customizable mode
            - Data overlay between different sensors
            - Graph notation
            - Alarm parameters
            - Customizable reporting capability
            - Calibration
            - Ability of 3D visualization of sensor locations and data using digital twinning concept
            - Ability of data prediction based on mathematical algorithms

    - Data Processing
        - Microsoft Excel
        - Python

- **Storage**
    - Data Storage on a virtual server managed by the IT Infrastructure team

- **Access methods**
    - Users would be able to run algorithms over data and get reports
    - Enhancement of visualization tools

- **Preservation**
    - Data and report storage in an organized and coherent manner
    - Reproducibility of data analyses

## References

Desogus, G., Quaquero, E., Rubiu, G., Gatto, G., & Perra, C. (2021). Bim and iot sensors integration: A framework for consumption and indoor conditions data monitoring of existing buildings. *Sustainability (Switzerland)*, *13*(8). https://doi.org/10.3390/su13084496

Eini, R., Linkous, L., Zohrabi, N., & Abdelwahed, S. (2019). *A Testbed for a Smart Building: Design and Implementation*. 1–6. http://arxiv.org/abs/1902.10268

## Literary and Linguistic Data Service (LLDS) [formerly OTA]

*Contact*: Martin Wynne, University of Oxford martin.wynne@ling-phil.ox.ac.uk

**Description**

- *Its nature and purpose*

LLDS is a repository service hosting the Oxford Text Archive collections and engaging in the ongoing collection of data outputs from research projects in literary and linguistic disciplines in the UK and internationally, funded by UKRI. Datasets are curated for long-term preservation and made available for re-use by researchers. LLDS represents both an ongoing development of the long-standing work of the Oxford Text Archive, and a node in a new national digital curation service. LLDS is the national repository for the CLARIN European Research Infrastructure Consortium in the UK.

- *Technological approach*

The repository makes use of the CLARIN DSpace turnkey repository solution. As well as providing the key repository functions, it offers secure cross-border authentication and authorization, enabling researchers in many countries worldwide to log in with their own institutional logins, supported by Shibboleth software, and making use of cross-border domains of trust established by the CLARIN Federation and eduGAIN. Persistent identifiers are provided by the handle system. Metadata records are harvested, aggregated and promoted by the CLARIN Virtual Language Observatory.

- *Future direction and goals*

Current projects are building new capacity and enhanced functionality in a number of areas, including:

  - **beyond text**: enabling the repository to take deposits of audio and video data and make these outputs available for re-use by other researchers;
  - **beyond download**: establishing additional functionality so that users can not only download resources, but also search them, cross-search collections, and cite, re-use and combine data in innovative ways;
  - **lower barriers to data deposits**: a simple and user-friendly submissions workflow for new datasets, backed by a helpdesk;
  - **enhanced reliability and availability**: establishment of more robust procedures for management, monitoring, failsafes, replication and disaster recovery, to be assured via certification with CoreTrustSeal and CLARIN.

**Current state**

- *Hardware*

Virtual server with 8Gb RAM, 1TB additional high-speed storage.

- *Software*

Ubuntu 20.04 LTS operating system, with CLARIN DSpace (a version of Dspace 5.0 modified and enhanced by the CLARIN community)

- *Storage*

The datasets in the repository are also being backed up on the Digital Preservation service (DigiSafe), the University of Oxford's cloud-based digital storage facility for research data. Large audio and video datasets are delivered from the downloads.ox.ac.uk

- *Data and content – type and size; rate of growth*

Datasets include digital editions of historical printed texts, language corpora, lexical resources and audio and video data recorded and collected for research purposes. The current data collections include 68,000 catalogue items, comprising c.600,000 files and using 146GB of storage. The rate of growth is hard to estimate as the new collections development policy in collaboration with the AHRC is yet to be initiated, but likely to include at least 50 items (c.100Gb) per annum.

- *Audiences, users, disciplines, subjects*

LLDS caters for literary and linguistic subject areas, broadly defined, which can in practice include any discipline which is making use of text either as the object of study or medium. There is a core of users in disciplines such as Linguistics (including Computational Linguistics), Literary Studies, Modern Languages and in text- and language-focussed areas of History, Theology, Area Studies, etc. Users are worldwide, and are primarily researchers in HEI but also include students, professionals, activists and the general public.

- *Access methods: discovery, analysis*

Download of datasets is currently the only functionality offered to access resources. Discovery is aided by metadata records being available via the CLARIN Virtual Language Observatory.

- *Place in research lifecycle*
- *Staffing: number and type*

Staffing levels are variable, depending on short-term project funding, although there are hopes for more long-term strategic planning, staffing and funding. The current levels are approximately as follows:

    - Management: c.0.5 FTE
    - Technical support: c. 0.5 FTE
    - User support: no dedicated effort, currently provided ad hoc by staff in other roles

**Issues and problems**

A current project is assessing the adequacy of the current use of deposit and user licences, and making recommendations for future good practice in this area.

**Future requirements**

- *Hardware*

Increased RAM will be necessary to deal with audio, video and multimodal resources. A development server will be required for building the new version of the repository based on DSpace 7.0, and a backup server, duplicating the functionality of the current live production server is required in order to assure high levels of availability for the online services.

- *Storage*

Increased storage on the server will become necessary to accommodate future new deposits. Connection with analysis software is likely to require duplication of datasets, possibly in numerous formats, to accommodate the requirements of the software, and also to ensure that dissemination and preservation copies of data are kept secure.

- *Software*

CLARIN DSpace will be updated to the DSpace 7.0 codebase in the next 12 months. A roadmap is being prepared in collaboration with the CLARIN community and consultants in Slovakia. We are also investigating ways of connecting data to analysis software tools, via the CLARIN Language Resources Switchboard and other methods.

- *Access methods*

It would be desirable to allow more users from more countries to be able to use secure authentication. This will involve extending the domain of trust. It is also necessary for UK IdPs to routinely trust CLARIN SPs and to release required attributes to trusted centres.

It is also planned to offer ways to connect data to APIs in order to explore, analyze and transform the data.

- *Preservation*

Increased storage in DigiSafe will become necessary to accommodate future new deposits. A plan is needed for updating and format conversion for various datatypes where necessary to ensure that datasets remain usable and viable.