Consolidating research data management infrastructure: towards sustainable digital scholarship

Abstract

The sustainability of digital research outputs, particularly in the Humanities where these frequently comprise archives of digital cultural heritage material, has always offered a challenge to the researchers and institutions who have responsibility for them. The amount of upfront care, effort and funding that goes into developing a research project during the active (and funded) research phase is rarely replicated within the post-project maintenance and curation of the delivered digital assets or archives. What often defines the sustainability of a research project and its archive is a combination of research method and expected life span for the digital collection. Innovation in research data design is often at the expense of its longevity. But this doesn't need to be so. The trade-off between longevity and functionality is a false dichotomy. Yet what is clear is that care and consideration in planning the research data storage or archive for a project can make a big difference. A data management plan that meets grant funder requirements is asked for many research projects, but is more than simply a funding document. Good research data management and ensures outputs are available online for years to come, and available for future research and innovation. This paper offers a practical insight to the methods being employed at the University of Oxford to support Digital Humanities scholars (and beyond) safeguard their digital legacy for future generations.

Authors

Megan Gooch https://orcid.org/0000-0002-3190-0509

Damon Strange https://orcid.org/0000-0002-5851-718X

Keywords

1. Introduction

1.1. Broader Context and the Growth of Digital Scholarship

The growth of digital humanities over the last 45 years has to some extent mirrored the increasing digitisation and digitalisation of society. Increasing numbers of researchers are using digital tools and methods to collect, analyse, interrogate, and understand humanistic problems, and this has benefitted the study of humanities.[1] A common form of digital humanities project is one in which the researcher or research team collects images, data and metadata about cultural heritage collections, such as books, manuscripts, archaeological finds, buildings or other heritage resource. Unlike those working within cultural heritage institutions who are bound to study and curate their own resources, these digital humanities researchers are well-positioned to study object (or other heritage asset) types beyond one institution, and many aim to create a comprehensive collection of all known examples of a particular type of object, text or building. Without access to the collections management or library systems used by cultural heritage institutions, researchers have often resorted to creating their own systems and infrastructures for data collection, analysis, publication and dissemination. These collections of data are often called archives by their creators or users, even

---

[1] (Su and Zhang 2021)

if trained archivists would not necessarily label them as such. The features they share with institutional archives are the collection and collation of material and metadata, but they often lack the specialist information or records management training, metadata skills and preservations services found in institutional archives. Digital humanities projects often have databases and web front ends for public access to their data, which work for collaboration and reuse of their data, but are troublesome to maintain in the quickly changing digital world.

Some major digital humanities projects have been around for decades, and concern over the sustainability of the research data and outputs has been growing at the same time.[2] As many of the pioneering digital humanities researchers now reach retirement, they no longer have access to the funding nor the institutional support for their sometimes decades-old projects. In some universities libraries have been the institutions to offer homes for completed projects or their data, in others, data is stored by departments or IT services.[3] Despite these best efforts, data horror stories abound of outdated websites, entire databases only accessible on outdated software or on a single laptop, or researchers constantly seeking more funding to maintain their life's work, which is often a valuable community research resource. Behind these stories is a common problem: the loss of high-quality research data.

Recently, discussions at national or international levels have taken place to examine the huge risk of loss of digital research data without investment in technical infrastructures that go beyond individual research projects.[4] Initiatives driven by the research community such as FAIR and CARE data, which started in the sciences and have been adopted by many humanists, have shown a drive to make data more open and useful.[5] At the same time open science or open data criteria driven by governmental or funding body policies have pushed researchers to consider their research data more explicitly. The need for research data management plans, often supported by library functions in universities, is something any researcher in the UK will now be aware of, due to funder mandates, in contrast to the situation 10 years ago when 'technical plans' in humanities projects were optional.[6] At funder and governmental level too, discussions and projects to secure digital infrastructures are ongoing. In digital humanities in the UK, the Towards a National Collection funding scheme of £18.9 million was launched in 2019, and more recent AHRC calls have targeted finding answers to the research data sustainability problem.[7]

It is within this context then, of the growth of digital humanities projects and the risks of loss of research data due to researcher retirement or technological obsolescence combined with the push and pull of data management compliance and the open and FAIR/CARE data movements that institutions are developing a variety of ways to sustain these data, and funders and governments are prioritising wider infrastructures to support this work.

1.2. The Increasing Need for Digital Sustainability at the Oxford University

The explosion of growth in the variety and scale of digital research projects over recent decades at the University of Oxford, and no doubt across other research-intensive universities has increased the risk of data loss or 'near-misses'. The authors work in a research support function in the Humanities Division, one of four academic Divisions at Oxford. Our work involves frequent discussions with

---

[2] (Liu, McKay, and Buchanan 2021)

[3] (Åhlfeldt and Johnsson 2015)

[4] (McGillivary et al. 2020)

[5] (Wilkinson et al. 2016); (Whyte et al. 2019); (Harrower et al. 2020)

[6] ('Towards a National Collection | Collections United' n.d.)

[7] (Chang 2021)

digital humanists, and we work to ensure data is not lost, and to assist researchers with data at risk. It is due to these discussions that a new service was developed and launched in 2021 with the explicit aim of supporting digital humanities research projects and their research teams. The Sustainable Digital Scholarship (SDS) service provides both a data repository and a consultancy service around research data management.

To estimate the scale of the data loss or 'near-miss' problem we looked at the SDS consultation records and discovered that between February 2021 – February 2023, we consulted with 124 Oxford-led research projects and that of these, 17 (13.7%) had either suffered data loss from an existing system or suffered a significant element of technical degradation from current hosting and solutions. To put this in a wider context, there were 21,679 new projects between 2012-2022 which received funding at the University of Oxford across the University, with expected variations in the spread of these projects across the Divisions and departments of the University, as well as year by year.[8] We do not know the true scale of data loss in arts and humanities, but recent work by the Software Sustainability Institute commissioned by the UK's Arts and Humanities Research Council has highlighted that 75% humanities research projects used software in their research and that only 12% researchers were depositing data in University or national data repositories, with low levels of sharing data.[9]

The anecdotal evidence at Oxford was strong enough for the Humanities Division to recognise that the research and cultural data needs of their researchers were not being met, and we can see that this is not a problem that affects one institution. Within this paper we chart the development of this data repository service which arose out of researcher demand for a system that could meet the needs of open and sustainable digital data whilst retaining the flexibility of the (often old) systems they had used to create their research data. We will address some of the technical problems we have encountered as part of this project, but also the human problems of the need for improved research data management processes and better education for researcher which are key to the survival rates of digital research outputs and archives.

2. The plight of relational databases in digital humanities

2.1. What is a Relational Database, and Why do Researchers Love them?

Simply put, "a relational database is a database whose logical structure is made up of nothing but a collection of relations".[10] The use of relational databases for storing, analysing, and interrogating information from many subject areas within the digital humanities and beyond is a very mature and well-established concept. The advantages to using relational databases are the ease of access to cheap (or free) database software such as MS Excel, MS Access or Filemaker Pro,[11] the quick learning curve for researchers to build a database, and the powerful results that can be generated by querying the complex relationships within such a system. It is therefore unsurprising that researchers have for decades placed their trust in databases; they offer a quick and often cheap technical solution, are easily adaptable, and they can deliver powerful results.

---

[8] (Research Services n.d.)

[9] (David de Roure, n.d.)

[10] (Harrington 2016), p.89

[11] Other database software packages are available, but these are the most common we encounter in our work with researchers. Whilst not a relational database package, we often find many researchers using Excel in database-like ways, or referring to their spreadsheets as 'the database'.

Indeed, a well-planned and developed database coupled with a web front-end is a fantastic way to curate and present data from a range of academic disciplines to a variety of audiences. We find researchers can create databases in which the data items interact with each other in fairly complex ways, and that they want their data to be searchable on a faceted and granular level. At the University of Oxford, we are home to a vast array of such technical accolades, and we even host a database of such databases in the field of Digital Humanities where many examples of the projects we discuss in this paper can be found.[12] Through our work with multiple researchers and many research datasets we see the commonalities in the research questions, tools and approaches, as well as in some of the technology deployed to collect, store, analyse and publish these data. However, individual researchers do not see so many projects in their daily work and quite often cut their own technical delivery path, utilising varied technology stacks and different developers who will influence database design from their own experiences. Partially this is as a result of digital humanities projects developed before sustainable solutions at a departmental, institutional or national level were available. The early digital humanities adopters who innovated with their technological solutions as well as their research questions bear the burden of sustaining their own bespoke systems. For more recent projects, it is often a lack of awareness of the options and technical solutions available, combined with senior digital humanists advocating for the bespoke path that had worked for them.

Researchers who have specialised in humanities are increasingly likely to have encountered some form of digital training during their education, and this technical expertise is often a factor in how a database is developed and whether external technical support is imported to supplement the researcher's domain expertise. At Oxford, support is available through a number of routes including from departmental or central IT Services staff, or this support may be through an external consultant or developer. There is a strong tendency for new solutions to mirror previous achievements by all involved. A researcher will value a technical solution that has worked for them before, or that they have used extensively, and IT staff or contractors will often specialise in one form of technical solution such as the relational database.

Reuse of common development technologies and techniques is in many ways an extremely sensible and cost-effective means to undertake research, especially when these systems use of open-source code to mitigate the risks of proprietary software access or ongoing software subscription costs. However, the pace of technological change means that technological inertia is a risk many researchers cannot afford to take. The pace of change is one of many risks to relational databases as a repository or archive solution for digital humanities projects, but there are others.

Relational databases rely on complex connections between data records and fields, which can create technical challenges both for their researcher owners but also for public users on web interfaces. In addition, we frequently see databases deployed as a technical solution for small or simple datasets, where a spreadsheet or simpler technical solution would be a more appropriate and sustainable solution. It is all very well to innovate in digital archives, but sometimes simplicity is the solution.

The larger or more complex a database becomes, the harder it is to maintain and sustain. This is especially prevalent with custom-built databases and web interfaces for research projects. Whilst a project is in its active, funded phase, hiring staff and having a team responsible for performing maintenance, upgrades and updates to a database is easily done. However, once the project comes to an end, it becomes harder to pay for necessary technical support. If researchers are lucky, their institution or department may be able to continue to support their data resource or archive, but more often than not, unfunded archives are at risk archives. The danger here is partly aesthetic, web

---

[12] ('People & Projects' n.d.)

standards and designs evolve rapidly and many digital humanities online databases begin to look dated quickly. But there are also serious consequences to the lack of ongoing technical support including lapsed security certificates, not meeting equality legislation, and degradation of the software and functionality.

Thus far we have discussed maintaining digital humanities databases as the minimum needed to keep valuable cultural heritage research data open and accessible. In practice, standing still is not an option, as researchers will want or need to take advantage of new functionality that emerges in digital technologies. Once again, projects developed or hosted outside of an institutional system are unlikely to receive updates or further investment unless tied to new research funding proposals. We have found that good will from local departmental IT officers, other contacts or a particularly tech-savvy researcher is the way many digital humanities projects and their archives remain viable and accessible, and that large amounts of staff time, including researchers, administrators and IT officers is spent maintaining projects rather than on core activities or new research projects.

Innovations in cultural heritage and humanities research data are not always necessary to deliver open and accessible archives. The relational database is an easy yet powerful tool for the collation, analysis and dissemination of digital humanities projects, but the flexibility and complexity of some databases is not always needed, and the more flexible or complex a database becomes, the harder it becomes to sustain. This sounds very bleak for the digital humanist researcher, so what innovations will be useful to sustain these archives?

2.2. Understanding Oxford's Digital Scholarship Landscape

At the University of Oxford, the heightened risk level for digital scholarship projects is compounded due to the scale of digital humanities research undertaken at the university and the length of time this research had been going on. Oxford was an early adopter of humanities computing which became known as digital humanities.[13]

Many of these early digital humanities outputs, within Oxford and more broadly in the field, have sadly since become condemned to a life as an unreadable file format or media type, archived on internal servers when life on the open web would be more appropriate. Some have even been lost forever. This issue has been at the forefront of recent funder policies and funding initiatives and have been highlighted by a recent report commissioned by the AHRC.[14] Identifying the root cause of why projects go offline will often be a result of several factors which are people, process or technology related. We've already discussed one factor in sustainability challenges with respect to the relational database. [15]

The European Research Council recommends against a bespoke data storage or archival approach, noting that "in contrast to public data repositories, these are generally not deposition databases, and as long as they depend on a single individual and/or funding source, long-term sustainability is challenging."[16] Nevertheless, in the absence of institutional repositories capable of supporting data collections that need to be accessed at a record (rather than file) level, that may be updated as new finds or interpretations emerge, or that have become community research resources, individual technical solutions have proliferated in the digital humanities as cost effective and easily constructed mini repositories for the data of individual research projects. But what is cost effective in the short

---

[13] (McKnight, Prag, and Madsen 2014)
[14] (Taylor et al. 2022)
[15] (Taylor et al. 2022)
[16] (ERC 2022)

term for a research project does not always equate to value for money institutionally when they have to be maintained for the long term.

At Oxford, to understand how widespread this problem of bespoke online databases was, a review of digital projects was undertaken by the Humanities Division. The Humanities Division comprises 14 academic faculties and units. The review identified 47 projects which were suitable as ‘pilots’ for a data repository that could deal with these types of semi-active, record level data.[17] We call these data warm data to contrast with the hot data of active day to day research and collaboration, and the cold data of an institutional archive. The review was not comprehensive, and its results can only be taken as reflecting a portion of the digital humanities projects requiring digital sustainability support. Within the scope of the review were important research projects that had produced collections of warm data but which were currently outside their grant-funding envelope, and which faced challenges to their digital sustainability for reasons as diverse as technical obsolescence (hardware or software), lack of continuity in project staff, Principal Investigators (PIs) retiring, or leaving the University, or with security risks within their project websites/databases due to changes in web standards and legislation such as GDPR and the Equalities Act which added additional obligations concerning personal data and digital access.

However, the review of bespoke databases demonstrated the troubling scale of the problem of unsustainable digital humanities projects and this, combined with the need to comply with funder requirements, a desire to make data open and FAIR, and the sudden digital shift of the pandemic highlighting the value of these online (or offline) caches of high-quality research data at risk. A solution was in progress at Oxford in the form of the Digital Humanities Sustainability (DHS) project, which sought a way to utilise the collective knowledge of many projects and collaborators and attempt to find a way to standardise the non-standard.

2.3. If we’re not building a database for each research project, then what are we doing?

With the database-as-archive such a strong methodology and expectation for digital humanities research at Oxford and elsewhere, a solution that bypasses the relational database is a radical and innovative suggestion, even if the technology is not at the cutting edge of AI and technological invention. The relational database is so strong a paradigm that many researchers will refer to all data or datasets as a ‘database’. There will always be a place for databases, and the ‘death of the custom-built database’ may be overstated, but as technological innovations develop, there is a need for researchers to move away from their reliance on one imperfect solution, and often a reliance on a single technology or IT officer. Today’s researchers, if they want to create the data and archives of the future simply cannot afford to create projects with a single point of failure. But what is the alternative?

The Sustainable Digital Scholarship (SDS) service was launched in February 2021 as both a support service and a technical solution which can aid digital research scholars at Oxford with existing or new digital humanities projects.[18] It is important to highlight the dual function of the SDS service is both a digital platform for research data as well as a consulting service for researchers’ queries about research data. Researchers who contact the service may be referred to colleagues managing other research data repositories within the University or external subject-specific repositories, and this collaboration with colleagues is an essential part of the service. Researchers may also be guided towards the SDS platform run on Figshare, an open access Software as a Service (SaaS) platform.[19]

---

[17] (Strange, Roissetter, Hurst and Madsen 2019)

[18] (‘Welcome to the Sustainable Digital Scholarship (SDS) Service’ n.d.)

[19] (‘University of Oxford Research Repository - Search’ n.d.)

This is because we understand that there is no one solution for research collection, storage and dissemination. We even still refer researchers to colleagues who can assist in creating bespoke relational databases where this is the appropriate solution to enable the researcher to answer their research questions.

The SDS digital platform is an instance of 'Figshare for Institutions'. The platform ensures digital humanities data are hosted in an open by default manner thereby creating inbuilt open access and FAIR data to research projects. Data is 'as open as possible, as closed as necessary' and can be restricted if there are requirements such as embargoes, personal data or copyright restrictions. An early project for the service, Around 1968: Activism, Networks, Trajectories tested our ability as a service and as a platform to deal with restricted data options due to the nature of the recent oral history data.[20]

The benefits of using this SaaS approach are that openness is the default option, and projects will meet funder criteria and FAIR principles as a result of adding project data to the SDS platform. By contrast, researchers who have sought to build their own technical solution will often face several challenges in building, adding and maintaining their research data in addition to making it open and FAIR. The option to build bespoke systems remains open to researchers at the University of Oxford, as some research questions will always require specialist and innovative software solutions. However, we have found that for many researchers utilising a commercially built and maintained platform will provide the support they need for their research data.

The Sustainable Digital Scholarship service is part of a range of research data management solutions and repositories within Oxford, some of which are bespoke builds and others usuing commercially available software,[21] but SDS is the only one that serves the needs of researchers with warm data projects. Digital sustainability is a large and growing issue for many institutions, with challenges being not just technical but also related to the cultural adoption of research data management by researchers.[22] Our project and research has shown that despite a wide variety of research questions, subjects, methodologies and approaches, the introduction of a shared and common infrastructure for warm data can deliver digital sustainability for researchers and their publics.

We fully believe that our approach to is fully complementary of the sentiment that "nascent and ever evolving, digital humanities needs open and inclusive platforms."[23] At the University of Oxford our novel use of an open access repository to build functionality that replicates some of the functionality of a relational database, whilst offering a lot of other benefits too – Digital Object Identifiers, usage metrics, longevity and guarantees of institutional support.

3. From bespoke builds to buying software solutions

3.1. Embracing Software as a Service (SaaS), Commercial off-the-shelf (COTS) and Cloud Computing

Oxford is not alone in moving from bespoke built solutions to using software developed and provided by third party providers for some of their systems. The vast majority of corporate systems universities and other large companies depend on are typically not on-premises, self-built

---

[20] ('Around 1968: Activism, Networks, Trajectories' n.d.)

[21] ('Research Data Oxford' n.d.)

[22] (Taylor et al. 2022)

[23] (van Zundert 2012)

technology stacks. The ease in which an organisation can mitigate various risks, reduce costs and often have a better overall product is to outsource the responsibility of building and maintaining it to another company who are experts in doing so. The open access and FAIR agenda often applied to software code, and open source software is one way researchers can mitigate the risks of system obsolescence: by having code and documentation that are intelligible to a range of software developers. Another way to build sustainability is by using software as a service.

The journey towards a more standardised approach to offering sustainability to digital humanities scholars at Oxford has taken some time. The foundations were laid by Professor Jonathan Prag (Principal Investigator), assisted by Christine Madsen (Project Manager) and Janet McKnight (Research Associate) when they delivered DHARMa (Digital Humanities Archives for Research Materials), a one-year project (August 2013 to August 2014). The findings within DHARMa provided the catalyst and 'researcher-led' evidence base and a strong recommendation that continuing with historical practices for sustaining digital humanities research needed to be reviewed.[24] With these recommendations at its core, the 'Digital Humanities Sustainability (DHS) Project' was approved and commenced in 2017, which led to the delivery of the Sustainable Digital Scholarship service launch in 2021.

During this journey to find the right technical solution for digital humanities research (and beyond), building a bespoke system was an option along with SaaS. Extensive researcher community consultations were held to gather feedback on system and feature requirements, with more than 400 requirements identified in total. These requirements were refined by the DHS Project team, until 109 remained. Requirements were then broken down by functional categories such as persistent identifiers, metadata creation, metadata enrichment, workflow, API interactions, search, browse, and find, as well as ingest and egress. The bulk of these functional categories were derived from specific examples and consultation with researchers responsible for delivering and managing digital humanities research projects. Figure. 2 provides a full overview of these category areas, presented within a Reference Architecture for a future DHS Solution, highlighting primary requirements in scope as part of initial implementation and requirements that may be deemed desirable in the future.

---

[24] (McKnight, Prag, and Madsen 2014)

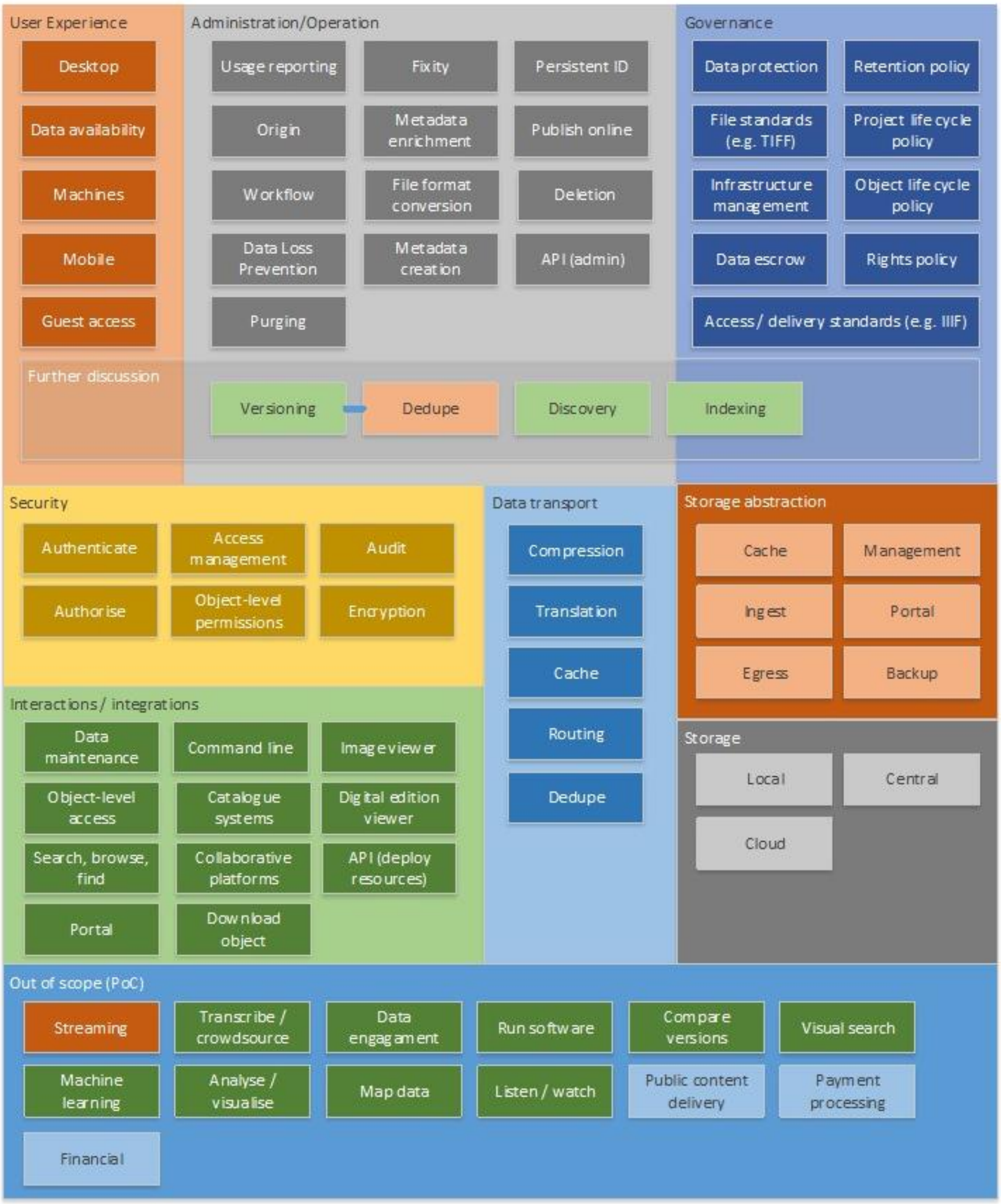


Figure 1: The DHS Reference Architecture which was developed to group features and requirements by key categories © University of Oxford.

Once we had a detailed understanding the project requirements, we carefully considered the options to self-build or to commission a technical system and solution. A key factor was an ambition that individual researchers, or those within a research project should be able to manage their own warm research data with minimal intervention from support staff beyond initial migration or on-boarding. As discussed above, we knew that researchers valued the freedom to manage their own data, as many had done with relational databases, and that technical reliance on one individual was a weakness of some bespoke systems. The ability to self-manage was considered important for the quality of the research data, the user satisfaction of the researchers, and also as a more sustainable service model that did not rely on a large team of support staff.

After reviewing the cost and risks of a bespoke build for the sustainability solution, the project team sought to understand what existing options could be used to meet the needs of this project. The well-honed requirements catalogue was deployed within a formal Invitation to Tender (ITT) process which sought bids from any system and solution providers including Figshare, who could potentially present a SaaS offer to meet our requirements.

Following the decision to proceed with Figshare, a swift implementation took place in parallel with work in the project team to design service management processes. Figshare allowed us to buy an off the shelf product, but also to adapt the system to our local users' needs. This allowed us to launch the Sustainable Digital Scholarship (SDS) service in February 2021 as both a support service and a technical platform which can aid digital research scholars at Oxford with existing or new digital humanities projects.[25]

3.2. Automating where possible, to get data online as quick as possible (and in a logical and useful manner!)

Automation of mundane processes with as little human interaction as possible is the great hope for AI in many realms of research, not least in research data management. Access has moved from consulting physical library books to access via digital resources in many research libraries, so much so that a flourishing sub-genre of digital libraries research from the later 90s to mid-00s no longer exists.[26] All research libraries are now at least partly digital libraries. And it is in this digital and digitalised context that single digital repositories or archives exist, and yet are often not indexed, catalogued or finable via the research libraries in the institutions that produced this valuable research. The theory of accessible and open online content, of digital access overcoming geographical, accessibility or temporal barriers is only useful if that content can be found.

The SDS service, like other repositories in Oxford, offers a findable home for research projects at Oxford which have been painstakingly curated and maintained. The process for moving these projects is part human and part machine. Each project migration begins with a consultation with a team member and a researcher, contains a number of approval points to check their project structure is as they wish and a final review to ensure they are happy with their new dataset.

One benefit in using Figshare is that it has a flexible, modern infrastructure that has a well-documented and featured API. This means that we are able to choose from a range of options to suit the various scenarios and requirements we have faced with project migrations to the platform. In the two years since the SDS service launched (February 2021 – February 2023) we have successfully created or migrated over 600,000 records to the platform. The two primary methods of migration involve deploying Figshare's Batch Migration Tool,[27] or utilising an open-source tool developed by David Banks, at the University of Sussex.[28] These tools have made it feasible to move research data onto the platform in bulk, creating hundreds of records in minutes with the simple upload of a spreadsheet of carefully cleaned metadata. The manual creation and curation of records has often been the alternative for many other digital humanities systems our researchers have used, and the speed in which SDS can import and visualise data online is an advantage for our users.

---

[25] ('Welcome to the Sustainable Digital Scholarship (SDS) Service' n.d.)

[26] For example (Arms 2001), (Chowdhury and Chowdhury 2003), (Fuhr et al. 2007), (Lesk 2005)

[27] ('New Administrative Batch Management in Figshare' n.d.)

[28] (Banks [2017] 2022)

As well as batch migration, we also trialled using an FTP Uploader with Figshare.[29] We have found this particularly effective and efficient for the migration of larger datasets onto the platform (e.g., 100GBs of data at a time). With the right skillset, and time to undertake the work, It is also possible to develop your own migration pipelines; we have seen one research group successfully utilise the Figshare API to manage the ingest and egress of their data independently of the platform's standard interface.

3.3. Efficiencies of scale

As a university service which has been carefully planned and delivered, the SDS team have been keen to develop and refine processes and workflows to ensure support to researchers is offered in an efficient manner. Business and customer management processes and workflows exist to manage interactions with researchers, however, as part of this paper we are focusing on how at the University of Oxford we have adopted processes to ensure we can move research project data onto a repository at scale and at pace.

The mobilisation time, from initial contact between research and the 'technical team' (SDS in our case) is far quicker than in more traditional research project and technical supplier relationships. The lead-in time and the amount of detail a developer needs to capture for a custom-built system or database would usually be much more involved than the information capture and planning process the SDS service needs to enact.

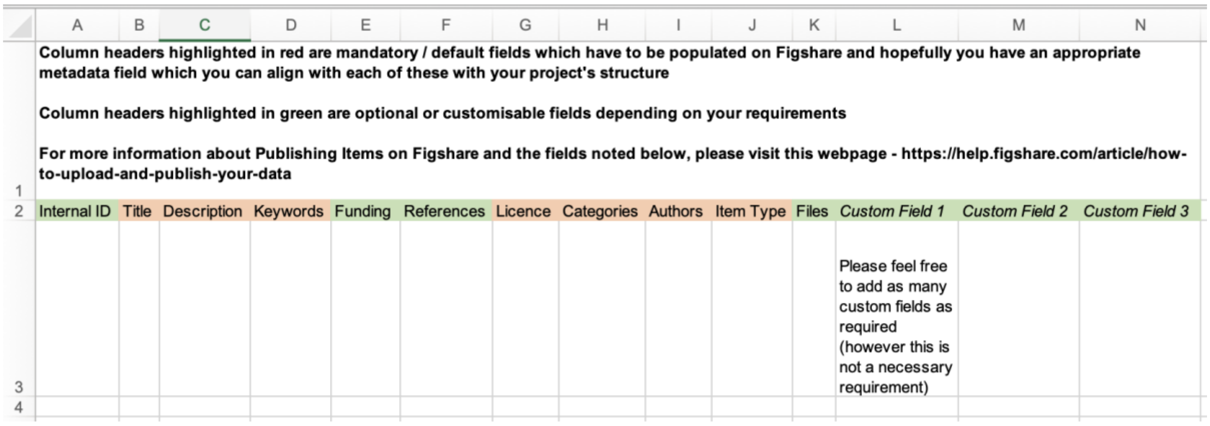
A typical engagement process with a researcher and their project and SDS when planning data migration of an existing collection of data, or for a new corpus they are in process of building would in most cases comprise:

1. Initial Engagement & Scoping Meeting – During this session the researchers provide some background and general context of their research, and what they are planning to do with their research data. The SDS team then share information on how this type of data could potentially be hosted, arranged, tagged, and made discoverable on the SDS platform. At this point agreeing custom metadata fields and tagging conventions for records would start to be defined. Also, the SDS team may share a metadata capture spreadsheet (see Figure 2), where researchers and their team collaborate to populate this as part of a planned data migration.
2. Metadata Checking and Cleansing, File Acquisition & Testing – The next step of the process is for the SDS team to evaluate the metadata and prepare it for the migration. As part of this phase the researcher supplies data assets, ranging from images, text, video and many other file formats. Each file (or multiple) would typically be associated to the metadata from each record (indicated by a row in the metadata capture spreadsheet). Figshare as part of their standard offer provide a staging and testing environment. The SDS team has found it useful to use this to creates a quick sample view to show the researcher how their project and data items would look on the platform, and how the metadata fields will be displayed.
3. Review and Update (pre-migration) – The SDS team provide access credentials to the researcher to view the sample view of their project alone or in tandem with a video call to provide an on-screen demonstration of how the records look and how a user may navigate through the collection (e.g., by using keyword tags / searching). During this process the researcher can provide any additional feedback the team need to take onboard before migrating the whole collection onto the platform.

---

[29] ('Upload Large Datasets and Bulk Upload Using the FTP Uploader, Desktop Uploader or API - a Help Article for Using Figshare' n.d.)

4. Data Migration – During the process of migrating data onto the platform the SDS team evaluate the best options (from the above-described techniques in section 3.2), plan and move the research data onto the platform. Each research project will have its own custom group (with personalised unique URL), that sits under their home department or faculty. The SDS team creates this, and then defines any custom metadata fields and applies any project-specific branding (such as header image) on the group landing page. The custom branding is in part cosmetic, but aids researchers with archives that are well known community resources in their field find and recognise that the data and project are the same trusted source. Once the group has been created and made live, we prepare to use an uploader tool which requires some mapping to correspond the custom metadata demanded by the researcher to the Figshare fields. For this process we use the metadata capture sheet below in Figure 2. One the preparation has been completed, the uploader deployed and is fairly efficient at both creating individual records on the Figshare side and uploading and attaching files to these records.
5. Review & Publication – Once the data has been ingested and records created on the platform, we undertake some quality assurance by random spot-checks of records to ensure they are displaying as intended. And, depending on the project and researcher, we sometimes invite a representative from the project to review the data items. If all are satisfied then the data is finally published.



Column headers highlighted in red are mandatory / default fields which have to be populated on Figshare and hopefully you have an appropriate metadata field which you can align with each of these with your project's structure

Column headers highlighted in green are optional or customisable fields depending on your requirements

For more information about Publishing Items on Figshare and the fields noted below, please visit this webpage - https://help.figshare.com/article/how-to-upload-and-publish-your-data

Internal ID	Title	Description	Keywords	Funding	References	Licence	Categories	Authors	Item Type	Files	*Custom Field 1*	*Custom Field 2*	*Custom Field 3*
											Please feel free to add as many custom fields as required (however this is not a necessary requirement)		

Figure 2: excerpt of the Sustainable Digital Scholarship service's basic research project metadata capture spreadsheet © University of Oxford.

As part of engagement with researchers we do encourage and refer to previously migrated research projects to identify metadata and tagging approaches which have proved successful. Although not every research project opts to adopt a large, customised metadata schema, we recognise that 'the use of metadata structures embedded in digital objects from the outset thereof are recommended as a starting point towards good preservation principle.'[30]

This stepped process is and can be a swift engagement period from initial contact to publication. In terms of actual engagement time and effort from a SDS team member for each project we'd estimate that between 2 – 10 days are spent in total on each. In reality the steps of the process are typically spread over a 6-12-month period. As a service we also have the flexibility to mobilise and

---

[30] ('The Use of Metadata and Preservation Methods for Continuous Access to Digital Data | Emerald Insight' n.d.)

reprioritise when required to migrate a project and its data onto the platform very quickly if required.

4. The benefits of shared infrastructures

We have touched on some of the individual benefits projects experience from hosting on a SaaS solution such as SDS. But there are strengths from having data accessible, searchable and discoverable across a platform that are only beginning to emerge, and that may prove to be the real benefit of having shared infrastructures. In SDS there are data from a variety of disciplines and subjects together on one platform, which is a technical feat given the diversity of projects and data structures they represent. We are interested to see if new research questions and projects arise from commonalities between projects within the platform, such as whether searching across early medieval or modernist literature projects creates new avenues for research.

Good customer service is the bedrock of what SDS aims to offer Oxford researchers, and we work with the lead researchers of migrated projects to learn from their experience and to promote the service to their peers. Figure 3 provides a view of the Featured Projects the SDS service has worked with to date (Feb 2023) and offers a 'shop window' and series of citable examples to other researchers. Having a relatable project or known academic using the service, has enabled us to assure new researchers to the service that working with us can lead to good outputs for their research data. As part of engagements and initial contact discussions with researchers, we will often call upon examples and screen-share how other research projects are using the SDS platform to host and curate their data. This ability to clearly demonstrate a working example of a similar project is an excellent way to help a researcher visualise and understand how their planned research data could be brought to life on the SDS platform.

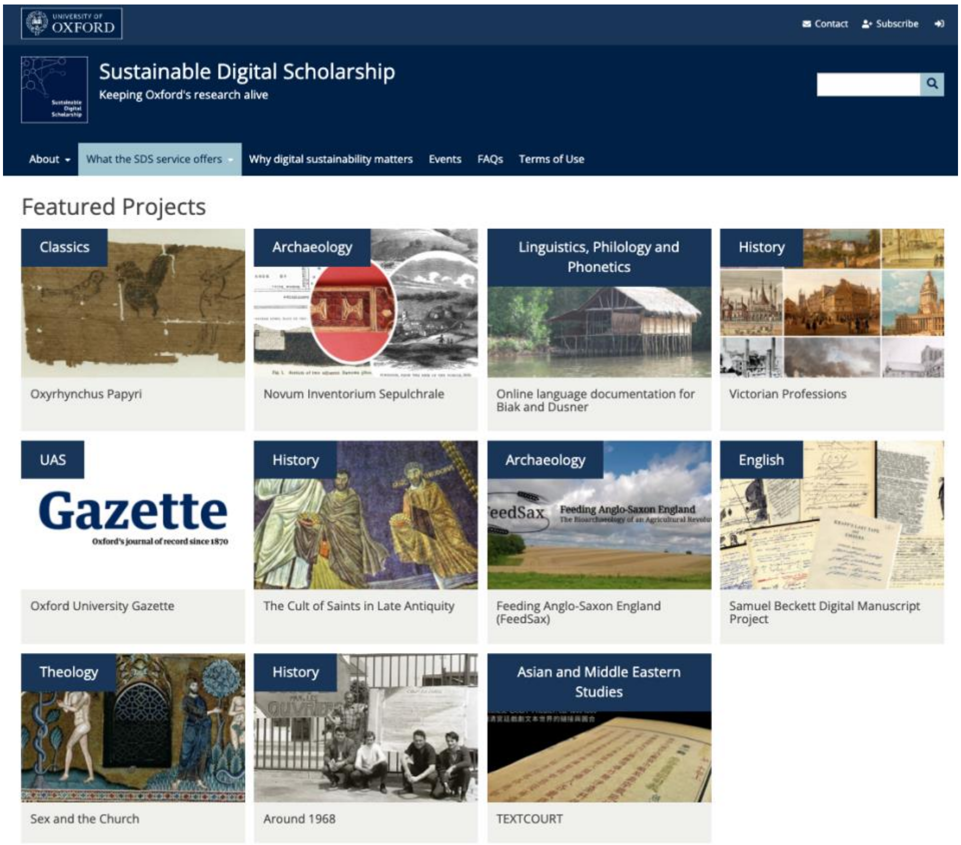

UNIVERSITY OF OXFORD
Contact
Subscribe
Sustainable Digital Scholarship
Keeping Oxford's research alive
About
What the SDS service offers
Why digital sustainability matters
Events
FAQs
Terms of Use
Featured Projects
Classics
Oxyrhynchus Papyri
Archaeology
Novum Inventorium Sepulchrale
Linguistics, Philology and Phonetics
Online language documentation for Biak and Dusner
History
Victorian Professions
UAS
Gazette
Oxford's journal of record since 1870
Oxford University Gazette
History
The Cult of Saints in Late Antiquity
Archaeology
FeedSax
Feeding Anglo-Saxon England
Feeding Anglo-Saxon England (FeedSax)
English
Samuel Beckett Digital Manuscript Project
Theology
Sex and the Church
History
Around 1968
Asian and Middle Eastern Studies
TEXTCOURT

Figure 3: Screenshot of the SDS featured projects webpage which acts as a ‘shop window’ to the SDS service © University of Oxford.[31]

Currently (February 2023) 20 of the 124 projects that the SDS service has consulted with have been fully migrated and supported onto the platform. The vast majority (72%) of these are from the Humanities Division, which is unsurprising as the SDS service was developed by the Division to meet the needs of its researchers, in particular its digital humanists. However, a recent review of research data management provision at the University shows that the type of hosting that the SDS offers for warm data is not unique to humanities or cultural heritage data.[32] At the University of Oxford, we have been expanding out the service to other academic areas and actively promoting it across the university.[33] More focused efforts have been made initially with the Social Sciences Division at Oxford, and there is interest and traction growing from a range of departments in this area. Indeed 18% of the 124 project consultations we have undertaken are from the Social Sciences Division. This would mean that the remaining 10% of research project consultations with are from the Mathematical, Physical and Life Sciences and Medical Sciences Divisions, which the research data management review had identified as having some warm data needs.[34] However, we are aware that there are barriers in the sciences due to data type and the potentially large sizes of data sets. Our service offering has primarily been developed for the sharing and dissemination of humanities and cultural heritage data which does not often have the limitations of personal and sensitive data found particularly in medical research. Furthermore, the scale and size of data hosting requirements from the sciences can often be in the multiple terabytes or petabytes. Theoretically the Figshare platform, is able to ingest large data as it is built on built on Amazon Web Services with scalable data hosting, but in reality, cost of hosting significant volumes of data in the cloud can become costly and it often not economically viable for researchers or institutions.

We hope that once we have spent some time on targeted engagement with scientific departments we will be able to have a migrated project that we can use as a reference point in our featured projects to show researchers from a similar academic background.

5. Conclusion

Every research project and every dataset is unique, but that does not mean that the infrastructures which support the archives of these data need to be. Standardisation of some aspects of infrastructure has been made possible for the SDS service by working with a Software as a Service provider rather than building a bespoke service. The SDS service is then able to work with researchers as consultants to advise the best location for their research data, and migrate some of those consulted researchers’ projects to the SDS platform based on Figshare’s software. Not every project we work with will chose SDS as there are many other repository services available at Oxford, but we hope that fewer than in the past will chose the path of building a bespoke relational database to store their data.

---

[31] (‘Featured Projects’ n.d.)

[32] (Chiarelli et al. 2022)

[33] (‘Future of SDS’ n.d.)

[34] (Chiarelli et al. 2022)

With every researcher we work with, our institutional knowledge of research data and platforms grows, and we are able to draw more data at risk from the shadows and into a place where the data can find a sustainable home.

Infrastructures for research are a rapidly growing and acknowledged issue, with institutions trying to keep pace with researcher needs and research outputs.[35] Getting the right archives or repositories is important but ‘sustainability through institutional maintenance of digital infrastructure is a large and potentially explosive burden.’’[36] Infrastructure need not be a burden. The significant advantages of utilising a SaaS provider offer benefits in terms of costs and the ability to move swiftly.

At the University of Oxford, researchers are free to choose the most appropriate archive for their data and this will continue. What is changing is that government and funder mandates around data management plans and open data deposits combined with researcher-led initiatives such as FAIR and CARE data means that researcher are more likely to need the services of a sustainable open and institutionally managed platform. SDS exists because it addresses a researcher need for warm data to be collected, analysed, published, updated and disseminated, and as a service we provide both consultations and an archive repository platform. It emerged to meet the needs of digital humanities researchers, often working with cultural heritage data, but has wider applications and benefits.

---

[35] (Taylor et al. 2022)

[36] (van Zundert 2012)

Åhlfeldt, Johan, and Maria Johnsson. 2015. ‘Research Libraries and Research Data Management within the Humanities and Social Sciences’. http://lup.lub.lu.se/record/5050462.

Arms, William Y. 2001. *Digital Libraries*. MIT Press.

‘Around 1968: Activism, Networks, Trajectories’. n.d. Accessed 18 October 2022. https://www.sds.ox.ac.uk/around-1968-activism-networks-trajectories.

Banks, David. (2017) 2022. ‘Figshare-Uploader’. C++. https://github.com/amoe/figshare-uploader.

Business, Agile. n.d. ‘Chapter 10: MoSCoW Prioririsation’. Accessed 18 October 2022. https://www.agilebusiness.org/dsdm-project-framework/moscow-prioririsation.html.

Chang, Tao. 2021. ‘Arts and Humanities Infrastructure Enabling Knowledge with Impact’. September 2021. https://www.ukri.org/blog/arts-and-humanities-infrastructure-enabling-knowledge-with-impact/.

Chiarelli, Andrea, Neil Beagrie, Lotte Boon, Ruth Mallalieu, Rob Johnson, Amy Warner May, and Rowan Wilson. 2022. ‘To Protect and to Serve: Developing a Road Map for Research Data Management Services’. *Insights* 35 (0): 4. https://doi.org/10.1629/uksg.566.

Chowdhury, Gobinda G., and Sudatta Chowdhury. 2003. *Introduction to Digital Libraries*. Facet Publishing.

David de Roure. n.d. ‘SHAPING DATA AND SOFTWARE POLICY IN THE ARTS AND HUMANITIES RESEARCH COMMUNITY’. *AHRC (Study For)*. https://www.ukri.org/wp-content/uploads/2022/10/AHRC-011122-SSIReport-ShapingDataAndSoftwarePolicyInTheArtsAndHumanities.pdf.

‘Digital Scholarship at the University of Oxford’. n.d. Accessed 28 February 2023. https://digitalscholarship.web.ox.ac.uk/home.

ERC. 2022. ‘Open Research Data and Data Management Plans Information for ERC Grantees’.

‘Featured Projects’. n.d. Accessed 18 October 2022. https://www.sds.ox.ac.uk/featured-projects.

Fuhr, Norbert, Giannis Tsakonas, Trond Aalberg, Maristella Agosti, Preben Hansen, Sarantos Kapidakis, Claus-Peter Klas, et al. 2007. ‘Evaluation of Digital Libraries’. *International Journal on Digital Libraries* 8 (1): 21–38. https://doi.org/10.1007/s00799-007-0011-z.

‘Future of SDS’. n.d. Accessed 28 February 2023. https://www.sds.ox.ac.uk/future-sds.

Harrington, Jan L. 2016. *Relational Database Design and Implementation*. Morgan Kaufmann.

Harrower, Natalie, Maciej Maryl, Timea Biro, Beat Immenhauser, and ALLEA working group E-humanities. 2020. ‘Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA Working Group E-Humanities - Digital Repository of Ireland’. https://repository.dri.ie/catalog/tq582c863.

Lesk, Michael. 2005. *Understanding Digital Libraries*. Elsevier.

Liu, Rui, Dana McKay, and George Buchanan. 2021. ‘Humanities Scholars and Digital Humanities Projects: Practice Barriers in Tools Usage’. In *Linking Theory and Practice of Digital Libraries*, edited by Gerd Berget, Mark Michael Hall, Daniel Brenn, and Sanna Kumpulainen, 215–26. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-86324-1_25.

McGillivary, Barbara, Beatrice Alex, David Beavan, Giovanni Colavizza, David De Roure, Federico Nanni, Thierry Poibeau, and Melissa Terras. 2020. ‘The Challenges and Prospects of the Intersection of Humanities and Data Science’. A White Paper from The Alan Turing Institute. The Alan Turing Institute. https://www.turing.ac.uk/research/publications/challenges-and-prospects-intersection-humanities-and-data-science.

McKnight, J, Jonathan Prag, and Christine Madsen. 2014. ‘DHARMa (Digital Humanities Archives for Research Materials)’.

‘New Administrative Batch Management in Figshare’. n.d. Accessed 18 October 2022. https://figshare.com/blog/New_Administrative_Batch_Management_in_Figshare/654.

‘People & Projects’. n.d. Accessed 28 February 2023. https://digital.humanities.ox.ac.uk/people-projects.

'Research Data Oxford'. n.d. Research Data Oxford. Accessed 18 October 2022. https://researchdata.ox.ac.uk/.

Strange, Damon, Blaine Roissetter, Christine Madsen, and Megan Hurst. 2019. 'Digital Humanities Projects Review & Sustainability Analysis'.

Su, Fangli, and Yin Zhang. 2021. 'Research Output, Intellectual Structures and Contributors of Digital Humanities Research: A Longitudinal Analysis 2005–2020'. *Journal of Documentation* 78 (3): 673–95. https://doi.org/10.1108/JD-11-2020-0199.

Taylor, Rebecca, Johanna Walker, Simon Hettrick, Philippa Broadbent, and David De Roure. 2022. 'Data and Software Policy in the Arts and Humanities Research Community: A Study for the AHRC'. Software Sustainability Institute. https://www.ukri.org/wp-content/uploads/2022/10/AHRC-191022-SSIRecommendations.pdf.

'The Use of Metadata and Preservation Methods for Continuous Access to Digital Data | Emerald Insight'. n.d. Accessed 28 February 2023. https://www.emerald.com/insight/content/doi/10.1108/02640471111125195/full/html.

'Towards a National Collection | Collections United'. n.d. Accessed 18 October 2022. https://www.nationalcollection.org.uk/.

'University of Oxford Research Repository - Search'. n.d. Accessed 18 October 2022. https://portal.sds.ox.ac.uk/search.

'Upload Large Datasets and Bulk Upload Using the FTP Uploader, Desktop Uploader or API - a Help Article for Using Figshare'. n.d. Accessed 28 February 2023. https://help.figshare.com/article/upload-large-datasets-and-bulk-upload-using-the-ftp-uploader-desktop-uploader-or-api.

'Welcome to the Sustainable Digital Scholarship (SDS) Service'. n.d. Accessed 18 October 2022. https://www.sds.ox.ac.uk/home.

Whyte, Angus, Claudia Engelhart, Daniel Bangert, Gabin Kayumbi-Kabeya, Simon Lambert, Mark Thorley, Ryan O'Connor, Patricia Herterich, and Joy Davidson. 2019. 'D3.2 FAIR Data Practice Analysis', December. https://doi.org/10.5281/zenodo.5362079.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Zundert, Joris van. 2012. 'If You Build It, Will We Come? Large Scale Digital Infrastructures as a Dead End for Digital Humanities'. *Historical Social Research / Historische Sozialforschung* 37 (3 (141)): 165–86.