The ADS Archiving Workflow

By Nicky Garland, Katie Green,

Olivia Foster and Jenny O'Brien

Version 1 - 17.05.2024

Please cite as:

Garland, N., Green, K., Foster, O. and O'Brien, J. 2024. The ADS Archiving Workflow. Version 1.0. University of York.

Introduction

This document describes the workflow that the ADS utilises to archive data into its collections. Below we describe the main steps of the archiving process. The ADS Archiving workflow includes both human-led and machine (i.e. automated) tasks. Archival procedures are reviewed and updated on a regular basis.

For a full description of these processes please see the ADS policies and procedures page and the Instruction for Depositors, both of which can be found on the ADS website.

The ADS archiving workflow follows the Open Archival Information System (OAIS) reference model, which provides a framework for the wider archival process. This model provides a broad workflow that begins with data creators and, following accession and preservation, leads to a dissemination version that can be accessed, queried and downloaded by a wide user base. Following this system allows our workflow to be understood and assessed by other data archivists to ensure high archiving standards.

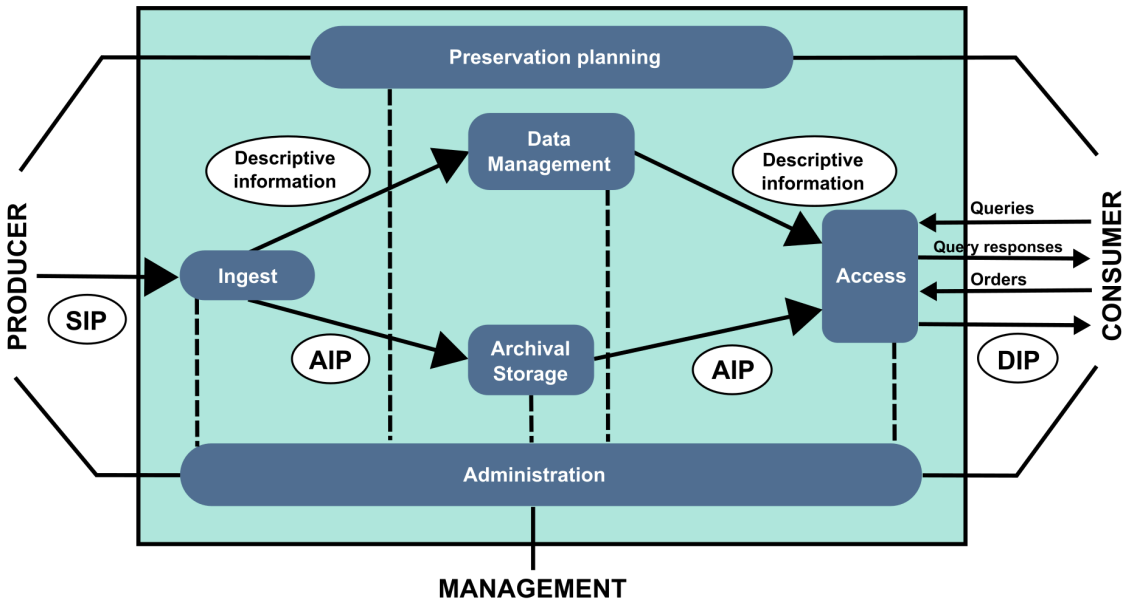

Figure 1: A diagram of the OAIS functional model diagram

Below is a diagram of the ADS Archiving workflow. Click on the corresponding heading in the table of contents to jump to the relevant sections of this document.

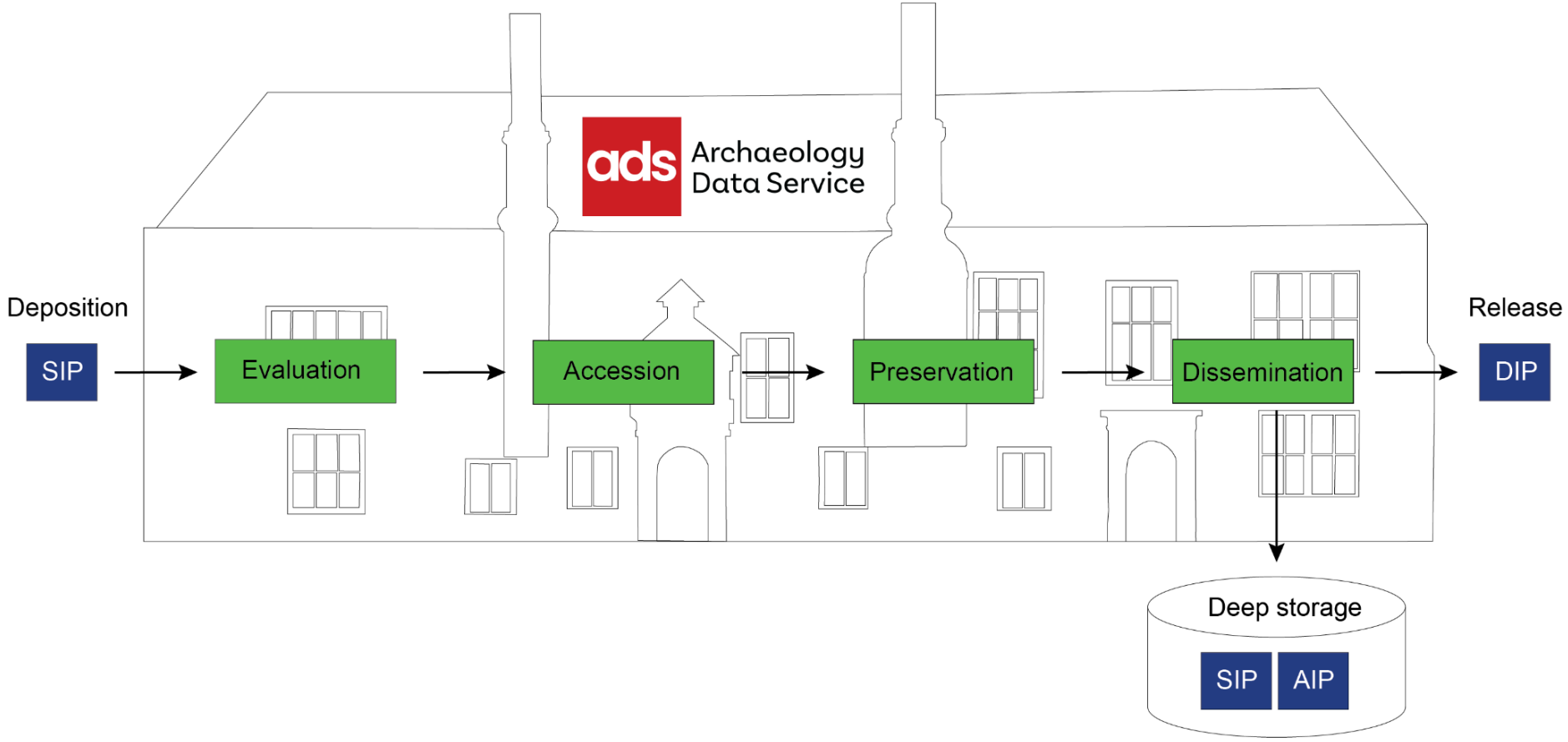

Figure 2: A diagram of the ADS Archiving Workflow

Table of Contents

Evaluation

The first stage of the ADS Archive workflow is the evaluation of the data that we receive. This evaluation is undertaken to see if the dataset is suitable to archive with the ADS.

Receive data from depositor

The first step of the evaluation stage is to receive data from our depositors. We ask that all depositors refer to the ADS Instruction for Depositors to ensure that the collection is correctly prepared before depositing data with us.

For small and medium sized projects, data can be submitted through our online submission process known as ADS-easy. For larger datasets, the data is submitted directly to the ADS collections team, often via email or a secure file transfer and in some instances by posting physical storage devices to the ADS.

Metadata

All data deposited with the ADS must be accompanied by appropriate metadata. Metadata is a set of data that describes and gives information about your dataset. There are different types of metadata including collection-level metadata, object metadata and technical metadata. This video provides a summary of the importance of metadata.

Part of the metadata required for both collections and objects is enough descriptive metadata to aid discovery. Without descriptive metadata, archive users would not be able to find the data or records they need. In general, descriptive metadata roughly falls into three main categories: 'What', 'Where', and 'When'. Each category explains what the content of the collection or digital object is, and the location and period of time that it focuses on. This metadata includes useful titles and descriptions, as well as keywords and phrases, using consistent terminology where possible.

This consistency ensures that the collections and objects can be found via ADS search interfaces, as well as any external catalogues that use the ADS catalogue records.

Metadata is assessed as part of the Assessment and Appraisal stage and improved during the Dissemination stage, particularly during the creation of the DIP.

Assessment and Appraisal

Once the data has been received, we then assess and appraise the dataset. This is an important step in the archive procedure as it ensures that all datasets the ADS receive are well formed, documented and meet all legal requirements.

The data is assessed by ADS archivists to make sure it meets a series of important criteria. The assessment criteria are as follows:

1. The data deposit contains no malware.
2. The digital objects are in correct formats as relevant to the type of dataset.
3. The data deposited has collection-level metadata.
4. All digital objects have core metadata.
5. Digital objects have additional technical metadata (if required).
6. Digital objects can be opened, are valid, and can be reused.
7. The data deposited has no sensitive data concerns.
8. The content is appropriate and complete.

Following the assessment process there are two possible outcomes:

- If there are no noted issues then the dataset can be accessioned;
- If issues are highlighted then the dataset is returned to the Depositor for correction or clarification.

This stage of our workflow is undertaken prior to the accession of any data into ADS systems. For full details of this process please refer to the ADS Assessment and Appraisal Policy.

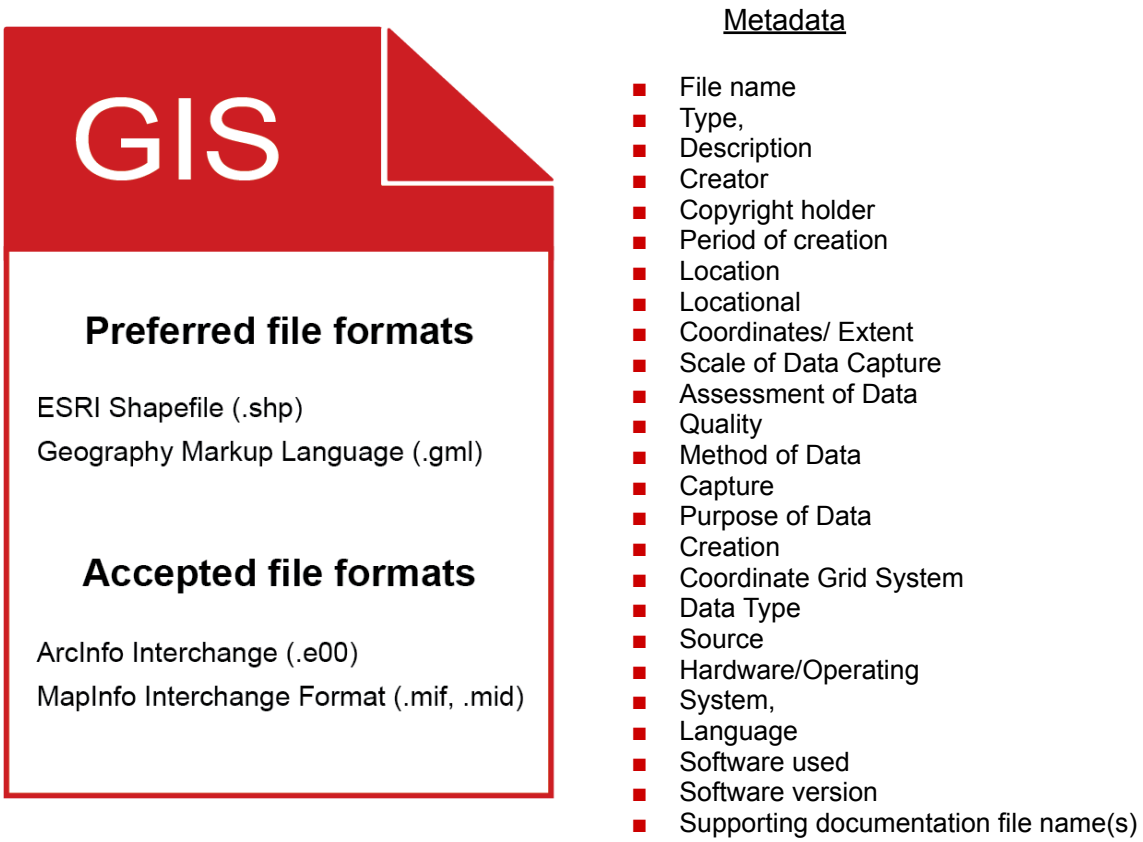

Figure 2: Example file format including metadata required: Geographical Information Systems

The Submission Information Package (SIP)

Once the assessment procedure is complete and any changes have been made by the depositor then the data archive is formally accepted to the ADS as a Submission Information Package, or SIP.

A SIP is defined as the data provided by a data creator or depositor, including any documentation that is required to facilitate archiving and reuse (i.e. metadata). This is only the first of three versions of the data that are created and archived by the ADS. An Archival Information Package (AIP) and Dissemination Information Package (DIP) are also created in this process. These are each detailed later in this workflow document

Accession

With the formal acceptance of a dataset by the ADS and the creation of a SIP, the process of accession can begin. Accession is the process by which a new dataset is added to the ADS collections.

Record the Collection

Accessioning the data begins with acknowledging receipt of the data from the depositors and creating an entry of that dataset within the ADS Collection Management System (also known as the CMS). The CMS is a large database that contains the details of all of the archives (or collections) that are held by the ADS.

This entry records the date of accession and the details of the depositor, as well as creating a digital checklist of all the steps that we need to follow to track the collection as it moves through the ADS workflow. Some of the collection-level metadata is also included in this record. At this point the dataset is also given an unique identifier that will be used to locate this data throughout this process.

Deposit licence

As part of the archiving process with the ADS, we ask that all depositors sign a 'deposit licence'. This licence is a contract between the depositor and the ADS and gives the repository permission to disseminate data on their behalf. This deposit licence allows the ADS to share the data but does not transfer the copyright for that data. The copyright is retained by the depositor. More information is provided on the ADS website.

Organise the data

The next step is to properly organise the dataset that we have received to ensure that it is compatible with the other archives that we hold.

First we ensure that all of the file names adhere to the guidance provided in our Instruction for Depositors. This guidance demonstrates best practice in file naming to ensure consistency across your dataset and so that people who look at your data can

easily understand what the files contain. The ADS uses an online system (ADS-Easy) that automates some of this process, however, not all datasets come to us in this way so in other instances this is amended by hand.

Next we check the structure of the data directory. There is no single way to structure your data, but we do ask depositors to put their data in a 'logical file structure' to ensure that all of the data is easy to find. At this stage we check the structure of the data and edit it if needs be.

Once this is all complete, then the dataset is then copied to the ADS data servers.

Preservation

This next stage of the archiving workflow is the process by which data is transferred from the depositors into an archive for long-term preservation and, later, dissemination. For full details about how we undertake these stages of preservation and access please refer to the ADS Ingest Manual.

Preparation of the Archival Information Package (AIP)

The next step of the process is to ensure datasets are suitable for long term preservation. This is accomplished by creating what is known as an Archival Information Package (AIP). This is the second (and a separate) version of the data that is accessed with the ADS. The AIP will contain both metadata that describes the structure and content of an archive as well as the actual data itself.

Normalisation

An important step in creating a version of the dataset that is suitable for preservation is Normalisation. Normalisation is the process by which each of the files within a dataset are converted to a single format according to their data type. The format for a particular data type (e.g. text files) is carefully chosen as we need to use a format that provides the best compromise amongst several characteristics (e.g. functionality, longevity, preservability) that help to preserve that data for the long term. The formats that we have chosen are set within individual ADS Data Procedures.

These conversions are done with two purposes in mind: preservation and dissemination. One version may be created with long-term preservation purposes in mind, while the other is created to allow us to disseminate the data as an archive on the ADS website (which is discussed below).

Migration (sometimes…)

Archiving data is an ongoing process and in some cases we may need to revisit archives to ensure that the formats that are used are suitable so that they can continue to be preserved. In some cases file formats become obsolete and unreadable

and it is up to us to update those files in our archive to later versions. This is the process of Migration.

Migration of files may not happen for every archive but it is important to ensure the longevity of the archives that we hold. Here at the ADS we track the file formats that we accept and migrate files where needed. In some cases we may receive data that includes the correct file formats, but the versions of this data are older than required for preservation. In this case we will migrate those files to a newer version as part of the normalisation process.

Create technical metadata

For every file within the dataset we need to create what we call 'technical metadata', i.e. information about the file type, size etc. At the ADS we use software called DROID to undertake this process automatically.

DROID is a software developed by the National Archives that identifies file formats (and the version of each format) for data in a batch. The dataset can be run through this software automatically to create technical metadata for each and every file in the dataset. Some manual checking is required to ensure that all files within the dataset are accounted for and have been assigned the correct data type.

Document, document and document some more...

An important part of the overall archival process is to document all the processes that have been undertaken and the reasoning behind those processes. In many cases these processes are standard but where we encounter unusual file formats or disorganised data some extra information may be required. In part this is the reason why we document our processes in detail in the ADS policies and procedures. All of this information is documented in our internal databases. Once the preservation procedure is complete and the AIP has been created we also document this part of this process.

Dissemination

The final stage of the archiving process is the creation of a final archive that is made available for all to access and use via the Archives section of the ADS website.

Creation of the Dissemination Information Package (DIP)

The third and final version of the archive is the Dissemination Information Package (DIP). This version is generated by the SIP and/or AIP and is made available to our users via the Archives section of the ADS website. The creation of the DIP is usually undertaken at the same time as the creation of the AIP, for reasons of efficiency, however, it may be a slightly different version of the dataset that is more accessible or user friendly. The processes needed to create the DIP are documented in the same way as the AIP.

Release

Following the creation of the DIP, the different versions of the dataset are checked by other archivists and the Collections Development Manager. A final check is made of the AIP and DIP to ensure that all procedures have been followed and that all the processes have been properly documented.

During this data checking process the following factors are considered:

Date of release

Once prepared and ready for release an agreement for the date of release is sought from the depositor. In some instances the release of the archive may be embargoed for a specific period of time. This may be because the archive contains sensitive information or that the depositors need to allow time for the full publication of any outputs from their research or investigation. If this is the case a date we will determine what can be published and establish a data with the depositor for the publication of the full archive.

Publicise archive

Once released the archive will be published on the ADS website and publicised through our communication channels (e.g. social media, newsletter). We also encourage depositors to publicise their dataset through their communication channels.

Persistent Identifier

At this stage each collection is assigned a Digital Object Identifier (DOI), also known as a persistent identifier, which can be used to locate the archive. The DOI provides a consistent and accurate way to reference digital objects, just in the same way you might reference a journal article or book.

Once each of these considerations have been assessed, the archive is officially signed off by the Collections Development Manager. Once this is complete then the interface to the archive is released online.

Interface

As part of the dissemination process an ADS collection interface is created for each dataset that we archive. You can search and access all of the archived collections using the ADS Archives Search tool. The interface is important as this is the way in which users view and access archived data on the ADS website. This is a typical layout used by the ADS (https://doi.org/10.5284/1118849).

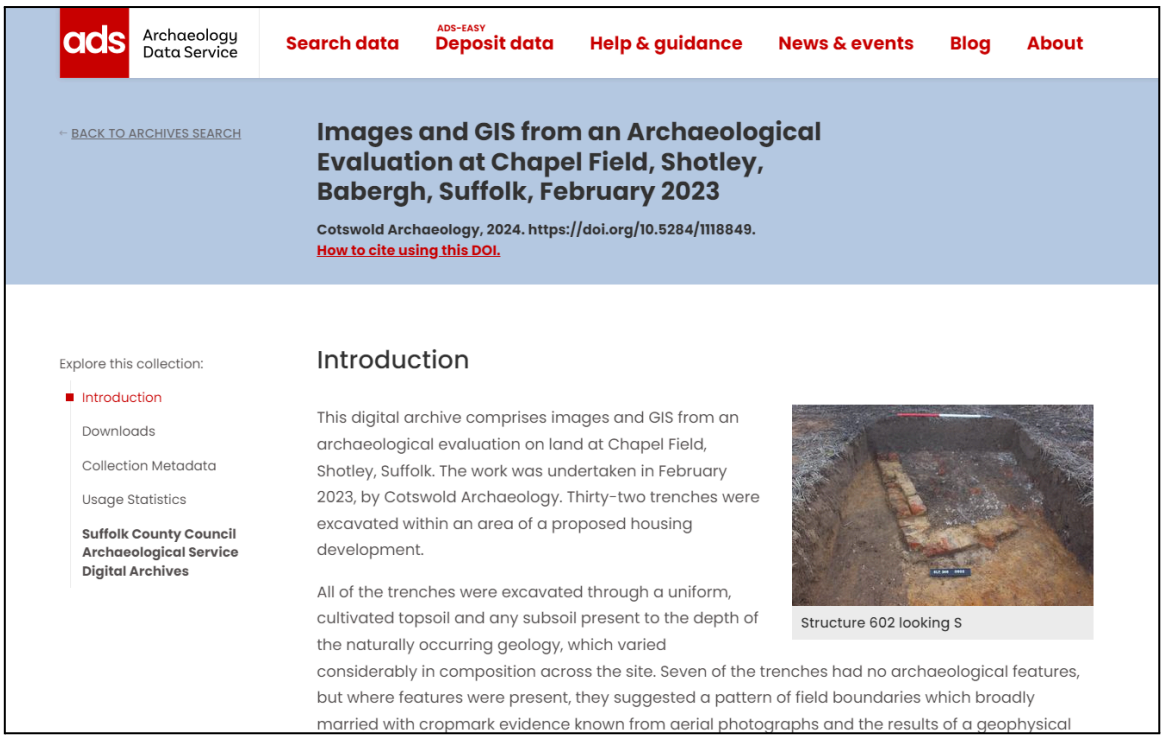

Figure 3: A typical layout for an ADS collection.

The interface will be designed to reflect the specific archive, however it will usually contain the following sections:

Introduction: The introduction page includes a short description of the archive, as well as where the data was collected from and by whom. This page also includes details about who to contact if you wish to find out more information about the collection. This information is shown on the left hand navigation bar under 'Primary Contact'.

Downloads: The downloads page is the part of the collection where you can directly access the contents of the archive that are available to download. The downloads are usually divided between data types (i.e. Reports, Images, GIS), but can be arranged thematically, or in any way that makes sense to the collection. Each downloadable file in the archive, or object, comes with its own metadata, which can be accessed from the 'Info' button.

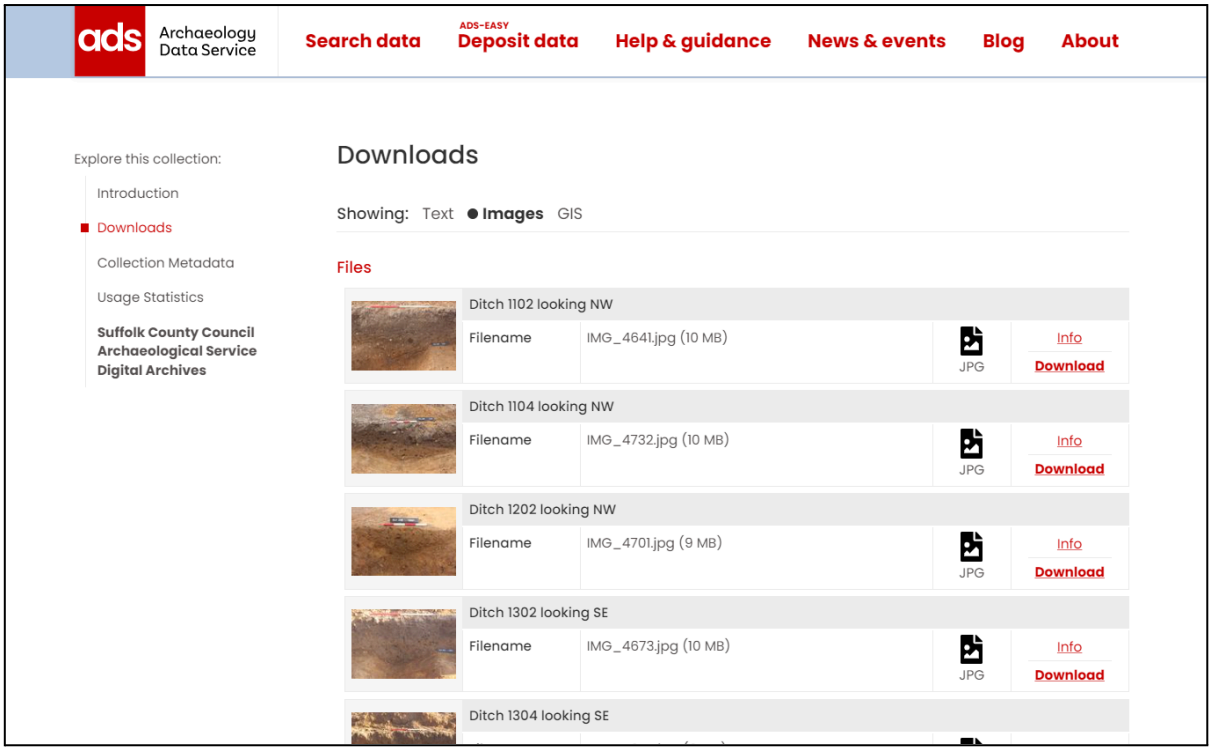

FIgure 4: Collection downloads page

Metadata: The metadata page provides all of the collection-level metadata, i.e. the data that describes the entire archive. This collection-level metadata typically includes information about geographic location, when the data was collected and by whom, as well as a breakdown of the numbers of different file types that are included in the collection. This data is important for facilitating the reuse of this data by others in their own research.

Usage Statistics: The final page provides a graphical interface to track the extent to which the collection has been accessed. The data collected includes:

- The number of times the collection has been visited.
- The number of times the collection has been downloaded.
- The number of page views within the archive.
- An overview graph for visits, downloads and page views.

This data is generated automatically and can be downloaded, either as an image of the graph (in several different formats) or as a .csv file to allow users to examine the figures themselves.

Deep storage

Following the completion of the AIP and DIP, including all relevant metadata and documentation, we send a copy off to our deep storage to ensure the long term preservation of the dataset.

This part of the process is usually undertaken once the dissemination version of the archive has been released for use by others. A backup of this data is stored off site to ensure that if disaster strikes we continue to maintain all of the data archived with us for the long term.

Summary

This document outlines the main stages of the ADS archiving process, extending from receiving data from depositors, through accession and preservation procedures, to finally disseminating the collection via the ADS website and sending archives to deep storage.

However, the work of a data repository is never done. As discussed above in the section on data migration, maintaining the archives is just as important as the initial archiving workflow. The ADS maintains and stores the original data for depositors and users to access and we maintain the long term preservation and integrity of the data. This includes migrating data but also updating our interfaces to update access and improve accessibility. The ADS also works to ensure that the dissemination version is accessible and available for reuse and encourages reuse of our collections by the heritage sector and beyond.

Glossary

Accession - The process by which a new dataset is added to the ADS collections.

AIP - The Archival Information Package. An archive ready version of the data that contains both the metadata that describes the structure and content of an archive as well as the actual data itself.

Collection - The term used by the ADS to describe datasets archived.

Collection level metadata - This is the type of metadata that records information at a broad level for the overall collection. This metadata may include project details (e.g. name), site codes and geographic location. This may also be referred to as Project-level metadata.

Dataset - A collection of related sets of information that is composed of separate elements but together forms part of the wider archive.

Deposit licence - A licence gives the repository permission to disseminate data on behalf of the depositor.

DIP - The Dissemination Information Package. The public version of the data that is made available to our users via the ADS website.

Interface - An interface is the structure and web page on which users view archives via the ADS website.

Metadata - A set of data that describes and gives information about other data.

Migration - Migration is the process by which file formats are updated to the most recent versions as they have become obsolete and unreadable.

Normalisation - Normalisation is the process by which the files within a dataset are converted to a single format to enable long term preservation and dissemination.

Open Archival Information System (OAIS) - A Reference Model that was developed to provide a broad consensus on the requirements for an archive or repository to provide long-term preservation of digital information.

SIP - The Submission Information Package. The data provided by a data creator or depositor, including any documentation (i.e. metadata) that is required to facilitate archiving and reuse.

Technical metadata - This is the type of metadata that provides information from specific files, including file type and size. In many cases we can use software (such as DROID) to undertake this process automatically.

Usage Statistics - This is the information that relates to how often users visit and interact with a particular collection.