

Hybrid Archives Model

A Model for the Joint Curation of Digital Resources Guidebook

Document Details

Document type: Project Report
Author: Gareth Knight & Hamish James
Draft / Version: Draft 2.0
Date of edition completion: 12/11/2004

Change History

1. First draft, 26/08/04, combining HJ work on WP1 and GK work on WP2 to form draft of introductory sections of final report
2. Further work on first draft in September 2004
3. November 2004, second draft

Contents

Contents	3
Introduction to the Hybrid Archives Project	4
Outline of the Hybrid Model	7
Deposit Agreement.....	8
Scenarios.....	8
Responsibility for Transferring Metadata and Data.....	10
Transferring Metadata.....	10
Transferring Data	10
Operational View of the Transfer Process	11
Subsidiary Requirements	12
Conversion to Full Deposit	13

Introduction to the Hybrid Archives Project

Traditionally, when either analogue or digital collections are lodged with an archive, the archive then takes full responsibility for preserving and providing access to the collection. The AHDS has found that *full deposit* of this nature is not always an attractive option for the institutions that provide the bulk of the resources held by the AHDS:

- Institutions may not wish to hand over control (as they see it) of their collection to a third party (and there is no legal requirement for them to do so)
- Institutions may wish to act as the primary disseminator for the collection in order to attract students and/or researchers, and to profile the work of the research team responsible for creating the collection
- There may be outstanding issues with copyright or other forms of intellectual property rights that create significant barriers to full deposit
- The collection may be dynamic – that is, still being extended and added to – where traditional archiving generally assumes a static, complete collection

Problems such as these make it difficult to transfer full responsibility for a digital resource from its original institutional home to an archive. Normal practice at the AHDS has been to avoid these types of problem by concentrating on acquiring resources whose owners have no objection to full deposit. However, as the I.T. infrastructure of UK higher and further education has developed, the ability of institutions to distribute their own resources has grown. It is now relatively easy to disseminate resources such as electronic texts, datasets and image collections across the internet and World Wide Web. Consequently, in recent years, there has been a rapid growth in the number of researchers who choose to become resource creator-distributors.

Rather than rely on traditional publishing or the deposit of unpublished material in an archive, research groups now often make their outputs directly available via the Web. Whilst most institutions provide back-up facilities for websites, this takes no account of the need to migrate systems and content after quite short spans of time, and is no substitute for a formal preservation strategy. Moreover, should the host institution no longer be able or willing to support the collection there is a real danger that the collection will no longer be available to the wider community and may be lost forever.

Full deposit is poorly suited to dealing with sophisticated web-based resources, because it relies on separating the content of a resource from its technical environment (the hardware and software used to access the content), and then depositing the content in an archive. Unfortunately, content and functionality are increasingly interwoven in web-based resources. Content, for example, may not be stored in the form it is presented to users, but rather it is compiled and processed on-the-fly in response to a request made by the user. In this situation, a traditional full deposit does not enable the archive to replicate the quality of delivery provided by the original website.

Duplicating the entire arrangement of servers, operating systems, security settings, application software and data needed to run the original website is much harder than just preserving the content, and many web-based resources are therefore not properly archived. Some of these problems can be overcome by depositing a copy of the content with an archive for preservation while the owning institution continues to provide access to the resource from its own website. However, this approach brings its own difficulties. It effectively leaves the master copy of the resource at the institution, creating a need for synchronisation between the two copies if changes are made, and a danger that changes made to the institutional copy will not be passed on to the archive. It also requires careful agreement of the conditions under which the archived copy can be distributed, and monitoring of these conditions, otherwise the archive may find it holds data that can never be released to users.

The Hybrid Archives Project was designed to investigate and address these problems by developing a new model for the joint curation of digital resources. In the hybrid model, responsibility for managing a resource is shared between an archive and the depositing institution, allowing the institution to provide sophisticated access to the resource, while the archive ensures the longer-term survival of the underlying content. Under the hybrid model, depositors will:

- Disclose agreed metadata at a detailed level of granularity for capture by the AHDS using either Z39.50 or OAI as appropriate, and an agreed preservation metadata set
- Make available, in a form and format to be agreed, content for preservation. This is likely to include documentation and data e.g. texts, images, databases, sound, moving images etc. plus explanatory documentation to enable informed use of the collection
- Sign a formal licence agreement that regulates the process, defines the rights and responsibilities of each party, and provides for a move to full deposit should the institution no longer be willing or able to continue to support the collection
- Move to full deposit should the institution no longer be willing or able to continue to support the collection
- Take responsibility for disseminating and supporting use of the collection

And the AHDS will:

- Capture rich, complex metadata and integrate it within AHDS search and retrieve systems for cross-searching with other AHDS collections, using Z39.50 and OAI technologies as appropriate
- Expose this metadata (along with metadata from full deposit collections) to appropriate Portals in the JISC information environment
- Capture content for preservation and integrate it into the AHDS preservation system
- Sign the formal licence agreement that regulates the process, defines the rights and responsibilities of each party, and provides for a move to full deposit should the institution no longer be willing or able to continue to support the collection
- Move to full accession of the collection should the institution no longer be willing or able to continue to support the collection
- Publicise the collection and promote awareness and use of the collection

The AHDS has long recognised that “the concepts of collecting for access and preservation need greater definition in a digital environment where there are a range of possible resource types and access arrangements”[1]. In 1998 the AHDS described a five level hierarchy of models for managing digital resources[2]. Only in the top two levels of this hierarchy, the archived and served models, is a resource actually transferred to the AHDS, and only in the top level does the AHDS take responsibility for preserving the resource (traditional full deposit). Resources managed using the served model are delivered to users by the AHDS, but not preserved by the AHDS.

The new hybrid model is effectively the reverse of the served model, creating a situation where the AHDS preserves a resource but does not normally disseminate it.

archived: the resource is archived by the AHDS and the AHDS intends to preserve and keep the intellectual content of the resource available on a long-term basis. The resource will also normally be disseminated by the AHDS unless special arrangements have been agreed with a depositor eg to restrict access for a specified period of time.

hybrid: the resource is accessioned, catalogued and preserved by the AHDS, but another institution retains primary responsibility for content and delivery.

served: the resource is accessioned, catalogued and disseminated by the AHDS but another institution has primary responsibility for content, maintenance and long-term preservation. This collection level may include 'mirrored' resources where a copy of a digital resource residing elsewhere is hosted by the AHDS to improve access, or resources held, maintained, or preserved by collaborating and commercial agencies, which are licensed and disseminated by the AHDS.

brokered: the resource is physically hosted elsewhere and maintained by another institution but the AHDS has negotiated access to it with a collaborating agency and includes metadata and links for the resource in its catalogue, or AHDS users are able to locate and cross-search, and in some circumstances acquire access to it.

linked: the resource is hosted elsewhere and the AHDS provides a web link pointing to it

at that location from its webpages. The AHDS has not accessioned that resource or negotiated a collaboration agreement with the agency which maintains it and has no control over the information or formal agreements for access to it.

finding aids: electronic finding aids and metadata held by the AHDS which will facilitate discovery and searching of digital resources. This metadata is associated with digital resources such as collections at the AHDS or elsewhere but may be stored, managed and maintained separately from them.

Table 1: Deposit Models at the AHDS

By preserving, but not disseminating resources, the hybrid model directly addresses the concerns of depositors who wish to continue to manage and be identified as the main disseminator of the resource, or who are unable to transfer responsibility for dissemination of the resource. By including resource discovery metadata in the AHDS catalogue, potential users will be able to locate hybrid deposits and then move to the institutional collection site to access the resource.

Digital resources are typically deposited with the AHDS after the completion of a research project. In this situation, the hybrid model is in fact very similar to the existing archived category. The AHDS has never claimed ownership of the resources deposited with it, and depositors have always been free to continue to disseminate their resources elsewhere. Currently, the AHDS archives a number of collections that are also available from other locations, and collections that are held for preservation only and are not available to users.¹ For these collections, the benefit of the Hybrid Archives project will be to provide a standardised approach that will replace the *ad hoc* arrangements made in the past with depositors who do not wish the AHDS to disseminate their resource.

Standard practice at the AHDS and other digital archives is to use a semi-manual approach to ingest; an expert member of staff, following general procedures and best practice guidance, makes decisions about the ingest process that are then implemented using a variety of software tools. This approach can handle a diverse range of resource types that are often not standards compliant, and it can be extended to cope with infrequent updates, but even a yearly update reveals the repetitious nature of the ingest process. As resources become more dynamic and complex there is a need for a more strictly defined, more automated, approach to ingest. The main focus of the Hybrid Archives project has been investigating how this might be achieved for the wide variety of digital resources that are deposited with the AHDS.

Deposit Frequency	Archived	Hybrid	Served
Once only deposit	✓	✓	✓
Yearly updates	✓	✓	✓
Monthly updates		✓	✓
Weekly, daily updates		✓	

Table 2: Deposit Frequencies and Models

¹ The AHDS discourages preservation only deposits, but may embargo collections for a fixed period of time.

Outline of the Hybrid Model

To create a workable model that addresses the problems discussed above, the Hybrid Archives project has concentrated on four aspects of the ingest of new collections into a digital archive: licensing and intellectual property rights, metadata harvesting and standards, methods for harvesting data, and requirements for long-term preservation.

The generic approach to full deposit taken by the AHDS involves a series of steps undertaken by the depositor and the AHDS to assemble, describe, transfer and preserve a digital resource:

- Depositor contacts the relevant AHDS Centre via the Depositing Advice Team
- Depositor produces copies of all the data and documentation files that make up the digital resource
- Depositor completes a data and documentation transfer form
- Depositor completes a catalogue form
- Depositor completes, and signs, a licence form
- Data, documentation and forms are transferred (physical media or network) to the relevant AHDS Centre
- AHDS Centre sends acknowledgement of receipt
- AHDS Centre prepares resource for preservation
- Resource is loaded into AHDS digital repository
- AHDS Centre sends confirmation that resource has been successfully archived

These steps are typically carried out in a semi-manual way, involving effort from both the depositor and staff at the AHDS. The hybrid model has been built around an assumption that while these tasks are necessary, they must be formalised to a greater extent, so that ingest can be automated.² In the sequence listed above, the key steps to automate are the transfer of metadata and data from the depositing institution to the AHDS. To automate these actions, the AHDS must understand how to establish communications across the network with each web accessible resource to be deposited, determine what metadata and data need to be transferred, extract this content and then confirm the transfer has taken place successfully. There are many possible ways of doing these tasks. The model assumes that an *ingest engine* at the AHDS will communicate with a *collection access point*, located at the institution, to manage the ingest process. Both the ingest engine and the collection access point can be thought of as a coordinated set of procedures, technical protocols and software, along with the staff needed to implement them. The exact implementation and operation of the ingest engine and collection access point will depend on the unique situation of each depositing institution.

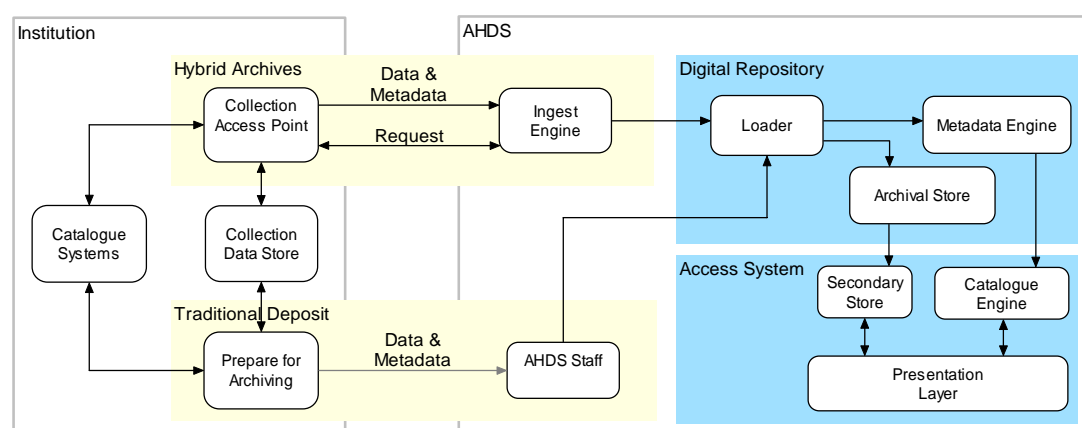


Figure 1: Full and Hybrid Deposit

² Although some full deposits are transferred to the AHDS electronically (as noted in work package 4, for example, 20 percent of data transferred to the Archaeology Data Service is done so electronically), the majority of deposits are still received on physical media. Electronic transfers that do occur are not highly automated, and they generally involve the use of email, FTP or HTTP downloads.

The Ingest Engine must provide the following minimum functionality:

- Support for harvesting metadata
- Support for transferring data
- File format conversion tools
- File integrity checking tools
- Preservation metadata extraction tools
- Collection access failure monitoring

The collection access point must provide the following minimum functionality:

- Support for publishing metadata to be harvested
- One or more methods for transferring data across the internet (e.g. FTP)

In some cases it will be desirable not to have to harvest the entire resource at regular intervals, but rather to harvest only when the resource is updated. In addition to the two minimum requirements, it will be useful if the collection access point can include a means of sending alerts to the ingest engine when data or metadata have been changed.

Combined, the collection access point and ingest engine will provide the ingest functionality defined in the OAIS ingest functional entity. The Submission Information Packages created by the ingest engine will be equivalent to those created using a semi-manual approach at each AHDS Centre.

Deposit Agreement

Before a digital resource can be accepted under the hybrid model, the AHDS and the host institution must formalise their relationship. The greater technical interoperability required by the hybrid model makes this stage more complex than it is for full deposits, which can generally be completed in a single data transfer session.

Before metadata harvesting and data transfer begin, the AHDS must establish that the depositing institution can meet the minimum technical requirements for running a hybrid deposit, and the institution and the AHDS must further decide on a particular set of technical options that will be used. These agreements will need to be written down as a schedule to the deposit licence signed by the depositing institution. Clarity over the agreed details of the hybrid deposit will be useful if the institution fails to make the digital resource available in the future, or if the AHDS needs to change the technical solution.

The AHDS uses a standard licence for most deposits[3]. The hybrid model of deposit uses a modified form of this licence that ensures that the AHDS can continue to preserve and make available any collection that the depositing institution ceases to support. Modifications include:

- Clauses that allow for the conversion of a hybrid deposit into a full deposit if the institution ceases to make the resource available
- A schedule that details the technical interoperability agreed between the AHDS and the institution
- Clauses that allow for the AHDS to require changes in the methods used for technical interoperability

Scenarios

The minimum requirements for an institution to use the hybrid deposit are the ability to publish metadata using OAI or Z39.50, and an agreement to provide data in a form and format agreed with the AHDS. Within these requirements, however, a hybrid deposit could follow a number of different scenarios depending on the available time and expertise at the institution and the amount of effort the AHDS is prepared to expand on the particular deposit. Tables 3 and 4 outline the ways in which a

hybrid deposit might be carried out given either high or low institutional involvement, and low, medium or high effort from the AHDS.

AHDS Effort (in additional to normal ingest procedures)		
low	medium	high
METADATA MAPPING AND SEARCH		
No mapping, full text search of original metadata is provided.	AHDS develops a Dublin Core mapping to support basic structured searching in addition to full text search	<i>Assumes the institution is using a recognised metadata standard.</i> The AHDS develops a full mapping to support advanced structured searching in addition to full text searching
DATA AND METADATA TRANSFER		
<i>Assumes the institution has already installed an OAI-PMH or Z39.50 provider for the collection.</i> The AHDS retrieves metadata via OAI or Z39.50 and data is retrieved by web harvesting. Transfer is initiated by AHDS following a fixed timetable.	<i>Assumes the institution has already installed an OAI-PMH or Z39.50 provider for the collection.</i> The AHDS retrieves metadata via OAI or Z39.50 and data is retrieved by customised web queries, designed to trawl underlying data sources. Transfer is initiated by AHDS following a fixed timetable.	AHDS assists institution to install necessary software for metadata harvesting and direct transfer of data from backend server applications such as databases. Transfer is initiated by AHDS following a fixed timetable.
INGEST		
Automatic tools at AHDS generate basic technical metadata and export received files to preferred formats for preservation. Errors are logged.	Automatic tools at AHDS generate basic technical metadata and export received files to preferred formats for preservation. Errors are investigated by AHDS staff.	AHDS staff customise automatic tools to create detailed technical metadata and export files to preferred formats for preservation. Errors are corrected by AHDS staff.

Table 3: Hybrid Deposit Scenarios with Low Institutional Effort

AHDS Effort (in additional to normal ingest procedures)		
low	medium	high
METADATA MAPPING AND SEARCH		
Institution works to convert or map metadata to AHDS standards, allowing full text and structured searching, using freely available information on the AHDS site.	AHDS provides advice to support institution as it works to convert or map metadata to AHDS standards, allowing full text and structured searching	AHDS develops a full metadata mapping in consultation with the institution to support advanced structured searching in addition to full text searching
DATA AND METADATA TRANSFER		
The institution installs software as recommended by the AHDS to support harvesting of metadata and transfer of data. Transfer is initiated by AHDS following a fixed timetable.	The institution installs software as recommended by the AHDS to support harvesting of metadata and transfer of data. Or, the institution alters workflow and devotes staff time to transferring collections manually. Transfer is initiated as appropriate by AHDS or institution	AHDS and institution cooperation to customise an automated transfer procedure for metadata and data. Transfer is initiated as appropriate by AHDS or institution
INGEST		

Technical metadata is created by the institution, and is used by automatic tools at the AHDS to determine preferred formats for preservation. Automatic tools export files to preferred formats and errors are logged.

Technical metadata is created by the institution, and is used by automatic tools at the AHDS to determine preferred formats for preservation. Files are validated on receipt by AHDS. Errors are corrected by AHDS staff and reported to institutional staff for more permanent solutions.

AHDS provides institution with tools to create detailed technical metadata, and to export files to preferred formats for preservation. Institution transfers exported, and validated, files to AHDS. Tools are modified by AHDS staff, or data is modified by institutional staff to correct errors.

Table 4: Hybrid Deposit Scenarios with High Institutional Effort

Responsibility for Transferring Metadata and Data

Transferring Metadata

A key point in the operation of the hybrid model is establishing whether the AHDS or the institution is primarily responsible for arranging the transfer of updates to metadata and data to the AHDS digital repository. Based upon the ability of the institution and the AHDS to identify updates and initiate the transfer process, one of three approaches could be taken to harvesting metadata:

1. **Push:** The institution initiates the transfer of metadata, and the metadata is 'pushed' to the AHDS using methods such as secure FTP or email. This approach is fairly unlikely to be used because OAI-PMH and Z39.50 do not support it (although enhancements have been proposed for an "Enhanced Kepler Framework" or OAI-P2P).
2. **Pull (timetabled harvesting):** The AHDS harvests metadata using OAI-PMH or Z39.50 according to a timetable. This is the most likely approach. Several techniques could be used to restrict the harvest to only updated metadata: 1) date stamp, 2) SETS (a user-defined key that identifies new metadata to be harvested in the OAI), or 3) some other form of generic index.
3. **Push/Pull (harvest on alert):** Metadata is harvested via OAI-PMH or Z39.50 as in the pull model, but each harvest is started when the institution sends an alert to the AHDS, which may also specify which records need to be retrieved. This approach would be appropriate where large volumes of irregularly modified metadata need to be harvested.

Either a pull or a push/pull approach are most likely to be used. These approaches reduce the demands on the institution, and ensure that the AHDS retains control over the process.

Transferring Data

Ideally, the process of transferring data will be automated and metadata led, so that analysis of the metadata harvested from the institution will indicate which parts of the underlying data have changed and need to be retrieved. However, there are situations where changes made to the data may not be reflected in the metadata.

Unlike metadata, options for transferring data are less clear-cut. The basic requirement is for an internet accessible machine-to-machine service that provides functionality to send requests for data and handle packages of data sent in response to requests. This might take the form of a webpage harvester, a scripted FTP program, a utility capable of sending customised query strings to the institutional resources web interface, or a number of other options. Ideally data transfer could make use of the existing facilities of the web interface to the institutions digital resource. For example, small updates might be transferred as UUENCODED data embedded in a XML file harvested via OAI-PMH, but additional software may need to be installed.

Operational View of the Transfer Process

The AHDS should possess suitable infrastructure to harvest and store metadata and data on a regular basis, as well as the appropriate skills to understand the institution's set-up arrangements. Figure 2 provides a high-level overview of the ingest engine functionality needed to manage the transfer of metadata and data between the depositing institution and the AHDS.

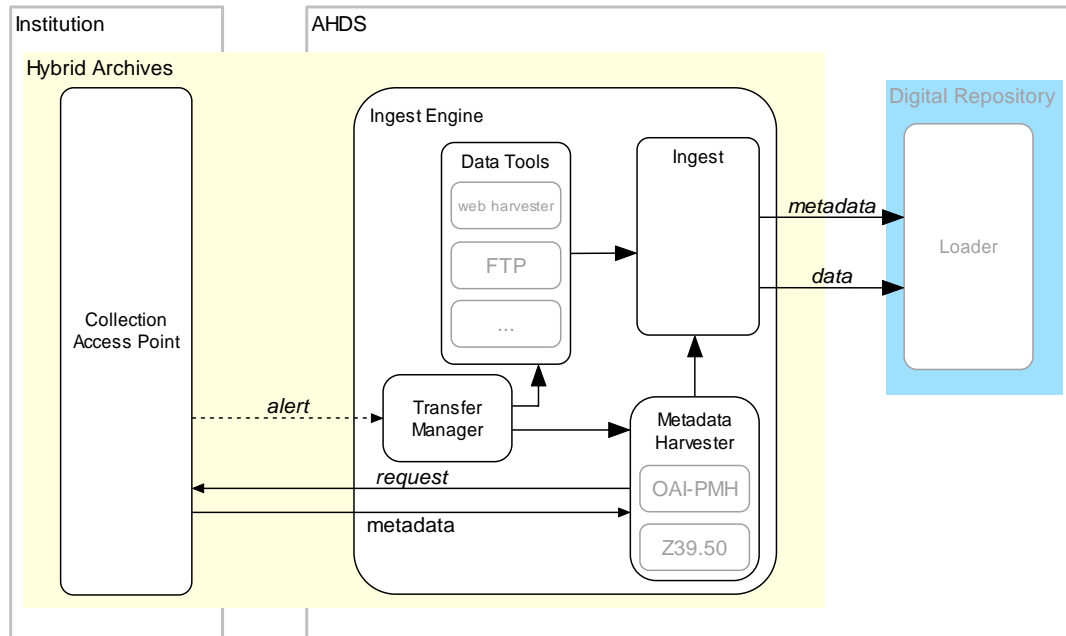


Figure 2: overview of the Ingest Engine

As a minimum, the Ingest Engine must: support the harvesting of metadata; support the transfer of data; convert files to a suitable file format; perform file integrity checks; generate preservation metadata; and monitor for collection failure. This functionality is provided by four entities:

1. **Transfer Manager:** The transfer manager is a database that stores administrative information needed to support, monitor and initiate the process of retrieving data and metadata from the institution. Necessary information may include:
 - Authentication details – passwords and encryption
 - Harvesting timetable– The regular intervals when data and metadata should be harvested
 - Access methods for each collection
2. **Data Tools:** The data tools handle the process of retrieving data from the institution. The data tools must be able to interact with differing institutional systems so the download manager will require support for multiple methods of access.
3. **Ingest:** Ingest provides the automated ingest processes to prepare both data and metadata for loading into the AHDS digital repository. This includes the automated conversion of file formats; creation of administrative and preservation metadata; and some method of error checking and monitoring to repair problems or request assistance.
4. **Metadata Harvester:** The metadata harvester reacts to requests from the transfer manager to harvest metadata using either OAI-PMH, Z39.50

In order to harvest data and metadata from the catalogue system and collection data store, the third-party institution must possess a Collection Access Point (figure 3).

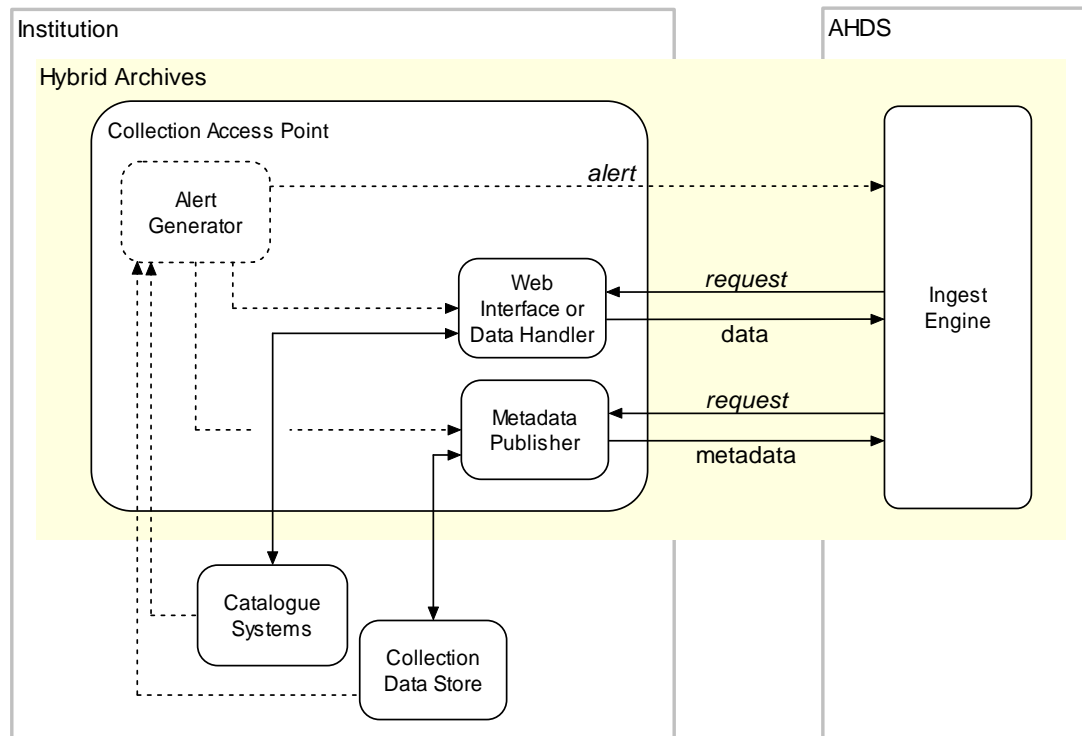


Figure 3: Overview of the Collection Access Point

To enable the harvesting of data & metadata, the Catalogue Access Point must make the following three services available through the Internet (and observe appropriate security precautions):

1. **Metadata Publisher:** An Internet accessible service for publishing metadata, either an OAI-PMH client, an OAI static repository and static repository gateway or a Z39.50 target.
2. **Alert Generator:** A tool that monitors the state of the data and metadata and sends an alert to the AHDS if it is changed. This functionality is unlikely to exist in the collection, and would have to be installed, therefore it should be viewed as optional.
3. **Web Interface or Data Handler:** An Internet accessible machine-to-machine service, that can receive requests for data and return the relevant data. A standard web server dealing with HTTP calls might provide this functionality, or it could be provided using an FTP client, a database connection or other technique. Ideally, existing facilities to deliver the collections content to users could be used, but it may need to be installed during the establishment stage of the Hybrid Archive model.

Subsidiary Requirements

- Name of primary and secondary contacts for each collection
- Tools for monitoring changes to webpages
- Procedures for recording a collection failure and informing the institution

Conversion to Full Deposit

The Hybrid Archive model assumes an on-going relationship between the AHDS and the depositing institution. However, the primary purpose of the hybrid deposit model is to ensure that if this relationship fails, the deposited resource survives in a usable state and can be preserved and made available by the AHDS. The relationship between the AHDS and the depositing institution may end for a number of reasons:

- The institution removes online access to the resource
- The institution is unable or refuses to meet changes in the minimum technical requirements required to implement the hybrid deposit model
- The institution explicitly revokes the licence governing the deposit
- The institutions IPR in the collection is challenged
- The collection is deemed complete, and no more updates are anticipated

If one of these situations occurs, then the AHDS must either move to convert the collection into a full deposit, or remove the collection from the AHDS digital repository. Establishing what has happened, and whether it is a permanent change or a temporary problem may take some time, so the first step will be to suspend updates to the collection, allowing time for the situation to be clarified.

The main threat that the Hybrid Archives model is being developed to counter is the loss of online access to a collection. This may occur for a range of reasons, and it will be important to establish the exact cause, particular establishing if the loss of access is permanent or only temporary. Problems ranging from hardware downtime to domain updates, may cause a collection to fail temporarily and it may be difficult to decide when a permanent failure has occurred.

A range of checks will be necessary:

- Check records for past downtime, notes on any anticipated downtime
- Attempt to telephone someone at the institution in question. If the manager cannot be contacted, try the Webmaster or other employees
- If there is no answer on the telephone, attempt to contact them via email. If there is a problem with their web site it is likely a problem will also be encountered when sending email to that domain. If available, contact employees using an alternative (possibly home) email address if one is known
- Depending upon the location of the institution, it may be useful to visit the building where they are located

Once the failure of a collection is confirmed, a waiting period will be necessary in which efforts should be made to get the collection reinstated.

- [1] *Managing Digital Collections: AHDS Policies, Standards and Practices*, Consultation Draft, December 1998, Neil Beagrie & Dan Greenstein
- [2] <http://ahds.ac.uk/collections.doc>
- [3] <http://www.ahds.ac.uk/depositing/licence.htm>