



An OAIS Based Approach to Effective Long-term Digital Metadata Curation

Arif Shaon (Correspondence Author)

Centre for Advanced Computing and Emerging Technologies (ACET), The University of Reading

PO box 68, Whiteknights Campus, Reading, UK

E-mail: a.b.s.shaon@rdg.ac.uk

Andrew Woolf

Science And Technology Facilities Council (STFC)

Rutherford Appleton Laboratory, Didcot, UK

E-mail: a.woolf@rl.ac.uk

The research is jointly financed by the University of Reading and the Science And Technology Facilities Council (STFC)

Abstract

Metadata has the proven ability to provide information necessary for successful long-term curation of digital objects. However, without curation metadata itself may deteriorate in terms of its quality and integrity over time. Therefore, a digital curation process needs to incorporate the curation of metadata along with that of data in order to ensure the accurate description of data over time. Unfortunately, no comprehensive method for effective curation of metadata for long periods of time is known to exist at present. Even the Reference Model for Open Archival Information System (OAIS), despite being the most comprehensive and widely adopted framework for long-term data preservation, fails to address the requirements of long-term metadata curation in a comprehensive and unambiguous manner. This paper presents an approach to efficiently curating digital metadata over the long-term that is achieved through articulating the metadata curation related ambiguities of the OAIS Reference Model. The approach essentially involves the use of a “Metadata Curation Model”, which is a specialised edition of the “Data Management” module of the OAIS Reference Model, dedicated to the purpose of long-term metadata curation.

Keywords: Metadata, Curation, Preservation, OAIS

1. Introduction & Motivation

Exponential increases in computing power and communication bandwidth have resulted in a dramatic rise in the volume of generated and published data within various complex information domains. This increasingly large volume of data needs to be preserved and made highly available (i.e. curated) over substantially long-periods of time in order to assist in high quality future research and experiments in both same and cross-discipline environments, as well as other productive uses of the data. However, the rapid growth of related technology and increased flexibility in their use also create a significant imbalance between the capacity for data generation and (long-term) data curation as the former is advancing significantly faster than the latter. And this imbalance in effect, poses the major challenge toward ensuring efficient and continued use of valuable data resources with their quality and integrity intact over the long-term.

Under the challenges set by the task of successful long-term data curation, the word ‘Metadata’ (i.e. further information about data) is becoming increasingly prevalent, with a growing awareness of the role that it can play in capturing information necessary for efficient functioning of different curation operations, such as data preservation and provenance tracking. For data preservation in particular, metadata can be used to record information required to reconstruct or at the very least understand the reconstruction process of digital resources on future technological platforms. However, without curation, metadata itself may deteriorate in terms of its quality and integrity over time. Therefore, a digital curation process needs to incorporate curation of metadata along with data in order to ensure an accurate description of data over time.

Over the past few years, several organised and arguably successful endeavours (e.g. NEDLIB, 2000 and CEDARS, 2002) have been made in order to find an effective solution for successful long-term data preservation. However, the territory of long-term metadata curation, although increasingly acknowledged, has yet to be conquered. In fact, in

most digital preservation or curation motivated workgroups and projects, the necessity of long-term metadata curation is deemed secondary, mainly due to the lack of awareness of the criticality of the problem. Even the Reference Model for Open Archival Information System (OAIS, 2002), despite being the most comprehensive and widely adopted framework for long-term data preservation, fails to address the requirements of long-term metadata curation in a comprehensive and unambiguous manner. As a result, no comprehensive method for effective curation of metadata for long periods of time is known to exist at present. This paper presents an approach that aims to fill the void for an efficient strategy for curating digital metadata over the long-term. The approach involves the use of a “Metadata Curation Model”, which is a specialised version of the “Data Management” module (OAIS, 2002) of the OAIS Reference Model, dedicated to the purpose of long-term metadata curation.

2. Metadata Defined

In light of its acknowledged role in the organisation of and access to networked information and importance in long-term digital curation, metadata may be defined as structured and standardised information that is crafted specifically to describe a digital resource, in order to aid the intelligent and efficient discovery and retrieval of that source, accurate verification of its integrity (e.g. provenance tracking) as well as its apposite use and effective preservation over time. In the context of digital preservation, information about the technical processes associated with a data preservation technique is an example of metadata.

3. Digital Metadata Curation Defined

As mentioned earlier, a digital curation process needs to incorporate curation of metadata along with data in order to ensure an accurate description of data. It is therefore necessary to have an understanding of the underlying notion of the term “Digital or Data Curation” before attempting to define Metadata Curation. The phrase “Digital Curation” has different interpretations within different information domains. For example, in the museum domain, which is one of the oldest curation environments, data curation covers three core concepts – data conservation, data preservation and data access. Access to data or digital information in this sense may imply preserving data and making sure that the people to whom the data is relevant can locate it - that access is possible and useful. Another interpretation of the phrase “Data Curation” may be the active management of information, involving planning, where re-use of the data is the core requirement (Macdonald and Lord, 2002).

Therefore, from a generic standpoint, long-term data or digital curation can be defined as the continuous activity of managing, improving and enhancing the use of a digital object (i.e. data) as well as its preservation over its life cycle and over time for current and future generations of users. This is to ensure that the suitability of a data object is sustained for its intended purpose or range of purposes, and it is available with its quality and integrity intact for efficient discovery and apposite re-use over the long-term.

In light of the above construal of digital curation, the term “Metadata Curation” may be defined as an inherent part of a digital curation process for the continuous management (which involves creation and/or capturing as well as assuring overall integrity of metadata amongst other things) and preservation of metadata records over their life cycles. This is primarily to ensure the suitability of metadata for aiding the long-term curation of a digital resource that it refers to, by facilitating the intelligent and efficient discovery and retrieval of that resource, along with accurate verification of its integrity (e.g. provenance tracking), its apposite (re-) use and effective preservation over the long-term.

4. Principal Requirements of Effective Digital Metadata Curation

The efficacy of metadata curation significantly relies upon satisfying a number of requirements. Although metadata curation requirements may be quite different according to the type of data described, the information outlined below attempts to provide a general overview of the main requirements.

- **Metadata Standards:** The very first step to successful long-term data and metadata curation is to employ a curation-aware metadata standard(s) or format that provide(s) necessary elements to capture sufficient information about both a data object and its associated metadata. Examples of such information include Representation Information (RI), annotations made to both data and metadata, information about changes made to data and metadata, amongst other things. Of particular note is the Representation Information about a data object, which is defined as the information required to enable access to the preserved digital object in a meaningful way (OAIS, 2002). The use of RI can be recursive, especially in cases where meaningful interpretation of one RI element requires further RI. This recursion continues until there is sufficient information available to render a digital object in a form the user base can understand.
- **Metadata Preservation:** Metadata curation requires metadata to be preserved along with data in order to ensure its accurate description over time. Therefore, it is necessary to devise or use a suitable long-term metadata preservation strategy that is also flexible for addition of further requirements.

- **Metadata Quality Assurance:** A long-term curation process should effectively ensure well-structured metadata records with an accepted (by the community or domain concerned) level of quality, notwithstanding any forms of evolution or changes in related technology or metadata requirements of the organisation(s) concerned. In general, quality of a metadata record is measured by the degree of consistency with and/or accuracy in reference to the actual dataset and conformance to some agreed standard(s). Therefore, appropriate quality assurance procedures or mechanisms need to be in place to eliminate any quality flaws in a metadata record and thereby ensure its suitability for its intended purpose(s).
- **Metadata Versioning:** It is essential for a metadata curation system to be able to distinguish between metadata in different states, which arise and co-exist over time by suitably versioning metadata information.
- **Metadata Annotation:** Annotation is a widely practiced means of adding value to data as well as establishing collaborative links between data producers and users. An efficient long-term metadata curation strategy will, therefore, need to facilitate annotation of both data and metadata (by data consumers/users), preserve and curate annotations made over the long-term.
- **Audit Trailing & Provenance Tracking:** A metadata curation process should ensure recording of information with the required granularity, and facilitate necessary means of tracking any significant changes (e.g. provenance change) to both data and metadata over their life cycles and determining their quality and integrity.
- **Metadata Search-ability:** Metadata needs to remain available and searchable to the potential users of the data objects or resources that they describe in order to aid the appropriate use of those data objects or resources over time. Therefore, search-ability of metadata in convenient and efficient manners is regarded as a crucial factor in successful long-term metadata curation, hence a part of its remit.
- **Metadata Policy:** A set of broad, high-level principles that form the guiding framework within which the metadata curation can operate, must be defined.
- **Access Constraints & Control:** Appropriate security measures should be adopted to ensure that the metadata records have not been compromised by unauthorised sources, thereby ensuring overall consistency in the metadata records.

5. The OAIS Reference Model as a Framework for Long-term Metadata Curation System

In the complex realm of digital preservation and curation, the Reference Model for an Open Archival Information System (OAIS) (OAIS, 2002) is perhaps the only standardised effort that attempts to provide answers to virtually all questions related to long-term digital curation and preservation. Unsurprisingly, it has proliferated rapidly through the digital preservation community and has been explicitly adopted by, or at least informed, many prominent digital preservation initiatives (e.g. NEDLIB, CEDERS, etc.). However, a problem with the OAIS model is the variable detail of the answers that it provides. While the OAIS provides very unambiguous and in-depth answers to some questions, others are only touched upon, in some cases requiring one to even assume the answers. Of particular note is the considerable level of ambiguity in description of the metadata curation technique presented in the model.

When attempting to build a metadata curation system for an OAIS-compliant digital archive, one faces the problem of having very vague definitions of different metadata curatorial functions. In fact, in order to outline the functional requirements of a metadata curation system, it would be necessary to extract and in some cases assume the definitions of such functions from some generalised description of related preservation functions as described within the OAIS model. As an example, the Data Management entity (Figure 1) of the OAIS model can be considered. This (according to OAIS, 2002) aims to provide necessary functions for populating, maintaining and accessing Descriptive Information, i.e. metadata about the digital objects being preserved. Based on the definition of this entity, one would naturally infer that “maintaining” metadata essentially includes curating it. However, when it comes to implementing such an entity in a curation system, owing to very vague definitions in the OAIS, one would also need to make further assumptions to develop a complete list of concretely defined functional requirements of metadata curation (e.g. metadata versioning, validation and so on).

Furthermore, while the OAIS model provides a generic overview of the ingest function, it does not however describe how and at what stage syntactic and semantic validity of metadata should be checked during the ingest process. Also, the OAIS does not specify how one would go about ensuring interoperability and coherence of metadata irrespective of the formats that it is submitted in, i.e. if any form of translation or mapping between different formats would be required, how and at what stage it should be done.

In addition, the model provides a macro view in which information objects are migrated to a future technological platform as part of the preservation measure employed, but it does not mention anything pertinent to the fact that metadata itself may need to be migrated to newer formats, versions and platforms to ensure its longevity and usability. In terms of handling updates and changes to both data and metadata, the OAIS model does not seem to take into

account (at least not in direct terms) how the system would facilitate annotating both data and metadata as well as storing, searching and retrieving the annotations made to data and/or its metadata.

The answer to this predicament is that the OAIS model is part of the user requirements that is only intended to serve as the starting point for the development process for an OAIS archive. The OAIS specification, in fact, clearly states that it is not a guideline for an implementation or a design. Consequently, building a dedicated system for metadata curation needs further specific requirements to be added on to the OAIS model, as described in the following section.

6. The Metadata Curation Model

As underlined in the previous section, the OAIS model contains certain ambiguities about metadata curation related functions. Therefore, having to build a dedicated OAIS-compliant system for metadata curation needs further specific requirements to be added on to the OAIS model, especially to its “Data Management” entity, which is essentially responsible for managing (i.e. curating) metadata in an OAIS archive. In other words, a specialised edition of the OAIS model is required for the design and implementation of an efficient long-term metadata curation system. The Metadata Curation Model (MCM) attempts to fulfil that requirement by articulating the metadata curation related ambiguities of the OAIS model and refining its “Data Management” entity, and thus making it metadata curation focused. From this perspective, the MCM should be regarded as *an OAIS-based solution to long-term metadata curation*.

Furthermore, long-term metadata curation requires a model that is efficient and precise in reflecting all core requirements of metadata curation (see section 4) as well as being extensible and adaptable to incorporate any future requirements. The Metadata Curation Model presented in this section endeavours to accomplish these objectives.

6.1 Overview of the Model

The curatorial functions designed in the MCM include metadata ingest, metadata versioning, metadata quality assurance, annotation of data and metadata, preservation of metadata, access to (e.g. querying, searching) preserved metadata, migration of metadata to new formats and tracking provenance of metadata. In addition, the use of a curation-aware metadata format (see section 4) is also incorporated into the design of the model and is essential for efficient and optimal execution of its curatorial functions.

The model is only focused on the curation of metadata and does not assume the responsibility of curation of the data that the metadata describes. As Figure 2 illustrates (compared to Figure 1), the model can be seen and implemented as one of the functional entities of the OAIS reference model.

A brief description of each of the entities in Figure 2 is given as follows:

6.1.1 Data Ingest

The **Data Ingest** entity directly refers to the **Ingest** (Figure 1) entity of the original OAIS model, with one significant addition. In short, the entity provides functions to accept **Submission Information Packages (SIPs – Note 1)** from Producers, extracts metadata and its corresponding meta-metadata from the actual digital object and prepares them for preservation and curation. Metadata together with its corresponding meta-metadata constitute a **Metadata Submission Package (MSP)**, which is an addition to the original OAIS design of the Ingest entity. At this stage, meta-metadata in the MSP may contain various information about a corresponding metadata record, such as information about metadata creator(s), metadata publisher(s), metadata format (e.g. Dublin core), metadata provenance, existing annotations made to the metadata and so on. It is not necessary for meta-metadata to have further information as both metadata and its corresponding meta-metadata are assumed to be in the same format (e.g. XML, Text) and changes made to meta-metadata should only be reflective of changes made to the metadata, which it refers to. Therefore, curation mechanism(s) applied to metadata should also suffice for its corresponding meta-metadata without the need for any further information. The MCM requires a curation-aware metadata format (Section 4) as the underlying format of the metadata and meta-metadata in a MSP.

In addition, the Data Ingest module is responsible for assigning unique session identifier to each data/metadata submission request, thus enabling data objects and their corresponding metadata records to accurately reference each other from their respective entities (i.e. **Archival Storage** for data objects and **Metadata Curation** entity for metadata record), during the submission. This is particularly useful when a curation system is dealing with multiple data submission requests at the same time as it eliminates the risk of a metadata record referencing a data object that it does not describe or vice versa. The Data Ingest module could also have suitable means for checking or scanning submitted files for corruption and virus infection.

6.1.2 Archival Storage

As with the Data Ingest entity, the **Archival Storage** is also a direct reference to the Archival Storage entity of the original OAIS model. Functions within this entity include receiving digital objects from the data ingest module and adding them to permanent storage, managing the storage hierarchy and migrating preserved digital objects to new media or platforms.

6.1.3 Preservation & Curation Planning

As a revised version of the OAIS-defined **Preservation Planning** module, the **Preservation & Curation Planning** entity monitors and provides periodic recommendations for both data and metadata preservation to ensure they remain accessible to the User Base (Note 2) over the long-term. These recommendations cover preservation techniques, metadata standards, curation policy and so on. This entity is also responsible for developing detailed Migration plans, software prototypes and test-beds to enable implementation of successful migration of both data and metadata.

6.1.4 Metadata Curation

The **Metadata Curation** entity essentially represents the Metadata Curation Model within an OAIS system/archive by implementing a range of different functions to efficiently curate metadata over the long-term. This entity also elaborates the vaguely defined metadata curation related functionalities (Section 5) of the Data Management entity as outlined in the original OAIS model and presents a complete list of suitably defined functional requirements of metadata curation. Figure 3 takes a closer look at the Metadata Curation entity in Figure 2.

The **Metadata Ingest** entity in Figure 3 is essentially the passageway for metadata to curation and preservation. In short, the entity receives a Metadata Submission Package (MSP) or **Metadata Update Package** (MUP – Note 3) from the Data Ingest entity, isolates meta-metadata from metadata and finally forwards them both to the **Metadata Quality Assurance** (QA) entity for validation. The isolation of metadata from its corresponding meta-metadata at this stage is only essential if they both are not adhering to the same format. The MCM supports two generic ways in which metadata in an MSP can be captured and ingested into the system:

- As pre-created manually (i.e. by human), i.e. the SIP contains pre-created metadata files.
- Using a combination of automatic (i.e. extracted from the data object using tools external to the system) and manual (i.e. created at the time of the SIP submission through the submission interface, which could be a web form or a standalone tool provided by the curation system) metadata creation methods to ensure adequacy and accuracy of metadata and thereby minimising metadata creation costs and efforts.

A similar approach should be applicable to creation of meta-metadata and ingesting it into the system. Figure 4 illustrates the functions of the Metadata Ingest entity.

The primary task of the **Receive MSP/MUP** function (Figure 4) within the Metadata Ingest module is to receive Metadata Submission Packages from the Data Ingest entity or **Metadata Update Packages (MUP)** from the Administration entity (section 6.1.6) and put them forward for quality assurance in the Metadata Quality Assurance entity. On receipt and if necessary, an MSP or MUP is disassembled into its constituent metadata and meta-metadata, both of which are then fed into the QA entity for validation.

This function also receives the outcomes of quality assurance operations (returned by the QA entity) on metadata and meta-metadata and informs their source entity (i.e. Data Ingest in case of MSP and **Administration** for MUP) accordingly. In the case of a MSP, should either metadata or meta-metadata fail to pass any of the quality checks, i.e. if the QA returns negative results, a re-submission request for the MSP is sent to the Data Ingest entity, which then forwards the report to the Producer entity. In any case, a full report detailing the outcomes of different functions of the Metadata Ingest entity, such as MSP/MUP disintegration and quality assurance, is sent to the Administration entity. The report sent to the Administrator entity is also used (along with a relevant session identifier) by the Archival Storage (Section 6.1.2) entity to ensure that only data objects with valid metadata records are stored for curation.

The **Extract Meta-Metadata** function in effect assists the Receive MSP/MUP function in extracting meta-metadata from metadata in a MSP/MUP if necessary (see above). This function is also responsible for subjecting meta-metadata to different quality checks through the QA entity. In fact, the tasks of handling Metadata and its associated meta-metadata are isolated and allocated to Receive MSP/MUP and Extract Meta-Metadata function respectively for overall greater efficiency but could well just be handled by one function if design simplicity is desired.

The **Metadata Quality Assurance (QA)** entity (Figure 3) is responsible for ensuring overall quality and validity of submitted metadata. Figure 5 presents the functions of the QA entity.

The **Metadata Crosswalking** function (Figure 5) ensures that submitted metadata and meta-metadata conform to the format(s) supported by the curation environment. This essentially involves translating or transforming metadata in unsupported format(s) to format(s) that is/are supported. As there is always the danger of potential data loss in mapping between metadata in different formats, i.e. “metadata cross-walking”, the need for such an operation will largely depend on the related policy of the system. For example, if the system's policy was to support only one particular metadata format, there would be no need for any metadata crosswalk. However, it would also imply that any existing metadata (in non-supported format(s)) would need to be rewritten in the supported format in question before it could be accepted by the system, which might be deemed impractical where a large amount of metadata is involved. More importantly, it would result in sacrificing interoperability with other related curation systems and metadata

formats.

A solution to this problem would be to maintain suitably formulated pre-created mapping rules for the metadata translation within the repository in order to minimise data loss. However, such a solution would also imply that every time any of the supported metadata formats was to undergo any significant changes, the mapping rules for that format would need to be updated, which would essentially require detailed examination of the new version of the format in question to ensure accuracy. In other words, the archiving or curatorial body concerned would need to have a dedicated team who would be responsible for monitoring changes in supported metadata formats and calculating/updating mapping rules. While this should not be a problem if the body in question has the necessary financial and technical resources to facilitate this over the long-term, an organisation with comparatively lower or limited curation budget might not find it cost-effective.

Nevertheless, while successfully transformed metadata and meta-metadata are passed on to the **Structural Validation** function, failure in metadata cross-walking results in both metadata and meta-metadata being discarded. In both cases, the final outcome of the transformation is reported to the **Generate Quality Assurance (QA) Result** function, which then re-directs it to the Metadata Ingest (Figure 4) entity. Metadata and meta-metadata in supported format(s), however, bypass this function and go straight to the Structural Validation function.

The **Structural Validation** function (Figure 5) checks syntactical or structural validity of metadata records (and associated meta-metadata) against the corresponding metadata format(s). Ideally, this function should be fully automated. This function also sends a report to the Generate QA Result indicating the outcome of the validation. While structurally invalid metadata and meta-metadata are discarded, structurally valid metadata records are forwarded to the Semantic Validation function.

The **Semantic Validation** function (Figure 5) facilitates semi-automatic ways of checking whether the values assigned to the elements in structurally valid metadata records comply with the actual content of the data object. This function could also include metadata cleansing in order to remove any noise or anomalies in metadata records (e.g. correction of spelling mistakes, grammatical errors) and thereby maintain the desired level of consistency across all metadata records being preserved. This function is also expected to make use of some controlled vocabulary (Note 4), if applicable, in order to check semantic validity of values in metadata records. Ideally, a curation system would maintain a controlled vocabulary server for the system's principal metadata format. For other supported metadata formats, however, the system could maintain a database of information (e.g. server URL, port number, etc.) required to connect to and use the appropriate vocabulary server. Alternatively, the users submitting metadata records could be provided with the facility for specifying such information at the time of submission.

In case of extremely erroneous and inconsistent metadata or meta-metadata records, which fail semantic validation, the function will be forced to discard both metadata and its corresponding meta-metadata and (as with the Structural Validation function) send an appropriate report to the Generate QA Result function. A semantically valid metadata record makes its way to the **Record Quality Assurance Event Info**, where associated meta-metadata is updated with information about different QA operations that the metadata has been subjected to.

The **Generate Quality Assurance (QA) Result** function (Figure 5) collates information about the outcomes of different QA processes, such as cross-walking, structural validation and semantic validation, generates report based on it and sends the report to the Metadata Ingest entity (Figure 4). The information collected by this function is also used as the QA Event Info (Note 5), which is recorded in the meta-metadata associated with the metadata records.

The **Record Quality Assurance Event Info** function records information about different QA processes (e.g. cross-walking) that metadata is subjected to, in its corresponding meta-metadata. QA Event info includes description of a process, changes made to metadata, tools used, date and time of the occurrence of the process and so on. QA Event info is essentially obtained from **Generate QA Result** function. Updated Meta-metadata is forwarded to the **Generate Metadata Versioning Package** function.

The **Generate Metadata Versioning Package** is the final QA stage for both metadata and meta-metadata before they are ingested into the Versioning entity (Figure 6). For successfully validated (and cross-walked if necessary) metadata and associated meta-metadata, this function obtains Representation Information (RI) for both the digital object (if it is not already included in the metadata) and its corresponding metadata from a trusted Representation Information Registry and updates both the metadata record (s) and its associated meta-metadata respectively with it. The task of acquiring RI for a digital object could also be performed at the time of, or before, data ingest and more practically before the metadata ingest as that is when file format and/or rendering software related information is computed (ideally using a suitable tool/software, such as DROID – Note 6) for the object. File format and software related information (e.g. extension name or rendering software name) is normally what is used by RI repositories to determine RI for digital objects. An example of such RI repository is the PRONOM technical registry (PRONOM, 2007). The use of a trusted repository for RI ensures authenticity and accuracy of the RI obtained, which in the long run ensures accurate

interpretation and use of the digital object in question.

Successfully validated and updated (with RI and QA event info) metadata records and meta-metadata collectively form a **Metadata Versioning Package (MVP)**, which is then forwarded to the **Metadata Versioning** entity. The structure of an MVP at this stage should be similar to that of an MSP.

The **Metadata Versioning** module as depicted in Figure 3 is responsible for assigning unique version numbers to metadata records (both newly submitted and updated versions of existing records) to represent their states at particular times. Figure 6 pictorially presents different functions of the Metadata Versioning Entity.

The **Receive MVP** function (Figure 6) accepts MVPs from the Quality Assurance function. For MVPs consisting of separate files containing metadata and meta-metadata, this function feeds the file containing metadata into the **Process Metadata Versioning** function, while the associated meta-metadata file moves across to the **Record Versioning Info** function and waits for the metadata to be versioned. An MVP consisting of a single file containing both metadata and meta-metadata is sent directly to the Process Metadata Versioning function.

The **Process Metadata Versioning** function (Figure 6) performs a version check on the metadata received from the Receive MVP function. In the case of a modified instance of an existing metadata record, this entails assigning a unique version identifier to the edited record as well as establishing and updating relationships between this version and other co-existing versions in the database.

In effect, the **Process Metadata Versioning** function performs the versioning task in collaboration with the **Assign Version Number** function. In case of a failure in accurately assigning version identifiers to metadata records, a failure report is sent to the Administration entity. Successfully versioned metadata records, however, move on to the **Generate Metadata Preservation Package** function (Figure 6).

The **Record Versioning Info** function (Figure 6) receives the newly assigned version number for a metadata record in transition and adds it to the meta-metadata of the record. Updated Meta-Metadata is then forwarded to the Generate Metadata Preservation Package function.

The **Generate Metadata Preservation Package** function (Figure 6) begins with accepting a metadata record and its corresponding meta-metadata from the Assign Version Number and the Record Versioning Info functions respectively. Subsequently, it creates **Metadata Preservation Package (MPP)** with updated metadata records and its meta-metadata, which is then forwarded to the **Metadata Management** entity for preservation.

The **Metadata Management** entity in Figure 3 can be regarded as the heart of the curation model as it is responsible for satisfying perhaps the most significant requirement of long-term metadata curation - the actual preservation and management of metadata. This entity is essentially responsible for executing the final phase of metadata's journey from ingest to storage. Figure 7 represents the functions of the Metadata Management entity.

The **Receive MPP** function (Figure 7) is primarily responsible for storing metadata records and associated meta-metadata in the database. This function begins with acquiring an appropriate unique storage identifier (or reference information) and version history (particularly important for updated data objects) for the data object that a MPP (received from the Metadata Versioning entity) refers to, from the Administration entity. The acquired data object identifier and version history are attached to the metadata record in the MPP (elaborated in the following paragraph) at a later stage. During a digital curation process, a metadata record may be required to provide accurate reference to or accurately identify the particular version of a data object that it describes, especially when queried by a Consumer. Ideally, this is facilitated by assigning automatically generated unique identifier(s) or reference(s) to valid data objects (i.e. the ones that have passed the necessary validation steps) and attaching the identifier(s) to their corresponding metadata records before they are stored in the designated storage media in the Archival Storage (Figure 2) and the Metadata Management entities respectively. This method of uniquely identifying data objects in a curation system is particularly useful for enabling search engines that execute user-submitted queries for (a) specific data object(s) against their metadata records, to accurately link each metadata record (returned in a search result) to the particular version of a data object that it is associated with. On the other hand, information about the version history of a data object is required to track changes and provenance of that object. Figure 8 provides an overview of the workflow between the Data and Metadata storing functions of a curation system. It should be noted that Figure 8 only presents the primary functions responsible for storing data and metadata and assumes the incorporation of other (i.e. intermediate) functions and/or entities (e.g. Metadata Validation) that the primary functions may depend on, in the system.

In general, after a data object has been successfully stored in the Archival Storage, its storage identifier is passed on to the Administration entity, which then stores it in the relevant data/metadata submission session. In addition, the function responsible for storing the data object also generates a detailed version history of the data object, which (i.e. version history) is also forwarded to the Administration entity to be stored in the relevant submission session. Conversely, for data objects that fail to validate and/or to be stored, the session contains a failure report. Therefore,

acquisition of the data object identifier for a MPP (that corresponds to the data object) is achieved by querying the Administration entity based on the relevant session identifier for the data/metadata submission. For invalid data objects, the Receive MPP function terminates by sending a failure report to the **Generate Report** function (Figure 7), which subsequently forwards the report to the Administration entity. Depending on the related policy of a curation system, a MPP at this stage may either be removed from the system or held temporarily until the Producer re-submits a valid data object for the MPP or the session expires or until a certain pre-defined period of time, whichever is the earliest.

In the case of a valid data object, the Receive MPP function acquires the corresponding identifier and version information of the data object from the Administration entity and then disintegrates the MPP into constituent metadata and meta-metadata. The metadata is subsequently stored along with its corresponding data object identifier and its version history in the database, while the meta-metadata is stored with the metadata versioning info attached to it during the versioning process in the Metadata Versioning entity. From the implementation perspective, the data object identifier and metadata versioning info could be stored in their corresponding columns of the metadata record table and meta-metadata table in the database respectively. The data object identifier and metadata versioning information are mainly used by the **Access** entity (see section 6.1.5) to identify and provide access to an appropriate data object and metadata record respectively, when a Consumer (Note 7) queries for the respective objects. The format of the database can be any known database format, such as relational, XML, and object oriented, whichever is deemed suitable by the curatorial organisation or body.

The **Administer Database** function (Figure 7) is responsible for creating and updating schema or table definition of the metadata database as well as any other database administration related task(s) as required. More importantly, this function performs migration of metadata with the help of the **Metadata Migration** function in order to keep pace with changes in related technology and formats. This function also conducts periodic checking of metadata in collaboration with the **Periodic Quality Assurance (QA)** function. This function entails periodically evaluating metadata to ensure its quality for intended purpose or a range of purposes and updating the metadata (if required) based on the outcome of the evaluation. This function is carried out in accordance with the curation policy imposed by the Administration entity.

Updated metadata resulting from Periodic QA or the Metadata Migration function is fed into the **Generate Metadata Update Package (MUP)** function (Figure 7), which retrieves corresponding meta-metadata from the database and uses them both to generate Metadata Update Package. Generated MUPs are fed into Metadata Ingest entity to be eventually stored in the Metadata Management entity.

The **Perform Queries** function (Figure 7) receives queries about metadata stored in the database from a Consumer via the Access entity, searches the database based on the queries and returns the result set to the Access entity, where the result set is presented to the Consumer.

The **Generate Report** function receives reports from the Administer Database function about different activities that it conducts, such as Database Updates, Periodic QA and Metadata Migration. These reports are sent to the Administration entity for reviewing and assessment purposes. This function is also responsible for notifying a Producer via the Administration entity (Figure 7 and 8) of a success or failure of storage of metadata and meta-metadata extracted from a Metadata Preservation Package that was received (by the Receive MPP function) from the Metadata Versioning entity. In addition, the Generate Report function responds to report requests from the Administration entity about other processes or functions of the Metadata Management entity.

6.1.5 Access

The **Access** entity is another adaptation of an OAIS defined module - the “Access” module in this case. This entity has been re-designed for the curation model with a view to reflect the role that metadata plays in searching and retrieving digital objects (that it refers to) under preservation in the Archival Storage. In effect, this entity is responsible for facilitating search-ability and tracking provenance of metadata that are core requirements of long-term metadata curation

6.1.6 Administration

The **Administration** entity is an adaptation of the Administration entity of the OAIS model (OAIS, 2002), with a number of added features, such as receiving metadata updates and annotations made to either data or metadata in the form of **Annotation Submission Packages (ASPs – Note 8)**, dealing with errors in metadata and digital objects reported by the Consumer; and generating **Metadata Update Package (MUP – Note 9)** (Figure 3) for curation and preservation. In effect, it is the Administration entity through which the MCM facilitates annotation of both data and metadata – a core requirement of long-term metadata curation. Of particular note is the approach employed by this entity (and by the MCM as a whole) for curating annotation that allows annotation to be made to both digital objects and its corresponding metadata as an external entity and treats it in isolation as part of existing metadata records of the object and its meta-metadata respectively. The advantage of this approach over the one that allows annotation to be

embedded or attached in the actual data object or metadata is that the former does not cause any violation of the edits related legal rights (e.g. Copyright) associated with the digital object while retaining the ability of the latter to make the annotation available to the consumer in a convenient manner. Typically, the users of the system would be provided with an annotation interface, which would allow them to select any particular context(s) of the digital object or the metadata record of their interest and add annotation(s) to that context(s). The interface would also facilitate searching, displaying and editing annotations made to the digital objects under preservation.

6.2 Applicability of the Model

The Metadata Curation Model may be applicable to any OAIS-based information preservation system or archive as well as any long-term data curation system, where metadata is preserved separately to the actual digital resource that it is associated with. In general, the model is applicable to any organisation that is responsible for making digital resources available over the long-term and actively acknowledges the role that metadata can play in efficiently fulfilling that responsibility. Moreover, the MCM is equally applicable to both the organisations that are looking to build new curation systems, and those aiming to incorporate curation-related functions into their existing non-curation focused systems. This is facilitated by the modular architecture of the MCM that enables it or any of its entities to be easily integrated into any existing metadata system to make it curation-aware. In addition, the model (or any of its components) may be extended or customised to incorporate domain/system specific functions and accommodate future curation requirements. The case-study below illustrates potential use of the Metadata Curation Model in the Science and Technology Facilities Council (STFC, 2007) data portal (Note 10).

The STFC operates for the UK research community several large scale scientific facilities that all generate large quantities of data. While the STFC provides a common way for discovering and accessing these multi-disciplinary data through a web-based data portal, there is currently no comprehensive measure in place to curate and preserve these data over the long-term. Without proper and efficient curation and preservation, these data could potentially become obsolete due to fast changing technologies and data formats. Therefore, considering the current status and increasingly large volumes of data managed, the STFC could benefit from an efficient long-term curation system and hence makes an ideal use case for the Metadata Curation Model.

A close inspection of the STFC's data management architecture (Figure 9) suggests that it would not be too difficult, at least in theory, to implement an efficient curation system for the STFC data.

The present data management architecture enables users to manage (e.g. edit, store) their data files on file servers at Cambridge and London through the central Storage Resource Broker (SRB) software and database at STFC (Blanshard, Tyer, Calleja, Kleese and Dove, 2004). The architecture also facilitates (via the Metadata Editor) creation of metadata about the data to make it discoverable to the users via the data portal. In order to transform the present architecture into a long-term data curation focused architecture, the first step would be ensuring availability of adequate, appropriate and good quality metadata about the data. This would require appending necessary metadata curation elements to the currently employed metadata format, i.e. the STFC Scientific Metadata Model (SSMM), which currently lacks the ability to record sufficient information (e.g. data/metadata provenance, Representation Information, meta-metadata) required for efficient long-term curation. Modification to the metadata format would in turn require the modification to the metadata database schema, which is based on the SSMM format.

The next step would involve the implementation of the Metadata Curation Model, which would incorporate the features of the data portal and the metadata editor as well as other curation features, such as provenance tracking and data/metadata annotation amongst other things. Therefore, a revised version of the data management architecture (Figure 10) would replace the STFC data portal and the metadata editor with the metadata curation component as it supersedes them. The implementation of the metadata curation model would also require implementation and/or employment of other services, such as a Representation Information repository and controlled vocabulary server.

The final and most challenging step would be developing a long-term preservation archive for the data. This step would require an in-depth assessment of the SRB and existing data storage mechanisms to determine whether it would be feasible (in terms of costs and effort required) to extend them to incorporate long-term preservation features or they would need to be replaced with more suitable technologies (therefore marked with "?" sign in Figure 10).

Moreover, in order to evaluate and demonstrate the underlying concepts of the MCM, a web-services based prototype system has been developed. The prototype system is available online at <http://www.metadata-curation.co.uk>

7. Conclusions

Efficient and effective long-term metadata curation is a key component of successful preservation, apposite enrichment and sustained accessibility of digital information in the long term. Unfortunately, no comprehensive method for effective curation of metadata for long periods of time is known to exist till date. The Metadata Curation Model aims to meet the necessity of an efficient metadata curation approach by combining the best features of existing long-term digital preservation strategies (i.e. the OAIS model) with a considerable degree of innovation. However, there is still a

great deal of scope for further advancement, as the suitability and efficiency of the MCM may only be accurately measured when implemented and tested in a fully-operational long-term digital curation system. Nevertheless, the approach presented in this paper may be regarded as a conceptually complete and scalable solution for long-term metadata curation that would benefit any discipline concerned with long-term data curation.

References

- Blanshard, L., Tyerl, R., Calleja, M., Kleese, K. and Dove, M.T. (2004). *Environmental Molecular Processes: Management of Simulation Data and Annotation*, Proceedings of the UK e-Science All Hands Meeting 2004, © EPSRC Sept 2004, ISBN 1-904425-21-6, [Online] Available: http://archive.niees.ac.uk/documents/AHM_dataman_2004.pdf
- CEDARS, (2002). CEDARS Project. [Online], 2002, Available: <http://www.leeds.ac.uk/cedars/index.html> (November 4, 2007)
- Macdonald, A. and Lord, P. (2002). *Digital Data Curation Task Force Report of the Task Force Strategy Discussion Day*, November 2002, [Online] Available: http://www.jisc.ac.uk/uploaded_documents/CurationTaskForceFinal1 (January 15, 2008)
- NEDLIB, (2000). Networked European Depository Library. 2000 [Online] Available: <http://nedlib.kb.nl/> (November 4, 2007)
- OAIS, (2002). *Reference Model for an Open Archival Information System (OAIS)*, CCSDS Blue Book. Issue 1. January 2002, [Online] Available: <http://public.ccsds.org/publications/archive/650x0b1.pdf> (January 3, 2008)
- PRONOM, (2007). The technical Registry PRONOM, The National Archive, 2007, [Online] Available: <http://www.nationalarchives.gov.uk/pronom/> (January 17, 2008)
- SRB, (2007). Storage Resource Broker, 2007 [Online] Available: http://www.sdsc.edu/srb/index.php/Main_Page (January 20, 2008)
- STFC, (2007). The Science & Technology Facilities Council (STFC), 2007, [Online] Available: <http://www.scitech.ac.uk> (January 20, 2008)

Notes

Note 1. A SIP contains three objects - data to be preserved, its associated metadata and information about the metadata itself, i.e. meta-metadata.

Note 2. The term “User Base” encompasses all identified potential consumers (e.g. human, software application etc.) to whom curated metadata is beneficial in terms of accurate interpretation and proper utilisation of the digital object that the metadata describes and/or refers to. The User Base is essentially an adaptation of the Designated Community as defined in the OAIS reference model (OAIS, 2002).

Note 3. A Metadata Update Package consists of existing metadata records and their corresponding meta-metadata with any significant changes made to them. Changes to existing metadata records occur due to amendments submitted by producer and/or as a result of different curation related activities, such as metadata migration.

Note 4. A standardised and structured list of pre-defined values for different elements within a metadata record that conforms to some agreed standard(s). These pre-defined values also represent true knowledge organisation schemes that define the metadata concept, specify the scope and the relationships among the concepts.

Note 5. Information regarding different quality assurance functions or processes within a curation system, such as metadata crosswalking, structural validation and semantic validation, that a metadata record has to pass through before it is declared valid. This information includes time of the function execution, changes it makes to the record, tools used and so on.

Note 6. DROID (Digital Record Object Identification) is an automatic file format identification tool developed by the National Archives, UK- <http://droid.sourceforge.net/wiki/index.php/Introduction> (4 November 2007).

Note 7. The role played by those persons or client systems that find preserved information of interest and access that information in detail (OAIS, 2002).

Note 8. An Annotation Submission Package is comprised of annotation made to a digital object and metadata about the annotation, e.g. name and affiliation of annotation, date annotation made, part of the digital object it refers to, type of annotation and so on.

Note 9. A Metadata Update Package consists of existing metadata records and their corresponding meta-metadata with any significant changes made to them. Changes to existing metadata records occur due to amendments submitted by producer and/or as a result of different curation related activities, such as metadata migration.

Note 10. STFC Data Portal - <http://tiber.dl.ac.uk:8080>

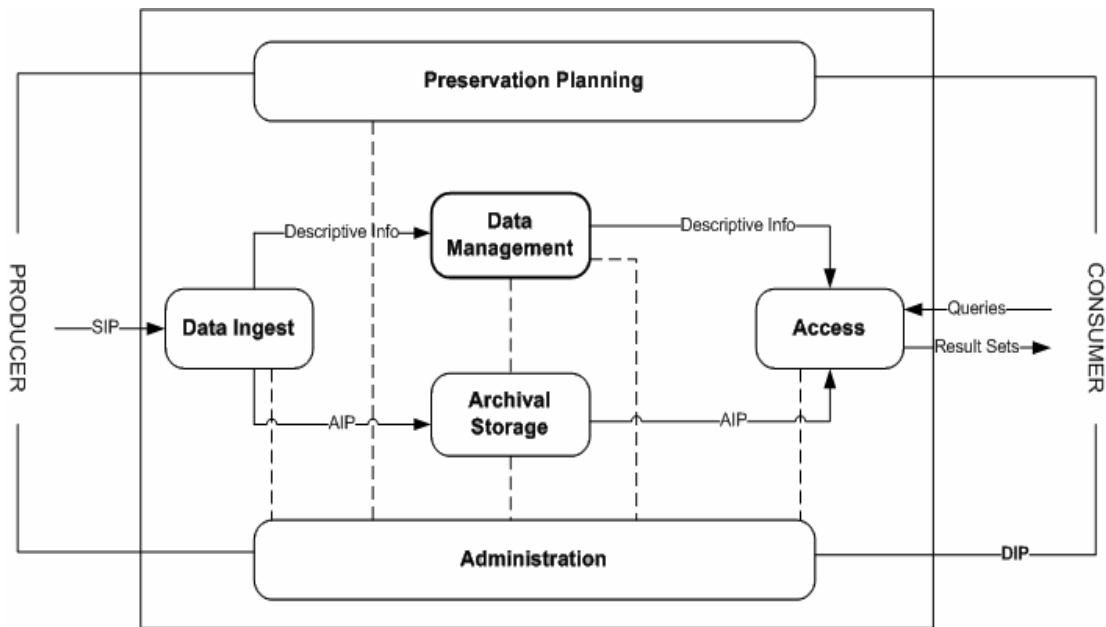


Figure 1. Functional Entities of the OAIS Reference Model [1]

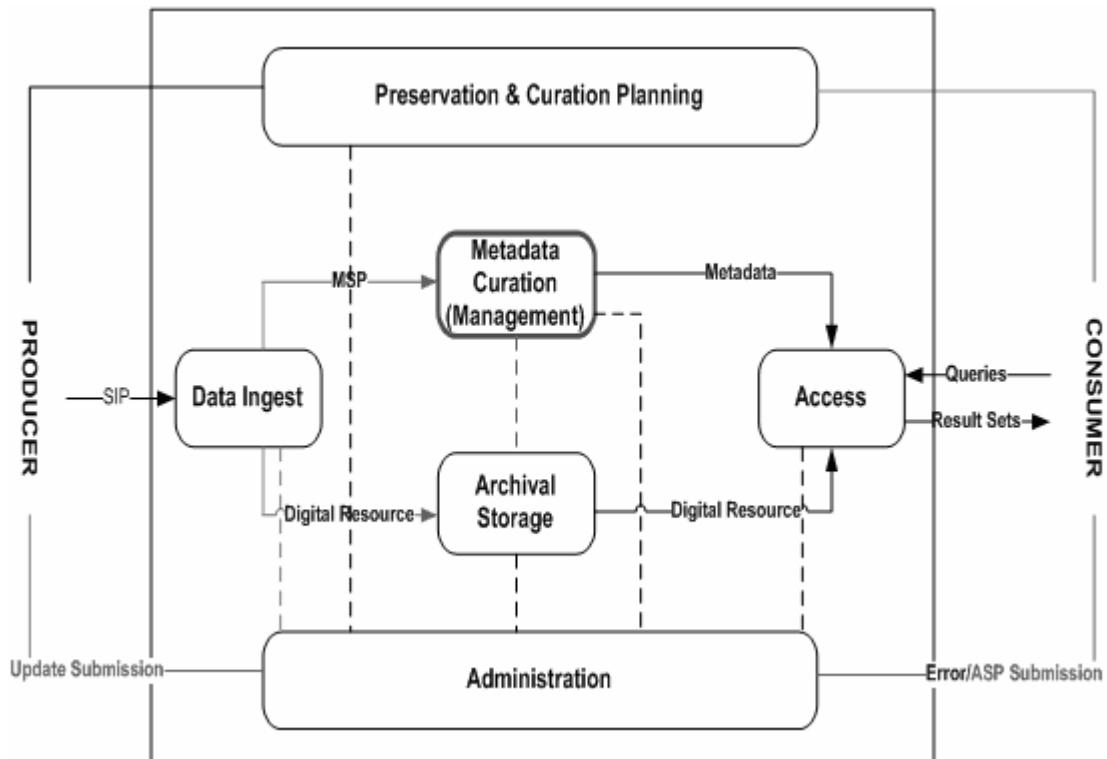


Figure 2. The Metadata Curation Model embedded in the OAIS Reference Model (highlighted in red)

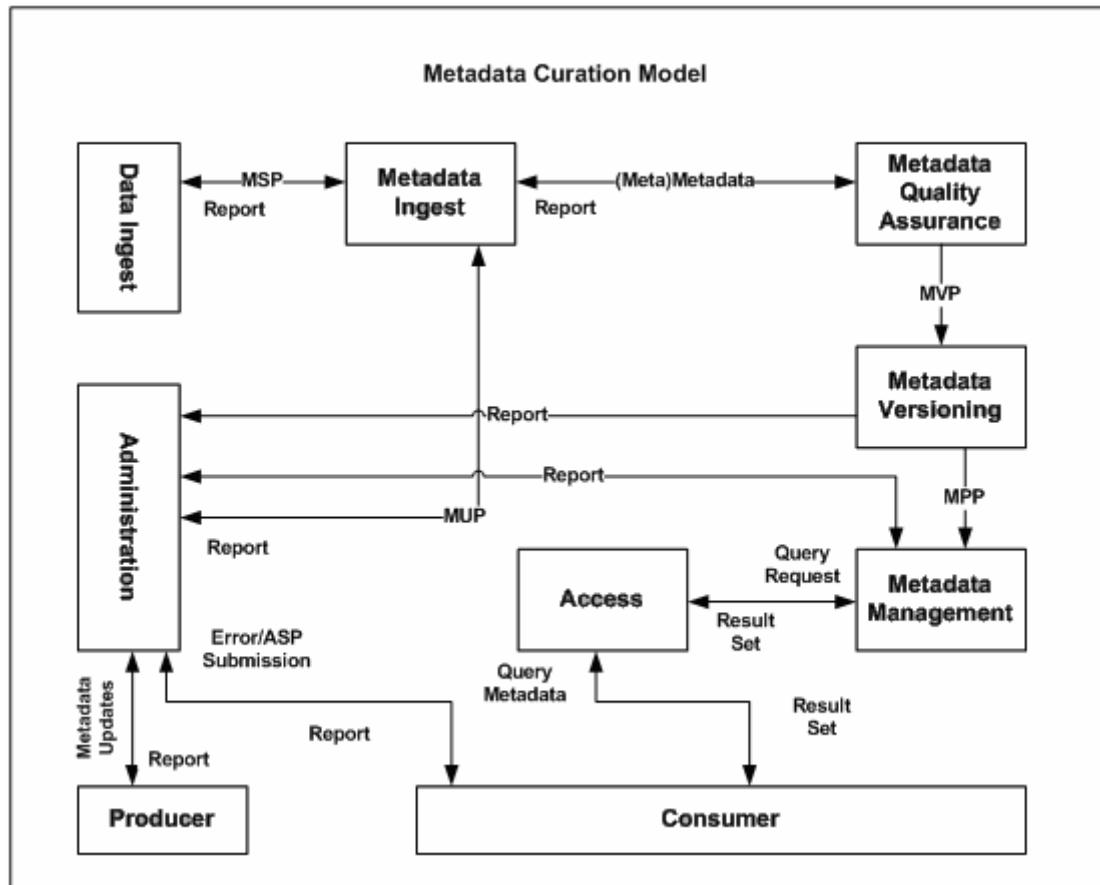


Figure 3. Functional Entities of the Metadata Curation Model

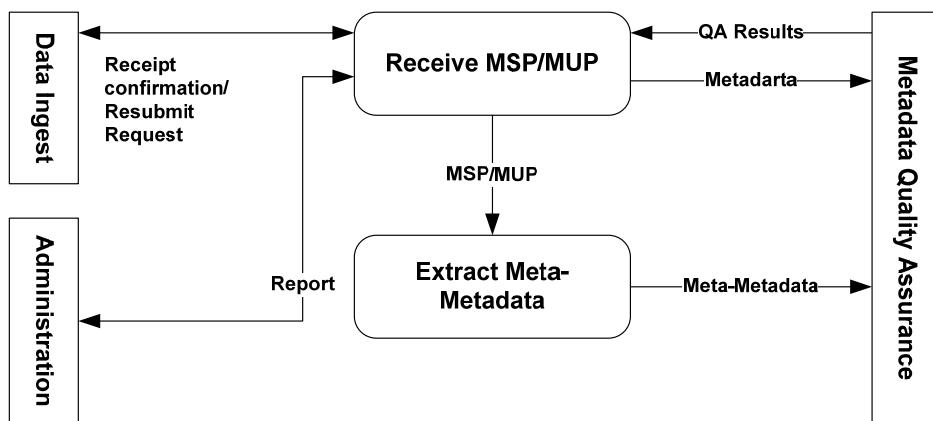


Figure 4. Functions of the Metadata Ingest Entity

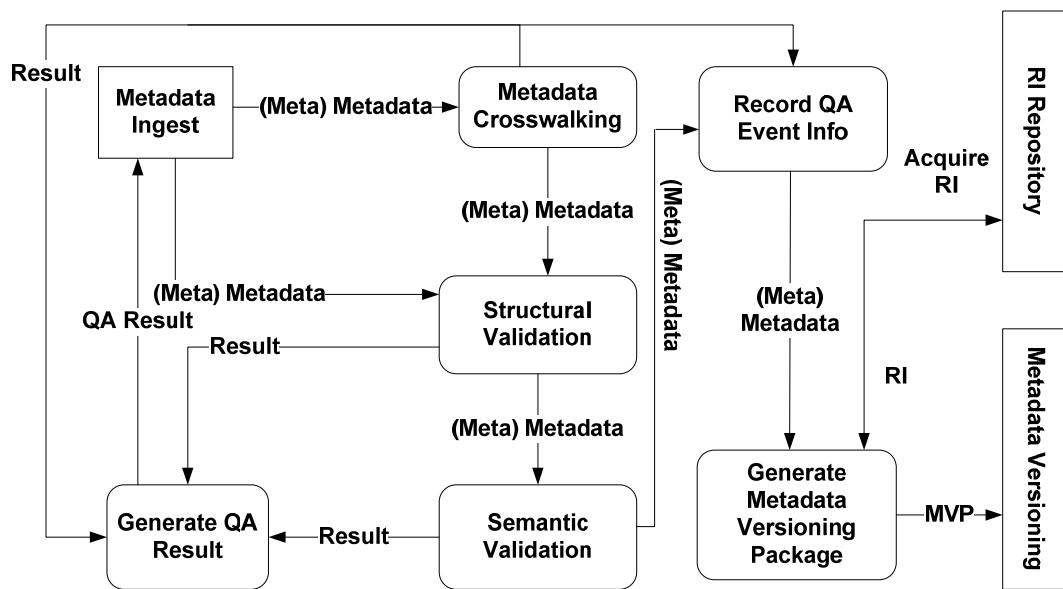


Figure 5. Functions of the Metadata Quality Assurance Entity

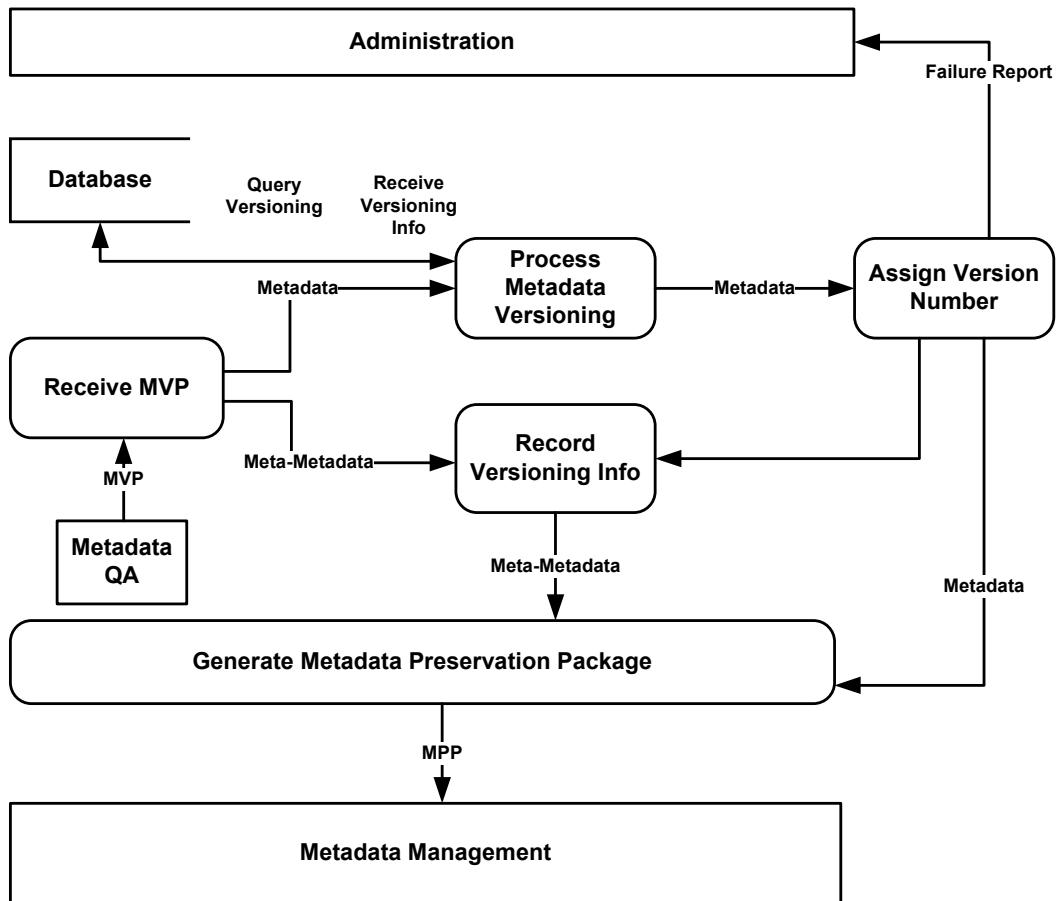


Figure 6. Functions of the Metadata Versioning Entity

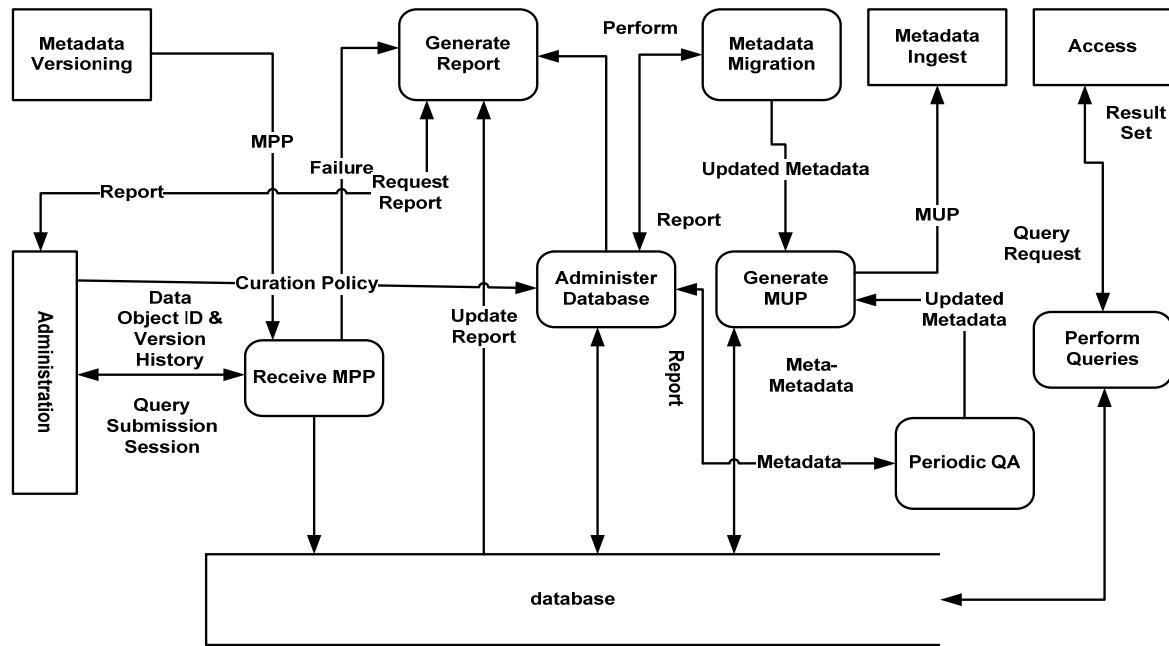


Figure 7. Functions of the Metadata Management Entity

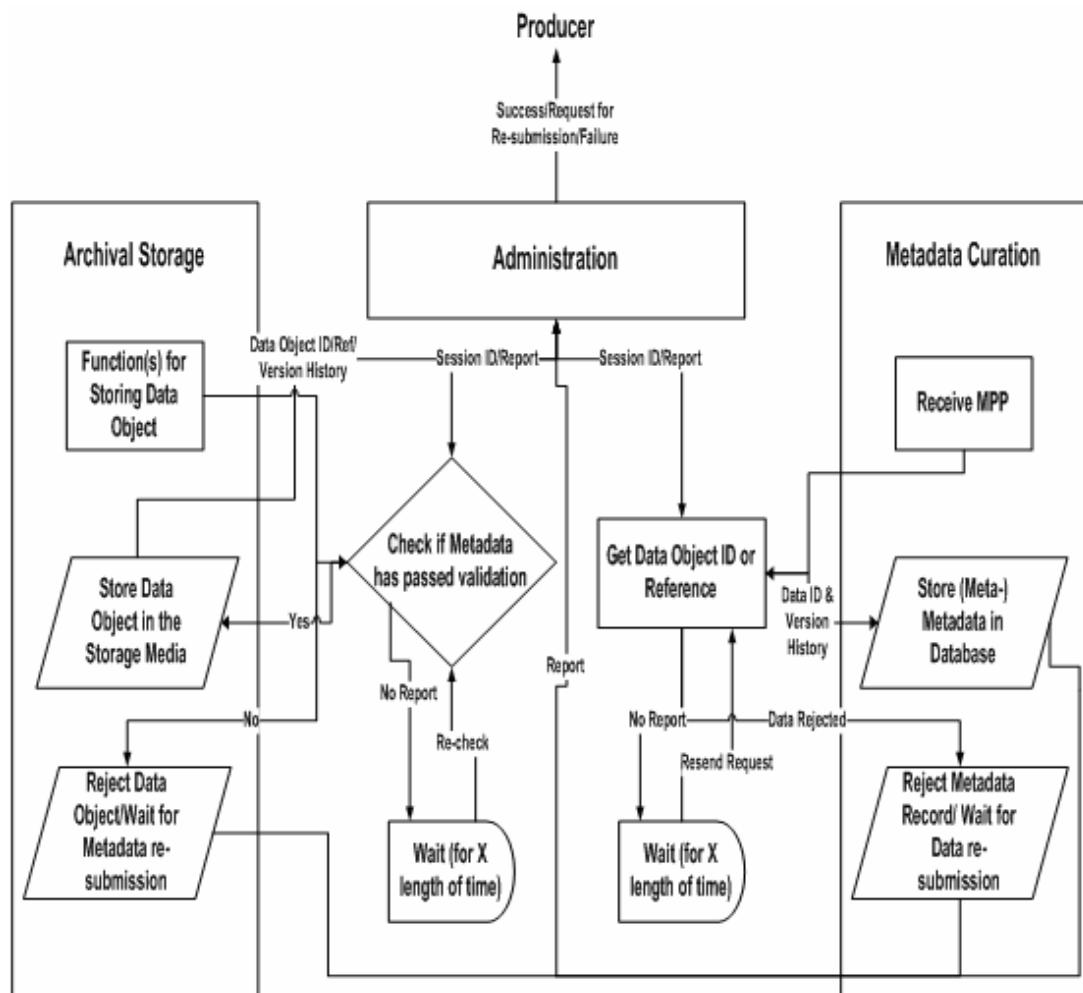


Figure 8. An Overview of the Data and Metadata Storing Process in a Curation System

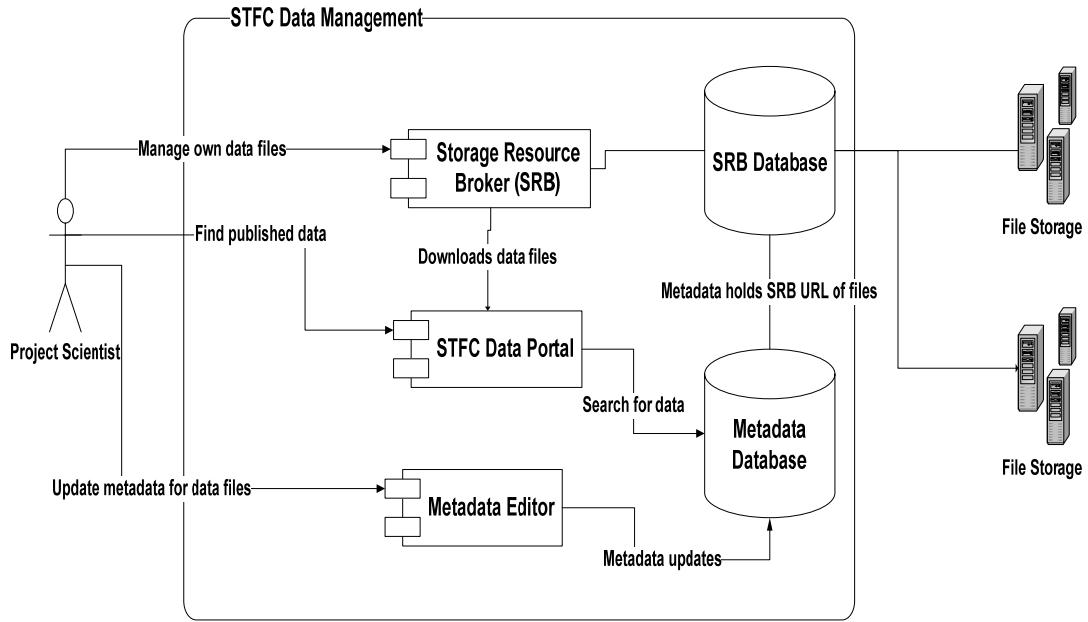


Figure 9. The STFC Data Management Architecture (Source: Blanshard, Tyer, Calleja, Kleese and Dove, 2004)

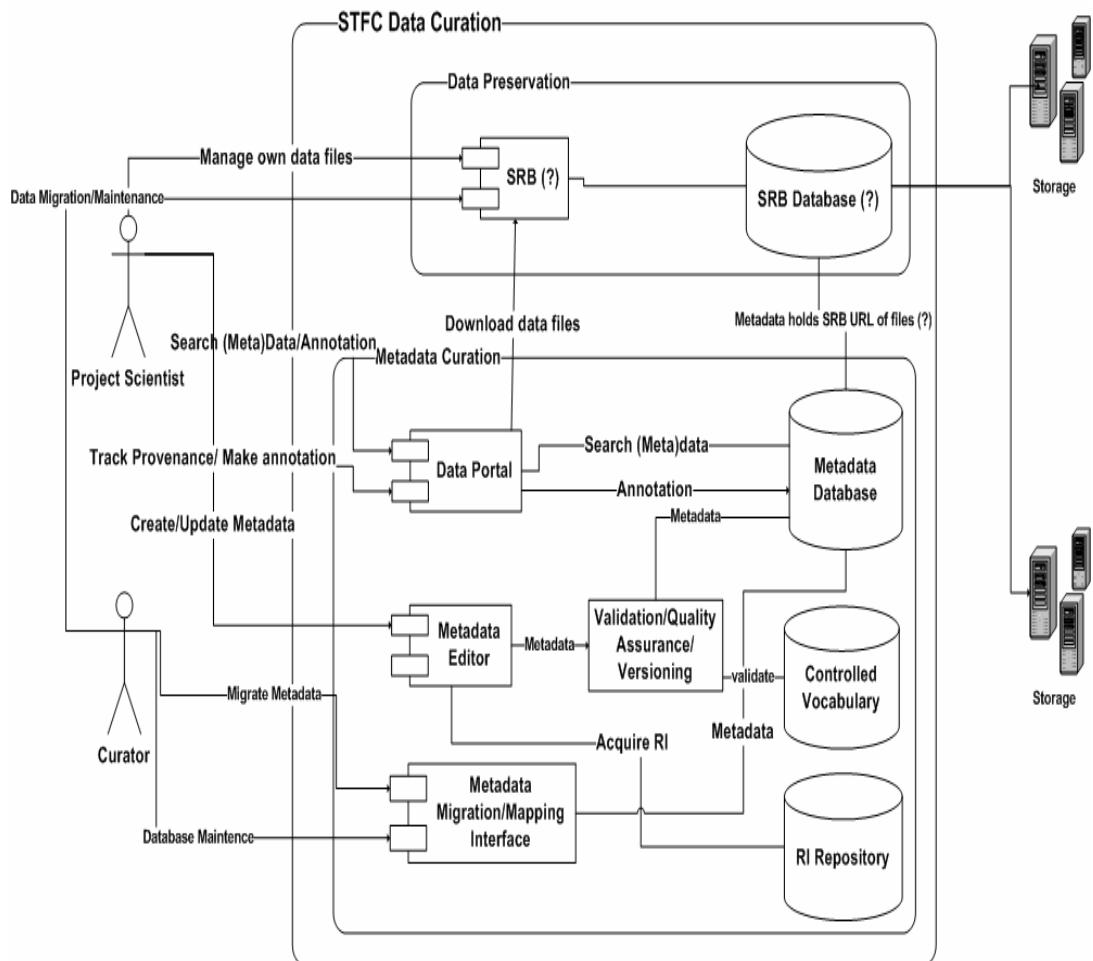


Figure 10. A Revised Version of the STFC Data Management Architecture with the Long-term Curation Features