

# A lifecycle model for an e-print in the institutional repository

## Document details

Project: SHERPA DP  
Work Package: 2.13  
Author: Gareth Knight  
Version: Version 1.0  
Document date: 13/02/2006  
Change history:

<i>Date</i>	<i>Version</i>	<i>Author</i>
13/02/2006	First version	Gareth Knight

---

## Contents

Introduction.....	2
Lifecycle of an E-Print.....	2
Responsibility for events in the e-print lifecycle .....	3
Conclusion.....	6
References .....	6

## Introduction

An archive is an organisation that is responsible for the long-term storage and maintenance of resources (RLG, 2002). Traditional archives expect that physical objects will be retained indefinitely and that active management will be necessary to ensure the physical object – a parchment, book or other intellectual work – can be maintained. These archives typically measure the lifetime of an object in decades, or even centuries. Digital archives, in comparison, are a recent development that has existed for less than 30 years. Equally, the lifecycle of a digital object is considered to be much shorter than the physical equivalent. It is considered optimistic if a digital record will last for 10 years over several successions of technology. Whatever the time period for curation, the creation of an information lifecycle is considered essential to understand the object and specify the period in which action should be taken to ensure it does not further degenerate. This document outlines a lifecycle model that may be applied to e-prints and identifies how responsibility may be allocated to different partners in the Sherpa DP disaggregated model.

## Lifecycle of an E-Print

The preservation of digital materials requires a continuous process of active management from the point of creation through to the point of withdrawal from the repository (if such an event occurs). The information lifecycle consists of a series of events occurring at various stages during the lifetime of the object. These establish a clear and distinct chronology of actions to be performed upon the object or a continuum of actions that cannot be distinguished from one another (Cedars, 2002). Several organisations have developed lifecycle models that, although useful for their own physical or digital holdings, do not cater for the specific characteristics of e-prints (Life project, 2005; Cedars, 2002). For institutional repositories and other types of e-print archive, the lifecycle model of an e-print, developed by James et al (2004) is considered useful for understanding the unique requirements of an academic research paper in the context of an institutional repository.

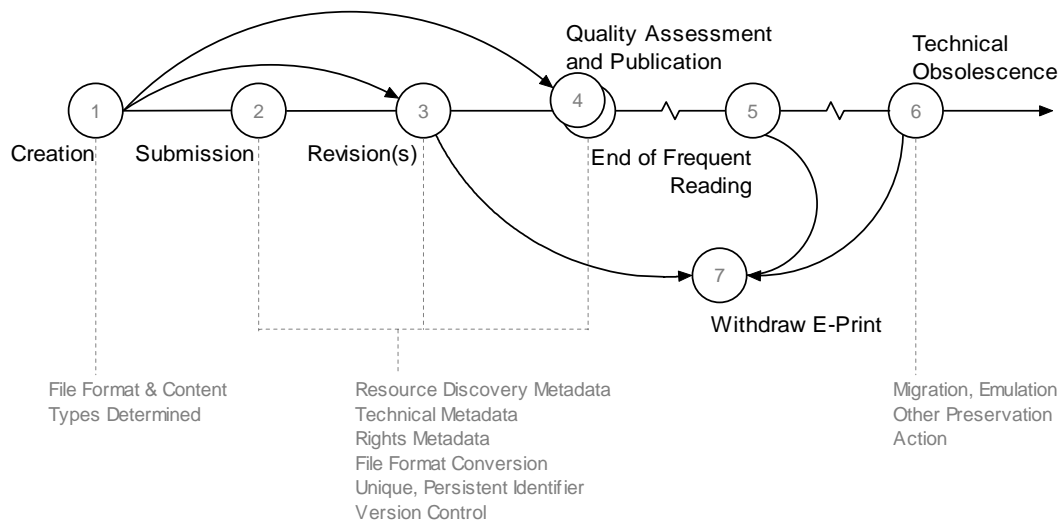


Figure 1: Lifecycle of an E-Print in an e-print archive, developed by James et al

The lifecycle model for e-prints establishes a linear interpretation of the relevant events that occur during the lifetime of an e-print. James et al (2004) outlines seven distinct stages:

1. Creation of the e-print as a digital object, when the file format (MS Word, RTF, etc.) and content types (text, images, etc.) are determined.

2. Submission of the e-print into the institutional repository, as well as the assignment of a persistent identifier, creation of resource discovery and administrative metadata necessary to manage the e-print.
3. Revision of the e-print by the depositor and subsequent submission of the updated e-print.
4. Formal quality assessment, performed by a publisher that result in the creation of a final version of the e-print. This final version may be deposited with the institutional repository or published elsewhere.
5. The value of a research paper is likely to decrease over time. The event that indicates the 'end of frequent reading' may be determined retrospectively by examining system logs to identify the number of times that an e-print has been downloaded.
6. The file format in which an e-print is held will, at an indeterminate point in the lifecycle, be rendered obsolete. Preservation action, such as migration or emulation will be necessary to restore the accessibility of the e-print.
7. An e-print submitted into the institutional repository may, if certain criteria are met, subsequently be removed. James et al (2004) identify three scenarios for withdrawal: 1) an early draft is replaced by a later draft; 2) the e-print is no longer read frequently; or 3) the file format is rendered inaccessible as a result of technological obsolescence.

The stages in the e-print lifecycle may be separated into active and non-active events. The active events are those performed by the creator/author, institutional repository and other entities (peer-reviewers, publishers) that result in some change to the status of the e-print. For example, the 'quality assessment and publication' stage is likely to produce a published post-print, derived from the pre-print that has previously been submitted. Non-active stages indicate events that occur through inaction and may be considered as part of the e-prints natural lifecycle. These are less tangible and difficult to identify until after the time has passed. For example, the 'End of frequent reading' stage can be identified when system logs indicate that only a small number of people have downloaded an e-print during the past six months. Similarly, 'Technical obsolescence' will occur when the file format is rendered obsolete and no action has been taken to migrate the content to other file formats.

### **Responsibility for events in the e-print lifecycle**

A business model for an institutional repository should identify relevant stages in the e-print lifecycle, decisions that must be made and the potential risks if action is not taken. The e-print lifecycle model may be refined to allocate responsibility to one of three parties - the Depositor, Institutional repository and the Preservation Service – in the Sherpa DP disaggregated service model.

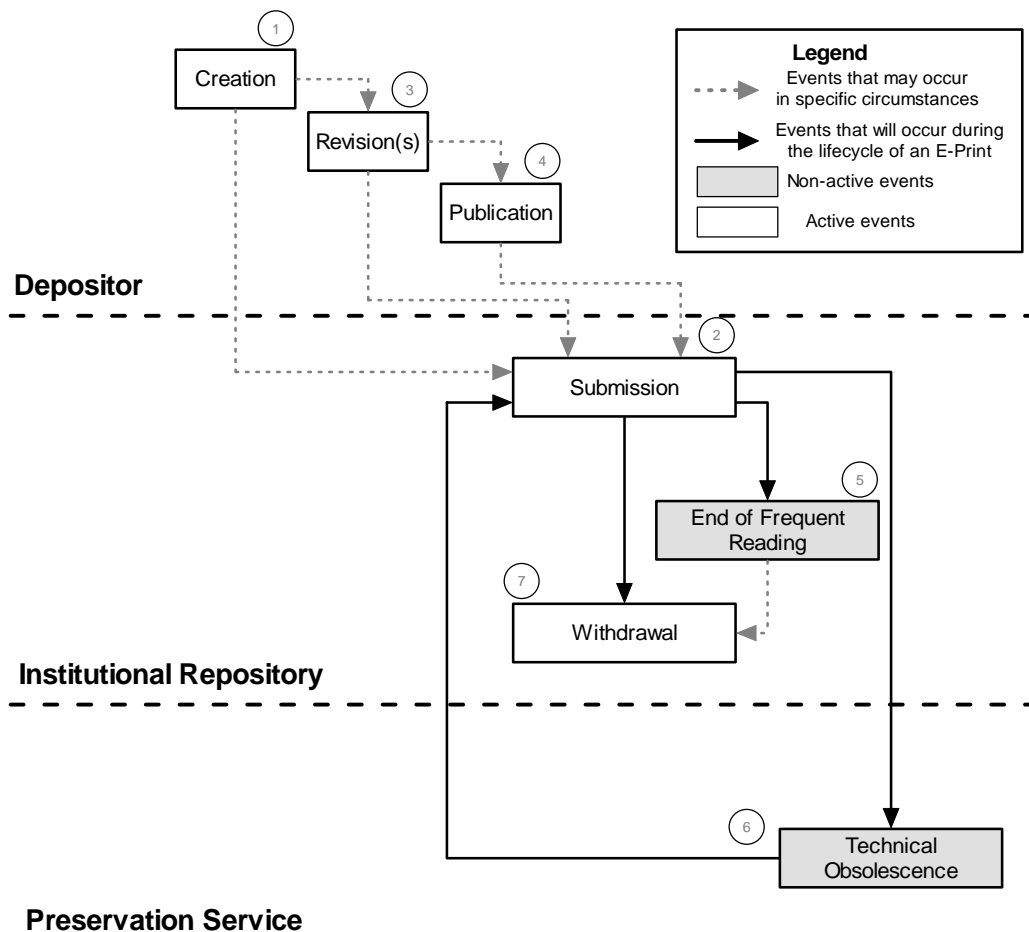


Figure 2: Allocation of responsibility for events in the E-Print lifecycle

### Depositor-driven activities

The Depositor is responsible for the creation, revision and submission of the e-print into the institutional repository. A depositor may be the author(s) of the e-print or a staff member allocated the task of scanning printed research papers or theses that have previously been deposited with the institution. At the creation stage, the depositor makes choices about which software package to create the e-print, the file format in which the content will be stored, and the type of content (text, images, etc.). For example, an author may write an e-print using Microsoft Word 9.0. Alternatively, repository staff may scan printed student theses, stored the images as TIFF and combine the pages into a single file using Adobe Acrobat. At submission, the depositor is responsible for the creation of metadata that will assist the discovery and administration of the e-print.

### IR-driven activities

The institutional repository operates as a facility an author may use to distribute their research paper to the wider research community. An appropriate infrastructure must exist to allow the repository to accept an e-print submitted by the depositor, hold it in an appropriate storage facility and deliver it to the researcher on request. On establishment, an institutional repository will develop policies and issue guidelines that an author must follow if they wish to deposit their research paper into the archive. These guidelines should identify the file formats (e.g. RTF, PDF, etc.) considered appropriate for submission into the institutional repository; file formats considered appropriate for distribution to the user community (e.g. PDF); and establishment of the legal responsibilities of the repository, through the creation of deposit licence. These policies should, in most cases, limit the amount of work that repository staff must perform on the e-print on submission.

The lifecycle model developed by James et al (figure 1) identifies five actions that must be taken by an institutional repository when accepting an e-print into the archive:

1. Allocate a unique, persistent identifier to locate the e-print within the repository.
2. Perform version control to identify previous revisions of the paper. Subsequent submission of pre-print revisions (event 3) or the post-print (event 4) should be noted and appropriate relation metadata created.
3. Create or refine resource discovery metadata.
4. Create or refine rights metadata
5. Convert deposited data to a file format considered suitable for distribution.

An additional action may also be identified:

6. Make the e-print available through OAI harvesting.

These actions are implemented by repository staff participating in the Sherpa DP project, according to the standards developed for use in their repositories. Most notably, the criteria for refinement of resource discovery and rights metadata, as well as the actions taken to convert deposited e-prints into a suitable distribution file format will vary between institutions. However, the final outcome should remain the same – on completion of the various processes in the submission entity, the institutional repository should have an e-print and associated metadata that is suitable for searching in their own repository software and harvesting by the preservation service.

### **Preservation Service-driven activities**

Repositories typically store a minimum amount of metadata necessary to catalogue digital objects. Few e-print repositories have encountered problems with long-term access to an e-print. This may be attributed to the relatively short time period since the first e-print repositories were established (15 years) and the relative ease with which text can be converted across multiple generations of software. To authenticate and provide long-term access to an e-print, further information is necessary to enable a preservation repository to describe the composition of a digital object, as well as software tools to allow its manipulation. The OAIS reference model uses the term 'Representation Information' (RI) to describe information that may be used to interpret the binary of a digital object as an information object. It may be separated into three information subsets:

1. Structure – describes the structural composition of a digital object and its relationship with other objects. This may be defined by an international standard, such as the HTML specification, or de-facto standards, such as HTML transitional.
2. Semantic – provides further information on the contents of a resource. Typically
3. Other representation information – describes further information stored on the standard, access and rendering software that may be necessary to interpret the digital object. For example, an e-print stored as a PDF may require a specific version of the Adobe software.

The preservation service is responsible for the preservation of e-prints made available by the institutional repository. Although institutional repositories often impose restrictions upon the data type, different degrees of conformance levels to format specifications (Knight, 2005) may require different treatment when maintaining the content. To record such information during the ingest process, the preservation service must possess the following procedures:

1. Identify and catalogue digital objects harvested from institutional repositories. The catalogue process should identify the technical characteristics of the e-print (file type, version), the significant properties that must be preserved, and any relationships created in the preservation repository (e.g. derivatives).
2. An ongoing assessment of the preservation risks to the e-print, with reference to the technical infrastructure of the e-print.

3. A plan for long-term preservation, detailing the options available to export the intellectual content of the e-print into a platform-independent format, and associate it with appropriate preservation metadata.
4. Identify a time period when the file format distributed by the institutional repository is considered to be obsolete and create a new version in an appropriate file format suitable for distribution. Connect to the institutional repository and transfer migrated data.

The programmatic identification of digital objects and the creation of relevant metadata is an important starting point for analysis. Analysis tools, such as Hove and the NLNZ Metadata Extractor Tool are useful tools to examine the technical composition of a digital object, including the conformance to a published specification, and are currently being used by several organisations to examine their digital holdings and are likely to be of use within the project. Distributed type registries, such as the Global Digital Formats Registry (GDFR), the National Archives' PRONOM and DCC Representation Information Registries (DCC-RR), may provide further information.

A policy to review the accessibility and preservation risks to e-print should identify a timetable and criteria for assessment, as well as a pragmatic method to resolve any issues. A risk assessment should consider three factors: the status of the file format specification (open, proprietary), the extent to which the data infrastructure is known and documentation, the availability of appropriate, preferably free software, to decode and view the intellectual content in different operating systems, and the availability of appropriate software tools to export the significant properties of the e-print. Proprietary file formats, such as Microsoft Word and Adobe PDF, are considered to present a high risk to the long-term accessibility of e-print content (James et al, 2004) and action should be taken to convert the content before the file format is rendered obsolete. Further investigation is necessary to develop a migration path that will convert the intellectual content of an e-print into a suitable file format, without modification to the layout of the document itself (Knight, 2005).

## Conclusion

The long-term preservation of their digital holdings should be a matter of concern for Institutional repositories and should be addressed as soon as possible. The lifecycle model created by James et al and developed within this document indicates how an e-print may be managed in a disaggregated service model. It is imperative that both parties – the institutional repository and the preservation service – establish the services they will perform and the timescale in which these actions will take place. Appropriate policies must be developed to specify appropriate formats for deposit and establish permission to preservation processes during the later stages of the e-print lifecycle, to ensure future actions are not unknowingly limited.

## References

The Cedars Project (2002), Cedars Guide to Digital Collection Management. Retrieved on December 20, 2005, from:

<http://www.leeds.ac.uk/cedars/guideto/collmanagement/guidetocolman.pdf>

The Life Project (2005), LIFE: Life Cycle Information for E-Literature. Retrieved on December 20, 2005, from: <http://www.ucl.ac.uk/lifeproject/>

James, H. et-al, (2004). Feasibility and Requirements Study on Preservation of E-Prints. Retrieved on December 20, 2005 from:

[http://www.jisc.ac.uk/uploaded\\_documents/e-prints\\_report\\_final.pdf](http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf)

Harnad, S. & Goodman, D. (2003). Online transactions ["Eprint versions and removals"]. Messages posted to American-Scientist-E-PRINT-Forum

<http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/2848.html>



Knight (2004) Report on Preservation Standards

[http://www.sherpa.ac.uk/documents/D4-5\\_Report\\_on\\_Preservation\\_Standards.pdf](http://www.sherpa.ac.uk/documents/D4-5_Report_on_Preservation_Standards.pdf)

RLG-OCLC (2002)

Trusted Digital Repositories: Attributes and Responsibilities. Retrieved on December 20, 2005

from: <http://www.rlg.org/legacy/longterm/repositories.pdf>