

# A Draft Of The E-Journal Archives Ingest Process

William R. Kehoe  
Digital Library and Information Technologies  
Cornell University Library  
March 2002

The list of questions has been prepared to guide the interaction between the E-Journal Archives and its depositors, with the goal of complying with the draft standard proposed in the Consultative Committee on Space Data System's [Producer-Archive Interface Methodology Abstract Standard](#) (1). This version of the draft is very general. As we gain experience with producing preservation metadata in the context of the interaction between data depositors and the E-Journal Archives, I expect that we will be able to be more specific about our procedures.

This document was prepared in Microsoft Word 2000. If you view it in Word's Outline View, you can expand or contract the topics and subtopics by clicking on the plus and minus signs to the left of the headings.

The numbers in brackets following the section headers refer to the parallel section in the CCSDS Producer-Archive Interface proposal. All quotes are taken from the same document.

## Preliminary Phase [3.1]

### First Contact [3.1.1]

- Who is the primary contact person for the Depositor?
- Who is the primary contact person from D-LIT for this project? From Preservation?
- Are there secondary contacts—technical or content experts, for example?
- The Depositor provides general information about its journal, “... *the aim of the project, its context, its schedule and its constraints.*” The E-Journal Archives provides an overview of its operation and standards as they apply to the collection.

### Preliminary Definition, Feasibility And Assessment [3.1.2]

#### *Information To Be Archived* [3.1.2.1]

- What exactly will be preserved?
- Are there open issues about what will be preserved? Can they be resolved?
- What complementary information exists to be preserved? Representation Information? Any PDI information?
- Who will access this information?
- How is it accessed? Is it restricted? Is there a required service level or constraints?
- For how long should the information be kept?
- What is the estimated cost at this point?

### ***Digital Objects And Standards Applied To These Objects [3.1.2.2]***

- What are the Data objects that will be kept, by category? Content Objects? Representation Information objects? Context objects? And so on.
- What are the standards used by the E-Journal Archives for the particular Content Objects and their Representation Information?
- *[Optional]* Does the E-Journal Archives have tools to aid in migration of the Content Objects or their metadata to the E-Journal Archives standards?
- What are the standards used by the Depositor for the Content Objects and the Representation Information?
- *[Optional]* Does the Depositor have tools to aid in migration of the Content Objects or their metadata to the E-Journal Archives standards?
- How well do the Depositor's files comply with the E-Journal Archives's standards?
- How much effort and cost will go into making the files compatible?

### ***Object References [3.1.2.3]***

- Are there are naming rules in the domain of the collection, in its legal provisions, or in its standards? If so, list them.
- Are other naming rules necessary? If so, define them.
- What will it cost to use the rules?

### ***Quantification [3.1.2.4]***

- How much data will be transmitted to the E-Journal Archives?
- How much data will be stored permanently—an estimate?
- How much storage will be needed during the ingest process?
- How much will all the storage cost?

### ***Security Conditions [3.1.2.5]***

- Should the transfer of the data from the Depositor to the E-Journal Archives be secure?
- Should the data be validated on transfer?
- Can you identify the tools or methods that will be used, if they are necessary?
- How much will security cost?

### ***Legal And Contractual Aspects [3.1.2.6]***

*“These aspects should be included in the Submission Agreement.”*

- What is the nature of the relationship between the Depositor and the E-Journal Archives? Is the data owned by the library? Or does it remain the property of an organization outside CUL?
- Who owns the intellectual rights to the material?
- Do the rights transfer from the Depositor to the E-Journal Archives?
- If so, what documents are necessary to legalize the transfer?
- If not, what rights does the E-Journal Archives have? Does it have the right to migrate the files to another format?
- Who has the right to access the material? What are the E-Journal Archives' obligations to provide or restrict access?
- If the information regulated in any way, capture those regulations in writing.
- For each type of digital object to be preserved, specify the standard to which it will be preserved. If there are specific tools that must be used, make them explicit.
- How much will complying with all of these aspects cost?

### *Transfer Operations* [3.1.2.7]

- Prepare a draft SIP.
- Discuss requirements, capabilities, and potential solutions for transferring the data.
- How much will the transfer cost?

### *Validation* [3.1.2.8]

- What tools and procedures does the E-Journal Archives have in place to validate or reject the data to be transferred?
- Will the E-Journal Archives' tools and procedures need to be modified to accommodate the data?
- Can the depositor's quality control processes be modified to make the validation process easier?
- How much will these changes cost?

### *Schedule* [3.1.2.9]

- "Define a preliminary schedule."

### *Permanent Impact On The Archive* [3.1.2.10]

- If the E-Journal Archives were to make a commitment to preserve the data, how would the commitment affect current and future planning for the Library's Central Depository as a whole? Areas for consideration include plans for storage, migration plans and capabilities, security plans, and plans for possible transfer of the data to another archive. This question is for the E-Journal Archives and Central Depository personnel alone.

### *Summary Of Costs, Risks* [3.1.2.11]

- What is a summary of the costs estimated to this point?
- What risks have been identified to this point?

### *Critical Points* [3.1.2.12]

- Which of the risks will cause the project to fail if they are not mitigated?
- Are there other risks that will cause a partial failure of the project?

## **3.1.3 Establishment Of A Preliminary Agreement**

- Is the project feasible for both the Depositor and the E-Journal Archives? This assessment should be made after both parties participate in creating a document that summarizes the findings of the Preliminary Phase. If the project will proceed, this document can be used later, in the Formal Definition phase, as the base document for a final Submission Agreement.
- Someone must make the official decision to proceed to the next phase. Someone from Digital Library and Information Technologies? The Library Management Team? [We'll have to decide on this policy issue later.]

## Formal Definition Phase [3.2]

### Setting Up The Organization [3.2.1]

#### Setting Up The Organization [3.2.1.1]

- Who will represent the Depositor and the E-Journal Archives during this phase? How will planning of the subsequent phases and document preparation be apportioned?
- Did the preliminary phase leave any points or issues to be decided or explored in detail during this phase? They must be made part of the current process.

### Formal Definition [3.2.2]

#### Information To Be Preserved And Model Of Data Objects To Be Delivered [3.2.2.1]

This activity defines the information to be transferred as precisely as possible. Together, the Depositor and the E-Journal Archives create a formal object model that represents the data to be preserved. Much of the necessary information will have been gathered during the Preliminary Phase. On completion, the model will serve as a checklist for data preparation prior to transfer and on receipt. Furthermore, as it will document the interrelationships among the many pieces of information needed for long-term preservation.

The authors of the Producer-Archive Interface Methodology Abstract Standard suggest:

*“3 main work stages are required to create this model:*

- 1. Description of the general objectives and project context, definition of all the Information Objects, definition of the coding, format, Information Object identifiers, in the form of a text document. All of these points have already been studied in the preliminary phase.*
- 2. Definition of the object classes associated with the aforementioned Information Objects, and creation of an associated dictionary to list these definitions.*
- 3. Construction of the formal model of the Archive Project.”*

#### General Project Context And Definition Of Information Objects [3.2.2.1.1]

- What is the context for the project and what information objects can be delivered?
- For each of the information object types, what format will the data be delivered in? What coding rules will be followed? What standards will be applied?
- How much data will be transferred? Total volume of Content Data? Maximum file size? Mean file size? Any other information that will aid in estimating storage costs and data management?
- What is the final definition of the identifiers to be used?
- Will the Depositor have to install any tools to help with data production, conversion, or documentation?
- Write a natural language description of the Information Objects, using all the semantic and syntactic information gathered so far. This description will be used to create a formal data dictionary and a formal model. . **This description will become part of the Submission Agreement.**

### *Creation Of A Dictionary [3.2.2.1.2]*

- What are the definitions of the classes of Data Objects? What information and attributes are associated with each class?
- Create the project's data dictionary. The Producer-Archive Interface Methodology Abstract Standard states that the dictionary should conform to the [Data Entity Dictionary Specification Language \(DEDSL\)](#) (2). An XML DTD mapping for the DEDSL is provided in a related [CCSDS Red Book](#) (3). . **This data dictionary will become part of the Submission Agreement.**

[NOTE: This might be a reasonable place for the E-Journal Archives to decide not to conform to the Producer-Archive Interface standard. These specifications were written in the context of the CCSDS's space data environment. Ongoing work by the OCLC/RLG Working Group on Preservation Metadata should inform our choice of "data entities" and the language we use to specify them. At a higher level of data organization, I haven't yet compared the DEDSL XML DTD with what we might do with METS. -wrk1]

### *Construction Of A Formal Model[3.2.2.1.3]*

- What view or views of the Data Objects will be most useful in for modeling the set of information to be delivered and preserved? At what level of granularity? Are there time-dependent views? *"This data or set of Data Objects are the basis for the definition of the SIPs."*
- Create a model or model, preferably using the Unified Modeling Language (4). This model will become part of the Submission Agreement.

## **Formalization Of Contractual And Legal Aspects [3.2.2.2]**

- Has the need for contracts between the Depositor and the E-Journal Archives been determined in the preliminary phase? If so, draw them up and **prepare the part of the Submission Agreement that documents them.**

## **Definition Of Transfer Conditions [3.2.2.3]**

### *Communication procedures [3.2.2.3.1]*

- By what means will the data be delivered to the E-Journal Archives?

### *Packaging [3.2.2.3.2]*

- How will the identity and structure of the mass of data transferred be described? In OAIS terms, what is the Packaging Information for the transfer?

### *Data Submission Session [3.2.2.3.3]*

- What will be considered a submission session? What will be transferred during a session? Will the data be delivered in a series of separate sessions with different types of data objects? What will be the time frame for a session or group of sessions? How will the Depositor and E-Journal Archives communicate information about the transfer?

### *Define the initial transfer test [3.2.2.3.4]*

- How will the transfer procedure be tested? What will be the test SIP (based on the data description, dictionary, and model described previously)? It should be constructed to test the integrity, performance, scaling of the transfer process.

### *Tools for the transfer [3.2.2.3.5]*

- What tools will be used during the transfer, if any?

### *Transfer procedures [not numbered]*

- Prepare the part of the Submission Agreement that describes the transfer procedures.

### **Validation Definition [3.2.2.4]**

- How will the E-Journal Archives validate the integrity, conformity, and completeness of the data?
- If rejection criteria will be based on the quality of the data transferred, what types of errors will the E-Journal Archives reject? This sort of validation might not be carried out immediately on receipt. It will include the automatic and manual quality control checks that will be made on the content.
- How will the Depositor and the E-Journal Archives communicate and negotiate about rejections and repeat transfers.
- How will the validation process be tested? Define a test SIP for validation and the tests that will be performed on it after the E-Journal Archives receives it. The SIP should be based on the data description, dictionary, and model described previously.
- Are there any tools that will be needed by the E-Journal Archives to perform the validation?
- **Prepare the part of the Submission Agreement that describes the validation procedures.**

### **Delivery Schedule [3.2.2.5]**

- What will be the delivery schedule? **This schedule will be part of the Submission Agreement.**
- What will happen if the delivery schedule isn't met?

### **Change Management During The Life Of An Archive Project [3.2.2.6]**

- What topics already discussed might change? For example, the results of the transfer and validation tests might make changes to the Submission Information Package necessary. A later need for format migration is another obvious change agent.
- How will intentional changes to archived data be handled? Changes to the preservation metadata?
- Will a time come when the Submission Agreement itself should be changed? How will it be done?
- **Prepare the part of the Submission Agreement that describes plans for change management.**

### **Feasibility And Assessment [3.2.2.7]**

- What will the archiving project cost?
- Will the project proceed?

### **Submission Agreement [3.2.3]**

Bring everything prepared during the Formal Definition Phase together. Review it all and make it official.

[NOTE: It seems that the following sequence suggested by the Producer-Archive Interface Methodology Abstract Standard should be reordered so that the transfer test is followed by the validation test. Only after the Depositor and E-Journal Archives have identified and fixed any problems that arise, should they begin transferring the actual SIPs. Because the transfer might occur over several sessions, validation management could occur subsequent to each session.]

## **Transfer phase [3.3]**

### **Carry Out The Transfer Test [3.3.1]**

- What does the transfer test show? Does the definition of the transfer in the Submission Agreement need to be modified?

### **Manage The Transfer [3.3.2]**

- Does the transfer happen as planned? Are the depositor and E-Journal Archives communicating about problems?

## **Validation Phase [3.4]**

### **Carry Out The Validation Test [3.4.1]**

- Does the transferred data meet the structural, syntactic, and semantic criteria defined in the Submission Agreement?

### **Manage The Validation [3.4.2]**

- Does the data pass quality control tests?
- Make it right, according to the Submission Agreement.

## **References**

1. Consultative Committee on Space Data Systems, *Producer-Archive Interface Methodology Abstract Standard*, 2001. <<http://ssdoo.gsfc.nasa.gov/nost/isoas/CCSDS-651.0-W-1.pdf>>
2. Consultative Committee on Space Data Systems, *Data Entity Dictionary Specification Language (DEDSL) - Abstract Syntax (CCSD0011)*, 2001. <<http://www.ccsds.org/documents/pdf/CCSDS-647.1-B-1.pdf>>
3. Consultative Committee on Space Data Systems, *Data Entity Dictionary Specification Language (DEDSL) - XML/DTD Syntax (CCSD0013)*, 2001. <<http://www.ccsds.org/documents/pdf/CCSDS-647.3-R-1.pdf>>
4. Object Management Group, *Unified Modeling Language (UML), version 1.4*, 2001. <<http://www.omg.org/technology/documents/formal/uml.htm>>