HARVARD UNIVERSITY
Information Technology
LIBRARY TECHNOLOGY SERVICES

Page Delivery Service:
Workflow and Checklist

The purpose of this document is to assist the collection curator in preparing materials for DRS deposit and delivery by the Page Delivery Service (PDS). Preparations involve decisions about:

- reformatting of the source materials,
- structure and navigation of the digital text,
- full text searching of the digital text,
- assigning persistent identifiers (URNs) to components of the digital text, and
- the agent responsible for depositing PDS-delivered files.

Contents

1   What is Page Delivery Service (PDS)?

The Page Delivery Service (PDS) delivers to a web browser scanned page images of books, diaries, reports, journals and other multi-page documents stored in the HUL Digital Repository (DRS). Documents delivered by PDS can be used in ways similar to their print counterparts, for example, browsed through a table of contents or viewed page-by-page.

PDS reproduces the structure of multi-page documents by organizing their contents into hierarchical levels. There are three basic levels of organization (although additional levels are possible): **citation (top) level**, **intermediate level**, and **page or leaf level**. These levels are sometimes referred to as **nodes**.

Details about a document's organization are captured in XML-based **structural metadata** files that are used by PDS to display the document and make it navigable.

Consult the PDS section of the OIS web site for additional information about PDS: http://hul.harvard.edu/ois/systems/pds/

2   Prerequisites to PDS use

2.1   Register as a DRS owner

Any Harvard organizational entity that is registered to use the Digital Repository Service (DRS) is eligible to use PDS. All of the digital files associated with digital objects to be delivered by PDS must be stored in DRS. Consult the list of registered owners for DRS to check if your organization is already registered. If not registered, read the DRS How to join instructions.

List of registered DRS owners: http://hul.harvard.edu/ois/systems/drs/owners.html

DRS "How to join" instructions: http://hul.harvard.edu/ois/systems/drs/join.html

2.2   Register to use the Name Resolution Service (NRS)

The major access points to your digital texts (delivered by PDS) will be assigned NRS persistent identifiers during deposit to DRS. You also have the option of assigning persistent IDs to individual pages of text *after* deposit by using the NRS Web Administrative Interface.

Your organization must have an active NRS authority path established in order to assign persistent identifiers. For more information about NRS and how to register, consult the NRS How to Participate procedures on the OIS web site.

Instructions for NRS registration: http://hul.harvard.edu/ois/systems/nrs/nrs-intro.pdf

Instructions for adding persistent IDS: http://hul.harvard.edu/ois/systems/nrs/addPID.pdf

See Section 5 for more information about assigning persistent IDs to PDS-delivered materials.

3 Making decisions about page reformatting

3.1 PDS file requirements

The Page Delivery Service can deliver page content as page image files, as plain text files, or both page images and plain text. The use of PDS requires that these image and text files, along with a structural metadata file, be stored in the DRS.

Page images of a document can be bitonal, grayscale, or color. PDS delivers page images in JPEG or GIF format, and can create delivery images from JPEG2000 or TIFF masters. The following table describes these options:

Page image deposited as:	PDS will deliver as:
JPEG2000	JPEG
JPEG	JPEG
GIF	GIF
TIFF	GIF

If more than one of these formats is available for a document, PDS will use this order of preference: (1) JPEG2000, (2) JPEG/GIF, (3) TIFF.

Plain UTF-8 encoded text for a document is required for DRS in order for the text to be searchable in PDS. Plain text files (sometimes called **full text files**) are created by using OCR (optical character recognition) during the page scanning process or by manually rekeying the text. These plain text files can be used only to power keyword searching in PDS or the plain text can be offered as a display option. It is also possible to have PDS offer plain text as the only display option for a document.

3.2 Reformatting issues to consider

These are some of the issues you will want to consider when preparing for text digitization. Discuss these with the reformatting lab that will handle your materials. Consult the Library Digital Initiative web site for more information about text digitization.

- Determine whether page images, plain text, or both need to be produced. If these already exist, do they meet PDS requirements? Consult PDS file requirements (above).

- If plain text needs to be produced (to enable searching), should it be generated by rekeying text or by OCR during the scanning process? Does your selected reformatting vendor handle plain text production in-house or subcontract this work out?

- What are your requirements for preparation, transfer, handling, and disposition of the source materials? Harvard conservators are available to advise on these issues.

- Do you want the OCR or keyed text visible in PDS? There is a Harvard METS option to disable text viewing.

3.3 Selecting a reformatting vendor

Materials delivered by PDS have relatively complex metadata requirements. While it is possible to produce this metadata manually, it would be best to employ a reformatting vendor with expertise at handling metadata.

OIS recommends the services of the HL Imaging Services (HL IS). The HL IS offers a web-based **Data Collection Tool** for recording metadata for PDS-delivered materials. See Section 6.2 for more information about this Data Collection Tool.

4 Making structural metadata decisions

It is important to understand the structure of the original material in order to make informed decisions about the structure of and access to the digital text. These decisions will be reflected in the structural metadata generated for your digital texts.

Use the Structural Metadata Worksheet in Section 4.6 to record your decisions about the hierarchical organization of your materials.

4.1 Document structure in PDS

PDS reproduces the structure of multi-page documents by organizing their contents into hierarchical levels. There are three basic levels of organization, although additional levels are possible: **citation level**, **intermediate level**, and **page or leaf level**. These levels are sometimes referred to as **nodes**.

Citation level represents the top of a document's hierarchy. For a single document, the citation level describes the whole document at the title level. Often, citation level is the first thing a user sees after following a reference from an online catalog or specialized web site. In PDS, the title of the citation level displays at the top of the screen.

For a collection of documents (e.g., a series of annual reports), the citation level would describe the collection. Depending on a document's complexity, the citation level will link directly to individual pages (page or leaf level) or to an intermediate level with more detailed links to contents.

Intermediate levels are used to represent the structure of a more complex document (or collection of documents). The intermediate levels usually consist of a list of contents. In PDS, the intermediate levels of a document or collection of documents display as a table of contents in the left-hand pane. In very hierarchical collections, such as an archive of journals, there may be more than one intermediate level (e.g., a journal collection consisting of individual volumes, issues, and articles). An intermediate level is optional and may not always be present.

Page (leaf) is the lowest level, representing individual page images.

4.2 Evaluate original source material

Evaluate the relationship between the physical format and the intellectual content of the original source material in order to determine access points for the digital object.

Example: Milman Parry Collection of Oral Literature (MPCOL)

Example of original source material: The Milman Parry Collection of Oral Literature, held by Harvard's Center for Hellenic Studies Library, includes handwritten notebooks of songs. The collector assigned each song a number and sometimes one song can span across multiple notebooks. As a result the physical format of the collection (the notebooks) do not exactly line-up with the intellectual content (the songs).

Notebook 1	Notebook 2	Notebook 3
Song 4 Song 5 Song 6	Song 7 Song 8 Song 9 Song 10	Song 10 cont'd Song 11 Song 12

4.3 Determine access points for the digital objects

Based on the evaluation of the original source material, determine the desired intellectual access points for the digital object. This decision will be reflected in the creation of metadata about the digital objects.

Example: Milman Parry Collection of Oral Literature (MPCOL)

There are two possible access points for the digital version of the MPCOL song notebooks. One option would be to **follow the physical format** of the collection so researchers would have the same access to the physical and digital versions of the material. The other option is to **follow the intellectual content** of the collection so researchers have access to individual songs and do not need to know which notebook contains the song.

Access by physical format (by Notebook):	4.3.1.1	Access by intellectual content (by Songs):
Notebook 1 Notebook 2 Notebook 3	OR	Song 4 Song 5 Song 6 Song 7 Song 8 Song 9 Song 10 Song 11 Song 12

There are advantages and disadvantages to each decision:

Access by Notebooks (physical format)

Advantages

1. Digital and physical format have the same access points

2. Less preparation and *flagging* will have to be done to the original source material for reformatting

Disadvantages

1. Digital version does not take advantage of the flexibility of the technology to deviate from the fixed physical format of the material

2. Researcher would have to know which notebook contains which songs

3. Songs that span across more than one notebook would be difficult to view

Access by Songs (intellectual format)

Advantages

1. Individual songs are easy to access

2. Researcher would not have to know which notebook contains which songs
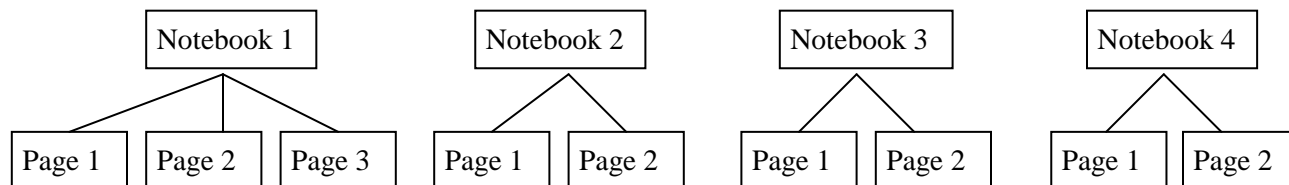
Disadvantages

1. It would be difficult to replicate the original collection and structure of the materials

2. Digital and original versions would have different access points

4.4 Determine digital object structure: simple or complex

For each object that will be delivered by PDS, collection managers need to identify the hierarchical organization of the object and supply a small amount of information about the object and its subsidiary parts. This information will allow PDS to reproduce the digital text and make it easy for the user to navigate through the text.

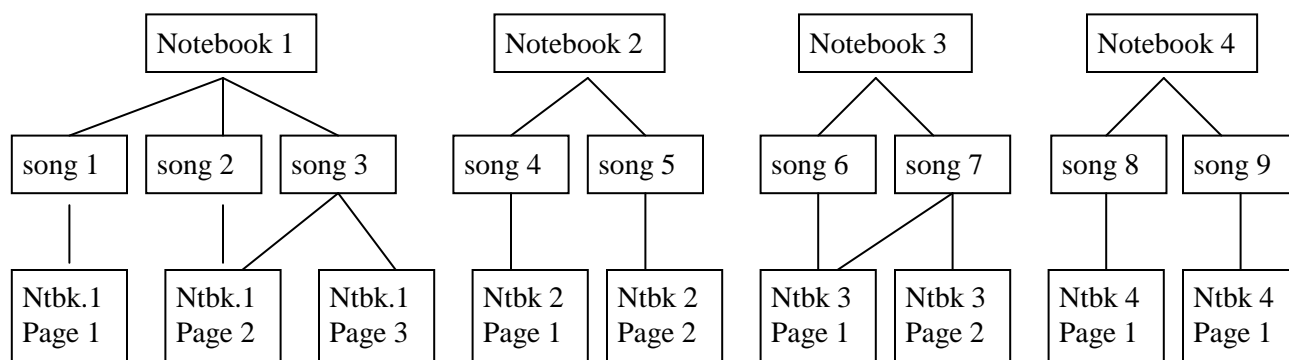The text may have a **simple structure** or a **complex structure.** A text with a simple structure provides one high level of access to its digital page objects. Text with a complex structure provides one high level of access (citation level) in addition to one or more intermediate levels of access (such as chapters or sections). PDS can support an unlimited number of intermediate access levels.

Example of simple structure: The Milman Parry collection could be organized as a simple object. In the diagram below, the single high level access points to this collection are the individual notebooks, which link to pages containing songs. This approach closely models the physical format of the collection. But if access by song is considered important, this approach is less desirable than the complex structure in the next diagram.

| Notebook 1 | Notebook 2 | Notebook 3 | Notebook 4 |

| Page 1 | Page 2 | Page 3 | Page 1 | Page 2 | Page 1 | Page 2 | Page 1 | Page 2 |

Text object with simple structure (one level of access)

Example of complex structure: the diagram below illustrates how the Milman Parry collection might be organized with several levels of access. Highest level of access is still by notebooks, but with an additional level providing access to individual songs.

| Notebook 1 | Notebook 2 | Notebook 3 | Notebook 4 |

| song 1 | song 2 | song 3 | song 4 | song 5 | song 6 | song 7 | song 8 | song 9 |

| Ntbk.1 Page 1 | Ntbk.1 Page 2 | Ntbk.1 Page 3 | Ntbk 2 Page 1 | Ntbk 2 Page 2 | Ntbk 3 Page 1 | Ntbk 3 Page 2 | Ntbk 4 Page 1 | Ntbk 4 Page 1 |

Text object with complex structure (multiple levels of access)

For complex objects, PDS will display each level of the hierarchy as its own table of contents.

4.5 Select PDS interface options

The web-based interface to PDS documents includes a few optional features that are largely controlled by the presence (or sometimes absence) of information in the document's METS file. The curator or project manager preparing materials for delivery by PDS should consider these features in consultation with the reformatting and deposit vendor.

4.5.1 Displaying plain text

PDS uses plain ascii text for a document to support keyword searching, but this text can also be offered as a display option. In the PDS interface, the user clicks the "View Text" button in the toolbar to display plain text. To de-activate the display of plain text, a flag must be set in the METS metadata file. If the plain text option is not available, the "View Text" button will be gray.

4.5.2 Offering "Go to page number"

The PDS interface provides the option of jumping to a specific numbered page in a document. This "Go to page number" option displays in the PDS toolbar:
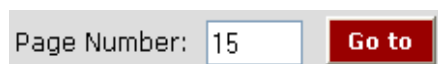
There may be instances in which the Go to page number option would provide confusing results for the user. Examples would be documents in which many or most pages are not physically numbered, or large groups of documents (such as journal runs) in which the same page ranges occur multiple times.

To de-activate the Go to page number option, a flag must be set in the METS metadata file. If this option is deactivated, the page number display and "Go to" button will not appear in the toolbar. Deactivating this option has no effect on availability of the "Go to" page sequence number option, which is mandatory.

4.5.3 Displaying related links

Inside the METS metadata file for a PDS document, it is possible to record descriptive information about the document, including links to information that is external to PDS. Examples would be a link to a description of the document in HOLLIS, to a finding aid in OASIS, or to a specialized web site with additional information.

In the PDS interface, this information displays in the **Related Links box** in the toolbar. Clicking on a link in this box will display the external information in a new browser window.

If descriptive metadata is not included in the METS file, the Related Links box in the PDS toolbar will be empty.

4.6 Structural metadata worksheet

Use this worksheet to record your decisions about the hierarchical organization of the digital objects that will be delivered by PDS.

Project title/description: _____

Simple Structure—One Level leading to the page image

I.

II.

III.

Complex Structure—Two Levels leading to the page image

I.
     1.
     2.

II
     1.
     2.

III
     1.
     2.

Complex Structure—Three Levels leading to the page image

I.
     1.
          a.
          b.
     2.
          a.
          b.
     3.
          a.
          b.

II.
     1.
          a.
          b.
     2.
          a.
          b.

5 Assigning persistent identifiers (URNs)

NRS persistent identifiers (also called URNs) can be assigned to objects delivered by PDS in order to ensure permanent access to these materials from online catalogs or special collection web sites. You can assign a URN to a PDS document as a whole (at the citation level) at the point of DRS batch deposit. URNs to individual parts of a PDS document must be assigned after deposit, using the NRS Web Administrative interface (known as NRS Web Admin).

The recommended practice is to assign a URN to the logical document as a whole (i.e., to its citation level). Assigning URNs to individual parts of a PDS document is not recommended unless there is significant intellectual or curatorial value to maintaining an access point to individual parts.

URNs requested during DRS batch deposit must be encoded within the DRS batch XML file. Consult the DRS Load Manual for more information about assigning URNs during batch deposit.

To request a URN after batch deposit, use NRS Web Admin. For this method, all digital objects, including page images and the structural metadata (METS) file, must be deposited in DRS before assigning the URN. See the document Creating an NRS Persistent Identifier for more information.

5.1 Assigning URNs to the citation level

A URN assigned to the citation level of a PDS document will take the user directly to the first page of that document (or group of documents). A URN to the citation level is usually assigned at the point that PDS files are deposited to DRS but can also be assigned after deposit, using NRS Web Admin.

Sample scenario. A diary will be scanned and its pages delivered by PDS. A URN will be assigned to the citation level because this is the point at which users will access the diary (from its description in an online catalog). The URN, to be assigned at batch deposit, will point to the citation level of the diary. Once created, this URN will resolve to a PDS URL with this format:

http://pds.lib.harvard.edu/pds/view/xxxxxxxx

where xxxxxxxx is the DRS ID for the diary's METS (structural metadata) file.

5.2 Assigning URNs to other levels

A URN assigned to a specific part of a PDS document that is below the citation level will take the user to an individual page in the document. This could be the first page of a section or any other page within the document.

A URN to any part of a PDS document other than citation level must be assigned after deposit, using NRS Web Admin. This method requires that all files associated with the PDS document, including page images and structural metadata (METS) file, be deposited first.

To capture the URL that will be associated with the URN, you will navigate through the PDS document to the page of interest and copy the URL for that page from the browser address window. The PDS URL should have a format similar to the following:

http://pds.harvard.edu:8080/pdx/servlet/pds?id=410820&n=5

where the parameter "id=" contains the DRS ID for the document's METS (structural metadata) file and the parameter "n=" contains the sequence number of the page. Note that the PDS URL that you capture may contain additional parameters that control other display aspects of the page (e.g., magnification level, rotation angle, etc.). These parameters are usually not important and can be left as is or deleted. If you remove these parameters, be sure to test the resulting URL to insure that it works in a browser.

Sample scenario. The Countway Library Department of Rare Books and Special Collections is digitizing historical Medical School images which will be accessed through an electronic finding aid in OASIS. In addition to providing links to the images, the finding aid will include links to specific pages of information in PDS about the Medical School from the University Archives digitized Harvard & Radcliffe Annual Reports. Through the use of links to both the Medical School and Archives collections, researchers can read about renovations to a Medical School building in a particular year and also view the associated images of the building.

To do this, the Countway staff member who creates the finding aid would browse the Annual Report web site to find a section of interest. Then, copy the URL to the first page in this section and use the NRS Web Admin to associate a URN with that URL. Finally, the staff member would insert the URN into the appropriate part of the finding aid so the user can link out to the Annual Report section about the Medical School.

Note about stability of URN-based access to PDS materials: OIS is obligated to maintain the PDS system and the URNs that reference PDS documents. If the PDS system is changed, OIS will programmatically modify PDS URLs that are associated with URNs. But, if you assign a URN to a PDS object that is not directly under your control, be aware that this URN will continue to work contingent on availability of the object. The owner of the PDS document might modify the structure or content of the document, or might withdraw it entirely. In these cases, you are obligated to periodically check that your URN to a PDS document still resolves correctly.

6 Planning for DRS deposit

PDS-delivered objects must be stored in the HL Digital Repository (DRS). The special structural metadata requirements of these objects require that they be deposited in batch. PDS Batches are created using the DRS Batch Builder software. PDS-related files are not eligible for deposit using the DRS Web Admin interface.

This section provides more information about batch deposit requirements and a description of the Multi-Page Metadata service offered by the HL Imaging Services (HL IS).

6.1 DRS batch deposit requirements

The descriptive, structural, and administrative metadata for a PDS document (or group of documents) are stored in a single XML file marked up according to the **METS schema** (version 1.3). Information about METS, including a link to the version 1.9 schema, is maintained by the Library of Congress.

The Harvard METS profile for PDS documents is available here:

http://hul.harvard.edu/ois/systems/pds/Harvard_METS_Profile_for_Page-Turned_Objects.doc

For more information about metadata requirements for batch deposit of PDS files to DRS, contact OIS by sending filling out PDS Comments Form.

In order for a document to be delivered by PDS, the following files must be deposited to DRS:

- Page image files,

- Plain text files (UTF-8 encoded text created by OCR or rekeying; for full text searching, if being used),

- A batch deposit file formatted in XML according to the DRS batch DTD, and

- A METS file of structural metadata, formatted in XML according to the Harvard PDS profile of the METS schema.

6.2 HL Imaging Services Multi-Page Metadata service

Although PDS structural metadata files can be created manually, it is easier to supply the information on a web-based form and have the XML tagging generated automatically. The Harvard Library Imaging Services (HL IS) offer a **Multi-Page Metadata service** to its customers that generates PDS structural metadata files from web form input.

6.2.1 Who can use the Multi-Page Metadata service?

This service is available to all customers of the HL Imaging Services.

6.2.2 How to get more information or to register?

If you want to register or need more information, contact the HL Imaging Services (hcl-dig@fas.harvard.edu or 617-496-9346).

7 PDS Worksheet

Keep this document for your records.

7.1 Action Item Check List

	Task/Decision	Completed?
1	Evaluate the original source material	
2	Determine access points for the digital objects.	
3	Determine the structure of the digital objects (simple or complex).	
4	Prepare for reformatting.	
5	Deposit to DRS.	
6	Provide access to specific pages if necessary.	
7	Register to use the HL IS Data Collection Tool	

7.2 Decisions

A.   Selection criteria

_____
_____
_____
_____
_____
_____

B.   Access points

_____
_____
_____
_____

C.   Structure – simple or complex?

_____

D.   What needs to be deposited?

_____
_____
_____
_____

E.   What persistent identifiers (URNs) need to be assigned?

_____
_____
_____
_____

Revised: April 24, 2013

8 PDS resources

PDS information on the OIS web site:
http://hul.harvard.edu/ois/systems/pds/

METS schema and documentation:
http://www.loc.gov/standards/mets/

Harvard PDS profile:
http://hul.harvard.edu/ois/systems/pds/Harvard_METS_Profile_for_Page-Turned_Objects.doc

Image Metadata Supplement (applies to page images)
http://hul.harvard.edu/ois/systems/drs/imagemetadata.pdf

NRS information
http://hul.harvard.edu/ois/systems/nrs/

DRS information
http://hul.harvard.edu/ois/systems/drs/

DRS Batch Builder
http://hul.harvard.edu/ois/systems/drs/bb.html