InDiPres: Preserving Indiana’s Memories

InDiPres Guidance Document & InDiPres Technical Appendix

Indiana State Library, Indianapolis 2016-2017

This project was made possible in part by the Institute of Museum and Library Services and the Indiana State Library
InDiPres: Preserving Indiana’s Memories

Guidance Document

Prepared by InDiPres Ingest Pathways and Workflows Working Group

William Knauth, Indiana State University, Chair
Sudha Anand, DePauw University
Jill Black, Indiana State Library
Carly Dearborn, Purdue University
Rachel Howard, University of Louisville
Cinda May, Indiana State University
Sam Meister, Educopia Institute/MetaArchive Cooperative
Stephen Patton, Indiana State University
Connie Rendfeld, Indiana State Library
Eric Schmidt, Butler University
Matt Schultz, Grand Valley State University
Deanna Ulvestad, Greene County Public Library, Xenia, Ohio

Indiana State Library, Indianapolis 2016-2017

This project was made possible in part by the Institute of Museum and Library Services and the Indiana State Library
# Contents

## Introduction

I. Local Management of Digital Content

1. Collection Control and Organization
2. Creating an Inventory Instrument for Local Digital Collections
3. Content Selection
4. Selection Criteria
5. Reviewing Content
6. Importance of Local Metadata

II. Collection Preparation for Ingest

1. Collection Preparation for Digital Preservation
2. Ingest Metadata for Bagger and the MetaArchive Conspectus at ISU

III. Ingest Workflow

1. Preparation of Content by InDiPres Members
2. Content Transfer to InDiPres
3. Content Preparation by Metadata Specialist
4. Ingest Preparation
5. MetaArchive Infrastructure
6. MetaArchive Tests and Digital Distribution
7. Content Restoration

IV. InDiPres Digital Preservation Policy Creation Guide

Part I: Policy Introduction
Part II: Policy Elements

V. Technical Appendix

Part I: Bagger Guide

1. Overview of Bagger/BagIt
2. Getting Started with Bagger/BagIt
3. Bagger Interface
4. Bagger Metadata
5. Data Upload and Bag Saving

Part II: MetaArchive Conspectus & Ingest Infrastructure
Introduction

Using Library Services Technology Act (LSTA) Digitization Special Grant funding provided by the Institute of Museums and Library Services (IMLS) as seed money, Indiana State University partnered with the Indiana State Library to establish Indiana Digital Preservation (InDiPres) in February 2016 to address the digital preservation needs of Indiana organizations, especially small to mid-sized under-resourced institutions that create, manage and provide access to content in a digital format. The mission of Indiana Digital Preservation is to collaboratively manage and sustain a low-cost, secure, and geographically distributed archive for the long term preservation of locally sponsored digital resources through its membership in the MetaArchive Cooperative Preservation Network. InDiPres is a community-based organization governed by its members, who share in the expense of digital preservation.

InDiPres Ingest Pathways and Workflows Working Group

The InDiPres Ingest Pathways and Workflows Working Group convened at the Indiana State Library in Indianapolis on October 14, 2016. Chaired by the InDiPres Metadata Specialist based at Indiana State University, the meeting brought together representatives from InDiPres member organizations, Educopia Institute, and the MetaArchive Cooperative. Participants discussed and reached practical decisions on appropriate avenues for the ingest of content contributed by a wide array of InDiPres members, the development of workflows to support the ingest pathways, and the creation of a Guidance Document to record best practices and workflows for InDiPres members and other interested parties.

I. Local Management of Digital Content

This section of the Guidance Document will provide information and suggestions on collection management and organization for InDiPres member institutions. Providing support assistance with digital collection management is an important component of digital preservation and of the InDiPres collaboration.

1. Collection Control and Organization

1.1 Importance of Organization and Management

A key piece of the InDiPres collaboration is to offer its members the guidance and best practices that empower staff to take physical and intellectual control of local digital collections. When developing and implementing policies for the creation, organization and management of digital content it is important that decisions are documented and followed for consistency. Writing a
local digital preservation policy and plan to address your institution’s collections and establishing it as part of your overall collection development or management document will ensure that digital preservation becomes an integral part of your institution’s mission. On the collection level, categorizing digital content, its format, extent, organization, available metadata, and storage locations are vital components of appropriate stewardship of digital assets for long term preservation. Creating an inventory instrument to record detailed information about your institution’s digital collections will assist you in tracking the characteristics and life cycle of its digital assets.

1.2 File Naming Guidelines

It is important to establish appropriate and accurate file naming conventions for content being digitized. Decide upon a short, simple set of naming practices and consistently apply them. Create file name guidelines that fit the needs of your individual institution and collections. Abbreviate when possible, but the more descriptive the file name the easier it will be to recognize the content being represented. Be consistent in structuring the elements of the file names such as the presentation of dates or proper nouns. Each InDiPres member is assigned a 2-3 letter organizational code followed by a dash that precedes the local file name. This code or acronym differentiates the content of each InDiPres member on the file level.

File Naming Best Practices

1. Do not use spaces, because they are not recognized by some software. Instead use underscores (file_name), dashes (file-name), no separation (filename), or camel case (FileName).

2. Avoid special characters: ~ ! @ # $ % ^ & * ( ) ` ; < > ? , [ ] { } ‘ “. For the bagger program including any of the special characters disrupts the program’s data packaging and will lead to a crash.

3. A good format for dates is YYYYMMDD (or YYMMDD). This makes sure all your files stay in chronological order.

4. Don’t make file names too long; longer names are more difficult to read in a directory system and can invite reading errors. Limit file names to 25 characters or less if possible.

5. For sequential numbering, use leading zeros to ensure files sort properly. For example, use “0001, 0002...1001, etc” instead of “1, 2...1001, etc.”

6. To manage drafts and revisions include a version number in the file name.
1.3 File Naming Examples

The three elements work together to locate a specific file by its institution, the parent collection, and its context within the original order of the collections materials:

1. **Acronym**: derived from title of InDiPres Member.

2. **Collection Name**: This should be the title of the individual collection.

3. **File Descriptive**: Any combination of number or letter organizational information used in file naming. This could be the page number from a volume, or sequence number from a photograph collection series, the date of the material, or anything of that nature.

*Example of blank format for InDiPres files:*

Acronym_CollectionName_XXX_ABC

*Examples of InDiPres file formats:*

SCPL_CivilWar_032_04
KCPL_CivilCourt_0011a
ALA_Peterson_2011_02
RHIT_Peddle_0034c

2. Creating an Inventory Instrument for Local Digital Collections

2.1 The Importance of the Collections Inventory

The inventory serves as an overall guide to the digital collections held by your organization and should be updated with new information as additional content is created, decisions are made, and/or actions taken. It is suggested that the inventory document be the responsibility of one individual in order to keep the data entry consistent. Develop an inventory instrument that makes sense for your institution and its needs. To create the inventory use software with which you are already familiar and which is easy to use. A spreadsheet program like Microsoft Excel or OpenOffice Calc are two useful basic software tools for collection inventory.

Inventory all the digital content currently held by your institution. This should include in process and completed digital collections. The goal of this effort is to ascertain and document what digital assets are present, their category, date of creation, extent, format, available metadata, and storage location.
Basic Example of an Excel Digital Collection Inventory

### 2.2 Categories of Content

Assign each digital collection a type category based on the nature of the collection content. Examples of these content categories such as “Photographs,” “Documents,” “Public Records,” “Oral Histories,” and “Audio” or “Video” are appropriate designations.

You will also want to note the origin of the digital collections and whether the content has an “analog” counterpart or if is “born digital.” Analog to digital content means that a given collection is a digital facsimile of a real physical set of materials. Born Digital on the other hand indicates that a collection was generated from its inception as digital data and does not have a corresponding physical version.

### 2.3 Date of Creation and Extent

An important piece of technical information for digital files is their age and total size. Include in the inventory the date each collection was digitized as well as its approximate data size. It is important for InDiPres Metadata Specialist to know the extent of a collection as this affects processing time and the collections interaction with MetaArchive. Also knowing how long digital data has existed is critical to judging its potential for degradation. The risk of degradation is an important factor in determining preservation priorities.

### 2.4 Available Metadata

Note the availability of both item and collection level metadata. Item level metadata will be packaged with the related data files for ingest into the preservation network; and thus, will be retrievable along with the content should file restoration be required. Collection level metadata is used to describe the collection as a whole at different points of processing.
2.5 Formats

Record the file formats of your digital collections. The MetaArchive Cooperative Preservation Network is platform agnostic and will accept all types of data. However, it is important to know the file formats of your data in case it is necessary to migrate from an obsolete platform into a new format. Using stable file formats in the creation of digital content helps ensure the longevity of the data. Examples of stable formats are: TIFF, JPEG, PDF and WAV.

2.6 Storage Locations

Add the location where digital collection data is stored to the inventory. This could be online, offline, or through the use of external media such as tapes, DVDs, or CDs. If you are using network storage specify the exact file path where the collection is stored. If you have multiple copies note where each file resides.

3. Content Selection

This portion of the InDiPres Guidance Document will offer assistance with this decision making. It will set out best practices and guidelines to consider in determining member preservation priorities when selecting content for preservation. Individual InDiPres members are considered to be the experts on their digital materials. It is up to each member to decide which of their digital collections require long term preservation, and the order in which they are submitted for ingest into the MetaArchive Preservation Network.

3.1 Risk Factors

Risk factors are threats to the integrity and accessibility of digital material. Digital content is far more fragile than its analog counterpart and can be permanently lost or damaged for a variety of reasons.

When deciding on what digital content to preserve, there are three major risk factors to consider. Weighing these factors within the framework of your institution’s total holdings allows you to determine the degree of risk that can be assigned to each and aides in prioritizing collections for preservation.

- **Change and Loss**: How vulnerable is the data and byte integrity of the digital collection? Is it possible or likely that the data could become corrupted? The individual bytes of data can corrode and lose their integrity over time if not checked and transferred at set intervals. In this way collection data on a hard drive can
become unusable even if the device is stored safely and undamaged by any other outside force. Data can also be deleted or altered intentionally or by accident.

- **Obsolescence**: A collection’s age, formatting, type of technology device, or access program can be a source of problems for digital content. Technology evolves quickly rendering software and playback devices obsolete. File formats and types also change over time and necessary access equipment or software may become difficult or impossible to locate. Digital content may require reformatting or emulation in a new application in order to retain access to the data.

- **Disaster Damage**: Are there any serious threats to the existence of this digital collection? Disaster damage can come in the form of fire, floods, electrical damage from storms or other threats from natural disasters.

### 4. Selection Criteria

In addition to creating an inventory, it is important to develop selection criteria to help you prioritize data for long term preservation. Developing selection criteria for digital content is similar to creating a collection development policy. Digital collections, selection criteria and the digital preservation plan should be represented in an institution’s collection development policy. Here are some things to consider:

- **Central Institutional Mission**: What is the primary mission of the institution in question? Which collections best support and reinforce that mission and its goals.

- **Collection Development Policy**: Would the preservation of a given digital collection support major elements of an institution’s Collection Development Policy?

- **Required to Keep**: Does this digital collections or its physical analogues have donation agreements or other legal requirements requiring the institution to keep and preserve the content?

- **Uniqueness**: Is the collection unique to its specific repository, or do other institutions hold duplicate digital or analogue copies? Digital collections which are present in other institutions may have a lower risk of loss and this affects their priority for preservation.

- **Significance**: How important might a given digital collection be to future users of the content? How does it relate to persons, events, or other features of importance either for the local region, institutional audience, or for United States as a whole?
• **Frequency of Use:** How much is the collection used by scholars, students, or members of the local community? Digital collections that are in higher demand and significance to the institutions patrons should be given higher preservation priority.

• **Age of Content:** The more time that has passed since the original digitization the more at risk the collection data can be for degradation. Depending on formatting and other circumstances this degradation can occur within 5 years. Additionally if the physical collection the digital scans were made from is very old, fragile, or at risk for damage this also should give the preservation of digital copies of this material priority for ingest into InDiPres.

• **Extent of Content Data:** How many gigabytes of digital content is present in each collection under consideration for preservation. Ingest batches preserved through InDiPres work best when in the area of 30-50 gigabytes. Larger collections can be accommodated by InDiPres, but these require the data to be divided into ingest units of the suggested size.

• **Value of Resource:** It is useful to consider the monetary value of digital collections and their physical parent collection. Consider also the investment in institutional time and resources invested in digitizing this content when making preservation priorities.

• **Who Owns Rights:** This is frequently a topic of concern with digital preservation and digitization. It is suggested that collections to which an institution indisputably holds the copyright be given priority for preservation both for intellectual property reasons and because it may represent a truly unique and valuable cultural heritage resource. The MetaArchive is a dark archive with zero content access outside of the content owner and the MetaArchive staff.

• **Digital Quality:** Consider the quality of the original digitization that created the collections in question. Is it high quality or low quality? Does it accurately capture the original material and is it a useful substitute for viewing the physical items? If digitization was done poorly and resulted in blurry, inaccurate, or low resolution digital copies, the poor quality should weigh against this material being digitally preserved.

5. **Reviewing Content**

An important point in the selection effort is to document the priority decisions and selection criteria made by InDiPres member institutions. It is important to do this to avoid a situation where the effort of this process becomes unavailable and has to be done again. An effective method of doing this, and one suggested generally for organizing digital collections, is to create a spreadsheet with each digital collection listed. Populate this list then with all the relevant information about file name, data size, storage location, metadata, and any other descriptive or
technical information that would be useful depending on the member institution’s unique circumstances. Decision points related to the content selection include the following:

- **Fits Criteria**: How well does a collection score on the several criteria of value and importance used for assessing and prioritizing digital collections for preservation? How to determine and apply this assessment is largely up to individual InDiPres members. An effective method might be to assign a numeric value to the desired criteria listed earlier in this section, rank a potential preservation collection by these metrics, and then use the combined score from this to determine preservation order.

- **Feasibility for Preservation**: What are the major obstacles to a specific collection being digitally preserved through InDiPres? Once the obstacles are identified, determine if they are surmountable or insurmountable.

- **Consistency of Content**: Take note of issues like file naming conventions used, level of description, quality of digitization, file formatting. If there is a great deal of variation on these points within a collection, it could make this material less of a priority for preservation.

### 6. Importance of Local Metadata

#### 6.1 Metadata Significance

The inclusion of appropriate metadata is equally important to keeping digital content useful and secure as the actual content files. Accurate and effective metadata is what will contextualize and explain preserved digital content that is retrieved from the MetaArchive network, if recovery becomes necessary.

#### 6.2 Inclusion of Collection & Item Level Metadata

Metadata is a critical element of digital preservation. When preparing collections to submit to InDiPres for ingest into the MetaArchive Preservation Network it is necessary to provide as much and as detailed a set of metadata as possible. This can include exported data from collection management systems, collection descriptions, and/or other types of contextualizing and descriptive information. Whether exporting metadata from a system or placing it in a document it is important to consider the long term usability of this data as well. Save metadata in a file format that is stable and will be usable long term. For example saving a spreadsheet with metadata in the .CSV file format type, or a written document as a .txt file will ensure the long term accessibility of this metadata information.
6.3 Minimum Metadata Elements

In cases where only limited metadata is valuable from InDiPres members, it is necessary to provide five basic metadata elements. These are the essential descriptions of a digital collection. This is the minimum for member digital content to be preserved. Without this metadata digital, content is likely to be difficult to reintegrate into an institution’s holdings.

*Essential Metadata Elements:*

- **Title:** The collection resource’s name
- **Format:** The file types of the digital information data for the collection
- **Rights:** A statement on the intellectual property rights of a collection
- **Date:** The temporal information about the collection
- **Description:** A short written account of the contents of the collection resource

<table>
<thead>
<tr>
<th>Title</th>
<th>LeFleur Whaling Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Collection Creator:</strong></td>
<td>Richard LeFleur, Archives of Nova Scotia Whaling Historical Society.</td>
</tr>
<tr>
<td><strong>Date Range:</strong></td>
<td>1895-1910</td>
</tr>
<tr>
<td><strong>Accession Date and Records:</strong></td>
<td>Accessioned 1987</td>
</tr>
<tr>
<td><strong>Digitization Date:</strong></td>
<td>Digitized May 2008</td>
</tr>
<tr>
<td><strong>Data Size:</strong></td>
<td>4.33 GB</td>
</tr>
<tr>
<td><strong>Subjects:</strong></td>
<td>Whaling; 19th Century Canada; Nova Scotia, Sailing Ships; Maritime Economy.</td>
</tr>
</tbody>
</table>

*Example of Collection Level Metadata*
Doll collection

Bisque doll

German bisque doll. Leather body, brown glass eyes, brown wig. Mouth is open with teeth. White dress. Doll is in good condition. Stands 26" (66 cm) tall. There is a mark on doll "AM 370".

Armand Marseille (Firm)


Germany

Bisque dolls; Dolls; Domestic Life

Physical Object

image/jpeg

Copyright to this resource is held by the Vigo County Historical Society. It may not be downloaded, reproduced, or distributed in any format without written permission of the Vigo County Historical Society.

http://visions.indstate.edu:8888/cdm/ref/collection/vchs/id/01.jpg Volume1\vchs-dc100.jpg /vchs/image/1.jpg

Example of Item Level Metadata exported from InDiPres member collection metadata in CONTENTdm

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W |
| Rose-Hulman Institute/Keusist House | Peddle, Juliet, 1899- Drawing &lt;br&gt; Rose-Hul 1/1/2005 United Sts Architect, Architect, image image/jpeg Digital Im print |
| Rose-Hulman Institute/Hereman House | Peddle, Juliet, 1899- The Herm &lt;br&gt; Rose-Hul 1/1/2005 United Sts Architect, Architect, image image/jpeg Digital Im print |
| Rose-Hulman Institute/Vigo County Seminary | Peddle, Juliet, 1899- Drawing &lt;br&gt; Rose-Hul 1/1/2005 United Sts Architect, Architect, image image/jpeg Digital Im print |
| Rose-Hulman Institute/Gilbert House | Peddle, Juliet, 1899- The vilbe &lt;br&gt;Rose-Hul 1/1/2005 United Sts Architect, Architect, image image/jpeg Digital Im print |

Rose Hulman Institute of Technology, Peddle Collection Metadata, exported from CONTENTdm and formatted in an Excel Spreadsheet

6.4 Include or Create Descriptive Text

In addition to metadata that is used in house by the institution, members should include: 1. A text description of each collection; and 2. A text description of the InDiPres member institution’s collections and mission. This information is used to describe and identify member collections during the phases of
content ingest. This material can be taken from finding aids or other collection descriptions created for other purposes.

LeFleur Whaling Collection, this material is a digital collection of primarily images related to historical whaling. Collection content includes 19th and early 20th century illustrations, etchings, and photographs and including a small (12) number of letters from whalers and sailors. Collection includes materials related to Richard LeFleur, a locally famous whaling captain and farmer. Collection created by the digitization of an analog collection owned and created by the Archives of Nova Scotia Whaling Historical Society.

Metadata Descriptive Text Example

II. Collection Preparation for Ingest

1. Collection Preparation for Digital Preservation

1.1 Differences among Member Collections

Digital collections being submitted by InDiPres for ingest into the MetaArchive Preservation Network represent a variety of formats and levels of description and organization. Content types so far have included JPEG and TIFF image files, as well as scans of text documents and records, metadata files for collections, and video files.

One of the strengths as well as sources of complication for content preparation for InDiPres is that its members have very different levels of resources. While this can create difficulties for the initial ingest stage of the collaboration, the digital preservation support and assistance InDiPres offers to medium and small cultural memory institutions is a critical feature of the collaboration and is integral for the goal of preserving documentation on the cultural and history of Indiana and its peoples.

1.2 Collection Ingest Pathways

InDiPres recognizes the differences among the membership and in order to accommodate and facilitate the transfer of digital collections for preservation, two methods of data transfer are offered. These two options are the web transfer of collection data directly onto the InDiPres NAS Staging Server, or the transfer of collections via hard drive. The decision of which transfer method to use is left up to members to decide based on their own needs and circumstances.

- If the NAS Staging Server is used, and email invitation will be sent to the InDiPres member’s account that will prompt the creation of a username/password. Once this is accomplished the member will have access to their own specific file within the NAS Staging Server, which will be designated with the member institution’s acronym.
- If a hard drive transfer is used, the member will have to contact the InDiPres staff at ISU to arrange the details of either an in person meeting to exchange hard drives, or postage of drive to ISU.

### 1.3 Steps of Collection Preparation by InDiPres Members

Collection preparation takes three primary forms;

- File naming and restructuring.
- Data wrangling.
- AU/Bag profile Description
- Transfer of data to ISU for ingest, via NAS Staging Server Interface or Hard Drive exchange.

### 1.4 InDiPres Member Acronyms

InDiPres members are assigned an institutional identifier consisting of a 2-4 letter acronym based on the title of the institution. For example, Indiana Digital Preservation would be IDP. This acronym is included in all the different stages of the ingest in order to differentiate each member’s collection. It is suggested that if possible this acronym be included as the first identifier for each individual file as a means to distinguish file ownership.

### 1.5 Recording File Name Changes

Should file renaming and restructuring be necessary, it suggested that these changes be recorded. This can be done in a spreadsheet or other document listing the original file designations along with their new reconfiguration.

This documentation is especially suggested if the original file names used in-house are not being changed, only those being ingested into the MetaArchive.

Situations when file names need revision:

- File names lack any significant description, i.e. are simply a numbered list of files.
- File names lack institutional acronym identifying their owner/creator.
- File name is excessively long, 25 or more characters.
- File names contain spaces or special characters.
- File naming structure is unclear and not useful in identifying collection content.
- File names are highly inconsistent or inaccurate.
2. Ingest Metadata for Bagger and the MetaArchive Conspectus at ISU

A more detailed picture of the Bagger and MetaArchive Conspectus workflow and creation can be found in the attached Technical Appendix document.

2.1 Bagger Metadata

The Bagger program is an integral piece of software for the ingestion of collection data into the MetaArchive LOCKSS preservation network. Bagger makes data packages called Bags of InDiPres member collection data. A description of this program and its functions can be found in the relevant portion of the Guidance Document Technical Appendix.

In addition a tutorial on Bagger can be viewed at the following hyperlink, https://www.youtube.com/watch?v=vK2aKQ64XCQ&t=16s.

This Bag creation process involves the drafting of a Bag Profile for each collection of data. This profile requires a number of key pieces of metadata about the Collection, Bag, and member institution. These pieces of metadata are supplied by the InDiPres member to the collection processor. Most of the elements are not complex and are already known through member participation in InDiPres. A profile is generated in Bagger and this data is included in a Bag when sent to MetaArchive for preservation. A Bag is also the location where Collection and Item level metadata is placed to describe digital collections content. This is a critical part of what makes
up a Bag and is described in detail in the Guidance Document section concerning collection metadata.

Bag Profile Metadata Elements:

Source Organization: Full official name of the InDiPres member institution
Organization Address: Mailing address of the InDiPres member institution
Contact Name: The primary liaison personnel for InDiPres at the member institution
Contact Phone: Phone number for Contact Name
Contact Email: Email address for Contact Name
External Description: A description of the collection which makes up the particular Bag
Bagging Date: The date the Bag was created
External Identifier: The file name of the Bag being created. This identifier must be constructed and recorded accurately and logically, as it will interface in a critical way with the MetaArchive Conspectus system.

2.2 MetaArchive Conspectus Metadata

MetaArchive has a system of member collection management systems called the Conspectus that supports its preservation ingest procedures. Within the Conspectus system are a number of levels of metadata that allow the entire ingest system to work:

Content Provider Metadata: These metadata describe InDiPres as a whole. This information includes the name of the organization as well as its acronym and plug-in prefix. It also lists the name of each member of InDiPres and the status of their Bag/AU content within MetaArchive.

Collection Details Metadata: Individual member information includes the member title, the archive for the collection, the plug-in prefix, Base_URL, and member description. These metadata are added for each InDiPres member institution in a specific page for each collection within the Conspectus system.

Bag/AU Metadata: Descriptive information about specific digital collections partially made up of the collection and item level metadata included in each Bag/AU, as well as the Bag profile metadata. It is also described by the Archival Unit Parameter values, which are pieces of technical description that allow the web harvesting of that specific collection for ingest. This information includes the Base_URL, path directory, and start directory. These three values have to be entered for each Bag/AU correctly for the MetaArchive system to be directed to harvest digital content and distribute it for preservation in the cooperative’s LOCKSS system.
• The Base_URL is the IP address for the NAS webserver owned by InDiPres.
• The path directory for InDiPres is a folder within the NAS webserver titled Ingest.
• The start directory is the exact file name of the Bag. If this filename is entered incorrectly, MetaArchive cannot harvest the content.

III. Ingest Workflow

This section outlines a schematic of the stages of content transfer and preparation necessary of ingest and preservation in MetaArchive.

![Schematic Example of InDiPres ingest workflow](image_url)

1. Preparation of Content by InDiPres Members

At the member institution staff will:

• Establish preservation priorities for collections according to individual circumstances and criteria of InDiPres.
• Review Collections and identify specific collections for preservation.
• If necessary organize and reformat file name conventions for collecting being preserved. Keep a record of any of these changes.
• Identify or create metadata for selected collection, any and all metadata for a given collection should be sent along with the digital content.
• Identify or create a description of each individual collection and include this with collection data.
• Identify or create a general description of the member organization and include this with collection data.

2. Content Transfer to InDiPres

Option 1: InDiPres NAS Staging Server Transfer

Once digital material is selected and prepared it must be moved to the InDiPres NAS Server, which serves as the web accessible point of ingest into MetaArchive.

• Within the NAS Server, designated InDiPres member’s folders have been created with a number value and the member institution acronym as identifiers.
• An email invitation has been sent out to all the InDiPres members allowing them access to their specified folder.
• Following the instructions in the invitation email, members must create a password and username to access the NAS.
• Once this process is complete, members will be able to sign into individual institutional accounts in the InDiPres NAS webserver and transfer content with this device’s web interface. This is an effective drag-and-drop process. InDiPres members may find it necessary to transfer via hard drives in some select cases.
• The method is flexible and up to what is accomplishable and feasible for members. What is paramount is that digital collection material be safely acquired and in a position to be ingested for preservation.

Option 2: External Hard Drive Transfer

• InDiPres member would need to place files on appropriate hard drive.
• Contact InDiPres staff and arrange a time and method exchanging the hard drive.
• Once at ISU the hard drive has the member’s collection data placed on the staging server and ingested.
3. Content Preparation by Metadata Specialist

Content preparation ideally should be accomplished by the creator and owner of the digital preservation collections. At this stage of the project, however, InDiPres is offering some support for members in preparing their content when needed.

4. Ingest Preparation

Once the digital collection from a member is prepared according to the above steps, it needs to be prepared for ingest. The first part of this process is packaging this data into the Bagger program.

- Bagger creates packages of associated digital content that interact well with the MetaArchive preservation infrastructure. The program also runs some basic data integrity functions such as generating checksums for digital files.
- All collection files as well as metadata files are included in a Bag.
- To create a Bag also requires the creation of a bag profile. This profile includes essential information about the institution and bag. This data is written by the Bagger program into a text file and included in the final bag with the original digital collection content.
- Once the Bag is created, it has to be placed in web accessible server space. The file name of the bag and the file it is contained in need to be configured in a logical way that allows for bag ingestion to be kept track of easily. Bag names should include the institutional acronym, name of the collection, and bagging date.
- This information also needs to be properly entered into portions of the MetaArchive infrastructure for ingest to occur correctly.

5. MetaArchive Infrastructure

The MetaArchive Conspectus system requires the following items to be created and configured correctly:

- An InDiPres content contributor page as the parent folder of a member’s collections
- A collection page for each institutional member
- Entry of collection and institution-level metadata within the Conspectus pages
- Creation of a collection plug-in.
- Creation of an HTML manifest page giving data collection permission to MetaArchive. This information must be included in the top level of the directory the MetaArchive web crawler data harvester is directed to.
• Individual archival units (AUs) need to be made within each collection page. These archival units must correspond to the Bags made from member data collections. The AU parameters must be configured correctly so as to be the same as the Bag name and data harvesting folder name.

6. MetaArchive Tests and Digital Distribution

When the above listed items are completed and in place, it is necessary to send an email to the MetaArchive support staff requesting a test ingest be run.

• This request initiates a test crawl on specified content AUs within a Collection Page. This test will identify any errors of problems present in how this unit is configured. If errors are present, this process will have to be repeated once the issues are identified and corrected.
• Following a successful test ingest, a preservation ingest will be initiated. This is the real time preservation offered by MetaArchive and will eventually comprise a set of 7 copies of collection data held in distributed locations around the United States.
• The copies maintained in this set of servers will be constantly checked against one another for errors or data corruption and if problems of this sort are identified the damaged portion will be corrected.
• This process provides a secure method of preserving valuable cultural content.

7. Content Restoration

• Once content is ingested into the preservation network, it is stored without outside access and is secure. In the event that the original data present at the member institution is lost or corrupted the copies kept safe by MetaArchive can be returned to the data’s owner/creator for restoration.
• In case of a case of disastrous data loss by member, contact staff at Indiana State Library.
• It is at this point that correct and effective metadata, file naming, and description will become critical to the utility of the restored content. If the restored material is not properly contextualized, it is likely to be very difficult to understand what it is.
IV. InDiPres Digital Preservation Policy Creation Guide

[This section incorporates content for a Digital Preservation Policy Workshop. This event was led by Matt Schultz of Grand Valley State University as part of the InDiPres deliverable on preservation policy creation. This workshop took place at the Indiana State Library in Indianapolis in September of 2016.]

1. Policy Introduction

1.1 Policy Guide Purpose

This section will provide guidance to InDiPres members in writing a specialized digital preservation policy document for their collections and institutions, as well as justification of the importance of having a documented preservation policy for digital collections. It will provide a guide to the major issues and considerations necessary to be aware of when conceiving a Digital Preservation policy. The end result will be a flexible policy statement which conforms to the each institutions unique circumstances and is an aid to the preservation of their digital materials. Members should use this Guidance Document section to craft their own policy relevant to their specific needs and circumstances.

1.2 Why Digital Preservation is Important

Ensuring the integrity and accessibility of digital assets of all kinds is a critical task for cultural memory institutions in Indiana. Online digital content is quickly becoming a vital method of collection outreach and visibility for cultural memory materials. This cannot be supported if the digital files of collection materials become inoperable or corrupted. In addition there is the cost of digitization of resources having gone to waste if this data is lost through lack of necessary preservation and curation.

1.3 What is a Digital Preservation Policy?

“Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time.”

(ALA, Medium Definition)

1.4 Importance of a Policy for Preservation

Having a formal written policy which describes and directs the work of digital preservation can be a great aid in successfully carrying out this complex but critical work. As with other types of policy document this will act as a bedrock formulation of what the goals, practices, and responsibilities are for enacting preservation of digital assets. Listed below are some of the benefits that a written policy can accrue for InDiPres members.
• Formal policies are in place for a wide variety of other elements of an institution’s work
  procedures and organization.
• Formalize informal decisions, and procedures
• Add transparency and explicitness to work priorities and procedures
• Clarify what is possible to accomplish
• Serves as a foundation for sustainable strategies
• Coordinate resources and efforts
• Provides clarity to roles and responsibilities within the institution
• Strengthen connections between practices and intuitional goals
• Communicates expectations and requirements across the institution

1.5 Policy Flexibility

An InDiPres members Digital Preservation policy should be a high level statement of intent rather than a
detailed guide. The specific strategies, procedures, and methods employed for preservation can change
and reorient over time. This area of information technology is rapidly changing and is likely to continue
to do so. For this reason a DPP does not need to be a precise guide or description of exactly what is
being done to preserve digital content. Think of it more as a general overview of the policy and
approach to this issue at the respective institution. This has the virtue of both allowing for brevity of
writing for the policy and reduces the need to re-write the policy as changes and developments occur.
Also a policy can be amended and developed over time, do not be too preoccupied with getting the first
effort perfect. It is better to have a draft that can be edited than to stall at an early stage of writing.

2. Policy Elements

This section will contain a guide to the major points a Digital Preservation Policy should address to be
effective.

2.1 DP Policy Purpose

• Key Aims of Preservation
• Audience for Policy
• How is document to be used?
• What form should it be in?
• Connection between policy and implementation
• Relation to preservation strategies.

2.1.1 Purpose: Example

The InDiPres Member institution is committed to preserving the rich history of the peoples, places, and
events of the county and the surrounding region. This commitment includes artifacts donated, collected,
and created in digital form. This policy works together with our collection policies to help guide the
institution’s curators in their work and ensure that digital materials are available for ongoing education
and outreach. The InDiPres Member recognizes that digital preservation is a dynamic field and
endeavors to stay abreast of best practices and appropriate use of technologies. This policy will be updated routinely to address any changes.

2.2 DP Policy Research

- Organizational Context and Circumstances?
- Existing Policies?
- InDiPres Member Stakeholders
- Needs of designated community
- What are current best practices?
- Content/Tone examples from other policies
- What is being done at similar institutions?

2.3 Mandate

2.3.1 Mandate: Example

The InDiPres Member Institution takes seriously its responsibility to the county, its citizen taxpayers, our member donors, and any granting agencies to be good stewards of the resources the institution devotes to digital preservation. The InDiPres Member will work with the Governing Board to approve all requests to allocate resources for this important area of work and will report regularly to the public on impacts for collection building and education/outreach.

2.4. Objectives: Example

Through this digital preservation policy the InDiPres Member seeks to:

- Preserve and make accessible the rich history of the area
- Demonstrate a commitment to preservation in the digital age
- Be faithful stewards of the public trust
- Develop successful partnerships with peers, funders, and technology vendors
- Grow our expertise and leadership

2.5. Operating Principles: Example

The InDiPres Member institution will:

- Engage best practices and adopt standards for digital preservation as capacity and resources permit
- Develop sustainability plans and processes to ensure ongoing support for digital preservation
- Make use of reliable, affordable, and proven technologies
- Commit to a healthy balance of internal and external expertise and services
- Preserve only those digital materials for which there is ownership and permission
- Take seriously the role of information security
2.6 Preservation Strategies

- Bit Preservation
- Checksums
- Digital Forensics
- Emulation
- Files Formatting
- Fixity Check
- Ingest into an Archive
- Format Migration

2.6.1 Preservation Strategies: Example

The *InDiPres Member* institution will responsibly accomplish digital preservation by:

- Inventorying digital materials
- Selecting digital materials in accordance with our collection policy
- Investigating needs for legacy digital media
- Identifying file formats and migration scenarios
- Recording digital signatures
- Securing sufficient digital storage and seeking redundancy
- Routinely auditing digital materials for integrity
- Producing access copies and metadata

2.7 Challenges

- Rapid Growth and Evolution
- Sustainability
- Content Provider Partnerships
- Enabling Full Preservation
- Flexibility
- Education

2.7.1 Challenges: Example

The *InDiPres Member* institution recognizes that there will be challenges for its digital preservation program, which include:

- Addressing changes to technology over time
- Balancing financial and other supporting resources to ensure sustainability
- Allocating staff and keeping staff trained
- Managing relations with partners and vendors
2.8 Roles & Responsibilities

When writing and undertaking a digital preservation initiative the areas of responsibility and authority should be as clearly defined as possible.

2.8.1 Roles & Responsibilities: Example

The InDiPres member institution understands digital preservation to fall under its mission as a cultural heritage institution. The daily operations of the digital preservation program will be administered by the Head Curator of Collections with assistance from volunteers and interns and oversight from the acting Director. The Director and the Governing Board will recommend and approve resources in furtherance of the program. Where necessary the Historical Society & Museum will work with external partners to leverage expertise and sharing of resources.

2.9. Collaboration: Example

The InDiPres Member institution values collaboration with peers, funders, and vendors as key to success for its digital preservation program. The HS&M works closely with St. Olaf’s College to garner student interns/volunteers and to secure some technology consulting services. The HS&M works closely with the Lake Wobegon District Library to mutually contract for commercial cloud storage and access services. We are grateful to the NHPRC and the NEH for their support of our digitization and conversion initiatives. ACME Conversion Services provides offsite digitization and conversion on a contract basis.
InDiPres: Preserving Indiana’s Memories

Technical Appendix

Prepared by InDiPres Ingest Pathways and Workflows Working Group

William Knauth, Indiana State University, Chair
Sudha Anand, DePauw University
Jill Black, Indiana State Library
Carly Dearborn, Purdue University
Rachel Howard, University of Louisville
Cinda May, Indiana State University
Sam Meister, Educopia Institute/MetaArchive Cooperative
Stephen Patton, Indiana State University
Connie Rendfeld, Indiana State Library
Eric Schmidt, Butler University
Matt Schultz, Grand Valley State University
Deanna Ulvestad, Greene County Public Library, Xenia, Ohio

Indiana State Library, Indianapolis 2016-2017

This project was made possible in part by the Institute of Museum and Library Services and the Indiana State Library
Part I: Bagger Guide

1. Overview of Bagger/BagIt

1.1. What is Bagger/BagIt?

BagIt is a file packaging format intended for storing and moving digital content in an organized fashion. A BagIt-formatted package is called a “bag,” and a bag is really nothing more than a directory with a particular structure and some key metadata stored in text files that provide information about the Bag’s content (file inventory, file checksums, payload size, etc.).

There are a few different tools for creating bags, some of which we’ll guide you through using in this document.

1.2. Organizing Your Bags

The BagIt format requires no special organization for the files you are packaging. All of your original file and folder names and hierarchies are preserved, and you are always free to Bag up your collections as is. Nevertheless, even though BagIt doesn’t prescribe a hierarchy or organization onto a group of files you should give some thought to how you can Bag your collection(s) at appropriate levels so that they can be received by another party as meaningful units. As you do so, you can label your Bags accordingly with locally meaningful information such as control numbers, unique identifiers, file/folder name conventions, etc.

1.3 Creating Bags Graphically Using Bagger

If you administer your collections directly on a system with a graphical interface, the easiest way to create a bag is to use Bagger. Bagger is an easy-to-use application for creating and verifying bags. It’s easy to run on Windows computers, though it can also work on Mac OS X and Linux systems with some finessing (see Bagger README files for details on this).

This guide uses version 2.1.3 of Bagger, but the latest stable release is always recommended.

1.4 Requirements & Installation

1.4.1 Setup

To install and open Bagger:


2. In the table on that page, click on the zip file for the most recent release of Bagger.

3. After the download finishes, extract the files to a location of your choice (if you’re not sure, choose your Desktop).

4. Open the folder you extracted bagger into in the last step.
5. If you’re using Windows, double-click the icon titled bagger.bat (MS-DOS Batch File) to launch the app. Note: Be sure to launch Bagger using the bagger.bat file rather than the bagger-x.x.x.jar file. If you launch the .jar file directly, Bagger will not have enough resources to handle large sets of data.

6. Exit Bagger by closing its window.

7. Bagger is now ready to use.

1.5 Bagger requires Java.

The Bagger Application needs to access the Java Runtime Environment (i.e. Java Runtime Environment 6) on the user's machine. For Linux/Ubuntu systems use OpenJDK Runtime Environment 6 (preferably the latest release).

If Java Runtime 6 is not installed or it is not set in the System Path, then alternatively the JAVA_HOME environment variable needs to be set in the bagger.bat (i.e. Windows) or bagger.sh (Linux/Ubuntu) files provided in the bagger-2.1.3 folder as follows:

i) WINDOWS (File Path has space)

```
SET JAVA_HOME="C:\Program Files\Java\jre6\bin"
%JAVA_HOME%\java.exe -jar bagger-2.1.3.jar -Xms512m -classpath spring-beans-2.5.1.jar;bagger-2.1.3.jar
```

ii) WINDOWS (File Path with no spaces)

```
SET JAVA_HOME=C:\jre6\bin
%JAVA_HOME%\java.exe -jar bagger-2.1.3.jar -Xms512m -classpath spring-beans-2.5.1.jar;bagger-2.1.3.jar
```

iii) Linux/Ubuntu

```
JAVA_HOME = /usr/java/jre/bin
$JAVA_HOME/java.exe -jar bagger-2.1.3.jar -Xms512m -classpath spring-beans-2.5.1.jar;bagger-2.1.3.jar
```

Note: The above steps are just examples and could be avoided if the Java Runtime Environment 6 is set in the System Path, where the path or name of the Java Runtime Environment folder could be different.

2. Getting Started with Bagger/ BagIt

2.1 Deciding How to Create Your Bag

There are two different approaches you can take when creating a bag, regardless of which tools you decide to use, and it’s very important to understand the distinction.
**Bag In Place:** Creating a bag “in place” means transforming the original folder where your data resides into a bag. This causes all of the file locations in that folder to change, which is problematic if those files are being hosted by a web server or any other software that expects the data to be in the same place it was before. For instance, consider if you’re bagging a folder named “Theses” which contains several PDF files. If you turn this folder into a bag in-place, those PDF files will be moved to a new folder named “data” within “Theses”. Be sure to understand the consequences this may have before trying it on live data.

**New Bag:** The alternative to creating a bag in-place is creating a brand new bag somewhere else, with a copy of the original data. This leaves the original data unmoved and unchanged, but keep in mind that this effectively doubles the hard drive space the data will need to occupy on the system. If you’re trying to bag up many gigabytes of content, you may need to consider the practical implications of duplicating that much data.

*Example Image* - The two red circles in the screenshot below indicate the two bagging options available in bagger. These circles highlight the Create Bag In Place, and Create New Bag options.
3. Bagger Interface

3.1 Launch Bagger

Interface Features - Note the Basic layout, with a menu ribbon at the top of the window, and below this a number of tools marked by large icons. The program has two information display areas, a Left hand space showing the bag data and related files, and a larger central area where bag profile metadata is entered and displayed.

Example Image - Screenshot of the initial menu interface for Bagger. There is no current bag, profile information, or data present. This is the basic layout of Bagger that will be use to package and check the integrity of InDiPres member’s collection data.
3.2 Bag Creation

Example Image - In this screenshot is the new Bag creation menu. This will create a blank bag profile that data to be bagged can be uploaded into, and which metadata fields be attached too.
3.3 Adding Bag Profile Metadata

First, add metadata to the bag. Please refer to the section listing and describing the potential metadata fields.

i. In the middle part of the window, in the “Bag Info” section, click on the drop-down menu to the right of the word “Standard” and select the tag you want to set. (If the tag you want to set is not in the list of available tags, uncheck the checkbox to the left of the word “Standard” and you will be able to type the tag name yourself.

ii. Type the tag’s value into the text box to the right of the tag’s name.

iii. Click the Add button to the right of the text box to finish adding the tag.

iv. Repeat the last 3 steps for any additional tags you wish to add.

4. Bagger Metadata

4.1 Using Bag Metadata

Bags profiles contain a number of metadata fields that help to describe the Bag’s contents and origin. This metadata is contained in a file named bag-info.txt. Some of the BagIt metadata fields are generated by the program itself as a way of predefining the set of metadata ensuring consistent use across bags. However others must be entered manually by the user. It may not be feasible to enter in all of the below metadata fields, however it is advisable to enter as much of the information as possible as this will make the data easier to recognize and use if it is necessary to restore it from the MetaArchive.

Below is a listing of potential metadata fields with a description.

4.2 Metadata Field Descriptions

-Source-Organization

The organization where the bag was made. (No abbreviations)

-Organization-Address

The address of the Source-Organization.

-Contact-Name

The name of the person responsible for the bag.

-Contact-Phone

Phone number of the person from Contact-Name.

-Contact-Email
Email address of the person in Contact-Name.

- **External-Description**
  A description of the bag’s contents for those outside of your organization.

- **Bagging-Date**
  The date the bag was created.

- **External-Identifier**
  A sender-supplied identifier for the bag.

- **Bag-Size**
  The size of the bag. This is usually created by the bagging tool.

- **Payload-Oxum**
  A byte count and file count signature for the bag. This is usually created by the bagging tool.

- **Bag-Group-Identifier**
  A unique name given to a group of more than 1 bag.

- **Bag-Count**
  This bag’s sequence number, if part of a group of bags.

- **Internal-Sender-Identifier**
  The ID assigned to this content internally at your institution, if any.

- **Internal-Sender-Description**
  A written description of the contents of the bag based on internal standards.
Example Image - The initial step of the bag metadata creation procedure can be seen in the screenshot below. The metadata fields seen in the central right of the interface need to be completed for the bag manifest information. This can be done as part of a premade profile or manually. This is what you will be using to navigate what is in the data bags if they are resent to you by MetaArchive after being lost or damaged. It is a ‘get out what you put in’ system, so populate the metadata fields with what you and you successors will find essential and useful to restoring lost data.
Example Image- Highlighted in the red oval is the menu used to populate a Bag profile with metadata. Remove the green check from the left hand box and manually type in the metadata field name in the left hand menu. Then in the right hand menu enter the metadata for each field. For example for the title metadata field, type Organization Title in the left hand box and the organizations name in the right hand box.
Example Image - Here the metadata fields for the bag are circled in red. These are unpopulated field.

5. Data Upload and Bag Saving

5.1 Saving Instructions

Depending on which bagging approach you’ve chosen to use, follow the appropriate instructions below in order to save the bag:

a. For in-place bag creation:
   i. Click the Create Bag In Place button on the toolbar. A dialog window will appear.
   ii. Click the Browse button to the right of “Select Data”. A file browser dialog will appear.
iii. Navigate into the folder you wish to transform into a bag.
iv. Click the OK button.
v. A message will appear, confirming that the bag has been saved. Click the OK button to dismiss it.

5.2. Creating A Bag Copy

i. Click the Create New Bag button on the toolbar. A dialog window will appear.
ii. Click the OK button.
iii. If you selected a profile, complete the metadata fields presented.
iv. In the left column of the Bagger window, click the green “+” button to the right of the “Payload” heading.
v. Navigate to and select the file or folder you wish to copy into the bag, then click the Open button.
vi. Repeat the previous two steps until all desired content has been added into the bag.

vii. Click the Save Bag As button. A dialog window will appear.
viii. Click the Browse button. A file browser dialog will appear.
ix. Navigate to the location where you wish to store the new bag, give it a name, and then click the Save button.
x. Click the OK button. (Do not change any additional settings.)
xi. A message will appear, confirming that the bag has been saved. Click the OK button to dismiss it.
Example Image - Now that the profile metadata fields are created and populated, the collection data can be uploaded into the bagger interface, then checksums and file packaging can begin. Click on the green plus button in the top right of the left hand menu which is circled in red.
Example Image- After clicking the green (+) button Bagger will bring up a file navigation interface. Find the location within your desktop system of the collection data you wish to bag press open to place in bagger. This will enter this data into the bag.
Once the bag data has been selected and in place in bagger it is time to save the bag. Do this by selecting the Save Bag button highlighted by the red circle. You will need to select a location to save the new bag and assign a name to the file.
Example Image- The next step in bag saving is to select where the bag will be saved, select you wish it to be holey or not, and which algorithms you wish to use. It is suggested that the sha256 algorithm be used.
Example Image- After the bag has been saved the Bagger program will begin writing the bag. Once this process is done the bag is almost complete.
Example Image - Finally, select and click the Validate Bag, and Is Bag Complete buttons. These will check the quality of the bag data and is the final step in bag creation.
Example Image - Here is an example of a complete saved bag and the bag file generated on the user's desktop.
Part II: MetaArchive Conspectus & Ingest Infrastructure

This guide will serve as an introduction to how the Conspectus system functions within the MetaArchive cooperative and how it facilitates the preservation of member content. Screenshot examples will be used to identify, describe, and explain the Conspectus’ different elements along with information on creating and formatting pieces of infrastructure for successful content preservation.

1. MetaArchive Context

The Conspectus is a part of the MetaArchive infrastructure used to organize and describe digital collection content. The Conspectus elements interface with the MetaArchive LOCKSS system used to enact distributed digital preservation. This organization system has multiple elements.

1.1 InDiPres Organization Hierarchy

InDiPres content is organized in the following fashion.

*Content Provider Page (InDiPres) > Collection (Members Institution) > AU (Members Collection(s) in bag format)*

The content organization begins with a Content Provider Page which must be created by the MetaArchive staff upon request from a member.

The choice was made to configure the Indiana Digital Preservation (InDiPres) Content Provider page as the parent directory for content.

Within this InDiPres Content Provider page are individual Collection pages for each member of the InDiPres grant project.

Each Collection page presents some basic metadata about the member institution. Such as a description, title, and archives category, base url for harvest, and designated plugin.

This includes two very important pieces of technical infrastructure, the Base URL and plugin page.

This Collection Detail page is also where one can view, create, and edit Archival Units. These are the blocks of digital content which MetaArchive crawls, ingests, and sends out copies of to its distributed digital preservation network.

There is a menu at the bottom of the Collection details page. This is the access point for adding Collection Metadata, which uses the standard Dublin Core elements either entered manually or uploaded as an XML page. This is also where the user can manage Archival Units through the tab with this title on the far right.
1.2 - Content Providers List

This page lists the MetaArchive members and their testing and preserved collection Archival Units. It also included information on member plugin pages, institutional abbreviations and allows access to each individual Content provider’s page.
1.3- InDiPres Content Provider Page

This is the individual page for the InDiPres collaboration. Members of InDiPres will have their specific institutions and collections AU’s organized and described from this central access point. At the bottom of the screenshot image can be seen four InDiPres members entered into the Conspectus.
1.4- Collections Page

This is an example of an individual Collection page within the Conspectus. The title and other descriptive metadata about the institutional member, webserver Base URL, and plugin page information must be established and entered manually. This is a critically important element for the ingest of digital content into MetaArchive.
1.5- Collection Metadata, Manual Entry

Metadata for each institution can be entered manually or by uploading an xml document. The suggested metadata schema to use is the Dublin Core Elements. It is also suggested that members creating this metadata enter all relevant metadata fields and ensure that the metadata entered is accurate and useful to their needs.
1.6 Collection Metadata XML upload option
1.7 Plugin XML Page
The plugin is a document used in the ingest process to direct the digital content harvesting from MetaArchive to the correct webserver location. It is formatted in HTML code. This plugin page is specifically configured by the MetaArchive to work with the Bagger data security and packaging program. All that is necessary is to enter in the webserver Base URL into the appropriate portion of the HTML document.

```xml
<?xml version="1.0"?>
<map>
  <entry>
    <string>plugin_config_props</string>
    <list>
      <org.lockss.daemon.ConfigParamDescr>
        <key>start</key>
        <displayname>Start</displayname>
        <type>1</type>
        <size>20</size>
        <definition>true</definition>
        <defaultOnly>false</defaultOnly>
      </org.lockss.daemon.ConfigParamDescr>
      <org.lockss.daemon.ConfigParamDescr>
        <key>base_url</key>
        <displayname>Base URL</displayname>
        <description>Usually of the form http://<journal-name>.com/</description>
        <type>3</type>
        <size>40</size>
        <definition>true</definition>
        <defaultOnly>false</defaultOnly>
      </org.lockss.daemon.ConfigParamDescr>
      <org.lockss.daemon.ConfigParamDescr>
        <key>path</key>
        <displayname>Path</displayname>
        <type>1</type>
        <size>60</size>
        <definition>true</definition>
        <defaultOnly>false</defaultOnly>
      </org.lockss.daemon.ConfigParamDescr>
    </list>
  </entry>
  <entry>
    <string>plugin_version</string>
    <string>2</string>
  </entry>
  <entry>
    <string>au_name</string>
    <string>"All reachable from BaseUrl/%s/%s", path, start</string>
  </entry>
  <entry>
    <string>au_start_url</string>
    <string>"%s/%s/%s", base_url, path, start</string>
  </entry>
  <entry>
    <string>au_manifest</string>
    <string>1,"%s/mabagmanifest.html", base_url</string>
  </entry>
  <entry>
    <string>au_crawl_depth</string>
    <string>999</string>
  </entry>
  <entry>
    <string>au_def_new_content_crawl</string>
  </entry>
</map>
```
1.8 Plugin Details Page

On this page the details of the Content Provider Plugin is displayed. Once a plugin HTML page has been formatted it is necessary to upload this document into the conspectus using the plugin tab to the left.

1.9 InDiPres Collection Parameters for AU

This is a very critical piece of the MetaArchive ingest process. The Archival Units (AUs) are the packages of data which the crawler has harvested from the institution’s webserver location. For data stored on this site to be successfully and correctly harvested into the collection AUs by MetaArchive the parameter values on this page must be entered correctly. This information is what

1. In the screenshot below one can see in the lower half of the page a menu titled Archival Units. This is where AUs are created, edited, and managed.
2. To begin select the green 'add AU' button. This creates a blank AU with the Base URL entered into the Collection details menu.
3. Now the parameter values for the AU must be entered. This is very important to format and enter correctly for the harvest and ingest of digital content to work successfully.
4. The Path and Start are the two parameter values the user needs to manually enter.
5. Path is the designation for the file directory path that one has placed the bag with digital content for preservation. For example within a webservser file directory, InDiPres_Staging_Folder could be used for path.
6. The start parameter is the specific file within the Path file where the MetaArchive crawler will begin harvesting content.
7. If multiple Bagged collection will eventually be staged and harvested it is recommended that an organized logical file naming system be created to keep track of what digital content is where and which path/start combination to use for different collections.

![Image of Manage Archival Units for Collection](image-url)
MetaArchive Manifest Page Screenshots

Blank Manifest Page

The Manifest page is a basic HTML document that contains the Base URL webserver location that content will be harvested from. This document also contains a short sentence giving the MetaArchive crawl program permission to harvest digital content from the webserver. This is necessary for any content to be ingested and preserved.
Manifest Page with HTML entered for Butler Pelton Collection with information field populated.
Create Manifest Page

COLLECTION_NAME

Content preserved by MetaArchive Network.

LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

Collection Info:

• Conspectus Collection(s): COLLECTION_NAME
• Institution: INSTITUTION
• Contact Info: CONTACT_PERSON

COLLECTION_DESCRIPTION

Links for LOCKSS to start its crawl:

• LINK_NAME - LINK_INFO
Part III. Metadata

1. Metadata Guide for InDiPres Preservation Workflow

1.1 Importance of Metadata for Digital Preservation

Metadata is a critically important element of digital preservation within the InDiPres project and the MetaArchive Collaboration.

The inclusion of appropriate metadata is as important an element of the overall Archival Unit as the digital collection content. This Metadata is what will contextualize and explain preserved digital content that is retrieved from the MetaArchive system if and when the original data is compromised.

The effort spent on making collection metadata complete, useful, and accurate will be directly proportional to how useful this metadata will be in the future and correspondingly how ineligible retrieved digital content will be.

There are several different types of metadata included or generated in the preservation ingest of digital content with MetaArchive. Each plays an important role in describing the information elements at differing levels and locations.

Figure: Graphic from MetaArchive showing Content and Metadata relation within Archival Unit
2. Metadata Types within InDiPres Ingest

2.1 Item and Individual Collection Level Metadata.

This level of metadata describes the individual collection being preserved within a Bag/Archival Unit. If available it would be best to include as detailed a level of metadata description as possible. A useful rule of thumb for this is to include the type and detail level of metadata at this point that will be most useful to your institution and collection inheritors in restoring lost content.

It is important to keep in mind that the MetaArchive is a dark archive and does not support any public access or searching capacity. So the level of metadata that would be needed to support this function is unnecessary.

The most effective metadata records that members could include are those which are currently used to manage and access this collection material at their home institution.

However it is also true that when preserved data is restored to an InDiPres member the metadata supplied for each collection AU is the same information that will need to be intelligible in order to understand and contextualized the restored content data. In other words what is put into the system is what members will get out of it when they are attempting to restore lost collection data.

At this stage of the InDiPres grant the Metadata specialist at ISU does provide assistance in creating a basic Dublin Core elements record for member collections that lack any in-house metadata.

If there is not collection metadata the ISU specialist will put together a document containing the information for all the relevant Dublin Core elements and include this in the collection bag/AU during the final ingest.

2.2 Types of Metadata used by InDiPres and the MetaArchive Preservation Network

Collection Level Metadata- Metadata of this type is included in a given Bag/Archival Unit, and in the information profile describing the data packaged in a Bag.

Item Level Metadata- Metadata of this type is included in a given Bag/Archival Unit.

Bag Metadata- Mixture of information about collection, member’s institution, Bag data and creation details, and staff contact information. Describes content in Bag, where it came from, when it was created, and the integrity of the collection’s data.

Conspectus Metadata- Contains collection and institutional level metadata, used to identify and describe members within the MetaArchive content organization system.

2.3 Example of Basic Metadata for Member Collection

Below is a list of Dublin Core elements created for the Pelton Collection from the InDiPres Members Butler University.
Title 1 – Butler University Special Collections

Title 2 – InDiPres Collections in the MetaArchive

Subject –

South Pacific Islands 18th and early 19th centuries, herbariology and botanists, Native Americans, First Edition American authors, African American poets, American westward expansion, American natural history, and early educational materials.

Description –

The content in this Conspectus entry for Butler University is comprised of digital collections selected from the university library Special Collections for preservation in the MetaArchive LOCKSS system under the InDiPres grant project at ISU. The digital material in these collections is primarily text and image files created from the physical holding at Butler University. Collections preserved here so far include the Pelton Collection of 19th century botanical specimen illustrations.

Type – Image, Text

Source X

Relation x

Coverage x

Creator – Butler University Library Special Collections

Publisher- Butler University Special Collections

Contributor x

Rights - http://rightsstatements.org/vocab/NoC-US/1.0/

Date – 18th-20th centuries

Format – Jpeg2000

Identifier x

Language – English, Latin, French, Malabar, German, Dutch.

Audience x

Provenance x

Rights Holder – Butler University Special Collections

Instructional Method x

Accrual Method x

Accrual Periodicity – Irregular

Accrual Policy – Closed
2.4 Metadata Creation in Bagger Application

This is metadata added to each content bag in the bagger program. There are a number of metadata fields which make up a bag’s profile.

This profile is displayed with each bag in the bagger program, and is included as a simple text file along with bagged content. The fields of metadata in the profile can apply to the collection content, the institution, or information on the contact information for institutions.

Some are required to be entered and attached to each ingested bag, while other are optional or my not apply to the specific collection in question.

This information is also present in the section of the Guidance document on using Bagger.

See the example figure table below for a list of each of these profile fields.

![Example Figure: Table of Bagger Metadata fields for inclusion in Bag profile.](image)

2.5 Conspectus Level Metadata
The Conspectus system within the MetaArchive site contains metadata elements which help to organize content and facilitate preservation ingest.

Content Creator Page- This section is the top level organizational unit of InDiPres within the MetaArchive conspectus. All member pages are collections within this parent entry. There is not much metadata present in this portion of the conspectus. All that is listed is the organization name, the selected acronym, and the listing of the prefix associated with the plugin page. This is also the access and display point for each of the collections and their metadata and Archival Units.

Collection Page

The Collection Page for individual members of the InDiPres project is the primary site for metadata in this portion of the ingest process. This is found in the Collection Details area at the upper portion of this page, and on the lower portion a section titled “Related items” which has an interface for metadata field information to be entered manually and to be uploaded as an XML document.

Collection Details- The metadata fields in this section are relatively simple to complete and include the Collection’s Title, the Archive category it falls under (note for InDiPres members all collections should be categorized in the GEN archive, the correct Bag-it Plugin must be selected for each collection, a base URL must be assigned for collection crawling and harvesting, and a short collection description added.

Related Items

There are two methods here for adding metadata about the member institution, manual entry and uploading an xml formatted document. The metadata schema used is the Dublin Core elements set. The list of potential metadata fields is extensive and some may not be relevant to the institutional member collections in question. One should use their best judgement in selecting what metadata fields to entry for this section of the conspectus.

Part IV. NAS Staging Server Documentation

1. ISU Staging Server Documentation

Outline of MetaArchive Webserver Use

The InDiPres grant project aims to preserve digital cultural memory content from member institutions in the state of Indiana. This is accomplished by partnering with the MetaArchive Cooperative, an organization made up of members of various academic and cultural institutions which pool resources and expertise for the purpose of offering excellent digital preservation capability.

The method of digital preservation used by MetaArchive and InDiPres is by uploading content (ingesting) into a distributed digital preservation network. This is a set of interconnected communicating server
storage caches distributed over an extensive geographic area. Each of these caches constantly check the integrity of the others and interact to replace any lost or degraded data.

2. Description of Ingest Preservation Steps from MetaArchive

The Content Owner prepares (or stages) content for preservation by:

- Making content accessible in a firewalled web hosted directory;
- Organizing content so that document files and metadata can be harvested together by LOCKSS caches; and
- The Content Owner prepares a collection description in the MetaArchive's Conspectus tool:
  - Gives the collection a title and archive designation;
  - Enters the source URL (base_url) for the web hosted directory (see above); and
  - Provides some descriptive metadata for the collection.

*MetaArchive graphic showing interconnected Cache communication.*
3. Webserver Infrastructure

In order for any of the above to take place it is necessary for a MetaArchive member or subsidiary to create and provide access to a Base URL from a firewall protected webserver. This is a dedicated web address with the capacity to store digital content and have access to the internet.

This webserver is the Base URL which is where MetaArchive’s crawler algorithm is directed to harvest digital content. A web crawler is a program which automatically visits a website and creates copies of the digital content found on that address. Hence the name, in the it crawls through web content piece by piece and gathers copies.

With the current Bagger supported crawl rules for MetaArchive this process is far simpler and more accessible than ever.

The Base URL is a critical piece of infrastructure for harvesting and preservation. It is places in the Conspectus Collection page of members as well as in the Manifest and plugin page documents. If this Base URL is not associated with a correctly configured webserver; or if the Base URL is entered incorrectly the MetaArchive system cannot find and harvest member collection data.

The MetaArchive Conspectus page allows for a test ingest procedure to ensure that any problems with the server or Collection/Content Provider page are present these can be identified and remedied before the final Preservation Ingest.

Hardware acquisition and technical configuration accomplished by InDiPres Co-Primary Investigator Stephen Patton, the Head of the Library Systems Department at ISU.

4.1 Potential Ingest Channels

There are currently two potential channels for content ingest, the first is a webserver hosted at ISU, and the second is a portable NAS webserver purchased with InDiPres funds.

- **ISU Staging Area- Webserver**
  - 120 GB of storage space, 100 GB functional storage. Bagger installed. Ingest occurs in MetaArchive folder directory with separate path/start combinations for each member collection and Bag/AU.

- **II- InDiPres NAS Server ~ 9 TB total**
  
<table>
<thead>
<tr>
<th>Folder</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>InDiPres Folder</td>
<td>4 TB</td>
</tr>
<tr>
<td>MetaArchive Folder</td>
<td>4TB</td>
</tr>
</tbody>
</table>

This webserver is purely for storage, member FTP transfer, and MetaArchive ingest. The NAS hardware and software does not support editing or manipulating data once it is uploaded. It is possible to move collection data from the InDiPres to the MetaArchive folder divisions and vice versa. Because of this it is necessary to complete content preparation work on either a separate ISU Staging Webserver or another workstation, in particular when Bagging content.

**III - Metadata Specialist Workstation**

- 930 GB computer system, with internet access to edit and configure MetaArchive Conspectus elements such as the necessary Manifest Page and Bagger Plugin. Also has FileZila and Bagger program installed, as well as Remote Desktop Connection access to the ISU Staging Server.

- Data flow and transfer during preparation and ingest work is accomplish using the File Transfer protocol program FileZila. This allows for data to be moved between webserver sites as well as in and out of the Metadata workstation.
4.2 Ingest Data Flow for ISU Webserver

Webserver is an effective ingest method, has been shown to function correctly
4.3 Ingest Data Flow for NAS InDiPres Webserver

The NAS server has a very high storage capacity and will be adequate for the ingestion plans for InDiPres’ moving forward. This device is also portable and could be relocated to other member Institutions in the future. This ingest pathway has also been successfully tested and has been used to ingested final preservation content.
4.4 NAS Server Configuration
VI. Appendix 2: Quick Start Collection Preparation Checklist

InDiPres Ingest Preparation: Quick-Start Checklist

This is a list of the essential digital collection preparation work necessary for InDiPres members to undertake in order for digital collections to be preserved effectively in the MetaArchive. These should be guidelines to follow in appraising, organizing, and describing digital collections prior to sending it to InDiPres for preservation with the MetaArchive. Adhering to these suggestions as closely as possible will maximize the utility of collection preservation for member institutions. Greater details for following these guidelines can be found in the InDiPres Guidance Document.

___ 1. Inventory Organization: InDiPres Members review their total digital collection assets and create an organized inventory of this material. This can be accomplished with a spreadsheet containing collection titles and essential information on digital collections. Some potential pieces of collection information: Content Categories, Date of Creation, Data Extent, Metadata Description, File Formats, In House Physical and Digital Storage Arrangements. This type of physical and intellectual control of an institution’s digital holdings is key to protecting and preserving digital material on a local level.

___ 2. Collection Appraisal/Priorities: Member holdings are examined and appraised across several categories of value and utility; from these decisions sets of digital collections are assigned a priority level for digital preservation. This priority ranking is then used to decide the timetables for InDiPres/MetaArchive preservation and staff attention in terms of the preparation and description. Listed below are some major points to consider when assigning a given priority for digital preservation.

Risk Factors: Change and Loss, Obsolescence, Disaster Damage.


___ 3. File Naming: Collections have their items assigned consistent and logical file names, if necessary. The file naming system is simple, descriptive, and free from spaces and special characters. It should contain the member acronym as the first identifier. Examples: SCPL_CivilWar_032_04; KCPL_CivilCourt_0011a; ALA_Peterson_2011_02.

___ 4. Metadata: Collection level and item level metadata associated with collection is included with data package sent to InDiPres. This metadata must adequately describe the digital collection’s content. It is critical to include this with each collection. Data preserved in the MetaArchive is only as useful as its description and context supplied by metadata. This is how digital collections will be identified when restored from MetaArchive in the future. If no metadata is natively present, create in a unique document the information for the following basic information fields. Include any and all collection metadata in the same folder as the digital items when placing it on the InDiPres NAS Server.

Title- The collection resource’s name.
Format- The file types of the digital information data for the collection.
Rights- A statement on the intellectual property rights of a collection.
Date- The temporal information about the collection.
Description- A short written account of the contents of the collection resource.
5. **Collection and Member Description**: A short text description of each digital collection and of the InDiPres member institution as a whole should be created or derived from a native source within the member institution; this must be transferred to InDiPres along with the rest of the data package.

6. **Collection Data Transfer**: InDiPres members will be invited to access their own designated folder in the InDiPres NAS ReadyCLOUD web-server. This invitation will contain a hyperlink for creating an account with a username/password combination. Once this is accomplished, content can be transferred through a simple drag and drop interface. Folders are named by member institution acronym. Please include all collection files in a single named folder when placing on the NAS.