

Web Archives Workbench

Managing the identification, selection, and
acquisition of web content

Leah Houser, OCLC



Web Capture for Gov Docs

■ Timing

- Regular capture
 - Official publications, Proceedings, Law
- Quick capture
 - Events & changes in administration, one-time publications

■ Flexibility

- Decentralized systems
- Inconsistent organization of data
- Roles & priorities

Web Archives Workbench

- OCLC Digital Archive Service
- Library of Congress NDIIPP Program
- UIUC / OCLC, Tufts, Michigan State, Arizona, Connecticut, Illinois, North Carolina and Wisconsin
 - Digital preservation research issues
 - Repository testbed
 - Software development based on Arizona Model for web harvesting


The Arizona Model: Archival Practice

- Identify Collecting Scope
- Provenance, Original Order and Context
- Macro Appraisal
- Analysis – Identification of Record Series
- Content Acquisition
- Arrangement
- Package and place in repository

WAW: Suite of Tools

Discovery Tool	Identify Collecting Scope
Properties Tool	Provenance and Context Macro Appraisal
Analysis Tool	Analysis Arrangement
Harvest Tool	Content Acquisition Placed in Repository

Discover


NDIIPP group, User1
Acknowledgments
Logoff

Discovery Tool
Properties Tool
Analysis Tool
Harvest Tool
Package Tool
Alerts
System Tools


Entry Points
Domains

Please enter the domain search term:
and View By: Scope
and Obsolete
Apply

Save/Update
Number of Results

	Scope			Domain	IP Address	Entity Assigned	Obsolete?	Delete?
	New	In	Out	Show By Page			Select All	Select All
Add	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text"/>				
Details	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	a257.g.akamaitech.net	205.161.4.156		<input type="checkbox"/>	<input type="checkbox"/>
Details	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	answers.hhs.gov	63.240.89.10		<input type="checkbox"/>	<input type="checkbox"/>
Details	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	apps.das.ohio.gov	66.145.134.51		<input type="checkbox"/>	<input type="checkbox"/>
Details	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	bmv.ohio.gov	156.63.96.228		<input type="checkbox"/>	<input type="checkbox"/>
Details	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	business.ohio.gov	156.63.96.228		<input type="checkbox"/>	<input type="checkbox"/>
Details	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	collisionboard.ohio.gov	156.63.96.228		<input type="checkbox"/>	<input type="checkbox"/>
Details	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	crc.ohio.gov	156.63.96.230		<input type="checkbox"/>	<input type="checkbox"/>
Details	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	das.ohio.gov	156.63.96.228		<input type="checkbox"/>	<input type="checkbox"/>
Details	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	disputeresolution.ohio.gov	156.63.96.228		<input type="checkbox"/>	<input type="checkbox"/>
Details	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	dublincore.org	132.174.1.71		<input type="checkbox"/>	<input type="checkbox"/>
Details	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	dw.ohio.gov	156.63.96.229		<input type="checkbox"/>	<input type="checkbox"/>
Details	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	find.ohio.gov	198.234.22.81		<input type="checkbox"/>	<input type="checkbox"/>
Details	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	frwebdate.access.ann.gov	162.140.64.234		<input type="checkbox"/>	<input type="checkbox"/>

Describe

 **NDIIPP group, User1** [Acknowledgments](#) [Logoff](#)

Discovery Tool Properties Tool Analysis Tool Harvest Tool Package Tool Alerts System Tools

Entities

Please edit the entity information on the form below:

Save Delete Cancel

Identity:

[Preferred Name:](#)

Ohio Department of Natural Resources

[Content Standard:](#)

[Key Name *:](#)

Natural Resources

[Content Standard:](#)

[Aliases:](#)

No aliases have been assigned.

Add Alias

[LCNA ID:](#)

[Local ID:](#)

Notes:

[Mandate/Authority:](#)

[Start Date \(YYYY-MM-DD\):](#)

[End Date \(YYYY-MM-DD\):](#)

[History:](#)

[Predecessors:](#)

All Entities

Fish and Wildlife
Minerals, Mining and Industry

Predecessors

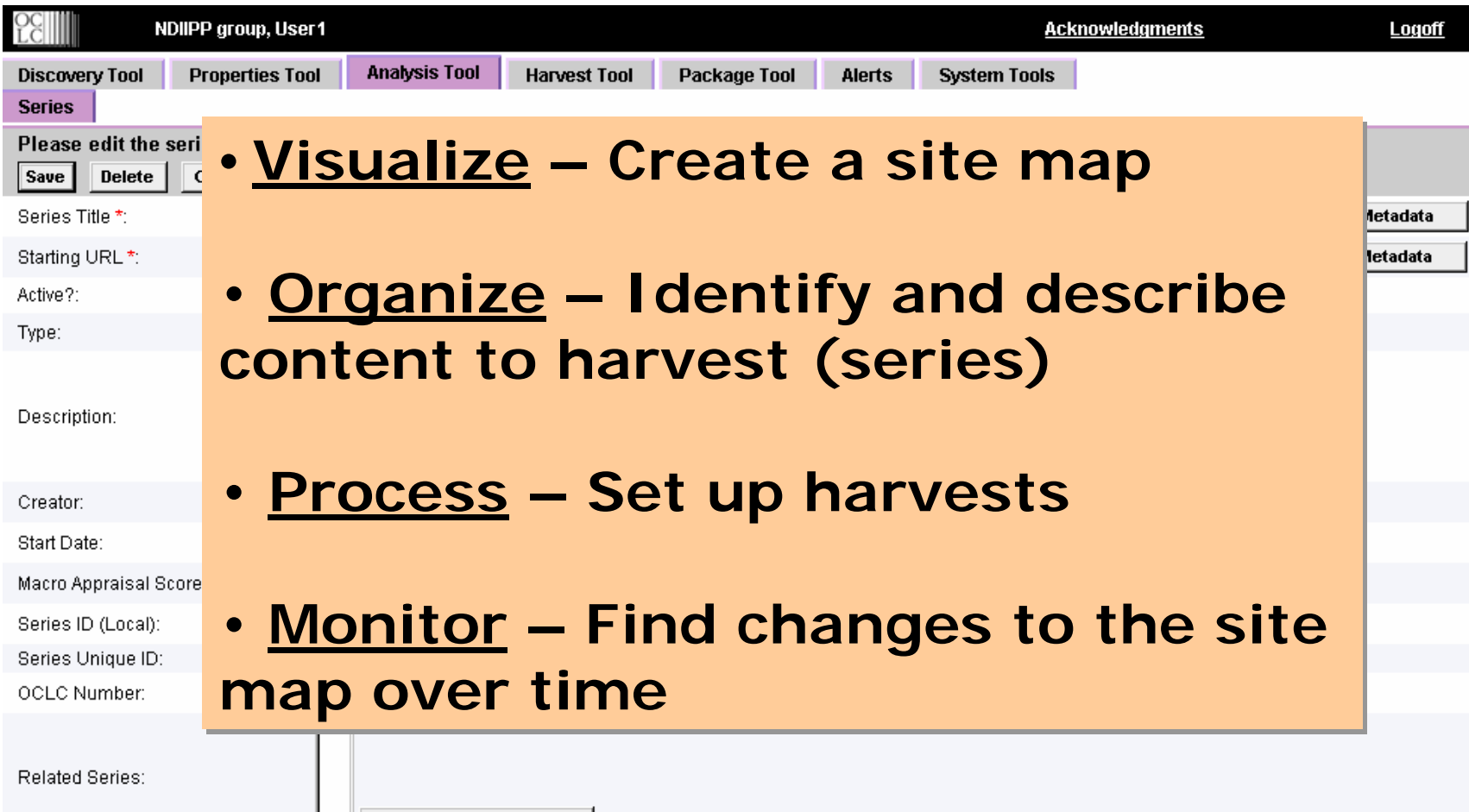
[Successors:](#)

All Entities

Fish and Wildlife

Successors

Analysis & select



The screenshot shows a web interface for the NDIPP group. The top navigation bar includes the OC LC logo, the user name 'NDIPP group, User1', and links for 'Acknowledgments' and 'Logoff'. Below this is a series of tabs: 'Discovery Tool', 'Properties Tool', 'Analysis Tool' (which is selected and highlighted in purple), 'Harvest Tool', 'Package Tool', 'Alerts', and 'System Tools'. The 'Analysis Tool' tab contains a 'Series' section with the prompt 'Please edit the series'. It includes 'Save' and 'Delete' buttons. Below this are several input fields: 'Series Title *:', 'Starting URL *:', 'Active?:', 'Type:', 'Description:', 'Creator:', 'Start Date:', 'Macro Appraisal Score:', 'Series ID (Local):', 'Series Unique ID:', 'OCLC Number:', and 'Related Series:'. On the right side of the interface, there are two 'Metadata' buttons. A large orange box is overlaid on the right side of the form, containing a bulleted list of actions.

- Visualize – Create a site map
- Organize – Identify and describe content to harvest (series)
- Process – Set up harvests
- Monitor – Find changes to the site map over time

Harvest & ingest

NDIIPP group, User1

Discovery Tool | Properties Tool | Analysis Tool | Harvest Tool | Package Tool

Packager

Please review the package information on the form below:

Edit DC Metadata | Extract Metadata | Create Package | Delete Pack

Save | Cancel

Package Status

Title:	Harvest for series Publications for Fish and Wildlife
Status:	In review
Aging Status:	
Starting Point:	http://www.dnr.state.oh.us/wildlife/Publications/pbli
Creator:	Fish and Wildlife
Created:	2006-08-29
Packaged:	
Ingested:	
Report?:	Yes
Harvest Metadata to WorldCat?:	<input checked="" type="checkbox"/>

Edit DC Metadata | Extract Metadata | Create Package | Delete Pack

Save | Cancel

Website

Starting Point: <http://www.dnr.state.oh.us/wildlife/Publications/pblast1.htm>

Include?

- ☒ <http://156.63.196.250>
- ☒ <http://img.constantcontact.com>
- ☒ <http://video.google.com>
- ☒ <http://www.dnr.state.oh.us>
- ☒ <http://www.ohiodnr.com>
- ☒ <http://www.ohiodm.com>

Ohio Division of Wildlife Publication List - Microsoft Internet Explorer...

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media

Address <http://webarchives.oclc.org/WAW/viewFile?key=11e814154d1b0f0f> Go

OHIO DEPARTMENT of NATURAL RESOURCES

OHIO
DIVISION OF WILDLIFE

Ohio Department of Natural Resources
Division of Wildlife

Home | Fishing | Hunting and Trapping | Wildlife

Online Services

BUY
Click to Purchase
LICENSES & PERMITS

YOUR CONTRIBUTIONS HELP
CLICK TO DONATE

Ohio Division of Publications List

Publication Category

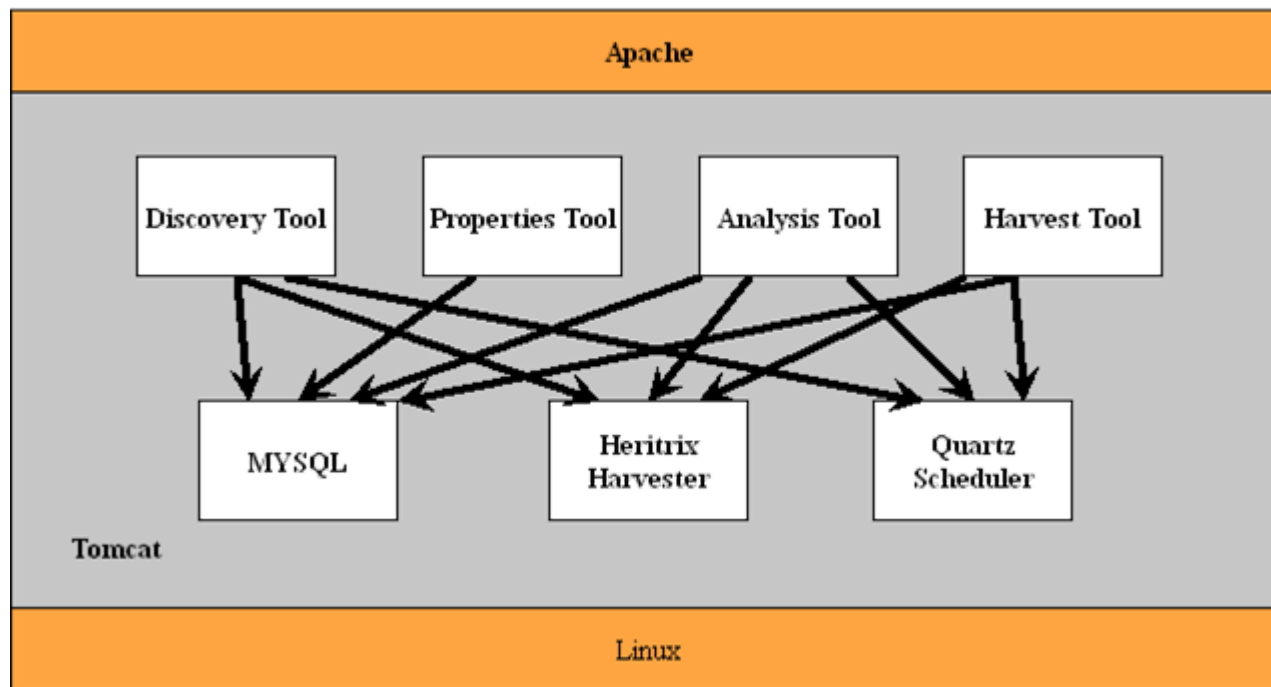
Life Histories: Birds, Fishes, Mammals	Diseases & Parasites
General Information	Birding in Ohio
	Laws/ Regulations

Hard copies are currently available at district and c

Birds
American Crow in Ohio

Fishing
Filleting

Web Archives Workbench



Web Archives Workbench

- One tool, many types of harvests to meet many needs
 - Entire sites
 - Logical pieces of sites (Series)
 - Individual documents from sites
 - Scheduled vs. one-time
 - Change detection
 - Metadata extraction

WAW features

- Robust capabilities for documenting the collection space
 - predecessors, successors, children, parents, etc.
- Inheriting metadata
 - applied to content/objects associated with particular domains
- Run locally or as a service
- Content can be automatically deposited into a reliable digital archive

WAW delivers ...

- Open Source Software
- Commercial service built on open source
- New web services
 - Metadata extraction
 - Metadata crosswalk
- Implementations of community standards
 - PREMIS
 - METS (new web & generic package profiles)

Thank you. Questions?

Web Archives Workbench

Managing the identification, selection, and acquisition
of web content

Leah Houser, OCLC
leah_houser@oclc.org

