Web Archives Workbench BETA

User Guide

Build version WAW-2.0.0.20061204-1405

Developed under the ECHO DEPository, an NDIIPP-partner project of the University of Illinois at Urbana-Champaign with OCLC and the Library of Congress
http://www.ndiipp.uiuc.edu
**Icons Legend**

The following symbols are used to highlight important information or special sections in this Guide.

**Alert.** Alerts you to additional logistical information or tips for working with the Web Archives Workbench, or to upcoming information later in the Guide.

**Definition.** Used to highlight terminology and additional ‘behind the scenes’ workings of the Workbench.

**Big picture.** Used specifically to highlight connections between different parts of the Workbench or different steps in the workflow.

**Under construction.** Information on upcoming builds of the Workbench.

**Checkpoint.** Used to recap and summarize steps at key points in the Workbench workflow.
Table of Contents

1. Background .......................................................... 5
   1.1. Introduction .................................................. 5
   1.2. The Arizona Model ........................................... 5
   1.3. The Web Archives Workbench ................................. 5

2. Getting Started ..................................................... 7
   2.1. How to Log In ................................................ 7
   2.2. Workbench Interface Overview ............................... 7
   2.3. Quick Harvest ................................................ 7
       2.3.1. Quick Harvest preview ................................. 7

3. Discovery Tool ...................................................... 9
   3.1. Introduction .................................................. 9
       3.1.1. Functions overview ..................................... 9
   3.2. Entry Points Tab .............................................. 11
       3.2.1. Add New Entry Points .................................. 12
       3.2.2. Update/Delete Entry Points ............................ 13
       3.2.3. Filter/Sort Entry Points ................................. 13
       3.2.4. Starting the spider .................................... 14
   3.3. Domains Tab .................................................. 15
       3.3.1. Reviewing new domains: assigning scope, marking obsolete, deleting .......................... 15
       3.3.2. Adding new domains manually; assigning Entities 17
       3.3.3. Filter/Sort Domains ...................................... 17
       3.3.4. Details screen ........................................... 19
       3.3.5. Search Domains ........................................... 19
   3.4. Discovery Tool Summary ..................................... 20

4. Properties Tool ..................................................... 21
   4.1. Introduction .................................................. 21
       4.1.1. Functions overview ..................................... 21
   4.2. Entities Tab .................................................. 23
       4.2.1. Update/Delete Entities .................................. 23
       4.2.2. Show Entity Hierarchy .................................. 24
       4.2.3. Add New Entities ......................................... 25
   4.3. Import/Export Entities ..................................... 34
   4.4. Properties Tool Summary ................................... 34

5. Analysis Tool ....................................................... 35
   5.1. Introduction .................................................. 35
       5.1.1. Functions summary ...................................... 36
   5.2. Series Tab .................................................... 50
       5.2.1. Add Series ................................................. 50
       5.2.2. Next Scheduled Harvest ................................. 51
1. Background

1.1. Introduction

OCLC’s Web Archives Workbench is a suite of Web archiving tools for identifying, selecting, describing and harvesting Web-based content for ingest into an external digital repository. It is based on an archival model for selecting digital materials for preservation developed by the Arizona State Library. An overview of this model and of the Workbench are provided below. The rest of this document is a quick reference guide for learning and working with the Workbench.

For more background, including information about the ECHO DEPository Project under which the Web Archives Workbench is being developed, see http://www.ndiipp.uiuc.edu.

1.2. The Arizona Model

The Arizona Model articulates a rationale and methodology for selecting digital materials for preservation, whether Web-accessible or not, as aggregates, rather than at the item level, based on archival principles, and using provenance, functional analysis, and context analysis to facilitate metatagging for retrieval.

This methodology differs from approaches utilized to date. Manual, item-level selection fails because information professionals cannot keep up with the enormous number of resources on the Web. A fully automated approach to capture all the Web results in substantive materials being buried under a mountain of ephemeral, redundant, or irrelevant information.

Instead, the Arizona methodology is based on an archival approach to the Web. In this approach, materials are managed as they are in paper-based archives: as a hierarchy of aggregates rather than as individual items. This approach reduces to a more practical size the sheer volume problem of preserving Web materials, while maintaining a scalable degree of human involvement. Developed by the Arizona State Library, it is the guiding model for OCLC’s Web Archives Workbench.

For more information see http://www.ndiipp.uiuc.edu, and browse to Core Activities > Selection rationale.

1.3. The Web Archives Workbench

The Web Archives Workbench is a suite of tools for identifying, selecting, describing and harvesting Web-based content. It bridges the gap between manual selection and automated capture of Web-based content by transforming collection policies into software-based rules and configurations. Based on the Arizona selection model, the tools will help information professionals implement Web collection policies, add metadata to harvested objects as aggregates, and package harvested objects for ingest into a digital repository. A beta version of the Workbench is currently in testing. Development of the suite continues through 2007.
The Web Archives Workbench suite:
- Discovery Tool
- Properties Tool
- Analysis Tool
- Harvest Tool

For more background information see [http://www.ndlipp.uiuc.edu](http://www.ndlipp.uiuc.edu), and browse to Core Activities > Tools development. The rest of this document provides a detailed overview of the latest Workbench beta version.
2. Getting Started

2.1. How to Log In

Open your browser and go to http://webarchives.oclc.org/WAW. Log in using one of the user names and the password provided to your institution by OCLC. If you have difficulties logging in, contact Erik Mayer by telephone at 800-848-5878 ext. 6072 or by e-mail at digitalarchive@oclc.org.

**NOTE** that though you have been provided with multiple user names, they all log into the same account. Multiple users may log in and use the Workbench at the same time. The only time a user "locks" functionality of the Workbench is by launching a Domain Spider, in which case all users will find the Domains Tab content read-only while the spider runs.

**NOTE** that the Workbench does not support all browsers. To avoid problems, please use Internet Explorer or Mozilla/Firefox. As with most web-based tools, using your browser's BACK button is not advisable and will often result in lost information. Use the navigation features built into the Workbench instead.

2.2. Workbench Interface Overview

![Web Archives Workbench Start Screen](image)

The next five sections of this Guide will walk you through a workflow for building an archive of web-harvested content based on the Arizona approach. Note that if you simply wish to schedule a one-time harvest of web content, you may choose to use the Workbench’s Quick Harvest feature.

2.3. Quick Harvest

2.3.1. Quick Harvest preview

To configure a Quick Harvest, click on the Quick Harvest tab in the second row under the Harvest Tool. You will be brought to a new screen where you
will enter some basic metadata and select spider settings. (Spiders are covered in Spider Settings (Section 8.3) and also within the Discovery, Analysis and Harvest Tool sections.)

**Figure 3: Harvest Tool > Quick Harvest tab**

![Quick Harvest tab](image)

After filling in the fields, click the *Harvest* button to launch the harvest spider. The initiated Quick Harvest will immediately appear in the list of harvests in the Harvest Tool Status tab, and your screen will be replaced by a blank Quick Harvest form, ready to create a new Quick Harvest. (If you do not wish to schedule another Quick Harvest, click the Status tab to view scheduled harvests.) Next steps for working with the harvested content are covered beginning in Section 7.1.

*Note* that Quick Harvest uses features discussed in the next three sections of this Guide, and will be more easily understood after reviewing these sections.
3. Discovery Tool

3.1. Introduction

In the Arizona Model, the first step in building an archive of materials from the Web is to identify the parts of the Web that have content you want to collect. This step takes place in the Discovery Tool.

The Discovery Tool helps to identify potentially relevant web sites by crawling relevant “seed” Entry Points to generate a list of domains that they link to. This is effectively a simple citation-analysis type of approach, working on the premise that on-topic sites might tend to point to other sites on a similar topic. The domains in the generated list are then manually evaluated as in-scope or out-of-scope, based on your subject interest and collecting policies. You can also enter in-scope domains here directly. At the end of this process you will have a list of domains that defines the sub-set of the Web relevant for your archiving purposes. From here, you can use the Properties and Analysis Tools to manage creator information about domains, and associate this information with scheduled harvests of content.

The Discovery Tool is divided into two functional areas:
- the Entry Points page, where you can enter ‘seed’ sites to be crawled by the Domain Spider to generate a list of domains that they link to;
- the Domains page, containing the list of domains generated by crawling the seed sites, and relevant domains entered manually.

3.1.1. Functions overview

Use the Discovery Tool to...

- Generate a list of domains by spidering ‘seed’ sites.
- Assign domains as in or out of scope.
- Add domains manually to domains list.
- Associate Domains with their creators (‘Entities’).

Key Workbench Terms

Domain: A server on the internet that may contain web content and is identified by a high level address. Thus http://illinois.gov/government/gov_legislature is a web site, and its domain is “illinois.gov”. Note that domains do NOT include http://.

Web site: A collection of related web content identified by a URL and stored on a domain. Note that in the Workbench, web sites must always be entered beginning http://.

Entry Point: A specific web site URL where a spider will begin to search for domains or collect web content. E.g. http://www.illinois.gov. Note that in the Workbench, Entry Points must always be entered beginning http://.

Entity: A term referring to the content creator responsible for a domain and its associated web sites (e.g., a government agency). An Entity may have multiple domains or web sites associated with it.
**Spiders:** A type of ‘bot’ or software agent that crawls the web by following hyperlinks. It is helpful to keep in mind that the Workbench uses spiders for several different purposes. The Domain Spider is used in the Discovery Tool to crawl “seed” sites to find out what other sites they point to. Other tools in the Workbench use different spiders, such as the Analysis Spider or the Harvest Spider, for purposes which are discussed later.

The diagram below graphically illustrates the stages the user goes through in order for the Discovery Tool to bring back domains, so that entities (creators of domains and their associated websites) may be recorded. “View/add” refers to what users may do when they are at the screens for the Entry Points Screen and the Domains Screen: that is, details about the entry points and domains may be viewed; more entry points and domains may be added. The outcome of the Discovery Tool process will be the creation, or recording, of entities. **Text that is highlighted in gray reflects where the user is (i.e., what screen is on display) in the archiving process (see below).**
Entry Points are URLs of known relevant web sites, which are entered here by you. They are used by the Workbench to identify additional potentially relevant domains from which you may want to collect content. You can also specify domains directly in the Domain function (3.3), with or without searching for new domains.

Sections 3.2.1 to 3.2.4 below provide additional guidance for working with the Entry Points page, following the functions identified in the diagram below.

Figure 4: Discovery Tool > Entry Points screen

3.2.1 Add new Entry Points (i.e., "seed" sites to be crawled by the Domain Spider).

3.2.2 Update Entry Points

3.2.3 Sort / filter / delete Entry Points

3.3.4 Starting the Spider

(Top of page contains a filter for managing view of Entry Points [see below].)

(Bottom of page contains table of Entry Points.)
3.2.1. Add New Entry Points

To add an entry point, click the Add button on the main Entry Points screen. The following page appears.

Figure 5: Discovery Tool > Entry Points Details screen > click Add button

Enter the following information:

- **Included**: Indicates whether this Entry Point will be included the next time the spider runs.
- **Entry Point**: Enter a fully qualified URL for the Entry Point (including `http://`).
- **Priority**: Assign a priority level for the Entry Point. A priority level is used to rate Entry Points by the estimated importance of their content, with 1 being the highest rating. When the Domain Spider is started, it will spider Entry Points in order of priority assigned.
- **Spider Settings**: Choose from available spider configurations listed in the drop-down menu. Spider Settings control how deep the Domain Spider will crawl for a particular Entry Point, how it traverses links, and other criteria.

Spider Settings in the pull-down menu are created and edited in the System Tools (see Section 8.3). If you wish to see more information about a setting, or create a new one, you can do this using the System Tools.

- **Notes**: Optional field to include any notes about this Entry Point.
- **Date Last Spidered / Problems / Robots** are not relevant for new Entry Points. For previously-entered Entry Points information will appear here to indicate when the site was last spidered, whether the Domain Spider encountered any problems, and whether Robot exclusion rules were followed.
You have now created an Entry Point, or “seed site.” From here, you can update and delete Entry Points, or choose to view a list of Entry Points in a particular order or according to selected characteristics. These features are covered next, as well as how to launch the Domain Spider which will crawl the list of Entry Points to discover the domains they link to.

3.2.2. Update/Delete Entry Points

From the main Entry Points tab users can indicate whether an Entry Point should be included in future web crawls by selecting the Yes/No radio buttons and clicking Update for each screen.

Unwanted Entry Points can be removed from the list by checking the Delete? box and clicking the Update button for each screen.

Entry Point addresses, notes and priority levels can be modified later by clicking the Details button.

3.2.3. Filter/Sort Entry Points

Filter: Filter the list of Entry Points using the Apply Filter button at the top of the main Entry Points screen. Select the criteria based on which Entry Points you wish to view and click Apply Filter.

Figure 6: Discovery Tool > Entry Points tab > view of filter options
3. Discovery Tool > 3.2. Entry Points Tab

Sort: Sort Entry Points alphabetically by Entry Point address, Priority or Date Last Spidered by clicking on the headings in the main table.

Figure 7: Discovery Tool > Entry Points tab > view of Sort options

Clicking the Start Spider button starts the Domain Spider’s crawl of all Entry Points tagged as Included (see 3.2.1), and using the Spider Settings assigned to each Entry Point. This will generate the Domains list, discussed next. Whether the Domain Spider is currently running is indicated in the top right of the Workbench screen.

Remember that available Spider Settings are edited and created in the System Tools, which are discussed in Section 9.

Note: up to 3 entry points are spidered at a time per priority level, starting with priority 1. When one priority level finishes completely, the next level is started.

You have now created one or more Entry Points, and have launched the Domain Spider. The domains identified by the Domain Spider’s crawl of the Entry Points will appear in the Discovery Tool > Domains tab. This feature is described next.
3. Discovery Tool > 3.3. Domains Tab

The second tab in the **Discovery Tool** is for the **Domains feature**.

The Domains feature contains the list of domains discovered by the Discovery Spider as it crawled the Entry Points discussed in the previous section. The main function of the Domain feature is to allow you to indicate which domains fall within the collecting scope of the digital archive.

Note that another important feature in the Domain tab is the ability to link a domain with Entity information managed in the Properties Tool (discussed in the next section). More information on Entities follows below.

Sections 3.3.1 to 3.3.3 below provide additional guidance for working with the Domains page, following the functions identified in the diagram below.

Figure 8: Discovery Tool > Domains screen

All new domains retrieved by the Domain Spider will be tagged as **New** domains. You will now review these new domains to decide which are in scope and which are out of scope.
Big picture recap: Here you are using the Discovery Tool to identify the subset of the web that may be relevant for your archiving purpose. Shortly you will see that Domains that are marked as ‘in scope’ can be associated with an Entity (i.e., creator) and that later, in the Properties and Analysis Tools, metadata associated with Entities can be inherited by content harvested from a particular web site. All of these are discussed further in upcoming sections.

Scope: The main function of the Domain feature is to define which domains that were discovered in a scan of Entry Points fall within the collecting scope of the digital archive. All new domains will be retrieved by the spider as New domains. Curators can specify whether a domain falls In or Out of the collecting scope by selecting the appropriate radio button. Domains marked In scope will be included for analysis in the Analysis Tool. Domains marked Out of scope will remain in the domain report, and will remain marked Out of scope if found again in a subsequent crawl of a different or the same Entry Point. In future versions of the Workbench, Out of scope domains will be automatically excluded from processing by the Analysis Tool.

Obsolete: Users can also flag domains as Obsolete, regardless of whether they are In or Out of scope. Obsolete domains are considered inactive, but curators may wish to continue analyzing these in-scope domains for changes. Obsolete domains will remain in the domain report list but will not be processed by the Analysis Tool unless they are considered In scope.

Delete: Unwanted domains can be removed from the list by checking the Delete flag and clicking Update for each screen. Note --- best to delete errors etc; if out of scope, mark out of scope instead so not rediscovered.

Note: Currently, domains that you delete may be rediscovered by subsequent crawls by the Domain Spider and reappear in the Domains list as ‘new.’ (This would happen, for example, if you were to run the Domain Spider again with the same settings and Entry Points.)

Note: To save changes to the Scope, Obsolete or Delete settings, you must click the Update button for each individual screen of domains that you work with.
3. Discovery Tool > 3.3. Domains Tab

3.3.2. Adding new domains manually; assigning Entities

New domains can be **added manually** to the Domains list by entering a domain address in the text field beside the *Add* button, then clicking the *Add* button. (Remember that domain addresses do not include `http://`) When you click the *Add* button you will be taken to a new screen that includes the items discussed above (Scope, Obsolete, etc), as well as a new option: *Select Entity*.

Note that the Add new domain screen is identical to the Details screen (3.3.4), which is used to edit information for existing Domains list entries.

**Select Entity** allows you to associate the domain with one or more Entities (or creator(s)). (You can assign multiple Entities by holding down the Ctrl key while making selections.) Entities are managed in the *Properties Tool*, and are discussed further in Section 4. It is useful to remember that currently the only place you can associate an Entity with a domain is in this Add/Details screen.

**Figure 9: Discovery Tool > Domains screen > click Add button**

![Figure 9: Discovery Tool > Domains screen > click Add button](image)

- **Please edit the domain information on the form below:**
  - **Domain:** Enter domain here (e.g., il.gov) if you did not already do so in the box beside the Add button. Remember the difference between domains and web sites (3.1.1).
  - **Select Entity:** Select Entity or Entities from the list, or click button, at right of box, to Add New Entity. (Entities are managed in the Properties Tool.)

3.3.3. Filter/Sort Domains

**Filter:** Domains can be filtered for viewing by clicking *Apply*. The two drop-down boxes provide a Boolean “and” argument to the filter.
Sort: Sort the domain report alphabetically by domain name, IP address or by assigned Entity by clicking on headings in the main table. Initially the column is sorted by domain name, ascending alphabetical. First click to sort (on the Domain heading) will display the column by domain name, descending alphabetical. Second click to sort will display the column by a reversed version of the domain name, ascending alphabetical. (E.g., gov.ny.www is the reversed version of the domain name www.ny.gov.) Third click to sort will display the column by reverse domain name, descending alphabetical. A fourth click to sort will return the column to domain name, ascending alphabetical. You may also sort by ascending or descending IP address and Entity Assigned by clicking those headings.

Figure 10: Discovery Tool > Domains screen > view of Filter options

Figure 11: Discovery Tool > Domains screen > view of Sort options
You may edit information for existing Domains list entries by clicking the Details button beside the entry. As already mentioned, the Details screen is identical to the Add screen. The Details/Add screen is currently the only place where you can associate an Entity (i.e., creator) with a domain.

Figure 12: Discovery Tool > Domain Details screen > click Details button beside Domain entry

You can search domains in the domains list by typing a domain search term into the search field and pressing Enter or pressing the Apply button. Search terms may be words included in a web site’s url or numbers from its IP address. A search for “.gov”, for example, will return all web sites in the list from the “.gov” domain.
You have now reviewed the Domains identified by the Discovery Spider, and decided which are in-scope for your collecting area. You may also have associated creator (Entity) information with in-scope Domains. More work with Entities is coming next.

The Discovery Tool helps to identify potentially relevant web sites by crawling relevant “seed” Entry Points to generate a list of domains that they link to. At the end of this process you will have a list of domains that defines the sub-set of the web relevant for your archiving purposes. From here, you can use the Properties and Analysis Tools to manage creator information about domains, and associate this information with harvests of content. More information about Entities is covered in the Properties Tool in the next section.
4. Properties Tool

4.1. Introduction

Another premise of the Arizona Model is that, as much as possible, metadata should only have to be entered once, and should be inherited by associated harvested objects. Metadata about content creators is managed in the Properties Tool.

The Properties Tool allows you to maintain information about content creators or 'Entities' (e.g., government agencies), associate them with the web sites they are responsible for, and enter high-level metadata about them that may be inherited by content harvested from their web sites. The Properties Tool also allows you to describe the relationships (e.g., parent/child) of Entities with one another, as well as enter other properties, such as contact information. Importantly, the Properties tool can also be easily used to create Analyses and Series from Entities' web sites. ('Analysis' and 'Series' definitions provided below.) Analyses, series, and harvests are then further managed using the Analysis and Harvest Tools.

The Properties Tool has one functional area:
- the Entities page, where you can create, edit and manage Entity (or creator) information, including associating Entities with web sites, and web sites with Analyses and Series.

4.1.1. Functions overview

Use the Properties Tool to...

- Create and manage list of content creators (called 'Entities').
- Assign metadata and other properties to Entities.
- Specify web sites that Entities are responsible for, and create Analyses and Series based on those web sites.

Key Workbench Terms

Entity: Refers to the content creator responsible for a domain and its associated web sites (e.g., a government agency). An Entity may have multiple domains or web sites associated with it.

Analysis: The purpose of enabling Analysis of a web site is to examine its structure—i.e., the directories comprising the web site. Through Analysis it is possible to discern whether there is valuable content in these directories and, if so, to identify those chunks of content. More details are provided in Section 5, which discusses the Analysis Tool in greater depth.

Series: Flexible aggregates of content that are loosely analogous to archival series, and which are used to drive the Workbench harvest operations. Series are established and managed using the Analysis Tool, then harvested and packaged in the Harvest and Packaging Tools. Series may also be established from within the Properties Tool. Note that currently, Entity metadata from the Properties Tool is made available for packaging with harvested content only if the Series is linked to an Entity web site from within the Properties Tool. More details provided in Section 5.
The diagram below describes the basics of what occurs in the Properties Tool. Once entities have been recorded in the Discovery Tool (i.e., entities have been associated with websites), users may describe them in the Properties Tool. This description process involves filling out a form on the Entity Details screen in the Properties Tool; the "View/Add" label below refers to viewing the information about the entity that was given in the Discovery Tool and adding more details by filling out the form in the Properties Tool. In addition, users may associate series with a website (or websites) and analyze the structure of the web site from which a Series is being derived. These functions are what “Add New Analysis” and “Add New Series” refer to. Analysis Details and Series Details are what the user fills in once analyses and series are added. Text that is highlighted in gray reflects where the user is (i.e., what screen is on display) in the archiving process (see below).
From the **Entities** page, you can create, edit and manage Entity (creator) information. Here we will also see how to associate Entities with web sites, and web sites with Series, and what this accomplishes (4.2.3.2).

Sections 4.2.1 to 4.2.3 below provide additional guidance for working with the Properties Tool > Entities page, following the functions identified in the diagram below. Note that the key *Add* feature is covered last (4.2.3) due to its complexity and that 4.2.1 – 4.2.2 assume an Entity has already been created.

**Figure 13: Properties Tool > Entities screen**

Users may delete entities by checking the *Delete* box and clicking the *Update* button for each screen.

The list of Entities can also be sorted according to *Key Name*, *Web Sites* or *Macro appraisal score* by clicking the headings in the main Entity list.

Users can update properties of existing Entities by clicking the *Details* button on the main entity list. The *Details* button will open the full *Entities* property window.
4.2.2. Show Entity Hierarchy

Users can view the parent/child hierarchy of assigned entities by clicking on the "Show Hierarchy" button. Clicking 'Show List' will return the view to the list of entities.
4.2.3. Add New Entities

To add new Entities, click the Add button on the main Entities screen:

Figure 16: Properties Tool > Entities screen > Add entities

After entering information click Save to save new entities. The rest of this section walks you through filling in the information on the Add new Entities screen.
4.2.3.1 Identity

Preferred Name: The non-repeating and unique name which serves as the principle name of the entity/content provider. The "main entry" for the record. (This could be the form established by AACR2.)

Content Standard
The content standard used to establish the name. E.g. AACR, Local, etc.

Key Name (required)
The key name is the form most people are likely to look for in an alphabetical list. The key name is the default name used for all lists of entities in the Workbench and is the only required field in an entity record.

Aliases
Other names by which the entity is known; may include legal names, abbreviations, successor names, and previous names.
4.2.3.2. Notes

**LCNA ID**
Link to Library of Congress Name Authority file.

**Local ID**
Locally defined identification code. Identifying codes are likely to be classification codes, similar to SuDoc (e.g., TRT = Transportation). If all an agency’s materials are classed under a Dewey code, it could be indicated here.

**Mandate/Authority**
Related statutes or organizational charters, programs, etc. granting particular authority to the entity.

**Start Date**
Date the entity was established. Dates must be entered in W3C-DTF (YYYY-MM-DD) format.

**End Date**
Date the entity was terminated or transformed. Dates must be entered in W3C-DTF (YYYY-MM-DD) format.

**History**
A textual administrative history of the current entity.

**Predecessors**
An entity(ies) that preceded the current entity. An entity record must already exist for the predecessor entity in order for it to appear in the selection list. Thus, if the predecessor entity name does not appear in the list, you must first add a new entity record for it. To select a predecessor entity, click on the entity name to highlight it. You can select more than one predecessor by using the Ctrl key.

**Successors**
An entity(ies) that succeed the current entity. An entity record must already exist for the successor entity in order for it to appear in the selection list. Thus, if the successor entity name does not appear in the list, you must first add a new entity record for it. To select a successor entity, click on the entity name to highlight it. You can select more than one successor by using the Ctrl key.
4. Properties Tool > 4.2. Entities Tab

Parents
An entity(ies) that is administratively superior to the current entity. An entity record must exist for the parent entity in order for it to appear in the list. Thus, if the parent entity name does not appear in the list, you must first add a new entity record for it.
To select a parent entity, click on the entity name to highlight it. You can select more than one parent by using the Ctrl key.

Children
An entity(ies) that is administratively subordinate to the current entity. An entity record must exist for the child entity in order for it to appear in the list. Thus, if the child entity name does not appear in the list, you must first add a new entity record for it. To select a child entity, click on the entity name to highlight it. You can select more than one child by using the Ctrl key.

Note: Parents, Children, Predecessors, and Successors to the current Entity can also be established by the Workbench based on the association in other entity's property definitions and are not explicitly stated in a description.

Public Notes
Information that is of value to the user of the content created by the current entity and is not placed in other entity record elements.

Web Sites
Web sites associated with the current Entity. Web sites should be entered as a fully qualified URL (including http://).

Add New Analysis: If you click on this button, you will be taken to a separate Analysis Details screen inside the Analysis Tool to associate the web site with a new Analysis. This screen allows you to enter Spider Settings and schedule the analysis. The Analysis Details screen is addressed more thoroughly in the Analysis Tool overview in Section 5.2.

Note: Analysis of a web site may also be done in the Analysis Tool. The advantage of activating Analysis while in Properties > Entities is the increase in type options, such as change monitoring.

Add New Series: If you click on this button, you will be taken to a separate Series Details screen inside the Analysis Tool to associate the web site with a new Series. This is a detailed screen, including Harvest Spider and scheduling settings, and a link to edit an associated Dublin Core record for the Series. The DC record will automatically be populated with metadata from this Entity Details screen when you create Series by clicking the Series button on the Entity Details screen. The Series Details screen is covered in detail in the Analysis Tool overview in Section 5.3.

Note: Setting up a series in Properties > Entities allows the user to create series based on a whole website. This approach is different in the Analysis Tool, where series may also be set up; that is, an Analysis series enables the user to create series based on pieces of a website.
The following screen shows the outcome of clicking on the buttons “Add New Analysis” and “Add New Series” and filling in the Analysis Details and Series Details screens, respectively, for these buttons.

Figure 19 Properties Tool > Entities Details screen > Enter web site > Analysis and Series button appear

Please note in particular the information below about different options for adding Series, and their implications for ease in associating Entity metadata with harvested content.

**Big picture: Understanding the link between Entities, Web Sites, Analyses, Series, and Harvests.** Remember that harvests of content in the Workbench are based on Series (except for one-time Quick Harvests), and that a Series is made up of a web site or part of a web site. By associating a web site with an Entity, then a Series with that web site, you provide a path by which metadata already associated with the Entity can also be associated with content harvested from that Entity’s web site. In addition, setting up an Analysis via Properties > Entities allows you to analyze the structure of the web site from which a Series is being derived. This approach also makes available more type options, such as change monitoring.

**The difference between accessing the Series Details screen from the Properties Tool VS the Analysis Tool.** You may associate Series with websites by going directly to the Series tab in the Analysis Tool, as will be seen in detail in Section 5. You may also, as we have just seen, associate Series with web sites from the Entity Details screen in the Properties Tool by clicking the Series button, which takes you to
the Series Details screen in the Analysis Tool. Existing Entity metadata associated with that web site is copied into the Series DC record if you arrive at the Series Details > DC record screen from the Entity Details screen in the Properties tool. If you go directly to the Analysis Tool > Series tab to create a Series, any existing metadata for an Entity associated with the web site in that Series will be automatically copied from the Properties Tool.

The specific Entity metadata copied to the Series DC record is: Entity Preferred Name [mapped to Creator field], Subject Headings [mapped to Subject], History [mapped to Description], Functions [mapped to Description], LCNA ID [mapped to UniqueIdentifier.LocalID], Website [mapped to Source.URI].

Currently, there are two ways of associating a Series with an Entity’s website: you may use the Properties tool to do this, whereby select Entity metadata can be copied automatically to a Series and thus to harvested objects, or—as will be explained in detail in Section 5—you may use the Analysis Tool. Once a Series has been associated with an Entity’s website, any changes to that Entity will not be reflected in the associated Series.

Figure 20: Creating Series from within the Properties Tool VS Analysis Tool

Domains
Domains associated with the current entity. Currently, these must be associated using the Discovery Tool > Domains tab > Details screen.

Associated Personal Names
Names of persons (and their corresponding dates) related to the current entity. Used as a potential access point.

Subject Headings
Subjects or topics that can be assigned at the entity level and will be inherited by the digital objects that are related to this entity. (See important notes on Entity metadata inheritance under ‘Web Sites’ in section 4.2.3.2.)
The Workbench is pre-installed with the GILS Jessica Tree. Users have the option of entering a "free-form subject heading" or selecting previously imported subject headings; the workflow for these options follows below. Alternatively, you can also use your own local terms by using the Import functionality (see 8.4.5 Import Subject Headings).

1) Click Add Subject Heading:

![Add Subject Heading](http://webarchives.oclc.org)

2) Next, a window will pop-up asking whether the user would like to enter a free-form subject heading, or to use previously imported subject headings:

If you opt to enter your own subject heading, type it in the box and click "OK":

![Add Subject Heading](http://webarchives.oclc.org)
Your subject heading then will appear as one of the Access Points:

If you select Cancel, then you may choose the appropriate subject heading from the drop-down list that appears beside the words “Jessica_Tree”:

3) Click Apply to save your selection.

Functions
A textual description of the functions of the current entity.

Contacts
Contact information for individuals related to the current entity. There is no limit to the number of contacts that can be related to an entity. To add a new contact:

4.2.3.4. Contacts
Click **Add Contact**. Enter the contact’s full name.

![Figure 22: Properties Tool > Entities tab > click Add button > click Add Contact](image)

After selecting **OK**, complete additional contact information displayed in the main form.

**4.2.3.5. Control**

**Macro-appraisal Score**

An alpha-numerical score for ranking the importance of the information created by this entity. The system will support numeric, alphabetical, and alphanumeric scoring systems that must be created and defined locally, then implemented within the Workbench.

**Type/Category**

The type of the current entity. (Use of this field will vary based on the context in which the tool is being used.) It could be used to indicate branch of government, level of government (state, local), type of organization (commercial, non-profit) or other appropriate category.

**Cataloger Notes**

Notes to be used by staff creating entity records.

You have now used the Properties Tool to create a record for an Entity (creator), including possibly mapping its relationships with other Entities, and adding metadata that can be inherited by content later harvested from web sites associated with the Entity. In the coming sections, we will look at setting up these harvests.
See System Tools (Section 9) for information about Importing/Exporting Entities.

The Properties Tool allows you to maintain information about content creators or ‘Entities’ (e.g., government agencies), and associate them with the domains and web sites they are responsible for. The Properties Tool also allows you to describe the relationships (e.g., parent/child) of Entities with one another, as well enter high-level metadata about them that may be inherited by content harvested from their web sites. Importantly, the Properties tool can also be used to create and associate Series with Entities’ web sites; Analyses may be started in the Properties Tool as well. Series and harvests are then further managed using the Analysis and Harvest Tools, which are discussed next.
5. Analysis Tool

5.1. Introduction

The Web Archiving Workbench incorporates an Analysis Tool to enable the analysis of websites. The analysis of a website allows you to look at the structure of the website and see what kind of content is represented by the file directory. This way chunks of content may be identified for harvesting.

A key premise of the Arizona Model is to balance human input and computer automation by managing archived web content in aggregates, rather than at an individual document level. These aggregates of content are analogous to archival series. In this tool, a ‘Series’ is loosely defined as any collection of web material that a user chooses to collect in one ‘bucket.’ The Series may be a whole web site or portion (e.g., only PDFs of Annual Reports) of a web site, or even one individual page or document from websites. Series can be established and managed using the Analysis Tool, then harvested and/or packaged in the Harvest Tool.

The Analysis Tool allows users to define Analyses of websites, define Series, associate metadata with these Series, hold Series objects, and schedule recurring harvests of web content. (For quick, unscheduled harvests, see the Quick Harvest option in the Harvest Tool menu.) The Analysis Tool includes visualization tools to aid in analyzing the structure and content of web sites, mainly in order to help decision-making on what content to harvest and how to divide it into series.

Note: Setting up a Series via the Analysis Tool allows users to specify directories of websites from where content should be collected on a scheduled basis. These “analysis-based” Series are able to collect only the new content added to the series as future analyses take place. If a Series is created directly from the Entity Details or from scratch, only a single website or document is allowed to be collected on a scheduled basis.

The Analysis Tool two functional areas:

• the Analysis page, which provides visualization tools to aid in content selection decision-making and in Series structure decisions. Here, too, a baseline analysis can be created against which to measure future website analyses.

• the Series page, where you can create, edit and manage Series, hold Series objects, and schedule harvests of Series.

Entity metadata and Series recap: There are two ways to associate a series with an Entity’s web site, whereby select Entity metadata can be copied automatically to a Series’ DC metadata record, and thus to harvested objects: 1) by using the Properties tool; and 2) by using the Analysis Tool (where Associated Entities may be entered). To create Series in the Properties Tool go to Entities > Add/Detail screen > Add Web Site > Add Series. Then see Series Detail screen discussion below in Section 5.3.5.)
5.1. Functions summary

Use the Analysis Tool to...

- Analyze the structure of a web site.
- Enter Associated Entities.
- Set a baseline analysis against which to compare future analyses.
- Adjust settings, such as for the spider and for change notification threshold.
- Define a “Series” for harvesting, with option to associate it with an Entity.
- Hold Series objects.
- Schedule harvests of Series.
- Note both the above can also be accessed via the Properties Tool (see note in 4.2.3.2)

Below is a diagram outlining the workflow within the Analysis Tool. The Analyses Screen and Series Screen are where analyses names and associated entities may be entered. Upon clicking on the Details button that appears in the Analyses List, you will be taken to a screen showing a form to fill in for scheduling the analysis; the same thing occurs upon clicking on the Details button that appears in the Series Screen. After scheduling analyses and allowing them to run (which also involves setting a baseline analysis for future analyses comparisons), it will be possible to make series and harvest content as that new series. **Text that is highlighted in gray reflects where the user is (i.e., what screen is on display) in the archiving process (see below).**
From the Analysis page, you can enter an Analysis Name and an Associated Entity.

**Note:** The “Analysis-based Series Details Screen” may be accessed while in the Analysis tab of the Analysis Tool, or it may be accessed while in the Series tab of the Analysis Tool. This is what the above dotted line, connecting “Analysis-based Series Details Screen” and “Analysis Details Summary Screen (Series View),” is meant to delineate. See Section 5.2.8. Analysis-Based Series Details and Section 5.3.5. Series Details.
5. Analysis Tool > 5.2. Analysis Tab

5.2.1. Add Analysis

After filling in the Analysis Name and Associated Entity [optional], click the Add button. A new screen (called the Analysis Details screen) will appear that will enable you to provide information so that you can run an analysis on a website or websites.

The Add Analysis button will take you to an Analysis Details screen, described below.

In the Analysis Details screen you will see the Analysis Name and Associated Entity [optional] you entered. Here is where you may set up and schedule your Analysis. See Figure 23 below. (Note: this screen is similar to the Series details screen. See section 5.3.5).

5.2.2. Last Run Date

Refers to the last time the analysis was run.

5.2.3. Analysis Details

To view and edit Analysis Details, click the Details button beside an existing Analysis entry. This screen is where the Add Analysis button also takes you.
The next section overviews the Analysis Details screen by taking you through some of the key steps for setting up a website analysis.

**Figure 24: Analysis tool > Analysis Details screen**

- **Analysis Name** and **Associated Entity** should be the same as what you entered on the previous Analysis page.

The Analysis Details screen is also where you enter the website that you wish to analyze, as well as where spider settings may be selected. If you have arrived at the analysis details screen by clicking on Add New Analysis in the details of an Entity record from the Properties Tool, the website(s) will already be entered.

The **Active** button should be checked if you want the Analysis tool to run an analysis. If left unchecked, the analysis will be dormant (i.e., though ready to take place, it will not run).

To **Schedule the Analysis** choose a Start Date and Time, a Recurrence Pattern (once, daily, weekly, monthly, etc.), and a Recurrence Range (select an end date).

The **Change Notification Threshold Setting** allows you to assign a change percentage:
Thus as future analyses run, if the website has changed 33% or greater (as has been input above in the screenshot), then the change will be noted, or logged, in the Alerts Tool. This means that you will need to check the Alerts Tool to gauge any possible alterations in a website’s structure.

Once the settings on the Analysis Details screen have been made, click Save. At this point, if you checked the Active box, an analysis run will take place and recur in accordance to the schedule you have set.

After you click Save, you will be returned to the Analysis Tools main page and see the Analysis Name, Associated Entity, and website filled in. If you arrived at this point via Properties Tool > Entity Details Screen (by clicking on the Add New Analysis button, then you will be taken back to the Entity Details Screen you came from.

To recap: The Change Notification Threshold Setting allows you to track changes in the structure of a website. To activate this setting, enter a number for a change percentage; that is, if you enter 10, then each time a website’s structure changes 10% or greater, it will be noted in the Alerts Tool. You must check the Alerts Tool to discern whether changes have occurred or not.

On the Analysis Tools > Analysis Details Summary screen, click on the Details button for the Analysis you have named and has completed its analysis of specified website(s). Next you will be taken to a screen that summarizes your settings, where you may also set a Baseline Analysis for future comparison. By designating an analysis as a Baseline Analysis, you are confirming, or approving, this analysis was good enough to be a standard against which to measure future analyses. You are essentially accepting what the site looked like according to the directory structure that was brought back (see bottom part of next screenshot, Figure 25).

Note: To add or define a series from the website structure (folders) that is returned, there must be a baseline. The tool will not allow you to start piecing together a series if there is no indication that the website is a good baseline. Furthermore, analyses are compared only with the baseline. If another analysis run suggests a better baseline, then you may replace the original baseline with it. (The previous baselines will be aged-off (deleted) by the system.)
At the bottom of the Analysis Tool > Analysis > Details screen you will see a graphic representing the directory structure of a website, with folders and web icons cascading downward. Once you have set a Baseline Analysis you can begin piecing up the site and defining a series from the folders, which signify web content, that have been brought back.

You may see some bright green question marks. A question mark signals an unknown series status. As you set up a series from these folders, you may click on a question mark, which will then be replaced by a dark green check mark; this means that you are assigning the folder beside the check mark to a series. When you do this, you will then see red “x” marks appearing elsewhere; an “x” mark beside a folder signals that the folder will not be included in the series. Re-clicking a dark green check mark will change it to a red “x”; clicking a red “x” will change it to a dark green check mark.
As you make a series, you may put multiple folders in the same series (i.e., a single series), as long as the series is associated with the same analysis (which leads back to the associated entity). Not only can you assign multiple folders to a series, you may also assign multiple domains to a series.

Figure 26: Analysis Tool > Analysis Details screen > Make Series button > Enter new series title

Defining Series recap: In selecting folders, you have created, or made, a series. Recall that in the Properties Tool, series may also be created, via the Entities tab, but that the series is created from an entire website. Creating a series in the Analysis Tool allows you to create series from pieces of a website.

5.2.8. Analysis-based series details

After you have made a series and named it, you may edit the source analysis (the analysis where the series originated), work with metadata about the series, activate auto ingests and harvests, and select what kind of objects (file formats) to harvest.
The next screenshot shows in full what the Analysis-based Series details screen looks like. This is where you will refine your series further. (See Sections 5.2.8. through 5.2.13.).

Figure 27: Analysis Tool > Series tab > Analysis-based Series Details screen
Note: the Series tab also has an Analysis-based Series Details screen (see Series tab details, Sections 5.3 through 5.3.5.).

5.2.9. Associated Entity

If there are additional entities you would like to associate with this analysis, they may be input here.

5.2.10. Edit Source Analysis

This button allows you to edit the analysis that the series came from.

5.2.11. Starting URL

Enter the URL from which content for your series will be harvested. More than one URL may exist.

5.2.12. Edit DC Metadata

This button opens a form for editing Dublin Core metadata for your series.

5.2.13. Extract Metadata

Besides editing metadata, you may want to extract it as well. When clicking this button you have the option of having the extracted metadata Dewey Decimal subject headings.

5.2.14. Active box

By default series are inactive. You must check the “Active?” box in order to activate a series.

5.2.15. Related Series

If there are any series related to the one you have just made, then you may input them here. If there are any existing Series already in the system, then they will appear in the box here and may be selected as Related Series.

5.2.16. Auto Ingest options

Auto Ingest options: In addition to auto-ingesting to the OCLC Digital Archive, you may also auto-ingest to the Hub and Spoke, if you have content that will be going to more than one repository.

5.2.17. Harvest Content as Individual Objects

Harvest Content as Individual Objects: This option means that the tool will go into one of the directories from the series you have created and make a harvest of each item. You can select what kind of item this will be, depending on its format: PDF, DOC, XLS, or all. So, if you have three items, this means you will have three harvests and thus three distinct digital objects.

If you have content in HTML, then the dependencies will be harvested. (“Dependencies” are anything needed to render a webpage back to the user, such as images, stylesheets, etc.) Single HTML pages are harvested with a spider setting of By Links, Depth of 0.

Once individual objects are harvested, metadata in Dublin Core is created for them. Thus, there will be Series metadata (from websites) and harvest metadata for individual objects.
The point of distinguishing a separate harvest has to do with access. If the box for Harvest Content as Individual Objects is left unchecked, then your content objects will be ingested and harvested together (imagine a clump of content), and metadata will be assigned to the objects as a group, rather than individually (which is what is typically desirable). This process will be repeated for the next scheduled harvest. For more details about harvesting content, see the Section 6.

This option enables the creation of granular metadata for each individual object. It means that information about an object will be ascertained and put into the form of metadata.

The Series View button (shown in the bottom half of the screenshot in Fig. 27) gives you access to a Tree View of your series. This means that you have a Baseline analysis set and have named (made) a series.

The next screenshot is what you see after filling in your Series details.

Figure 28: Analysis Tool > Analysis Details Summary screen
As seen above, the name of your series appears in blue to the right of the directory structure, and an Edit Series button appears on the left, allowing you to remove or include folder as needed for your series.

Figure 29: Analysis Tool > Series View > Analysis Details Summary screen

Harvest Content From New Series: this button is now included as part of the Series buttons, to enable harvesting materials from the series you have made. The Harvest Content button appears only after a new series has been created (which, in the case of the screen above, is Sample Series).

Workflow recap: In the Analysis Tool, you have now set the parameters for analysis (named the analysis, and input an associated entity, the URL(s) for
the website(s), spider settings, and a change notification threshold setting) and scheduled the analysis. From this you defined a series (you may define as many as you wish) and set a series harvest schedule. The next step will be setting an analysis schedule. This analysis will tell you about the harvested new content resulting from the series harvest schedule.

5.2.20. Pre-Comparison Review

The next screenshot shows the form for setting the analysis schedule, also known as a pre-comparison review.

5.2.21. Comparison of Analyses

Next, to view a comparison of the analyses you have created (baseline and saved analyses), check the boxes for them and click on the View button (see above screenshot, bottom part of Figure 29). The screenshot below displays the screen that is returned after clicking on the View button.
Figure 31: Analysis Tool > Analysis Details Summary screen (comparison of analyses)

The screenshot above shows how you may go about comparing the analyses you have set up. As before, during series set-up and baseline analysis determination, you will see bright green question marks denoting unknown series status. (The heading “Assign?” means: do you want to click on a question marks and assign it to a series.)

In addition, if there is missing content, this information will be seen in red. Any new content will be denoted by a dark green color, as the next screenshot shows:
Note: If you choose, you may create a new series from this comparison of analyses. Just click the Make Series button once you have assigned folders (shown in the directory structure) for a new series.

5.2.22. Node Type

At the top of the graphic for the directory structure, you will see an option to select a **Node Type**. "Node Type" refers to the filter that allows a user to view All, New, Missing, or (just) Series content.
If you wish, when in the Analysis Tool, you may bypass an Analysis-based Series and go straight to the Series tab of the Analysis Tool. Doing so will mean that you will not be able to analyze your websites (i.e., compare them in order to help you decide what kind of content you want to harvest from a website.)

From the **Series** page, you can create, edit and manage Series, hold Series objects, and schedule harvests of Series.

Sections 5.3.1 to 5.3.5 below provide additional guidance for working with the Analysis Tool > Series tab, following the functions identified in the diagram below.

**Figure 32: Analysis Tool > Series tab > Series screen**

After filling in the Series Title and Starting URL fields, click the **Add** button. A new form (called the **Series Details** screen) will appear that will allow you to enter information describing your series.
5.3.2. Next Scheduled Harvest

5.3.3. Update/Delete Series

5.3.4. Filter/Sort Series

5.3.5. Series Details

The Add Series button will take you to the Series Details screen described below in 5.3.5. Please go to this section for a detailed tour of the screen (Figure 23).

Shows when the next Series Harvest will take place.

To remove unwanted Series, check the Delete box next to the series entry and click the Update button.

To filter Series, enter a keyword in the text box and click the Apply Filter button. To remove a filter, delete text from the box and reapply the filter.

Series can also be filtered according to the Active status.
Series can be sorted by Series Title, Starting URL, Active status, Macro Appraisal Score, or Creator (entity) by clicking the column titles above the Series list.

To view and edit Series Details, click the Details button beside an existing Series entry. This screen is where the Add Series button also takes you. The next section overviews the Series Details screen, broken into several screenshots.
Series Title and Starting URL should be the same as they appeared in the previous screen. (If arrived at this screen by clicking the Add Series button in the Entity Details screen, the Series Title will be blank, ready for you to fill in.)

**Edit DC Metadata button**

Opens a form to create a Series Dublin Core Metadata record.

The Series DC metadata will be included in the final package for ingest into a digital repository and can be automatically included in OCLC WorldCat. (Note that this feature is not activated in the Workbench. OCLC will send notification when it is turned on.) The DC record will contain automatically
imported Entity metadata in certain circumstances. (See Entity metadata and Series recap in Section 5.1 above.)

The next screen shows where the user is taken after clicking on the Edit DC Metadata button. The layout is similar to the screen for assigning subject headings to Entities.

Figure 35: Analysis Tool > Series tab > Series-level Dublin Core Metadata Editor screen (after clicking Edit DC metadata button)

As in the Entities tab, subject heading assignment on this screen draws on the GILS Jessica Tree.

Extract Metadata button

In addition to editing metadata in the Analysis Tool, you may also extract metadata.
If you select Extract Metadata, a window will pop up asking the user to choose between Dewey numbers or just extracted metadata.

Clicking “OK” brings back Dewey numbers, as seen in the screen shot below:

Figure 37: Analysis Tool > Series tab > Series Details screen (after clicking Extract DC Metadata button)

Extracted metadata showing Dewey decimal numbers as subject headings.
Clicking “Cancel” takes you to this screen, which simply shows the extracted metadata:

Figure 38: Analysis Tool > Series tab > Series-level Metadata Editor screen (after clicking Extract DC metadata and then clicking Cancel [for no Dewey number form]
Series DC Metadata Reminder: Any extracted metadata will have “Extracted: “ in front of it.

Edit & Extract Metadata Summary: The Analysis Tool enables editing of the Series DC Metadata record. Alternatively, users may have metadata extracted (based on title and other content information). The extracted metadata also include Dewey Decimal subject headings.

Active
An inactive Series will not be harvested until the Active check box is checked. Active series will be harvested according to the Series Schedule. To discontinue harvesting of a series, uncheck the Active check box.

Type
At present only the Snapshot option (i.e., a capture of the web site at this particular point in time) is available. In a later version of the Workbench, Change Monitoring will also be available.

Description
Enter a narrative description of this series.

Figure 39: Analysis Tool > Series tab > click Details button (cont’d)

Creator
Either type in a name for the creator responsible for the series in the top field, or choose an Entity in the lower field. You can either choose a pre-defined Entity from the drop-down menu or add a new Entity by clicking the Add New Entity button.

Entity metadata inheritance reminder: Currently, existing Entity metadata is not automatically copied from the Properties Tool when a Series is created or associated with an existing Entity using this page. Instead it must be re-keyed in the Series DC record. (See 4.2.3.2 > Web Sites for details on Entity metadata inheritance.)

Start Date
Click on the calendar to choose a date with which to associate the creation of the series. Dates will be represented in the W3C-DTF (YYYY-MM-DD) format.

Macro Appraisal Score
If you wish, assign a value according to your own local macro appraisal score system to distinguish the relative importance of this series.
Series ID
Assign a local identifier to the series.

Figure 40: Analysis Tool > Series tab > click Details button (cont’d)

<table>
<thead>
<tr>
<th>Series Unique ID:</th>
<th>OCLC Number:</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Related Series:</th>
<th>Add Related Series</th>
</tr>
</thead>
</table>

Series Unique ID
Currently disabled.

OCLC Number
If this series has an OCLC Number assigned, enter it here. This will associate the harvested content with a record in WorldCat. (If you have questions about OCLC Numbers and associating content in WorldCat contact OCLC staff.)

Related Series
Series may be connected to other Series through the Related Series button. When prompted, enter the title of a series that you wish to establish a relationship with.

Obsolete
Like an obsolete domain, an obsolete series is considered inactive, but still of enough interest to warrant keeping an eye on.

Auto Ingest to OCLC Digital Archive
If Auto Ingest to OCLC Digital Archive is checked, the Series will be ingested immediately upon completion of the harvest. If you wish to review harvested content and metadata for a Series in the Harvest Tool, leave this box unchecked.

Auto Ingest to Hub and Spoke
If Auto Ingest to Hub and Spoke is checked, the Series will be ingested immediately upon completion of the harvest. As with Auto Ingest to OCLC Digital Archive, if you wish to review harvested content and metadata for a Series in the Harvest Tool, leave this box unchecked.

Harvest Metadata to WorldCat
In a future version of the WAW, if Harvest Metadata to WorldCat is checked, Series Dublin Core records will be collected on a periodic basis via OAI.
Harvesting and imported into OCLC WorldCat. This feature is currently not functional, but users will be notified when it is activated.

**Figure 41: Analysis Tool > Series tab > click Details button (cont’d)**

Spider Settings
Select which Harvest Spider settings will be used by this Series. Harvest Spider Settings can be configured in System Tools-Spider Settings (See Section 8.3).

**Schedule the Series**
The Series Schedule will determine when the Workbench will send a Harvest Spider to collect content. Within the Series Schedule users can specify:

- **Activity Start**: Select the date and time to begin harvesting content.
  *Note that this time always set to the Eastern time zone. In future releases of the Workbench, you will be able to specify a time zone.*

- **Recurrence Pattern**: Determines how frequently the Harvest Spider will collect content.

- **Recurrence Range**: Users may specify that the Series will continue to harvest content on a regular basis or establish a date when harvesting should end.

- **Harvest History**: Holds all Series scheduling information: the dates on which a harvest was scheduled, when it was started, and when it was completed.
You have now used the Analysis Tool to create or edit a Series, including associating it with a specific beginning location on the web (and possibly Entity) and scheduling harvests of the Series. In the coming sections, we will see how to review and package the results of these harvests.

The Analysis Tool allows users to define archival Series, associate metadata with these series, hold Series objects, and schedule recurring harvests of web content. It also provides visualization tools to aid in analyzing the structure and content of websites; this kind of analysis aids in making decisions on what to harvest and how to create further series based on an analysis. The Harvest History provides scheduling information for Series harvests. (For quick, unscheduled harvests, see the Quick Harvest option in the Harvest Tool menu.) Harvesting and packaging activities are then monitored and managed in the Harvest Tool, discussed next.
6. Harvest Tool

6.1. Introduction

The Harvest Tool lists all harvests within the Workbench, including Series harvests scheduled using the Analysis Tool as well as Quick Harvests. It is used to monitor their status, and to provide an opportunity to review and modify the harvest before packaging it up and ingesting it into a repository. Harvest reviews may be of single-object harvests or of multiple-object harvests, depending on whether the option to harvest content as individual objects was selected in the Series details screen of an Analysis-based Series. The Harvest Tool also offers a separate Quick Harvest feature, which schedules one-time harvests of content based on a URL inputted directly into the Harvest Tool. After harvests are complete and reviewed (optional), at which time extra metadata may be assigned, the harvested content may be ingested.

The Harvest Tool is divided into two functional areas:
- the Status tab, where you can review the status of harvests, review completed harvests, see which entities are associated with them, assign extra metadata to the harvested content, and ingest harvested content;
- the Quick Harvest tab, where you can schedule a one-time non-series-based harvest.

When reviewing the status of your harvest in the Status tab, you may set filters for viewing your harvests. These filters are accessible via four pull-down menus at the top of the screen (to the right of “View by,” see Figure 40).

Key Workbench Terms
Harvest. A 'harvest' refers to the initial gathering of content from a web site using the Harvest Spider.
6.1. Functions summary

<table>
<thead>
<tr>
<th>Use the Harvest Tool to…</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Monitor the status of harvests scheduled in the Analysis Tool.</td>
</tr>
<tr>
<td>• Delete completed harvests.</td>
</tr>
<tr>
<td>• Review completed harvest content, whether single-object or multi-object, prior to ingest.</td>
</tr>
<tr>
<td>• Review completed harvests; if desired, edit metadata and/or include/exclude content.</td>
</tr>
<tr>
<td>• Ingest harvested content into a repository.</td>
</tr>
<tr>
<td>• Launch a one-time Quick Harvest.</td>
</tr>
</tbody>
</table>

**Repository support.** At present the only option is for ingest is into the OCLC Digital Archive. In a future version of the Workbench, users will have the option of exporting into the intermediary repository structure known as the Hub and Spoke.

Below is a diagram that outlines the basic workflow within the Harvest Tool. The Harvest Statuses Screen of the Harvest Tool provides a list of the series that you have set up to be harvested and the harvest status for each series (Ready for Review, Saved, or Failed). As in other tools there are Details screens in the Harvest Tool: one for reviewing a single-object harvest and another for reviewing a multi-object harvest. Once necessary details have been filled in (such as for editing/extracting metadata and where to ingest—OCLC Digital Archives or to Hub and Spoke), the content will be packaged for ingest. **Text that is highlighted in gray reflects where the user is (i.e., what screen is on display) in the archiving process (see below).**
6. Harvest Tool > 6.2. Harvest Tool screen Status Tab

All scheduled harvests defined in the Analysis Tool and in the Quick Harvest tab in the Harvest Tool will be listed in the Harvest Tool screen Status Tab. Here you can see the status of harvests, delete harvests, review contents of completed single-object and/or multi-object harvests, and ingest the harvested content.

Harvests initiated by the Analysis Tool will have a Title that begins with “Series: ” Harvests initiated by the Quick Harvest tool will display the title assigned in the Quick Harvest form.

Sections 6.2.1 to 6.2.3 below provide additional guidance for working with the Harvest Tool screen > Status tab, following the functions identified in the diagram below.

Figure 42: Harvest Tool screen > Status tab

Under “Status of Review,” if it says “Saved,” this means that the harvest has been ingested. Either the ingest will result in error or, if successful, show date of ingest.
6.2.1. Filter/Sort Status

Users can filter the Status, Aging, Entity, and Series lists according to the current harvests by selecting options in the various pull-down menus and clicking Apply. Note that the top row (showing Status and Aging information) combine with the second row (Entity and Series) using a Boolean “and.” The list of scheduled harvests can also be sorted according to Title, Status (scheduled, started, complete, etc), Aging, Starting Point (URLs), Associated Entity, Harvest Completed Date, or Ingested Date by clicking the headings in the main list of scheduled harvests. Lists of completed harvests can also be sorted in this way.

Headings are further defined below.

Title
The title assigned to a harvest in the Analysis Tool or Quick Harvest tab.

Status of Harvest
The status of a harvest in the harvest execution workflow. Possible statuses are...

- **Ready for review:** Harvest was successfully completed and now awaits review prior to ingest.
- **Pending:** Harvest was initiated by the Quick Harvest feature and is set to start momentarily.
- **Harvest in process:** Harvest was successfully scheduled in a Series or Quick Harvest, and is in the process of collecting web content. When this is successfully completed, status will change to Ready for review.
- **Harvest failed/cancelled/interrupted:** Harvest was successfully scheduled in a Series or Quick Harvest, and was started, but did not successfully complete. Please contact OCLC to report unexplained failures.
- **Saved:** Harvest was successfully completed and has now been successfully ingested into a repository. This completes the Workbench workflow. At this juncture, users may review the material and ingest again, if desired. The “Status of Ingest” column will show the date of successful ingest or the ingest status

Aging
The completed harvests listed in the Harvest Tool will remain in system memory for thirty days. If the user does not successfully ingest a given harvest into the OCLC Digital Archive before a ten (10) day countdown ends, the harvested content will be permanently deleted.

Starting Point
The starting point URL of the harvest, as assigned in either the Analysis Tool or Quick Harvest tab.
6. Harvest Tool > 6.2. Harvest Tool screen Status Tab

**Associated Entity**
The entity, usually the creator of the harvest, as defined in either the Analysis Tool or Quick Harvest tab.

**Harvest Date Completed**
The date on which the harvest stopped (whether successful or not), if applicable.

**Status of Ingest**
If an ingest has been initiated, this column will show either the date that the content was ingested or the current status of the ingest. “Ingest” may refer to Digital Archive or Hub and Spoke (H & S) status.
Possible statuses for Digital Archive (DA) are…
- **Ingest Pending**: Ingest is set to start momentarily.
- **Ingest in process**: Content is currently being ingested into the repository.
- **Ingest error**: Ingest into the repository began, but failed and is not complete. Please contact OCLC for help with ingest errors.

Possible statuses for Hub and Spoke (H & S) are...
- **Packaging Pending**: Packaging is set to start momentarily.
- **Packaging in process**: Content is currently being packaged into the Hub and Spoke format.
- **Packaging error**: Packaging for Hub and Spoke failed. Please contact OCLC for help with packaging errors.

Ingest of harvested content may be enabled by clicking on a checkbox in the Ingest column. In this column, more than one checkbox may be clicked if more than one ingest is desired. A similar functionality is behind the Ingest button on the Details screen (accessed by clicking on “Details” on the Harvest Tool > Status tab screen)—except that multiple ingests cannot be done there. In addition, on the details screen for multi-object harvests, there will be checkboxes for ingesting to the OCLC Digital Archives ("Ingest to DA"), or to the Hub and Spoke ("Ingest to H&S").

There is also a checkbox for extracting metadata, if you so choose. The point here is that user may ingest everything for a multi-object harvest at once; users may also decide which object gets ingested where (that is, an object within the multi-object harvest may be selected for ingest to the DA, while another object in that harvest may go to the H&S).

**Note:** There are two ingest destination options: ingests may be into the OCLC Digital Archive (abbreviated DA in the screenshots) or into the Hub&Spoke intermediary repository architecture (abbreviated H&S in the screenshots), which prepares the harvest for ingest into other repositories such as DSpace and Eprints. You may select both options if you wish your harvested content to be ingested in both the DA and the H&S.
6.2.3. Delete Harvests

Unwanted harvests can be deleted by checking the Delete box and selecting Update for each screen.

6.2.4. Completed Harvest Details

Clicking on Details launches a new screen that displays a record about that particular completed harvest entry. Also note that all completed harvests, regardless of status, will age off the system in thirty days.

Two kinds of screens: Users will see one of two kinds of screens—either a screen for a single-object harvest or a screen for a multi-object harvest. Which screen you see depends on how you configured the content to be harvested when scheduling the harvest in the Analysis Tool.

Below is a description of what the Details screens look like for a single-object harvest and a multi-object harvest that are ready for review. Screenshots of the Details screen for these two kinds of object harvests will then follow, each image accompanied by an explanation of the key functionalities on the screens.

- **Ready for review** entries have a Details screen with buttons for Edit DC Metadata, Extract DC Metadata, Ingest (including where to ingest—Digital Archive or Hub and Spoke) and Delete.
  
  - Single-object harvest: The Details screen for a single-object harvest also shows an editable Website Tree (details in next section) for reviewing the harvested content. This makes sense, as a ready for review status means a harvest has been successfully completed in the Harvest Tool, and is now awaiting ingest. The Review screen is where these ingest steps are initiated.
  
  - Multi-object harvest: The Details screen for a multi-object harvest shows the Object Paths for the objects harvested, as well as buttons for editing DC metadata and extracting metadata for them. In addition, in the Harvest Status report itself, the file types that were harvested are listed. Note: A multi-object harvest is a harvest of individual objects (potentially objects of mixed file formats).

- **Saved** entries have been ingested into a repository at least once. They will continue to have the same options as a Ready for review entry and may be ingested as many times as desirable (to various repositories in the future). Saved entries will be aged off by the system after thirty days, or may be deleted manually.

The Details screens for a single-object harvest and for a multi-object harvest now follow. They display a ready for review completed harvest entry listed in the Harvest tab.
6. Harvest Tool > 6.2. Harvest Tool screen Status Tab

The above screenshot shows the destination and metadata options for a single-object harvest, which typically is content from a single website.

Figure 43: Harvest Tool > Status tab > Details Screen > Single-object harvest

The screenshot below shows the Details screen for a multi-object harvest. A multi-object harvest is a harvest of different kinds of objects—hence (as displayed in the upper part of the screen) the variety of File Types Harvested. This is different from a single-object harvest, which is a harvest of one type of object (such as a whole website). Note the various checkboxes for ingesting to the DA and the H&S, and for extracting metadata.
6. Harvest Tool > 6.2. Harvest Tool screen Status Tab

Figure 44: Harvest Tool > Status tab > Details screen > Multi-object harvest

A multi-object harvest contains individual objects of various file formats.

Selecting Harvest Metadata to WorldCat will indicate that DC Metadata should be included in WorldCat. Although you may check or uncheck this box, currently the Harvesting of Metadata to WorldCat feature is not activated in the Workbench. OCLC will send notification when it is turned on.

A final opportunity to edit the Dublin Core record, which has been copied from the Series that was used to schedule this harvest, or from the basic metadata used to set up a Quick Harvest. For setting up Series DC Metadata, See 5.2.5 (Analysis Tool > Series tab > Series details screen) section for details.
6.2.7. Extract DC Metadata

As in the Analysis Tool, DC metadata may be extracted automatically. Users are also given the option to have extracted metadata returned in Dewey numbers (click “OK” to enable this option).

The next figure shows a screenshot of a form filled with extracted metadata in Dewey numbers.

Figure 45: Harvest Tool > click Extract Metadata button (click OK – brings back Dewey number form) > Harvest-level Dublin Core Metadata Editor screen
6. Harvest Tool > 6.2. Harvest Tool screen Status Tab

Clicking on “Cancel” will return simply the extracted metadata for the harvested content:

Figure 46: Harvest Tool > click Extract Metadata (click cancel – no Dewey number form)
Harvest-level Dublin Core Metadata Editor screen

Reminder: As with the Extract Metadata button in the Analysis Tool, any extracted metadata will have “Extracted: ” preceding it.
To ingest this harvest into a repository:

- View the Details screen again for the now completed harvest, which will show a status of Ready for review or Saved. The Details screen now offers buttons for Delete and Ingest. Click Ingest OR
- From the Harvest Tool: Status Tab, select the checkbox under the Ingest column for the desired completed harvest and press Save/Update.

**Note:** You may wish to review content included in a completed harvest prior to ingesting in a repository. For 'ready for review' or 'Saved' entries, the Details screen includes a Web Site Tree (located below the Save etc buttons), where this content can be reviewed. See 7.2.2.5 below.

For ‘Ready for review’ and ‘Saved’ entries only, the bottom half of the Details screen displays a tree representing content included in the package. Using the Website Tree, users can identify content to be included or excluded from the completed package sent to a repository.

**Figure 47:** Harvest Tool > Status tab > click Details button beside 'ready for review' or 'saved' entry

Browsing the Web Site Tree.
To view content included within a domain or directory, select the + box (+) to expand the tree. Your starting point URL will appear in bold in the tree. Please see the next screenshot below.
Exclude/Include Content
By default all content collected by the harvest is included in a package. Included content is indicated with a green check mark (✓).

To exclude content, click the green check mark, which will become a red X (X)

Checking/unchecking marks in the tree includes/excludes all children below the level selected. E.g. to exclude all content from a particular directory it is not necessary to select each individual file or subdirectory, simply select the top level directory and exclusion will automatically be applied to lower level content.

Be careful if you use the option to Exclude content. Note that if your Starting Point URL (i.e., -- the point at which the Harvest Spider entered the web site which makes up this Series) appears deeper in the tree, it is possible to inadvertently exclude it. Your content will still be ingested into the repository, but there will be no starting point for access. (This is analogous to putting items into a box, but magically deleting the lid. It is sealed up, with no point of entry.) It is also possible to delete dependencies – content that is necessary to render a web page to the user such as images or style sheets – which you should be careful to do only intentionally.

To save changes to the Web Site Tree without creating a ingesting, select the Save button. To exit the Web Site Tree without saving changes, select the Cancel button.

Next steps
To ingest a harvest into a repository, go to its Details screen from the Harvest tab main screen. The Details screen for Harvested items provides buttons for Delete and Ingest. Alternatively, in the Harvest Tool, Status Tab, select the checkbox under the Ingest column for the desired completed harvest and press the Save/Update button.

You have now used the Harvest Tool to review successfully completed harvests, and ingest them into a repository. This Ingest of harvested content and its associated metadata is the final main workflow step in the Workbench.
6.3. Harvest Tool Summary

The Harvest Tool allows you to initiate the final harvesting and ingest steps.

You have now used the Harvest Tool to review the status of scheduled harvests created using the Analysis Tool > Series function, as well as Quick Harvests. As harvests are executed, you can review their contents and metadata before ingesting the harvested content, also through the Harvest Tool.

6.4. Quick Harvest Tab

As an alternative to a scheduled harvest, which are created in the Analysis Tool, you may also launch a Quick Harvest, or one-time harvest of web content, using the Harvest Tool. To configure a Quick Harvest, click on the Quick Harvest tab in the second row under the Harvest Tool. You will be brought to a new screen where you will enter some basic metadata and select spider settings. To learn more about configuring the Harvest Spider, see Spider Settings (Section 8.3).

Figure 48: Harvest Tool > Quick Harvest tab

After filling in the fields, click on Harvest to launch the harvest spider. The initiated Quick Harvest will immediately appear in the list of scheduled harvests in the Harvest Tool Status tab and you will now see a blank Quick Harvest form, ready for a new harvest.

6.5. Workbench Recap

This completes the overview of the four main tools which make up the Web Archives Workbench workspace. The remaining two tools, Alerts and System Tools contain services and features that interact with the four main tools. These are discussed next, in the final section of this Guide.
6.5.1. Workbench functions summary

<table>
<thead>
<tr>
<th>Tool</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery Tool</td>
<td>Comprising the Entry Points and Domains tabs, the Discovery Tool helps to identify potentially relevant web sites by crawling relevant &quot;seed&quot; Entry Points to generate a list of domains that they link to. At the end of this process you have a list of domains that defines the sub-set of the web relevant for your archiving purposes. From here, you can use the Properties and Analysis Tools to manage creator information about domains, and associate this information with harvests of content.</td>
</tr>
<tr>
<td>Properties Tool</td>
<td>Comprising the Entities tab, the Properties Tool allows you to maintain information about content creators or 'Entities' (e.g., government agencies), and associate them with the domains and web sites they are responsible for. The Properties Tool also allows you to describe the relationships (e.g., parent/child) of Entities with one another, as well enter high-level metadata about them that may be inherited by content harvested from their web sites. Importantly, the Properties tool can also be used to create and associate Series with Entities' web sites. Series and harvests are then further managed using the Analysis and Harvest/Package Tool.</td>
</tr>
<tr>
<td>Analysis Tool</td>
<td>Comprising the Series tab, the Analysis Tool allows users to define archival Series, associate metadata with these series, and schedule recurring harvests of web content. Harvesting activities are then monitored and managed in the Harvest Tool.</td>
</tr>
<tr>
<td>Harvest Tool</td>
<td>Comprising the Harvester and Quick Harvest tabs, the Harvest Tool lists all harvests within the Workbench, including Series harvests scheduled using the Analysis Tool as well as Quick Harvests. It is used to monitor their status, initiate the final harvesting and ingest steps for the completed harvests tracked in the Harvest Tool, including reviewing harvest contents and metadata before ingest. This is the final step in the Web Archives Workbench workflow. It also offers a separate Quick Harvest feature.</td>
</tr>
</tbody>
</table>
7. Alerts

7.1. Introduction

The Alerts Tool contains a collection of reports and alerts on Workbench processes. Note that this area is in an early stage of development.

The Alerts Tool contains one functional area:

- the **Domain alerts** page identifies domains discovered by the Domain Spider which suffer from known errors.

Use the Alerts Tool to...

- Review reports and alerts on Workbench processes. Currently there is a Domain alert available.

The diagram below reflects the workflow that is found in the Alerts Tool. **Text that is highlighted in gray reflects where the user is (i.e., what screen is on display) in the archiving process (see below).**

Unlike other tools in the Web-Archiving Workbench suite, the **Alerts Tool does not have a separate Alerts Details screen**. The button for Details on the Alerts Screen will link back to the appropriate tool screen, depending on what kind of alert is listed. For example, if the alert is a spider problem that prevented an analysis from completing successfully, then the Details button for this alert will link back to the Analysis Details screen for that particular analysis. (See first alert listed on next screenshot.)
7.2. Domain Alerts

Domain alerts identify domains discovered by the Domain Spider which suffer from known errors. For example a domain that is no longer active will report an error. Users can use the Domain alerts to remove unwanted domains from the Discovery Tool, or indicate that they need to identify a new domain for content that has moved to a new domain.

7.2.1. Filter/Sort Alerts

Domain Alerts can be filtered according to whether domains are/are not obsolete. Click Apply Filter to apply the filter.

7.2.2. Delete Alerts

To remove unwanted Alerts, check the Delete check box and select the Update button.

7.3. Alerts Tool Summary

The Alerts Tool contains a collection of reports and alerts on Workbench processes. The kinds of alerts you may monitor or find results for are listed below:

- New domain discovered by a domain spider run
- New node (folder) found during analysis
- Content harvested and auto-ingested
- Content harvested from series and auto-ingested
• DNS error for domain
• Analysis comparison found missing folders
• Analysis change threshold met
• Harvest started but did not complete successfully
• Analysis started but did not complete successfully
• Domain spider started but did not complete successfully
• Harvest started but timed out
• Analysis started but timed out
• Domain spider started but timed out
• Error ingesting content to the OCLC Digital Archive
• A saved analysis run will be overwritten by a new run within the next seven days or less.
• Error packaging content.
8. System Tools

8.1. Introduction

The System Tools tab contains a number of behind-the-scenes functions that affect and report on activities of the five main tools of the Workbench.

The Systems Tools are divided into four functional areas:

- the Audit Log page, which displays recent Workbench activities and events;
- the Spider Settings page, where you can configure default Domain, Analysis, and Harvest spider settings, and create additional Domain, Analysis, and Harvest spiders with custom settings.
- the Import/Export page, which you can import or export a variety of metadata commonly used in the Workbench.
- the Reports page, which generates reports on activities of the main five Workbench tools. This area is in development, and currently offers a view of in-development Entity and Series reports.

8.1.1. Functions summary

<table>
<thead>
<tr>
<th>Use the Systems Tool to…</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Configure spider settings and create new spiders.</td>
</tr>
<tr>
<td>• Import/export metadata used in the Workbench.</td>
</tr>
<tr>
<td>• View logs of recent activities and reports.</td>
</tr>
</tbody>
</table>

The diagram below displays the workflow of the Systems Tool. **Text that is highlighted in gray reflects where the user is (i.e., what screen is on display) in the archiving process (see below).**
The Audit Log displays recent Workbench activities and events, showing events in the system when they happened. Examples of events include user logins, beginning and ending of spider activities, and the beginning and ending of content harvests.

The Audit Log tab

8.2. Audit Log tab

8.3. Spider Settings tab

8.3.1. Harvest Off Domain

8.3.2. Honor Excluded Domains

New options in spider settings:
1. Harvest off domain (i.e., go somewhere else after going however many levels it’s been set) and 2. Honor excluded domains (i.e., domains pulled from Discover Tool list), during the crawl. Default is set to “no”, which will limit the spider to crawling the starting point domain.

This setting relates to the Discovery Tool Domains list. When “yes” is selected, the spider will restrict itself to not traverse onto domains listed as “out of scope” in the Domains tab of the Discovery Tool. Any domains encountered during the crawl that are not in the Domains list in the Discovery Tool will automatically be collected because the scope has not been set for them.
8. System Tools > 8.3. Spider Settings tab

Clicking a spider entry’s Details button on the System Tools > Spider Settings page brings up the following screen.

Figure 51: System Tools > Spider Settings Details screen

Spider Settings Object Name: The name assigned to these Spider Settings. This name will appear elsewhere in the WAW where spiders can be selected.

What Type of Spider do these Settings Apply To?: Associate spider settings with Domain, Analysis or Harvest web crawls. The Domain spider is used in the Discovery Tool, where it crawls the “seed sites” that the user has input and generates a list of domains, or the sites that the spider points to. The Analysis Spider is used in the upcoming Site Analysis Tool, and the Harvest Spider for all harvests scheduled in the Series tab or the Quick Harvest tab. The Harvest spider is launched in order to harvest content from inside a seed site (or Entry Point) domain.

How deep do you want to spider (levels)?: Select the depth from Entry Point address that you want the spider to survey. Increasing the number of levels increases the time required for the spider to survey entry points, but may produce a broader number of domains. For Harvest Spiders (which only collect content from within the Entry Point domain), the number of levels to be spidered only applies to the domain specified in the Entry Point.

Depth of 0 (zero): For By Links, the spider will not follow any links away from the entry point page. For By Path, the spider will not move down any directory levels from the entry point.

Unlimited Depth: Even though you can choose unlimited depth, keep in mind that...
there is currently a maximum time limit for the spider and it will stop when that time is reached.

**Spider Time Limit:** Enter the number of hours and minutes that the spider will be allowed to run. At the end of the specified time the spider will cease surveying. The maximum time limit you may enter is currently **167 hours 59 minutes**. Please contact OCLC if you find that this maximum time limit is too short.

**Robot Exclusions:** Indicate whether the spider should honor instructions specified by the Robot Exclusion Protocol. More information about robot exclusions is available at [http://www.robotstxt.org/](http://www.robotstxt.org/)

**How should the spider traverse the site?:**

**By Links:** From the entry point, the harvester follows all links, collecting content from anywhere within the domain. You specify how many links away from the entry point the harvester goes by using the depth setting.

**By Path:** Sometimes By Links harvests too much content. Use the By Path method to limit your harvest based on the directory structure of the web site. From the entry point, the harvester moves through subdirectories collecting the linked content they contain. You specify how many subdirectories down from the entry point the harvester goes by using the depth setting.

Note: OCLC recommends By Links as a better default option for beginners, as By Path can be more limiting.

**Duplicating Spiders**

To create a new spider based on current settings, select Save As. A pop-up box will allow you to assign a new name to the settings.
The Web Archives Workbench allows users to import or export a variety of metadata commonly used in the Workbench. Using the import/export features, users can add a list of Entities, export Entities defined in WAW, include Subject Headings, or a list of known domains for use in the Discovery Tool.

**8.4. Import/Export Tab**

**8.4.1. Import Entities**

To import a list of entity records from an existing source:
1) Create a tab delimited file in which:
   a. Each row represents information related to one entity,
   b. The first two rows of data are blank (or contain label information; these rows will be ignored upon Import), and
   c. The elements are in the following order:
      - Local ID
      - Local ID Code
      - Preferred Name
      - Key Name
      - Alias1
      - Alias2
      - Alias3
      - MacroAppraisal Score
      - History
      - Authority/Mandate
      - Public Notes
      - BLANK FIELD (this field MUST be left empty)
      - Begin Date
      - End Date
      - Type/Category
      - Functions
      - Subject Heading1

![Figure 52: System Tools > Import / Export tab > Import/Export screen](image)
Subject Heading 2
Subject Heading 3
Associated Personal Names1
Associated Personal Names2
Associated Personal Names 3
Website1
Website2
Website3
Cataloging Notes
Contact Name
Contact Title
Contact Address
Contact Address2
ContactCity
ContactState
ContactZIP
ContactPhone
ContactFax
ContactEmail

2) The first two rows of data must be empty
3) Save the file to a local drive
4) Click the Properties Tab
5) Click the Import/Export Tab
6) Next to Entities Import, Click Browse and select your tab delimited file
7) Select Import

8.4.2. Exporting Entities

To export a tab delimited file of your existing entity records:

1) Next to Entities Export, enter a name for the file to be exported. Just enter the filename and an extension. Do not enter the desired path at this time.
2) Click Export.
3) The system will create the file; you will be asked where you would like to store it locally.
4) The exported file is a tab delimited file that can be imported into other software programs for easier viewing and editing.

8.4.3. Import Domains

To import a list of domains from an existing source to the Discovery Tool:

1) Create a tab delimited file in which:
   a. Each row represents information related to one domain
   b. The first two rows of data are blank (or contain label information; these rows will be ignored upon Import), and
   c. The elements are in the following order:

   Domain Name ([www.mysite.org](http://www.mysite.org)) (required)
   Scope (options for this field are “New” “Out” “In” – please note that this field is case sensitive!)
   Obsolete (include “True” if the domain is obsolete; “False” if the domain is not obsolete)
   Notes
2) The first two rows of data **must** be empty
3) Save the file to a local drive
4) Click the System Tools Tab, then select the Import/Export Sub-Tab
5) Next to Domain Import, Click Browse and select your tab delimited file
6) Select Import

8.4.4. **Export Domains**

To export a tab-delimited file of your existing domains:
1) Next to Domains Export, enter a name for the file to be exported (filename and extension only – file path is not needed)
2) Click Export. *Please note, if you have many existing domains it will take some time to for the system to export them*
3) The system will create the file; you will be asked where you would like to store it locally.
4) The exported file is a tab delimited file that can be imported into other software programs for easier viewing and editing.

Tab-Delimited File will be formatted in the following order:
Name, Reverse Domain, IP, Scope, Notes, Entities (all associated)

8.4.5. **Import Subject Headings**

The Workbench comes pre-loaded with the GILS subject thesauri. You may load in one additional set of subject headings/topics.

To import subject headings into the Workbench:
1) You must have the Edit Token in order to import data to the Workbench.
2) The import subject headings functionality supports a hierarchical structure of headings/topics (like the GILS Tree). In order to create this hierarchy in a tab delimited file, the data you want to import must be formatted as follows:
   Level 1
   . Level 2
   . . Level 3
   . . . Level 4
(note that this formatting convention is a `<period>` followed by a `<space>`)  
An example would look like:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Agriculture and food production</td>
</tr>
<tr>
<td>2</td>
<td>Agricultural finance</td>
</tr>
<tr>
<td>3</td>
<td>Agricultural statistics</td>
</tr>
<tr>
<td>4</td>
<td>Agriculture pages for kids</td>
</tr>
<tr>
<td>5</td>
<td>Aquaculture</td>
</tr>
<tr>
<td>6</td>
<td>Crops</td>
</tr>
<tr>
<td>7</td>
<td>Fruits and vegetables</td>
</tr>
<tr>
<td>8</td>
<td>. . Corn</td>
</tr>
<tr>
<td>9</td>
<td>. . . Cranberries</td>
</tr>
<tr>
<td>10</td>
<td>. . . . Grains</td>
</tr>
<tr>
<td>11</td>
<td>Farmlands</td>
</tr>
<tr>
<td>12</td>
<td>Fisheries</td>
</tr>
<tr>
<td>13</td>
<td>Irrigation</td>
</tr>
<tr>
<td>14</td>
<td>Livestock</td>
</tr>
<tr>
<td>15</td>
<td>Pesticides</td>
</tr>
<tr>
<td>16</td>
<td>Soil erosion</td>
</tr>
<tr>
<td>17</td>
<td>Business and industry</td>
</tr>
<tr>
<td>18</td>
<td>Banking</td>
</tr>
<tr>
<td>19</td>
<td>Business and professional licenses</td>
</tr>
<tr>
<td>20</td>
<td>Business related licenses</td>
</tr>
</tbody>
</table>

3) When your data is formatted, create a tab delimited file.
4) Save the file to a local drive.
5) Click the Properties Tab.
8. System Tools > 8.5. Reports Tab

6) Click the Import/Export Tab.
7) Next to Subject Headings Import, click Browse and select your tab delimited file.
8) Input the name you want the system to use for your subject headings.
9) Select Import.

You must store and update your subject heading files locally. When you have made changes in the local file that you want to upload to the Workbench, you must re-import the file. Re-importing replaces your current subject headings file with the new, updated subject headings file.

As development of the rest of the Workbench continues, the Reports section will offer additional reports on its activities. Currently two in-development reports on Entities and Series are listed. These reports will become fully functional in later releases of the Workbench.

Figure 53: System Tools > Reports tab > Reports screen

8.6. Systems Tools summary

The System Tools tab contains a number of behind-the-scenes functions that affect and report on activities of the five main tools of the Workbench. Most importantly, it contains the Spider Settings page, where you can configure default Domain, Analysis, and Harvest spider settings, and create additional Domain, Analysis, and Harvest spiders with custom settings. The System Tools also includes a metadata import/export utility, and in-development reports and audit logs of Workbench activities.
9. Appendix

9.1. Additional resources

The following resources provide additional information relevant to the Web Archives Workbench.

1. Dublin Core Metadata Initiative
   http://dublincore.org

2. Web Archives Workbench background materials
   http://www.ndlipp.uiuc.edu
   Go to Core Activities > Tools Development

3. Digital Preservation Pathfinder
   http://www.ndlipp.uiuc.edu
   Go to Resources > Digital Preservation Pathfinder.