

BEN(BioSci Education Network)
Metadata Harvesting Quick Start Guide
April 2003

Introduction

The purpose of the metadata harvester is to provide a portable plug-in application that will allow BEN Partners to automatically send their BEN metadata records to the BEN server. The Metadata Harvester has two components which we shall call the Reaper, which resides on the Partner web server, and the Thresher, the plug-in which resides on the central BEN Portal server.

This document provides an overview of the Reaper, including installation and configuration. It was written for the release of Reaper v2.0.1.

Installation

The Reaper can be installed on either a Windows or Unix system. The installation package can be downloaded from the BEN Project Site (http://www.biosciednet.org/project_site/). Installation should be done by someone with administrative privileges on the system being used, and preferably by somebody familiar with Perl.

The README.htm file and, if installing on Windows, the win32_installation_readme.htm file, included in the installation package, provide step-by-step instructions on installing the Reaper and its required components.

Configuration

To understand how to utilize the metadata harvester, it is first necessary to understand the BEN Metadata Specification and how it is mapped to your collection. The specification is available via the BEN Project Site. Each required metadata field should correspond with a data field in your collection. It is important to make sure that the records in your collection use the appropriate vocabulary and specifications as defined in the BEN Metadata Specification. For instance, when mapping a collections field to the Educational.Context metadata field, you should make sure that all of the values for that field fall within the defined vocabulary list.

Once the mapping has been completed on a conceptual level, then it is necessary to explain that mapping in a format that the Reaper software can understand. This is done with an XML map file. The map file translates the information in your database into an XML record that can then be sent to the BEN portal. An example map file is provided with the Reaper download, and should provide a good starting point for mapping your database fields. More information on mapping relational databases to XML is available at <http://www.rpbouret.com/xmldbms/index.htm>.

Activation and Usage

Once the Reaper has been installed and configured, it may be tested using the included test script (note: this will only test that the software has been installed correctly; it will not verify that you have mapped your database fields correctly). When you are satisfied that the Reaper is working and the mapping of your database collections fields is satisfactory, then you may activate the harvester, either by running the `harvesterActivate.cgi` script or by notifying BEN of the URL where your `reaper.cgi` program is located.

After this point, the BEN portal will begin sending out requests to your Reaper for metadata records in your collection. After the initial harvest, the Reaper will send requests for only records that have been added or modified since the last harvest. The Reaper should not require maintenance unless changes are made to the collections database.

Requirements

Perl 5.8 or higher

Ability to install Perl modules, preferably using CPAN (Unix) or PPM (Windows)

Ability to run Perl over the web via CGI or a Perl-capable web server

SQL database storing metadata records that are compliant with the BEN-1.0 Metadata Specification

Software Specification and Components

The Metadata Harvester operates on the following model: The Thresher will periodically send out Open Archives Initiatives (OAI) Protocol (http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html) compliant HTTP requests requesting metadata updates to the Reapers running on the web servers of the BEN Partners. The Reaper application, a Perl script running on the web server of the BEN Partner Society, will access recently added or modified metadata records of the Partner Society from a SQL database and will convert these records into XML data. The XML data will then be returned via HTTP, again in OAI-compliant form, over the Internet to the Thresher on the BEN Server. The Thresher will then parse the XML, making sure it is well-formed and valid. The Thresher will then use the XML data to recreate the metadata record in the central BEN database. The metadata record will show up in the search engine when a separate application has updated its collections file with the information from the database.