OAI Toolkit – Version 0.6

Release Notes

May 26, 2009

© eXtensible Catalog

University of Rochester

1325 Mt. Hope, Suite 102

PO Box 278960

Rochester, NY 14627-8960

http://www.extensiblecatalog.org

How to preserve existing settings from previous version when upgrading

1. **Extract** the **importer package** in the OAIToolkit directory. Make a **back-up of the previous OAIToolkit** folder to make sure you do not lose important files.

2. Since the storage structures haven't changed, you don't need to reindex your records, and you can use your existing MySQL database and/or Lucene index. **So maintain the "lucene_index" folder in the OAIToolkit folder.**

3. The best method to preserve your existing settings, if you extract the content into a new directory on your disk, is to copy your existing configuration files from the older directory to the new one **except for the OAIToolkit.log4j.properties and OAIToolkit.directory.properties** file. **Use these new configuration properties file** as a template which is created in the home directory of the OAIToolkit folder.

As usual after doing the required changes to it, **copy** it to the to CATALINA_HOME\bin, where CATALINA_HOME is the default folder of the Tomcat Installation. In Windows it would be C:\Program Files\Apache Software Foundation\Tomcat 6.0\bin

4. **Use all the other files and folders from the new OAIToolkit importer and server packages**. Use the sample scripts as a template to make changes to customize the scripts according to your ILS.

What's NEW in this release

- **Modified the Marc4j source code (which is embedded in the OA**I Toolkit) to improve error reporting. When the OAI Toolkit is processing a file with multiple raw MA**R**C records into MARC XML records, and it encounters a problematic or invalid record, it now does a better job of reporting which record has an issue, and what that issue is.

- Convert and load process error logging messages have been improved to be clearer and more **understandable.** This change addresses the issue reported here: http://code.google.com/p/xcoaitoolkit/issues/detail?id=1

- **This release includes a document which describes how to use the OAI Toolkit convert and load logs to remove the errors in the MARC records (**http://xcoaitoolkit.googlecode.com/files/Debugging_Librarian_Convert_Load_logs_for_errors_in_MARC_records.pdf**).** You could find this file in the Downloads section of the XC Toolkit Google Code page.

- **Logging mechanism** of the OAI Toolkit has been **improved** to have **one more log file** (**lucene_dbStatistics.log)** in addition to the four available. This new log file captures information about the current state of the lucene index, including the **number of records, their types, and the count of deleted record**. This log file is written to only when a provided utility script is run, called "**lucene_dbStatistics.bat**" (on **Windows)** and "**lucene_dbStatistics.sh**" (on **Linux**). This change addresses the issue reported here: http://code.google.com/p/xcoaitoolkit/issues/detail?id=13.

- **New feature: When an ILS or other repository had provided data that has been loaded into an OAI Toolkit instance, it is possible that some of the data will become obsolete over time. If the data is deleted within the ILS or other repository, a mechanism is needed to report those records to the OAI Toolkit. This way the OAI Toolkit can pass on that deleted record information to OAI**

Harvesters down the line. A new option exists in to tell the OAI Toolkit about specific records that must be deleted.

The OAI Toolkit supports by having the parameter ***-fileof_deletedrecords*** which when used with the normal convert and load or just load while importing, would store that file's records **as deleted** in the **Lucene/MySQL Database**. An administrator would need to export a set of deleted records (since the last update) and then load them into the OAI Toolkit using this parameter, in order for the OAI Toolkit to have knowledge that these items have been deleted.

Deleted script is also provided which calls convert and load with this parameter appended. The script is called "**convertload_as_deleted.bat" for Windows and "convertload_as_deleted.sh" for Linux** This change addresses the issue reported here: http://code.google.com/p/xcoaitoolkit/issues/detail?id=12

- OAI Toolkit now **supports folders and sub-folders inside the "marc" folder** when importing raw MARC files. All files in the folders and sub-folders would be successfully **imported.** This change addresses the issue reported here: http://code.google.com/p/xcoaitoolkit/issues/detail?id=4
- OAI Toolkit Manual has been updated to include changes for this version (0.6).
- The SVN repository has been restructured to have **folders inside the sample_scripts** folder, so that it could be used to distribute **sample scripts** for the range of ILSs that XC will support. Currently, only sample scripts from University of Rochester Voyager are provided.
- **Sample script "extract_error_recordno.bat" for Windows** and "extract_error_recordno.sh" for **Linux** have been provided to automate the process of extracting the error record numbers from the librarian_convert.log. This will assist in the process of finding the problematic records and addressing the issues.
- Corrected the way that the OAI Toolkit reports available formats via OAI-PMH. The metadataFormat name was changed from "marc21" to "marcxml" in the metadataFormats.xml file in the resources directory of the OAI Server Package.

Bug Fixes

- **Harvest Slow-down bug** when **Lucene Database** is used has **been solved**. The **OAI Toolkit response time** to any requests made for the records **now roughly remains constant**, which **earlier** it used to **increase logarithmically** as the size of the Database increased. Thus now the harvests from the OAI Toolkit would be efficient and faster than before. (http://code.google.com/p/xcoaitoolkit/issues/detail?id=3)
- **Naming of the directories** was **different** when a user would **run convert and load together or separately**. This **bug has been rectified** and now it is consistent directory naming in the OAI Toolkit. (http://code.google.com/p/xcoaitoolkit/issues/detail?id=2).
- **Proper Validation of the URL parameters for the OAI Server** bug has been rectified so that it shows the appropriate message when any incorrect parameters have been supplied to it. (http://code.google.com/p/xcoaitoolkit/issues/detail?id=10).
- The configuration file "**OAIToolkit.directory.properties**" now **supports the slashes which are operating system dependent**. Therefore inside the file, the folder location can use (/) front-slash for Linux and (\) back-slash for Windows, and everything would still work fine in OAI Toolkit. (http://code.google.com/p/xcoaitoolkit/issues/detail?id=14).

- Corrected the way that the OAI Toolkit reports available formats via OAI-PMH. The metadataFormat name was changed from “marc21” to “marcxml” in the metadataFormats.xml file in the resources directory of the OAI Server Package.

- Log4j was initialized before it should be actually, and therefore **OAI Toolkit gave un-necessary log4j warnings**, **when still the logs were working** fine. This has been **rectified** so that the user **does not get a wrong impression** of the logging system. (http://code.google.com/p/xcoaitoolkit/issues/detail?id=15).

Known Issues/Feature Requests

If you see anything on the known issues list that you would like the XC team to prioritize, or if you know of an issue that is not on the list, please report it in the Issue Tracker on the XC website.

- The OAI Toolkit is not yet capable of processing item data; we are working on gathering information from partners on how item data is available, and what item data is available, before we implement this feature (http://code.google.com/p/xcoaitoolkit/issues/detail?id=7).

- There has been a request to make documentation available in HTML; this is being reviewed to determine how documents should be formatted when put into HTML (http://code.google.com/p/xcoaitoolkit/issues/detail?id=8).

- The ability to import Dublin Core records may be added to a future release (http://code.google.com/p/xcoaitoolkit/issues/detail?id=5).

- Right now the OAI Toolkit uses Struts version 1, but at some point we would like to shift it to Struts version 2. Although it is not impacting much at this moment. (http://code.google.com/p/xcoaitoolkit/issues/detail?id=9).

Reporting Bugs/Requesting Features

Bug reporting and feature requesting can be done through the Google Code website, using the login credentials of a Google Account. The issue tracker is located at http://code.google.com/p/xcoaitoolkit/issues/list. Please use this form to report bugs as well as difficulties with the documentation, or to request features.

Contact

Please contact Shreyansh Vakil at svakil@library.rochester.edu with any questions or concerns.

Previous releases

- Release 0.5 (March 27, 2009)
    - **What’s new in this release**
        - We have made the **OAI Toolkit Code as open source software via the Google Code**. You can access the project via (http://code.google.com/p/xcoaitoolkit/). There are links to some of the important files like the OAI Toolkit Manual and the Release Notes on the main page of the project. There is a section called the

Downloads in there by which one could access the various packages which would help to install the Toolkit.

- The **bug fixes and feature requests** of this 0.5 version point to the **Google Code website issues** section. The **earlier version links of bug fixes and feature requests** of the OAI Toolkit which pointed to the XC Website are **no longer valid**.
- OAI Toolkit Manual has been updated to accommodate for the current 0.5 version.
- **We have improved the Logging mechanism of the OAI Toolkit** so that it is now easier to identify the errors in the source marc files. There are now 3 log files inside the OAI Toolkit which are as follows:

  a. Librarian_convert.log

  This file is very useful for the librarians who are using the log files to find the location of the errors in the raw MARC files after the convert process has been finished.

  We are working on improving the logging for this step more by trying to solve some of the bugs in Marc4j and thereby participating in the marc4j open source code. We are changing the source code and committing the same back to the open source Marc4j code

  b. Librarian_load.log

  This file is very useful for the librarians who are using the log files to find the location of the errors in the MARC-XML files after the loading process. They could see these errors , locate the record and rectify the record data using this.

  c. Programmer.log

  This file would be used in general for the normal programmers who would like to see the Debug/Info statements and know in general the processing of the OAI Toolkit. It would also detect the errors if any in the OAIToolkit.

  d. Toolkit.log

  This file basically duplicated all the statements from the librarian.log and the programmer.log files. If one wants to see all the logs, this is the log file he should see.

  The logging mechanism in place is now also able to identify the type of MARC record where the error is. There are also some cosmetic changes in the logs to improve the formatting of it and make it more user-friendly and readable.
- **The *-ignore_repository_code* flag was rectified**. It governs the merging of 001 and 003 behaved as the opposite as we documented it in the last OAI Toolkit version. The default value is false, which means, that the toolkit should merge 001 and 003, but I forgot to add an exclamation sign (!), and the toolkit did the opposite: merge if the user don’t want to merge. It works fine now.

  - **In the interest of differentiating the weird characters in the Leader and other fields, two flags** were added to the OAI Toolkit. Those are:
  - `translate_leader_bad_chars_to_zero`

    This flag as the name suggests replaces the weird/bad characters in the **Leader field** of the raw MARC files to zero instead of the weird character.
  - `translate_nonleader_bad_chars_to_spaces`

    This flag as the name suggests replaces the weird/bad characters in the **Non-Leader fields** of the raw MARC files to a single space instead of the weird character.

    If a person wants to remove the weird characters in both the Leader as well as the other fields he would have to use both of these flags from now on.

    This feature request was proposed by Chris Delis from CARLI.
  - **Also means that the `–replace_weird_characters` flag is longer in existence now.** As the above two flags have been added, the `–replace_weird_characters` flag **does not exist**.
  - The SVN repository has been restructured to contain the older versions in the /tags directory and the working version inside the /trunk directory.

- **Bug Fixes**
  - The **cache directory** which was not getting created automatically is now fixed in the program and it does get created when the toolkit is installed.
  - Earlier a **Null Pointer Exception when merging 003 and 001** in records without 001 did come, but now the Toolkit handles it properly.
  - There were some **speed issues with the import**. As the import continued, it seemed to become slower. But that bug is fixed and the **speed has improved considerably**.
  - There was a bug in the OAI Toolkit which gave a "**Communications link failure" error** while **harvesting the records** from OAI Toolkit after period of inactivity experienced by it. It does not happen anymore now, and the records could be harvested without this error.

- **Known Issues**
  - There are some bugs inside the **open source Marc4j source code** which sometimes does not allow proper identification of the location of the error in the raw MARC files. Efforts are being put in so that we could participate in the open source changes and then commit it back to the author. This would ensure less effort from the librarian in locating the error (http://code.google.com/p/xcoaitoolkit/issues/detail?id=1).
  - Performing convert and load as one step results in a different directory structure than when performing convert and load as separate steps (http://code.google.com/p/xcoaitoolkit/issues/detail?id=2).

        - Allowing tier architecture of directories for the source folders when the convert process reads the raw-MARC records. This would ensure even files within the sub-folders inside the source folder having raw-MARC during convert and MARCXML during load process would be taken by the OAI Toolkit. It doesn't support this now (http://code.google.com/p/xcoaitoolkit/issues/detail?id=4).

        - The OAI Toolkit is not yet capable of processing item data; we are working on gathering information from partners on how item data is available, and what item data is available, before we implement this feature (http://code.google.com/p/xcoaitoolkit/issues/detail?id=7).

        - The OAI Toolkit will store the URL of the corresponding NCIP Toolkit, but this feature is not needed at the moment (http://code.google.com/p/xcoaitoolkit/issues/detail?id=6).

        - There has been a request to make documentation available in HTML; this is being reviewed to determine how documents should be formatted when put into HTML (http://code.google.com/p/xcoaitoolkit/issues/detail?id=8).

        - The ability to import Dublin Core records may be added to a future release (http://code.google.com/p/xcoaitoolkit/issues/detail?id=5).

        - When a client harvests large collections, the Lucene response time gets progressively longer. For example, when harvesting the University of Rochester's 6.5 million MARC records, a client can harvest 1.3 million records in the first hour of the harvest but only 450,000 records in the last hour (this test was run on a desktop machine, and not on a server) (http://code.google.com/p/xcoaitoolkit/issues/detail?id=3).

        - The Proper validation of the URL parameters of the OAI Server to ensure appropriate message display bug was reported and needs to be solved yet (http://code.google.com/p/xcoaitoolkit/issues/detail?id=10).

        - Right now the OAI Toolkit uses Struts version 1, but at some point we would like to shift it to Struts version 2. Although it is not impacting much at this moment. (http://code.google.com/p/xcoaitoolkit/issues/detail?id=9).

        - Processing files of records as though they were deleted. This feature request would then have the records no longer available for further discovery in the ILS. (http://code.google.com/p/xcoaitoolkit/issues/detail?id=12).

        - Feature Request for creating an embedded test set and writing new OAI verbs to handle the harvesting of that test set. (http://code.google.com/p/xcoaitoolkit/issues/detail?id=11).

- Release 0.4 (December 23, 2008)

    - **What's new in this release**

- We have upgraded to the newest version of Marc4j (version 2.4). With this update, the toolkit can skip records with problematic headers and continue processing with the next record (as opposed to crashing).

- **We introduced a new step to modify MARC records in between the convert and load steps.** This step transforms the MARCXML files with XSLT stylesheets. We created some sample stylesheets that you may use,or you may create your own XSLT files. You can also apply multiple files. The order or transformation of how multiple stylesheets is processed will be the same as the order in which they are listed in the command line parameter.

  Usage:

  ```
  -modify my_stylesheet.xslt
  ```

  or if you want to transform with multiple files, you should list files in quotation mark, separated with space:

  ```
  -modify "one.xslt two.xslt three.xslt"
  ```

  You can find our sample stylesheets in the *xslt* directory. **You may skip the modify step if you do not want to use it.**

- **In the interest of conserving time, we have added a `–production` flag.** If your process of exporting from your ILS and importing to the OAI repository is stable, you don't need the temporary MARCXML files. With `–production` flag you can chain the different steps in memory. The process will be quicker, because you don't need to read and write multiple files from and into the disk. This flag is effective only if you chain any two steps.

  Sample:

  ```
  -convert –modify my_stylesheet.xslt –load -production
  ```

- **In the interest of conserving disk space, we have added a `–delete` flag.** If you want to delete temporary files you can use the `–delete` flag.

- **By default, the program now automatically splits large files of MARC records into separate files of 10,000 records each.** Previously, this would have been accomplished by using the `–split_size` parameter with the number 10000. If you want to modify this value, use the `–split_size <size>` parameter. If you do not want to split up your MARCXML files, use a negative number or zero as the value associated with the parameter.

- **To make the toolkit work in a consortial environment, we have changed the default behavior to require that both the 001 and 003 (repository code) fields of each record be filled in.** This allows records from multiple sources to be pulled into the OAI Toolkit, as the toolkit will now identify records as unique using the 001 in conjunction with the 003 (to avoid collisions when using just the 001 field). This does mean that you must have the 003 field defined accurately (for example one 003 code for each school). **If you do not want to use the 003 field in identifying records and wish to continue using the OAI Toolkit as you previously have been, use the parameter `–ignore_repository_code`.** Further, if you do not have the 003 field filled in, the toolkit can fill it in for you when you import MARC records. **To fill out the 003 field in each record with a specific code, use the following:**

```
-default_repository_code <code>
```

  - We removed the version numbers from the directory names inside of the zip files. If you aren't sure what version of the software you have, the release notes will now be packaged in the downloads.
  - **XML 1.1 is no longer available.** To create this feature, we had to modify Marc4j, and because we have upgraded Marc4j, that modification is no longer in the OAI Toolkit. If you would like to use XML 1.1, please create a feature request on the XC website for it.

- **Bug Fixes**
  - "create_user.sh doesn't work in linux" bug reported by Chris Delis from CARLI has been fixed using the fix that Chris sent to us.
  - "-log not interpreted in importer" bug is now fixed so that logs are actually sent to the location specified by the parameter value.
  - The `–replace_weird_characters` flag did not replace weird characters in the indicators of data fields; it now does.
  - If a record crashed the processing of a MARC file, and the problem was in the Leader, the OAI Toolkit did not report the location of the problem. Now, Marc4j skip these records, and they do not cause the OAI Toolkit to crash. Since we do not know the control number of those files, the OAI Toolkit logs the last successfully read record's control number, which should help you to find the bad records. We are currently discussing with partners what the proper solution for dealing with the bad data is.
  - "Error MARCXML files saved are not well-formed" bug is fixed by the upgrade of Marc4j.
  - "Timeout on last OAI Request" bug was a problem with the last OAI request automatically sending a new request without checking that there are any more record available. This does not occur any more.
  - "No action instance for path/oai-request could be created" bug (reported by Notre Dame) is now fixed.
- **Known Issues**
  - The OAI Toolkit is not yet capable of processing item data; we are working on gathering information from partners on how item data is available, and what item data is available, before we implement this feature .
  - The OAI Toolkit will store the URL of the corresponding NCIP Toolkit, but this feature is not needed at the moment.

        - There has been a request to make documentation available in HTML; this is being reviewed to determine how documents should be formatted when put into HTML.
        - The ability to import Dublin Core records may be added to a future release.
        - Currently, the XC website does not facilitate downloading old versions of the software
        - It has been requested that the “replace bad characters” feature differentiate between characters in the Leader and in the fields
        - When a client harvests large collections, the Lucene response time gets progressively longer. For example, when harvesting the University of Rochester’s 6.5 million MARC records, a client can harvest 1.3 million records in the first hour of the harvest but only 450,000 records in the last hour (this test was run on a desktop machine, and not on a server).

- Release 0.3 (November 12, 2008)
    - **What’s new in this release**
        - This release contains one bug fix, as detailed below. Also, the downloads for this release accidentally replaced the downloads for the 0.2 release shortly after 0.2 was released. Therefore, some people who thought they were downloading the 0.2 release may have actually downloaded this release.
        - If you experienced the bug that is fixed by this release, you should upgrade to 0.3. Otherwise, it is acceptable to continue using version 0.2.
        - Appendix II of the OAI Toolkit Manual has been updated to reflect changes in how we are exporting data from Voyager. These changes are minor and should not affect our partners, but may be useful to review if you are using Voyager. Changes are in Appendix II: ILS Connector section, pages 39 – 42 (XcExport script section). Additionally, a reminder that you must create the cache directory manually if you are using cacheDir has been added to the manual (page 30).
    - **Bug Fixes**
        - IndexOutOfBoundsException issue reported by CARLI is fixed. This problem was caused by subfield separator (’$’) character in content of subfield.
    - **Known Issues**
        - Currently, all known issues are scheduled to be addressed in the 0.4 release of the OAI Toolkit.
        - University of Buffalo reported a problem with the OAI Toolkit failing during the convert process. We believe this bug is specific to the MARC records being used by UB and has not been reported by any other institution.
        - Notre Dame reported a problem with the server piece of the OAI Toolkit where any link from the starting page results in an error message. There is a possibility that this has also been fixed in this release, but that is presently unknown. This has not been reported by any other institution.

        - University of Rochester reported a problem with the OAI Toolkit ceasing to respond when the resumption token ends in "|2360000". This has not been reported by any other institution.

    - **Important note about cache**

        - In the manual and the release notes for release 0.2, we did not explicitly write that the user should create a directory for cache (it is referenced as *cacheDir* in manual). In the future, the OAI Toolkit will create that directory if it does not exist, but in this release you should create it manually before launching Tomcat.

- Release 0.2 (October 9, 2008)

    - **What's new in this release**

        - Simplified XML processing on the server side. Instead of using high-level XML manipulation library (jDOM), we use simple character and regular expression base replacements. This drastically improved the speed of response creation.

        - Predictive caching : Since the OAI harvesting is a sequential process, the server can predict the next most probable request parameters based upon the resumption token. When the response is starting to be downloaded by the client, the toolkit starts to collect the response to the next predicatable request. When ready, it saves the response to a file and then when the client issues the new request it can be served from the cache file. The administrator should setup the cache directory by the *cacheDir* parameter in the *<Tomcat>/bin/OAIToolkit.directory.properties* file, and a maximal cache file lifetime (in minutes) in the configuration page.

        - The log messages concerning the MARC data errors are clearer for librarians. There is a new class which translates the messages that come from the MARCXML schema based validation process of MARCXML records to those terms, which are in the MARC standards (tag, subfield, indicator etc.)

        - Three configurable options to handle non-standard, "weird" characters in MARC records as feature request have been created:

            1. create XML 1.1 with -xml_version_11 command line parameter and leave the characters as they are

            2. replace weird characters with zero (at the Leader) and spaces (elsewhere) with -replace_weird_characters command line parameter

            3. filter out these records (default behavior)

        - All these new features are detailed in the updated OAI Toolkit Manual.

    - **Bug/error fixes**

        - The null resumptionToken issue is fixed. This problem occurred at the first usage of OAIToolkit.

        - The *luceneDir* and the storageType configuration settings were missing from the sample configuration files, and they caused null pointer exceptions.

- Release 0.1 (September 8, 2008)

- Initial release