

# Fred 2.0

## Phase I: Library Of Congress Authorities Files

*Open Catalog Liberation Council, Provisional ALA*

*22nd December 2006*

- **i Dedication**

Fred 2.0 is dedicated to the memory of  
Frederick G. Kilgour (Jan. 6, 1914 - July 31, 2006)  
*Distinguished Research Professor Emeritus  
School Of Information And Library Science  
University of North Carolina at Chapel Hill*

This phase of the project is dedicated to the men and women at the Library of Congress and outside, who have worked for the past 108 years to build these authorities, often in the face of technology seemingly designed to make the task as difficult as possible.

Save The Time Of The Cataloger

- **ii Summary**

Using a custom agent, we were able to harvest 6.95 million authority records using the publicly accessible interface to the Library of Congress authority files located at [authorities.loc.gov](http://authorities.loc.gov).

Retrieved records have been converted into MarcXML

Accented characters have been converted into NFC (Composed Normal Form).

Initial checks against [authorities.loc.gov](http://authorities.loc.gov) indicate that the retrieved data faithfully reflect that on the original system; however these checks are still only preliminary.

Cross checks against Classification Web have revealed some inconsistencies. For this reason, we are releasing this data for research purposes only. This data is **not** suitable for production use.

- **iii COPYRIGHT NOTICE**

These data are works of the United States Government and as such are not subject to copyright within the United States. (17 U.S.C §105).

The Library of Congress has copyrighted these data for use outside the United States. Contact the LC for permission prior to use or distribution of this data outside the United States. [www.loc.gov—mds.html](http://www.loc.gov—mds.html)

These data were obtained via the public web interface to the authorities files located at [authorities.loc.gov](http://authorities.loc.gov).

“All authority information in Library of Congress Authorities is available free of charge via this Web site ([authorities.loc.gov](http://authorities.loc.gov)). Users do not have to register or request permission to search, save, print, or email the LC authority records. The only limitation is that authority records may only be saved, printed or emailed one at a time.”[authorities.loc.gov—auth-faq.htm](http://authorities.loc.gov/auth-faq.htm):

- **iv WARNING NOTICE**

These data have several known inconsistencies with the reference source. They are suitable for research purposes only.

These data **MUST NOT** be used to prepare records for any Cooperative Cataloging Program unless verified against another source of data.

These data are presented as is and without any warranty, expressed or implied. Any use is at your own risk

- **v Methodology: Harvesting**

The data were harvest via web interface to voyager, using a custom written pipelined HTTP agent.

This approach allows requests to be made over wide area networks with minimal delays due to latency, and with greatly reduced load upon both client and server.

Requests were generated for sequential blocks of authority records. These requests were synthesized using standard techniques for constructing Voyager URLs. Generating requests programmatically allowed for greatly improved pipelining, and removed the agent from the scope of any robots.txt restrictions.

“A robot is a program that automatically traverses the Web’s hypertext structure by retrieving a document, and recursively retrieving all documents that are referenced.” [www.robotstxt.org—faq.html](http://www.robotstxt.org/faq.html)

MarcXML was recovered directly from the displayed XHTML. Using this avoids generating a request to save, print or email a marc21 encoded record, halving the number of requests processed by the server, more than doubling throughput.

Minimal processing was performed at this stage:

- The leader was extracted from the 000 pseudo-field, and extended to 24 characters by appending ‘o’

- Non-XML character entities were expanded (&rsquo; and &quot;)

- No character re-encoding was performed at this stage; composing diacritics were left as is.

- **vi Methodology: Operations**

To avoid impacting interactive users, the agent was deployed only during the late night hours and over weekends.

Monitoring was performed using the Java Management Extensions (JMX) `java.sun.com—javamanagement`. Metrics were created to monitor throughput and response time. Adjustments were made in an effort to keep response times within a few hundred milliseconds of the unloaded response time.

During the final stages of operations it was discovered that certain records were could not be retrieved, and that reference to these records appeared to cause sessions with Voyager to become unresponsive. Once this became apparent, we attempted to exclude those records from retrieval.

- **vii Analysis: Database Structure**

For purposes of analysis, the records were loaded into a relation database

Four tables were created.

The `marc_xml` table contains the unprocessed MarcXML, keyed by `001` record number. This table contains entries for all authority records retrieved.

The `MarcRecord` table contains the control fields, along with the `1XX` heading field associated with the record. Records containing more than one heading field were discarded at this stage of the processing. This table is also keyed by the `001` record number. This table contains entries for all authority records retrieved.

The `MarcField` table contains one entry for each field, keyed by `001` record id, and index of the field within the record. This table contains entries for all authority records retrieved.

The `MarcSubfield` table contains one entry for every subfield, keyed by `001` record id, field index, and subfield index within the record. This table was only populated for records with tags greater than `140`.

- **viii Analysis: Validation Checks**

Records were selected at semi-randomly, and compared to those on the `authorities.loc.gov` system. The records retrieved matched those chosen. These tests were purely a cursory check looking for systemic flaws in the transfer.