Results of *L'Année philologique online* OpenURL Quality Investigation

Mellon Planning Grant Final Report
February 2009

Adam Chandler
Cornell University

Note: This document is a subset of a report sent to Mellon on work I did with Professor Eric Rebillard and David Ruddy. (See http://cwkb.org for more detail the other parts of our grant.)

1. Restatement of original problem statement

Since 2004 *L'Année philologique online* has been Open-URL compliant. Each record contains a link which can be processed by a link resolver. When the patron clicks on the link, the OpenURL carries the data about the item to the link resolver of the library. The resolver compares the data with what is held in the library's collection and presents the available options in a results page. For a book, there is a link to the library's catalog card; for an article, ideally this is a link directly to the full-text of the article.

However, many OpenURL links fail. One strategy would be to develop linking partnership with online journals in order to log and monitor the links. However, rather than developing bi-lateral, direct linking to resources, we believe it is more cost effective to invest in metadata enhancement work that will improve OpenURL success rates. OpenURL was developed to get beyond the problems of bi-lateral linking: the costs associated with the implementation and long-term maintenance of these links, and unreliable access for off-site users, an ever growing problem for libraries.

Adam Chandler, Database Management and Electronic Resources Librarian, Library Technical Services, Cornell University Library, developed initial recommendations for

metadata improvement based on a manual review of a sample set of 126 OpenURLs generated by *L'Annee*. This report identified many typical metadata problems that cause OpenURLs to fail: malformed dates, volume and issue numbers combined into one field, reliance on the pages element instead of the start page element for linking, lack of identifiers, etc. Such a report is extremely useful to L'Annee because it precisely identifies the critical failure points where improvement efforts can most profitably be focused. However, performing such a manual review of all the OpenURLs generated by *L'Annee*, or any other vendor, would be prohibitively expensive and time consuming.

We propose exploring the feasibility of developing a fully automated OpenURL evaluation process. Such a system would accept OpenURLs and return scores based on a set of evaluation metrics. These scores would allow resource providers to see precisely where their OpenURLs were weakest, letting them target metadata improvement efforts in the most cost-effective manner.

We ultimately envision a community recognized index for measuring the quality of OpenURL links from content providers. An OpenURL quality investigation with L'Annee would serve as a proof of concept of this idea, providing a common reference point from which to begin a wider conversation about OpenURL quality among librarians, publishers, NISO, and OCLC (the OpenURL Maintenance Agency). Our work would build on similar efforts by Baden Hughes, who developed metrics to measure the quality of Dublin Core metadata submitted to the Open Language Archives Community ["Metadata Quality Evaluation: Experience from the Open Language Archives Community." 7th International Conference on Asian Digital Libraries, ICADL 2004, Shanghai, China, December 13-17, 2004. Proceedings, pp. 320-329].

2. Findings

After considerable analysis and experimentation with different ways of creating metrics, the conclusion is that our original assumption about how to measure quality was only partially correct. We started with the assumption that we could give each element in the incoming OpenURL a pass or fail rating, then attribute points to each of the elements. The total of the points would be the score for the OpenURL. There are several problems with this approach:

1. It is impossible to give a pass/fail grade for many if not most of the elements in an OpenURL. For example, *atitle* and *title*, two of the most common elements. Titles are endlessly variable. How can we say if a particular pattern is correct?
2. It does not provide OpenURL providers with information about the particular data patterns that are present in their OpenURLs.
3. It does not effectively take into account the elements used by link resolvers to link to full text providers. Each full text provider requires a separate proprietary URL when linking to their content, sometimes referred to as the "link-to" syntax. Without such a weighting of OpenURL element importance, providers do not know where to get the most value for their metadata clean-up buck.

With an eye towards building them into a slightly altered feedback and evaluation model for OpenURL quality, our response is to create three new constant values that can then be applied against the OpenURLs sent from an individual provider such as *L'Année philologique online*.

1. The set of elements used most frequently by full text content providers when attempting to link into their sites from library link resolvers, based on a review of over 200 link-to syntaxes configured in the Cornell link resolver. From this analysis we learned that the following elements are required or recommended most frequently: title, spage, volume, issue, date, aulast, issn, atitle, DOI. (Cornell uses the Web Bridge link resolver product from Innovative Interfaces, Inc.)
2. The frequency of use of particular OpenURL elements contained in the OpenURLs sent to the Cornell link resolver over six months (July – December 2008).
3. The particular string patterns within elements that appear most often in incoming OpenURLs.

This revised approach requires the three constants above in order to generate two different reports for an OpenURL provider, one for element frequency by genre (e.g., article), and a second report that lists string patterns present within element values. We turn now to the details of our two reports.

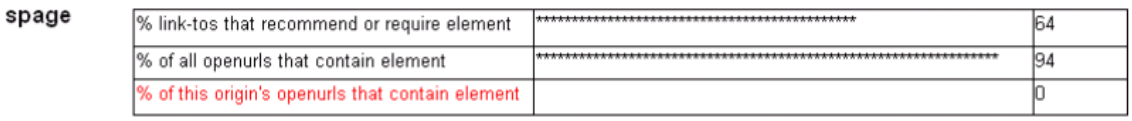
Genre Element Report

The genre report establishes a benchmark so an OpenURL provider can compare the frequency of elements in their OpenURLs against those of their peers. Based on our six months of data from the Cornell link resolver, the most frequently used genre type exchange is article, so in our prototype we have developed that article genre report first.

article	102465
book	2243
bookitem	325
conference	56
issue	9
journal	1617
proceeding	1249
report	3
unknown	2946
TOTAL	110913

Table: Genre frequency in OpenURLs sent to the Cornell link resolver, July – December 2008.

To illustrate how this metric works, we will now look at one element within the report in more depth, the OpenURL element called spage (or rft.spage in OpenURL 1.0). What do the numbers mean in the spage figure below?



spage			
	% link-tos that recommend or require element	**********	64
	% of all openurls that contain element	**********	94
	% of this origin's openurls that contain element		0

Figure: spage element use report for *L'Année philologique online*

% link-tos that recommend or require element:

In our analysis of more than 200 Cornell link resolver link-to syntaxes, 64% of them recommend or require that the OpenURL spage element be included. Here is an example of a link-to syntax that uses the spage element. This is how we link to BioOne.

http://www.bioone.org/perlserv/?request=get-document&issn=#@ISSN#&volume=#@VOLUME#&issue=#@ISSUE#&page=#@SPAGE#

The link-to syntax can be parsed into the following components:

host: http://www.bioone.org/perlserv/?request=get-document
issn: &issn=#@ISSN#
volume: &volume=#@VOLUME#
issue: &issue=#@ISSUE#
spage: &page=#@SPAGE#

In other words, BioOne needs the spage element to link to the correct article within their corpus.

% of all openurls that contain element:

In our analysis of all the OpenURLs sent to the Cornell link resolver over a six month period, we found that 94% of them in the article genre include the spage element. Aside from title, this is the most common element we found in incoming OpenURLs.

% of this origin's openurls that contain element:

Origin here means one particular OpenURL provider. In our analysis of the OpenURLs from *L'Année philologique online* 0% contain the spage element. In our prototype user interface we have highlighted this in red because it is problematic. The combination of the two constants (% link-tos, and % all OpenURLs that use the spage element) indicates clearly that OpenURLs sent from *L'Année philologique online* have a lower chance of

being resolved to full text than peers. Most full text providers (64%) need the spage element to successfully resolve the incoming request. Given that 94% of all OpenURL senders include it, it is reasonable therefore to expect that *L'Année philologique online* could, through some effort, also include it.

A complete genre element report is included as a screenshot in the prototype, below.

Element Patterns Report

The patterns report is based on the same six month data set. In contrast to the genre report's emphasis on element frequency, however, this report digs deeper into the string patterns found within individual element values. For each OpenURL element included in the element patterns report we create a handful of regular expressions for identifying the particular string patterns present in the data. Take for example the volume element pattern report, included below.

element	pattern name	regular expression	% from origin with this pattern	% from all origins with this pattern
volume	ROMAN	/^[IVXLCDM]+$/	30	0
	ROMAN-ROMAN	/^[IVXLCDM]+-[IVXLCDM]+$/	20	0
	NUMBER	/^\d+$/	20	99
	NUMBER-NUMBER	/^\d+-\d+$/	10	0
	NUMBER (NUMBER-NUMBER)	/^\d+ \(\d+(-\d+)*\)$/	10	0
	other		10	1

Figure: volume element string pattern report for *L'Année philologique online*

The volume field has six pattern categories, five regular expression matches plus "other" (for everything that does not match one of the five regular expressions). In the case of *L'Année philologique online*, 50% (30+20) of the instances of that field contain roman numerals. This is in sharp contrast to all OpenURL providers: in our sample data set, none of the volume elements contained roman numerals. In fact, 99% of them contain only a number. The value of this kind of report is it tells the OpenURL provider, in this case, *L'Année philologique online*, that their use of the volume field is different than all others sampled. One implication of this is that link resolvers are probably optimized for handling Arabic numerals when they are in the volume field, not Roman numerals, although their use is legal in the element, which could, and probably does, lead to errors when they are passed along in a link-to syntax at the full text content provider end of the process.

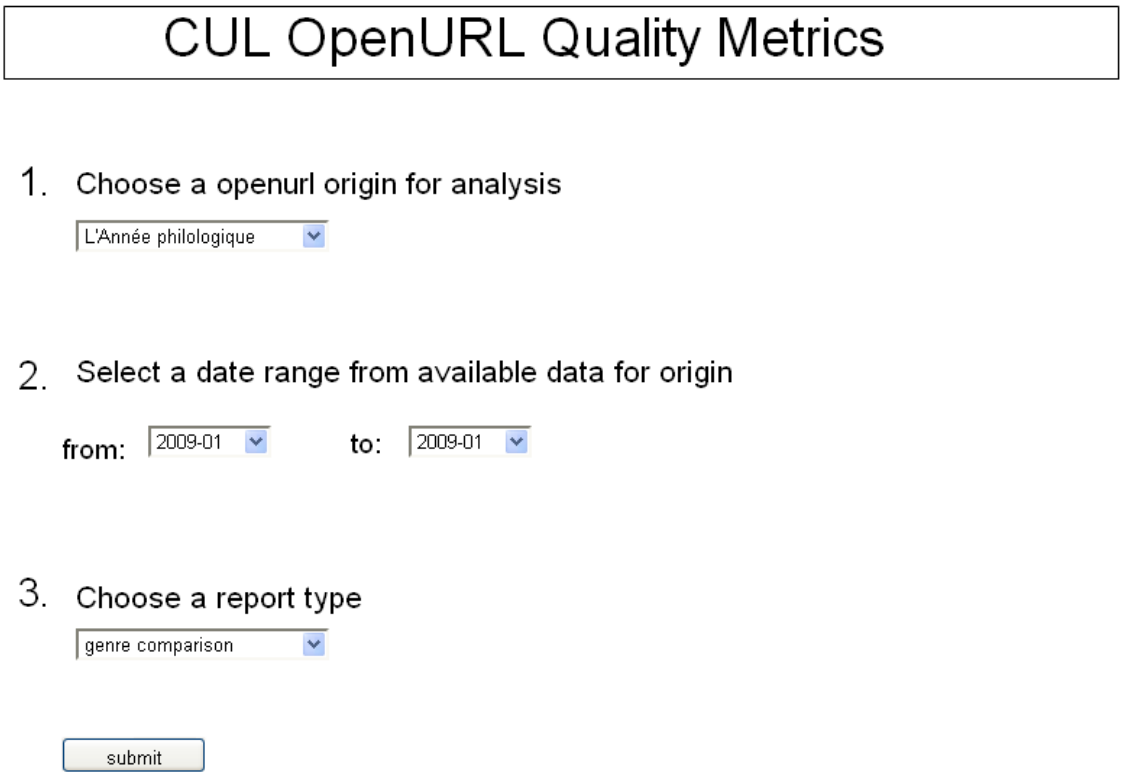
Implications

We believe we now have an easy to comprehend, scalable OpenURL quality model. Building on our work here we could easily extend it to include reports for all the genres

and elements. We believe implementation of a such a system as a stand alone service would have significant value to libraries, OpenURL providers, link resolver vendors, full text content providers, and most important of all, library patrons, by filling a critical gap in the OpenURL protocol: objective, empirical, transparent feedback for supply chain participants.

Prototype User Interface

The screenshots below illustrate the user interface for this service.



Report: Genres

CUL OpenURL Quality Metrics

Genre Comparison Report for **L'Année philologique**

genre = article

title

% link-tos that recommend or require element	**********	64
% of all openurls that contain element	**********	97
% of this origin's openurls that contain element	**********	100

spage

% link-tos that recommend or require element	**********	64
% of all openurls that contain element	**********	94
% of this origin's openurls that contain element		0

volume

% link-tos that recommend or require element	**********	61
% of all openurls that contain element	**********	90
% of this origin's openurls that contain element	**********	95

issue

% link-tos that recommend or require element	**********	60
% of all openurls that contain element	**********	86
% of this origin's openurls that contain element		0

date

% link-tos that recommend or require element	**********	48
% of all openurls that contain element	**********	95
% of this origins openurls that contain element	**********	95

aulast

% link-tos that recommend or require element	**********	47
% of all openurls that contain element	**********	93
% of this origins openurls that contain element	**********	95

issn

% link-tos that recommend or require element	**********	35
% of all openurls that contain element	**********	66
% of this origins openurls that contain element		0

atitle

% link-tos that recommend or require element	**********	35
% of all openurls that contain element	**********	98
% of this origins openurls that contain element	**********	80

DOI

% link-tos that recommend or require element	**********	14
% of all openurls that contain element	**********	17
% of this origin's openurls that contain element		0

about contact

Report: Element Patterns

CUL OpenURL Quality Metrics

Element Patterns Report for **L'Année philologique**

element	pattern name	regular expression	% from origin with this pattern	% from all origins with this pattern
aulast	SimpleLastName	/^[A-Za-z]+$/	25	95
	SimpleLastNamePlusInitialS	/^[A-Z][a-z]+([A-Z]\.)+$/	40	0
	other		35	5
date	YYYY	/^\d{4}$/	40	86
	YYYY-YYYY	^\d{4}-\d{4}$	40	1
	other		20	13
title	words without punctuation	/^[A-Za-z]+$/	30	9
	other		70	91
atitle	words without punctuation	/^[A-Za-z]+$/	20	26
	other		80	74
volume	ROMAN	/^[IVXLCDM]+$/	30	0
	ROMAN-ROMAN	/^[IVXLCDM]+-[IVXLCDM]+$/	20	0
	NUMBER	/^\d+$/	20	99
	NUMBER-NUMBER	/^\d+-\d+$/	10	0
	NUMBER (NUMBER-NUMBER)	/^\d+ \(\d+(-\d+)*\)$/	10	0
	other		10	1
spage	NUMBER	/^\d+$/	0	97
	NUMBER-NUMBER	/^\d+-\d+$/	0	1
	STRING WITH NUMBER	/[A-Za-z].+\d/	0	2
	other		0	0
issn	NUMBER-NUMBER	/^\d+-\d+$/	0	90
	NUMBER	/^\d+$/	0	0
	other			10
issue	ISSUE IS NUMBER	/^\d+$/	0	97
	ISSUE HAS STRING	/[A-Za-z].+/	0	2
	other		0	1

Figure: OpenURL Quality Relational Data Model