6-2009

Ingesting TEI metadata into Encore at the University of Nebraska-Lincoln: TEI/Encore Task Force Report--University Libraries, June 2009

Charity Martin
University of Nebraska-Lincoln, charity.martin@library.tamu.edu

Stacy Rickel
University of Nebraska-Lincoln, srickel1@unl.edu

Laura Weakly
University of Nebraska-Lincoln, lweakly2@unl.edu

Elaine L. Westbrooks
University of Nebraska-Lincoln, elainelw@email.unc.edu

Ingesting TEI metadata into Encore at the University of Nebraska-Lincoln

TEI/Encore Task Force Report – University Libraries
June 2009

Charity Martin {cmartin3@unl.edu}, Stacy Rickel {srickel1@unl.edu}, Laura Weakly {weakly2@unl.edu}, and Elaine L. Westbrooks (editor) {ewestbrooks2@unl.edu}

ABSTRACT

In January 2009, a library task force was formed to inform the digital asset management group about which dates from a TEI record should be used in an Encore record. While investigating this issue, the taskforce encountered a host of other issues that they did not anticipate but were addressed in this report.

This report documents the key problems that emerged as a result of the *Cather* and *Lewis and Clark* harvesting.[1] Before and after metadata is ingested quality control is necessary. Responsibility for quality control rests with all library departments. The key recommendations for Encore quality control are:

1. Establish criteria to determine the level that records (project, collections, and sub-collections) should be harvested for each project in Encore.
2. CDRH and TS should establish a policy that will articulate what CDRH projects should be cataloged.
3. TS should review the use of format codes (e.g. PRINT MATERIAL, COMPUTER FILE, THESIS/DISSERT) in the catalog code records.
4. We do not recommend that the TEI encoding be edited or enhanced so that it will be harvested different in Encore.
5. Going forward, all project teams working with TEI should be aware of OAI harvesting as well as Encore aggregation. Furthermore, project teams should attempt to create metadata that would aggregate according to fields indexed by Encore—without compromising the integrity of the metadata (TEI or otherwise) or the project
6. Utilize Encore's community tagging mechanism

The TEI to DC Crosswalk is included but it is constantly changing. The crosswalk should not be cited because it is customized for UNL's projects only. Questions about the crosswalk should be directed to Laura Weekly in the Center for Digital Humanities Research.

[1] *Birds of Nebraska* <http://libtextcenter.unl.edu/birds_of_nebraska/> was the third project harvested in Encore but it did not present major challenges. This is due to the fact that the metadata are consistent and far less complex than the items that make up *Lewis and Clark* and *Cather.*

Introduction

In January 2009, a library task force consisting of Stacy Rickel, Laura Weakly, and Charity Martin was formed at the request of Dee Ann Allison to inform the digital asset management group about which dates should be used in an Encore record. During the task force's investigation, the following 2 questions became evident:

- How should the library map TEI to the OAI-compliant Dublin Core (DC) fields for OAI/Encore?
- What TEI elements should be harvested for OAI/Encore?

While investigating these 2 questions, the taskforce encountered a host of issues that they did not anticipate:

- Displaying both harvested MARC and OAI records for the same item in Encore can be confusing and redundant.
- It is not clear at what level the CDRH Collections/projects and the items within those collections should be cataloged. Currently the MARC records describing CDRH collections/projects are inconsistently held in the catalog as well as Encore.
- Now that we know that our digital content will be mapped to OAI and harvested by Encore, how does that knowledge inform the way we encode dynamic CDRH collections/projects in the future?
- What kind of workflow is necessary to ensure that new CDRH collections (which vary) are harvested, correctly and consistently in Encore? What policies should be in place to handle projects that heavily use SQL? Such projects require a different workflow than those that are solely TEI or EAD.
- In *The Willa Cather Archive* and *Journals of the Lewis and Clark Expedition*, what should be done, if anything, to improve the metadata (mapped from TEI to DC) so that Encore users can retrieve relevant and meaningful records?
- Information is not properly coded in the MARC records that are being harvested into Encore from the catalog. Encore faceting expose cataloging code flaws that typically are hidden in the catalog.
- There is little to no quality control of metadata records once they are harvested. Who is responsible for the quality control and who should decide how to implement change?

In order to address the initial questions, the taskforce was forced to investigate seven additional issues. Indeed the charge proved to be more complicated and nuanced than initially thought.

Goals and Objectives

The goal of this report is to:

1. Share the metadata crosswalk (see appendix)
2. Highlight the key problems that emerged as a result of the *Cather* and *Lewis and Clark* harvesting[2]
3. Recommend how to improve the quality of metadata once it is harvested in Encore i.e. post-harvest quality control.
4. Recommend what to change on the TEI to DC mapping to improve metadata when it is being harvested i.e. pre-harvest quality control.

[2] *Birds of Nebraska* <http://libtextcenter.unl.edu/birds_of_nebraska/> was the third project harvested in Encore but it did not present major challenges. This is due to the fact that the metadata are consistent and far less complex than the items that make up *Lewis and Clark* and *Cather.*

High quality metadata will expose users to diverse content that is relevant, while larger quantities of metadata will expose the depth and breadth of our collections (images, databases, books, journals, finding aids) that they might not otherwise encounter outside of Encore.

Background

The Willa Cather Archive and the Journals of Lewis and Clark Expedition

The first project indexed and aggregated in Encore was the *Willa Cather Archive* <http://cather.unl.edu/> which consists of "Writings," "Letters," Images, Multimedia, and other content. Encore is harvesting approximately 2600 images from ContentDM (Willa Cather Images Gallery), 30 items from the catalog, 2 items from the Digital Commons, and 35 texts from *Cather*. The MARC record for *Cather* has been harvested from the catalog as well as records for a 'sub-collection' of the "Journalism Archive" known as "*Amusements*" which is a column authored by Cather in the *Nebraska State Journal*. *Cather* is dynamic and content continues to be added. In the fall of 2008, Stacy Rickel worked with Andrew Jewell to harvest metadata from the *Cather Archive*. The biggest challenge is the fact that the same content is harvested twice; once from the catalog (MARC) and the other from the online collection (TEI).

The second project indexed and aggregated in Encore was *The Journals of the Lewis and Clark Expedition* http://lewisandclarkjournals.unl.edu/index.html which consists of "Journals," "Maps," "Images," "Multimedia," and other content. There are sub-collections within *Lewis and Clark* that are also harvested. Rickel worked with Laura Weakly to harvest metadata for Encore. Encore is harvesting approximately 1500 texts and audio from *Lewis and Clark*, 3 from Digital Commons, and 20 from the catalog. The biggest challenge is the fact that metadata is frequently being pulled from the wrong part of the TEI record. The result is different content that is harvested may produce the very same metadata; therefore the content cannot be distinguished until one clicks "Go to this Record" link in Encore. The TEI metadata is automatically mapped into DC so additional tweaking of the mapping to make the DC metadata more descriptive and meaningful to Encore users would be the best option.

In the case of *Lewis and Clark* and any other TEI-based projects, the encoding is complex and someone must decide when to describe an entire collection or when to describe a single image or article. **This complexity is difficult, and sometimes impossible to map to a DC record.** Given the thousands of records harvested thus far, it is amazing that there are so few errors.

Encore

Encore is a metadata aggregator,[3] which also includes the Metadata Harvesting Protocol of the Open Archives initiative (OAI-PMH).[4] In order for any metadata (TEI, EAD, MARC) to be ingested by Encore it must be OAI compliant. The most commonly used metadata schema harvested by OAI is Dublin Core. In addition, the metadata must be well formed according to Innovative's (III) indexing standards for Encore.

In order to automate the creation of the DC records from the *Cather* and *Lewis and Clark* projects, Rickel used a combination of PERL scripts and eXtensible Style sheet Language Transformation (XSLT) to extract

[3] A Metadata aggregator is a tool that gathers heterogeneous metadata (e.g. Dublin Core, MARC, TEI, EAD) from a variety of sources (Website, Digital Collection, Catalog, Finding Aid) and presents it in one place. The benefits of such a tool include providing better access to digital collections that might not otherwise be promoted and exposed through their native channels. However, the benefits can be mitigated by inconsistent and weak metadata.

[4] See The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) http://www.openarchives.org/pmh/

metadata from the relevant TEI tags of the eXtensible Markup Language (XML) documents, and map those tags to the DC fields required by OAI-PMH.

Most of the scholarly websites created and maintained by CDRH use the Text Encoding Initiative (TEI) for the encoding of documents.[5] The TEI metadata was mapped to the DC metadata schema so that it can be harvested by the OAI-PMH. A TEI header, which contains 4 sections, can contain hundreds of elements/tags; most of which are optional. On the other hand, simple DC contains only 15 elements; all elements are optional and repeatable.

Harvesting TEI Metadata in OAI/Encore

A number of problems have emerged because the rich metadata of *Cather* and *Lewis and Clark* was never intended to be 'reduced' to a simple metadata schema such as DC (for OAI) and aggregated within a tool such as Encore. In other words, data are lost when TEI metadata is expressed as DC metadata. In addition, Encore aggregation brings a variety of disparate metadata records together which are out of context. OAI works well with metadata that is simple and describes one item; one metadata record for 1 item. In contrast, TEI metadata is rich and is designed to describe the physical item as well as the digital item in different elements of one metadata record. So in many respects, it is not always possible to have 1 metadata record for 1 item. Although these problems have been well document in the literature, they are not prohibitive.[6]

The Encore User Experience

The taskforce investigated how users may discover 'records' from *Cather* and *Lewis and Clark* in Encore. Although each project has its unique challenges, in both cases, the retrieval of relevant results was difficult at times. In fact, the results may be confusing and not meaningful to users.

The mapping of the TEI to DC is automated so we can expect a number or errors that are unavoidable. The metadata may be valid but inaccurate. Once metadata is harvested in Encore, our job is not finished. Ongoing quality control of Encore metadata is necessary. The first three problems that are articulated in this report are largely due to the fact that there is no workflow or policy for quality control.

Problem #1 - "Redundant" Records

Encore is harvesting MARC records for *the Willa Cather Journalism Archive,* a subset of *Cather*, from the catalog (MARC) as well as the project online (TEI). For example one can discover Cather's *Amusements* articles published in the *Nebraska State Journal*. However, as demonstrated in Figure 1 and 2 below, the user will retrieve 2 different records; one is harvested from the catalog (MARC) and the other is harvested from TEI (DC). The records harvested from the catalog are displayed first and one can gain access to the item on the website by clicking "Willa Cather Journalism Archive" (Figure 1). For the same records harvested from TEI, the user can gain access on the website by clicking "Go to this Record" (Figure 2).

[5] TEI was founded in 1987 and is widely used by libraries, and museums, for the marking up of scholarly texts. It includes sets of tags that are tailored to particular genres, such as prose, poetry, and drama.
[6] Future Directions in Metadata Remediation for Metadata Aggregators http://www.diglib.org/aquifer/dlf110.pdf

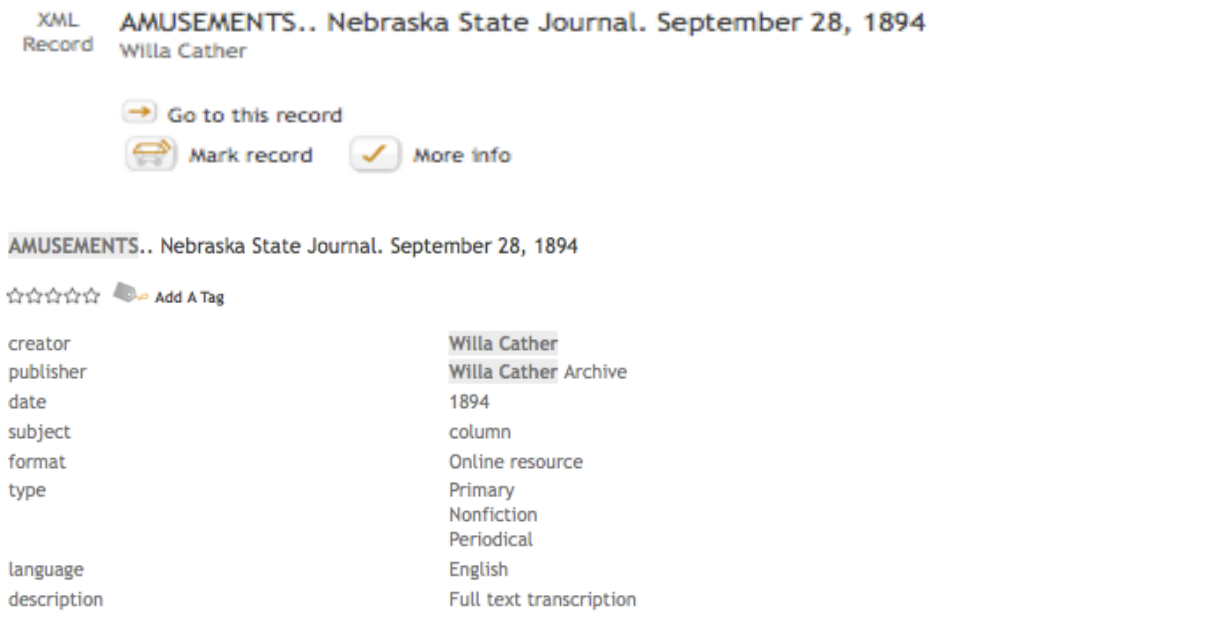

Figure 1. This record is harvested from *Cather*.

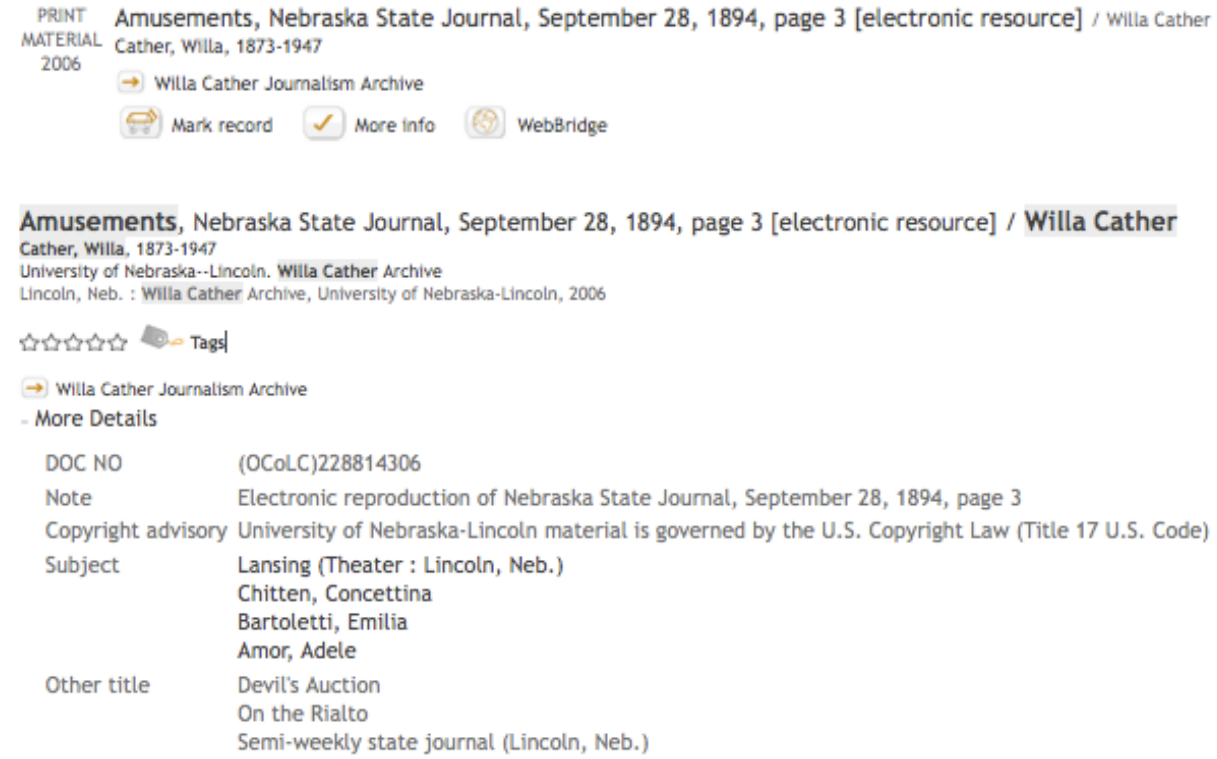

Figure 2. This record is harvested from the catalog. There are more access points in this record

Problem #2 – Flawed MARC encoding (AACR2)

It is inevitable that some CDRH projects will have MARC records in the catalog as well as records harvested from the TEI. So it is reasonable to expect some overlap. Whether or not there is overlap, there is no policy to address the way that MARC metadata is being displayed in Encore.

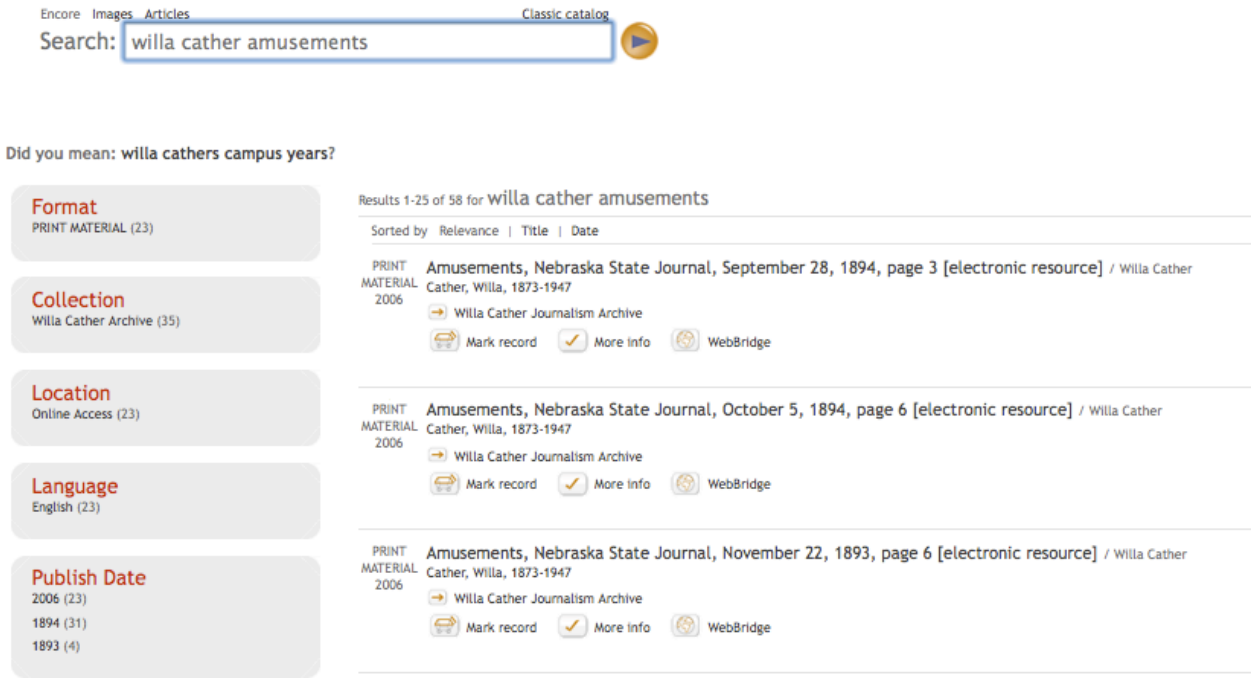

Figure 3. In a search for "Willa Cather Amusements" an Encore user will not be able to effectively limit records by Format, Collection, Location, Language, or Publish Date because the metadata that makes this faceting possible is not always present. In addition, this metadata may be AACR2-compliant but still not adequately represent the electronic editions of these works.

The figure above demonstrates that an Encore user might not be able to discover a relevant record for an electronic resource harvested from the catalog (i.e. a MARC record) because it would be sorted as "Print Material" and the publish date would be "2006." According to AACR2, this encoding is absolutely correct, yet a user may not expect to find a digitized 1894 Nebraska State Journal article classified as "Print Material" with a 2006 "Publish Date." We cannot take advantage of Encore's faceting if the MARC record encoding is flawed (due to inadequacies of AACR2) and confusing. In addition, a user would expect that the resources categorized in "Print Material" are different from those in the "Willa Cather Archive." In this case there is no difference. **Figure 3** demonstrates that there are 23 *Amusements* articles cataloged and there are 35 being harvested from *Cather*. This indicates that there are 12 *Amusements* that have not been cataloged yet. The Location facet would suggest that only 23 *Amusements* articles are available online, when in fact all of them (35) are online. Interestingly enough, only the records harvested from the catalog can be categorized or limited by Location. The Publish Date would indicate that 23 were published in 2006 while 4 were published in 1893 and 31 in 1894. According to AACR2, the 23 digital editions of the Cather articles were published in 2006. So there is nothing "wrong" with the MARC encoding. But the end result is simply confusing.

Some of these problems can be easily remedied. We recommend that CDRH and TS create a policy that will articulate what CDRH projects should be cataloged and harvested. More importantly a workflow should be established. There is agreement that the project-level record (e.g. *Cather*) should be cataloged so that it is part of WorldCat as well as our catalog and Encore. However, there is no policy that articulates how sub-collections within a project should be handled. Although a one-size-fits-all solution is unlikely, documentation should be created to at least increase the consistency across projects. With *Cather* it must decided whether or not both types of records are necessary for "Amusements." Removing or editing the TEI does not appear to be a feasible option. Another viable option would be to adjust the mapping or enhance the DC metadata once it is harvested; this method can be resource intensive (tweaking the XSLT and/or mappings) depending on the volume of records. Even though the records harvested from the catalog are better, the most expeditious option would be to remove *Amusements* MARC records from the catalog to prevent duplicate harvesting. Technical Services and CDRH should decide what the most appropriate remedy should be. Also, CDRH must keep technical services informed when items are added to projects so that the catalog records can be created in a timely manner.

The faceting problem cannot be easily remedied because it essentially exposes the problems inherent in AACR2. We recommend that Technical Services conduct a thorough review of these issues and propose a policy that is consistent with departmental priorities and resources. Local cataloging practices, which do not necessarily adhere to national rules, should be considered. The new cataloging code, RDA[7] may address some of these issues, but its implementation may be years away.

Problem #3 - 'collection' level metadata does not fully describe individual resources uniquely

More problematic than redundant records is harvesting too little metadata from rich TEI records. There is nothing wrong with the TEI metadata but when it is reduced to DC, data are lost. In addition, the TEI metadata that contains critical access points cannot always be mapped to DC and harvested. The TEI metadata is meant to be considered in context. Creating a DC record often removes the metadata from the context that adds value to the item.

For example, **Figure 4** demonstrates one problem that users encounter when searching for *Lewis and Clark* content in Encore. Although the two records are identical, the first record is a surrogate for the Mandan Plains Native people while the second record is a surrogate for the Flathead Native people. There is 1 metadata file that is populating the <title> fields for 34 unique *Lewis and Clark* resources.

[7] Resource Description and Access (RDA) is a standards effort to develop cataloging rules that would supersede the Anglo-American Cataloging Rules, 2nd edition (AACR2). It is scheduled for release at the end of November 2009.

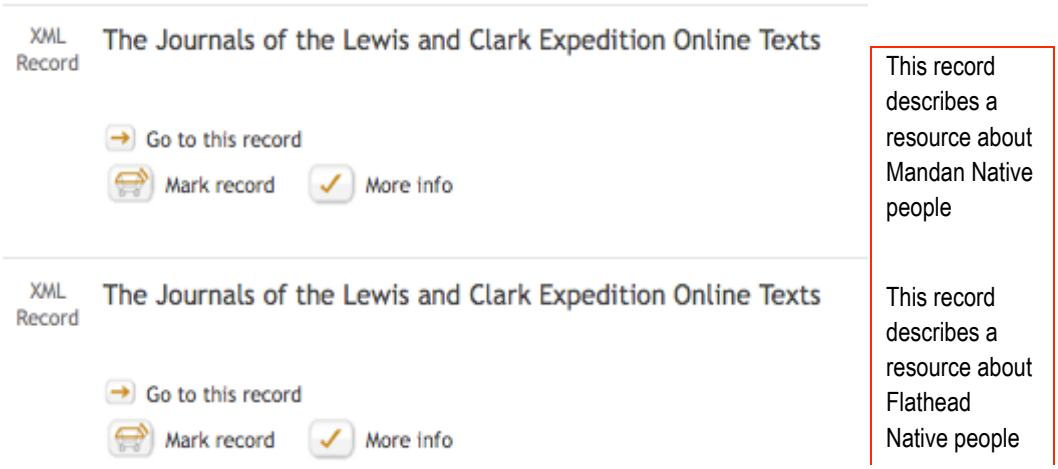

Figure 4. OAI-PMH is harvesting the same metadata element [<title>] even though the content is different.

In other words there is no TEI record that uniquely describes the Mandan, instead there is 1 record for 34 resources. This would be like cataloging a chapter without reference to the book, or, more accurately, like cataloging a column from a newspaper without reference to the newspaper. The contextualizing information was not mapped to the DC elements harvested, forcing a reassessment of the TEI to DC mappings. So the 34 online texts are described with the same metadata record, see **Figure 5** below.

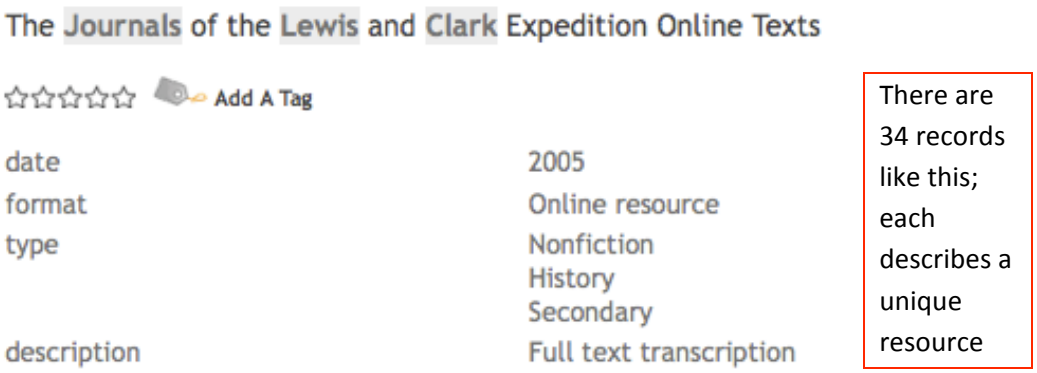

Figure 5. This is the metadata for the Mandan as well as 33 other resources; so identical metadata is being harvested from the TEI to describe different resources. The harvested metadata will lead a user to different resources.

How can we harvest metadata that can distinguish one resource from another? Editing the TEI metadata would be resource intensive thus unfeasible. However the TEI to DC mapping via XSLT can be customized and tweaked to improve description and access. These types of customizations will be handled on a case-by-case basis.

Another problem encountered was the exclusion of key metadata elements such as <title>. In **Figure 6**, the <title> element is empty so a user would not discover this record even if she searched by the terms "gallery of the open frontier." The record is 2 pages long in Encore.

Problem #4 – Automatically created metadata can be valid but inaccurate; errors are inevitable.

Quality control will pick out the rare cases such as this one below (**Figure 6**):

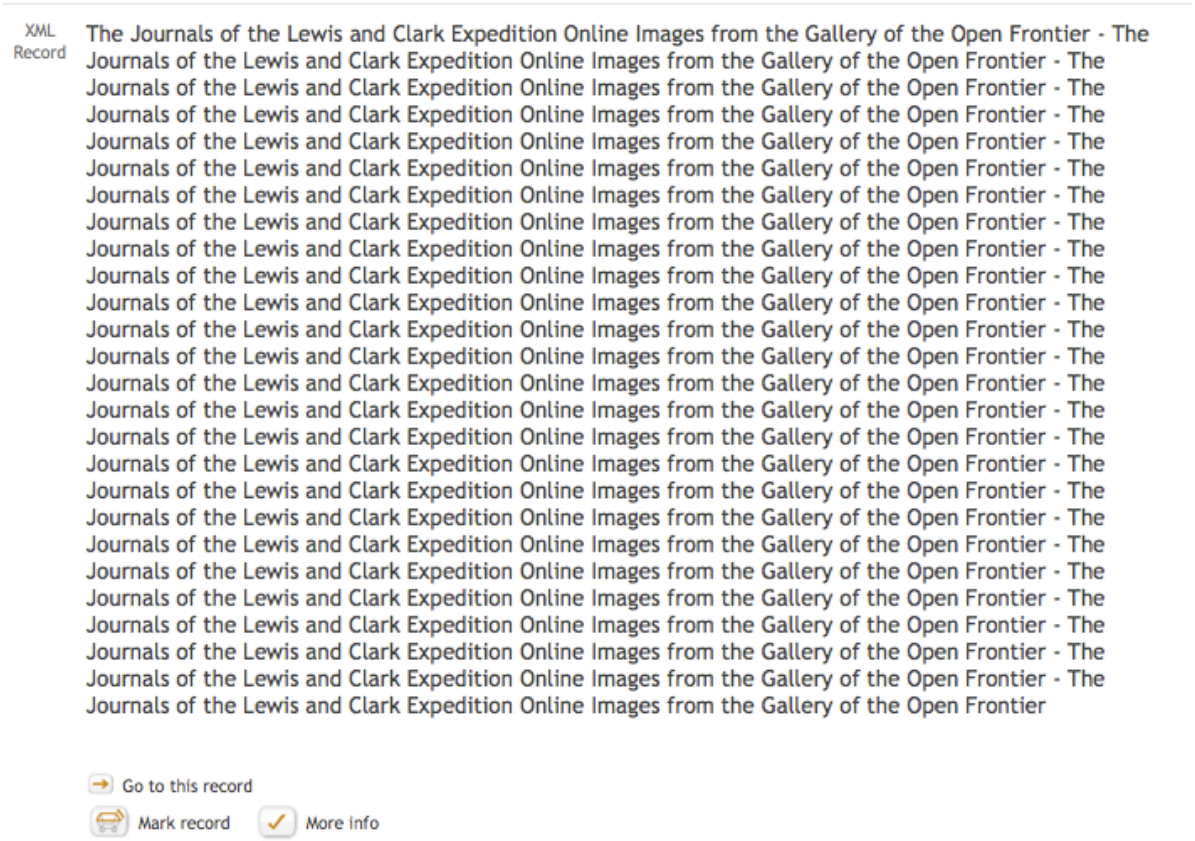

Figure 6. This record does not include the <title> element and it appears that one element is being pulled from numerous images into one element of the DC record. The result is a long record that is not descriptive or helpful to users.

To make matters worse, there are nearly 100 keywords in this record, which are not shown in **Figure 6**. Essentially the metadata for every image is included in this single record. We recommend that the mapping for the "Images" record should be edited. If having access to individual images is important, then the mapping should be customized for all the images that are part of the gallery. Another option would be to include these gallery images in ContentDM; thus it will end up being harvested in Encore the same way *Cather* images have become part of ContentDM. However this option still necessitates customization because descriptive metadata would be extracted from a different part of the TEI record. Because content such as images, are not necessarily owned by UNL, this possibility may be limited by rights issues. **These issues should be sorted out before the items are harvested.**

TEI to DC Crosswalk

The crosswalk (See Appendix) provided by III and tweaked by UNL appears be a good starting point for mapping the TEI to OAI compliant metadata for Encore. However, the way the TEI documents are encoded is not always consistent between -- and sometimes within – projects. This is a metadata normalization

problem that is typically unavoidable. Two catalogers do not catalog the same book identically and the same cataloger my catalog a book differently on two different occasions. **The bulk of our concerns about the TEI documents being harvested in Encore center on the following question: What level should CDRH projects be harvested?** What should we consider a record for the catalog and what should we consider a record for Encore?

Problem #5 – Dates in TEI metadata

Another problem that could confuse users would be the dates associated with the resources harvested in Encore. The date mapping became an issue in DC because TEI projects allow for the encoding of multiple date fields to record the provenance of the record. There are at least five date fields recorded in the TEI metadata for *Lewis and Clark*:

1. Date of the last online update
2. Date of the original online publication
3. Date of the publication of the source material
4. Date of the original publication of the source material and
5. Date of the original creation of the source.

Which of these dates should be harvested by OAI-PMH and then aggregated by Encore? How will the aggregated metadata influence the ordering within Encore? The task force agreed that the date of online publication and original creation are the two dates that are most likely to be of interest to the user. For instance, the metadata of *Lewis and Clark* should reflect the 2005 online publication date and the year the work was written, 1805. Rickel then added two date mappings/fields to the harvested metadata to improve the date sorting in Encore. The metadata scripts were customized to deal with the date problem in *Lewis and Clark* and *Cather* to improve the metadata. It is likely that once new projects are added, it will be necessary to create more customizations. The programmer should be working closely with the project curator or CDRH staff on this quality control process.

Problem #6 - Simple DC Elements in Encore

There are concerns about the metadata elements that validate according to OAI-PMH. Only simple metadata elements will validate according to the OAI-PMH validator. So, for example, in *Lewis and Clark*, these simple DC elements are indexed instead of the qualified elements. **Figure 7** demonstrates how useful information is lost in this metadata conversion:

TEI Metadata	**Simple DC Metadata**
<author>William Clark</author>	<dc:creator>William Clark</dc:creator>
<editor>Gary Moulton**</editor>**	<dc:creator>Gary Moulton</dc:creator>

Figure 7. Qualified metadata elements in TEI distinguish creators who are authors from creators who are editors, translators, curators, and so on. In this record, the simple DC metadata does not distinguish editors from authors; both are encoded as creators.

The fact that information is "stripped" out forces us to make decisions about who should be considered a creator and who is a contributor. This can be quite difficult and even a curator who is intimately familiar

with a project may not like being forced to make such decisions. These distinctions may seem innocuous on the surface but indeed they can be extremely important as more content is added to Encore.

Recommendations

Encore, as a tool to enhance access to a diverse array of resources, (images, theses, books, databases, images) has great potential to collocate our diverse assets for our users far better than the catalog could do. However, its beauty can also be a curse. Now our users are exposed to much more content than ever and if the metadata quality of harvestable records is not high, the user may end up confused. When *Cather*, full of its complexities, and embedded layers of information, was created it was never intended to be 'reduced' to DC and then aggregated with thousands of other resources that are not related to *Cather*. We have shown examples that demonstrate how the context and data are lost. **These problems are not prohibitive.** We have capable professionals who write scripts that convert hundreds of metadata records at one time. We also have capable professionals who can edit the mappings to produce more meaningful metadata. In sum, any of the problems can be mitigated with good (not necessarily perfect) metadata and customized mapping. This process is known as quality control.

The plan of action for improving the harvesting and indexing of TEI into Encore rests on costs and benefits. The problem comes down to the time and expertise needed to automate the metadata conversion as well as the creation and subsequent tweaking of the scripts/style sheets for harvesting data. We also have to decide what is good enough to meet the needs of our users.

The taskforce recommends that the library take the following action:

1. Establish criteria to determine the level that records (project, collections, and sub-collections) should be harvested for each project in Encore. Because of the complexity of CDRH projects, their inclusion in Encore requires thoughtful consideration on a project-by-project basis. Technical services and CDRH should work together closely on this process.
2. CDRH and TS should establish a policy that will articulate what CDRH projects should be cataloged. In addition establish a workflow that will result in the production of quality records in a timely manner.
3. TS should review the use of format codes (e.g. PRINT MATERIAL, COMPUTER FILE, THESIS/DISSERT) in the catalog code records. When the records are harvested in Encore, users can take advantage of Encore's format faceting feature. We recommend that this type of cleanup should only be completed as a batch process-- globally—rather than manually. Manually editing MARC records is resource intensive- thus unfeasible.
4. It appears that every project in CDRH will need a customized crosswalk to enable harvesting. CDRH, TS, and CORS will collaborate on metadata mapping customizations. In other words, at some point, someone intimate with the project itself will need to be involved in the decision making process of the data harvesting, if not the coding that generates the data itself.
5. TEI encoding was designed specifically to describe humanities data. Projects like *Cather* and *Lewis and Clark* will continue to exist primarily as a free standing website complete with its own search interface and context. We do <u>not</u> recommend that the TEI encoding be edited or enhanced so that it will be harvested different in Encore.

6. Future TEI projects in CDRH? The Libraries should create a plan in regard to TEI/Encore harvesting for new digital projects. Going forward, all project teams working with TEI should be aware of OAI harvesting as well as Encore aggregation. Furthermore, project teams should attempt to create metadata that would aggregate according to fields indexed by Encore—without compromising the integrity of the metadata (TEI or otherwise) or the project. Being proactive in creating Encore-harvestable metadata would greatly reduce the amount of customization needed to produce useful Encore records. Furthermore, projects should be aware of wider online implications, in which the contextualization inherent in well-planned websites is not available for individual files.
7. The Libraries should always be engaged in discussions with III about the direction of Encore and how it should be used in the future.
8. Utilize Encore's community tagging mechanism. On one hand, inputting tags would not require a high level of expertise, but on the other, this method could require close cooperation between those who know the content and those doing the tagging. Community tagging provides addition access points for our users.

Conclusion

The bottom line is that everyone is the library is responsible for quality control. Technical Services is not the only department responsible for Encore's metadata quality. It is critical that the faculty and staff engage Encore to really learn how it works. With this intense exploration, one will learn its strengths and weaknesses and contribute to the improvement of the tool and improve its value. All library departments need to collaborate to create viable policies and workflow to improve the effectiveness of Encore.

APPENDIX

Mapping *Lewis and Clark* and *Cather* TEI to Dublin Core for loading into Encore (June 15, 2008)

DC	Current mapping	NOTES
Title	Lewis and Clark: <teiHeader><fileDesc><titleStmt><title type="main"> <teiHeader><fileDesc><titleStmt><title type="sub"> Cather: <teiHeader><fileDesc><sourceDesc><bibl><title level='a'> <teiHeader><fileDesc><sourceDesc><bibl><title level='m'> <teiHeader><fileDesc><sourceDesc><bibl><date>	Original recommendation from Innovative: <teiHeader><fileDesc><titleStmt><title>
Creator	Lewis & Clark: <teiHeader><profileDesc><handList><hand scribe=**?**> (value of scribe attribute based on certain conditions Cather:<teiHeader><fileDesc><titleStmt><author>	<teiHeader><fileDesc><titleStmt><author> followed by (in separate element(s)) <teiHeader><fileDesc><titleStmt><editor>
Date	Lewis & Clark:))<teiHeader><fileDesc><editionStmt><edition><date> for the digitized date and <text><body><date value=**?**> (value of the value attribute is original publication date) Cather:<teiHeader><fileDesc><sourceDesc><bibl><date>	Original recommendation: <teiHeader><sourceDesc><editionStmt><bibl><date>(if available) followed by (in separate element(s))<teiHeader><fileDesc><editionStmt><edition><date> A date associated with an event in the life cycle of the resource. Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [Date and Time Formats, W3C Note, http://www.w3.org/TR/NOTE-datetime and follows the YYYY-MM-DD format.
Description	Full text transcription	Think of this as a "notes" field. It is an account of the content of the resource. Description may include the abstract or TOC. There doesn't seem to be a good data element to map for this? Maybe every record could get a blanket description? Thus far "Full text transcription" is being used in LC & WCA
Format	"XML" based on original recommendation	This is hard coded but not compliant with DC standard. Format is the physical or digital manifestation of the resource. Typically, it includes the media-type or dimensions of the resource. Examples of dimensions include size and duration. Format may be used to determine the software, hardware or other equipment needed to

		display or operate the resource. EX: “image/gif,” “4MB,” “Macromedia Flash,” See http://www.iana.org/assignments/media-types/
Identifier	URL to the resource	URL to find the resource in Tamino
Publisher	Lewis and Clark: <teiHeader><fileDesc><publicationStmt><distributor> Cather: <teiHeader><fileDesc><publicationStmt><distributor>	The entity responsible for making the resource available. Original recommendation: <teiHeader><fileDesc><publicationStmt><distributor>
Language	English	A language of the intellectual content of the resource.
Subject	Lewis & Clark: <keywords><term> Cather: <keywords><term>	The topic of the content of the resource. Typically, a Subject will be expressed as keywords or key phrases or classification codes that describe the topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme. There doesn’t seem to be a good data element to map for this? Maybe every record could get a blanket set of subject elements based on criteria known to you?
Rights	Lewis & Clark: not sent through (flagged for <availability><p>, but commented out) Cather: not sent through (flagged for <availability><p>, but commented out)	Original recommendation: <teiHeader><fileDesc><publicationStmt><p> Information about rights held in and over the resource. Typically a Rights element will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights.
Type	Lewis and Clark: “Nonfiction,” “History,” “periodical,” “Secondary,” “Primary, ”Paratext” Cather: Had the same as Lewis and Clark, but commented out recently.	LC & WCA has been hardcoded with the following terms: “Nonfiction,” “History,” “Secondary,” “Primary,” Paratext” but in order to be DC standard it should be “Text,” “Image,” “audio for LC The list is available http://dublincore.org/documents/dcmi-type-vocabulary/
Contributor	Lewis & Clark: <monogr><editor> Cather: <biblFull><editor> and <biblFull><titleStmt><respStmt>	An entity responsible for making contributions to the content of the resource, this could be a person, institution. This is typically used when primary responsibility is unknown
Source	Lewis and Clark: not supplied Cather: recently added <teiHeader><fileDesc><sourceDesc><bibl><title level='m'> <teiHeader><fileDesc><sourceDesc><bibl><date> <teiHeader><fileDesc><sourceDesc><bibl><biblScope>	A Reference to a resource from which the present resource is derived. The present resource may be derived from the Source resource in whole or part.

Coverage	Not supplied	The extent or scope of the content of the resource. Coverage will typically include spatial location (a place name) or temporal period (date range) or jurisdiction. We should select a value from a controlled vocabulary (LCSH or Thesaurus of Geographic Names)
Relation	Not supplied	A reference to a related resource; this can be used to show how an image relates to a collection or how a collection relates to a project. "IsVersionOf," "IsPartof," "ISReferencedBy"