Using Schematron for Analyzing Conformance to Best Practices for EAD, TEI, and MODS

(and some other thoughts on workflow tools)

Jenn Riley
Metadata Librarian
Indiana University Digital Library Program

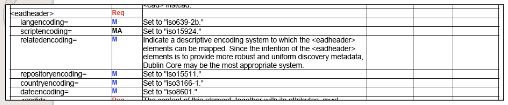
Consistency is a challenge

- Document-centric XML (TEI, EAD) is very difficult to create consistently
- Some common tools to help:
  - Schema/DTD validation
  - Tag libraries
  - XML templates
  - Example documents
  - Keyboard macros
  - Detailed encoding guidelines
- These are not enough!

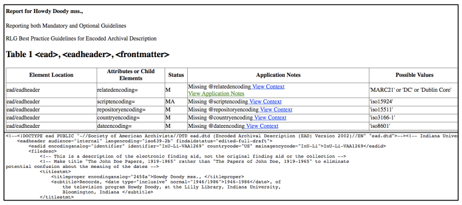
Another possible tool layer

- Machine validation of a file against local encoding guidelines
- Can only go so far, but that far is extremely helpful
- Indiana University implemented using:
  - Schematron (http://www.schematron.com)
  - <oXygen /> plugin architecture

Inspiration: RLG EAD Report Card



		<ead> instead.			
<eadheader>	Req				
langencoding=	M	Set to "iso639-2b."			
scriptencoding=	MA	Set to "iso15924."			
relatedencoding=	M	Indicate a descriptive encoding system to which the <eadheader> elements can be mapped. Since the intention of the <eadheader> elements is to provide more robust and uniform discovery metadata, Dublin Core may be the most appropriate system.			
repositoryencoding=	M	Set to "iso15511."			
countryencoding=	M	Set to "iso3166-1."			
dateencoding=	M	Set to "iso8601."			
<eadid>	Req	The content of this element, together with its attributes, must			



Report for Howdy Doody mss.,

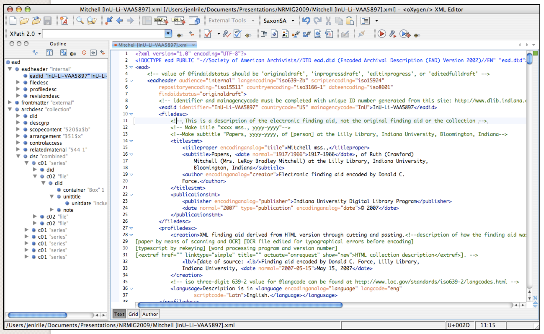
Reporting both Mandatory and Optional Guidelines

RLG Best Practice Guidelines for Encoded Archival Description

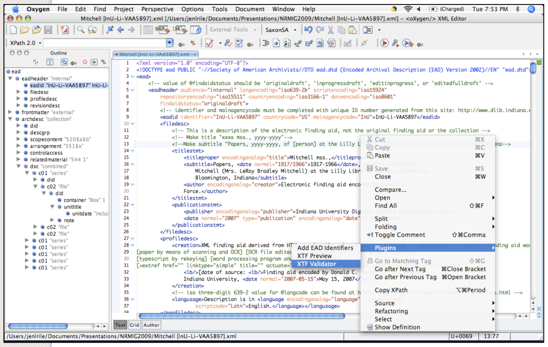
Table 1 <ead>, <eadheader>, <frontmatter>

Element Location	Attributes or Child Elements	Status	Application Notes	Possible Values
ead/eadheader	relatedencoding=	M	Missing @relatedencoding View Context View Application Notes	'MARC21' or 'DC' or 'Dublin Core'
ead/eadheader	scriptencoding=	MA	Missing @scriptencoding View Context	'iso15924'
ead/eadheader	repositoryencoding=	M	Missing @repositoryencoding View Context	'iso15511'
ead/eadheader	countryencoding=	M	Missing @countryencoding View Context	'iso3166-1'
ead/eadheader	dateencoding=	M	Missing @dateencoding View Context	'iso8601'

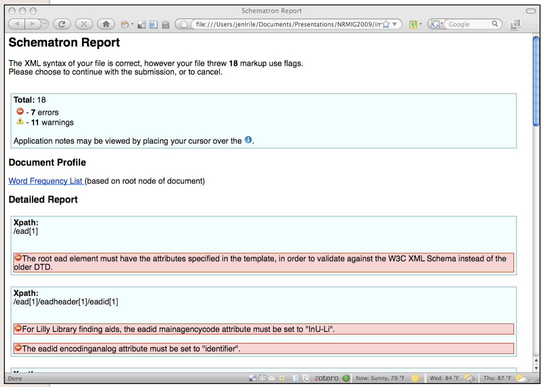
```
<!--<!DOCTYPE ead PUBLIC "-//Society of American Archivists//DTD ead.dtd (Encoded Archival Description (EAD) Version 2002)//EN" "ead.dtd">--><!-- Indiana Unive
    <eadheader audience="internal" langencoding="iso639-2b" findaidstatus="edited-full-draft">
        <eadid encodinganalog="identifier" identifier="InU-Li-VAA1269" countrycode="US" mainagencycode="InU-Li">InU-Li-VAA1269</eadid>
        <filedesc>
            <!-- This is a description of the electronic finding aid, not the original finding aid or the collection -->
            <!-- Make title "The John Doe Papers, 1919-1965" rather than "The Papers of John Doe, 1919-1965" to eliminate
potential confusion about the meaning of the dates -->
            <titlestmt>
                <titleproper encodinganalog="245$a">Howdy Doody mss., </titleproper>
                <subtitle>Records, <date type="inclusive" normal="1946/1986">1946-1986</date>, of
                    the television program Howdy Doody, at the Lilly Library, Indiana University,
                    Bloomington, Indiana </subtitle>
            </titlestmt>
```



```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ead PUBLIC "-//Society of American Archivists//DTD ead.dtd (Encoded Archival Description (EAD) Version 2002)//EN" "ead.dtd"
<ead>
    <!-- value of @findaidstatus should be 'originaldraft', 'inprogressdraft', 'editinprogress', or 'editedfulldraft' -->
    <eadheader audience="internal" langencoding="iso639-2b" scriptencoding="iso15924"
        repositoryencoding="iso15511" countryencoding="iso3166-1" dateencoding="iso8601"
        findaidstatus="originaldraft">
        <!-- identifier and mainagencycode must be completed with unique ID number generated from this site: http://www.dlib.indiana.e
        <eadid identifier="InU-Li-VAA5897" countrycode="US" mainagencycode="InU">InU-Li-VAA5897</eadid>
        <filedesc>
            <!-- This is a description of the electronic finding aid, not the original finding aid or the collection -->
            <!-- Make title "xxxx mss., yyyy-yyyy"-->
            <!--Make subtitle "Papers, yyyy-yyyy, of [person] at the Lilly Library, Indiana University, Bloomington, Indiana-->
            <titlestmt>
                <titleproper encodinganalog="title">Mitchell mss.,</titleproper>
                <subtitle>Papers, <date normal="1917/1966">1917-1966</date>, of Ruth (Crawford)
                    Mitchell (Mrs. LeRoy Bradley Mitchell) at the Lilly Library, Indiana University,
                    Bloomington, Indiana</subtitle>
                <author encodinganalog="creator">Electronic finding aid encoded by Donald C.
                Force.</author>
            </titlestmt>
            <publicationstmt>
                <publisher encodinganalog="publisher">Indiana University Digital Library Program</publisher>
                <date normal="2007" type="publication" encodinganalog="date">© 2007</date>
            </publicationstmt>
        </filedesc>
        <profiledesc>
            <creation>XML finding aid derived from HTML version through cutting and pasting.<!--description of how the finding aid was
[paper by means of scanning and OCR] [OCR file edited for typographical errors before encoding]
[typescript by rekeying] [word processing program and version number]
[<extref href="" linktype="simple" title="" actuate="onrequest" show="new">HTML collection description</extref>]. -->
                <lb/>[date of source: <lb/>Finding aid encoded by Donald C. Force, Lilly Library,
                Indiana University, <date normal="2007-05-15">May 15, 2007</date>
            </creation>
            <!-- iso three-digit 639-2 value for @langcode can be found at http://www.loc.gov/standards/iso639-2/langcodes.html -->
            <langusage>Description is in <language encodinganalog="language" langcode="eng"
                    scriptcode="Latn">English.</language></langusage>
```



```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ead PUBLIC "-//Society of American Archivists//DTD ead.dtd (Encoded Archival Description (EAD) Version 2002)//EN" "ead.dtd"
<ead>
    <!-- value of @findaidstatus should be 'originaldraft', 'inprogressdraft', 'editinprogress', or 'editedfulldraft' -->
    <eadheader audience="internal" langencoding="iso639-2b" scriptencoding="iso15924"
        repositoryencoding="iso15511" countryencoding="iso3166-1" dateencoding="iso8601"
        findaidstatus="originaldraft">
        <!-- identifier and mainagencycode must be completed with unique ID number generated from this site: http://www.dlib.indiana.e
        <eadid identifier="InU-Li-VAA5897" countrycode="US" mainagencycode="InU">InU-Li-VAA5897</eadid>
        <filedesc>
            <!-- This is a description of the electronic finding aid, not the original finding aid or the collection -->
            <!-- Make title "xxxx mss., yyyy-yyyy"-->
            <!--Make subtitle "Papers, yyyy-yyyy, of [person] at the Lilly L                                  na-->
            <titlestmt>
                <titleproper encodinganalog="title">Mitchell mss.,</titlepro
                <subtitle>Papers, <date normal="1917/1966">1917-1966</date>,
                    Mitchell (Mrs. LeRoy Bradley Mitchell) at the Lilly Libr
                    Bloomington, Indiana</subtitle>
                <author encodinganalog="creator">Electronic finding aid enco
                Force.</author>
            </titlestmt>
            <publicationstmt>
                <publisher encodinganalog="publisher">Indiana University Dig
                <date normal="2007" type="publication" encodinganalog="date"
            </publicationstmt>
        </filedesc>
        <profiledesc>
            <creation>XML finding aid derived from HT                                   ding aid was
[paper by means of scanning and OCR] [OCR file edite
[typescript by rekeying] [word processing program and
[<extref href="" linktype="simple" title="" actuate="
                <lb/>[date of source: <lb/>Finding aid encoded by Donald C.
                Indiana University, <date normal="2007-05-15">May 15, 2007</
            </creation>
            <!-- iso three-digit 639-2 value for @langcode can be found at h                    s.html -->
            <langusage>Description is in <language encodinganalog="language"
                    scriptcode="Latn">English.</language></langusage>
```



Schematron Report

The XML syntax of your file is correct, however your file threw **18** markup use flags.
Please choose to continue with the submission, or to cancel.

Total: 18
- **7** errors
- **11** warnings

Application notes may be viewed by placing your cursor over the ⓘ.

Document Profile

Word Frequency List (based on root node of document)

Detailed Report

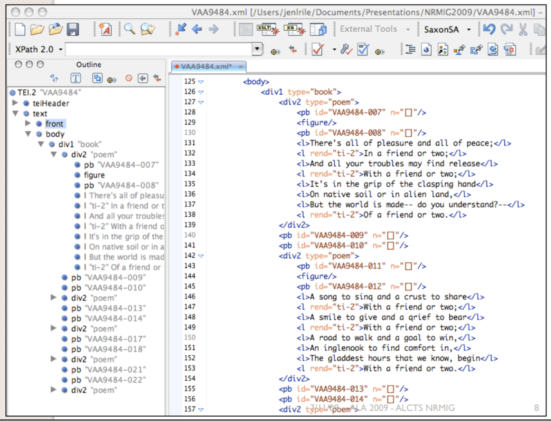
Xpath:
/ead[1]

The root ead element must have the attributes specified in the template, in order to validate against the W3C XML Schema instead of the older DTD.

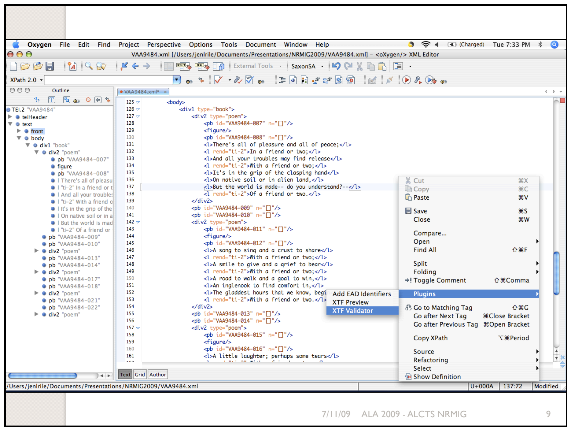
Xpath:
/ead[1]/eadheader[1]/eadid[1]

For Lilly Library finding aids, the eadid mainagencycode attribute must be set to "InU-Li".

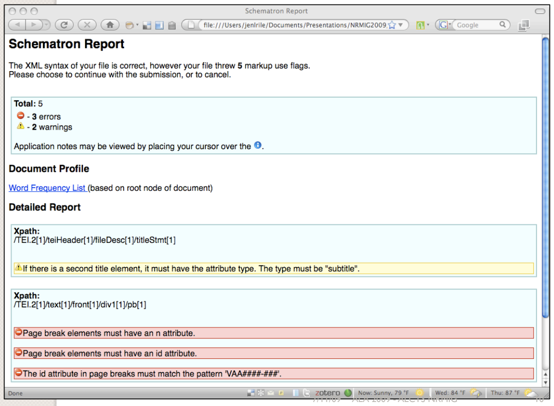
The eadid encodinganalog attribute must be set to "identifier".



```
<body>
    <div1 type="book">
        <div2 type="poem">
            <pb id="VAA9484-007" n="□"/>
            <figure/>
            <pb id="VAA9484-008" n="□"/>
            <l>There's all of pleasure and all of peace;</l>
            <l rend="ti-2">In a friend or two;</l>
            <l>And all your troubles may find release</l>
            <l rend="ti-2">With a friend or two;</l>
            <l>It's in the grip of the clasping hand</l>
            <l>On native soil or in alien land,</l>
            <l>But the world is made-- do you understand?--</l>
            <l rend="ti-2">Of a friend or two.</l>
        </div2>
        <pb id="VAA9484-009" n="□"/>
        <pb id="VAA9484-010" n="□"/>
        <div2 type="poem">
            <pb id="VAA9484-011" n="□"/>
            <figure/>
            <pb id="VAA9484-012" n="□"/>
            <l>A song to sing and a crust to share</l>
            <l rend="ti-2">With a friend or two;</l>
            <l>A smile to give and a grief to bear</l>
            <l rend="ti-2">With a friend or two;</l>
            <l>A road to walk and a goal to win,</l>
            <l>An inglenook to find comfort in,</l>
            <l>The gladdest hours that we know, begin</l>
            <l rend="ti-2">With a friend or two.</l>
        </div2>
        <pb id="VAA9484-013" n="□"/>
        <pb id="VAA9484-014" n="□"/>
        <div2 type="poem">
```



```
<body>
    <div1 type="book">
        <div2 type="poem">
            <pb id="VAA9484-007" n="□"/>
            <figure/>
            <pb id="VAA9484-008" n="□"/>
            <l>There's all of pleasure and all of peace;</l>
            <l rend="ti-2">In a friend or two;</l>
            <l>And all your troubles may find release</l>
            <l rend="ti-2">With a friend or two;</l>
            <l>It's in the grip of the clasping hand</l>
            <l>On native soil or in alien land,</l>
            <l>But the world is made-- do you understand?--</l>
            <l rend="ti-2">Of a friend or two.</l>
        </div2>
        <pb id="VAA9484-009" n="□"/>
        <pb id="VAA9484-010" n="□"/>
        <div2 type="poem">
            <pb id="VAA9484-011" n="□"/>
            <figure/>
            <pb id="VAA9484-012" n="□"/>
            <l>A song to sing and a crust to share</l>
            <l rend="ti-2">With a friend or two;</l>
            <l>A smile to give and a grief to bear</l>
            <l rend="ti-2">With a friend or two;</l>
            <l>A road to walk and a goal to win,</l>
            <l>An inglenook to find comfort in,</l>
            <l>The gladdest hours that we know, begi
            <l rend="ti-2">With a friend or two.</l>
        </div2>
        <pb id="VAA9484-013" n="□"/>
        <pb id="VAA9484-014" n="□"/>
        <div2 type="poem">
            <pb id="VAA9484-015" n="□"/>
            <figure/>
            <pb id="VAA9484-016" n="□"/>
            <l>A little laughter; perhaps some tears</l>
```



Schematron Report

The XML syntax of your file is correct, however your file threw **5** markup use flags.
Please choose to continue with the submission, or to cancel.

Total: 5
- **3** errors
- **2** warnings

Application notes may be viewed by placing your cursor over the ⓘ.

Document Profile

Word Frequency List (based on root node of document)

Detailed Report

Xpath:
/TEI.2[1]/teiHeader[1]/fileDesc[1]/titleStmt[1]

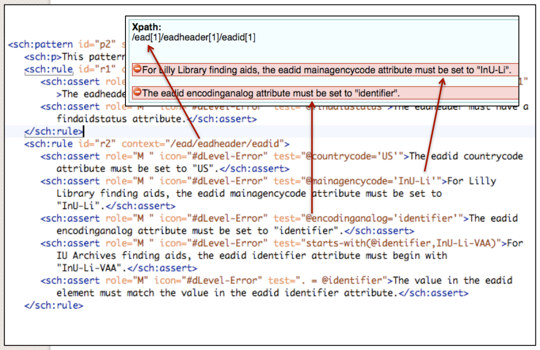
If there is a second title element, it must have the attribute type. The type must be "subtitle".

Xpath:
/TEI.2[1]/text[1]/front[1]/div1[1]/pb[1]

Page break elements must have an n attribute.

Page break elements must have an id attribute.

The id attribute in page breaks must match the pattern 'VAA####-###'.



Xpath:
/ead[1]/eadheader[1]/eadid[1]

For Lilly Library finding aids, the eadid mainagencycode attribute must be set to "InU-Li".

The eadid encodinganalog attribute must be set to "identifier".

```
<sch:pattern id="p2" s
   <sch:p>This pattern
   <sch:rule id="r1" c
      <sch:assert role
         >The eadheade
      <sch:assert role
         findaidstatus attribute.</sch:assert>
   </sch:rule>
   <sch:rule id="r2" context="/ead/eadheader/eadid">
      <sch:assert role="M " icon="#dLevel-Error" test="@countrycode='US'">The eadid countrycode
         attribute must be set to "US".</sch:assert>
      <sch:assert role="M " icon="#dLevel-Error" test="@mainagencycode='InU-Li'">For Lilly
         Library finding aids, the eadid mainagencycode attribute must be set to
         "InU-Li".</sch:assert>
      <sch:assert role="M " icon="#dLevel-Error" test="@encodinganalog='identifier'">The eadid
         encodinganalog attribute must be set to "identifier".</sch:assert>
      <sch:assert role="M " icon="#dLevel-Error" test="starts-with(@identifier,InU-Li-VAA)">For
         IU Archives finding aids, the eadid identifier attribute must begin with
         "InU-Li-VAA".</sch:assert>
      <sch:assert role="M" icon="#dLevel-Error" test=". = @identifier">The value in the eadid
         element must match the value in the eadid identifier attribute.</sch:assert>
   </sch:rule>
```

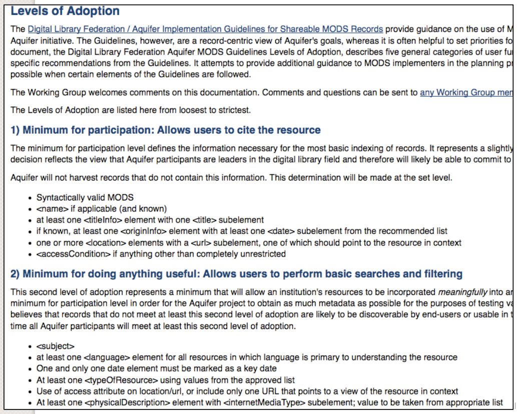
More info on Schematron

- ISO/IEC 19757 - Document Schema Definition Languages (DSDL) - Part 3: Rule-based validation – Schematron.
- Be careful! http://www.schematron.com has the primary specs; http://schematron.com is for a particular company's tool using them. (Weird.)
- This is the page you want:

schematron
A language for making assertions about patterns found in XML documents

Using a Schematron file

- Schematron home page provides two distributions:
  - One for XSLT 1.0 processors and one for 2.0 processors
  - Each includes a set of three stylesheets to be used in turn on the Schematron file
  - Result of this processing is a stylesheet to be run on your XML instance document
- IU implementation wraps this all up into an <oXygen /> plugin written in Java
- You could also pipe them together with a shell script, a Windows .bat file, etc……



Levels of Adoption

The Digital Library Federation / Aquifer Implementation Guidelines for Shareable MODS Records provide guidance on the use of M
Aquifer initiative. The Guidelines, however, are a record-centric view of Aquifer's goals, whereas it is often helpful to set priorities fo
document, the Digital Library Federation Aquifer MODS Guidelines Levels of Adoption, describes five general categories of user fu
specific recommendations from the Guidelines. It attempts to provide additional guidance to MODS implementers in the planning pr
possible when certain elements of the Guidelines are followed.

The Working Group welcomes comments on this documentation. Comments and questions can be sent to any Working Group mer

The Levels of Adoption are listed here from loosest to strictest.

1) Minimum for participation: Allows users to cite the resource

The minimum for participation level defines the information necessary for the most basic indexing of records. It represents a slightly
decision reflects the view that Aquifer participants are leaders in the digital library field and therefore will likely be able to commit to

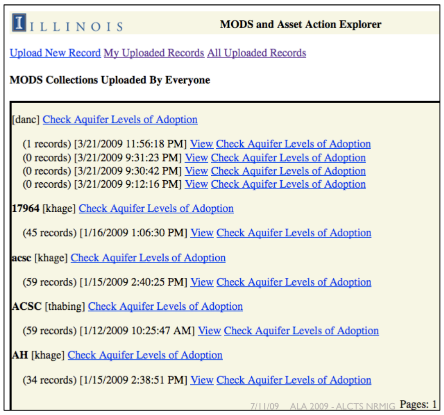
Aquifer will not harvest records that do not contain this information. This determination will be made at the set level.

- Syntactically valid MODS
- <name> if applicable (and known)
- at least one <titleInfo> element with one <title> subelement
- if known, at least one <originInfo> element with at least one <date> subelement from the recommended list
- one or more <location> elements with a <url> subelement, one of which should point to the resource in context
- <accessCondition> if anything other than completely unrestricted

2) Minimum for doing anything useful: Allows users to perform basic searches and filtering

This second level of adoption represents a minimum that will allow an institution's resources to be incorporated *meaningfully* into ar
minimum for participation level in order for the Aquifer project to obtain as much metadata as possible for the purposes of testing va
believes that records that do not meet at least this second level of adoption are likely to be discoverable by end-users or usable in t
time all Aquifer participants will meet at least this second level of adoption.

- <subject>
- at least one <language> element for all resources in which language is primary to understanding the resource
- One and only one date element must be marked as a key date
- At least one <typeOfResource> using values from the approved list
- Use of access attribute on location/url, or include only one URL that points to a view of the resource in context
- At least one <physicalDescription> element with <internetMediaType> subelement; value to be taken from appropriate list



ILLINOIS **MODS and Asset Action Explorer**

Upload New Record My Uploaded Records All Uploaded Records

MODS Collections Uploaded By Everyone

[danc] Check Aquifer Levels of Adoption

(1 records) [3/21/2009 11:56:18 PM] View Check Aquifer Levels of Adoption
(0 records) [3/21/2009 9:31:23 PM] View Check Aquifer Levels of Adoption
(0 records) [3/21/2009 9:30:42 PM] View Check Aquifer Levels of Adoption
(0 records) [3/21/2009 9:12:16 PM] View Check Aquifer Levels of Adoption

17964 [khage] Check Aquifer Levels of Adoption

(45 records) [1/16/2009 1:06:30 PM] View Check Aquifer Levels of Adoption

acsc [khage] Check Aquifer Levels of Adoption

(59 records) [1/15/2009 2:40:25 PM] View Check Aquifer Levels of Adoption

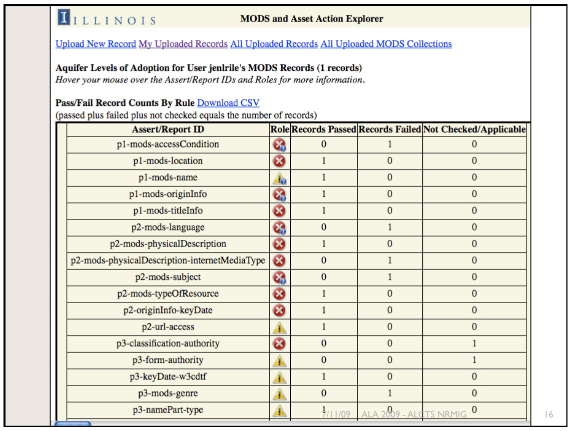
ACSC [thabing] Check Aquifer Levels of Adoption

(59 records) [1/12/2009 10:25:47 AM] View Check Aquifer Levels of Adoption

AH [khage] Check Aquifer Levels of Adoption

(34 records) [1/15/2009 2:38:51 PM] View Check Aquifer Levels of Adoption

Pages: 1



ILLINOIS **MODS and Asset Action Explorer**

Upload New Record My Uploaded Records All Uploaded Records All Uploaded MODS Collections

Aquifer Levels of Adoption for User jenlrile's MODS Records (1 records)
Hover your mouse over the Assert/Report IDs and Roles for more information.

Pass/Fail Record Counts By Rule Download CSV
(passed plus failed plus not checked equals the number of records)

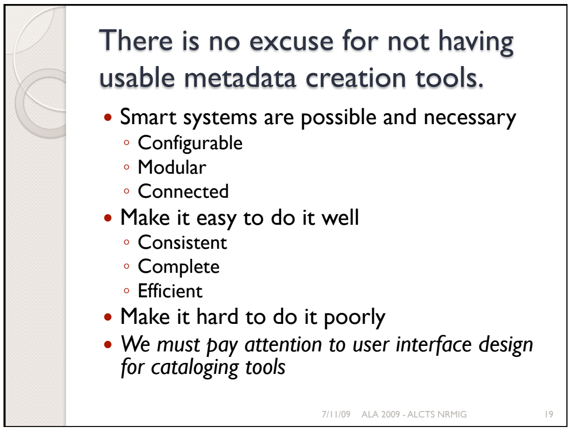
Assert/Report ID	Role	Records Passed	Records Failed	Not Checked/Applicable
p1-mods-accessCondition	⊗	0	1	0
p1-mods-location	⊗	1	0	0
p1-mods-name	⚠	1	0	0
p1-mods-originInfo	⊗	1	0	0
p1-mods-titleInfo	⊗	1	0	0
p2-mods-language	⊗	0	1	0
p2-mods-physicalDescription	⊗	1	0	0
p2-mods-physicalDescription-internetMediaType	⊗	0	1	0
p2-mods-subject	⊗	0	1	0
p2-mods-typeOfResource	⊗	1	0	0
p2-originInfo-keyDate	⊗	1	0	0
p2-url-access	⚠	1	0	0
p3-classification-authority	⊗	0	0	1
p3-form-authority	⚠	0	0	1
p3-keyDate-w3cdtf	⚠	1	0	0
p3-mods-genre	⚠	0	1	0
p3-namePart-type	⚠	1	0	0

Let's step back

- How can better tools revolutionize metadata creation workflows?
  - Promoting consistency
    - This is hard and not something that humans are generally good at
  - True interoperability between systems
    - *Without futzing!*
- We spend too much valuable human time doing repetitive and low-value tasks as part of descriptive workflows

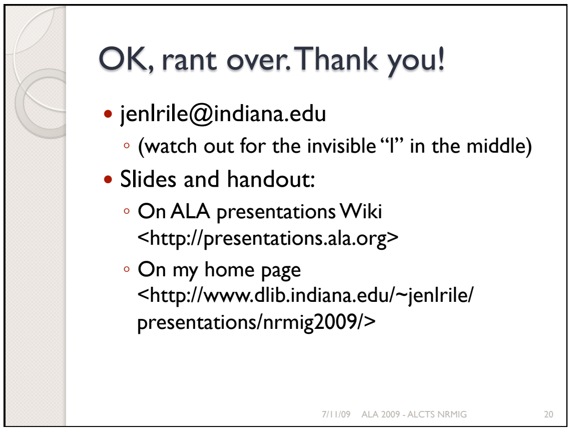
Were do we go from here?

- Make better use of available technologies
  - Automating
  - Streamlining
  - Validating
- We can and *must* do our jobs better and more efficiently, with the help of better tools
  - Providing comparable services with less
  - Creating a convincing argument for more?



There is no excuse for not having usable metadata creation tools.

- Smart systems are possible and necessary
  - Configurable
  - Modular
  - Connected
- Make it easy to do it well
  - Consistent
  - Complete
  - Efficient
- Make it hard to do it poorly
- *We must pay attention to user interface design for cataloging tools*



OK, rant over. Thank you!

- jenlrile@indiana.edu
  - (watch out for the invisible "l" in the middle)
- Slides and handout:
  - On ALA presentations Wiki <http://presentations.ala.org>
  - On my home page <http://www.dlib.indiana.edu/~jenlrile/presentations/nrmig2009/>