

# **pre-checker checks**

how to look at EML

# **big outline**

why would I want to look at eml?

what can I see?

what is expected?

how can I fix it? (manually)

how can I fix it? (script)

the EML Manager

# why look at eml?

Even if my site's EML is generated programmatically, and I never need to manually edit the EML, I still want to know

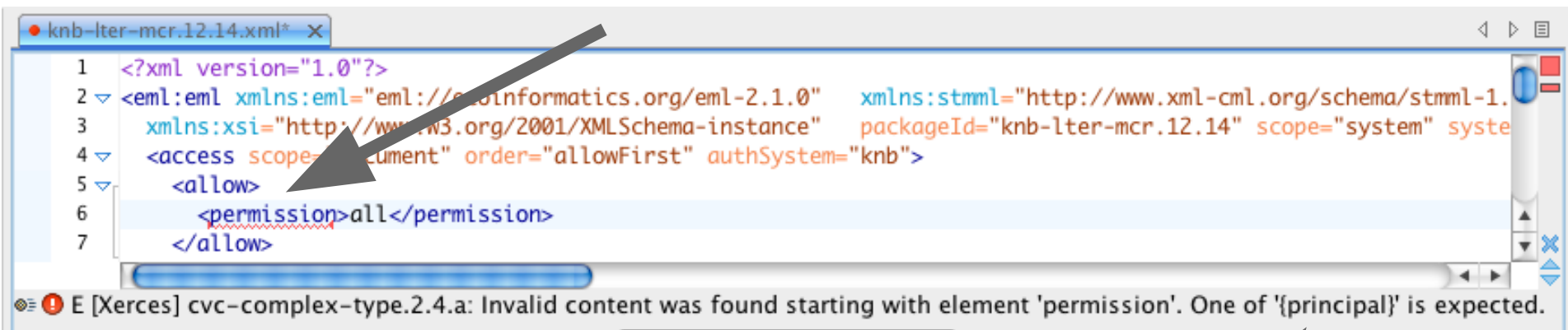
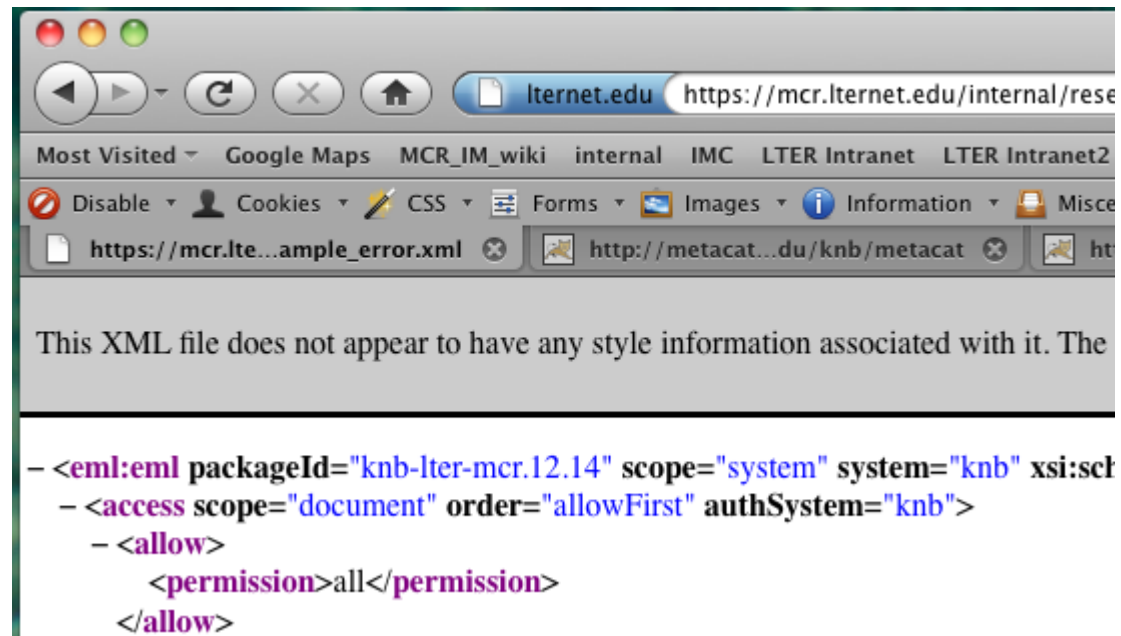
- \* is it valid?
- \* what is missing?
- \* what is causing harvest errors?
- \* what do I want the EML product to look like?

(the requirements for the eml-generating code)

# is it valid?

- \* browser will not tell if eml is valid

- \* xml editor will



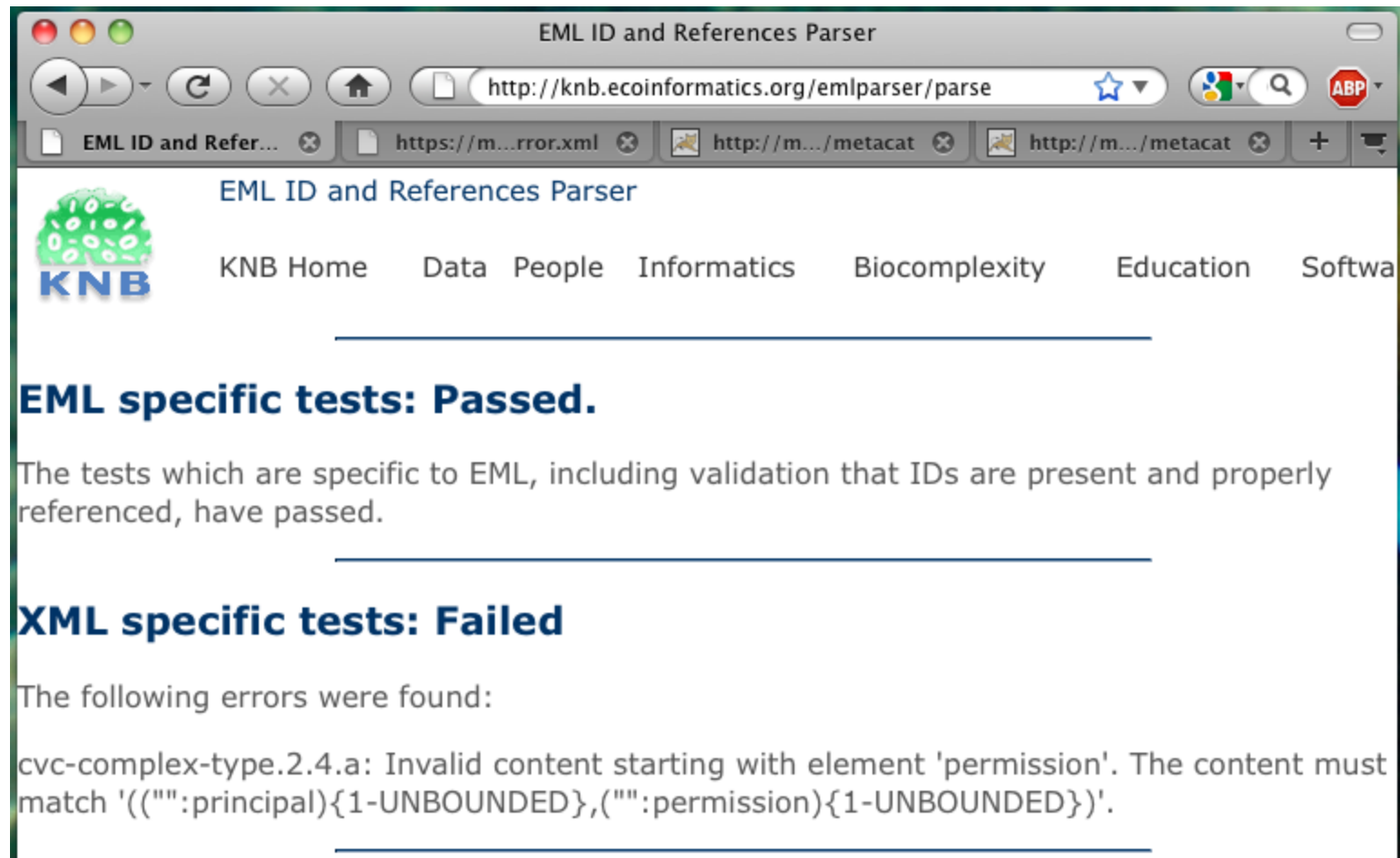
- \* knb parser will...

<principal>uid=MCR,o=lter,dc=ecoinformatics,dc=org</principal>

# is it valid?

\* knb parser will find errors

<http://knb.ecoinformatics.org/emlparser/>



The screenshot shows a web browser window titled "EML ID and References Parser". The address bar displays the URL <http://knb.ecoinformatics.org/emlparser/parse>. The page features the KNB logo and a navigation menu with links: KNB Home, Data, People, Informatics, Biocomplexity, Education, and Software. The main content area displays the results of two tests:

- EML specific tests: Passed.**  
The tests which are specific to EML, including validation that IDs are present and properly referenced, have passed.
- XML specific tests: Failed**  
The following errors were found:  
cvc-complex-type.2.4.a: Invalid content starting with element 'permission'. The content must match '("("principal){1-UNBOUNDED},("":permission){1-UNBOUNDED})'.

parsing



schema  
validation



# parsing separate from validation

parsing:

- \* ids must be unique

Make attribute id's unique.  
(One way is to qualify  
attribute id with table id.)

```
<attribute id="long.year" s  
  <attributeName>year</at
```

```
<attribute id="wide.year" s  
  <attributeName>year</att
```

- \* id references must be present

```
<unit>  
  <customUnit>gramsPerUnitPerHour</customUnit>  
</unit>
```

customUnit is a reference  
to an id

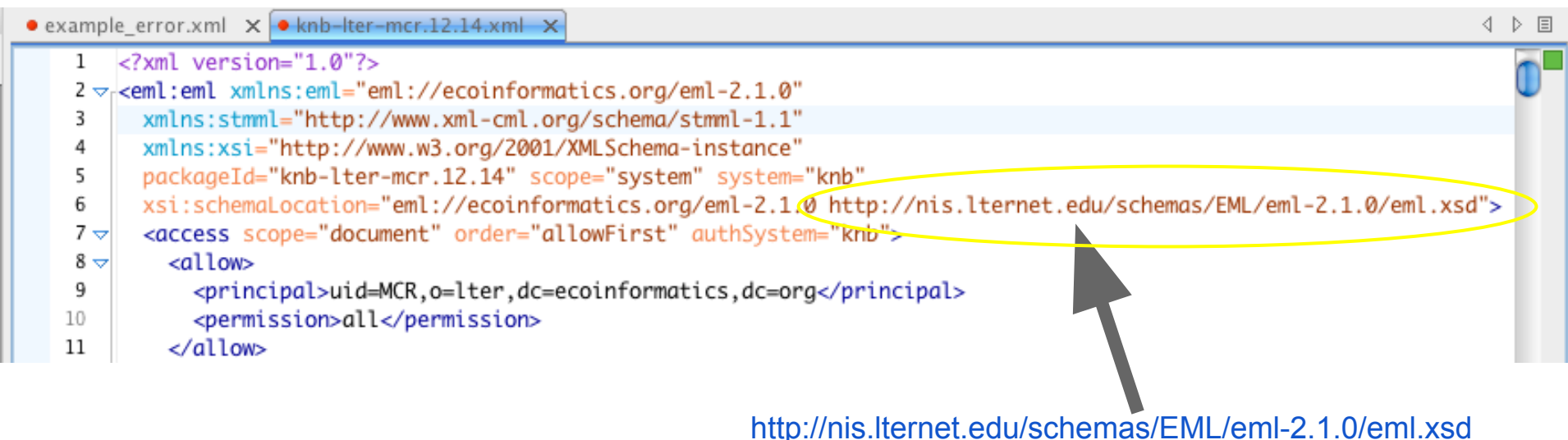
```
<additionalMetadata>  
  <metadata>  
    <unitList>  
      <unit id="gramsPerUnitPerHour"
```

This might not be a  
good unit.

...but parser service errors are not as friendly as xml editor error messages

\* especially if schema is accessible

When the **eml.xsd** file is reachable, error messages and help hints will have useful content.



```
1 <?xml version="1.0"?>
2 <eml:eml xmlns:eml="eml://ecoinformatics.org/eml-2.1.0"
3   xmlns:stmml="http://www.xml-cml.org/schema/stmml-1.1"
4   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
5   packageId="knb-lter-mcr.12.14" scope="system" system="knb"
6   xsi:schemaLocation="eml://ecoinformatics.org/eml-2.1.0 http://nis.lternet.edu/schemas/EML/eml-2.1.0/eml.xsd">
7 <access scope="document" order="allowFirst" authSystem="knb">
8   <allow>
9     <principal>uid=MCR,o=lter,dc=ecoinformatics,dc=org</principal>
10    <permission>all</permission>
11  </allow>
```

<http://nis.lternet.edu/schemas/EML/eml-2.1.0/eml.xsd>

# is it valid?

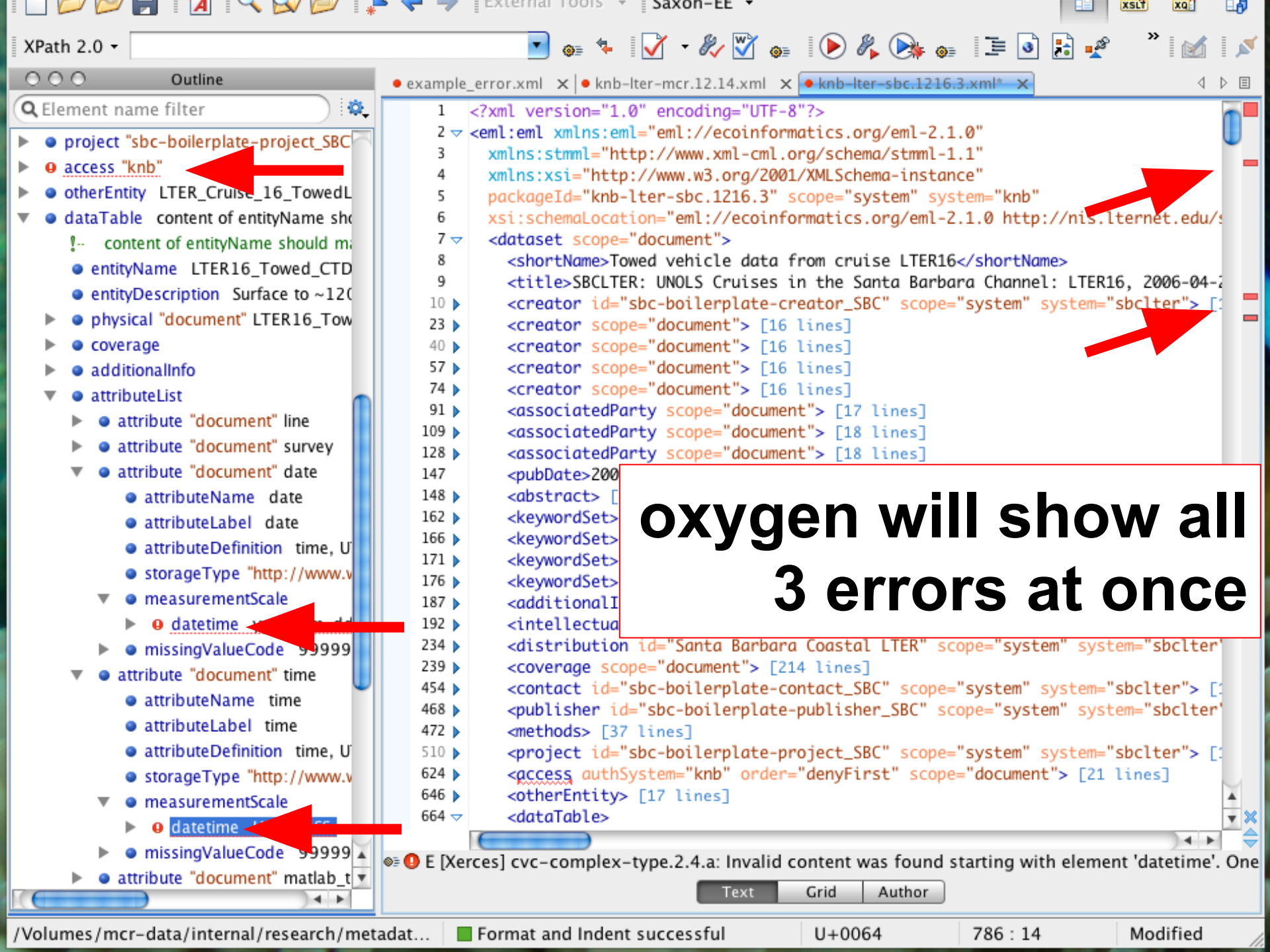
- \* harvest or EML Manager will find errors

Insert failed: cvc-complex-type.2.4.a: Invalid content was found starting with element 'permission'. One of '{principal}' is expected.

- \* but only one-at-a-time







# big outline

- ✓ why would I want to look at eml?
- ✓ what can I see?

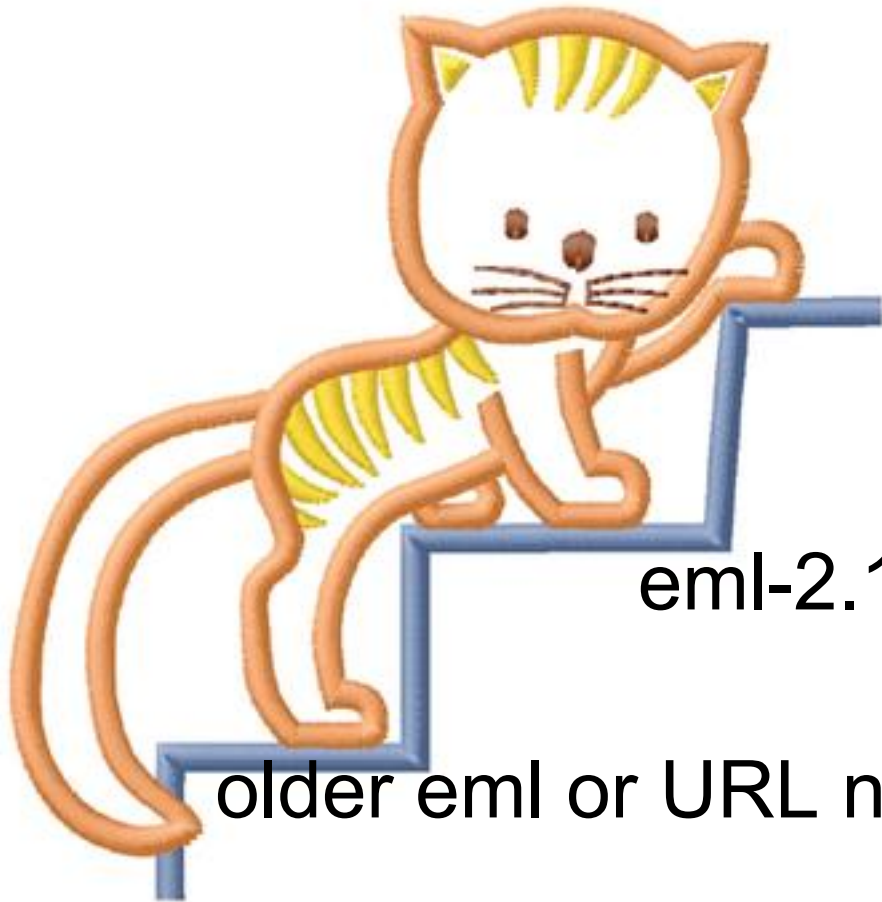
what is expected?

how can I fix it? (manually)

how can I fix it? (script)

the EML Manager

# what is expected?



workflow-ready

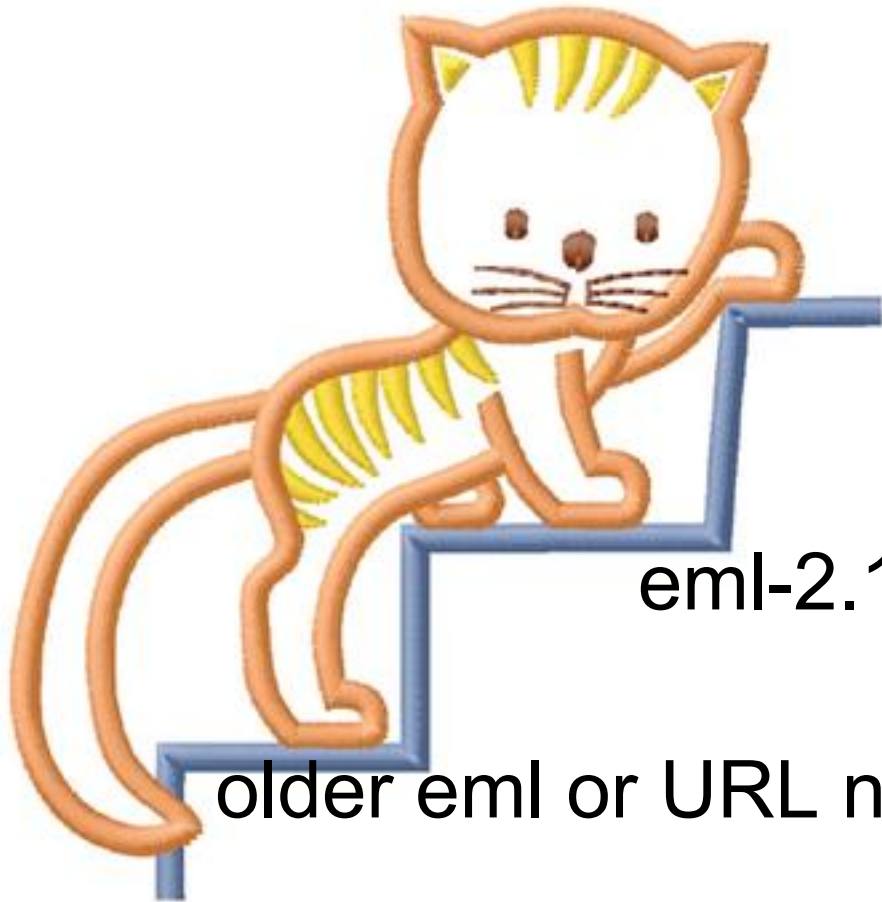
pasta-ready

eml-2.1.0 with data URL

older eml or URL not data download

metadata at site but not in metacat

# what is expected?



workflow-ready

pasta-ready

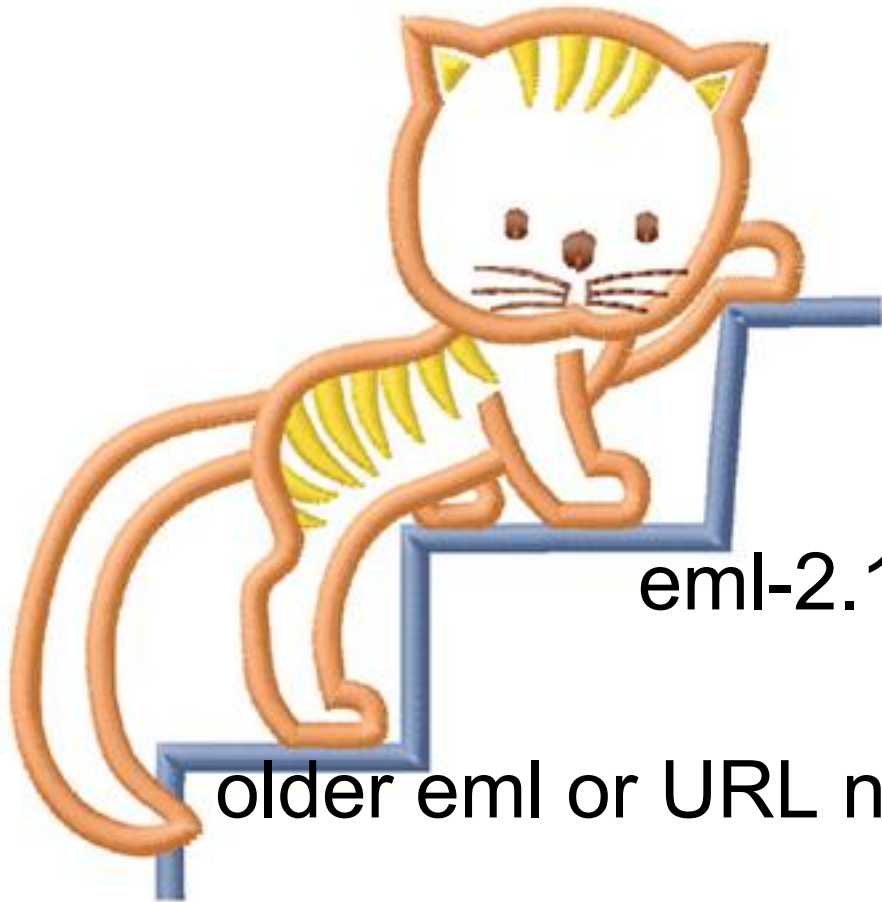
test congruence

eml-2.1.0 with data URL

older eml or URL not data download

metadata at site but not in metacat

# where am I starting?



workflow-ready

pasta-ready

eml-2.1.0 with data URL

older eml or URL not data download · · · today

metadata at site but not in metacat

# a good metric

- \* informs the process
- \* is a natural outcome, not artificial

Let **data use** drive standards

# why do we care?

metacat accommodates bad practice  
pasta does not

to some extent, bad practice has resulted in the  
"perception of data not being available"

short term solution: accommodate variation  
long-term solution: converge practice

# **inventory cannot show intent**

Given a dataset with no data description and/or  
no data access URL...

how to discern "Type II" intentional restriction

from

merely under-described data?



# Type II vs under-described

## Data Set Citation

Moorea Coral Reef LTER and Alldredge A of Moorea Coral Reef LTER. **MCR LTER: Coral Reef: Water Column: Particle sedimentation on the Forereef, Back Reef and Fringing Reef.**

knb-lter-mcr.12.13 (<http://metacat.lternet.edu:8080/knb/metacat/knb-lter-mcr.12.13/lter>).

Metadata download: [Ecological Metadata Language \(EML\) File](#)

## Offline Distribution Info:

Medium: This is an offline dataset (LTER Network Type II). Data may be made available through arrangement with the dataset owner.

## Data Set Owner(s):

Organization: **Moorea Coral Reef LTER**

Address: Marine Science Institute,  
University of California, Santa Barbara,  
Santa Barbara, CA 93106-6150 USA

Web Address: <http://mcr.lternet.edu/>

Individual: **Alice Alldredge**

Organization: Moorea Coral Reef LTER

Position: Principal Investigator

Address: Ecology, Evolution and Marine Biology,  
University of California, Santa Barbara,  
Santa Barbara, CA 93106-9610 USA

Phone: +1(805)893-3997 (voice)

Phone: +1(805)893-4724 (fax)

Email Address: [alldredg@lifesci.ucsb.edu](mailto:alldredg@lifesci.ucsb.edu)

Web Address: <http://www.lifesci.ucsb.edu/eemb/faculty/alldredge/>

## Abstract:

This data package contains measurements of the sedimentation rate of particulate matter to the seafloor on the forereef, backreef and fringing reef of the north shore of Moorea, French Polynesia, during 2 to 4 seasons per year. Measurements include the accumulation rate of particulate organic carbon (POC), particulate organic nitrogen (PON), dry mass, and particulate inorganic carbon on the bottom. Sampling began in August, 2005. Samples were collected for a 24 hour period in triplicate using sediment traps of plastic matting, 1 cm high and 200 cm<sup>2</sup> in area placed at random on the seafloor.

\*\*\* This is an offline dataset (LTER Network Type II). Data may be made available through arrangement with the dataset owner. \*\*\*

## Keywords:

Lets see what is in the eml that makes that appear there...

This is the <offline> element.

This text is just in the abstract. (Not ideal!)

# explicitly offline or restricted as Type II

offline example at dataset level (knb-lter-mcr.12.13)

```
168     </intellectualRights>
169     <distribution>
170       <offline>
171         <mediumName>This is an offline dataset (LTER Network Type II).
172           Data may be made available through arrangement with the dataset owner.</mediumName>
173       </offline>
174     </distribution>
175     <coverage scope="document">
```

restricted example at dataTable level (knb-lter-mcr.4002.11)

```
656     </dataFormat>
657     <!-- gauntlet plus internal dir; this is Type II data. -->
658     <distribution scope="document">
659
660       <online>
661         <onlineDescription>
662           This is an offline dataset (LTER Network Type II).
663           Access with MCR internal account login only until publication.
664           Data may be made available through arrangement with the dataset owner.
665         </onlineDescription>
666         <url function="download">
667           >http://mcr.lternet.edu/data/db/script/database/dbrequest.php?docid=knb-lter-mcr.4002
668         </url>
669       </online>
670     </distribution>
671
672   </physical>
673   <attributeList>
```

# use url function attribute

The url element in eml has an attribute, function, which indicates whether the url points to data, or to information about the data. **Use the function attribute\*.**

`<url function="download">`

- \* streams data

- \* may present browsers with a form

`<url function="information">`

- \* anything other than just-the-data

\*If not specified, default "download" is assumed. Be explicit.

<url function="**download**"> or <url>

- \* zip ok if ONLY the data is zipped by itself
- \* may contain header (as specified)
- \* url returning data described in <physical>

<url function="**information**">

- \* directory
- \* data catalog entry
- \* alternative format of metadata
- \* zip of data plus other files
- \* interactive query tool

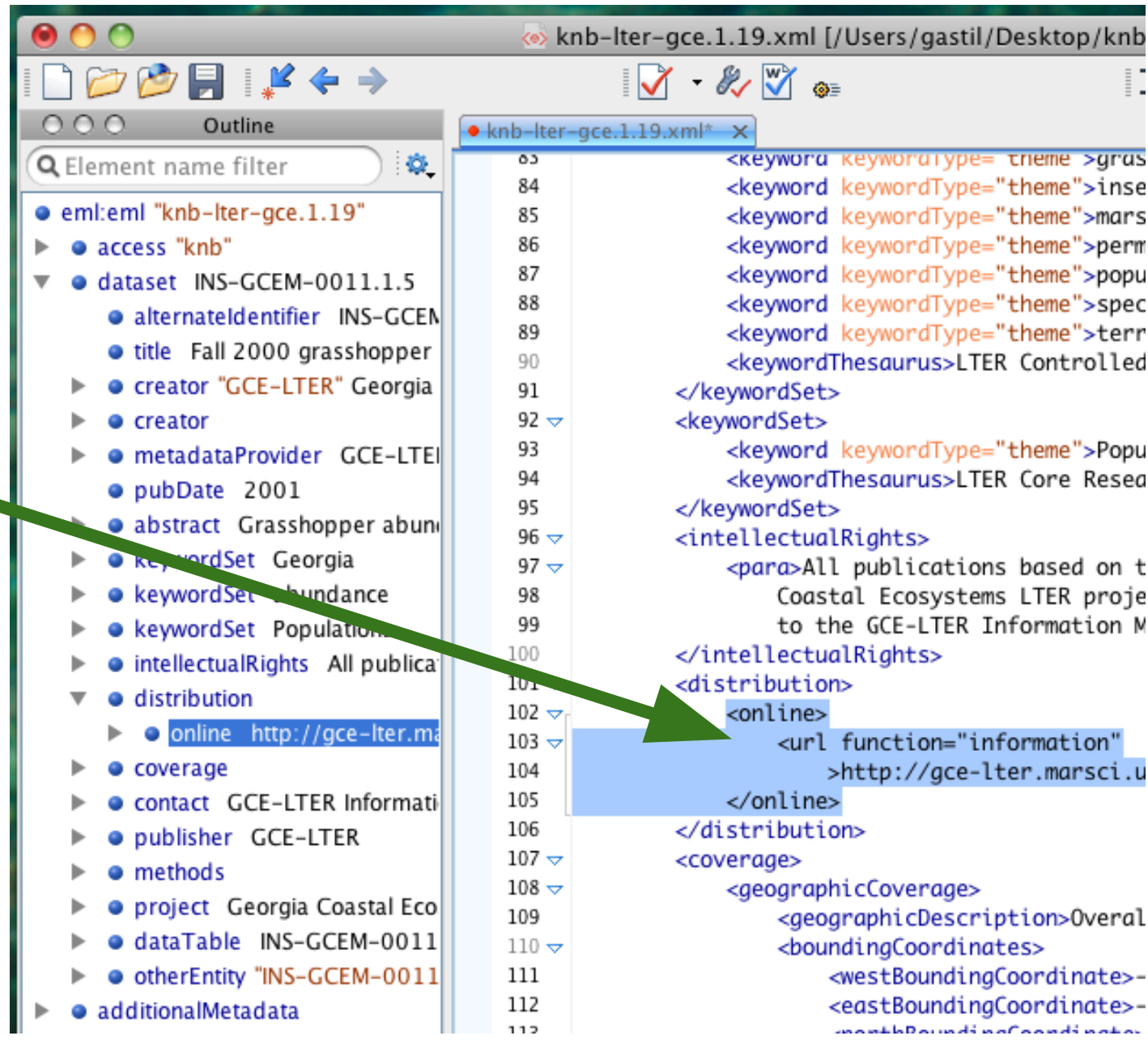
# where the url goes in the eml

- \* data urls go at the "dataTable level"  
includes any entity
- \* informational urls go at the "dataset level"

(It isn't quite that simple, but that captures 80% of it.)

# where the urls go in the eml

<url function="information">  
at dataset level



# where the urls go in the eml

`<url function="download">` at dataTable level

- ▶ keywordSet Populations
- ▶ intellectualRights All publica
- ▶ distribution
- ▶ coverage
- ▶ contact GCE-LTER Informati
- ▶ publisher GCE-LTER
- ▶ methods
- ▶ project Georgia Coastal Eco
- ▼ dataTable INS-GCEM-0011
  - entityName INS-GCEM-0
  - entityDescription Main d
  - ▼ physical INS-GCEM-001
    - objectName INS-GCE
    - size "kilobyte" 8.08
    - characterEncoding AS
    - ▶ dataFormat
    - ▼ distribution
      - online Spreadsheet
      - onlineDescripti
      - url "download"
      - access "kn
    - ▶ methods
    - ▶ attributeList
      - numberOfRecords 350
    - otherEntity "INS-GCEM-0011
  - ▶ additionalMetadata

```
607 <entityDescription>Main data table for data set INS-GCEM-0011</entityDescription>
608 <physical>
609   <objectName>INS-GCEM-0011_1_5.CSV</objectName>
610   <size unit="kilobyte">8.08</size>
611   <characterEncoding>ASCII</characterEncoding>
612   <dataFormat>
613     <textFormat>
614       <numHeaderLines>5</numHeaderLines>
615       <numFooterLines>0</numFooterLines>
616       <recordDelimiter>\r\n</recordDelimiter>
617       <numPhysicallinesPerRecord>1</numPhysicallinesPerRecord>
618       <attributeOrientation>column</attributeOrientation>
619       <simpleDelimited>
620         <fieldDelimiter>,</fieldDelimiter>
621         <quoteCharacter>"</quoteCharacter>
622       </simpleDelimited>
623     </textFormat>
624   </dataFormat>
625   <distribution>
626     <online>
627       <onlineDescription>Spreadsheet comma-separated value (CSV) text file with a fi
628       <url function="download"
629       >http://metacat.lternet.edu/das/dataAccessServlet?docid=knb-lter-gce.1.198
630     </online>
631   </distribution>
632   <access authSystem="knb" order="allowFirst" scope="document">
633     <allow>
634       <principal>uid=GCE,o=lter,dc=ecoinformatics,dc=org</principal>
635       <permission>all</permission>
636     </allow>
637     <allow>
638       <principal>public</principal>
```

# how to get examples

any eml doc in metacat can be downloaded and examined as an example.

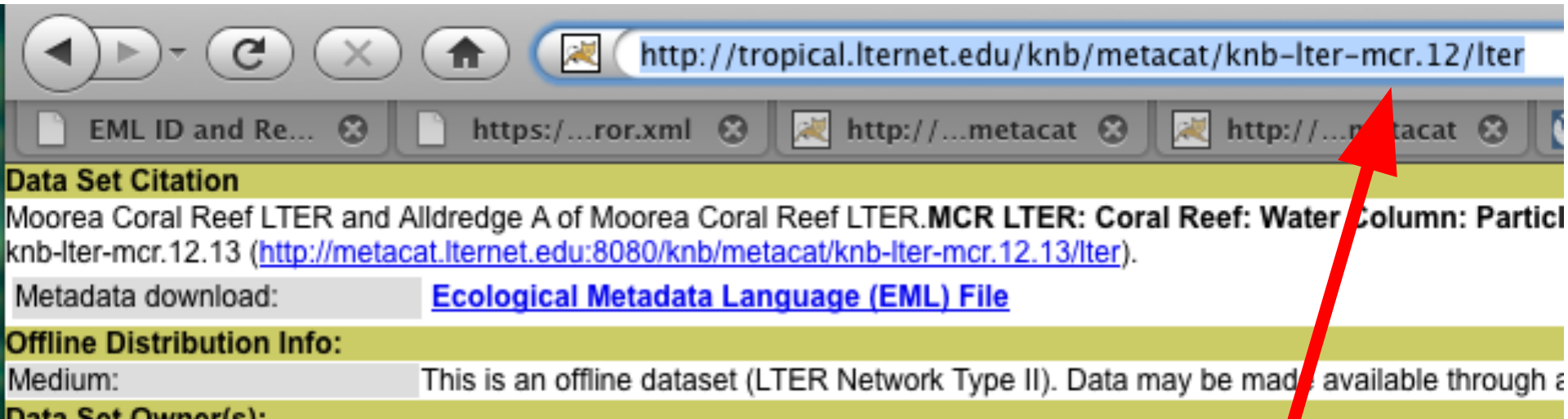
all you need to know is the docId

`knb-lter-abc.identifier number.revision`

(.revision optional)



# how to open any eml from metacat



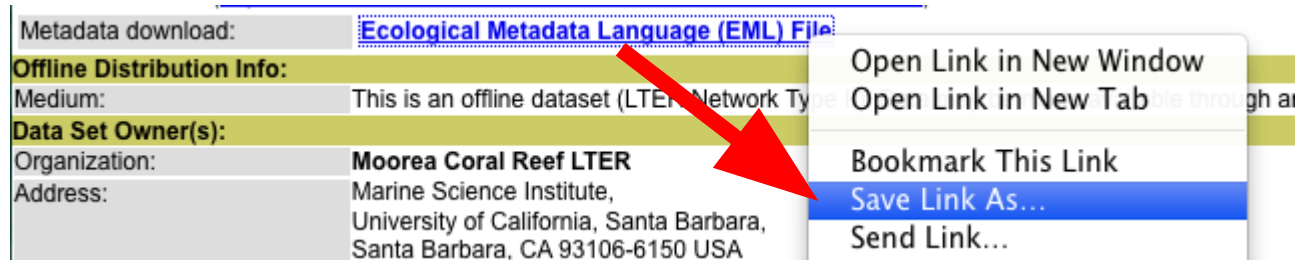
<http://metacat.lternet.edu/knb/metacat/knb-lter-mcr.12/lter>

Edit the scope and identifier to any one you want. ie **sbc.1216** instead of **mcr.12**

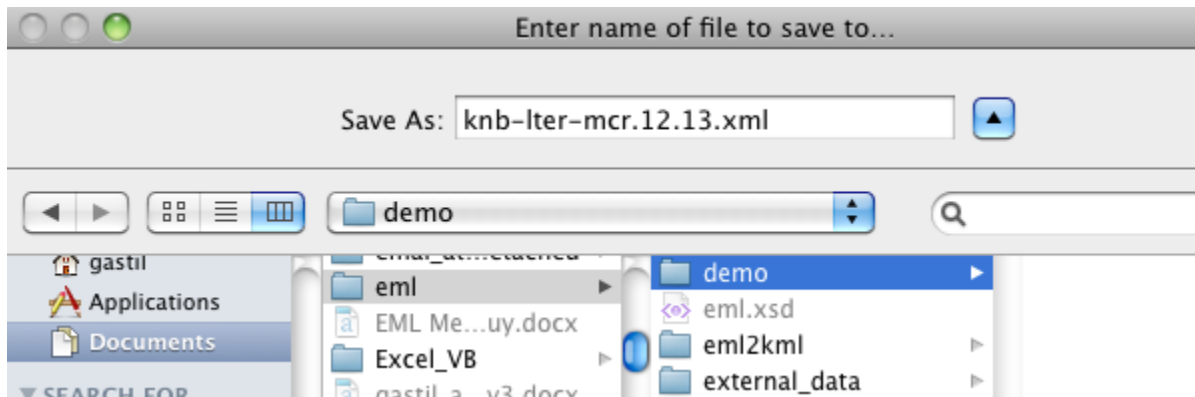
If no revision specified, returns latest.

# how to open any eml from metacat

Either download from browser...  
right-click... Save Link As...



save to local file...

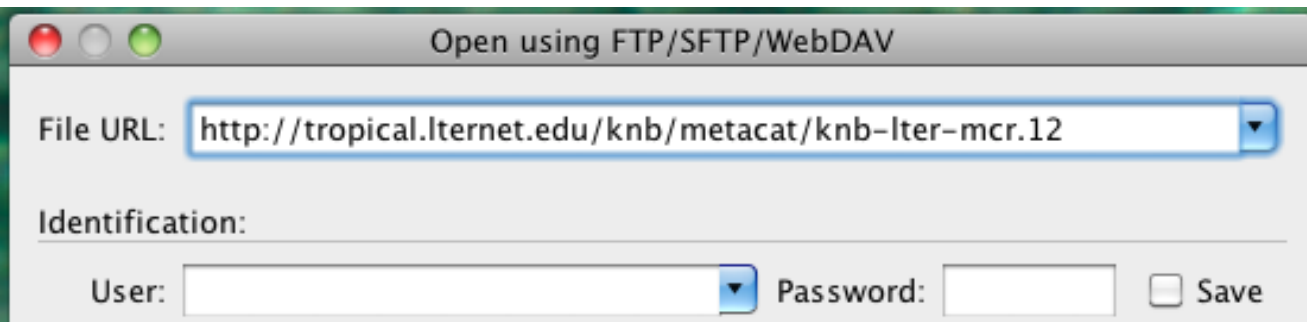
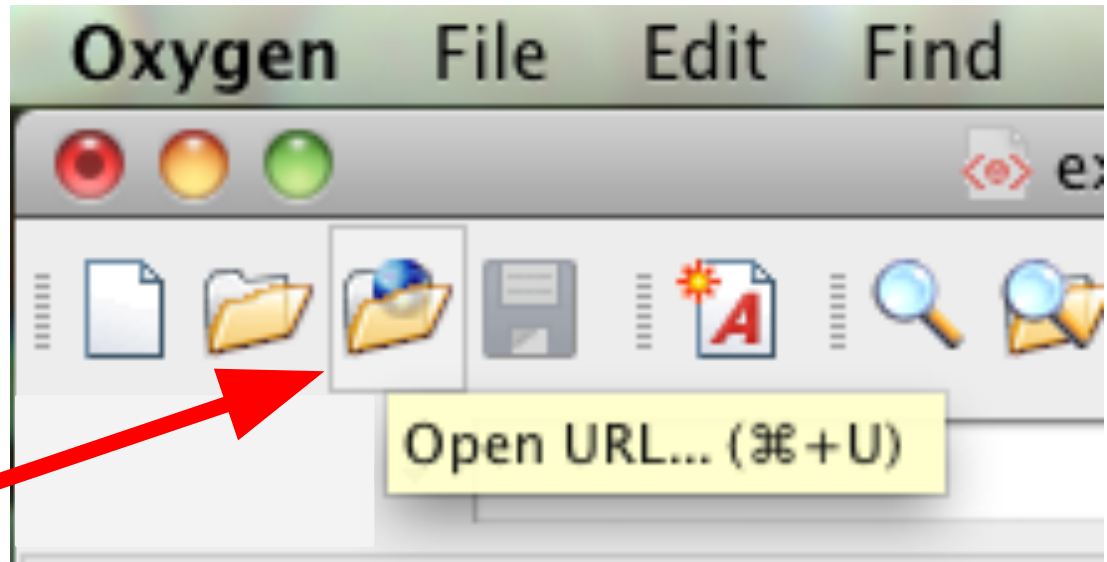


... and *then* open in xml editor. Or...

# how to open any eml from metacat

Or in one step...

Open URL  
in <oXygen>  
xml editor

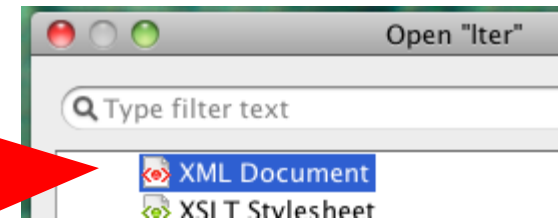


no user or  
password needed  
since eml is  
readable by user  
public.

<http://tropical.lternet.edu/knb/metacat/knb-lter-mcr.12>

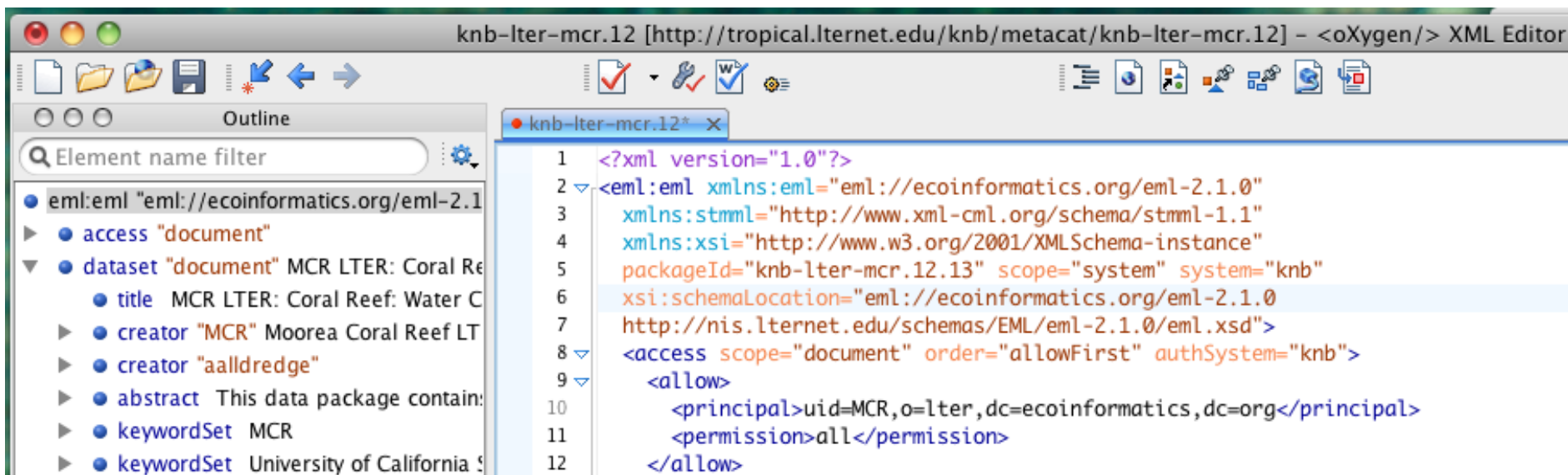
Notice no /lter  
at end of URL

select file type



# how to open any eml from metacat

... and the eml file opens in <oXygen>



To save locally, use File --> Save As...

and name it knbn-lter-xyz.id.rev.xml for your site.id.revision

# how to open any eml from metacat

same result either method.

# the upgrade path

1. version 2.0.1 --> 2.1.0
2. move data url to dataTable level  
(unless it is an information url)
3. completeness *(NOT covered today.)*
4. congruence *(NOT covered today.)*

How do I know if a dataset is WORTH upgrading? *(NOT covered today.)*

# big outline

✓ why would I want to look at eml?

✓ what can I see?

✓ what is expected?

how can I fix it? (manually)

how can I fix it? (script)

the EML Manager

# how can I fix it (manually)

- ✓ how to open any eml doc from metacat
- ✓ is it valid

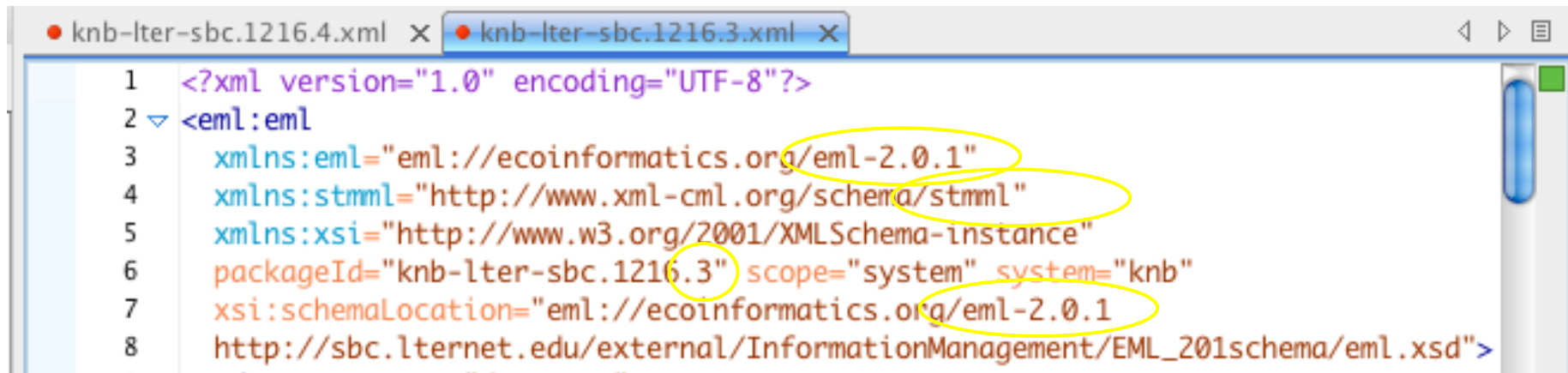
upgrading eml 2.0.1 --> eml 2.1.0

1. root element
2. accessible xsd ✓
3. for each red mark... until all green (valid)
4. check parsing



# upgrading eml 2.0.1 --> eml 2.1.0

root element before

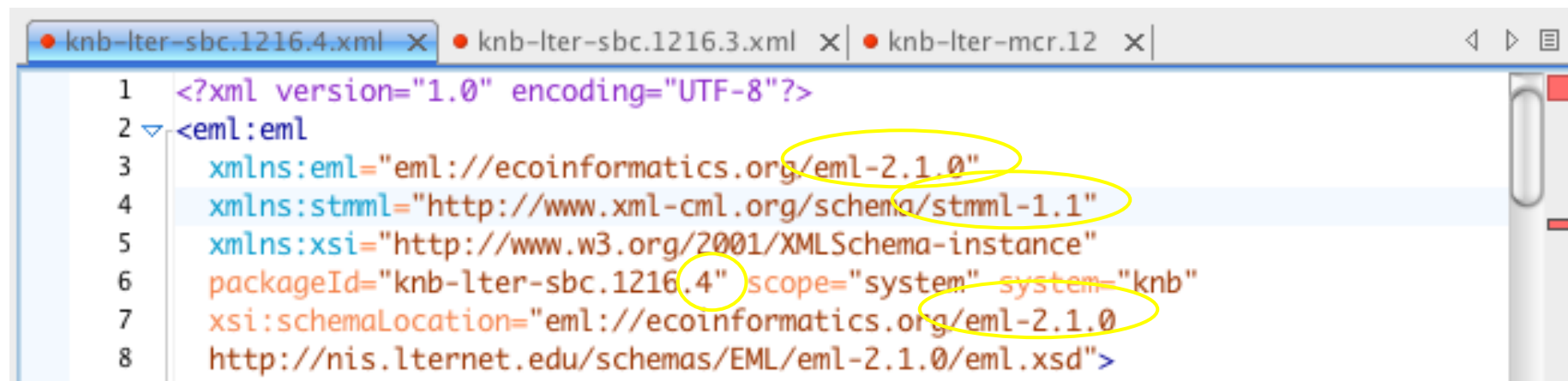


A screenshot of an XML editor window with two tabs: 'knb-lter-sbc.1216.4.xml' and 'knb-lter-sbc.1216.3.xml'. The 'knb-lter-sbc.1216.3.xml' tab is active. The XML content is as follows:

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <eml:eml
3   xmlns:eml="eml://ecoinformatics.org/eml-2.0.1"
4   xmlns:stmml="http://www.xml-cml.org/schema/stmml"
5   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
6   packageId="knb-lter-sbc.1216.3" scope="system" system="knb"
7   xsi:schemaLocation="eml://ecoinformatics.org/eml-2.0.1
8   http://sbc.lternet.edu/external/InformationManagement/EML_201schema/eml.xsd">
```

Yellow circles highlight the following values in the original image: 'eml-2.0.1', 'http://www.xml-cml.org/schema/stmml', 'knb-lter-sbc.1216.3', and 'eml://ecoinformatics.org/eml-2.0.1'.

root element after



A screenshot of an XML editor window with three tabs: 'knb-lter-sbc.1216.4.xml', 'knb-lter-sbc.1216.3.xml', and 'knb-lter-mcr.12'. The 'knb-lter-sbc.1216.4.xml' tab is active. The XML content is as follows:

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <eml:eml
3   xmlns:eml="eml://ecoinformatics.org/eml-2.1.0"
4   xmlns:stmml="http://www.xml-cml.org/schema/stmml-1.1"
5   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
6   packageId="knb-lter-sbc.1216.4" scope="system" system="knb"
7   xsi:schemaLocation="eml://ecoinformatics.org/eml-2.1.0
8   http://nis.lternet.edu/schemas/EML/eml-2.1.0/eml.xsd">
```

Yellow circles highlight the following values in the original image: 'eml-2.1.0', 'http://www.xml-cml.org/schema/stmml-1.1', 'knb-lter-sbc.1216.4', and 'eml://ecoinformatics.org/eml-2.1.0'.

# upgrading eml 2.0.1 --> eml 2.1.0

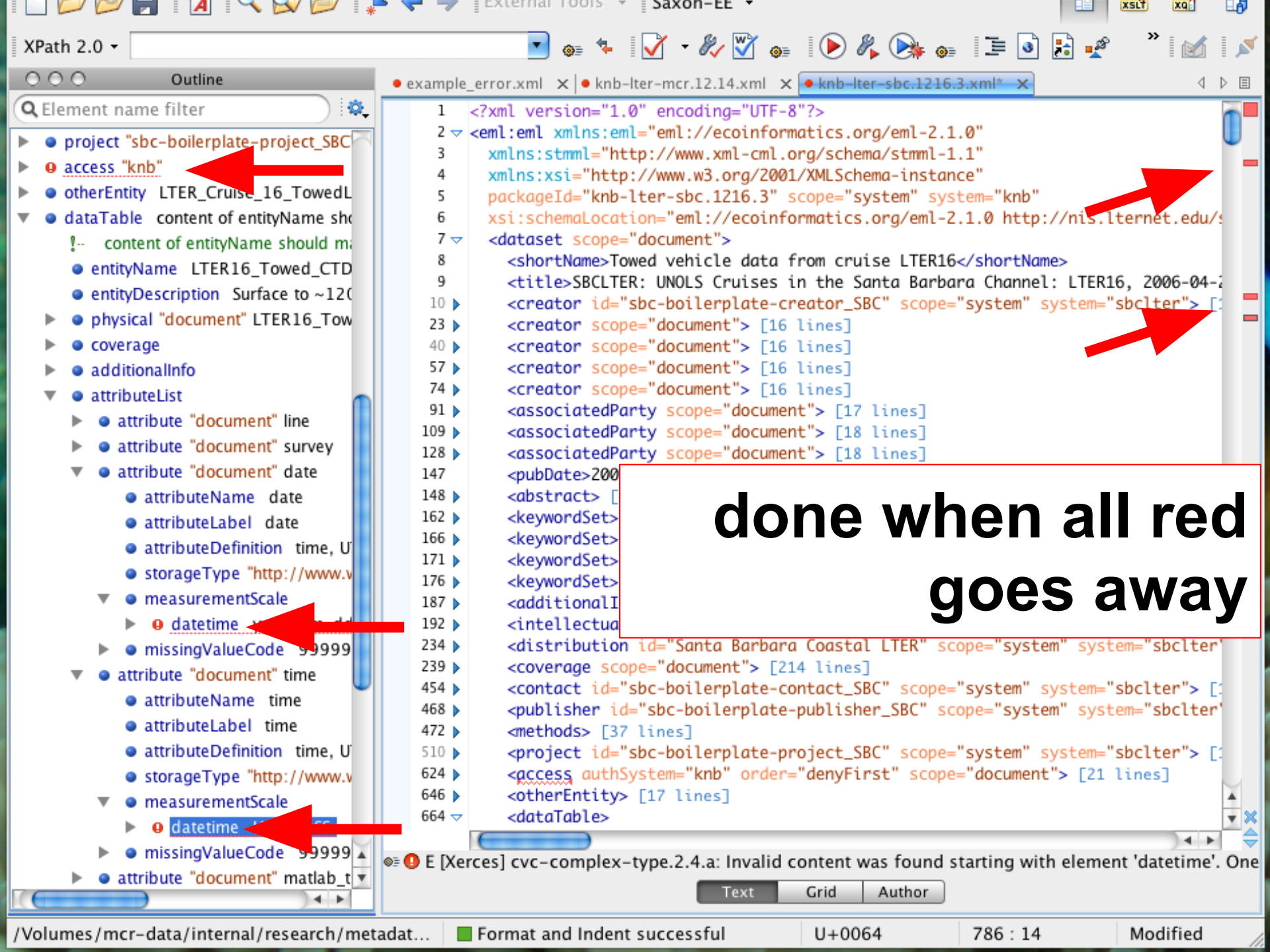
1. move <access> up above <dataset>

2. `<datetime>` becomes `<dateTime>`

3. <method> becomes <methods>

4. no empty tags; no self-closing tags  hard part

If the eml 2.0.1 was complete,  
upgrade is mechanical



done when all red  
goes away

# big outline

- ✓ why would I want to look at eml?
- ✓ what can I see?
- ✓ what is expected?
- ✓ how can I fix it? (manually)  
how can I fix it? (script)

the EML Manager

## **how can I fix it? (script)**

Even if my eml is generated from a program, it helps to know what I want the end product to look like.

To know what to ask the script to do, it helps to upgrade at least one eml document manually, first.

Then compare the script-generated eml for that same dataset to the manually edited version.

## how can I fix it? (script)

Often xslt is used at least in the end step.

Any non-required element, if it has no content, wrap the xslt with for-each, or choose/when, or if. Only write out elements that have content.

If an element is required and has no content, that content needs to be added, ie to database table or input file.

## how can I fix it? (script)

... more xslt

Remove <literalLayout>

Use <listItem> if that was the intent.

Trim leading and trailing whitespace from content and avoid introducing unintentional whitespace.

Ask a site that has already upgraded their xslt to see their refactoring.

# how can I fix it? (script)

after fixing xslt... view the output

View eml document in <oXygen>,

use the handy wrapper-indenter thingie



To compare before and after, use

Tools -->





# compare eml docs

before

after

Diff Files

file:/Volumes/mcr-data/internal/research/metadata/sbc/knb-lter-sbc.1216.3.xml

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <eml:eml
3   xmlns:eml="eml://ecoinformatics.org/eml-2.0.1"
4   xmlns:stxml="http://www.xml-cml.org/schema/stxml"
5   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
6   packageId="knb-lter-sbc.1216.3" scope="system" system="knb"
7   xsi:schemaLocation="eml://ecoinformatics.org/eml-2.0.1
8     http://sbc.lternet.edu/external/InformationManagement/EML_201sche
9   <dataset scope="document">
10     <shortName>Towed vehicle data from cruise LTER16</shortName>
11     <title>SBCLTER: UNOLS Cruises in the Santa Barbara Channel: LTER
12     <creator id="sbc-boilerplate-creator_SBC" scope="system" system
13       <organizationName>SBCLTER</organizationName>
14       <address scope="document">
15         <deliveryPoint>Santa Barbara Coastal LTER</deliveryPoint>
16         <deliveryPoint>Marine Science Institute</deliveryPoint>
17         <deliveryPoint>University of California</deliveryPoint>
18         <city>Santa Barbara</city>
19         <administrativeArea>California</administrativeArea>
20         <postalCode>93106-6150</postalCode>
21         <country>United States</country>
22       </address>
23       <onlineUrl>http://sbc.lternet.edu</onlineUrl>
24     </creator>
25     <creator scope="document">
26       <individualName>
27         <givenName>Mark</givenName>
```

file:/Volumes/mcr-data/internal/research/metadata/sbc/knb-lter-sbc.1216.4.xml

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <eml:eml
3   xmlns:eml="eml://ecoinformatics.org/eml-2.1.0"
4   xmlns:stxml="http://www.xml-cml.org/schema/stxml-1.1"
5   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
6   packageId="knb-lter-sbc.1216.4" scope="system" system="knb"
7   xsi:schemaLocation="eml://ecoinformatics.org/eml-2.1.0
8     http://nis.lternet.edu/schemas/EML/eml-2.1.0/eml.xsd">
9   <access authSystem="knb" order="denyFirst" scope="document">
10     <allow>
11       <principal>uid=sbc,o=LTER,dc=ecoinformatics,dc=org</principal>
12       <permission>all</permission>
13     </allow>
14     <allow>
15       <principal>public</principal>
16       <permission>read</permission>
17     </allow>
18   </access>
19   <dataset scope="document">
20     <shortName>Towed vehicle data from cruise LTER16</shortName>
21     <title>SBCLTER: UNOLS Cruises in the Santa Barbara Channel: LTER
22     <creator id="sbc-boilerplate-creator_SBC" scope="system" system=
23       <organizationName>SBCLTER</organizationName>
24       <address scope="document">
25         <deliveryPoint>Santa Barbara Coastal LTER</deliveryPoint>
26         <deliveryPoint>Marine Science Institute</deliveryPoint>
27         <deliveryPoint>University of California</deliveryPoint>
```

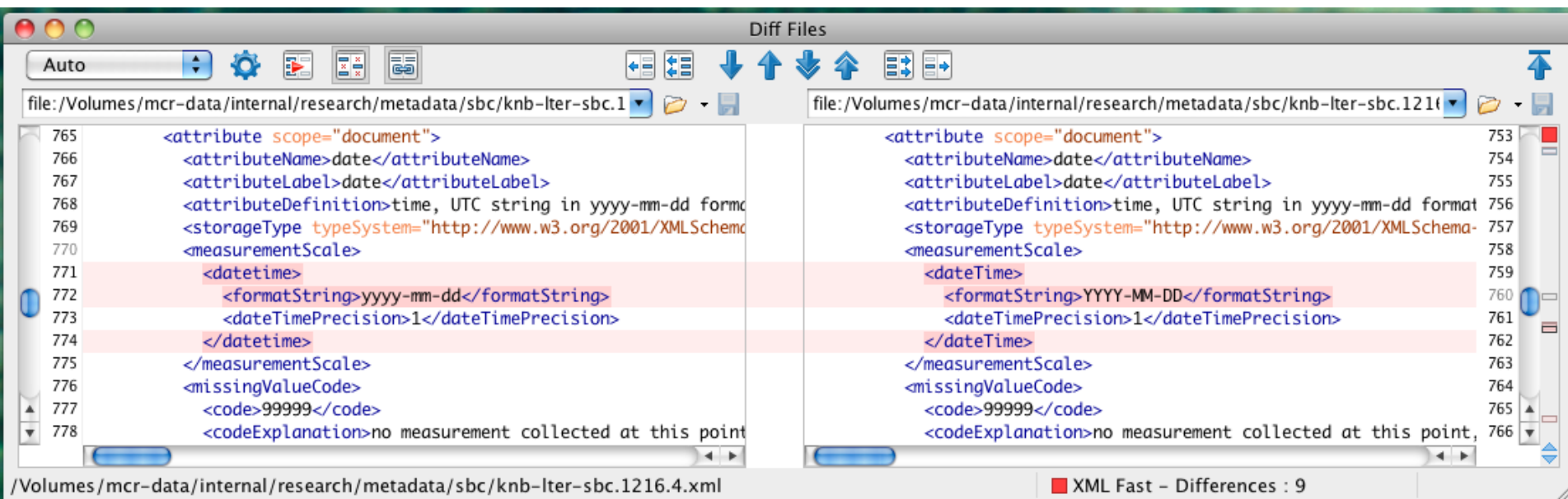
/Volumes/mcr-data/internal/research/metadata/sbc/knb-lter-sbc.1216.4.xml

XML Fast - Differences : 9

# compare eml docs

before

after



That one was as easy as it gets.  
Your mileage may vary.

# big outline

- ✓ why would I want to look at eml?
- ✓ what can I see?
- ✓ what is expected?
- ✓ how can I fix it? (manually)
- ✓ how can I fix it? (script)

the EML Manager

# is it ready for metacat?

is it **valid**?

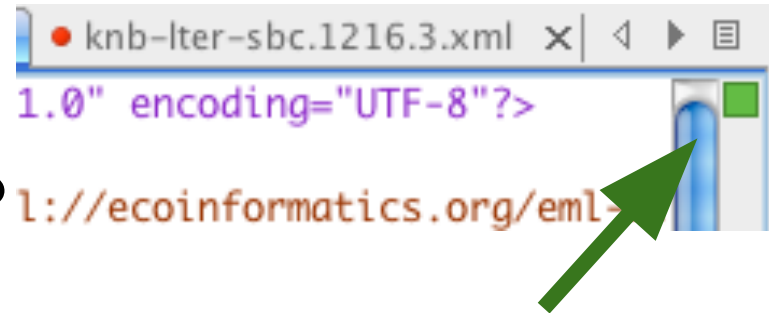
- \* green box in <oXygen>?
- \* or passes knb validator

does it **parse**?

- \* passes knb parser

does it pass the self-service one-doc-at-a-time checker?

- \* coming soon



# the knb parser

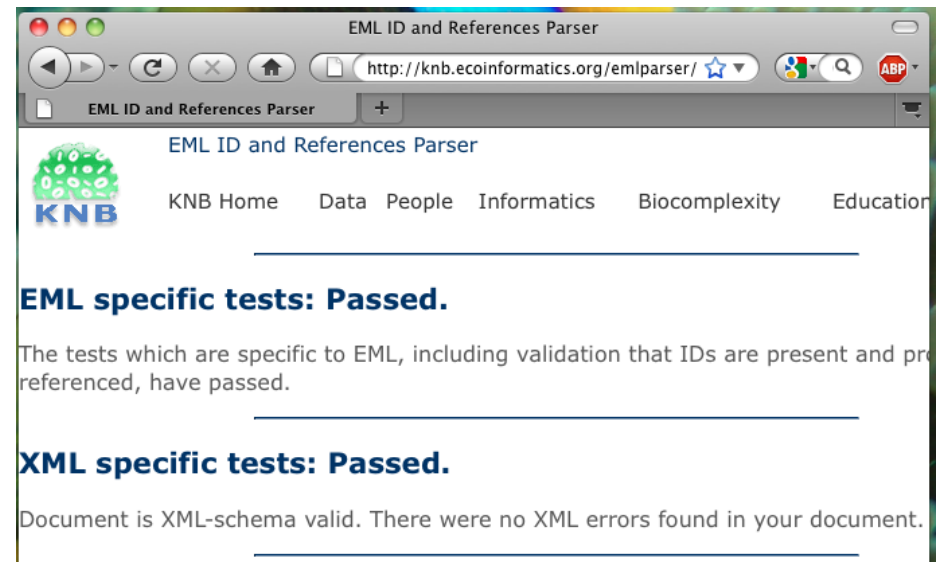
<http://knb.ecoinformatics.org/emlparser/>

File Name:

1

2

3



**why wait?**

The **EML Manager**  
is an alternative to the Harvest.



If you just want to  
insert or update  
one EML doc  
**NOW.**



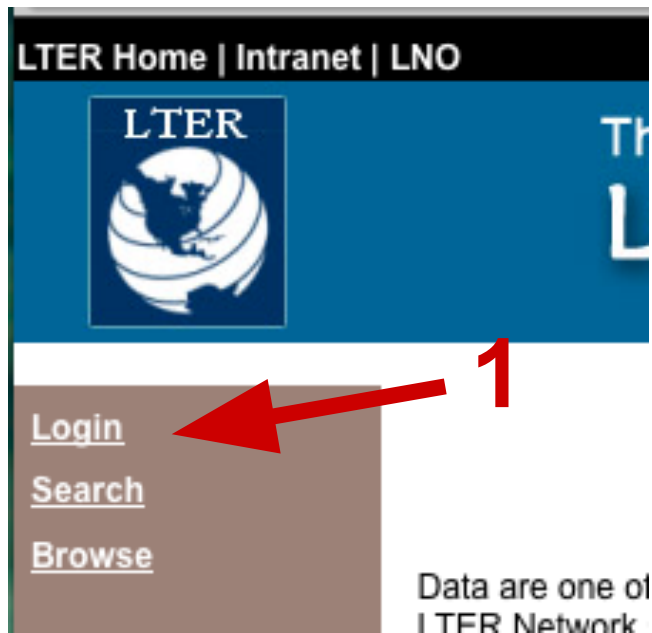
Some cats get impatient.

# EML Manager

interactive, manual harvest of one eml doc

1. log into metacat **as your site** not yourself

<http://metacat.lternet.edu/das/lter/>

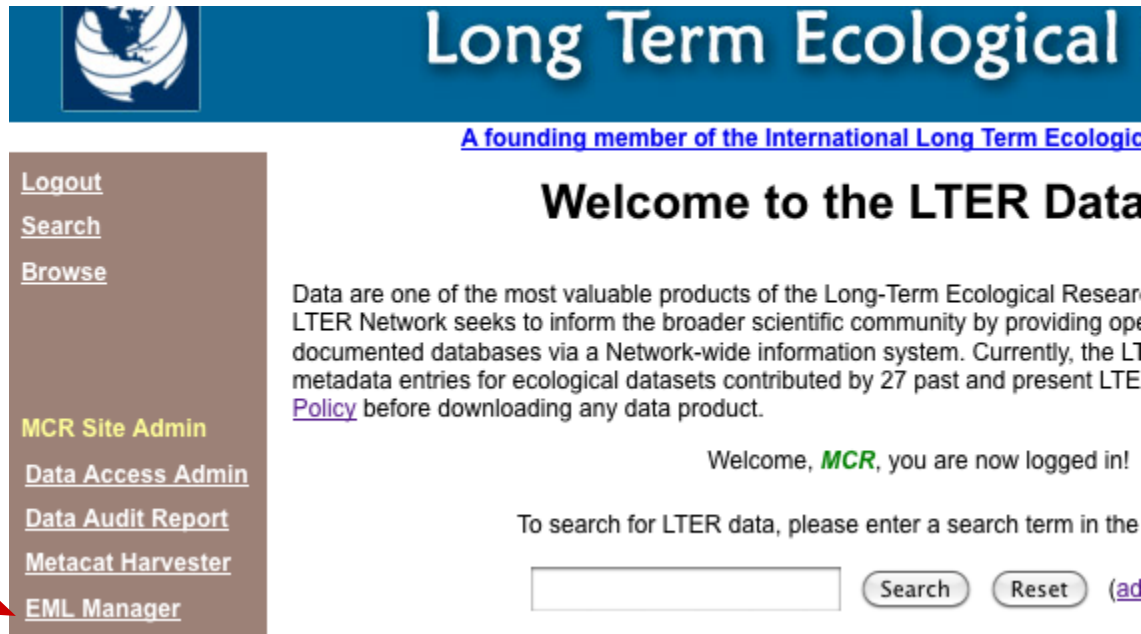


The screenshot shows a login form with the following fields and controls:

- User name:** A text input field containing the text "MCR". A red arrow points from the text "site acronym" to this field.
- Affiliation:** A dropdown menu with "LTER" selected.
- Password:** A text input field with masked characters ".....".
- Login** and **Reset** buttons.



# EML Manager



Long Term Ecological

[A founding member of the International Long Term Ecological](#)

**Welcome to the LTER Data**

Data are one of the most valuable products of the Long-Term Ecological Research. The LTER Network seeks to inform the broader scientific community by providing open access to documented databases via a Network-wide information system. Currently, the LTER Network has over 100 metadata entries for ecological datasets contributed by 27 past and present LTER sites. Please read the [Policy](#) before downloading any data product.

Welcome, **MCR**, you are now logged in!

To search for LTER data, please enter a search term in the

[\(ad](#)

## Data Portal EML Manager

**Insert** To *insert* a new EML document, enter a DocID and a valid EML File, and click Insert.

**Update** To *update* an existing EML document, enter a DocID and a valid EML File, and click Update.

**Delete** To *delete* an EML document, enter only a DocID (revision optional), and click Delete.

For more information on - [DocIDs](#)

DocID:

EML File Upload:

3



# harvesting

- ✓ when eml generation is "in production"
- ✓ when reasonably sure eml is valid & parses
- ✓ when updates are mechanical

then harvest

see <http://im.lternet.edu/node/418>

# The good-enough data package

