Best Practices for PASTA

Technical Aspects - 2013

2013 May 6/7

Take home message

Most likely, whatever problem you see has been seen before.

Ask Sven or Margaret (2013 consultants)

Production vs Staging

Production: the public facing PASTA and NIS Data Portal

portal.lternet.edu

Staging: a testing environment for LTER information managers

portal-s.lternet.edu

Best Practice:

look at your EML transformed in HTML before you harvest to production, please

Put your harvest-list in portal-s FIRST

an EML consultant is available to review a package in portal-s with you

Attribute Names

Best practice:

attributeName should be something that can act as a variable, ie, keep them alphanumeric, short but descriptive, avoid spaces, lowercase is suggested!

attributeLabel can use other characters, descriptive and expressive.

Good practice:	Bad:
n_and_n pri_prod	N&N pri prod, grams/m^2/h

Characters

Best practice:

- Declare UTF-8 in your EML prolog, and stick to it
- Be aware that copy-paste from Microsoft can cause problems

New Check coming will alert submitters to problems

Examples of packages in staging

Example of EML with a character that is illegal HTML:

https://portal-s.lternet.edu/nis/metadataviewer?packageid=knb-lter-sbx.1.1

Same EML, no illegal characters (better …):

https://portal-s.lternet.edu/nis/metadataviewer?packageid=knb-lter-sbx.1.2

Finished:

https://portal-s.lternet.edu/nis/metadataviewer?packageid=knb-lter-sbx.1.5

Evaluate vs. Upload

Evaluate mode shows you more checks

You can evaluate your entire harvest list

Why evaluate first?

Ensures higher quality in production PASTA

You have more control

Iterative editing is faster

DOIs

Each package has a DOI (Digital Object Identifier)

- Permanent
- Assigned hourly
- DOIs are NOT in the EML
- DOI resolver points to portal, "resource map" display (more later)
- what the general public will use, e.g., to cite in a paper

Examples. the doi itself starts with ‘10.’

usually given: DOI:10.6073/pasta/d6032b2b81a44b20511627ba90254fa4

with the resolver:

http://dx.doi.org/10.6073/pasta/d6032b2b81a44b20511627ba90254fa4

PASTA resource map

The manifest (URLs) for all the parts of the package in PASTA

Example for knb-lter-nin.1.1:

Data entity(s): https://pasta.lternet.edu/package/data/eml/knb-lter-nin/1/1/
DailyWaterSample-NIN-LTER-1978-1992

XML Metadata: https://pasta.lternet.edu/package/metadata/eml/knb-lter-nin/1/1

Report: https://pasta.lternet.edu/package/report/eml/knb-lter-nin/1/1

the Resource Map itself: https://pasta.lternet.edu/package/eml/knb-lter-nin/1/1

Status updates

XML storage for PASTA

Usability engineer to work with the Data Portal Tiger Team

Questions?

Down in the Weeds

More details for those interested

Character sets

What is the site responsibility, what is the code's?

PASTA Code

Standardized on UTF-8, which supports all known languages.

Include a check for the quality engine to alert submitters to non-uft8 chars. (warn)

Sites

Declare UTF-8, and make sure EML complies.

How?

- Avoid copy-paste from Microsoft products
- If your ‘system’ does allow copy-paste, it should export in UTF-8

Examples

- test docs in portal-s:

- a list of the offending characters (from Windows) http://www.cs.tut.fi/~jkorpela/www/windows-chars.html

Status of checker reports

You need to know how many checks to expect.
number of checks to look for: 27
13 Dataset-level
14 dataTable *(6 for otherEntity, spatialRaster, spatialVector)*

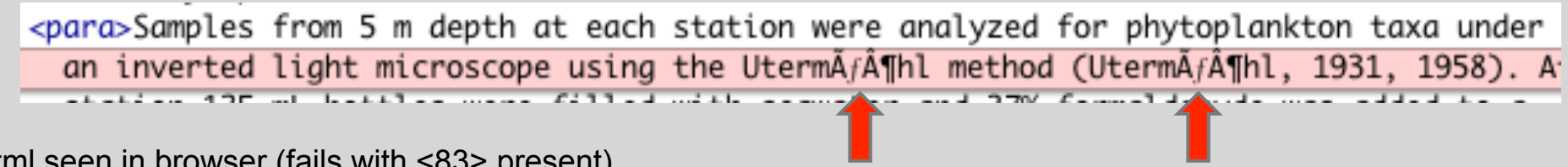
what if you see fewer?
explore why.

list new checks that have been suggested, or the next ones that will be implemented

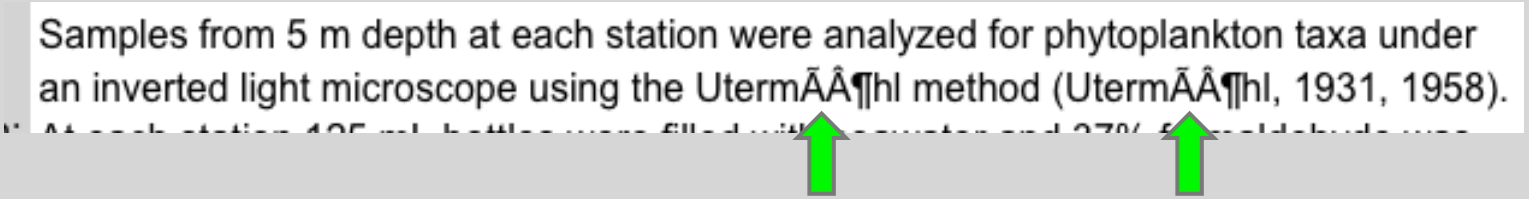
- Character sets

odd characters (weeds)
Example: <83> = *f* = 302 203

oXygen



html seen in browser (fails with <83> present)



vi

od -c

```
[gastilbuhl@adelie sbx]$ grep "an inverted light microscope using" knb-lter-sbx.1.1.xml | cut -c78-86 | od -c
0000000   e   r   m 303 203 302 203 303 202 302 266   h   l  \n
0000016
[gastilbuhl@adelie sbx]$ grep "an inverted light microscope using" knb-lter-sbx.1.2.xml | cut -c78-86 | od -c
0000000   r   m 303 203 303 202 302 266   h   l   ,     \n
0000015
```