Elog - Support #7651

Import IOTA/ASTA (NML) ECL elog entries and comments

01/15/2015 04:21 PM - Kyle Hazelwood

Status:	Assigned	**Start date:**	01/15/2015
Priority:	High	**Due date:**	04/30/2015
Assignee:	Kyle Hazelwood	**% Done:**	0%
Category:	Database	**Estimated time:**	48.00 hours
Target version:		**Spent time:**	56.00 hours

Description

Once a IOTA/ASTA logbook exists we will import the old NML ECL entries and comments. This may take sometime to implement.

Related issues:

Related to Elog - Support #6284: Transfer PXIE ECL logbook entries into elog	**Assigned**	**05/15/2014**
Related to Electronic Collaboration Logbook - Bug #7959: API fails on unsanit...	**Resolved**	**02/27/2015**

History

#1 - 01/15/2015 04:36 PM - Kyle Hazelwood

- Related to Support #6284: Transfer PXIE ECL logbook entries into elog added

#2 - 01/19/2015 09:12 AM - Kyle Hazelwood

- Estimated time set to 48.00 h

Here's the link to the ECL XML API.

#3 - 01/19/2015 01:29 PM - Kyle Hazelwood

As of today there are 8266 NML ECL logbook entries to import.

The scheme for importing all the entries is:

- Retrieve all the entry ids that exist using link: http://dbweb4.fnal.gov:8080/ECL/nml/E/xml_search?o=ids
- Retrieve each entry individually using link: http://dbweb4.fnal.gov:8080/ECL/nml/E/xml_get?e=<entry id>
- Parse the ECL API entry XML into an ElogECLEntry object (with sub-objects for tags, comments, and attachments)
- Append the log name to the ElogECLEntry object
- Convert the ElogECLEntry object into an ElogEntry object using a toElogEntry() method in the ElogECLEntry class. This method should resolve the Elog log, file, author, and category properties
- Send the ElogEntry object to the db. This will give the entry a recent entry id but will preserve the timestamp so sorting will not be affected

I need to do the following:

- Create an ElogECLEntry class
- Create an ElogECLTag class
- Create an ElogECLComment class
- Create an ElogECLAttachment class
- Create a method in ElogDB to import the ECL entries using ElogECLEntry objects. This method needs to write the timestamp in the db instead of letting the db pick the time.
- Create a robot thread to add these entries at a reasonable throttle time. We could possibly keep this running for sometime to pick up any entries that are put in the ECL NML logbook after the initial import

#4 - 01/19/2015 02:00 PM - Kyle Hazelwood

Thankfully the ECL XML entry API provides urls to the file attachments!

#5 - 02/09/2015 01:31 PM - Kyle Hazelwood

- File ECLUsers.xlsx added

Unfortunately ECl doesn't require their account usernames correlate to any other service (such as the services account). Their usernames don't necessarily match the services accounts that the elog uses. This makes importing entries much more difficult. Thankfully Chip Edstrom has created a spreadsheet of usernames for me to use to match users. Also, some ECL users have multiple usernames. The spreadsheet does include their email address which will be helpfull.

#6 - 02/09/2015 02:07 PM - Kyle Hazelwood

Good example with comments: http://dbweb4.fnal.gov:8080/ECL/nml/E/xml_get?e=8279

#7 - 02/09/2015 02:08 PM - Kyle Hazelwood

Good example with files: http://dbweb4.fnal.gov:8080/ECL/nml/E/xml_get?e=8000

#8 - 02/09/2015 02:53 PM - Kyle Hazelwood

- File ECLEntryComments.PNG added

- File ECLEntryFiles.PNG added

#9 - 02/09/2015 03:07 PM - Kyle Hazelwood

ECL offers a "related to" feature that essentially links entries related to each other. We've been wanting to incorporate this feature into our elog for sometime but haven't found the time. The NML elog uses this feaure slightly so I think we could afford to wait until the feature exists in our elog and then manually link entries.

#10 - 02/10/2015 01:12 PM - Kyle Hazelwood

I need to create a field in the entries and comments tables to indicate whether the entry/comment was imported to the elog from another source for statistical reasons.

#11 - 02/10/2015 01:13 PM - Kyle Hazelwood

Is it important to save the ECL entry id?

#12 - 02/10/2015 02:32 PM - Kyle Hazelwood

ECL must use the entry ids for all logbooks because the per log entry ids are not sequential.

#13 - 02/10/2015 02:35 PM - Kyle Hazelwood

ECL supports multiple "Tag"s which I think are analagos to our "Subject" tag. We only allow one "Subject" tag per entry. We will have to just take the first tag and ignore the rest. At a glimpse this will not affect many entries.

#14 - 02/25/2015 02:17 PM - Kyle Hazelwood

I've added the "is_imported" field to elog tables "comments", "files" and "entries". No more db table manipulation should be necessary for the import.

#15 - 02/26/2015 08:02 PM - Kyle Hazelwood

Apparently ECL doesn't sanitize their XML... http://dbweb4.fnal.gov:8080/ECL/nml/E/xml_get?e=8350

#16 - 02/26/2015 08:08 PM - Kyle Hazelwood

Kyle Hazelwood wrote:

> Apparently ECL doesn't sanitize their XML... http://dbweb4.fnal.gov:8080/ECL/nml/E/xml_get?e=8350

Chip put quotes in the file caption! Hope this doesn't mean that the ECL API will fail whenever someone puts quotation marks in their entries.

#17 - 02/26/2015 11:08 PM - Kyle Hazelwood

Prior to ~2011, ECL attachments would use the word "None" in place of a valid index number. 2012 and on the index numbers start at 0 apparently. This is very confusing for parsing.

#18 - 02/27/2015 12:31 AM - Kyle Hazelwood

- File ElogECLImportThread_FailedEntries.log added

Tested out the ECL import thread tonight... its not ready.

- Of the ~7700 entries that need to be imported ~800 failed.
- Of the ~800 that failed ~60 failed during retrieval on the ECL server due to poorly sanitized XML. A redmine issue has been made for ECL.
- The remainder of the entries that failed did so because the script did not account for a change in the ECL data format for attachment index "numbers". Index numbers were assumed to be integers which all are after ~2011, prior to that they can be the word "None".

#19 - 02/27/2015 12:44 AM - Kyle Hazelwood

- Related to Bug #7959: API fails on unsanitized XML added

#20 - 02/27/2015 11:10 PM - Kyle Hazelwood

- File ElogECLImportThread_FailedEntries2.log added

I handled the ECL entry attachment String|Integer problem with an XML adapter class. After doing this I only have 70 poorly formatted XML entries to deal with.

#21 - 03/02/2015 06:49 PM - Kyle Hazelwood

ECL offers the ability to use pre-made forms in their logbooks. NML didn't use these forms very much. There appears to be only one form for NML and only about one dozen entries were made using this form. None of these forms have been used in the NML logbook since 2011. These entries will not transfer over very well.

#22 - 03/02/2015 08:05 PM - Kyle Hazelwood

http://dbweb4.fnal.gov:8080/ECL/nml/A/xml_category_list gives me the list of catgories used in the NML logbook (the xml is not sanitized). There are 83 categories as they are named. I can probably split this ?/?/? categories into far fewer categories that can be selected in combination.

#23 - 03/02/2015 08:14 PM - Kyle Hazelwood

- File ECLNMLCategories.xml added

#24 - 03/02/2015 09:13 PM - Kyle Hazelwood

All the NML categories exist in the elog exactly as they are spelled in ECL. I've marked all these categories inactive to force the IOTA ASTA group to parse which categories they want to keep or split up before they can tag entries with them again.

#25 - 03/02/2015 10:29 PM - Kyle Hazelwood

- File deleted (ECLUsers.xlsx)

#26 - 03/02/2015 10:32 PM - Kyle Hazelwood

Kyle Hazelwood wrote:

> Unfortunately ECl doesn't require their account usernames correlate to any other service (such as the services account). Their usernames don't necessarily match the services accounts that the elog uses. This makes importing entries much more difficult. Thankfully Chip Edstrom has created a spreadsheet of usernames for me to use to match users. Also, some ECL users have multiple usernames. The spreadsheet does include their email address which will be helpfull.

I've checked the usernames for the ECL NML logbook versus the elog. There are 175 ECL NML log users.

- 98 exist in the elog and need no further action
- 7 exist in the elog as a different username
- 1 user has two accounts in ECL
- 69 do not exist in the elog and must be added manualy before the entries can be imported

#27 - 03/02/2015 10:33 PM - Kyle Hazelwood

- File ECLUsers.xlsx added

#28 - 03/04/2015 07:24 PM - Kyle Hazelwood

- File ECLUsers.xlsx added

All the ECL NML log users exist in the elog now. There are 8 users that exist in both the elog and ECL with different usernames. Luckily, of these 8 users only 1 has contributed to the ECL NML log. I'll add these comments by hand afterwards.

I'll do a few more tests and probably transfer the entries next week. The transfer will probably last a couple days so that the elog and ECL servers don't really notice the traffic.

#29 - 03/04/2015 07:25 PM - Kyle Hazelwood

- File deleted (ECLUsers.xlsx)

#30 - 04/08/2015 11:24 PM - Kyle Hazelwood

I successfully added a couple NML entries to the elog "Testing" logbook tonight using the new import scheme. I'll try a few more unique entries before importing all entries into the IOTA ASTA log.

#31 - 04/08/2015 11:28 PM - Kyle Hazelwood

Kyle Hazelwood wrote:

> All the ECL NML log users exist in the elog now. There are 8 users that exist in both the elog and ECL with different usernames. Luckily, of these

8 users only 1 has contributed to the ECL NML log. I'll add these comments by hand afterwards.

I'll do a few more tests and probably transfer the entries next week. The transfer will probably last a couple days so that the elog and ECL servers don't really notice the traffic.

It has taken longer than expected to work out the bugs. The new import client is multi-threaded so the import should be much faster (about 1K entries per minute).

#32 - 04/12/2015 12:29 AM - Kyle Hazelwood

I imported one hundred entries to the test log with much success. However, I did notice a few of the entries imported were not formated very well. This appears to be a flaw in the ECL XML API. The ECL XML API does not preserve any line breaks. So, entries that appear formatted with line breaks in the ECL client are missing those line breaks when imported.

Example:
http://dbweb4.fnal.gov:8080/ECL/nml/E/show?e=8390 vs http://dbweb4.fnal.gov:8080/ECL/nml/E/xml_get?e=8390

#33 - 04/12/2015 06:29 AM - Kyle Hazelwood

I tried a few hundred more entries.

Problems:

- Auralee Morin (amorin01) does not have a services account so an elog account can not be made for them. Any entry or comment authored by them will fail to be imported.
- When attempting to upload large .zip files a HTTP 502 error is being thrown. I may need to increase the timeout.
- There are a few unique file extensions that are not allowed in the elog (i.e. .dmc?).
- TIFF files will need to be supported in the elog. These images are used often in the NML ECL log.

#34 - 04/13/2015 01:31 PM - Chip Edstrom

> Problems:
>
> - Auralee Morin (amorin01) does not have a services account so an elog account can not be made for them. Any entry or comment authored by them will fail to be imported.

Auralee Morin is now Auralee Edelen, and the amorin01 is her services user name. She has been added as a user with the appropriate information.

> - There are a few unique file extensions that are not allowed in the elog (i.e. .dmc?).

The .dmc is a proprietary scripting language... I don't expect that there will be many of these. It may work best to just make a list of attachments for odd file names and then add them manually after the move in a zip file if there aren't too many.

#35 - 04/15/2015 05:19 AM - Kyle Hazelwood

- File Elog_NML_import.log added

- File Elog_NML_import_failed.log added

- File ElogImportClient.PNG added

Most entries were imported to the IOTA ASTA log this morning.

- 7137 entries uploaded successfully
- 630 entries failed
- Average import rate 15.3 entries/sec
- Failure percentage 8.1%

The amount of failed entries was greater than anticipated. The vast majority (~500 entries) failed due to incorrect category name mapping. I'll try to fix this problem and import these entries in the coming weeks. The remaining entries failed mostly due to not having any entry/comment text which is mandatory for the elog or from file types that are not allowed in the elog. The entries and comments without text will probably never get imported. The not allowed file types will just have to wrapped in a zip and imported later.

#36 - 04/16/2015 02:15 AM - Kyle Hazelwood

- File Elog_NML_import_failed_ids.log added

Files

ECLEntryComments.PNG	22.9 KB	02/09/2015	Kyle Hazelwood
ECLEntryFiles.PNG	77.1 KB	02/09/2015	Kyle Hazelwood
ElogECLImportThread_FailedEntries.log	2.11 MB	02/27/2015	Kyle Hazelwood

ElogECLImportThread_FailedEntries2.log	7.83 KB	02/28/2015	Kyle Hazelwood
ECLNMLCategories.xml	4.02 KB	03/03/2015	Kyle Hazelwood
ECLUsers.xlsx	26.7 KB	03/05/2015	Kyle Hazelwood
Elog_NML_import.log	1.57 MB	04/15/2015	Kyle Hazelwood
Elog_NML_import_failed.log	1.2 MB	04/15/2015	Kyle Hazelwood
ElogImportClient.PNG	50.3 KB	04/15/2015	Kyle Hazelwood
Elog_NML_import_failed_ids.log	3.68 KB	04/16/2015	Kyle Hazelwood