ACE 2008: Cross-Document Annotation Guidelines (XDOC)

Version 1.6 – 2008.05.06

Linguistic Data Consortium

http://projects.ldc.upenn.edu/ace/

Overview

The objective of the Automatic Content Extraction (ACE) series of evaluations is to develop human language understanding technology to provide automatic detection and recognition of key information about real-world entities, relations and events in source language text, and to convert that data into a structured form.

For 2008, the ACE evaluation will not only include within-document entity and relation detection and disambiguation for Arabic and English, but also include cross-document global integration and reconciliation of information.

For cross-document and cross-language entity and relation disambiguation tasks, system output will be evaluated only for person (PER) and organization (ORG), and only for documents in which the "target entities" are mentioned by name. Target entities refer to a carefully selected list of pre-defined entities of interest.

LDC/sponsors have identified 50 target entities, and a 400 document corpus, which have the following features:

- each entity should be mentioned 5-100 times in the corpus
- some examples of the alias issue, for instance, Carlos the Jackal aka Ilich Ramírez Sánchez
- some examples of spelling and orthographic variations, for instance, Khadafi example
- some examples of the confusable entities issues, for instance, the six Michael Jordans
- some entities that appear in both English and Arabic data sets

The purpose of the Cross-Document (XDOC) task is to globally coreference these 50 ACE entities, and all ACE relations which contain them, over the 400 document corpus. These 400 documents will have been previously ACE annotated, so all ACE entities and relations will already be coreferenced within the documents.

The XDOC task will be to find the target ACE entities and relations, and coreference them globally, i.e., not within but **between** documents.

The XDOC annotation tool, called Callisto, will allow us to do this effectively, by allowing searching and global co-reference of ACE annotations. The procedure for this annotation is outlined below.

1 Entity and Relation Cross-Document Coreference

The three main tasks for XDOC annotation are:

- Global Entity co-reference on 50 target entities
- Global Relation co-reference on relations containing 50 target entities
- Global Entity co-reference on PER and ORG arguments in those relations that are not one of the 50 target entities

We will also do Global Entity co-reference (GEDR) on the remaining PER and ORG entities in a post-hoc process a few weeks after the main ACE 2008 delivery has been completed.

1.1 Global Entity Co-reference on 50 Target Entities

The procedure for Global Entity co-reference on the 50 target entities is:

- Open the Callisto annotation tool
- Open name profile/list of candidate files for your assigned name (AWS will do this)
- Open an anchor file for your assigned name
- Find your assigned name in the anchor file.
- Search the 400-document ACE annotated corpus for every possible name variant of that entity.
- For each search result, view the document to see if that entity is referring to your target entity.
- If the search result entity is referring to your target entity, link it to your target entity
    - Watch out for confusable entities of your target entity (a different entity with the same name). Do not link a confusable entity to your target entity.
- If you find a confusable entity that falls under the criteria in section 1.1.2, repeat from step 3.
- If you find a mention of your assigned name that is missed, wrongly tagged, or not co-referenced with other mentions in that ACE annotated file, **do not** globally co-reference it. Please mark that name “Broken” in AWS, and include in the comment field a description of the error and the file name it occurs in.

1.1.1 Entity Profiles

During the Structured Data Exploration task, “Entity profiles” were compiled for the 50 target entities. These entity profiles are intended to be a complete list of all names which can refer to that entity, so the annotator can find all the mentions of that entity in the corpus.

It is possible some variants in the corpus are not in your entity profile. If you find a name mention not in your entity profile, make sure to do a search for this variant as well.

The entity profiles consist of:

- **Entity Handle (main name)**
- **Aliases and Nicknames**
- **Transliterations**
- **Orthographic Variants**
- **Knowledge/Fact about Entity**

You may need to search online for more information about your assigned name if your entity profile seems sparse.

1.1.2 Confusable Entities

The Knowledge/Fact is given to help you differentiate each entity from other similarly named entities.

If you find your ***entity handle*** refers to multiple entities in the corpus, you must globally co-reference all of the entities during that assignment, ***even if the entity profile fact refers to only one of the entities.***

If the entity handle was “George Bush”, you would need to perform GEDR on both “George W. Bush” and “George H. W. Bush” entities, even if the fact was “the 43rd U.S. President”.

However, if the *entity handle* was “George W. Bush”, you would **not** have to perform GEDR on both George Bushes, because that name only refers to one of them.

If you are not sure whether you need to annotate multiple entities for an assigned name, ask your supervisor.

2 Callisto Annotation Tool

2.1 Starting Callisto

Callisto can be run by opening a terminal and running the following command:

```
/ldc/bin/callisto.sh &
```

2.2 Setting up Callisto

In order to work properly, you must configure Callisto for your account.

Run Callisto for the first time with the above command, and then close it. This will create a text file called "callisto.prefs" in a subdirectory of your home directory called ".callisto".

Before starting annotation, you will need to close the tool, and edit the callisto.prefs file by opening a new terminal window and running the following command:

```
emacs .callisto/callisto.prefs &
```

This opens the callisto.prefs file in the emacs text editor.

Add the following line into the file:

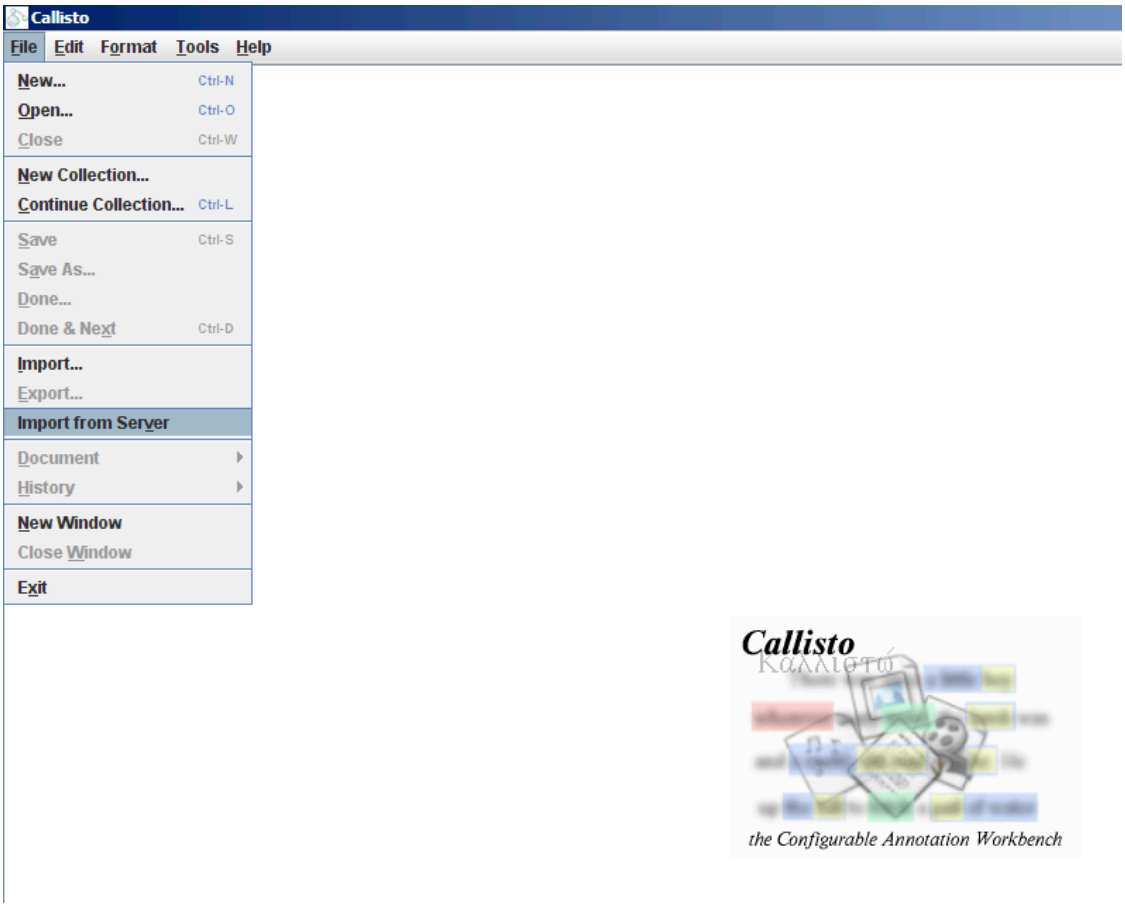
```
jawbFrame.showEdnaImport=true
```

Go to “File” and click on “Save current buffer”, and close this file. Then you can open Callisto again and start annotation.

2.3 Importing files from Edith database

Callisto requires that you start XDOC annotation with a file mentioning your assigned entity. That file serves as an anchor to which you will link the other mentions in the corpus.

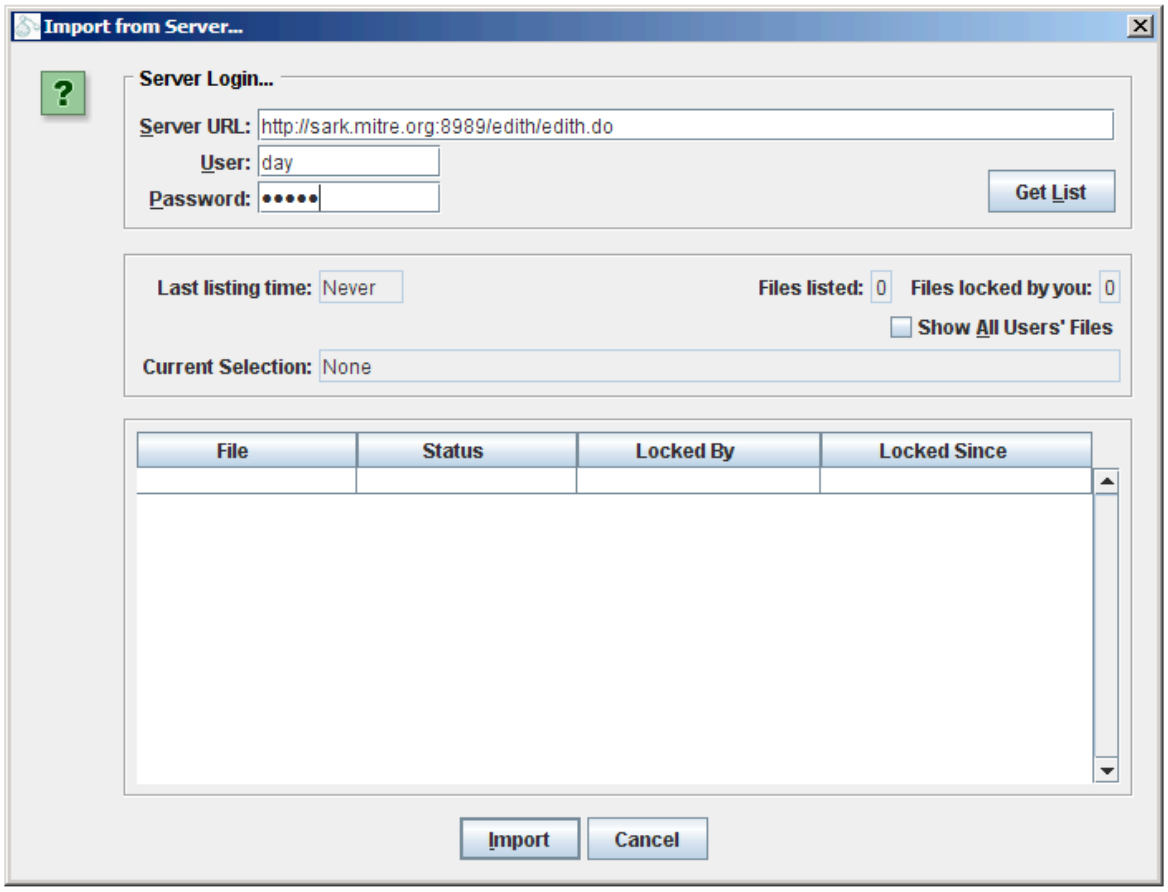
Your assigned entity profile has a list of candidate anchor files at the bottom. Go to the File menu and select "Import from server":



Input the Server URL. The English Server URL is listed below. The Arabic URL will be provided in email or on the wiki.

English Server URL:

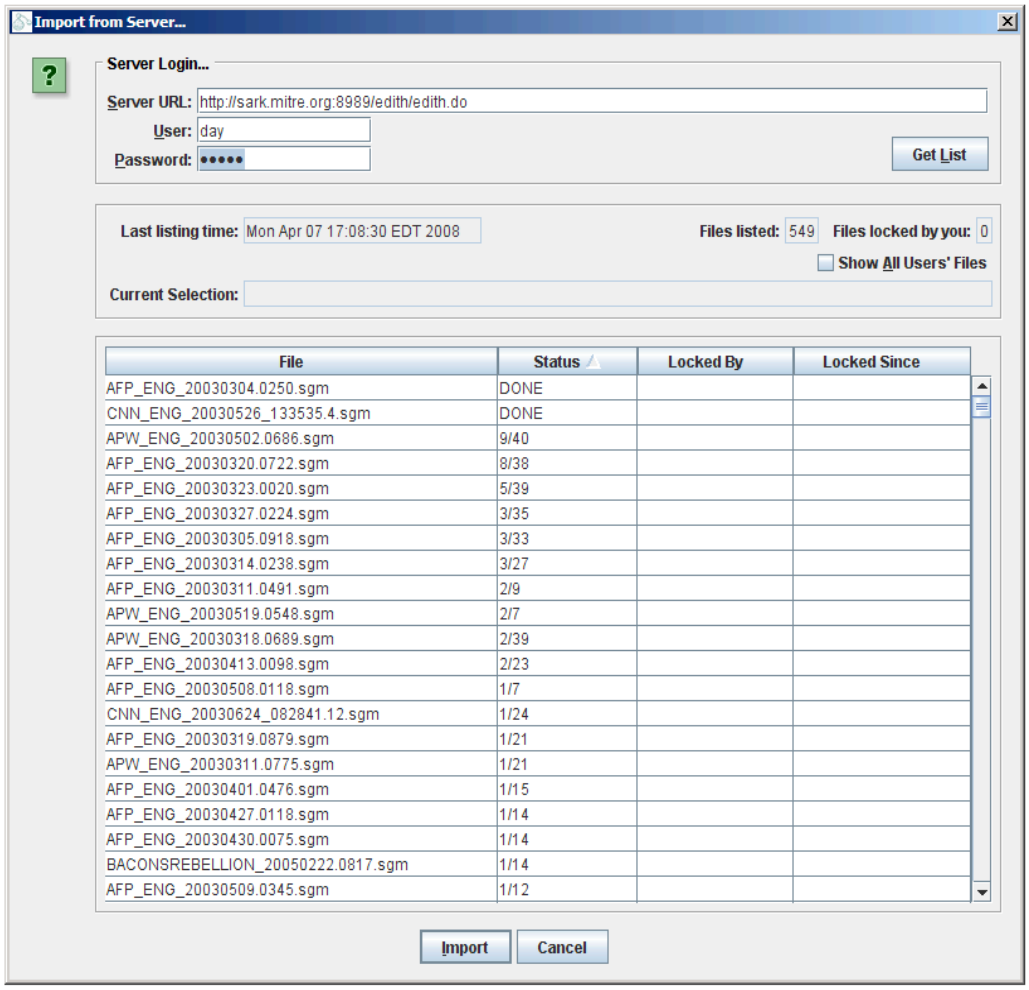
http://blooie.ldc.upenn.edu:8080/edith/edith.do



Enter your XDOC username/password and hit “Get List”

Within the same session, your XDOC password will be remembered, so next time you can just hit “Get List” again.

“Get List” will bring up the list of all the files in the corpus that can be imported:



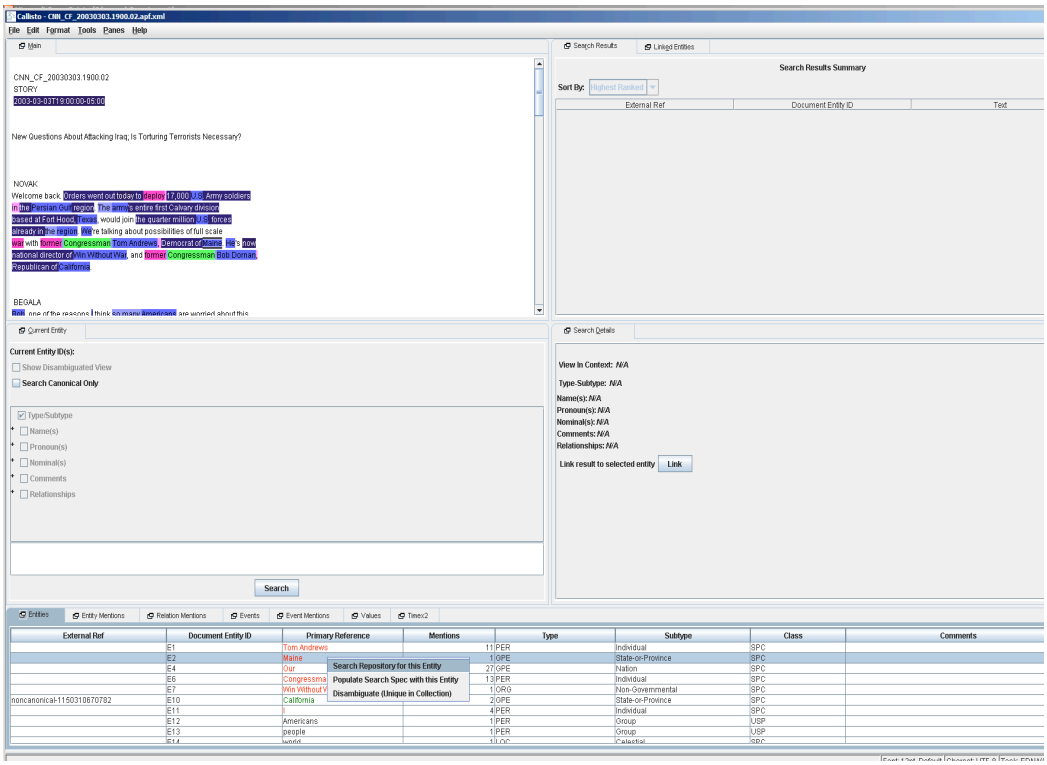
Click on any column header to sort. The “Status” column shows how many entities have been annotated out of the number of entities in that file.

Find the file you want, then select that file’s row and hit “Import”

If you don’t have any candidate files available for importing, tell your supervisor immediately.

2.4 Annotating Files

When you import a file, it will open the annotation function window. You will see the text of the anchor file on the left, and the Entities tab at the bottom, which contains a table of all the ACE annotated entities in the document.

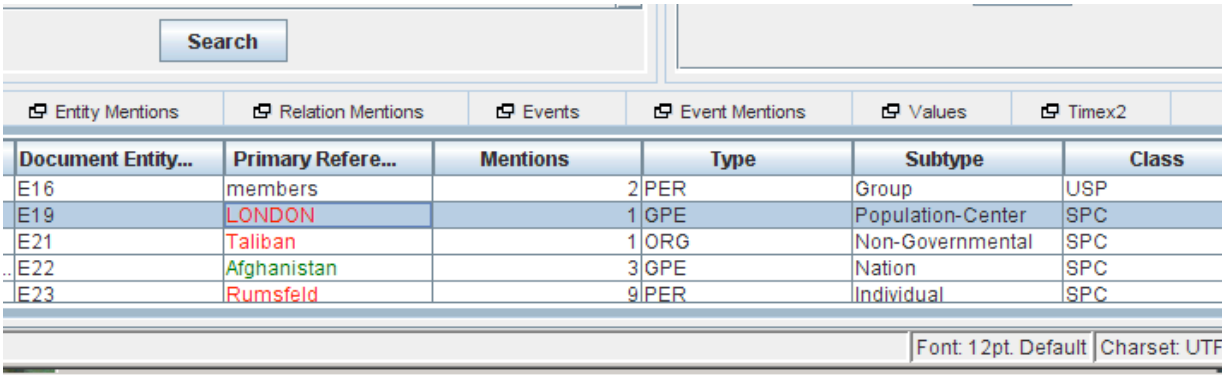


You can sort by type, primary reference, name, etc., by clicking on the header of each column.

Entities in red text have not yet been given an External reference (global) ID.

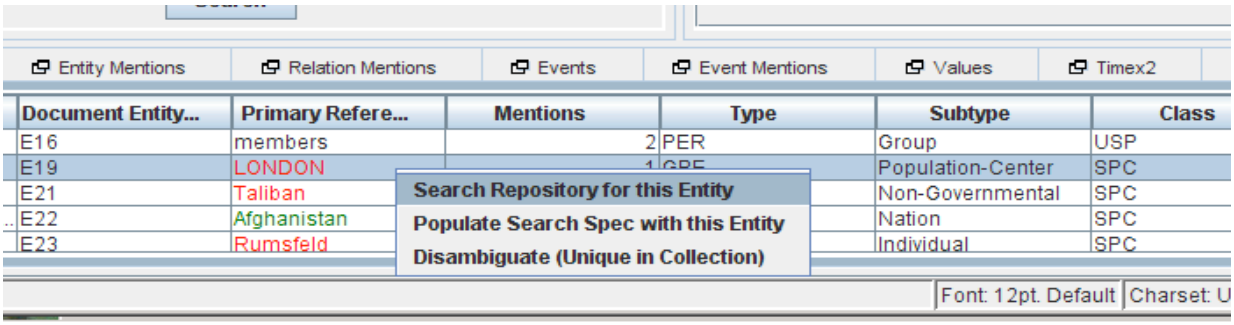
If your assigned PER or ORG is green, someone else has already linked that entity. In that case, mark the name “Broken” and inform your supervisor.

If your PER or ORG name is in red, double click on it:



Document Entity...	Primary Refere...	Mentions	Type	Subtype	Class
E16	members	2	PER	Group	USP
E19	LONDON	1	GPE	Population-Center	SPC
E21	Taliban	1	ORG	Non-Governmental	SPC
E22	Afghanistan	3	GPE	Nation	SPC
E23	Rumsfeld	9	PER	Individual	SPC

ACE08_XDOC_1.6.docThen right-click:



Select “Populate Search Spec with this Entity”. This will automatically fill the search term field with the NAM mentions, and let you manually add more terms before executing the search.

2.4.1 Manual Search Terms

You can manually add in additional search terms into the text box below the NAM mentions. Searching for part of a name will find any names that contain that string. For instance, a search on “dou” will find “Mohammed Al-Douri”.

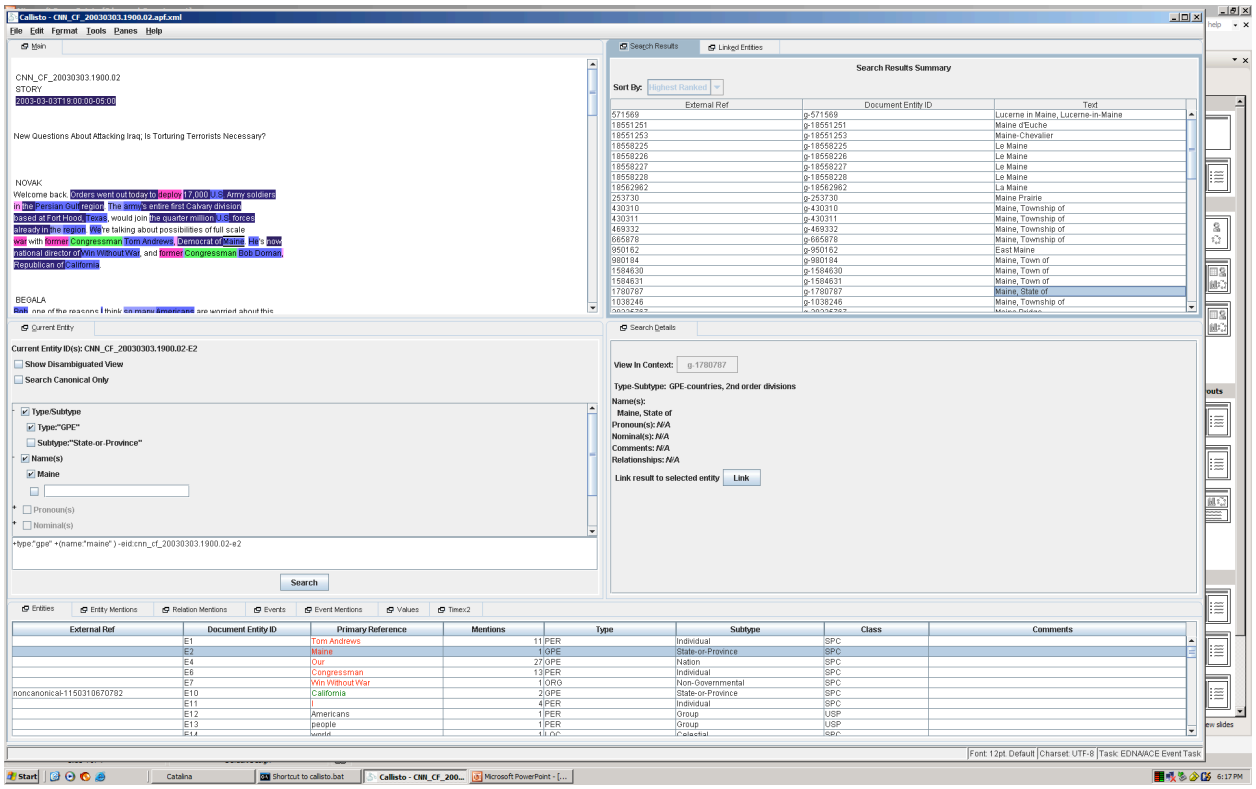
Use this function to search for all variants in your entity’s name profile. Make sure to do a manual search for the NAM variants from your anchor file, because the automatic search function can miss some mentions.

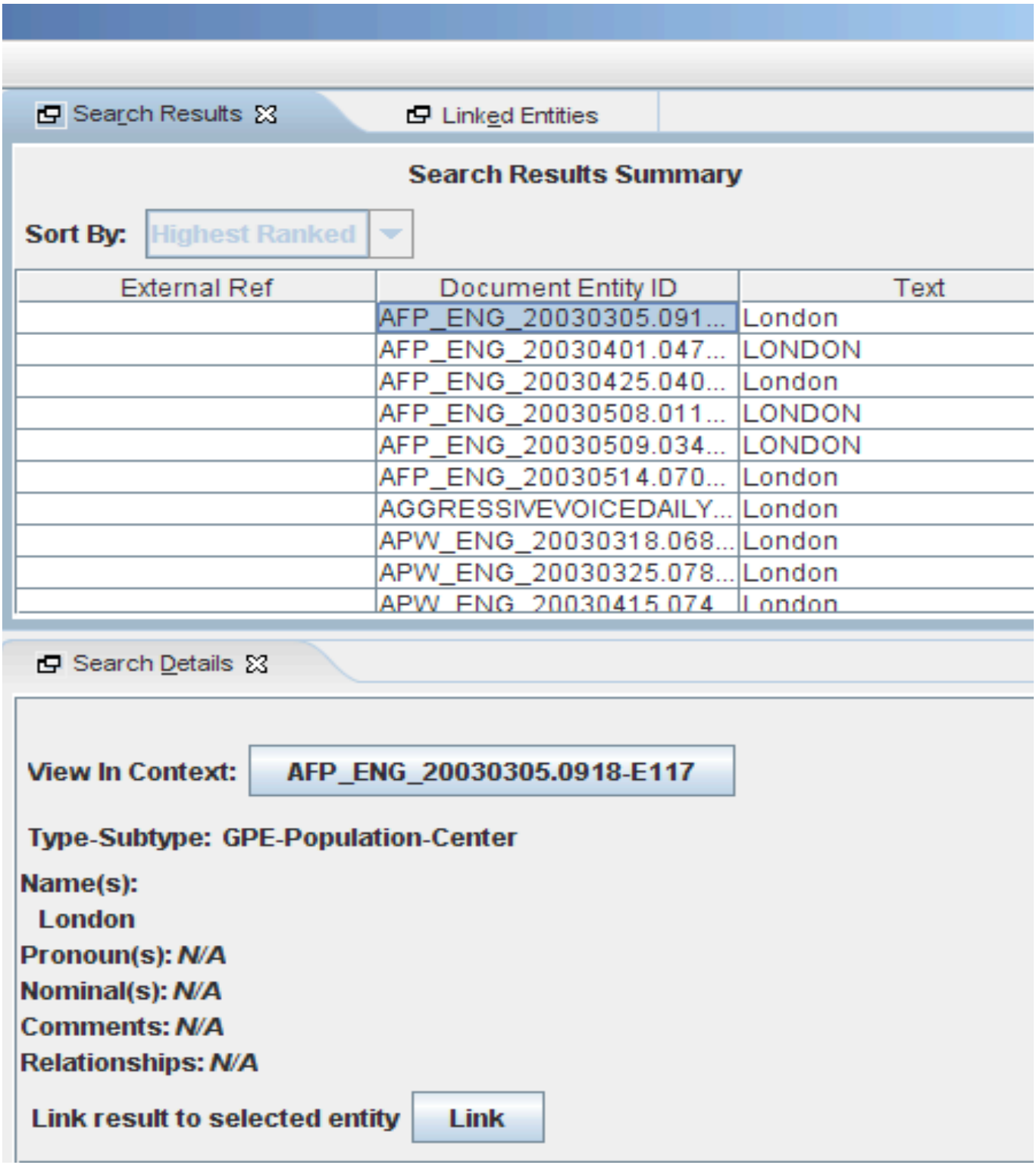
If you want to search not just on PER or ORG NAMs, you need to unselect “Type/Subtype”, or select the type you want to include.

These searches can be useful to find names spelled completely wrong, like “leo clinhinghof-hoffer” for “Leon Klinghoffer”, which can occur in transcribed speech.

Leon Klinghoffer was often referred to as an “American tourist”, so you could search for “tourist” under NOM to find articles that didn’t match your NAM searches.

When you have entered the terms you wanted, hit “Search”.



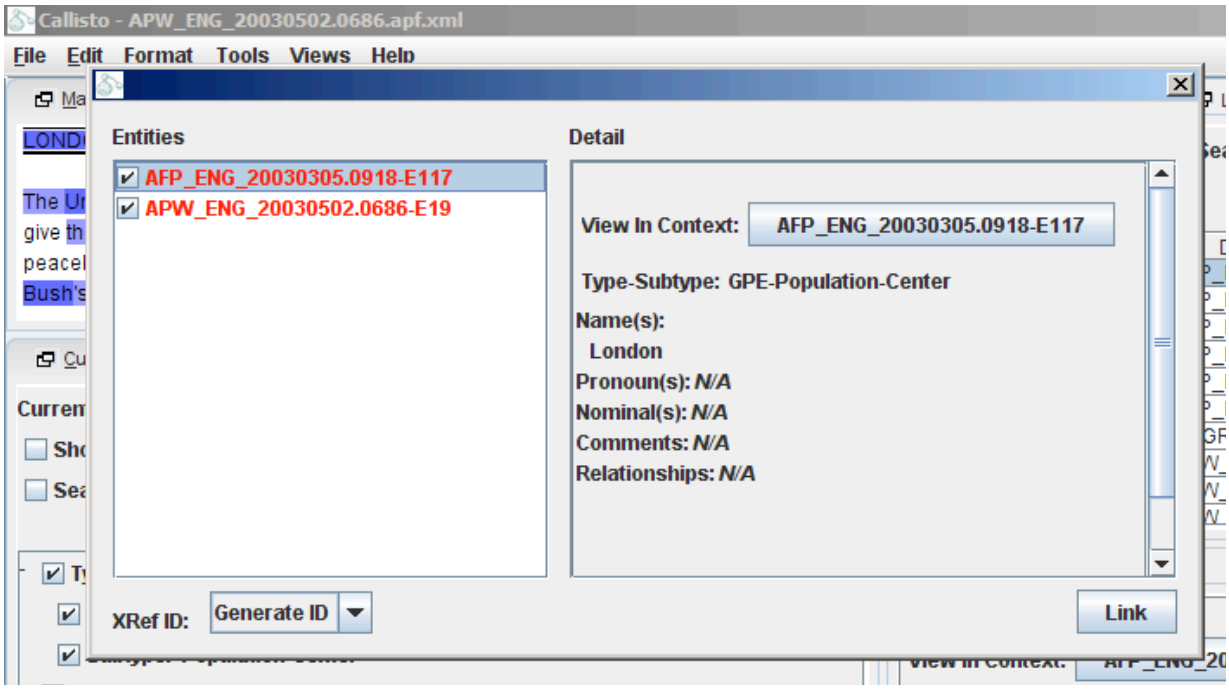


If an entity in your search result has an External Ref ID, it has already been globally co-referenced.

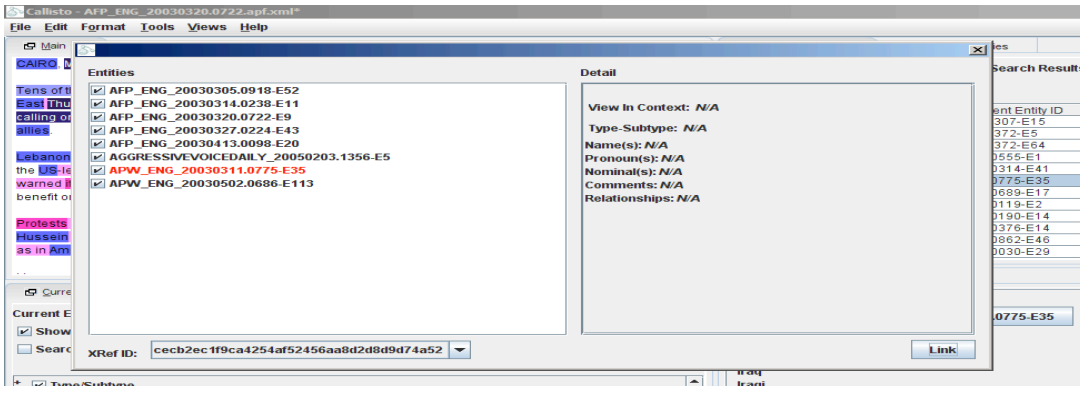
You can see from the Text column which mentions were a match for your search.

Select the row and click on “View in Context” to see the entity mentions in the file context. They will be highlighted in red text.

If the entities are the same, close the document and select “Link” to link that entity to your “anchor” entity. The tool will generate an External Ref (Global) ID for that entity the first time it is linked.

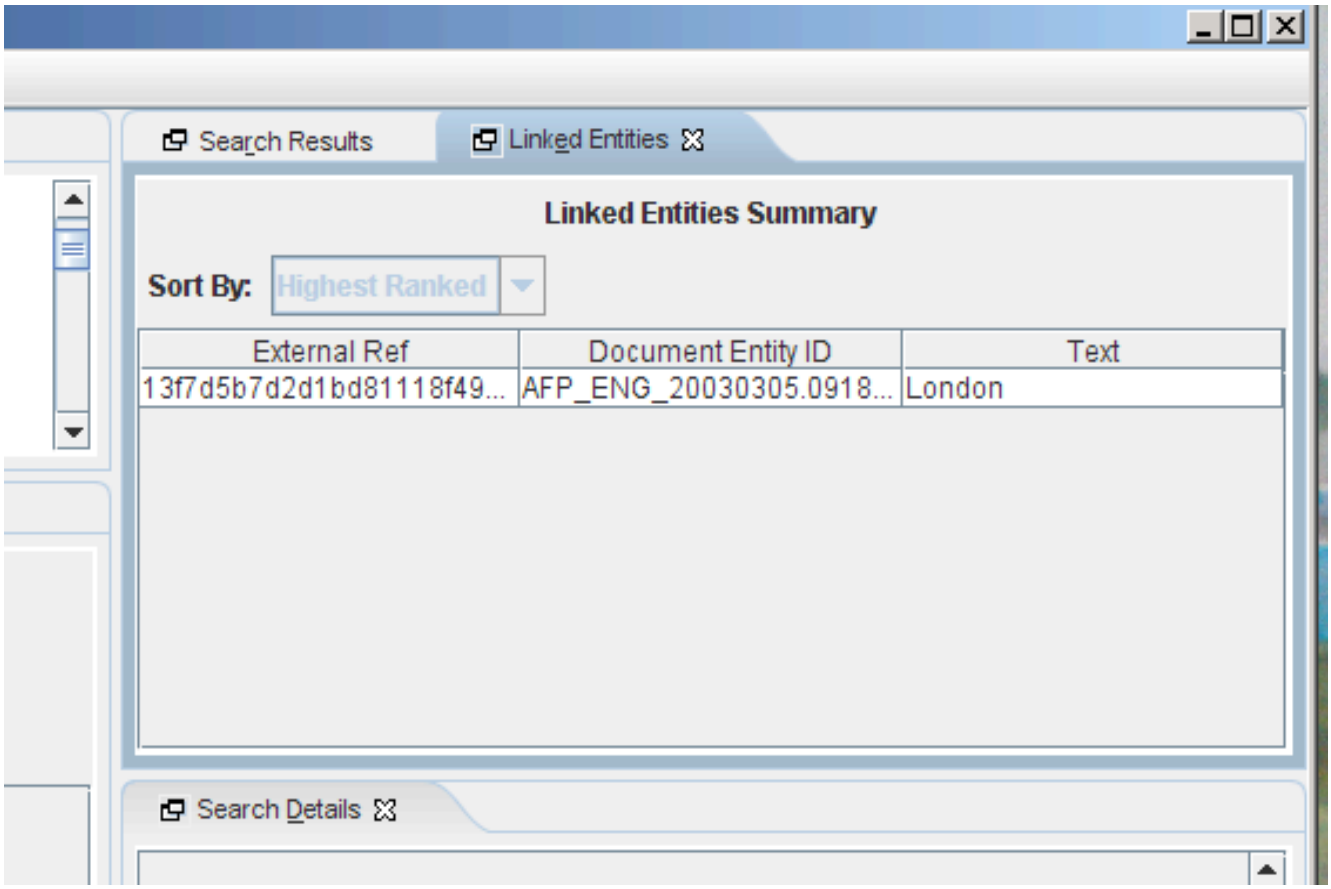


When it already has an External Ref (XRef) ID, it will show you that ID and the other entities already belonging to it.



If you don’t want to link the entity, you can just close this window.

If you do select Link, that document result will be moved from the Search results tab to the “Linked Entities” tab.



If you link an entity accidentally, you can select this row and hit “Unlink”.

When you are confident there are no more mentions of your target entity in the corpus, then you are done with your collection.

If you did not find any other mentions of your target entity in the corpus, right-click on your entity and select the “Disambiguate in the Corpus” option. This will give the entity an External Ref ID, without linking any other mentions to it.

Now select File -> Close to close the file. The tool saves automatically, so you never need to use “Save”.

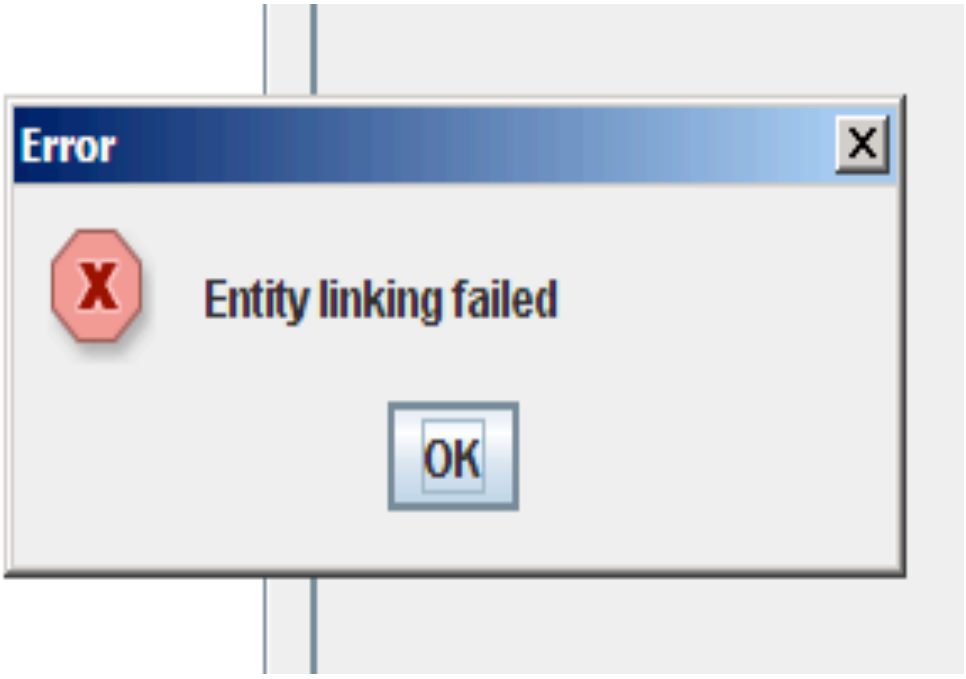
After you select “Close”, it will ask you if you want to release the lock on this file. Make sure to click "Yes" to release the file if you are done with that nACE08_XDOC_1.6.docame.

To get your next assignment, hit “Stop” on AWS as normal, and say Yes to a new name assignment.

Then go to “File > Import from Server” again in Callisto, hit “Get List”, and find the anchor file for the new name.

2.5 Entity Linking errors

Due to the fact that multiple annotators can simultaneously search and annotate with the Callisto/Edna task, it is possible that two or more annotators might inadvertently attempt to link to the same entity at the “same time.” If Callisto suspects this behavior is happening, this error message will appear.



If this error message happens to you, report it to your supervisor.

You may keep working on the name as long as your mention is correctly tagged in the ACE file. If it’s not, mark the name as “Broken” and move on.