## 

[Overview](#h.m0jbr21r411e)

[Base requirements](#h.a8g6npfhfwkm)

[Prerequisites](#h.ub9eyiv32rus)

[Suggested Workflow Processes](#h.h0p1phuwbx36)

[To add new MODs data](#h.yutctk9aciyd)

[To correct the expression type of an expression in the catalog](#h.p1oerlf6emq)

[To remove an invalid MODs file for an expression (but to keep the expression)](#h.nhjio4uep07a)

[To remove an invalid expression](#h.k9nxzd3ope8u)

[To add a work without MODS data](#h.z6vztppszro9)

[To add a textgroup without MODS data](#h.on33vmbynqfk)

[To add missing MADS data for a textgroup already in the catalog](#h.hzgq2xownkt2)

[To correct or update MODS or MADS data which doesn’t impact URN assignment/identity](#h.pkgk51qt6pav)

[Automatic Feed Update Process](#h.2qhsoa65k5lx)

[Discussion Notes and Questions](#h.y52zsy6a98b4)

## Overview

The following document describes a suggested workflow and approach for maintaining the catalog data in the interim period between the initial release of the catalog and the time when we’re ready to open it up more broadly for direct user contributions.

## Base requirements

1. Once we publish a urn we can’t reassign or delete it but we can redirect it.
2. We need to be able to add MODs for versions which either aren’t yet cataloged or are in the inventory of versions but don’t yet have MODs data. These MODs records might be host-level mods files with multiple versions, or single version level mods files.
3. We need to be able to add MADs files for textgroups whose authors don’t have them yet.
4. We need to be able to correct any of the data in the catalog.

## Prerequisites

1. Before release, we put the aggregated MODs and MADS data into a GIT repository ([*catalog\_data*](https://github.com/PerseusDL/catalog_data)).
2. [*catalog\_data*](https://github.com/PerseusDL/catalog_data) contains one CTS text inventory file per work,0 or 1 MADs files per textgroup and 0 or more MODS files per version. Not all versions have MODS files. A version may also have multiple MODS files (e.g. in the case where individual volumes have their own MODs data).

[tlg0012.mads.xml]

tlg0012.tlg001.cts.xml

[tlg0012.tlg001.perseus-grc1.mods1.xml]

Note, the *catalog\_data* may contain:

* + MODs records which have been incorrectly assigned to versions
  + versions which are incorrectly categorized as edition or translation
  + versions without MODS records
  + duplicate (i.e. invalid) versions
* We create a separate GIT repository location ([*catalog\_pending*](https://github.com/PerseusDL/catalog_pending)*)* for future MODS data to be processed.
* We create a CITE collection of the authors in the catalog (*author\_collection*). Maintained manually for now via the CITE Collection Editor.
  + Preliminary Author collection data is at <https://www.google.com/fusiontables/DataSource?docid=1JKDi1OHvxWoh1w38mnQfUDey2pB_nx3UnRITvcA>
  + This table was populated by the import file at <https://raw.github.com/PerseusDL/catalog_data/master/citecoll/mads.cite.import.20130613.csv> (created by running [parsemads.xquery](https://github.com/TuftsUniversity/perseus_catalog/blob/master/data/src/xquery/parsemads.xquery))
  + Changed needed:
    - ~~the collection name should be changed from urn:cite:perseus:primauth to urn:cite:perseus:author because it will eventually contain all authors not just those of the primary source texts~~
    - mads file path should probably be a direct link to the location of the file in github
* We create a CITE collection of the textgroups in the catalog (*textgroup\_collection)*. This is maintained manually for now via the CITE Collection Editor.
  + Preliminary Textgroup collection data is at <https://www.google.com/fusiontables/DataSource?docid=1I0Zkm1mAfydn6TfFEWH2h0D3boAd4q7zC-4vuUY>
  + This table was populated by the import file created from the release feed at <https://raw.github.com/PerseusDL/catalog_data/master/citecoll/cattg.import.20130613.csv>
    - import file created by running the following steps against the work feeds:
      * [textgroupfeed.xq](https://github.com/TuftsUniversity/perseus_catalog/blob/master/data/src/xquery/textgroupfeed.xq) to create a composite text inventory of all works
      * [ti\_to\_cattg.xsl](https://github.com/TuftsUniversity/perseus_catalog/blob/master/data/src/xslt/ti_to_cattg.xsl) to create import file from composite text inventory
  + Changes needed:
    - has\_mads and mads\_possible fields should be replaced by a set of many-to-one relationships of author urn to textgroup urn
* We create a CITE collection of the works in the catalog (*work\_collection)*. This is maintained manually for now via the CITE Collection Editor.
  + Preliminary work collection data is at <https://www.google.com/fusiontables/DataSource?docid=1PQY6nVHZV8Ng42-qrbiLKXwDz9XoNlStKi3xKfU>
  + This table was populated by the import file created from the feed at <https://raw.github.com/PerseusDL/catalog_data/master/citecoll/catwk.import.20130613.csv>
    - import file created by running the following steps against the work feeds:
    - [textgroupfeed.xq](https://github.com/TuftsUniversity/perseus_catalog/blob/master/data/src/xquery/textgroupfeed.xq) to create a composite text inventory of all works
    - [ti\_to\_catwk.xsl](https://github.com/TuftsUniversity/perseus_catalog/blob/master/data/src/xslt/ti_to_catwk.xsl) to create the import file from the composite text inventory
* We create a CITE collection of the versions in the catalog (*version\_collection)*. This is maintained by an automated process.
  + Preliminary version collection data is at <https://www.google.com/fusiontables/DataSource?docid=1STn9raQzWZDeIC4f_LHuLDUPPW3BDkpFfyKrKtw>
  + This table was populated by the import file created from the feed at [https://github.com/PerseusDL/catalog\_data/blob/master/citecoll/catver.import.20130613.csv](https://github.com/PerseusDL/catalog_data/blob/master/citecoll/catver.import.20130613.csv?raw=true)
  + import file created by running the following steps against the work feeds:
    - [textgroupfeed.xq](https://github.com/TuftsUniversity/perseus_catalog/blob/master/data/src/xquery/textgroupfeed.xq) to create a composite text inventory of all works
    - [ti\_to\_catver.xsl](https://github.com/TuftsUniversity/perseus_catalog/blob/master/data/src/xslt/ti_to_catver.xsl) to create the import file from the composite text inventory
  + Changes needed:
    - has\_mods field may not be sufficient. We need to support a many-to-one relationship between version and mods file (e.g. for multi-volume works)

## Suggested Workflow Processes

### To add new MODs data

1. Check for textgroup level entry in *textgroup\_collection* and add if not present
2. Check for textgroup level MADs file in [*catalog\_data*](https://github.com/PerseusDL/catalog_data) and add if not present (unless textgroup doesn’t have a MADS file as in the Bible, Homeric Hymns, etc.)
3. Check for work level entry in *work\_collection* and add if not present
4. If adding a MODS file for an version which is already inventoried in the catalog (i.e. one which has already been assigned a CTS urn), then you MUST include the version URN at the version level in the MODS file. Otherwise, if no version level URN is found in the MODs file it will be assumed that this is a new version and a new URN will be created in the opp-[lang] sequence.
5. *Pending:* exact process for adding additional mods files for an already catalogued version which has more than one mods file. The process should accept a new mods file for an expression which is already catalogued and has multiple mods files and the eventual interface should allow moving mods files from old incorrect location to new one and renaming them.

### To correct the version type of an version in the catalog

1. Add MODs file for the correct version to [*catalog\_pending*](https://www.google.com/url?q=https%3A%2F%2Fgithub.com%2FPerseusDL%2Fcatalog_pending&sa=D&sntz=1&usg=AFQjCNF553sg1kQysCabo9fSxkbuqOLtLg)
2. Update the entry for the invalid version in the *version\_collection* to flag it as invalid
3. After corrected version is processed, update entry for the invalid version in the *version\_collection* to point at new version
   1. if a valid MODs file existing for the version that is being redirected which should be added to the set of MODs files for the new version, set the value of the mods\_status flag to ‘aggregate’. Otherwise set the value of the mods\_status flag to ‘invalid’

### To remove an invalid MODs file for an version (but to keep the version)

1. delete the MODS file from the [*catalog\_data*](https://github.com/PerseusDL/catalog_data) repo.

### To remove an invalid version

1. delete the MODS file (if it exists) from the [*catalog\_data*](https://github.com/PerseusDL/catalog_data) repo.
2. Update the entry for the invalid version in the *version\_collection* to flag it as invalid

### To add a work *without* MODS data

1. Check for textgroup level entry in *textgroup\_collection* and add if not present
2. Check for textgroup level MADs file in [*catalog\_data*](https://github.com/PerseusDL/catalog_data) and add if not present
3. Add work level entry to *work\_collection*

### To add a textgroup *without* MODS data

1. Add textgroup level entry to *textgroup\_collection*
2. Add textgroup level MADs file in [*catalog\_data*](https://www.google.com/url?q=https%3A%2F%2Fgithub.com%2FPerseusDL%2Fcatalog_data&sa=D&sntz=1&usg=AFQjCNFUcHjd2hRIp6X9Luc3rEObQDmSJg) (if applicable)

### To add missing MADS data for a textgroup already in the catalog

1. Add textgroup level MADs file in [*catalog\_data*](https://github.com/PerseusDL/catalog_data) (if applicable)

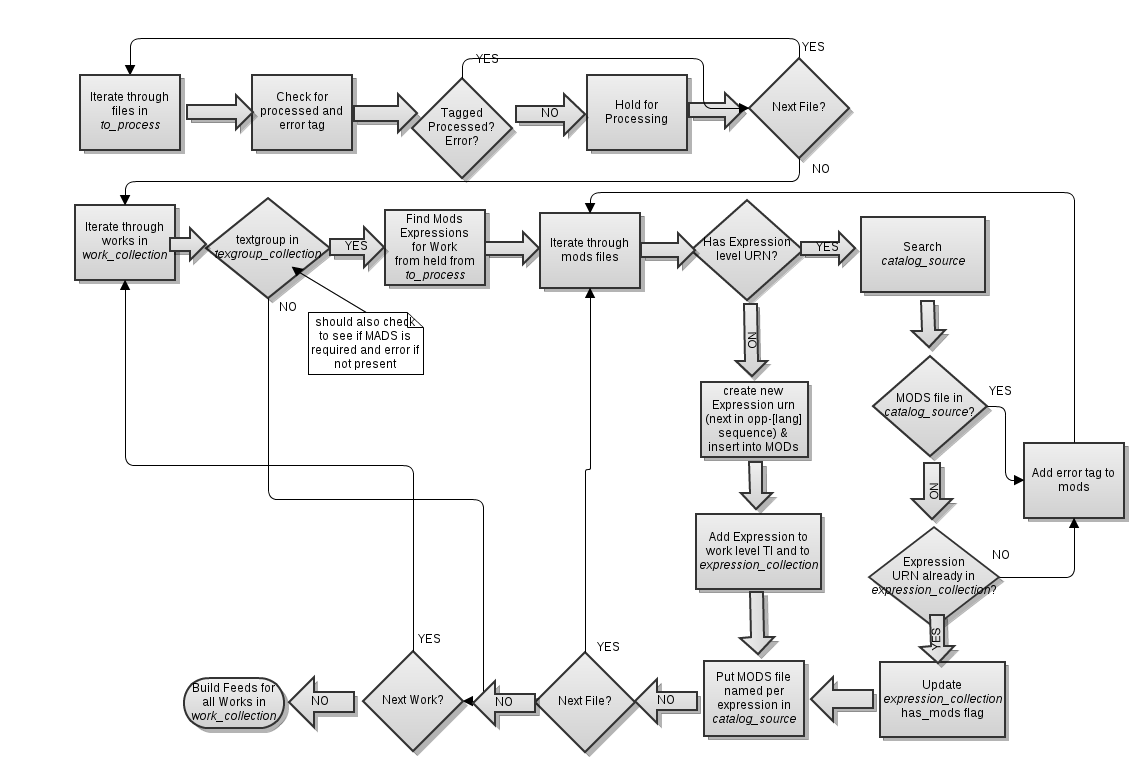
### To correct or update MODS or MADS data which doesn’t impact URN assignment/identity

1. Update the corresponding file in [*catalog\_data*](https://github.com/PerseusDL/catalog_data)[(n](https://github.com/PerseusDL/catalog_pending)ote eventually this could be done through Perseids with the catalog interface can providing a direct link to edit the file there, but see also the discussion below.)

## Automatic Feed Update Process

An automated process will look for updated data and regenerate feeds according to something similar the following process flow.

Note that corrections need to be made to the flow depicted below to support many-to-one relationships between textgroup urn and author urn and between version urn and mods file



## Discussion Notes and Questions

The CITE collections for textgroups, works and versions are to a certain extent redundant with the corresponding tables in the mySQL database. It’s not yet clear to me whether or not these should be the same source. I see the CITE collections as providing us a way to manage the editing workflow for the catalog without impacting the runtime database. It may at some point be desirable to merge these two, along the lines of what Anna has proposed, but I think we need to collect a little more information about the types of contributions people will want to make to this data first. In particular, I think answers to the following questions are not clear:

1. Do we want to maintain all data that may appear in the MODs and MADs files in database fields, and just construct these as separate disseminations, or is there value in retaining the MODs and MADs files as the canonical source format for this data?
2. What problems does allowing field-based editing of this data solve and what problems might it introduce?
3. Is there data currently in the MODs and MADs files that should only be populated automatically ? For example, it seems like perhaps the list of versions in the MADS files would be a good candidate for this.
4. In what other standard catalog formats besides MODS and MADS are people going to want to contribute data?
5. Are the fields suited for indexing and display of the data for the catalog interface the same as those needed for data entry? Right now, some of the fields in the catalog database concatenate one or more fields from the source MODs and MADs data. I think we need to put more time into database design before these tables could fully replace the need for retaining the MODs and MADs data as the canonical source.

We also need to take into account:

* requirements and workflow for Lost Content Item/Fragmentary Authors
* requirements and workflow for secondary sources and related items

f