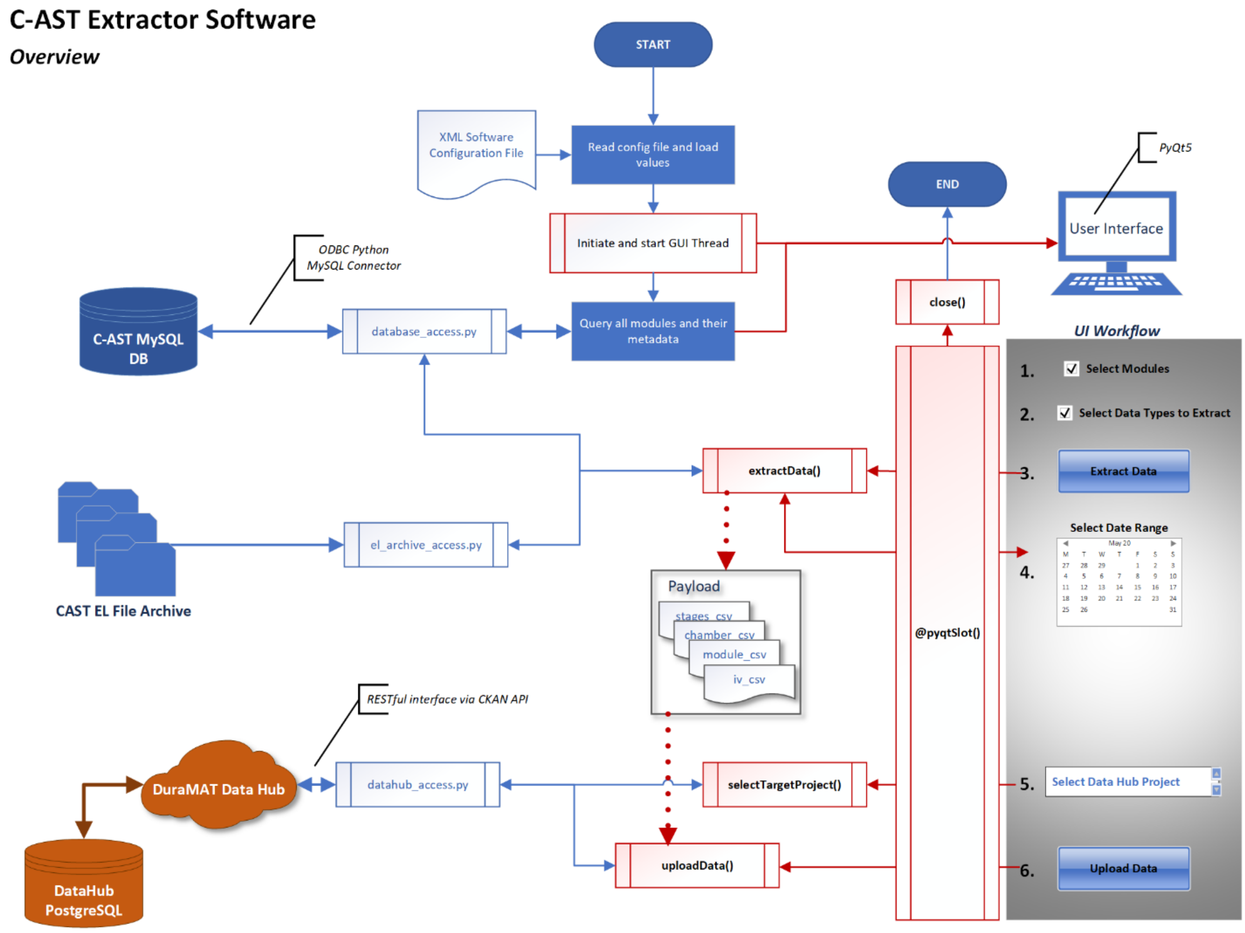
C-AST Extractor Software

Overview

The C-AST Extractor software provides a programmatic method to extract experimental and module data from the instrumentation database at NREL and push those datasets up to an EMN data hub. Initially the software is focused on delivering the data to the DuraMAT Data Hub, but the code can be quickly refactored to access other data hubs. The software is written in python using a variety of easily accessible python modules. The interface is written using PyQT5; a python-based wrapper around the C++ native, cross-platform Qt library. All needed software files are available in the github code repository (see below), or through pip for the outside modules.

Process

The software initiates an extraction of data from the C-AST instrument database keyed on the user selected modules that have been under test. The extracted data is marshalled into a series of CSV files and is set up as a payload to be pushed into the data hub. Additionally, the software will scan and identify in-situ EL images, taken of a selected module, and prepare each as part of the upload to the data hub.



Getting the Software

The software package is available in the internal NREL github repository at:
https://github.nrel.gov/rwhite/cast_extractor

Setting up the Software

While not required, it is often advisable to build up a python virtual environment to run the code. Details on how this is setup can be found at:

venv: https://docs.python.org/3/library/venv.html
Anaconda: https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html

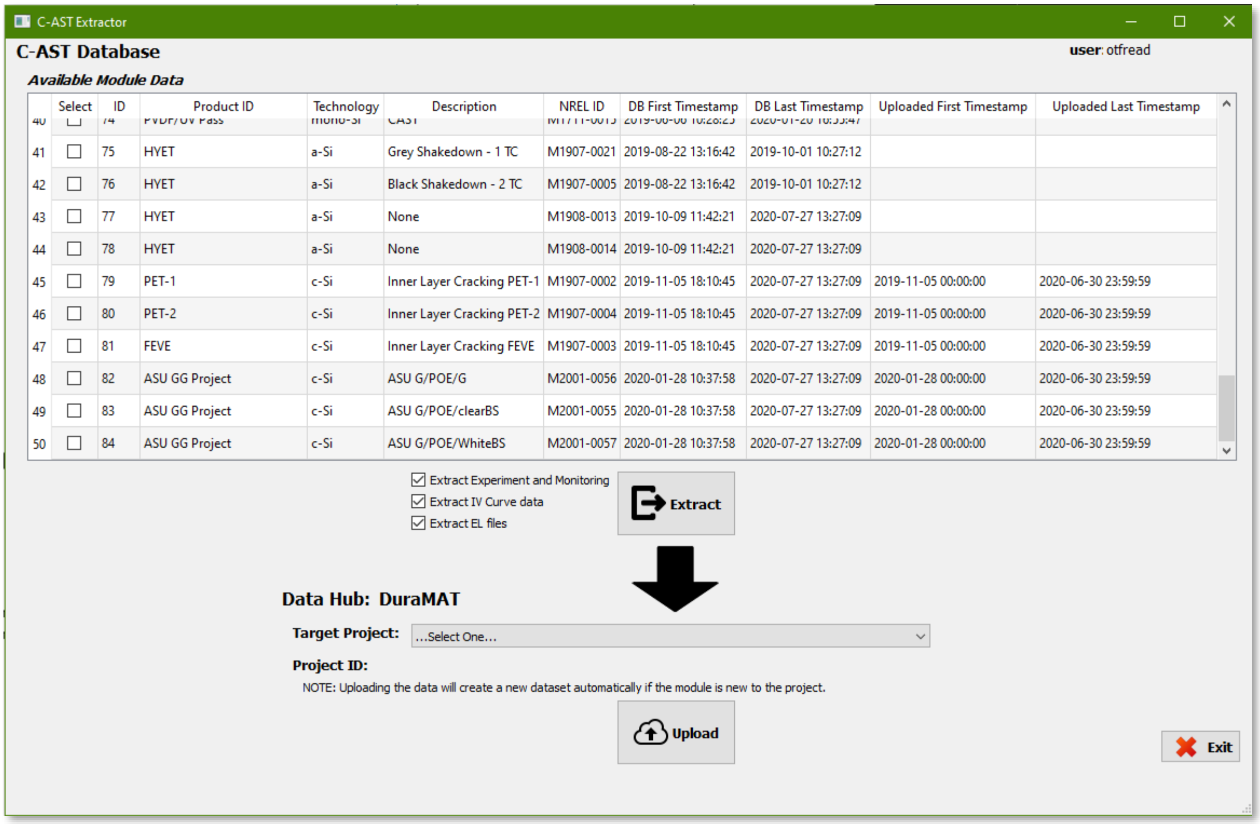
1. Activate virtual environment from a console (if using).
2. Run setup.py after downloading software. The following modules will need to be installed to make use of the software:
   - PyQt5 – provides the GUI
   - xml – used to decode the software configuration file on startup
   - mysql.connector – provides ODBC to the C-AST MySQL database
   - pandas – used to order queries from the database in preparation to create CSVs
   - urllib – provides connection means to the data hub RESTful interface
   - requests – also used for working with RESTful interfaces
   - fnmatch – easy tools for working with filenames and directory file discoveries
   - json – used to encode and decode json files
   - pprint – provides a method to make print output into user friendly formats
3. In same directory as the software make a new folder called ".access" this directory should be at the same level as /src
4. Open the access_config_template.xml file. Fill in the values (red) in the file with the credentials you and locations you wish to use.

```
<?xml version='1.0' encoding='UTF-8'?>
<CAST_Extractor>
    <Software dir='path/cast_extractor/src'/>
    <Database host='1lv14otfdb01.nrel.gov' u='username' p='password' database='combitestingdb'/>
    <Datahub name='DuraMAT' url='https://datahub.duramat.org/' username='your data hub username
here' api_key='your UUID API key here' user_email='your email'/>
    <DataArchive name='EL' dir='path/CASTRawData/EL'/>
    <TmpFileDir dir='path/cast_extractor/outbox/'/>
    <HistoryFile filename='path/cast_extractor/logs/history.csv’ />
</CAST_Extractor>
```

5. Save the config file into the .access directory
6. Change directory to src/
7. To run: `python main.py -f path/to/access_config.xml`
   (Be sure the python version you call is the one from the virtual environment if you are using one.)

How to Use the Software

Once started, there is usually a delay before the GUI becomes visible. The software needs to perform a series of queries to populate the main window widgets. Once the query is received and data organized, it will display the main GUI window.



C-AST Database

user: otfread

Available Module Data

	Select	ID	Product ID	Technology	Description	NREL ID	DB First Timestamp	DB Last Timestamp	Uploaded First Timestamp	Uploaded Last Timestamp
40	□	74	PVDF/UV Pass	mono-Si	CAST	M1711-0013	2019-06-06 10:28:23	2020-01-20 10:55:47		
41	□	75	HYET	a-Si	Grey Shakedown - 1 TC	M1907-0021	2019-08-22 13:16:42	2019-10-01 10:27:12		
42	□	76	HYET	a-Si	Black Shakedown - 2 TC	M1907-0005	2019-08-22 13:16:42	2019-10-01 10:27:12		
43	□	77	HYET	a-Si	None	M1908-0013	2019-10-09 11:42:21	2020-07-27 13:27:09		
44	□	78	HYET	a-Si	None	M1908-0014	2019-10-09 11:42:21	2020-07-27 13:27:09		
45	□	79	PET-1	c-Si	Inner Layer Cracking PET-1	M1907-0002	2019-11-05 18:10:45	2020-07-27 13:27:09	2019-11-05 00:00:00	2020-06-30 23:59:59
46	□	80	PET-2	c-Si	Inner Layer Cracking PET-2	M1907-0004	2019-11-05 18:10:45	2020-07-27 13:27:09	2019-11-05 00:00:00	2020-06-30 23:59:59
47	□	81	FEVE	c-Si	Inner Layer Cracking FEVE	M1907-0003	2019-11-05 18:10:45	2020-07-27 13:27:09	2019-11-05 00:00:00	2020-06-30 23:59:59
48	□	82	ASU GG Project	c-Si	ASU G/POE/G	M2001-0056	2020-01-28 10:37:58	2020-07-27 13:27:09	2020-01-28 00:00:00	2020-06-30 23:59:59
49	□	83	ASU GG Project	c-Si	ASU G/POE/clearBS	M2001-0055	2020-01-28 10:37:58	2020-07-27 13:27:09	2020-01-28 00:00:00	2020-06-30 23:59:59
50	□	84	ASU GG Project	c-Si	ASU G/POE/WhiteBS	M2001-0057	2020-01-28 10:37:58	2020-07-27 13:27:09	2020-01-28 00:00:00	2020-06-30 23:59:59

☑ Extract Experiment and Monitoring
☑ Extract IV Curve data
☑ Extract EL files

Data Hub: DuraMAT

Target Project: ...Select One...

Project ID:

NOTE: Uploading the data will create a new dataset automatically if the module is new to the project.

Overview

The initial query retrieves a list of all the available modules in the C-AST database. The data is displayed as a table across the top of the GUI. In addition, there is a log file keeping track of uploads that have been performed to the data hub. This file is also versioned as part of the github repository of the code. The file method was chosen as an initial effort to provide a tracking method without relying on new tables in the C-AST database or in trying to read all possible projects within the data hub that might have C-AST data in it. At a later time, this process could be revisited.

Across the top of the table is the headers for all the base module information. DB First Timestamp and DB Last Timestamp are created by searching for the max and min values of any data taken for a particular module, across all the database tables. The Uploaded First Timestamp and Uploaded Last Timestamp are retrieved form the logs/history.csv file. This marks the first and last timestamps for data uploaded to the data hub for this module.

Selection checkboxes are provided in the first column, allowing the user to pick one or more modules' data to upload.

The checkboxes under the table allows the user to choose what data type is to be part of the payload to be pushed to the data hub. By default all the data types are selected

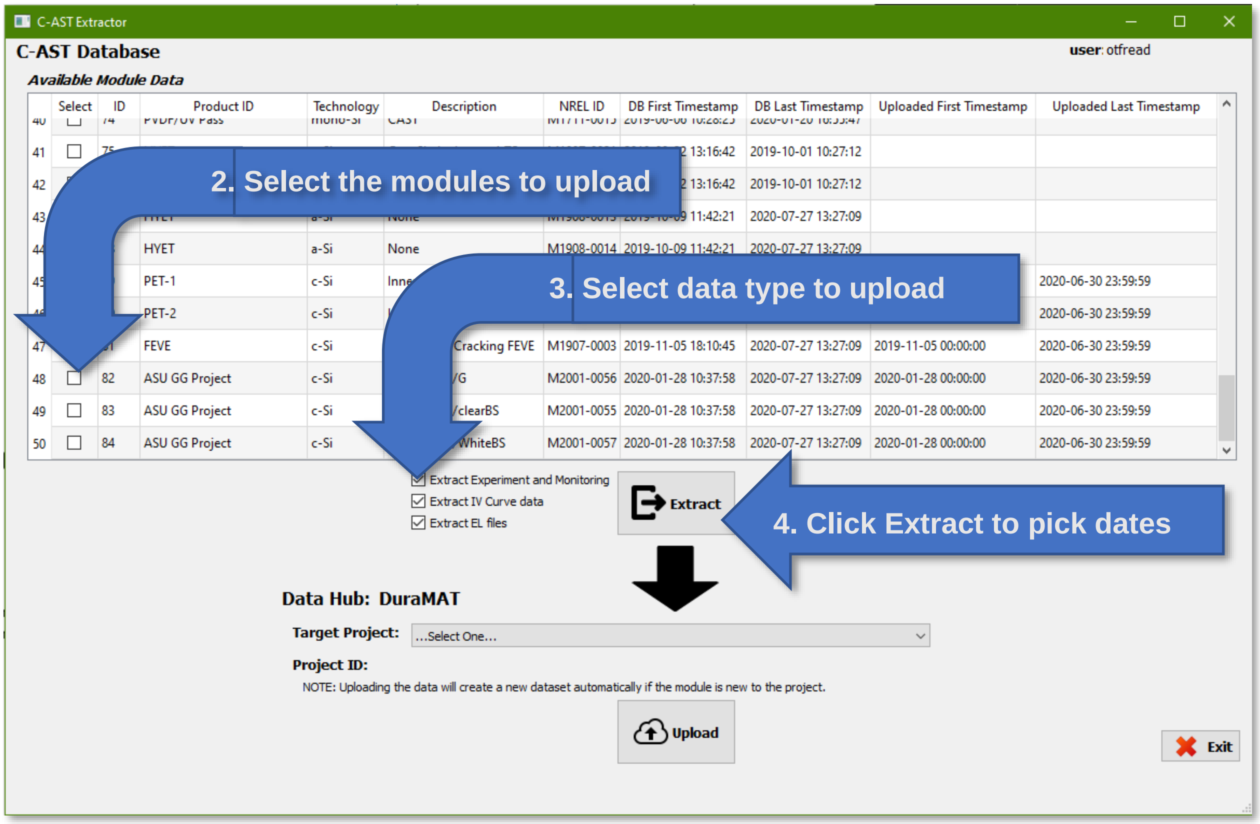
The large “Extract” button is used to initiate the extraction of all data products and their marshalling into the data hub payload.

The large black arrow is an indicator that will turn green on a successful preparation of data. If the arrow is red, then an error has occurred in the extraction process.

The lower part of the GUI is for the data hub push. There is a pull-down menu that lists all available projects the user credentials (in the access_config.xml file) can access. Once that choice is made the final button, “Upload”, will allow the prepared data to be pushed to the selected project.

Walkthrough

1. Start the software (as above, #6)



C-AST Database

user: otfread

Available Module Data

	Select	ID	Product ID	Technology	Description	NREL ID	DB First Timestamp	DB Last Timestamp	Uploaded First Timestamp	Uploaded Last Timestamp
40	□	74	PVDF/UV Pass	mono-Si	CAST	M1711-0013	2019-06-06 10:28:23	2020-01-20 10:33:47		
41	□	75	[illegible]	[illegible]	[illegible]	[illegible]	[illegible] 13:16:42	2019-10-01 10:27:12		
42	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible] 13:16:42	2019-10-01 10:27:12		
43	[illegible]	[illegible]	HYET	a-Si	None	M1908-0013	2019-10-09 11:42:21	2020-07-27 13:27:09		
44	[illegible]	[illegible]	HYET	a-Si	None	M1908-0014	2019-10-09 11:42:21	2020-07-27 13:27:09		
45	[illegible]	[illegible]	PET-1	c-Si	Inne[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2020-06-30 23:59:59
46	[illegible]	[illegible]	PET-2	c-Si	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2020-06-30 23:59:59
47	[illegible]	81	FEVE	c-Si	[illegible] Cracking FEVE	M1907-0003	2019-11-05 18:10:45	2020-07-27 13:27:09	2019-11-05 00:00:00	2020-06-30 23:59:59
48	□	82	ASU GG Project	c-Si	[illegible]/G	M2001-0056	2020-01-28 10:37:58	2020-07-27 13:27:09	2020-01-28 00:00:00	2020-06-30 23:59:59
49	□	83	ASU GG Project	c-Si	[illegible]/clearBS	M2001-0055	2020-01-28 10:37:58	2020-07-27 13:27:09	2020-01-28 00:00:00	2020-06-30 23:59:59
50	□	84	ASU GG Project	c-Si	[illegible] WhiteBS	M2001-0057	2020-01-28 10:37:58	2020-07-27 13:27:09	2020-01-28 00:00:00	2020-06-30 23:59:59

☑ Extract Experiment and Monitoring
☑ Extract IV Curve data
☑ Extract EL files

Data Hub: DuraMAT

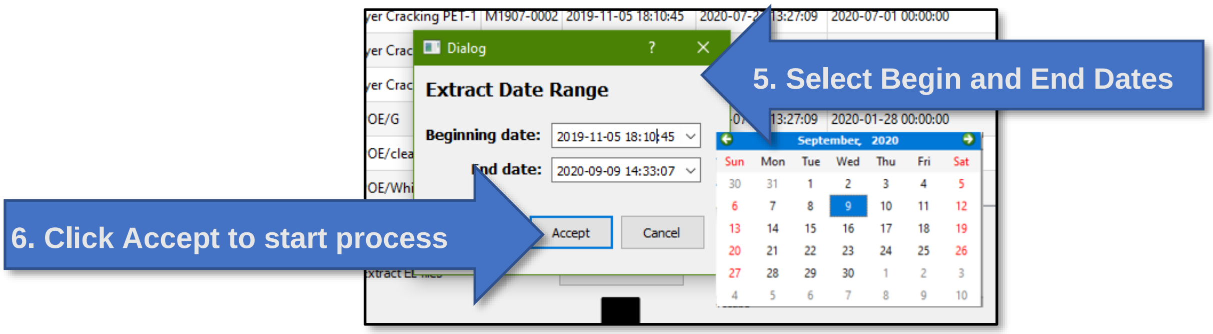
Target Project: ...Select One...

Project ID:

NOTE: Uploading the data will create a new dataset automatically if the module is new to the project.

Upload
Exit
2. Select the modules to upload
3. Select data type to upload
4. Click Extract to pick dates

A dialog box to pick the date range for the module(s) data will appear. A range must be selected before it will continue. By default, it will choose the start and end dates based on the first and last timestamps for the module(s) data in the database. Clicking the upside-down carat in each date field will bring up a calendar interface to make it easier to pick dates. *NOTE: To keep things consistent, choose the time for the* `Beginning Date` *as 00:00:00 and the time for the* `End Date` *as 23:59:59*. This gives a nice concise wrapper for the data. Users can always choose any time that is needed. Before selecting, it is good to examine the previous upload date ranges (Uploaded First Timestamp and Uploaded Last Timestamp) columns, to avoid picking any overlapping time period. This would duplicate data in the files being uploaded, with those already on the data hub.



Dialog

Extract Date Range

Beginning date: 2019-11-05 18:10:45

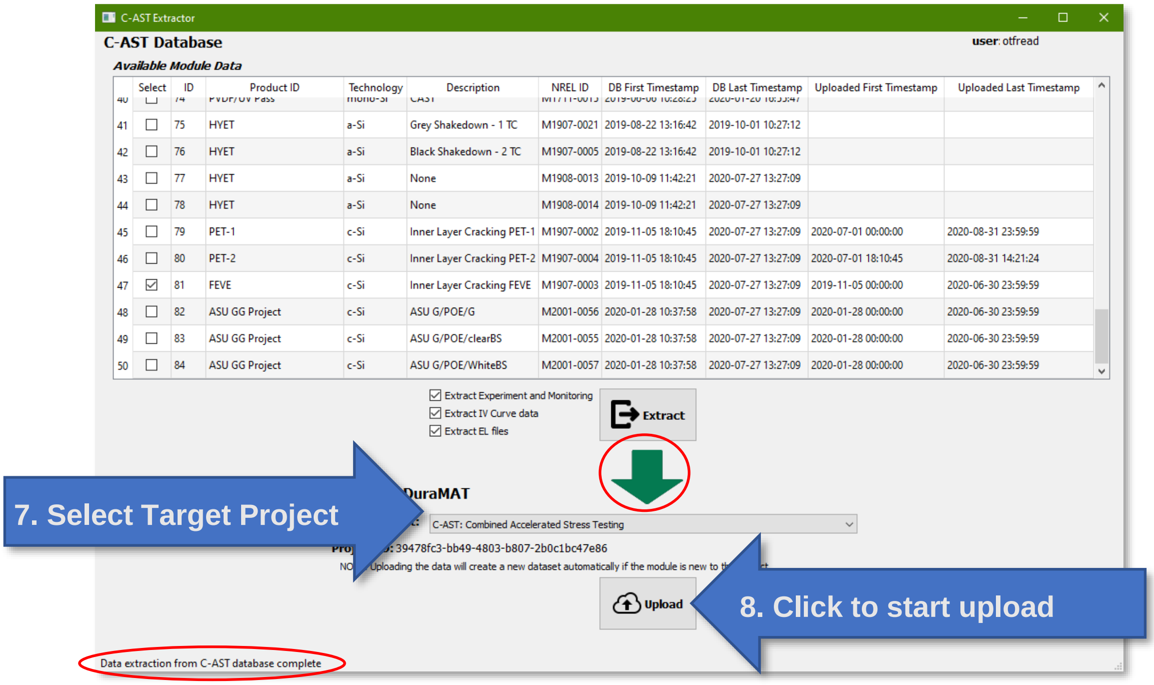
End date: 2020-09-09 14:33:07

Accept | Cancel

5. Select Begin and End Dates

6. Click Accept to start process

The extraction process can take quite a while to perform, depending on number of modules chosen, the data types selected, and the network bandwidth to the computer running the software. Once extraction is completed the arrow on the GUI will change color. Green for successful extract and red if there is an error. Also, at the bottom left of the GUI a success or error message will be displayed on the status bar (circled in red).



C-AST Extractor

C-AST Database — **user**: otfread

Available Module Data

	Select	ID	Product ID	Technology	Description	NREL ID	DB First Timestamp	DB Last Timestamp	Uploaded First Timestamp	Uploaded Last Timestamp
40	□	74	PVDF/UV Pass	mono-Si	CAST	M1711-0015	2019-06-06 10:28:25	2020-01-20 10:33:47		
41	□	75	HYET	a-Si	Grey Shakedown - 1 TC	M1907-0021	2019-08-22 13:16:42	2019-10-01 10:27:12		
42	□	76	HYET	a-Si	Black Shakedown - 2 TC	M1907-0005	2019-08-22 13:16:42	2019-10-01 10:27:12		
43	□	77	HYET	a-Si	None	M1908-0013	2019-10-09 11:42:21	2020-07-27 13:27:09		
44	□	78	HYET	a-Si	None	M1908-0014	2019-10-09 11:42:21	2020-07-27 13:27:09		
45	□	79	PET-1	c-Si	Inner Layer Cracking PET-1	M1907-0002	2019-11-05 18:10:45	2020-07-27 13:27:09	2020-07-01 00:00:00	2020-08-31 23:59:59
46	□	80	PET-2	c-Si	Inner Layer Cracking PET-2	M1907-0004	2019-11-05 18:10:45	2020-07-27 13:27:09	2020-07-01 18:10:45	2020-08-31 14:21:24
47	☑	81	FEVE	c-Si	Inner Layer Cracking FEVE	M1907-0003	2019-11-05 18:10:45	2020-07-27 13:27:09	2019-11-05 00:00:00	2020-06-30 23:59:59
48	□	82	ASU GG Project	c-Si	ASU G/POE/G	M2001-0056	2020-01-28 10:37:58	2020-07-27 13:27:09	2020-01-28 00:00:00	2020-06-30 23:59:59
49	□	83	ASU GG Project	c-Si	ASU G/POE/clearBS	M2001-0055	2020-01-28 10:37:58	2020-07-27 13:27:09	2020-01-28 00:00:00	2020-06-30 23:59:59
50	□	84	ASU GG Project	c-Si	ASU G/POE/WhiteBS	M2001-0057	2020-01-28 10:37:58	2020-07-27 13:27:09	2020-01-28 00:00:00	2020-06-30 23:59:59

☑ Extract Experiment and Monitoring
☑ Extract IV Curve data
☑ Extract EL files

Extract

DuraMAT

C-AST: Combined Accelerated Stress Testing

Proj…: 39478fc3-bb49-4803-b807-2b0c1bc47e86

NO… Uploading the data will create a new dataset automatically if the module is new to th…

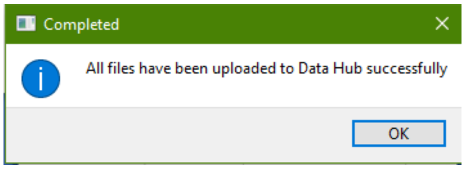
Upload

Data extraction from C-AST database complete

7. Select Target Project

8. Click to start upload

Once the upload process begins it may take a bit of time to complete, but it should be much quicker than the extraction process. Once it is complete a dialog box will appear to indicate if it was a successful upload or if there was a problem.



Completed

All files have been uploaded to Data Hub successfully

Once complete the GUI clears and the user may make additional uploads or may shutdown the software if all extraction and uploads are done.

The Data Extraction Process

During the data extraction the software makes a query into the C-AST instrumentation database using the module id's as the initial key. From that query the base information is found for each module to populate the main window table. Once the actual extraction begins it performs a series of queries using the selected module IDs. The first is to find all experiments for that module(s) within the data range chosen. It will then use each discovered experiment ID as a key to access the recipe for the experiment (stages) and all the data in the chamber monitoring tables in the database. Two files will be generated:

- exp_*number*_stages.csv
- exp_*number*_monitors_*startDate-endDate*.csv

Where *number* is the experiment number as listed in the database and *startDate* and *endDate* are the values chosen from the GUI by the user.

Using the module ID, the system will then make another query to the database to pull all the module monitoring data within the same date range. This will generate a single file:

- mod_*moduleID*_monitors_*startDate-endDate*.csv

If the data type "Extract IV measurements" has been chosen, then additional tables will be queried to pull out the calculated IV summary data and the actual IV curve data points. These will be merged into a single file, where each record represents a point in time and all the IV measurements at that point. There could be 1-> n records over the selected date range. Once assembled, the file generated is:

- mod_ *moduleID*_iv_measure_*startDate-endDate*.csv

If the data type "Extract EL Files" is chosen, then the software will access the otfdataserver file system, and target the directory where the EL files are stored (Note: This is one of the configuration elements in the access_config.xml file). It will scan the directory for EL files with the module ID and creation dates in the date range requested. The software will build up a list of filenames for the payload. During the upload process it will merge the metadata information for each image, found in the similar named csv file, with the actual image. This means the payload will eventually be just the images and not the csv files in the EL file archive.

The payload files are available to review, if needed, before upload. Prior to activating the upload, the user can see the files in the `cast_extractor/outbox` directory.

The Upload Process

During startup the list of available projects that can be accessed by the user credentials in the configuration file will be queried from the Data Hub, along with the Universal Unique Identifier of the project (UUID). When the user selects the project for upload from this list, the UUID will be displayed on the GUI under the pulldown menu. Should researchers wish to be able to access this data programmatically, then they may share this UUID with others. This provides the target address of the project and is accessible by software, assuming the new user has the credentials for access.

Once the user clicks upload, the software begins to assemble each element of the payload. This data will either be added to an existing dataset or the software will create a new dataset. Depending on the data type selected there could be 1 to 2 datasets per module in the project

- If the module has never had any data uploaded, a new dataset will be created. The dataset will be given a name of the *Module ID*. This will contain all the experiment data, chamber monitoring data, module monitoring data and module IV data
    - If the user has chosen EL files, then a separate dataset called "*Module ID* EL Files" will be created and all EL files will be pushed into it.
- If the module has had previous data uploaded, then the new data will be added to the existing datasets. Each upload would be visible through the start and end date tags on the filenames.

When the upload process is complete and verified, the copies of the payload file are deleted from the `cast_extractor/outbox` directory.

Improving the Software

All software for the extractor tool is in the github repo, including the historical upload text file. If a user wishes to contribute, they should feel free to clone the repo and work on the code. It is important that the contributor work in a branch and upon completion of that development, the contributor push only that branch up and submit a pull request for merger to the main production branch. After a code and compatibility review the code will be merged.

Problems

If you are seeing errors or problems, feel free to contact the lead developer at any time to see if the issue can be resolved or if we need to correct bugs in the code.

Robert.White@nrel.gov or 303-384-7802