ENCHILADA BASICS MANUAL

April 13, 2008
Complied by Deborah Gross

Enchilada can be downloaded from: http://www.cs.carleton.edu/enchilada/

Enchilada is created by Musicant, et. al, at Carleton College, and is funded by NSF-ITR II – 0326328. You can download two versions:

- A version that works with the full installation of SQL Server, a Microsoft database application which must be purchased.
- A version that works with SQL Server Express (included), a free version of the Microsoft database, which can handle databases no larger than 4 Gb.
- Installation instructions are provided at the website, for both versions.

Basic structure of Enchilada:

- When you launch Enchilada, two windows open
  - The Enchilada window, where work is done and data is displayed.
  - The Output window, where messages and logs are displayed. You can usually just minimize this and ignore it.
- Data in Enchilada is stored in "Collections" – think of these as folders of data, which can contain sub-collections. Operations carried out on higher level collections will include the data in sub-collections.
- The main Enchilada window includes
  - Pull-down menus at the top
  - Tabs below them which correspond to commonly used functions from the menus
  - A collection-tree at the top left, in which the collections and sub-collections can be displayed
  - A list of aggregated time-series data at the bottom left, which includes any aggregated time-series
  - A main pane, which includes the main information that Enchilada displays, including
    - Displaying individual particle spectra
    - Displaying time-series data
    - Displaying information about individual collections
    - Displaying information about clustering results
  - Various pop-up windows designed for specific tasks

Goals of This Document:

Currently, this document is a bare-bones "how-to" guide. It will be expanded with more information about the current features and descriptions of more features, as they are created. Be patient and please send corrections or comments to Deborah Gross (dgross@carleton.edu).

How to do the following in Enchilada:

1. Load data into Enchilada

a. Load ATOFMS data by dataset.
   ATOFMS data can be loaded into Enchilada directly from the raw TSI-format datasets in two ways: Through the File | Import Collection | From ATOFMS data menu and through the "Import ATOFMS Data" tab at the top of the Enchilada screen.
   i. Select the import method of your choice.
   ii. Click on the .par file button in the top-left cell.
   iii. Navigate through the popup to the .par file saved in your datafile and double-click on the .par file.
   iv. Click on the mass calibration entry next to the .par file, and navigate to the appropriate .cal file for your dataset. Double-click on the file to select it.
   v. Click on the size calibration entry next to the .cal file, and navigate to the appropriate .noz file for your dataset. Double-click on the file to select it.
   vi. Enter values in the cells to the right of the size calibration file, inputting the minimum height (Min. Height), minimum peak area (Min. Area), and minimum relative area (Min. Rel. Area) that defines the minimum criteria for a peak to be detected.
   vii. Check the box at the far left to use the Auto Calibrator (recommended).
   viii. Click on the far left cell in row two to add an additional dataset. By default, the calibration and peak-detection values will be filled in as you had done in the previous row.
   ix. *Recommended*: click the button in the lower left of the window to create a parent collection for all incoming datasets. When prompted, enter a name for the collection – a name for the experiment – and enter a comment. This will create a top-level folder with each dataset as a sub-folder below.
   x. Click OK to start the import.
   xi. A progress bar will be displayed at the top of the screen. You cannot use Enchilada while it is importing data.

b. Load ATOFMS data with bulk import.
   This accomplishes the same result as the process described in part a, for loading ATOFMS dataset by dataset, without having to enter the data in each cell.
   i. Using Excel or another program that can save .csv files, generate a file with one row per dataset, with the following data in the row (separated by commas). If using Excel, enter values in individual cells and save the file as a .csv file.
   ii. Each row/line of the file should include each of the following items, in order, separated by commas. To see the format, just start the process for importing a single dataset – the .csv file should follow the same format as the import table:
      1. The full path of the .par file for the dataset.
      2. The full path of the .cal file for the mass calibration file for the dataset.
      3. The full path of the .noz file for the size calibration file for the dataset.
      4. The value for the minimum peak height for peak detection.
      5. The value for the minimum peak area for peak detection.
      6. The value for the minimum relative peak area for peak detection.

iii. Save the file as a .csv file.
iv. Choose File | Import Collection | From ATOFMS data (with bulk file).
v. Navigate to the .csv file you saved that contains the information about the data you wish to load and open it.
vi. Enchilada will ask you a few questions:
    1. Whether to create a parent collection (i.e. a top-level folder) into which to import the data.
    2. Whether to use the Auto-calibrator for the data (recommended).
vii. A progress bar will be displayed at the top of the screen. You cannot use Enchilada while it is importing data.

c. Load data of the form (time,data)
This is the way to import data from instruments that provide data of the form (time,data). The time and date information must be concatenated into a single text value (or in Excel, into a single cell). This can be done through the File | Import Collection | From txt file or from the "Import Time Series" tab.
    i. Select File | Import Collection | From txt file or click on the "Import Time Series" tab.
    ii. Navigate to the text file to import.
    iii. Click OK.
    iv. A progress bar will be displayed at the top of the screen. You cannot use Enchilada while it is importing data.

d. Load Q-AMS data.
This is the method to import AMS data from standard – not high resolution – datasets. This can be done through File | Import Collection | From AMS data or by selecting the "Import AMS data" tab.
    i. Select File | Import Collection | From AMS data.
    ii. Click on each cell in a row to identify the dataset, timeseries, and masstocharge text files to import.
    iii. Click on the next row to import another set of data.
    iv. Click OK.
    v. A progress bar will be displayed at the top of the screen. You cannot use Enchilada while it is importing data.

2. Manage Collections in Enchilada.

Datasets and results from analyses are stored in "Collections" within Enchilada – each folder of data or analysis results is itself a collection. Collections are designed to contain all the units of data in sub-collections, so analysis carried out on a top-level collection will be carried out on each of the unique data elements in all of its sub-collections. See specific analysis tasks below for details of those.

a. Move a collection (and all of its sub-collections) by selecting it and choosing Edit | Copy. Then click on another collection into which you want to copy it and choose Edit | Paste. Be patient if you are moving a lot of data.
b. Delete a collection by selecting Collection | Delete Selected. All data and results contained in sub-collections will also be deleted.

c. Delete a collection and adopt the sub-collections into the next higher collection by choosing Collection | Delete Selected and Adopt Children.

3. View ATOFMS spectra in Enchilada.

a. Once ATOFMS dataset has been imported, the spectra can be viewed in one of two ways.
   i. If the raw data (the .amz files) are stored at the same path as they were when the data was loaded, the spectra can be viewed as raw (calibrated) spectra, showing all detected datapoints (spectrum view).
   ii. Whether or not the raw data is available, the data can be viewed as a bar-plot of detected peaks in which the height of the bar is proportional to the area of the detected peaks in the spectrum (peaks view).

b. To display a spectrum:
   i. Click on a collection in the top-left pane of the main Enchilada menu (the Collection tree).\
   ii. Information about the individual particles contained within the collection will be displayed in the main Enchilada pane, in the “Particle List” tab, with each row corresponding to a particle. The list is populated in groups of 1000 particles, with the range of particles displayed, and the total number in the collection, shown at the top right of the pane.
   iii. Click on a row in the particle list and click the “Analyze Particle” button at the bottom of the pane.

c. Within the Spectrum Display window:
   i. The positive mass spectrum is displayed on the top (in red) and the negative spectrum is displayed on the bottom (in blue), as peak area vs. *m*/*z* (peaks view) or as arbitrary intensity vs. *m*/*z* (spectrum view).
   ii. You can change the type of spectrum displayed (peaks or spectrum view) by clicking the radio buttons on the lower right of the window. If you don’t have the raw data in the path that Enchilada expects, only the “peaks” view is possible. Clicking on the “spectrum” view returns an error.
   iii. The *m*/*z* axis scale in both spectra can be zoomed. It is typically too broad initially.
      1. Zoom in by clicking in the spectrum and dragging the mouse pointer horizontally to define the x-axis range in which to zoom. The ends of the selected region must both be within the spectrum. Both the positive and negative spectra will zoom. The y-axis will auto-scale.
      2. Zoom out by clicking on “Zoom Out” at the bottom of the window, to zoom out one step, or the “Zoom → Default” button to return to the original spectrum display.
   iv. The detected peaks (positive and negative, sorted ascending) and the associated height, area, and relative area of the peak are displayed in the top right pane in the window.
      1. “Location” means *m*/z.

2. If you mouse over a peak, it is selected in the list (but the list will not jump to display the selected peak, so you may have to scroll down to find it).

v. You can click through the particles in the given collection by clicking on the "Previous" or "Next" buttons at the bottom of the pane.

vi. Interpret spectral labels in Enchilada: The middle-right and lower-right panes in the spectrum display window relate to peak labels. This is a problematic section of the program, so use it with caution. This feature can be disabled by unchecking the "Label Peaks" box in the lower right corner of the window.

1. The top pane suggests ions which have the same nominal mass as the *m*/*z* on which the mouse sits. The code knows the following to determine this:
    a. Lists of commonly detected ions.
    b. PAH molecular weights.
    c. The molecular weights of all atoms in the periodic table, including their natural isotope abundances.
2. Ion identification will be made based on both parent and isotope peaks, for example a peak at *m*/*z* 37 will be identified as C3+ and C3H+, with the C3+ peak actually referring to the $^{12}C_2{}^{13}C^+$ peak.
3. IMPORTANT NOTE: The code has a problem in that the isotope peaks are not used correctly. Using the example above, if there was a spectrum which included a peak at *m*/*z* 37 but no peak at *m*/*z* 36, the code will still suggest C3+ as a possible identification. It doesn't know to confirm the presence of the most abundant isotope when suggesting less abundant species. BE CAREFUL AND CHECK ALL SUGGESTED ASSIGNMENTS YOURSELF. There are also errors in the list.
4. In the lower right "Signatures" pane, specific ions in the list can be excluded as potential ion formulae by unchecking them. This can be done separately for positive and negative ions, and limits are saved with each particle.

4. **Cluster ATOFMS mass spectra.**

ATOFMS spectra can be clustered using a variety of algorithms. They are all accessed through the same menu/window.

a. Select a collection of ATOFMS data to cluster.

b. Select Analysis | Cluster.

c. In the resulting window, select the algorithm of choice:

i. Art-2a

ii. K-clustering

1. K-means
2. K-medians

d. Select "Normalize data"

e. Depending on the algorithm selected, you will need to make further selections:

i. Art-2a

1. Pull down the desired distance metric from the menu:

a. Euclidean Squared

b. City block

c. Dot Product

2. Enter values for the learning rate, the vigilance factor, and the maximum number of passes to take through the algorithm.

a. **Vigilance Factor:** 1.0

b. **Learning Rate:** $\alpha = \frac{1}{\text{expected \# points per cluster}} = \frac{1}{m/|S|}$ where *m* is the number of points, and |*S*| is the expected number of clusters. In general, if the user does not have a good intuitive sense of how many clusters to expect, an "arbitrary but small" learning rate (such as 0.01) can be used to run a dummy pass to get an estimate of |*S*|.

c. **Number of Passes:** 50. If it converges, it will stop on its own. If it has high error at 50 passes, it should be redone with more. This is a challenge with this algorithm.

3. Art-2a will generate an additional “unclassified” cluster for those particles which are not within one of the determined clusters.

ii. K-clustering

1. Pull down the menu to select the desired algorithm (K-means or K-medians) and the desired distance metric.

a. K-means/Euclidean Squared

b. K-medians/City bloc

c. K-means/Dot Product

2. Enter the number of clusters (K). All particles in the data being clustered will be assigned to exactly one of the K clusters.

3. Check “refine centroids” to pre-cluster a subset of the data, to determine good starting cluster centroids to initiate the clustering. Note that in the K-means/Euclidean squared algorithm, refining the centroids produces a result that has a greater distance between particles and centroids than does a clustering run that does not include “refine centroids.”

iii. Insert a comment about the data and/or the clustering run, which will be used to label the collection containing the results.

iv. Advanced …????

v. Click OK to start clustering. This could potentially take a long time (many hours). You cannot use Enchilada (or do much on your computer) while it is clustering.

1. All algorithms can cluster datasets of ~10^5 particles.

2. K-means is the only algorithm that has been tested clustering ~$2x10^6$ particles. It ran overnight on a standard desktop PC.

f. Clustering results can be found as follows:

i. The cluster centers, which can be viewed as individual spectra, in the “peaks” view, are found in a new collection at the top level, labeled “Centers:…” and including comment and data info in the name. The cluster centers are located here in order that they not be treated as individual particles in future clustering or other analysis.

ii. The particles which are assigned to each resulting cluster are found in a new set of nested collections within the top-level collection that was clustered.
    1. A new collection will be created which has a name including the algorithm used and the parameters set.
    2. The top level collection contains one sub-collection for each cluster obtained, numbered 1 – K (where K is the number of clusters obtained).
        a. The "Particle List" tab displays all the particles that were clustered.
        b. The "Collection Information" tab includes a great deal of useful information, all as text, which can potentially be quite long:
            i. The top includes information about the clustering run, including specific information about particles that contain no data, number of clustering passes, etc.
            ii. The second section includes the average distance between each particle and its centroid, calculated on each pass of the clustering algorithm. The clustering stops when the decrease in distance reaches a pre-set value. The bottom of this section states the "average distance of all points from their centers on final assignment," which is a measure of the success of the clustering run.
            iii. The next section, labeled "Peaks in Centroids," reports the population of each cluster, as a number of particles.
            iv. The next section, labeled "Centroid 1" through "Centroid K," reports the *m*/*z* and average area values for each centroid. These data can be copied and pasted just by selecting the text and copying using regular Windows commands. It can be pasted into another program for graphing if desired.
    3. The population of the cluster centers can be displayed at any time by selecting a cluster collection.
        a. Particles within the sub-collections can be viewed and displayed as described above for individual particle spectra, though the "Particle List" tab. If "Previous" and "Next" are used to click through multiple spectra, only those within the collection (the cluster) are displayed.
        b. The *m*/*z* and average area values for the centroid can be found by selecting the "Collection Information" tab, referred to as key and value, respectively. These data can be copied and pasted just by selecting the text and copying using regular Windows commands. It can be pasted into another program for graphing if desired.

iii. Visualizing cluster population-grams can be done through the Analysis | Visualize menu.
  1. Select a collection that contains a single cluster in the Collection-tree.
  2. Choose Analysis | Visualize.
  3. A graph that displays the population of particles with a given peak area at a given *m*/*z* value in the collection will be displayed.
    a. Each pixel in the main graph (positive and negative) is colored in gray scale, with black indicating more particles, and is located at a point in the (*m*/*z*, peak area) graph, indicating the relative number of particles with that combination of values.
    b. Under each graph, there is a foreshortened plot that is a frequency distribution of the appearance of peaks at differing *m*/*z* within the cluster.

5. Aggregate time series data

These features can be used in two ways, to collect data into user-defined time bins, which can then be visualized and exported:

a. To collect ATOFMS data into user-defined time-bins by counting particles within a collection within each time bin
  i. Select the collection to aggregate in the Cluster-tree.
  ii. Click the "Aggregate Selected" button under the Collection-tree.
    1. In the Aggregate Collections window, enter a name/comment for the aggregated data.
    2. Confirm that the correct collections are selected.
    3. Select the relevant information for the aggregation on the right side of the window.
      a. Whether the data value should be combined by summing or averaging the data value.
      b. Whether a "particle count" time series should be generated, in which the number of particles in the given collections in the time-bins is the output. This can be combined with other output.
      c. Whether or not to aggregate specific *m*/*z* values. If this is selected, you have two options
        i. Enter specific *m*/*z* values in the box below, which will be aggregated by summing or averaging, as selected above. Separate multiple values with commas. For negative ions, include the negative sign.
        ii. Leave the box blank to aggregate all *m*/*z* values.
    4. Select the time basis from the bottom portion of the window. This can be done in two ways:
      a. Match to a particular collection.

b. Manually enter start, stop, and width information for the time bins.

b. Datasets of the (time,data) type can be aggregated into different, user-defined, time bins. This is currently supported only for time-bins that are larger than the time between data points. Interpolation is not included. This can be done by using the methods described above for single-particle data, but there is no meaning to aggregating specific *m*/*z* values.

c. Visualizing aggregated time-series data.

   i. Select a collection from the “Synchronized Time Series” pane in the lower left portion of the main window.

   ii. The right side of the main window will display a time-line of the selected aggregated collection.

   iii. A second aggregated collection from the same parent collection can be overlaid with this timeline by selecting it from the list labeled “2^{nd} Sequence” at the top of the pane.

      1. The parent collected is listed but should not be selected.
      2. Click “refresh” after changing which sequences are displayed.
      3. If two sequences are displayed, an x-y plot of “2^{nd} Sequence” vs. “1^{st} Sequence” is displayed below, with a linear regression and r^2 value.

   iv. Conditions can be applied to the displayed collections, such that data will only be displayed that match the defined conditions (eg. that the value for a particular collection must be greater than a constant).

      1. If desired, enter up to two conditions from the pull-down menus at the top of the pane.
      2. Click refresh to update the plot.

   v. Time-series data can be exported as a .csv file by clicking the “Export to .csv” button.

      1. You will be prompted for a location to save to and a file-name.
      2. Note that time bins that contain no data are currently skipped, so that if the data is graphed in Excel, sections of the time series which should be blank will in fact be connected. This can be fixed in Excel (and should be fixed in Enchilada).

6. Back up databases and management Enchilada

There can only be one database open in Enchilada at a time. It can be backed up and restored, which is the method of choice for moving a database to another computer or exchanging it with another user.

a. In Enchilada, go to the File menu and click Backup/Restore.

b. Add a location to backup to, click the location that has appeared on your list of Backup locations, then click the backup to selected button.

c. Go to Backup/Restore again, click your backup location and then click Restore from Selected.