7.5 OpenClinica Data Extract File Format

OpenClinica Data Extract File Format

When data contain non-ASCII characters, you may encounter character viewing issues on extracted files. Here is the rundown:

- CDISC ODM XML 1.3 Full with OpenClinica extensions
  - Converts into Decimal values with Character Entity marker (&#), i.e., １２３４５
- CDISC ODM XML 1.3 Clinical Data with OpenClinica extensions
  - Converts into Decimal values with Character Entity marker
- CDISC ODM XML 1.3 Clinical Data
  - Displays as expected - see note below if Windows
- CDISC ODM XML 1.2 Clinical Data with OpenClinica extensions
  - Displays as expected - see note below if Windows
- CDISC ODM XML 1.2 Clinical Data
  - Displays as expected - see notes below if Windows
- View as HTML
  - Displays as expected - see notes below
- Excel Spreadsheet
  - Tab delimited text file. This can be displayed with workaround
- Tab-delimited Text
  - Displays as expected - see notes below if Windows
  - If opening with Microsoft Excel, see this workaround
- SPSS data and syntax
  - Tab delimited text file. This can be displayed - see notes below if Windows
- Datamart in a downloadable format
  - Currently not fully compatible - see notes below
- Datamart
  - Currently not fully compatible - see notes below
- Discrepancy Notes CSV Export
  - Converts into Hex values with Unicode Escape marker (u), i.e., uff11uff12uff13uff14uff15
- Discrepancy Notes PDF Export
  - Currently not compatible 17230

File Encoding Issue

If you encounter issues viewing UTF-8 characters where they are expected to display correctly, you may need to specify the encoding.

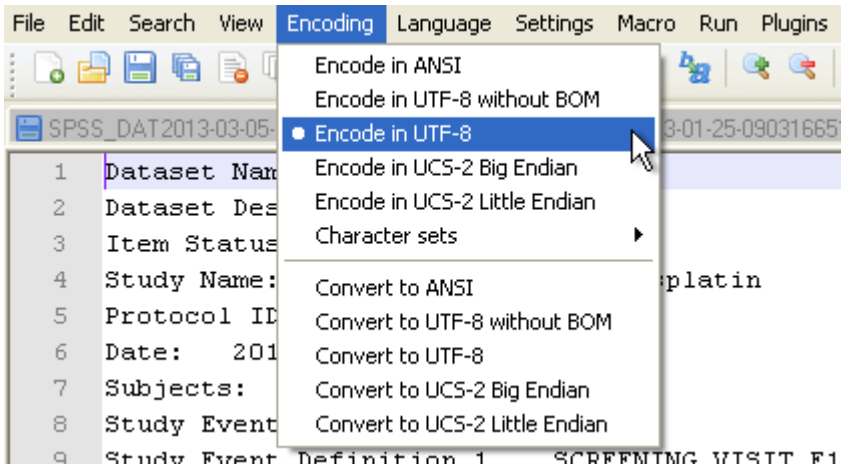
For example, if you open OpenClinica exported **HTML** file and see garbled corrupted characters, you need to set your browser encoding to UTF-8 to view those characters correctly.

Opening files on Windows machines

The native file encoding on Linux, Unix, Mac OS and popular databases such as SQL and Oracle is UTF-8 (Big Endian); however, Microsoft Windows' native file encoding is UTF-16LE (Little Endian). Depending on your text editor, this can become an issue because Java runs on OpenClinica server are UTF-8, not UTF-16LE.

If you open a file that contains non-ASCII characters and the file itself does not declare the encoding at the file binary header, the OS will try to determine with which encoding the file is written. Non-Windows OSes have an UTF-8 character map library in its OS level to determine the character map when opening the file, while Windows does not.

If you see garbled UTF-8 characters in ODM 1.3, ODM 1.2-Ext, ODM 1.2, Tab-delimited Text, and SPSS .dat files, you may need to **SaveAs** with the file encoding specified to UTF-8. Popular text editor such as Notepad++ (Win) and TextWrangler (Mac) will enforce encoding declaration at the file binary header level.



BOM Option

BOM (Byte Order Mark) can be critical on Windows environment. Unicode on Linux, Unix, Mac OS and popular databases such as SQL and Oracle is UTF-8, which is Big Endian byte order by default; Windows chooses UTF-16 Little Endian byte order.

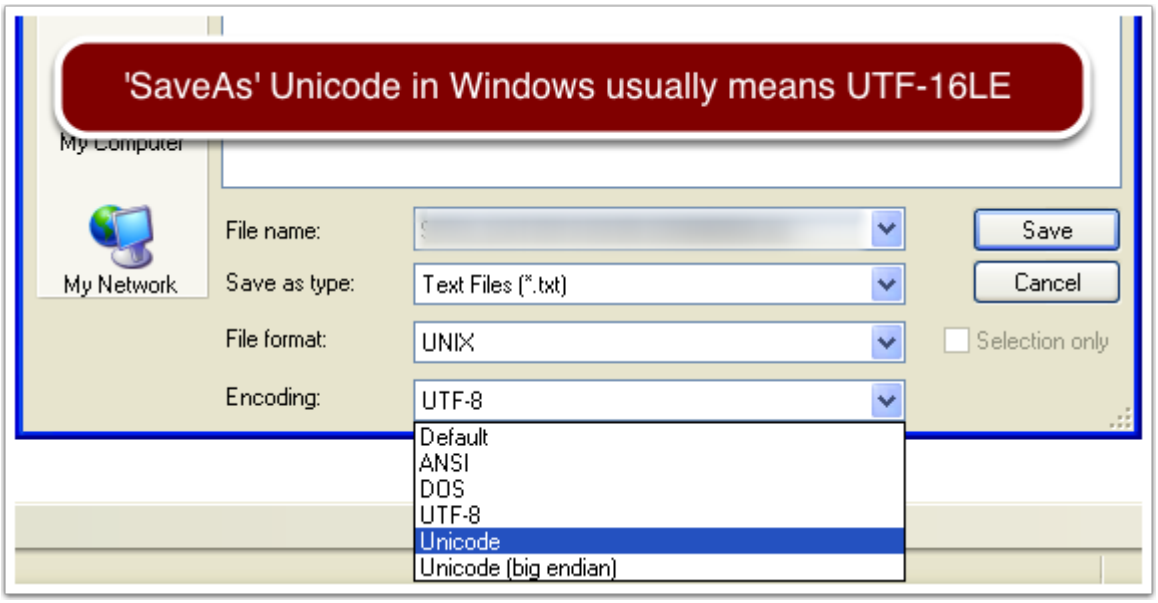
Your text editor should give you the option to SaveAs 'UTF-8 with BOM' and 'UTF-8 without BOM'. In our experience, this is somewhat hit or miss. Logically, it should work better with BOM but sometimes it seems to confuse Windows. You may need to experiment with the option of 'with' and 'without' BOM to find which option works on your Windows environment.

Excel issue

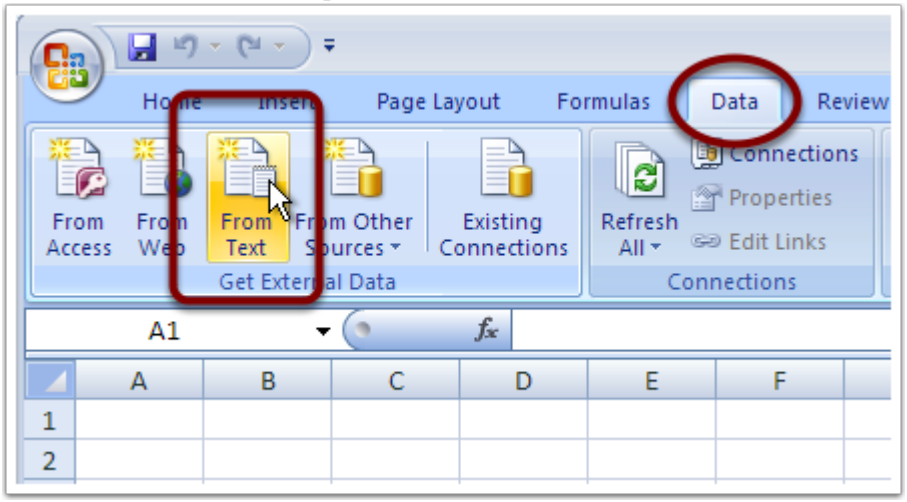
Not only does Excel not offer an encoding option, it doesn't seems to understand UTF-8 encoding. Even on a Mac OS platform, where UTF-8 is the native encoding, Excel cannot display non-ASCII characters unless file is encoded in UTF-16LE.

Workaround 1

1. Open the .xls file with a text editor of your choice
2. SaveAs UTF-16LE encoding

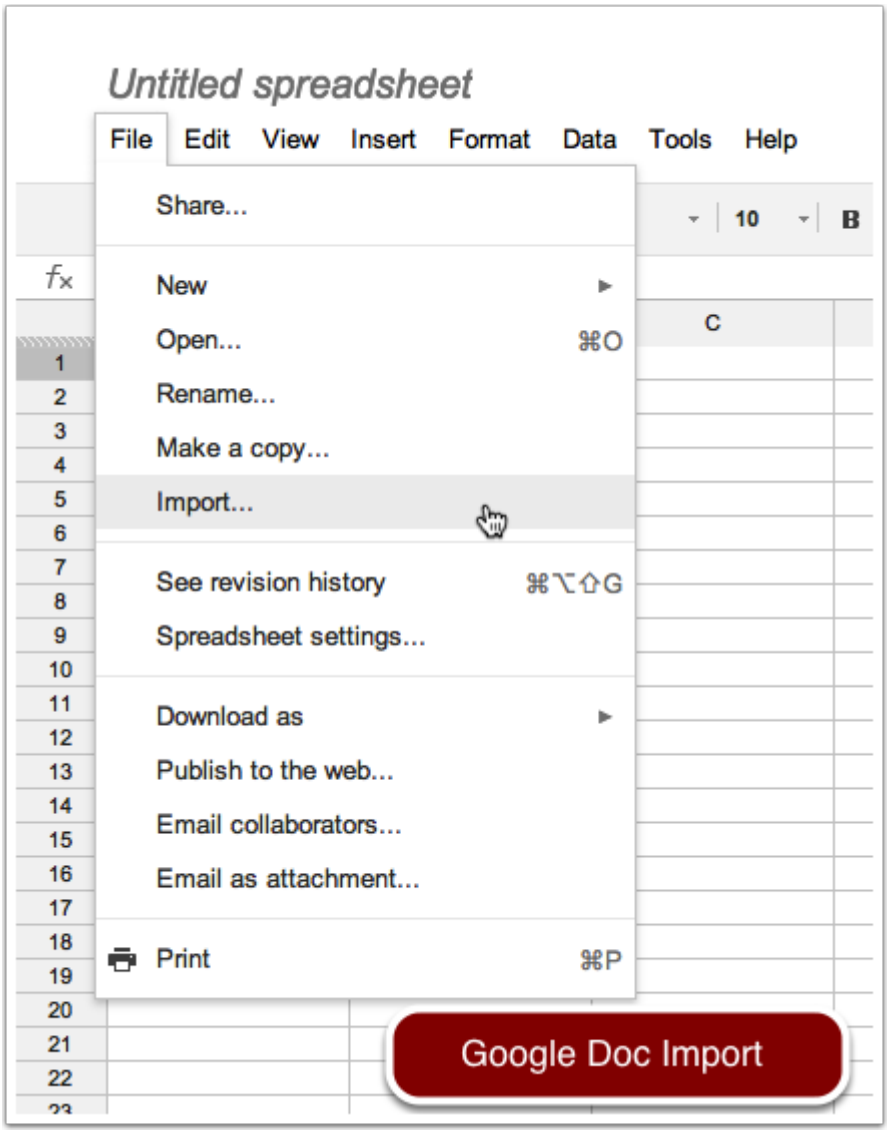


3. Open with Excel application
    - If .tsv instead of .xls, import the data into a new Excel file

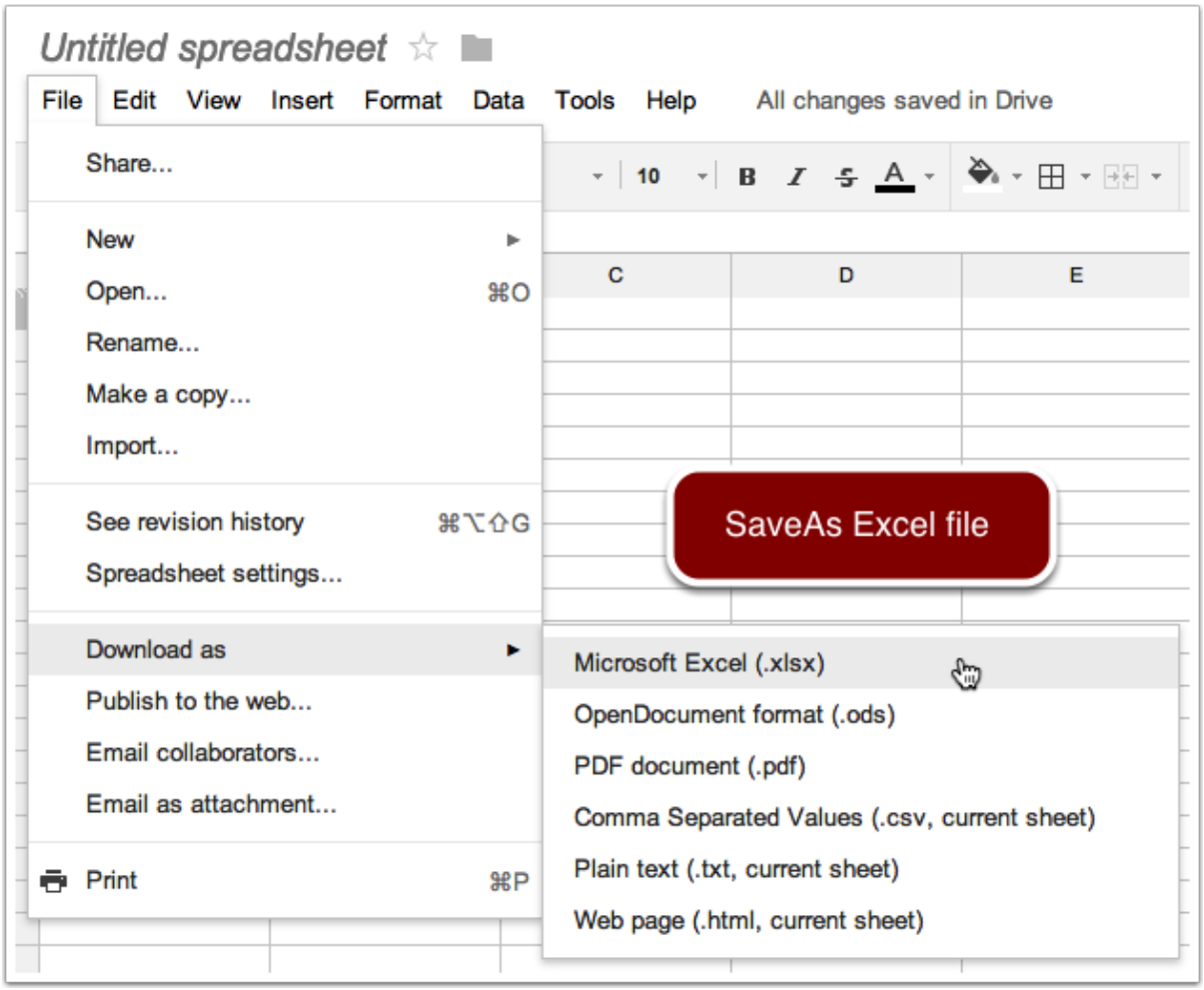


Workaround 2

1. Import to Google Doc Spreadsheet
    - Google Doc is native UTF-8 and correctly identifies UTF-8 encoded files



2. Select 'Download As' ? Microsoft Excel
    - Google Doc successfully embeds the encoding declaration binary header when the file is re-saved to your local directory



3. Open with Excel application

Troubleshooting

- If you see white boxes, e.g., ??????????
  - This is an indication of font problem. Your OS may not have Unicode mapped fonts.
  - This is a typical issue with Windows XP
  - You need to obtain Unicode font and install on your Windows
    - Arial Unicode Font is Microsoft default Unicode font
- If you see one or more white boxes in a recognizable i18n string, e.g., ????
  - This usually means incomplete Unicode font is assigned, often seen when the properly encoded file is opened with Excel
  - Select All and reassign known working Unicode font

Data Mart Issues

Currently, OpenClinica Data Mart function converts non-ASCII characters used for Table Names and Column Names into underscore character to avoid possible database issues.

It was designed this way for occasional non-ASCII character appearances among ASCII characters in a string, such as European word with accented characters. It was never meant for 100% non-ASCII string such as Asian languages.

If 100% non-ASCII string, all the entries become a series of underscore characters, which ends up with duplicated Table/Column names. We are hoping to resolve this issue as soon as possible.

On the other hand, data will not be affected by this. You can have Unicode characters in data string, and Data Mart will work as expected.

In summary:

- Any string that becomes a Table Name needs to be ASCII such as CRF name and Item Group name.
- Any string that becomes a Column names needs to be ASCII such as Multi-select Response Text and Item Name.
- Study Name becomes a series of underscore characters if non-ASCII

If a series of underscore characters become duplicated entries, Data Mart in a Downloadable Format output file will error when importing to Postgres. On the other hand, Data Mart extract operation silently stops during the operation without error message, leaving the data output incomplete (17249).

Data Mart in a Downloadable Format on Windows

Even if your .sql output file from Data Mart in a Downloadable Format does not contain any offensive underscore characters, remember Windows may require you to modify the file encoding as discussed above. This is not an issue when Postgres/pgAdmin III is running on Mac OS and/or Linux OS.