2006-03-17

Best practices for producing datasets

Formats Group, Deep Blue

Best practices for producing datasets

Version 1.0, 17 March 2006

Datasets, important to many areas of research, get produced and shared in a wide variety of applications and formats. To assure that we can maintain their usability into the future, we will require substantial descriptive information about the dataset and the use of standard file formats for the data itself.

Note also that **when your datasets involve human subjects there are serious ethical and legal considerations to address** before depositing them in Deep Blue. If your dataset contains information about human subjects, **you must** consult with the Institutional Review Board <http://www.irb.research.umich.edu/> before depositing your work to assure that you have taken the appropriate measures to de-identify the data you've collected.

To create preservable datasets

UM's Institute for Social Research has prepared a *Guide to Social Science Data Preparation and Archiving*. Available at http://www.icpsr.umich.edu/access/dataprep.pdf, it identifies all the major issues for creating well-formed and preservable datasets. Chapter 5, "Final Project Phase: Preparing Data for Sharing" gives detailed instructions on what you need to do, and is broadly applicable to

More generally, the following formats are appropriate for numeric data:

- For data held in a statistical package: **SPSS - portable (.por)** or **system (.sav)** file
- Data held in a database: **Delimited text with SQL setup** (tab delimited or comma separated)
- Data held in a spreadsheet: **Delimited text** (tab delimited or comma separated)
- Textual data: **Rich text format (RTF), XML, SGML, plain text**

General recommendations

Keep the following considerations in mind regardless of the method you use or the type of file you generate:

- What other applications are capable of opening or manipulating your files? If only the application that generated the file can open it, Deep Blue can only make an "As-is" (Level 3) preservation commitment, and while future researchers would have access to such files, without the original application they may not be able to use them or convert them to a form they can work with.

- Are you using
    a) off-the-shelf software (e.g. consumer-oriented products typically available at retail stores);
    b) open source software (available through places like SourceForge), or;
    c) custom software written by and for specialists?

    Typically as you move from (a) to (c), it becomes more important to check whether you can export your files to a standard format. This export may result in some loss of the features the originating application provides, but that's in exchange for the increased assurance that your file is usable by others, both now and in the future. For off-the-shelf and open source software, Deep Blue can usually provide at least "Limited" (Level 2) preservation support, where we will monitor the files they generate and make an effort to transform them when a significant risk to access is imminent.

- What export formats (available under "Save As..." or "Export" functions) are available in the applications you typically use to create and maintain the file?

    Typically, the more the better, and in each specific case of preparing a file to deposit, be sure to ask yourself the following question: Can I save my results in a format that is supported in Deep Blue at the "Highest Level" (Level 1) without significant loss of function?

    Example: If you have generated a spreadsheet, can you save it in tab-delimited form (.txt) rather than Excel's native form (.xls) without significant loss of function? If so, you should save it that way, since we can commit to both preservation and migration of the .txt form of your work, and ensure that it's available on many platforms and via many programs both now and in the future.

Questions?

If you have any questions, please contact us at deepblue@umich.edu and we will be happy to help you.