# NLS data advice sheet

## Overall:

Text collections:

* If exploring the text collections, I’d recommend downloading the text files rather than the ALTO XML – although, you may find the images useful for design and to understand what the texts originally looked like
* Consider OCR issues. Some collections have had their OCR cleaned up; others haven’t. Some basic clean-up using regular expressions could be helpful, or some recognition that not all of the text is accurate. Given the scale of the texts, however, trends should still emerge in the texts despite any OCR inaccuracies. More on OCR: <https://en.wikipedia.org/wiki/Optical_character_recognition>

Metadata collections:

* Information about MARC is here: <https://www.loc.gov/marc/> and here: <https://en.wikipedia.org/wiki/MARC_standards>

## Encyclopaedia Britannica:

If you want to get a general feel for what the Britannica looks like, without downloading the image files, take a look at: <https://digital.nls.uk/encyclopaedia-britannica/archive/188936619>

When you download the dataset, you will find we have included a few too many files (I need to fix this in the dataset soon!). If you consult the inventory csv file in the zip file, you can work out which files to remove before beginning your analysis:

* The two third editions are essentially the same: they can be distinguished in minor bibliographic ways but the actual content should not vary. This was simply a mistake with digitisation: two were picked up from the shelf and digitised, rather than one. You can use one of them and discard the other.
* The ‘supplements’ were a way to update many articles without producing an entirely new edition, so they appear in a much smaller number of volumes than the full edition. I believe these are actually duplication in the digitisation process, also, so if you choose to use them you would only need to explore one of them for each edition. Given the time you have I would suggest removing these files, though! There’s some more information about these supplements here if you’d like to know more: <https://en.wikipedia.org/wiki/History_of_the_Encyclop%C3%A6dia_Britannica#Supplement_to_the_fifth_edition,_1824_(later_known_as_the_supplement_to_the_fourth,_fifth_and_sixth_editions)>
* Also, you’ll see there are two first editions: the 1773 edition used unsold sheets of the Edinburgh 1771 edition, with new preliminaries and a new preface. Therefore the actual encyclopaedia articles shouldn’t vary from the 1771 first edition, so you can discard the 1773 edition.
* You might find this useful for some context to the data: <https://en.wikipedia.org/wiki/History_of_the_Encyclop%C3%A6dia_Britannica>

## Broadsides

* This collection is unexplored, but similar in content in many ways to the Chapbooks collection: <https://data.nls.uk/data/digitised-collections/chapbooks-printed-in-scotland/>, which was topic-modelled by an academic; <https://data.nls.uk/projects/scottish-chapbooks-as-music/>
* More about this collection (also known as ‘Word on the Street’) here: <https://digital.nls.uk/broadsides/>

## Spiritualist Newspapers

* It’s worth making sure you understand what Spiritualism was/is before starting to visualise this collection, so you can think of suitable questions to ask of the data. Eg. <http://www.victorianweb.org/victorian/religion/spirit.html>

## A Medical History of British India

* More contextual information here: <https://digital.nls.uk/indiapapers/>