**<title>Cleaning the Zoo: Preparing raw text for a corpus using regular expressions</title>**

John Williams, Portsmouth University

This talk/demonstration is based on the experience of the Language of Citizen Science research team at the University of Portsmouth in downloading and preparing data from the Zooniverse online forums (www.zooniverse.org) for use as corpora. The format and implementation of the forums was such that the data could not be downloaded as html files using standard tools, but only by automating the copy and paste function to create plain-text files. This resulted in a lot of unwanted material in the files: sidebars, metalanguage, boilerplate text... Cleaning up the files involved some painstaking work identifying recurring patterns, expressing them as regular expressions, then applying the Unix 'sed' (search-and-replace) command in shellscripts to eliminate the unwanted text. Judicious use of regular expressions also allowed us to add some limited metadata in angled brackets, such as the poster of each message (which in turn allowed us to generate separate corpora for each poster).

Each corpus-building project will raise its own specific challenges in preparing and cleaning up data. No two projects are identical in this regard. However, many of the principles are the same across projects. It is therefore hoped that the Unix scripts and commands we devised, soon to be in the public domain, will serve as models for future corpus researchers.

Anyone with a Linux computer, a Windows bash shell, or a Mac (?) will be able to follow the demo on their own device.