Data Curation in the School of Informatics

Version 1.0 – 5/12/14

Craig M Strachan

Background

Currently, we have three active development projects, all assigned to myself, related to aspects of data curation. These are:

- 307 – Produce a MHR data asset register – requirements and design phase
- 290 – Research Data Audit
- 194 – Data Archiving – produce report

It will immediately be seen that projects 307 and 290 have much in common, though 290 has a far greater scope. Both are concerned with the auditing of a dataset, all research date in the case of 290 and all Medium and High risk data in the case of 307. Both require that decisions be made regarding what information is to he recorded for each type of data and both require that some kind of data asset register is available to record the information once collected. Project 307 specifically covers just the design of these areas whereas project 290 covers the entire audit from design and requirements capture, implementation of the Data Asset Register, the actual gathering of information and its entry into the data asset register. It could be argued that the scope of this project is too wide ranging and that it would have made more sense to split it into at least two projects, one covering the requirements and design phase in a similar manner to 307 and the other covering the actual data capture. The third project, 194, is intended to produce a report on how the archiving needs of the School can be met. What these needs actually are depends on what data we hold. Some datasets such as home directories have clearly defined definitions and retention periods but others, such as research data and MHR data do not. In short, before we can identify the archiving needs of the School, we need to know what data we have, including the outcome from the other two projects.

Progress on all of these projects has been slow, in large part because of the need to see if the centrally provided data curation services in development would prove suitable for our needs. In an effort to ensure that this was indeed the case, the School engaged heavily in the design and planning stages of the central Data Asset Register and Archive service with what seemed like promising results. Regrettably, the decision was recently taken to put both these services out for further consultation leaving us with no prospect of a central Data Asset Register and Archiving service becoming available in a reasonable time frame.

The need for a DAR became all the more pressing with the release of the College of Science and Engineerings action plan on Computing Services and Data Security. Amongst much else, section 4.1 of this document proposes specific action at School level to:

produce initial registers of datasets, websites, servers and services, their primary locations and ownership, highlighting the most sensitive or vulnerable for special attention

It will be seen that this proposal encompasses all of the work covered by projects 307 and 290 (and a lot more besides).

Since it makes little sense for Schools to do this work in isolation, CCPAG have set up a working sub-group to "*work out the best way to define the nature of the register of datasets in such a way that it should be agreeable to all Schools in the college"* of which the current author is a member. It is to be sincerely hoped that the activities of this sub-group will extend beyond its original remit to encompass the actual defining of the nature of the register and indeed its implementation. More may become clear after the initial meeting of the sub-group which is currently looking likely to happen

early in the new year.

Recommendations

These projects have already been significantly delayed by a desire to avoid duplicate effort by waiting for the outcome of external projects. Any further delay is undesirable, especially in view of the new external pressure from College in this area. But it would be equally undesirable to go ahead on our own, lose the chance to benefit from the efforts of other Schools in the College and potentially end up with something which does not meet the College's needs. It seems that we must once more wait, this time on the outcome of the deliberations of the CCPAG sub-group. I therefore suggest that project 307 which most closely matches the aims of the sub-group remains active to record the effort spent by this author in being a member of the sub-group and in carrying out any actions allocated to him by the sub-group and that projects 290 and 197 are suspended until the outcome from the sub-group becomes clear.