# Project 3.8

# AMASED: Access methods for analysing sensitive data

Lead: University of Bristol
Collaborators: London Metropolitan University, Content Mine, British Library, F1000 Research
Stakeholders: data providers, researchers, research libraries, publishers etc.

## Summary

We propose the adaptation and implementation of the open-source software DataSHIELD (developed for securely analysing biomedical data) to circumvent key obstacles preventing or limiting the open analysis of digital datasets in the humanities and academic publishing. Irrespective of discipline, data access and analysis barriers result from a range of considerations:

- ethical-legal restrictions surrounding confidentiality and the sharing of, or access to, disclosive data;
- implications of extensive professional investment, intellectual property issues or licensing conditions constraining unconditional access to raw data;
- the physical size of the data is a limiting factor.

DataSHIELD (www.datashield.ac.uk) was born of the requirement in the biomedical and social sciences to co-analyse individual patient data (microdata) from different sources, without disclosing identity or sensitive information. Under DataSHIELD, raw data never leave the data provider and no microdata or disclosive information can be seen by the researcher. The analysis is taken to the data - not the data to the analysis. It provides a flexible, modular, open-source solution ideally placed to grow a broad user and development community.

DataSHIELD has been successfully piloted in two European biomedical studies and is of proven value in the biomedical and social science domains, but its potential utility is wider than this. Issues of data sharing and confidentiality – particularly relating to extensive professional investment, IP and licensing – are also of critical importance in the humanities and to publishers.

There are three project phases:
Phase 1: Scoping - months 1-3
Phase 2: Proof of concept development - months 4-7
Phase 3: Evaluation and Implementation - months 8+

Final Project Outcomes:

1. The integration of the ICT-RD data cleaning tool (led by London Metropolitan University, Data Spring project 3.3) will enhance user and data provider experience of DataSHIELD as a whole - data cleaning is fundamental to quality analysis but is often disclosive, and working with ICT-RD we will develop an automated data cleaning function that returns non-disclosive data cleaning information to the analyst and will flag up important potentially disclosive issues with the data provider

2. Implementation of DataSHIELD at the British Library (BL): Using digitised books held by the BL as a test corpus, we will develop a DataSHIELD proof-of-concept for a range of text analyses across datasets presenting divergent challenges to access and interpretation. BL Digital Research has strong ties with humanities' researchers undertaking such work, as the prototype is deployed we will draw on these to establish the range of useful analyses that DataSHIELD will enable.

3. Implementation of DataSHIELD with F1000 Research: Using sensitive academic paper data held by F1000 Research as a test corpus, we will adapt the existing DataSHIELD infrastructure to enable readers and reviewers to flexibly explore the datasets underpinning published articles without having to have physical access to the actual data. This is important when articles are published based on data from studies that have their own immutable governance constraints that mandate formal oversight of data access provision - *e.g.* most of the UK's major cohort studies