Focus group 1 - Software workshop 16 Jan 2017

Problem(s) that the group wants to solve: Different data types in different disciplines

Standards; handling of the data in a reproducible way; types of data; storage of data

Data file types: text files, matlab files, images, Excel sheets

Large scale data brings about problems with how to handle it

Diversity of data: genomics; flow cytometry, whole systems thinking in engineering; consumer behaviour

Need the development of code to allow researchers in the same field to analyse data in the same way.

Sometimes researchers are reluctant to share data for fear of mistakes being found by peers e.g. misdiagnosis in the field of medicine. Implications of the errors. In the case of patient samples/data, researchers are not allowed to share this data.

How to define the standards? FAIR data - laborious for researchers to implement practically even though researchers would benefit from high quality data that abides by these standards.

Define required fields and make these mandatory and then have optional fields - difficult to standardise across different disciplines

Software developers are often computer scientists not experts in that field of research; more communication is required between the computer scientists and the clinical researchers for example.

Suggestions of solutions: Have centralised repository and allow people to rate the data; Wikipedia style of crowdsourcing for a solution; Researcher needs an incentive to generate high quality data analysis with comments; Educating from the undergrad level about data reuse; Journals to help by cracking down on the datasets rather than just accepting the results at face value to ensure that the datasets meet the specifications/requirements

Facilitator: Stephen Eglen
Quick wins:

Example of working standard: Flow repository where raw data files of flow cytometry is uploaded and it gets a score based on quality of data (example of how this might be implemented in wider repositories: https://eudat.eu/events/webinar/fair-data-in-trustworthy-data-repositories-webinar); Information checklist in the form of an Excel spreadsheet format; CDF/HDF5 stores a diverse range of data types

Summary: