**Data Management Plan**

**1. What data will you acquire during the project:** Please describe the type of data you will generate (for example ‘flow cytometry data’) as well as file formats and data volume.

This project will generate DNA sequencing data for the ITS nuclear marker of seven species of Rosaceae, Angiosperms. The main data type and their formats are:

- DNA sequences (fastq)

- Metadata associated with samples (txt)

- Processing scripts (Shell, python)

We will require 1 Tb of working storage which has been costed for.

**2. How will you store the data:** Please describe how you will store and organize your data, what metadata will you capture in what form. Explain how you will document the data during the duration of the project

Active research data will be stored in UoE DataStore. It will be accessible by collaborators and fully back-up regularly.

Raw data will be deposited in GenBank for long term preservation and backed up in UoE DataVault

Metadata will be captured following DarwinCore metadata standards (https://dwc.tdwg.org) for the samples sequenced

We will use an electronic laboratory notebook, such as Benchling to document the experimentation work. It will be accessible by collaborators and fully back-up regularly.

**3. How will you share the data:** Please describe the strategies for data sharing, licensing and access information.

Both raw and pre-processed data will be shared in publicly available domain specific repositories such as Gene Expression Omnibus (GEO) by NCBI. For long term preservation and to obtain a DOI all data will be made available at the UoE DataVault service.

The data will be released at the time of publication. If not published it will be released after three years. All data will be made available under CC-BY licensing. All code will be available under MIT licensing