Publishing data at the UoE

Mario Antonioletti, EPCC
mario@epcc.ed.ac.uk

Data storage at the UoE

- Funding bodies, journals, etc are encouraging/demanding:
  - publishing of research data
  - production of data management plans
    - https://www.dcc.ac.uk/dmponline
  - making your data FAIR, etc
- Offerings described here are mostly for UoE staff/students
  - When you leave the institution access stops within a period of time
  - In a few instances published data may remain
- Other institutions will have similar offerings
- Will talk about another way to publish your data

UoE centrally provided spaces (I)

- OneDrive for Business as part of the O365 UoE offering
  - Students get 0.1 TB
  - Staff get 1TB
- SharePoint
  - Share documents, discussion boards, …
- MediaHopper
  - Curriculum based content
- Microsoft Stream
  - Sharing Teams recording
  - Non curriculum video content
- UoE GitLab (also SubVersion) (https://git.ecdf.ed.ac.uk)
  - Really intended for software development
  - Can be used for developing/versioning small amounts of data (MBs)
    - Never ever put sensitive data in a git repo
  - No quotas at the moment

UoE centrally provided spaces (II)

- DataStore
  - 500 GB for staff and PG research students, need to be on EdLan
  - Secure – important if your data is sensitive
  - Can pool quotas to create group spaces
  - Can allocate space to DataSync (later)
  - Can buy more space (start at 1TB)
- DataSync
  - Share data via https://datasync.ed.ac.uk
  - 20 GB more if you contribute data from your DataStore
  - Available to Staff and Research PGs

Data Share (https://datashare.ed.ac.uk/)

- DataShare:
  - Uses Dspace (open source framework for digital repositories)
  - Targets open data
  - Long term storage (for ever, survives if you leave UoE)
  - Up to 100 GB/dataset (> 20 GB require special measures for upload)
    - Up to 200 files in a deposit
  - Can mint DOIs for your data

DataVault (https://datavault.ed.ac.uk)

- DataVault (not free for > 100GB (total))
- Meant for larger and/or confidential datasets
- May need to be on the UoE VPN to access
- Support for deposits of up to 10 TB/project
- Data owner/PI must be a member of UoE
- Can embargo data (only made available after a period of time)

EIDF (Edinburgh International Data Facility)

- Can publish large data sets
  - 100s TBs – free up to a given threshold
  - Need to apply for a project to publish data that needs to be approved
    - https://portal.eidf.ac.uk
  - Data can be ingested via a number of mechanisms
    - The data is made available as read-only S3 buckets
  - Data S3 links are published in a customised CKAN instance
    - https://catalogue.eidf.ac.uk
    - Allows metadata to be associated with each dataset
    - A dataset can have one or more downloadable resources with S3 links

- Do not have to be a member of UoE to use

- Data remains if you leave the UoE

- Other services offered: Data Science Cloud, GPUs, Cerebras 2s, High memory (18 TB) server, …

Other non-institutional repositories

- Figshare – 20 GB max for file (free) , mint DOIs
  - Figshare+ 20GB to 10 TB Datasets (not free)
- Zenodo – 50GB, mint DOIs
- Discipline related repositories
  - https://www.nature.com/sdata/policies/repositories

Conclusions

- Hopefully, you already knew about most of these offerings
- Have covered ways to publish/share your data I am aware of
  - Most are free to a given threshold
  - All bar one, require UoE affiliation
  - Two will persist if you leave the UoE
- Can rely on other non-institutional repositories