DiRAC-3

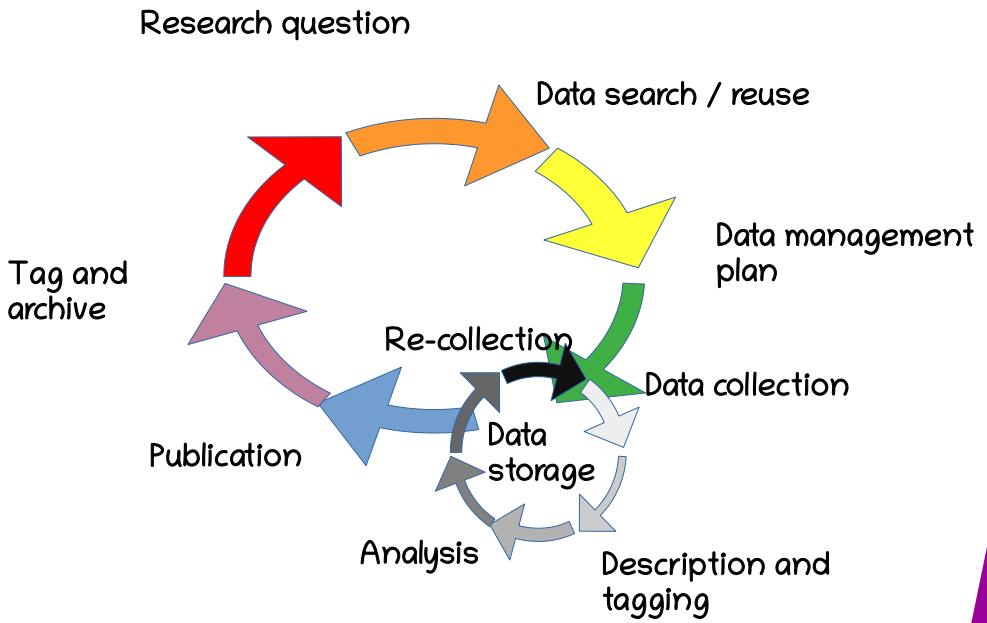
The new Data Curation Service

DiRAC Day
10th September 2020
Alastair Basden

Durham University
Institute for Computational Cosmology

What is data curation

- Organisation of data
  - Metadata labelling
- Integration of data
- Preservation of data
- Accessibility of data
  - Searching for data
- Presentation of data



Metadata

- Describes data
  - Manual and automatic tagging
  - Machine and human readable
- Summarises information in data
- Used to understand data
- Used to search data
- As important as the data itself
- Examples might include:
  - What the data represents
  - How the data was created (reproducibility)
  - Key search terms
  - etc

FAIR principals

- The DCS will ensure that data are:
  - Findable
    - Aided by rich searchable metadata with persistent unique ID
  - Accessible
    - Standard open free universal communications protocol
  - Interoperable
    - Allowing data to be integrated with other data
    - Readable and translatable by multiple applications and workflows
  - Reusable
    - Reuse should be promoted by making data easily accessible
    - Well described, accurate and relevant attributes
    - Meeting domain-relevant community standards

Current DiRAC data strategies

- Project PIs are responsible for their data
  - If a project ends, they are expected to remove it from DiRAC systems
  - But STFC are unlikely to provide large in-house storage
    - No long-term funding
- Sharing data is the responsibility of PIs
  - Usually through group access permissions on the DiRAC file systems
  - Not easily searchable
  - Not accessible for non-DiRAC users
  - Structured as created by the project researchers

New DiRAC Data curation

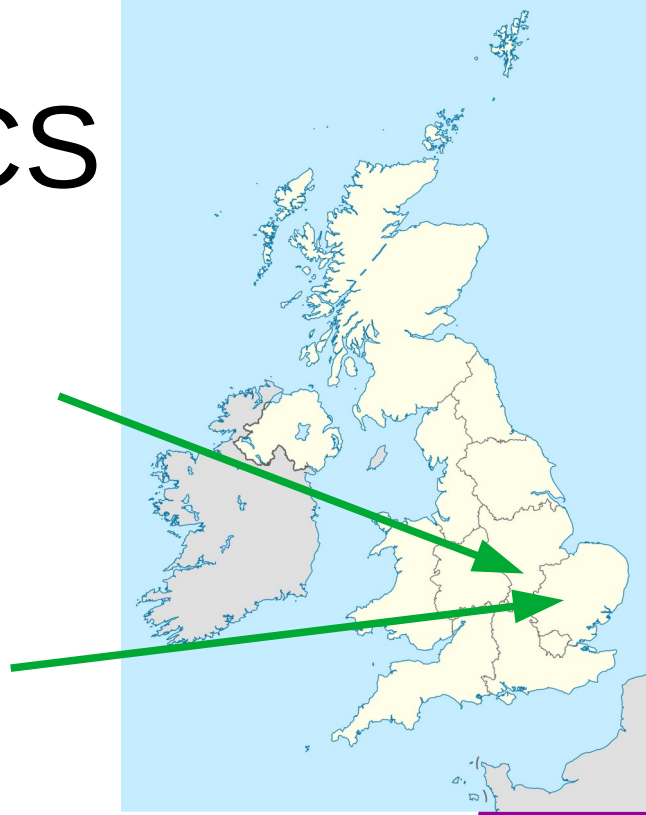
- A long-term repository for data
  - Up to 15 years identified by PIs as necessary
  - Accessible, searchable, navigable, comprehensive
  - FAIR
- Stage 1:
  - Site-specific
- Stage 2:
  - DiRAC-wide

Memory Intensive DCS

- The DiRAC Memory Intensive service at Durham
  - COSMA
  - High memory per core for cosmological simulations
- Current data curation service
  - Tape archive (zero-watt long-term storage)
    - Using a Spectra T950 library and Atempo software
  - Virgo database
    - Providing readily searchable data and metadata (SQL), web interface
- New data curation service
  - Expansion of existing facility
  - Initially:
    - ~10PB bulk storage, 1PB fast storage, 20PB tape storage: User interface to tape archives, website for metadata tagging
  - ~2PB database storage: Improved access, including JupyterLab and possibly SciServer
  - Mid-term low energy online object storage

Data Intensive DCS

- At Cambridge and Leicester



Extreme Scaling DCS

- At Edinburgh

Stage 2: DiRAC-wide DCS

- Linking individual site deployments
  - A joined-up DCS
  - Hardware and software agnostic
- Add data and (eventually) metadata searchable across DiRAC sites
  - Facilitating access and re-use of simulation outputs

DiRAC-wide DCS

- Ceph object store
  - Cross-site replication or erasure coding
  - Federated authentication (e.g. UKAMF)
  - Modest hardware for a trial period
- Transparent way to move data between DiRAC sites
  - Save a file, open it at another site
- Resilience
  - But not against accidental deletion/ransomware
- Does not address metadata or FAIR access initially
  - Will be added after a trial period

Questions?

- Researchers:
  - Please consider what you want to be able to do with your data in the future – there is still time to feed this into the DCS