

UCL Institutional Infrastructure for FAIR Data

James A J Wilson

March 2020

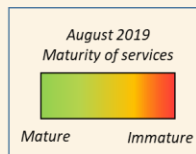


- UCL's Research Data Policy explicitly states that “This policy is intended to ensure that research data created as part of the research process are compliant with the FAIR principles”
- Data should be “as open as possible, as closed as necessary”
- “It is the policy of UCL that following primary use (e.g. publication) or when research data is archived for long term preservation, these data will be made available in the most open manner appropriate. Unless covered by third party contractual agreements, legislative obligations or provisions regarding ownership, it is advised that UCL research data be provided using a Creative Commons CC0 waiver”

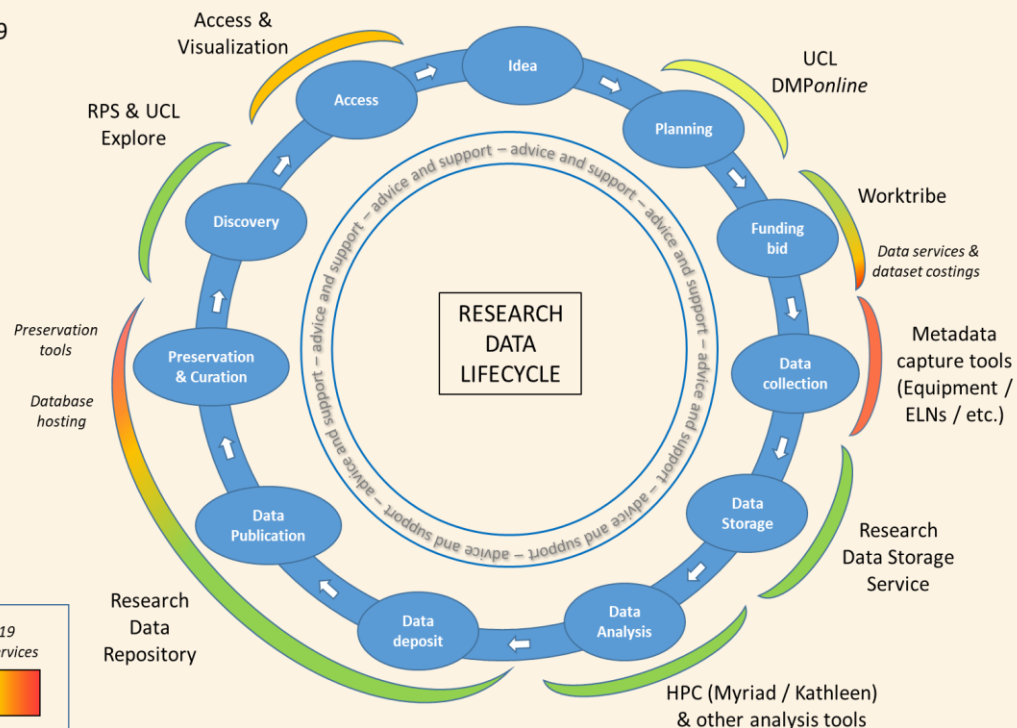
August 2019

RDSS

RDR



- UCL is developing a suite of services to support Open Science and good research data management across the research life-cycle
- Research Data Services (RDS) offers:
 - Research Data Storage Service (for use during projects)
 - Research Data Repository (for long-term preservation and publication)
- Research Data Management team (in Library) offers:
 - Data management planning
 - Advice and guidance
 - Metadata review for the Repository

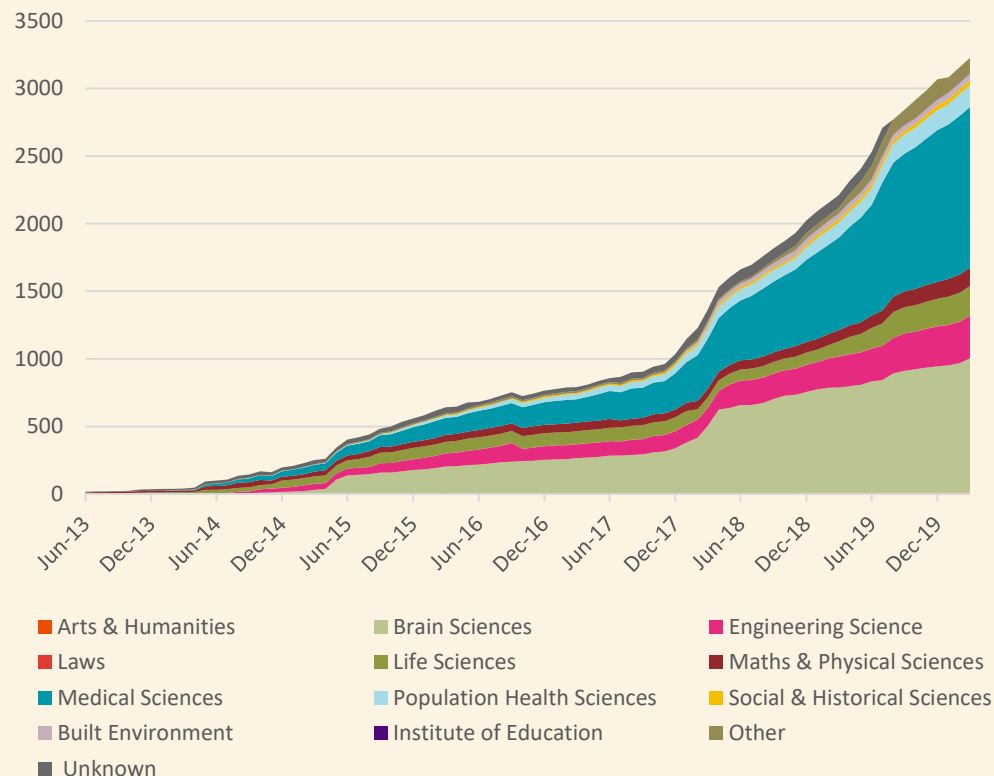


RDSS - Research Data Storage Service

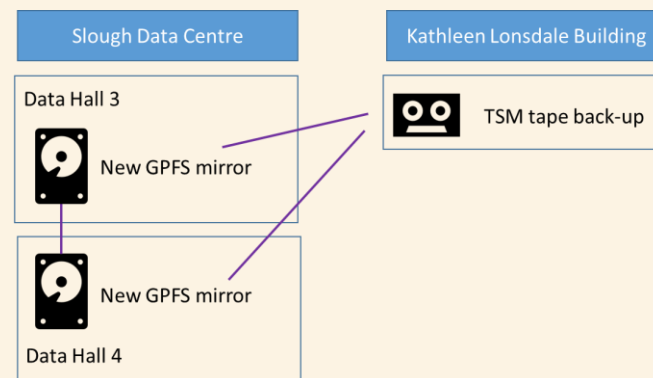


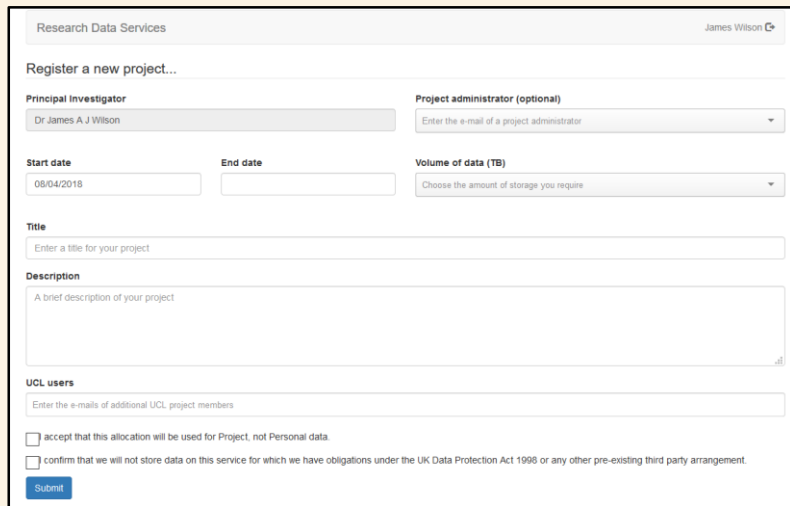
- Designed to support effective research data management during active phase of a research project
- Project-based structure
 - Shared storage between project members
 - Caters for both funded and 'unfunded' projects
- Medical Imaging and Genomics data are particularly large-scale users
- Largest project at present is 230 TB
- Largest single file > 3TB

Service usage (TB) by UCL faculty



- GPFS-based service
- All data mirrored between two separate data halls at Slough
- Co-located with Myriad HPC/HTC facilities
- Nightly back-up to tape
- Storage can be mounted as a local drive
- Simple metadata assists future data curation
- Free project allowance to 5TB for up to 5 years
 - Additional capacity costs £50 per TB per year



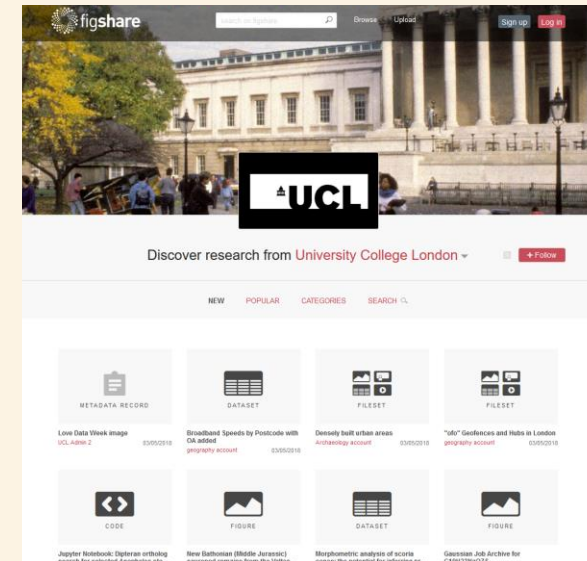
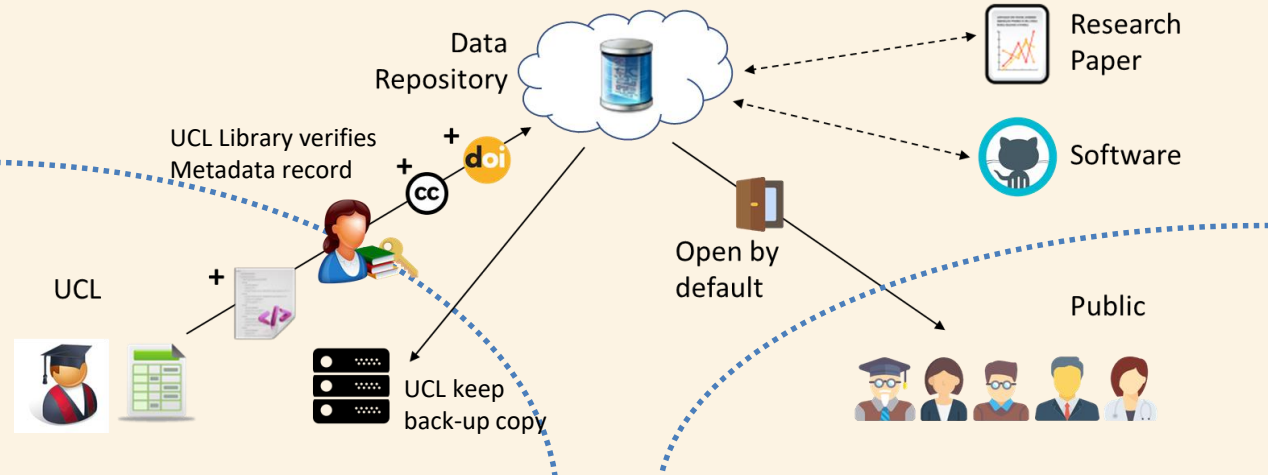


The screenshot shows a web form titled 'Research Data Services' with a user profile 'James Wilson'. The main heading is 'Register a new project...'. The form is divided into several sections: 'Principal Investigator' with a text field containing 'Dr James A J Wilson'; 'Project administrator (optional)' with a dropdown menu; 'Start date' with a text field containing '08/04/2018'; 'End date' with an empty text field; 'Volume of data (TB)' with a dropdown menu showing 'Choose the amount of storage you require'; 'Title' with a text field containing 'Enter a title for your project'; 'Description' with a large text area containing 'A brief description of your project'; and 'UCL users' with a text field containing 'Enter the e-mails of additional UCL project members'. At the bottom, there are two checkboxes: 'accept that this allocation will be used for Project, not Personal data' and 'confirm that we will not store data on this service for which we have obligations under the UK Data Protection Act 1998 or any other pre-existing third party arrangement'. A blue 'Submit' button is located at the bottom left.

- Users sign up via registration form: <https://storageadmin.rd.ucl.ac.uk/>
- Project-based system, with a shared storage allocation for named members of a project
- Projects can consist of 1 or more UCL staff, and must include a 'Principle Investigator'. Can also include nominated administrators
- 'Self-service' interface
- New projects should cost data storage into funding grants
- *Not* the same as Group FileStore ('S' drive) or other UCL Storage Services!

RDR - Research Data Repository

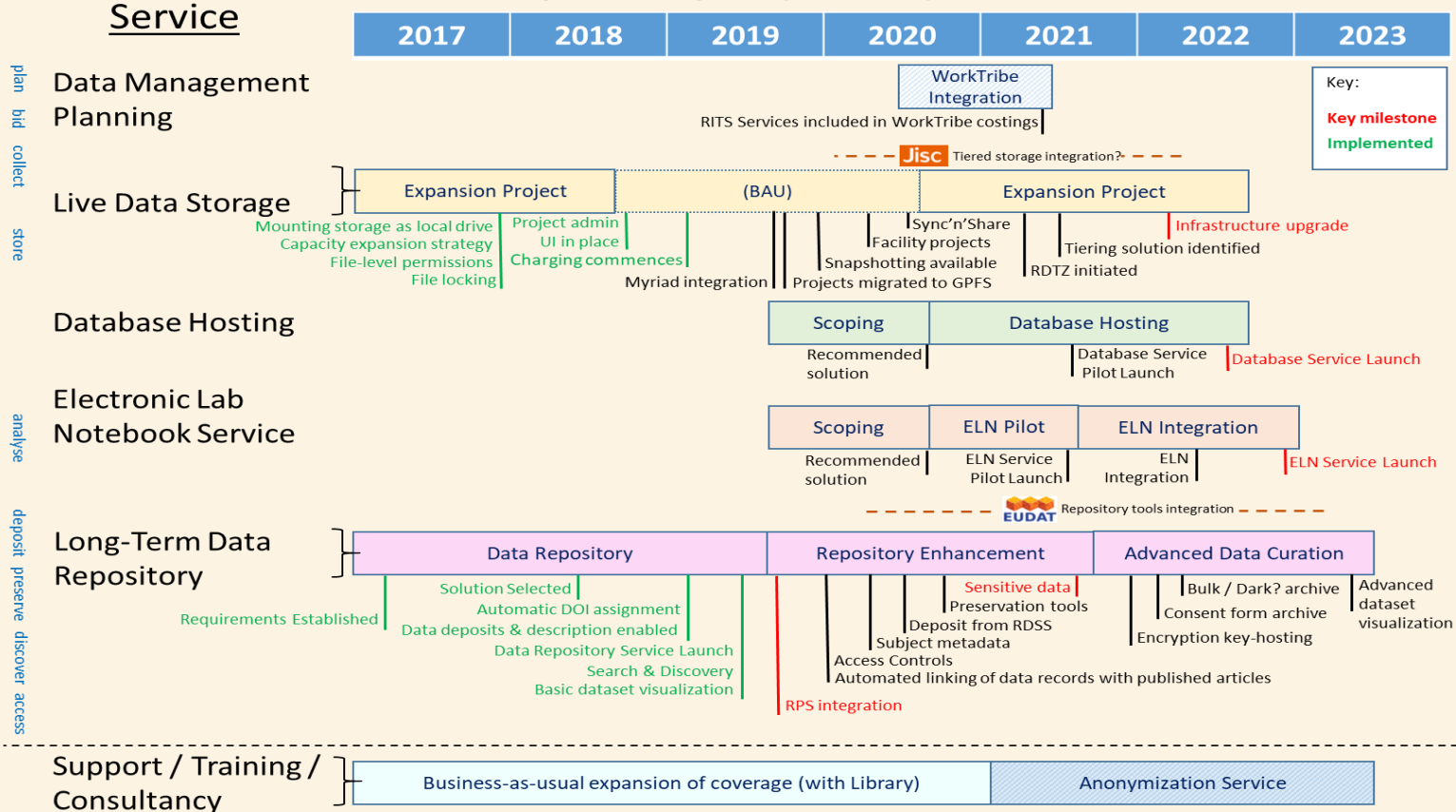
- 'Figshare' interfaces for deposit, discovery, visualization & access
- Long-term storage and curation for significant data
- Data publication, with licence and DOI (Digital Object Identifier) assignment
- Pricing: free initially (within reason)
- Launched June 2019



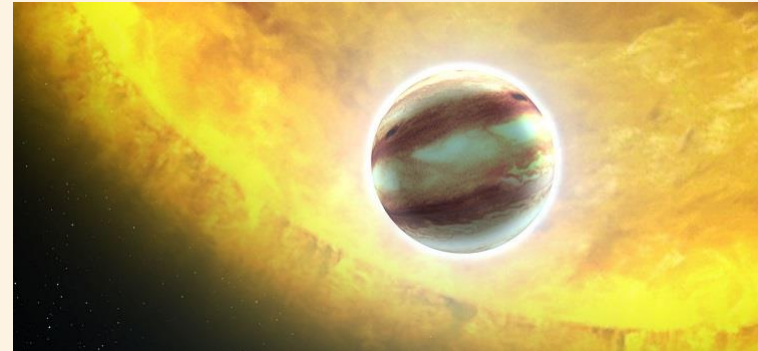
- 2-year project (until end of 2021)
 - a) Improved sensitive data handling
 - b) Long-term preservation tools
 - c) Integration between RDSS and RDR
 - d) Improved linking between research outputs (including integration between RDR and Symplectic elements - UCL's Research Outputs Management System)

- Almost 1,000 projects now using RDSS storage
- Repository sub-sections & projects
 - e.g. Ultrasound Metrology Repository (which sits under the Department of Medical Physics & Biomedical Engineering)
- Metadata mapping for Environmental Digital Solutions for People and the Planet (PLANET)
- Co-ordination and technical leadership for The Republican Antiquarians Research Database
- Federating UCL's Research Data Repository with EUDAT B2FIND
- Active Directory should be able to begin enabling RDSS access for UCL collaborators *without* formal honorary membership during coming months

RDS Proposed Projects (and BAU) Timeline

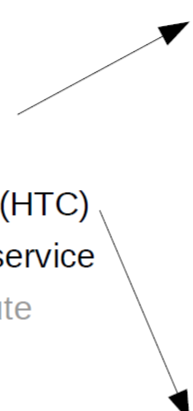


- Research Computing –
High Performance Computing facilities for intensive data crunching
- Research Software Development –
Can advise on and help create good re-usable code for research projects
- Data Safe Haven –
UCL's facility for analysing sensitive data
Currently undergoing a refresh to improve compute effectiveness
- Research Applications –
The team that looks after the Research Publications Service (RPS), UCL Discovery, IRIS, and the Research Equipment Catalogue



RITS's Legion Cluster is providing UCL astrophysicists with the computing firepower they need to help pinpoint exoplanets capable of supporting alien lifeforms

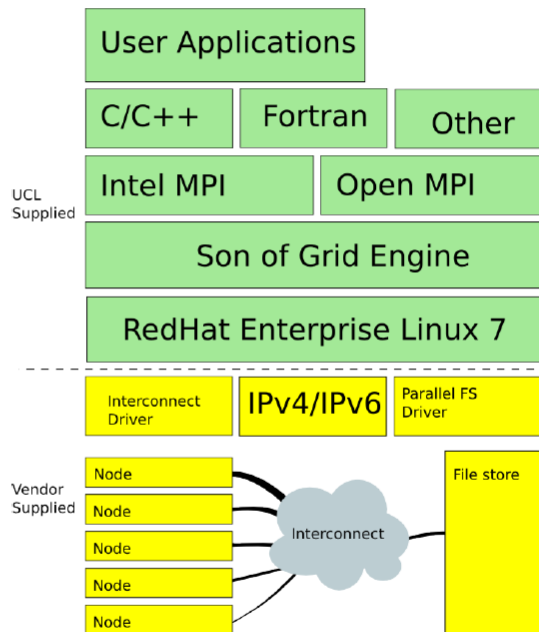
What we do:

- UCL only services:
 - **Grace, Kathleen** → High Performance Computing (HPC)
 - **Myriad** → High Throughput Computing (HTC)
 - **Aristotle** → Interactive teaching Linux service
 - **DSH** → secure data storage and compute (not currently under RC control)
 - National services:
 - **Thomas** (Tier 2 MMM hub)
 - **Michael** (Faraday Institution)
 - Parallel
 - Single job spans multiple nodes
 - Tightly coupled parallelisation usually in MPI
 - Sensitive to network performance
 - Currently primarily chemistry, physics, engineering
 - High throughput
 - Lots (tens of thousands) of independent jobs on different data
 - High I/O
 - Currently, primarily biosciences, physics, computer science
 - In the future, digital humanities
- 

COMMON software stack across RC controlled services

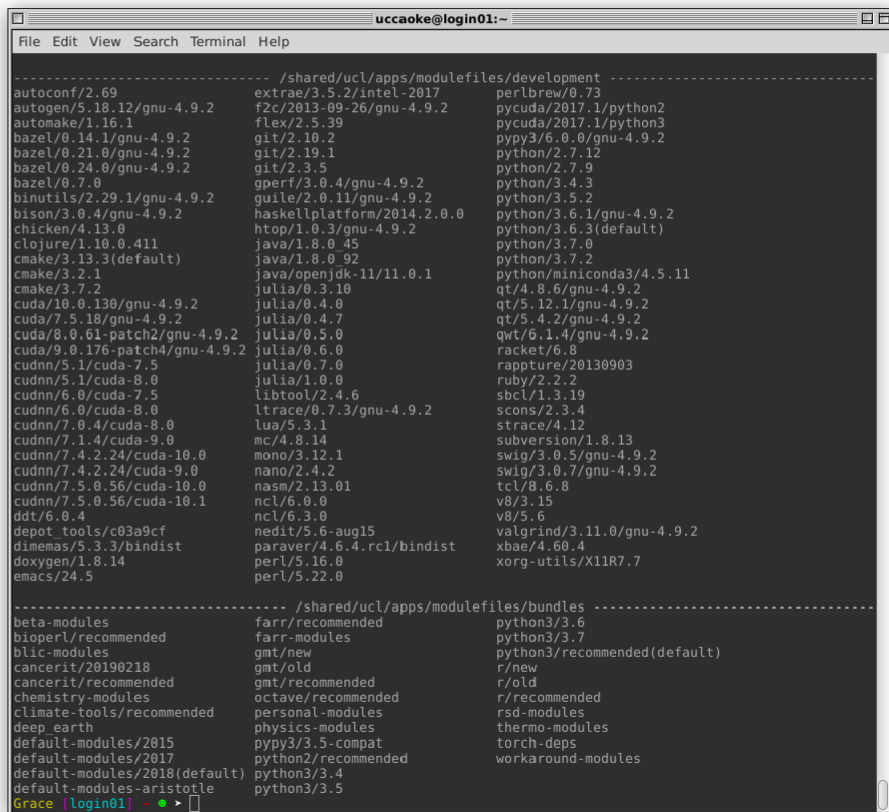
Common software stack

- Deployed across all our resources (inc Thomas + Michael)
 - ~750 user applications + development tools, presented through environment modules
 - Scripts + data from one machine can be run “seamlessly” on another
 - Same interface presented to users
 - **AUTOMATED**



Common software stack

- This is **not** a stack “just for traditional HPC users” (Fortran/C/C++)
- Supports Python (Cpython, Anaconda, PyPy), R, Julia, Perl (+ Bioperl), Java, Clojure, Common Lisp, Scheme, Mono (.Net), Lua, Go, Racket, Ruby, JavaScript, Matlab...
- ML tools like Tensorflow (GPU, MKL variants), Caffe, OpenCV...
- Allow departmental sysadmins access to install specialist applications centrally!



```
uiccaoke@login01:~  
File Edit View Search Terminal Help  
----- /shared/ucl/apps/modulefiles/development -----  
autoconf/2.69      extrae/3.5.2/intel-2017      perlbrew/0.73  
autogen/5.18.12/gnu-4.9.2  f2c/2013-09-26/gnu-4.9.2    pycuda/2017.1/python2  
automake/1.16.1      flex/2.5.39                 pycuda/2017.1/python3  
bazel/0.14.1/gnu-4.9.2   git/2.10.2                 pypy3/6.0.0/gnu-4.9.2  
bazel/0.21.0/gnu-4.9.2   git/2.19.1                 python/2.7.12  
bazel/0.24.0/gnu-4.9.2   git/2.3.5                  python/2.7.9  
bazel/0.7.0          gperf/3.0.4/gnu-4.9.2      python/3.4.3  
binutils/2.29.1/gnu-4.9.2  guile/2.0.11/gnu-4.9.2     python/3.5.2  
bison/3.0.4/gnu-4.9.2    haskellplatform/2014.2.0.0  python/3.6.1/gnu-4.9.2  
chicken/4.13.0         htop/1.0.3/gnu-4.9.2       python/3.6.3(default)  
clojure/1.10.0.411      java/1.8.0_45              python/3.7.0  
cmake/3.13.3(default)   java/1.8.0_92              python/3.7.2  
cmake/3.2.1            java/openjdk-11/11.0.1     python/miniconda3/4.5.11  
cmake/3.7.2            julia/0.3.10               qt/4.8.6/gnu-4.9.2  
cuda/10.0.130/gnu-4.9.2  julia/0.4.0                qt/5.12.1/gnu-4.9.2  
cuda/7.5.18/gnu-4.9.2    julia/0.4.7                qt/5.4.2/gnu-4.9.2  
cuda/8.0.61-patch2/gnu-4.9.2  julia/0.5.0                qwt/6.1.4/gnu-4.9.2  
cuda/9.0.176-patch4/gnu-4.9.2  julia/0.6.0                racket/6.8  
cudnn/5.1/cuda-7.5       julia/0.7.0                rappture/20130903  
cudnn/5.1/cuda-8.0       julia/1.0.0                ruby/2.2.2  
cudnn/6.0/cuda-7.5       libtool/2.4.6              sbcl/1.3.19  
cudnn/6.0/cuda-8.0       ltrace/0.7.3/gnu-4.9.2     scons/2.3.4  
cudnn/7.0.4/cuda-8.0     lua/5.3.1                  strace/4.12  
cudnn/7.1.4/cuda-9.0     mc/4.8.14                  subversion/1.8.13  
cudnn/7.4.2.24/cuda-10.0  mono/3.12.1                swig/3.0.5/gnu-4.9.2  
cudnn/7.4.2.24/cuda-9.0   nano/2.4.2                 swig/3.0.7/gnu-4.9.2  
cudnn/7.5.0.56/cuda-10.0  nasm/2.13.01               tcl/8.6.8  
cudnn/7.5.0.56/cuda-10.1  ncl/6.0.0                  v8/3.15  
ddt/6.0.4              ncl/6.3.0                  v8/5.6  
depot-tools/c03a9cf      nedit/5.6-aug15            valgrind/3.11.0/gnu-4.9.2  
dimemas/5.3.3/bindist   paraver/4.6.4.rc1/bindist  xbae/4.60.4  
doxygen/1.8.14          perl/5.16.0                xorg-utils/X11R7.7  
emacs/24.5             perl/5.22.0  
----- /shared/ucl/apps/modulefiles/bundles -----  
beta-modules      farr/recommended          python3/3.6  
bioperl/recommended  farr-modules             python3/3.7  
blic-modules       gmt/new                  python3/recommended(default)  
cancerit/20190218   gmt/old                  r/new  
cancerit/recommended  gmt/recommended          r/old  
chemistry-modules   octave/recommended       r/recommended  
climate-tools/recommended  personal-modules         rsd-modules  
deep-earth          physics-modules           thermo-modules  
default-modules/2015  pypy3/3.5-compatible     torch-deps  
default-modules/2017  python2/recommended       workaround-modules  
default-modules/2018(default)  python3/3.4  
default-modules-aristotle  python3/3.5  
Grace [login01] ~ ● ➤
```

- researchdata-support@ucl.ac.uk
- <https://www.ucl.ac.uk/research-it-services/services/research-data-services>
- Research Data Storage Service:
<https://www.ucl.ac.uk/isd/services/research-it/research-data-storage-service>
- Research Data Repository:
<https://rdr.ucl.ac.uk/>