

Data Week Online 2020

Making data work for everyone

bristol.ac.uk/golding

Data Week Online 2020

The Jean Golding Institute

- A central hub for data science and data-intensive research
- One of 5 University of Bristol research institutes
- Connect multidisciplinary experts across the University and beyond
- Events, training, funding, Ask JGI, The Alan Turing Institute

Our priorities

1. Societal challenges
2. Data visualisation
3. Reproducibility & data governance
4. Fundamental research

Making data work for everyone



bristol.ac.uk/golding

Data Week Online 2020

Date	Event	Speaker
Monday 15 June	Data science and COVID 19 & Data Week Introduction	Kate Robson Brown, JGI Director
Monday 15 June	Intermediate Python	Advanced Computing Research Centre
Tuesday 16 June	Talk: Working at and with The Turing Institute: experiences as a Fellow	Jon Crowcroft, Turing Fellow & University of Cambridge
Tuesday 16 June	Talk: increasing engagement with data	Michael Green, Luna 9
Tuesday 16 June	Introduction to data analysis in Python	Advanced Computing Research Centre
Wednesday 17 June	Do you want to be a data Rockstar?	Luke Stoughton, The Information Lab
Wednesday 17 June	Applied data analysis in Python	Advanced Computing Research Centre
Thursday 18 June	Talk: New data on COVID-19 is undermined by old statistical problems	Gibran Hemani, University of Bristol
Thursday 18 June	Managing sensitive research data: from planning to sharing	Library Research Services
Thursday 18 June	Introduction to deep learning	Advanced Computing Research Centre
Friday 19 June	Deep Learning for Health and Life Sciences	Valerio Maggio, University of Bristol
Friday 19 June	Tour of the Tidyverse	Max Kronborg, Mango Solutions
Friday 19 June	Best practices in software engineering	Advanced Computing Research Centre

Making data work for everyone

bristol.ac.uk/golding

Managing sensitive data: from planning to sharing

Research Data Service

Library Research Support

bristol.ac.uk

Outline

1. Overview of issues
2. Planning a project with sensitive data
3. Data collection and storage
4. Platforms for sharing sensitive data
5. Preparing sensitive data for sharing
6. Q&A

What is sensitive data?

Any type or format of data:

- ▶ Qualitative
- ▶ Quantitative
- ▶ Code
- ▶ Spreadsheets...

Types of sensitivity:

- ▶ Ethical (relating to identifiable people or at-risk species)
- ▶ Commercial (IPR)
- ▶ National security
- ▶ Legal

What is a data management plan?

A data management plan (DMP) should cover data creation, documentation, storage, preservation & sharing plans

Most funders require one as part of application process

Key areas for sensitive data are **storage** and **sharing**

Addressing these in your DMP will help you meet GDPR, DPA, funder and publisher requirements



What's covered in a DMP?

Data collection	What new data is being collected/what existing data is being reused? What formats and volume of data will be collected? How will data quality be maintained?
Documentation	How will data be described and organised? What schemas/formats/standards will be used?
Storage	Where will data be stored during the project? Who will have access to it?
Preservation	What data will be kept after the project ends and in what formats? Where will data be stored after the project ends? For how long?
Sharing	What data will be shared after the project ends and in what formats? When will it be shared? Where will it be made available? What access restrictions will be applied?
Responsibilities and resources	Who is responsible for data collection, quality assurance, curation, and sharing? What resources are required to carry out data management tasks?





DMPOnline

DMPONLINE Home Public DMPs Funder requirements Help

Welcome

DMPOnline helps you to create, review, and share data management plans that meet institutional and funder requirements. It is provided by the Digital Curation Centre (DCC).

Join the growing international community that have adopted DMPonline:

-  17,622 Users
-  203 Organisations
-  23,083 Plans
-  89 Countries

Some funders mandate the use of DMPonline, while others point to it as a useful option. You can [download funder templates](#) without logging in, but the tool provides tailored guidance and example answers from the DCC and many research organisations. Why not sign up for an account and try it out?

Sign in Create account

* Email

* Password

Forgot password?

Remember email

Sign in

- or -

Sign in with institutional credentials (UK only)

DMPOnline has templates for all major UK/EU funding bodies

<https://dmponline.dcc.ac.uk/>

bristol.ac.uk

Legislation: GDPR & Data Protection Act 2018

GDPR & DPA 2018 came into force in May 2018 and will **still apply** after Brexit

Legislation sets out data subjects' (research participants') rights with regards to their personal data, and your responsibilities if you are processing that data



Personal data:

Name	Online identifiers	Cultural/social identity
Address	IP addresses	Economic information
ID numbers	Physical attributes	...

'Special category' data:

Race	Trade union membership	Health
Ethnic origin	Genetic	Sex life
Politics	Biometrics (used for ID)	Sexual orientation
Religion		



Lawful bases for data processing

Anyone collecting personal data requires a lawful basis to do so:

1. Consent
2. Contract
3. Legal obligation
4. Vital interests
5. **Public task**
6. Legitimate interests



The University Charter gives our researchers a legal basis for performing research in the public interest:

“To make provision for Research and to furnish Scientific Advice for public purposes and for these objects to enter into such arrangements with other institutions or with public bodies as may be thought desirable.”

University of Bristol Charter, 3(17)

<http://www.bristol.ac.uk/university/governance/constitutionaldocs/charteractsstatutesordinances/>

Data subject rights

1. The right to be informed of the collection and use of their personal data
2. The right to access their personal data
3. The right to have inaccurate or incomplete information corrected
4. The right to have their personal data erased
5. The right to request that you restrict the ways in which their personal data is processed
6. The right to a copy of their personal data in a portable, machine readable format
7. The right to object to processing of their personal data
8. The right to be informed of any automated individual decision-making or profiling, and the right to challenge such decisions

Some rights are abrogated if you are processing data for statistical or research purposes – rights in red are **not**

Exemptions for research

Processing data for research purposes grants certain permissions and exemptions:

- Special category data may be processed
- Some data subject rights may be set aside

Key principles:

1. Data minimisation – collect no more than you need
2. Anonymise as much as possible as early as possible
3. Do not carry out automated decision-making without ethical approval

Participant rights:

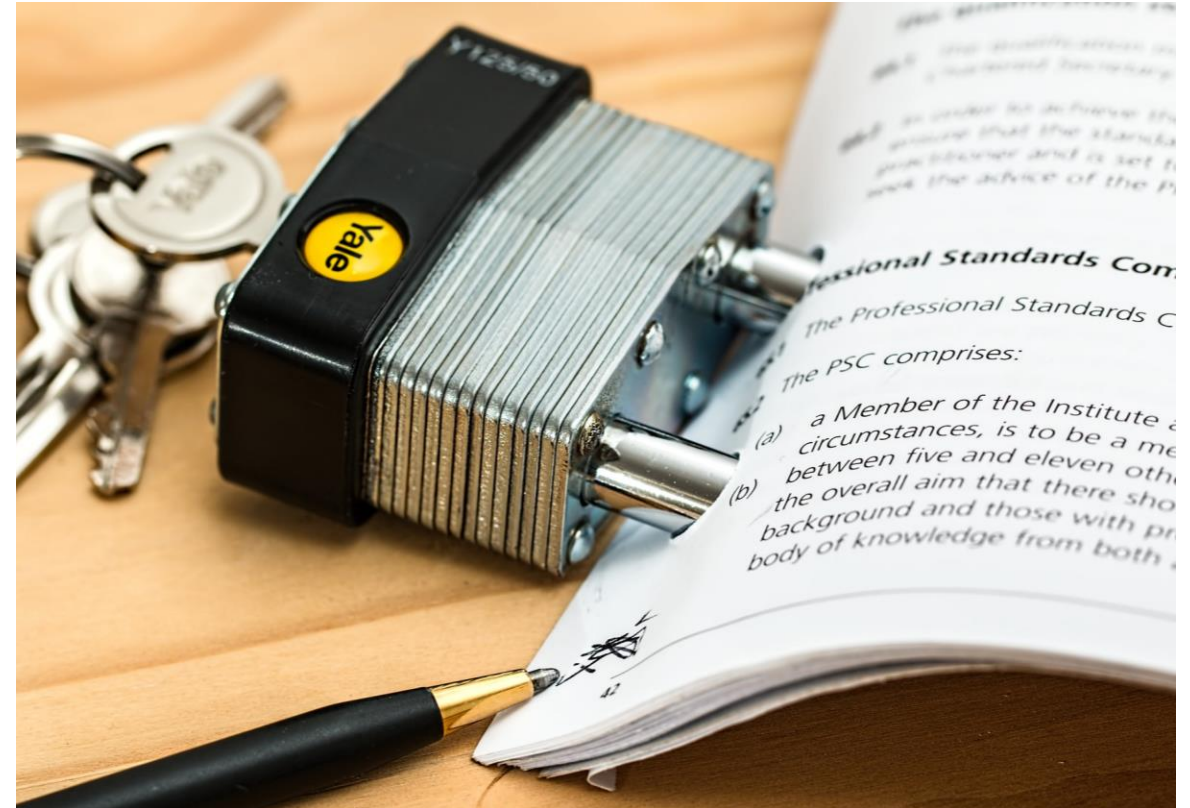
1. Inform participants of the collection and use of their personal data
2. Inform participants of any automated decision-making that happens as a result of data collection and how this can be challenged
3. If participants can be identified in published results, they will retain the right to access their personal data on request

Documenting data processing

You should record in your DMP and in privacy notices:

1. Legal basis for data processing and condition for processing special category data
2. What personal and special category data is being collected
3. Who will have access to it
4. How long it will be kept for
5. Who it will be shared with

See ICO checklist for what to include in privacy notices: <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-be-informed/>



Privacy impact assessment

High risk projects or projects using new technologies should conduct a Data Protection Impact Assessment (DPIA), also known as a Privacy Impact Assessment (PIA)

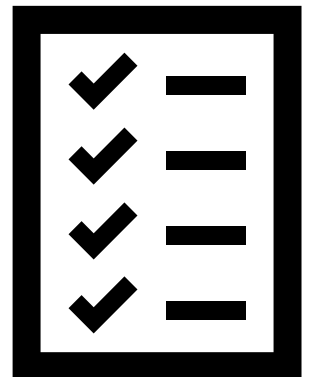
‘High risk’ means projects involving:

- special category data including medical or criminal records
- systematic profiling of participants
- systematically monitoring public places
- large scale medical research (large in terms of number of participants, volume of data, time scales, or geographical area)
- biometric or genetic data
- vulnerable groups (children, the elderly, or those with mental/physical disabilities)

Contact the University’s Information Governance Manager if you think a PIA may be required:

<http://www.bristol.ac.uk/secretary/data-protection/guidance/privacy-impact-assessment/>

bristol.ac.uk



Consent

Getting the consent form wrong can **seriously** impact data sharing and manuscript publication

ETHICAL CONSENT TO PARTICIPATE IN A STUDY \neq CONSENT FOR DATA PROCESSING UNDER GDPR

Key issues:

1. Separate *research data* and *personal information needed for administrative purposes*
2. Be clear on what will happen to both types of data after the study ends, including how long they will be kept for, and who will have access to them
3. Be clear on what will happen to data if a participant withdraws

Why does this matter?

“I understand that my data will be kept confidential and that only the research team will be able to access information about me.”

“I understand that the data may be used by other researchers after the project ends, for work unrelated to this project.”

**What data is being referred to in each statement?
Will the participants understand this?**

bristol.ac.uk



Example consent form clauses

Personal data:

“I understand that health professionals and members of the research team will have access to my personal information, including contact information and other direct identifiers.”

Research data:

“I understand that after the study, the research data (results) may be made available as ‘open data’. This means the data will be publicly available and may be used for purposes not related to this study. However, any personal information that could identify me will be removed or changed before files are shared with other researchers or results are made public.”

“I understand that study results research data (results) may be seen and used by other researchers, for ethically approved research projects, but that any personal information that could identify me will be removed or changed before files are shared with other researchers or results are made public.”

UoB templates available at <http://www.bristol.ac.uk/secretary/data-protection/gdpr/transparency-wording/>

Third party data

Third party data = data collected and owned by someone else, e.g. NHS Digital

Terms of use will be laid out in the data licence or data sharing/data transfer agreement

Contact RED Contracts, IT Services and the Secretary's Office for advice if you are planning to use third party sensitive data

bristol.ac.uk



UK Data Service



Collecting sensitive data

Sensitive data must be secure from the point of collection:

- ▶ Use encrypted devices for data collection in the field and transfer to secure storage as soon as possible
- ▶ Use approved transcription, scanning/digitisation and videoconferencing services:
 - ▶ <http://www.bris.ac.uk/infosec/uobdata/transcription/>
 - ▶ <http://www.bristol.ac.uk/print-services/document-services/>
 - ▶ <https://www.bris.ac.uk/it-services/advice/uobonly/recordsmgt>
 - ▶ <https://www.bristol.ac.uk/telephones/conference-solutions/video-conferencing/>
 - ▶ BlueJeans is recommended for conducting interviews with research participants
 - ▶ Other services can be used to communicate with collaborators
- ▶ Bristol Medical School REDCap: platform for secure online collection of survey data (brms-redcap@bristol.ac.uk)
- ▶ Store personal data for administrative purposes separately to research data

Collecting sensitive data continued

Minimize collection of disclosive data/identifiable information:

What are the
minimum
variables
needed?

What level of
precision is
needed?

Can you mask
identifiers at the
point of
collection?

Storing sensitive data

Recommended storage locations:

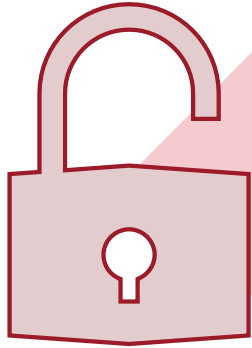
- ✓ Research Data Storage Facility
<http://www.bristol.ac.uk/acrc/research-data-storage-facility/>
- ✓ Any networked & backed up UoB drive
- ✓ NHS servers
- ✓ UoB Microsoft OneDrive

Not recommended:

- ✗ Other cloud service
- ✗ Personal storage (local PC, USB drive, CD...)

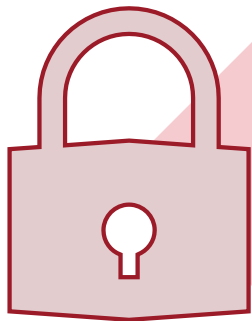


Data sharing options



Open data:

- Anonymised, low risk data
- Files available for download without restriction
- Licences may be applied to limit re-use (e.g. CC-BY-NC-SA) but are hard to enforce



Restricted/controlled/safeguarded data:

- Anonymised but some risk remains, or unable to be sufficiently anonymised
- Metadata/descriptive record freely available online
- Data access is via an independent, transparent, appealable process
- Terms of re-use are contractually enforced in a data access agreement

Data sharing platforms

UoB research data repository

- ▶ data.bris: <https://data.bris.ac.uk/data/>
- ▶ Offers open, restricted and controlled access
- ▶ Supported by Research Data Service

Commercial repositories

- ▶ figshare, Dryad, Zenodo, OSF, Mendeley Data...
- ▶ Usually only offer open access
- ▶ Some are integrated with manuscript submission process, depending on publisher
- ▶ Usually no restrictions on data format but may restrict data size



Data sharing platforms continued

Disciplinary repositories

- ▶ UK Data Service, GenBank, Protein Data Bank, Crystallography Open Database...
- ▶ Some offer controlled access (UKDS)
- ▶ May require certain data formats/structures/documentation
- ▶ Usually have good user support
- ▶ See <https://www.re3data.org/> for a searchable list

Warning



Preparing data at the end of a project can be very time consuming!

- ▶ Minimize what you collect
- ▶ Anonymise as early as possible as far as possible
- ▶ Include data curation costs in your budget

Organising your data

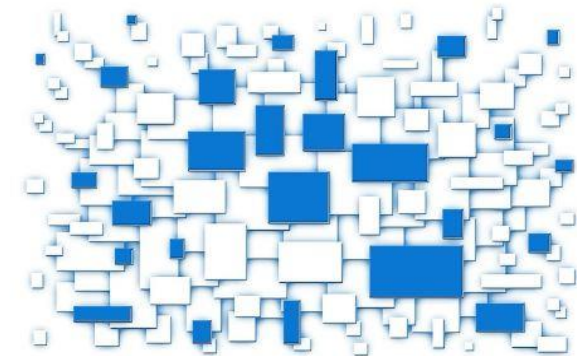
How will end users navigate and use your dataset?

- ▶ Consistent, logical folder structures and file names
- ▶ One minute guide to file organisation: <http://bit.ly/2Nr3Tfj>
- ▶ Think about which file formats are suitable for long-term use
- ▶ UK Data Service list of recommended formats:
<https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats>

bristol.ac.uk

One minute guide to file organisation

Version 2.2 March 2019



University of Bristol

Research Data Service



Documentation

1. Readme file

- ▶ File inventory
- ▶ Software needed to open files
- ▶ Other dependencies
- ▶ Column headers, row labels, and units of measurement for tabular data

2. Data dictionary

3. Descriptive keywords

4. Data and metadata standards

- ▶ SNOMED-CT, OME-XML, DDI...



See fairsharing.org for a searchable database of standards

bristol.ac.uk

Disclosure risk



Disclosure = identification of individual research participants



Can happen via **direct** and **indirect** identifiers

Direct identifiers: names, postcodes, national insurance number, NHS number, images

Indirect identifiers: occupation, household size, age, gender, free-text questionnaire responses



Combining indirect identifiers may be enough to identify participants

Specific risks for different data types



Quantitative data

May contain direct and indirect identifiers

Small sample size



Qualitative data

Images of participants

Metadata/documentation may also contain identifiers

Third parties mentioned in interviews



Other risks

Data linkage (between studies and to external datasets)

Additional indirect identifiers available from study protocol, etc. (e.g. geographical area)

Managing disclosure risk: before you start

- ▶ Keep a **log** of any changes made to the dataset
- ▶ Assess the dataset before applying any techniques – look for outliers and combinations of variables that could be disclosive
- ▶ Think about the likely environment the dataset will be used in – what other knowledge are users likely to have (e.g. geographical area)?

Formal anonymisation

- ▶ Replace names with pseudonyms
- ▶ Text redaction
- ▶ Blurring faces and other identifiable details in videos and images
- ▶ Break linkages between related datasets
- ▶ Software can help with these tasks, e.g. UKDS text anonymisation helper tool

<http://data-archive.ac.uk/curate/standards-tools/tools>

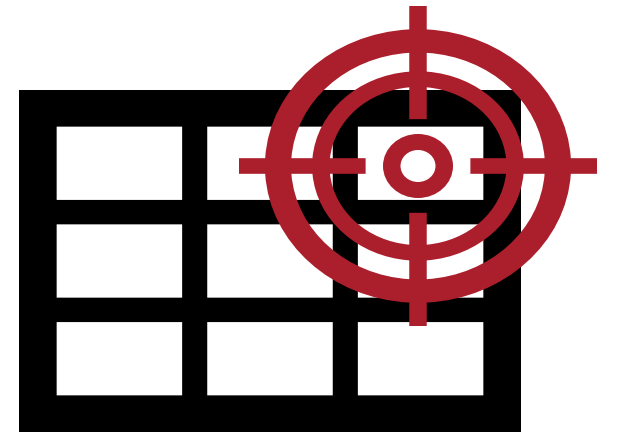
Examples:

Interview transcript: <https://www.ukdataservice.ac.uk/media/604650/anonymisationtranscriptexample.gif>

Quantitative data: <https://www.ukdataservice.ac.uk/media/604739/anonexample2.gif>



Statistical anonymisation



k-anonymisation

- ▶ At least k records have the same combination of indirect identifiers
- ▶ Threshold of 3 or 5 is common
- ▶ Apply techniques below to 6-7 variables at a time and see how the k-value changes – but don't use more than 3 techniques at a time
- ▶ Software packages can automate this to an extent

Techniques

- ▶ Aggregate variables
- ▶ Top-bottom coding
- ▶ Rounding
- ▶ Cell or value suppression

Software packages:

- ▶ R package sdcMicro: <https://cran.r-project.org/web/packages/sdcMicro/index.html>
- ▶ μ -ARGUS: <http://neon.vb.cbs.nl/casc/..%5Ccasc%5Cmu.htm>
- ▶ ARX: <https://arx.deidentifier.org/downloads/>

Functional anonymisation

Control the data environment:

- ▶ Apply an end user licence/data access agreement
- ▶ Offer access to data via a data haven or data enclave

data.bris

- ▶ Standard licence is the [Non-Commercial Government Licence for public sector information](#)
- ▶ For sample data access agreements, see <http://www.bristol.ac.uk/staff/researchers/data/publishing-research-data/publishing-data-under-access-restrictions/>



Controlled release may be required even for sampled, anonymised datasets if they contain large numbers of demographic attributes: in 5% Public Use Microdata Sample of US census data, **15 attributes would render 99.98% of MA residents unique** ([Rocher et al. 2019](#), doi: 10.1038/s41467-019-10933-3)

Research Data Service support

1. Advice on DMPs, consent forms, secure data collection
2. Encrypted voice recorder loan
3. Disclosure risk assessment
 - ▶ Checks for direct identifiers
 - ▶ Basic statistical anonymisation
4. Controlled data release
 - ▶ Data Access Committee
 - ▶ Data Access Agreements

Contact us at data-bris@bristol.ac.uk

Summary



Plan ahead – write a DMP!



Get your consent forms right



Include resources for data curation in your project plan and funding applications, e.g. staff time, specialist software



Think about likely uses for your data and what levels of security and anonymisation make sense



Seek advice if you're not sure!

Guidance and tools

UoB guidance: <http://www.bristol.ac.uk/staff/researchers/data/dealing-with-sensitive-data/>
<http://www.bristol.ac.uk/secretary/data-protection/gdpr/gdpr-and-research/>

UK Data Service: <https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation.aspx>

UK Anonymisation Network: <https://ukanon.net/ukan-resources/>

Irish Social Science Data Archive: <http://www.ucd.ie/issda/aboutus/anonymisation/>

SDC for Microdata practical guide: <https://sdcpractice.readthedocs.io/en/latest/>

ICPSR Recommended Informed Consent Language for Data Sharing:
<https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/conf-language.html>

Help available



Library Research Support

Research Data Service

Website: <http://www.bristol.ac.uk/staff/researchers/data/>

Email: data-bris@bristol.ac.uk

RED

Contracts: <http://www.bristol.ac.uk/red/contracts/>

Research Governance: <http://www.bristol.ac.uk/red/research-governance/>

Commercialisation: <http://www.bristol.ac.uk/business/research-commercialisation/>

IT Services

Faculty support teams: <https://www.bristol.ac.uk/it-services/locations/fits>

ACRC: <https://www.bristol.ac.uk/acrc/research-data-storage-facility/>

Jean Golding Institute

Website: <https://www.bristol.ac.uk/golding/>

Email: ask-jgi@bristol.ac.uk

Faculty Research Ethics Officers

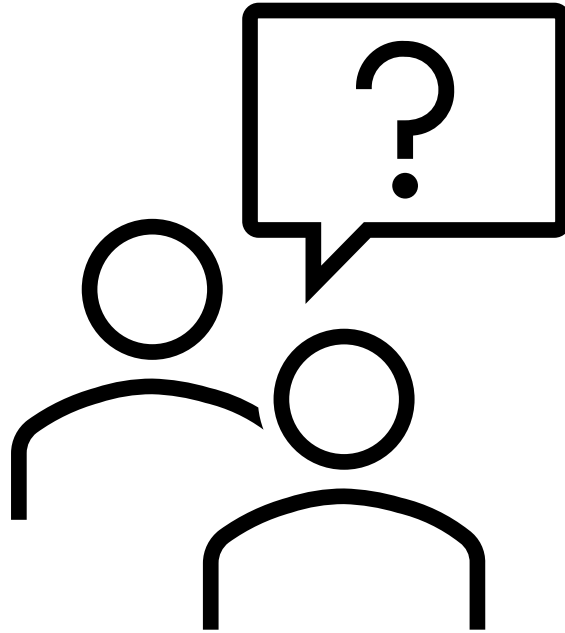
<http://www.bristol.ac.uk/red/research-governance/ethics/uni-ethics/>

Secretary's Office

<http://www.bristol.ac.uk/secretary/contact/>

bristol.ac.uk

Q&A



Data Week Online 2020

Share your participation

#Dataweekonline2020



@JGIBristol

jgi-admin@bristol.ac.uk

bristol.ac.uk/golding

Keep in touch

bristol.ac.uk/golding