Building Research Data Management Capacity in LMICs: Development and implementation of the RESPIRE Data Champion's Handbook

We take you to the heart of research data management and give you all the essentials you need: managing, protecting, and sharing data, concisely, practically, and without the jargon, so your datasets survive, your participants' contributions count, and your research reaches further than you planned.

Version 0.9
5 Sep 2026

Led by: THE UNIVERSITY of EDINBURGH | UNIVERSITI MALAYA

FUNDED BY NIHR | National Institute for Health and Care Research

UK International Development
Partnership | Progress | Prosperity

Stay Updated

This is Version 0.9 (05 Sep 2026) of the handbook.

Good data practice is constantly evolving. The RESPIRE Data Champions Initiative regularly updates this handbook with new case studies, updated legal guidance, and refined templates.

Ensure you are using the most current guidance.

Scan the QR code below or visit https://edin.ac/4ybicu3 to download the latest version of **Building Research Data Management Capacity in LMICs: Development and implementation of the RESPIRE Data Champions Handbook**.

Building Research Data Management Capacity in LMICs: Development and implementation of the RESPIRE Data Champion's Handbook

We take you to the heart of research data management and give you all the essentials you need: managing, protecting, and sharing data, concisely, practically, and without the jargon, so your datasets survive, your participants' contributions count, and your research reaches further than you planned.

Version: 0.9/ 05 Sep 2026

Revision History

Draft Versions	Date	Changes / Updates
0.1	Oct 2024	Initial draft outline
0.3	Dec 2025	Working Group contributions integrated, Restructuring of the chapters
0.4-0.6	Jan-Aug 2026	Addition of the chapters (Ethical use of AI, IP Policy, & tools/ software), incorporated illustrations and screenshots
0.7-0.9	Sep 2026	APA reference audit, streamlined flow, adding

Funding: The NIHR Global Health Research Unit on Respiratory Health (RESPIRE) is funded by NIHR 16/136/109 and NIHR132826 using UK international development funding from the UK Government to support global health research. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the UK government.

Illustrations including the cover page design: Tapas Kumar Mohanty created using Krita version 5.2.9
Formatting and publication design: Tapas Kumar Mohanty and Kitty Flynn

Authorship & Contributions (CRediT)

A Collective RESPIRE Publication Developed by the RESPIRE Data Champions Initiative

Lead Authors

Tapas Kumar Mohanty[1]; Laura Evans[1]; Jayakayatri Jeevajothi Nathan[3]; Chun Lin[1]; Fahmidur Rahman[4]; Farhia Azrin[5]; Maham Zahid[6]; Anuj Kumar Pandey[2]; Harshpreet Kaur[2]; Chanaka Karunarathne[7]; Dulshan Jayasinghe[7]; Beh Hooi Chin[3]; and Akindra Kariyawasam[7].

Affiliations

1. The University of Edinburgh, Edinburgh, UK
2. King George's Medical University, Lucknow, India
3. Department of Primary Care Medicine, Universiti Malaya, Malaysia
4. ARK Foundation, Bangladesh
5. International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b), Bangladesh
6. The Initiative, Riphah International University, Islamabad, Pakistan
7. University of Sri Jayewardenepura, Sri Lanka

Technical and Strategic Review

RESPIRE Platform III Co-leads

- **Tathagata Bhattacharjee**, London School of Hygiene & Tropical Medicine, London, UK
- **Christopher J. Weir**, Edinburgh Clinical Trials Unit, Usher Institute, The University of Edinburgh, Edinburgh, UK

Additional technical reviewers

- **Hani Syahida Salim**, Universiti Putra Malaysia, Malaysia
- **Laetitia Bracco**, University Libraries - Université de Lorraine,
- **Uddhavi Kand**, KEM Hospital Research Centre (KEMHRC), Pune, India

Working Group Contributors

- **Sandeep Bhujbal**, KEM Hospital Research Centre (KEMHRC), Pune, India
- **Hana Mahmood**, Neoventive Solutions, Pakistan
- **Genevie Fernandes**, Stakeholder Engagement Research Fellow
- **Kitty Flynn**, RESPIRE Communications Manager
- **Deesha Ghorpade**, Pulmocare Research and Education Foundation (PURE), India
- **Sapna Madas**, Pulmocare Research and Education Foundation (PURE), India
- **Abhishek Sable**, Pulmocare Research and Education Foundation (PURE), India
- **Murugan R**, Christian Medical College (CMC), Vellore, India

RESPIRE Data Champions Initiative Contributors

Bangladesh

- **Shakhawat Hossain Rana**, ARK Foundation
- **HM Zafor Sharif**, Bangladesh Primary Care Respiratory Society (BPCRS)
- **Md Shariful Islam**, Child Health Research Foundation (CHRF)
- **Kazi Sarmad Karim**, Fasiuddin Khan Research Foundation (FKRF)
- **Nabidul Haque Chowdhury & Tajkia Rumman Worthy**, Projahnmo Research Foundation (PRF)

Bhutan

- **Mimi Lhamu Mynak, Tshering Gyaltsen,** and **Amrita Rai,** Jigme Dorji Wangchuck National Referral Hospital (JDWNRH), Thimphu, Bhutan

India

- **Shilpa Ambalkar** and **Dhananjay Raje**, MAHAN Trust
- **Bharat Choudhari**, KEM Hospital Research Centre (KEMHRC), Pune

Indonesia

- **Tri Mulyani**, Universitas Padjadjaran (UNPAD)

Malaysia

- **Ng Wei Leik**, Universiti Malaya (UM)

Pakistan

- **Syed Yahya Shera**, Neoventive Solutions
- **Amjad Hussain**, The Aga Khan University (AKU)
- **Ahmad Kakakhel** & **Saqib Mustafa** The Allergy & Asthma Institute, Pakistan (AAIP)
- **Saeed Ansaari**, The Initiative

Sri Lanka

- **Lakshani Kaushalya**, **Damilth Nissanka**, **Nishadi Rangana**, & **Ayesha Saumyamali**, University of Peradeniya
- **Gayani Kanchana**, University of Sri Jayewardenepura

Statement on the use of AI tools:

During the preparation of this work, the authors used **the following AI tool/model and version** to assist with structuring, proofreading and formatting manuscript drafts, and with developing ideas for selected illustrations. Some illustrations began as original hand-drawn sketches, which were refined with AI as an intermediate step and then substantially redrawn and finalised by the authors using **Krita version 5.2.9**. The authors reviewed, edited and verified all AI-assisted material and take full responsibility for the originality, accuracy and integrity of the final publication. AI tools are not listed as authors, and their use is disclosed in accordance with applicable WAME and ICMJE guidance.

- *Nov-Dec 2024 and Jul- Sep 2026: OpenAI's ChatGPT-GPT-3.5, GPT-4o, & GPT-5.6 Luna*
- *May–Sep2026: Edinburgh Language Model (ELM)(https://information-services.ed.ac.uk/computing/comms-and-collab/elm) - Llama 3.3, Gemini 3.1 Pro, Claude Sonnet 4.6, & GPT 5.6 Sol,*

Disclaimer:

Please note that, except for REDCap and the tools provided by Edinburgh, any tools or software mentioned in this handbook are not currently provided by the University of Edinburgh (UoE) for use in research. If there is a need to use these tools or software, the following steps must be taken: (1) Contact the IS Helpline and (2) Consult the Data Protection Office to inquire about its use. The use of third-party platforms is generally considered a risk, especially when dealing with personal or sensitive data. The approval process for external systems can be lengthy and may incur additional costs. Therefore, unapproved tools or software cannot be used for RESPIRE-funded research at this time. If you are outside of the RESPIRE collaboration, please consult your data protection officer, who could help you with recommendations and advice by conducting data protection impact assessments.

Acknowledgement

This handbook belongs to the Data Champions who made it necessary. It grew out of the questions they asked, the problems they solved in their own study sites, and the practices they worked out together across seven countries. Everything in these pages was tested somewhere first.

We are grateful to Md. Saiful Hoque (Fasiuddin Khan Research Foundation, FKRF, Bangladesh), Divas Kumar (King George's Medical University, KGMU, India), and Saleem Abbasi (Neoventive Solutions, NS), former Data Champions whose early work helped define what the role could be.

And special thanks to: **Dominique Balhar**ry (Global Respiratory Health Research Manager), Morag Edwards (Unit Operations Coordinator), **Laura Gonzalez Rienda** (Community Engagement and Involvement Research and Operations Assistant), **Emily Healy** (Senior Research Administrator),

We extend our gratitude to **Simon Smith, Kerry Miller, Jennifer Daub, Robin Rice, Dominic Tate, Joy Davidson, Edward Wallace, Jay Evans, S.N. Singh, Kevin Ashley, Agnes Jasinska, Brianna Johns, Pen-Yuan Hsing, Felix Ritchie, Elizabeth Green, Harpreet Sing, Arindam Ray, and Ruth McQuillan** for their mentorship, guidance and valuable advice during the development of the RESPIRE Data Champions Initiative and this handbook.

A handbook of this kind is only as good as the people willing to say what did not work. We thank everyone who did.

Executive Summary

Research data are the foundation of credible, ethical and reusable research. When data are poorly planned, inconsistently recorded, inadequately protected or insufficiently documented, years of work can be lost, and participants' contributions may not achieve their full value. Good research data management prevents these outcomes by establishing clear responsibilities and practical procedures throughout the research lifecycle.

Building Research Data Management Capacity in LMICs: Development and implementation of the RESPIRE Data Champion's Handbook is a practical guide for researchers, study coordinators, clinicians, field workers, principal investigators and newly appointed Data Champions. It does not assume specialist knowledge of data science or computing. Instead, it explains the habits, decisions and tools needed to collect, organise, protect, analyse, share and preserve research data responsibly.
The handbook was developed through the NIHR Global Health Research Unit on Respiratory Health (RESPIRE), led by the University of Edinburgh and Universiti Malaya. It draws on the experiences of 35 Data Champions working across partner institutions in Bangladesh, Bhutan, India, Indonesia, Malaysia, Pakistan and Sri Lanka, alongside UK colleagues. Their work demonstrates that effective data stewardship can be developed through practical experience, a learning-by-doing ethos and institutional support.

The Data Champion Initiative

Data Champions are researchers who act as local points of contact for research data management. They connect study teams with investigators, data managers, ethics committees, information-security specialists and institutional governance services. The role does not transfer accountability away from the wider research team. Instead, Data Champions help ensure that responsibilities are understood and that agreed practices are followed.

Their core responsibilities include:

- supporting the development and regular review of Data Management Plans;
- establishing clear folder structures, file-naming conventions and documentation;
- promoting data quality through validation, audit trails and error reporting;
- protecting personal and sensitive information through appropriate access controls, pseudonymisation and encryption;
- supporting compliance with ethics approvals, institutional policies and funder requirements;
- coordinating responsible data sharing, preservation and secure destruction;
- identifying problems early and sharing solutions across sites; and
- helping colleagues build confidence and skills in data management.

The initiative is decentralised: knowledge and responsibility are embedded within partner institutions rather than concentrated in one coordinating centre. This strengthens local capacity, supports more equitable collaboration and allows

participating institutions to retain and share expertise beyond the life of an individual project.

Practical guidance across the data lifecycle

The handbook follows research data from planning to reuse.

- **Before data collection**, teams should prepare a living Data Management Plan that identifies what data will be generated, who is responsible for it, where it will be stored, how it will be protected and quality-assured, and what will happen to it after the project. Ethics, consent, intellectual property, data protection and cross-border transfer requirements should be addressed at this stage.
- **During collection**, teams should use tested instruments, agreed data standards and secure storage locations. Validation rules should identify missing, inconsistent or implausible values as early as possible. Consent records and direct identifiers should be stored securely and separately from research datasets.
- **During processing and analysis**, original data should be preserved as read-only files. Cleaning and transformation should take place on copies, with every significant change recorded. Data dictionaries, ReadMe files, analysis scripts and audit trails should enable someone outside the original team to understand how the data were produced. Pseudonymised data remain personal data and must continue to receive appropriate protection.
- **During sharing and preservation**, teams should assess disclosure and re-identification risks, complete the supporting metadata, use suitable repositories and apply an appropriate licence. Data should be made as open as possible and as restricted as necessary. FAIR data are Findable, Accessible, Interoperable and Reusable, but FAIR does not always mean publicly open. Sensitive data may require managed or controlled access.

Multi-site and low-resource research

The handbook recognises that research conditions differ across locations. Unreliable power, limited connectivity, shared devices, seasonal access problems, language differences and changing legal requirements can all affect data management. Practical responses include offline-first collection, encrypted local storage, scheduled synchronisation, alternative collection routes for digitally excluded groups and harmonised data standards across sites.

The RESPIRE Methodologies Network also demonstrates the value of co-designing guidance with researchers, communities, policymakers and technical specialists. Effective methods must reflect local realities rather than assume that approaches developed in well-resourced settings will transfer unchanged.

Security, openness and emerging technologies

Responsible openness depends on effective security. The handbook covers minimum-necessary access, strong unique passwords, multi-factor authentication, encryption,

secure backups, phishing awareness and incident reporting. It also provides guidance on spreadsheets, databases, REDCap, OpenRefine, qualitative data, open-source software and institutional repositories.

Artificial intelligence can assist with drafting, summarisation, analysis and other research tasks, but it can also produce false information, reproduce bias and expose confidential data. Identifiable or sensitive participant data must not be entered into unapproved AI systems. Researchers must verify AI-generated content, retain human accountability and disclose the tool, model or version, date and purpose of use where required.

The central message

Good data management does not depend primarily on expensive technology. It depends on planning, clarity, consistency, documentation and a willingness to ask questions. A clear Data Management Plan, secure storage, reliable backups, harmonised variables, accurate metadata and responsible sharing can determine whether a dataset is lost, misused or preserved for future benefit.

The handbook therefore offers both a working reference and a model for institutional capacity-building. Its aim is simple: to help research teams protect participants, produce trustworthy evidence and ensure that valuable data remain understandable, secure and useful long after a study has ended.

> ***Important:*** *Tools, legal requirements and institutional services change over time. Researchers should follow current ethics approvals and consult their institutional information-security, legal and data-protection services before using external tools or transferring sensitive data*.

How to read this handbook

Not sure where to start?	→ Read Chapter 1 first
If you are new to the role of Data Champion?	→ Read Chapters 1, 2, 3 in order
Specific problem to solve?	→ Go straight to the relevant chapter
Stumbled upon jargon or unfamiliar words?	→ Check the Glossary at the back
Looking for a template?	→ Go to Chapter 10

Contents

Chapter 1: About the Handbook

This handbook is for you, regardless of your background or where you are starting from. You might be a newly nominated Data Champion wondering what you have signed up for. You might be a researcher who has never thought much about data management, or a study coordinator, clinician, field worker, or principal investigator looking for practical guidance to share with your team.

You do not need a background in data science or computer science. Everything in this handbook is explained with examples from the kind of research you are already doing.

If you have picked up this handbook, you belong here.

***A quick piece of advice**: Read the section Before We Begin in Chapter 3 before anything else technical. It explains why data governance and data management planning matter, and once you understand the why, everything else makes more sense.*

1.1 How to use this Handbook

Think of this handbook less like a textbook and more like a field guide, something you carry with you, consult when a question arises, and return to as your experience grows. Each chapter stands on its own. Use the roadmap above to find what is most relevant to you right now. At the back, you will find a **Glossary** of every technical term used; if you come across a word you do not recognise, look it up there first.

1.2 The 'Anyone Can Do This' Ethos

Good data practice is not the exclusive territory of specialists. It does not require expensive software, advanced qualifications, or years of technical training. It requires care, consistency, and the willingness to ask questions. The most important data management skills, like thinking ahead, documenting decisions, protecting sensitive information, and keeping things organised, are habits, not technical abilities. Anyone can learn them.

Most of the Data Champions who contributed to this handbook started exactly where you are now. Through working together and trying things out, they built real expertise, and then shared it. This handbook is the product of that collective effort.

1.3 The Journey of Learning by Doing

There is no shortcut to becoming a confident Data Champion. The skills in this handbook are developed through practice, through writing your first Data Management Plan and improving it, through making a mistake with a file naming convention and fixing it, through asking a question in a meeting and getting an answer you did not expect.

Mistakes are not failures. They are part of the process. Every problem encountered at one site becomes a lesson shared across the network.

1.4 What Is a Data Handbook?

A data handbook is a practical guide that tells the people working on a research project how to collect, manage, protect, share, and preserve the data they generate.
It is not a policy document for compliance officers. It is a working tool for the people actually handling data day to day.

Think of it as the answer to the question every new team member should be able to ask: *"How do we do things here?"*

1.5 Why Is One Needed?

Without shared guidance, the same project can produce data that is:

- Stored in different places and different formats across sites
- Documented so poorly that nobody outside the original team can make sense of it
- Handled in ways that breach data protection law
- Lost entirely when a key team member leaves

A data handbook prevents this. It creates one agreed set of practices that applies consistently across every site, every team, and every stage of the research.

> **"If you want people to do something, make it easy."** — Richard Thaler

1.6 The RESPIRE Data Champion Initiative and Why Data Management Shapes Its Impact

What is RESPIRE?

Funded by the National Institute for Health and Care Research, RESPIRE is a Global Health Research Unit focusing on respiratory health in Asia.

> "For the things we have to learn before we can do them, we learn by doing them."— *Aristotle*

The NIHR Global Health Research Unit on Respiratory Health (RESPIRE) at The University of Edinburgh was funded as part of the NIHR Global Health Research Programme in 2016/17. After a

successful first phase with the four partner countries of Bangladesh, India, Malaysia and Pakistan, RESPIRE was awarded further funding to carry out research until 2027, expanding the partnership to include organisations in Bhutan, Indonesia and Sri Lanka.

Led by the University of Edinburgh and the Universiti Malaya, we aim to deliver low-cost, scalable policy and clinical interventions to reduce respiratory disease and death in Asia.

The term respiratory disease covers a range of conditions which affect the lungs and breathing. These conditions are one of the three leading causes of death in Asia and globally. Despite this, knowledge and awareness of respiratory diseases is low, and they do not get the priority attention they need from health systems, policymakers and funding agencies.

Supporting Platforms of RESPIRE

The supporting platforms of RESPIRE increase the ability of our partners to plan, undertake and implement the findings of research in their countries.

The supporting platforms focus on:

I. working with the people who are affected by RESPIRE's research, 'Stakeholder Engagement & Community Engagement and Involvement'
II. increasing the number of health professionals and researchers who are trained to undertake high-quality research, 'Education and capacity building'
III. maximising the uses that research data can be put to in safe and secure ways, 'Open Science, Data & Methodologies', and
IV. enabling access to resources through 'Digital Health and Innovation' and 'Knowledge Mobilisation Hub'.

The supporting Platform III coordinates the RESPIRE's Data Champions initiatives. Based throughout our study sites, the Data Champions enable research teams to maintain the highest standards of data management.

They implement best practices, facilitate the sharing of structured and anonymised data, and actively participate in activities to ensure compliance, while collaborating with other Data Champions to identify and resolve challenges. They also engage in cross-learning to share expertise and exchange best practices to broaden the collective understanding of data management and open science.

1.7 How a Data Handbook influences the impact

The connection between a handbook about data management and the lives of patients in any of the seven LMICs might not be immediately obvious. Let us make it explicit.

RESPIRE's impact depends on the quality and credibility of its evidence. That evidence is only as strong as the data. And data is only trustworthy if it has been collected

consistently, documented accurately, protected appropriately, and shared in a way that allows others to verify and build on it.

- Every time a Data Champion writes a clear ReadMe file, they make data more usable.
- Every time they set up an encrypted backup, they protect the privacy of each participant.
- Every time they harmonise a variable across sites, they make cross-country analysis possible.
- Every time they deposit a dataset openly, they extend the reach of the research.

> *"The goal is to turn data into information, and information into insight."*
> *— Carly Fiorina, Former CEO of Hewlett-Packard*

Chapter 2: Meet the Data Champions

Behind each of the datasets generated in RESPIRE is a person who ensures the integrity and that the information is treated with care, accuracy, and respect. The RESPIRE Data Champions are a network of researchers working across seven South Asian countries as well as Southeast Asian countries to promote good data practice in global health research.

2.1 Why were the Data Champions identified?

To promote consistent and high-quality research data practices across seven Southeast Asian countries and the UK, each RESPIRE study site nominated a Data Champion.

Their purpose is clear: to ensure that every project maintains the highest standards of data management, privacy, and governance while promoting a culture of openness, collaboration, and ethical research.

The initiative promotes sustainable research practice through three complementary approaches:

Decentralisation

The Data Champions Initiative uses a decentralised approach by embedding Champions within partner institutions. This ensures that data quality, privacy and governance principles are applied where research takes place, rather than being concentrated within a single institution.

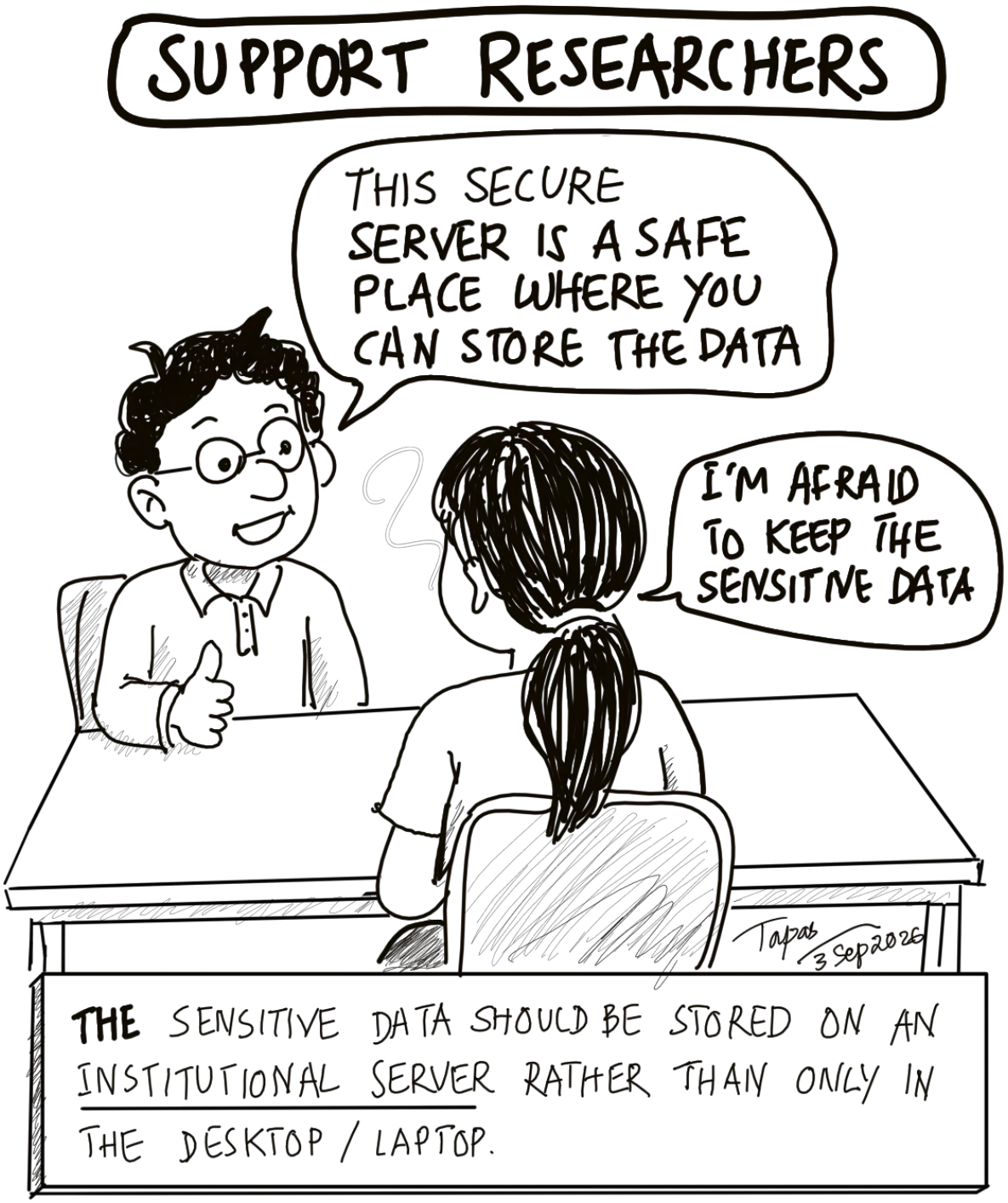
Institutional capacity-building

The initiative strengthens institutional capacity by equipping researchers across partner organisations with practical knowledge and skills in data management, privacy and governance. These skills remain within each institution, where Data Champions can support and train colleagues, extending the initiative's impact.

Equity and sustainability

Distributing expertise across partner institutions reduces reliance on a central team and supports the long-term sustainability of the initiative. It also advances equity within large research collaborations such as RESPIRE by enabling all partners to develop, retain and share the knowledge required to contribute meaningfully to research.

2.2 Key Responsibilities:



Data Champions act as the **primary point of contact** for all data management-related activities within their site or project. They work closely with investigators, data managers, and institutional governance teams to ensure that research data are well-planned, well-documented, and well-protected. Their main responsibilities include:

- **Implementing best practices** outlined in the **Data Management Plan (DMP)** and updating it as the project evolves.
- **Monitoring and reporting challenges** during monthly RESPIRE Data Champion meetings, ensuring that solutions are shared and collectively refined.
- **Facilitating data sharing**, ensuring that all datasets are structured, anonymised, and deposited appropriately at the end of the study.
- **Contributing to the Independent Data Monitoring Committee (iDMC)** by supporting data quality reviews and addressing issues of consistency or missing data.
- **Supporting compliance** with relevant ethics approvals, institutional requirements and funder requirements.
- Providing **appropriate tools and training** for effective data management.
- **Documenting** data-collection and data-cleaning procedures.

2.3 How were the Data Champions identified?

The selection was done by inviting nominations from all RESPIRE partner institutions. In total, **35 Data Champions** were identified across partner sites in Bangladesh, Bhutan, Malaysia, India, Indonesia, Pakistan, and Sri Lanka.

2.4 Benefits of Engaging with Data Champions:

While this is an **honorary position** and does not carry monetary compensation, it brings significant professional and academic benefits.

2.5 What Data Champions gain?

- **Orientation and capacity-building opportunities** in data management and governance.
- **Invitations to cross-learning workshops, online courses, and conferences**, including expert sessions led by leaders in Open Science and Data Privacy.
- **Opportunities for co-authorship** in peer-reviewed publications, in accordance with **ICMJE authorship criteria**.
- **Visibility and recognition** through dedicated RESPIRE communications, newsletters, and the official Data Champions webpage.

In addition, selected Champions have also been mentored and successfully received the **RESPIRE Capacity Building Fellowship**, which supports leadership and mentoring roles in research data governance.

2.6 How do the Data Champions work together?

Data Champions collaborate through **monthly coordination meetings** and **specialised working groups**.:

Working Group	Focus Area
WG-1	Methodologies Network and Stakeholder Engagement
WG-2	Data Management and Governance
WG-3	Handling Varied Data Types and Privacy Legislation
WG-4	Data Storage, Encryption, and Sharing

Each group develops guidance, templates, and case studies for its theme, contributing to the **Data Champion Handbook Series**. These working groups ensure horizontal collaboration across disciplines, encouraging Champions to learn from each other and co-develop solutions, including developing this handbook.

2.7 Looking ahead: 'Learning by doing'

The Data Champion initiative is evolving as a **"learning by doing" journey**.
Most Champions began as researchers with a strong interest in good data practice rather than as trained data stewards. Through hands-on experience, peer discussions, and expert-led sessions, they are now actively shaping how data management and governance are understood and applied within RESPIRE.

This practical approach allows every Champion to **learn, experiment, and improve together,** turning day-to-day challenges into shared learning opportunities. Mistakes become lessons, and lessons become improvements in the next version of a Data Management Plan or a revised metadata template.

Over time, this collective effort will:

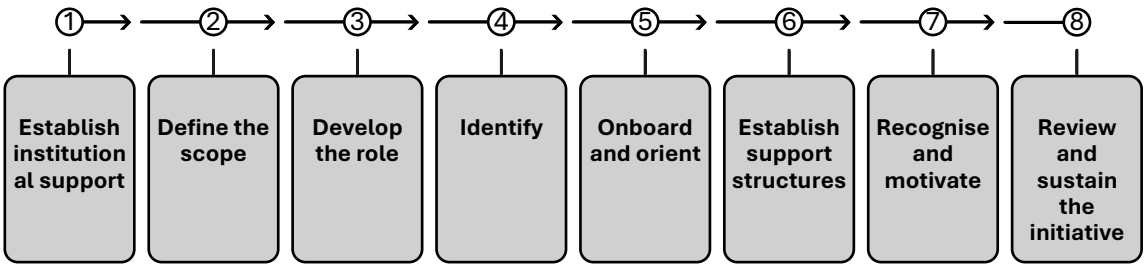
- Build **confidence** in handling complex data management tasks;
- Foster **peer learning and mentoring** within and across study sites;
- Strengthen local **data governance systems**; and
- Create a sustainable **community of practice** that values ethics, openness, and continuous improvement.

The **Data Champions' Handbook** itself is part of this evolving process, which is a co-created, living document that captures lessons, tools, and reflections from Champions across RESPIRE.

2.8 How to Set Up a Data Champions Model

You can replicate or adapt the following steps to establish a Data Champions initiative within your institution or research network.

Implementation pathway



Step 1: Establish institutional support

Establish the initiative with institutional support and endorsement from senior leadership, supported by a clear statement of its purpose and intended outcomes.

- Secure institutional endorsement and clarify why the Data Champions initiative is needed and what it aims to achieve.
- Identify the institutional structures or leaders who will support and oversee the initiative.
- Ensure that Champions have protected time to undertake their responsibilities. Even a modest commitment of **one to two hours per week** can help make the role feasible.

Step 2: Define the scope of the initiative

Define where and how the Data Champions model will operate.

- Decide whether the initiative will cover a single study, department, institution, or a multi-country research collaboration such as RESPIRE.
- Use the scope to determine the number of Champions required.
- Define the governance structure, including how Champions will be coordinated and supported.

Step 3: Define the Champion role

Develop a clear role description so that Champions understand what is expected of them.

- Describe the responsibilities and expected contributions of a Data Champion.
- Specify approximately how much time they are expected to commit, for example, **two to four hours per month**.
- Clarify who Champions report to and who will provide support.
- Explain what Champions can gain from participating, including opportunities for learning, professional development, recognition, and networking.

Step 4: Identify and select Champions

Identify people who are genuinely interested in data quality, good research practice, and research integrity.

- Look beyond IT specialists and Data Scientists; focus on **curious, committed researchers who are keen to do things properly**.
- Consider people who are well connected within their teams and able to encourage good data management practices among colleagues.
- Select individuals who have sufficient interest, time, and institutional support to take on the role.

Step 5: Onboard and orient Champions

Provide new Champions with the knowledge, resources, and support they need to begin the role confidently.

- Provide a copy of this handbook, adapting it to the local context where appropriate.
- Introduce Champions to the person responsible for coordinating the initiative.
- Give Champions an opportunity to review relevant templates and tools, such as a Data Management Plan (DMP).
- Encourage them to consider the question: **"What do I actually do as a Data Champion?"**
- Clarify how the role applies to their own research setting and responsibilities.

Step 6: Establish a support structure

Create mechanisms that allow Champions to learn from one another and obtain help when they encounter challenges.

- Hold regular coordination meetings, preferably monthly.
- Establish opportunities for peer learning so that Champions can ask questions and share experiences with one another, rather than relying only on the coordination team.
- Establish a clear support pathway for issues that are beyond a Champion's knowledge, role, or authority.
- Provide access to appropriate expertise when more specialised advice is required.

Step 7: Recognise and motivate Champions

Recognise Champions for their contribution, particularly because the role is often honorary.

Consider:

- Acknowledging Champions by name in relevant publications and reports.
- Providing certificates of participation that can be included in professional portfolios.
- Offering opportunities for co-authorship when ICMJE authorship criteria are met.
- Highlighting Champions through institutional communications.
- Encouraging and supporting fellowships or other professional development opportunities for outstanding Champions.

Step 8: Review and sustain the initiative

Review the Data Champions programme at least annually to ensure that it remains useful, supported, and relevant.

Ask:

- Are Champions clear about their responsibilities?
- Do they have sufficient time and support?
- What has worked well, and what has not?
- What challenges are Champions encountering?
- What should be changed or added in the next version of the handbook?

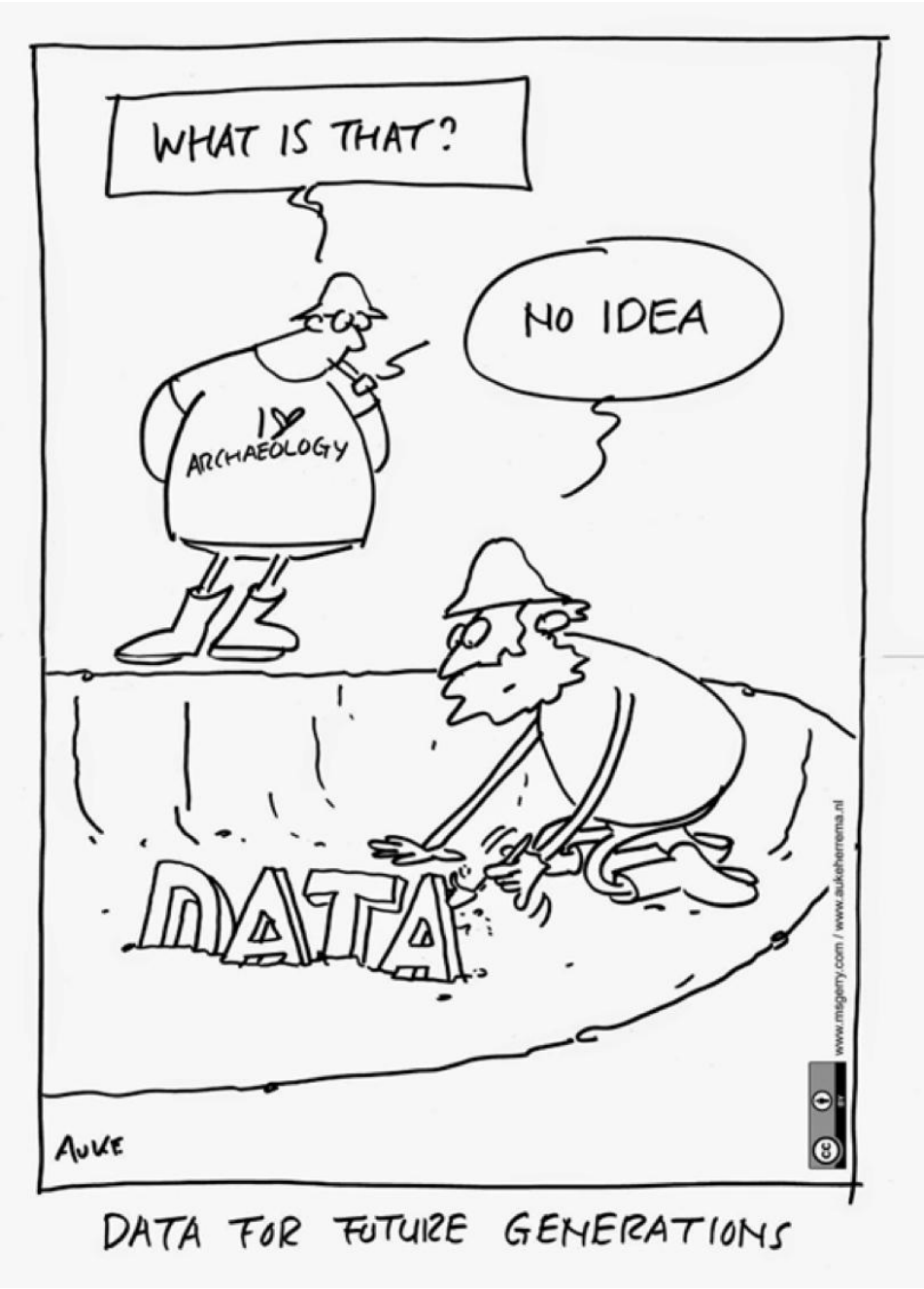
The RESPIRE Data Champion initiative is itself a work in progress. As Champions share their experiences and lessons honestly, the initiative can continue to evolve and improve.

Chapter 3: Foundations

This chapter covers the basic concepts every Data Champion needs to know, what a novice would need. It is organised around the four stages of the research data journey: planning, collection, analysis, and sharing. Technical terms are defined as they appear; a full glossary is at the back of the handbook

3.1 Before we begin: Why are Data Governance and Data Management planning so important?

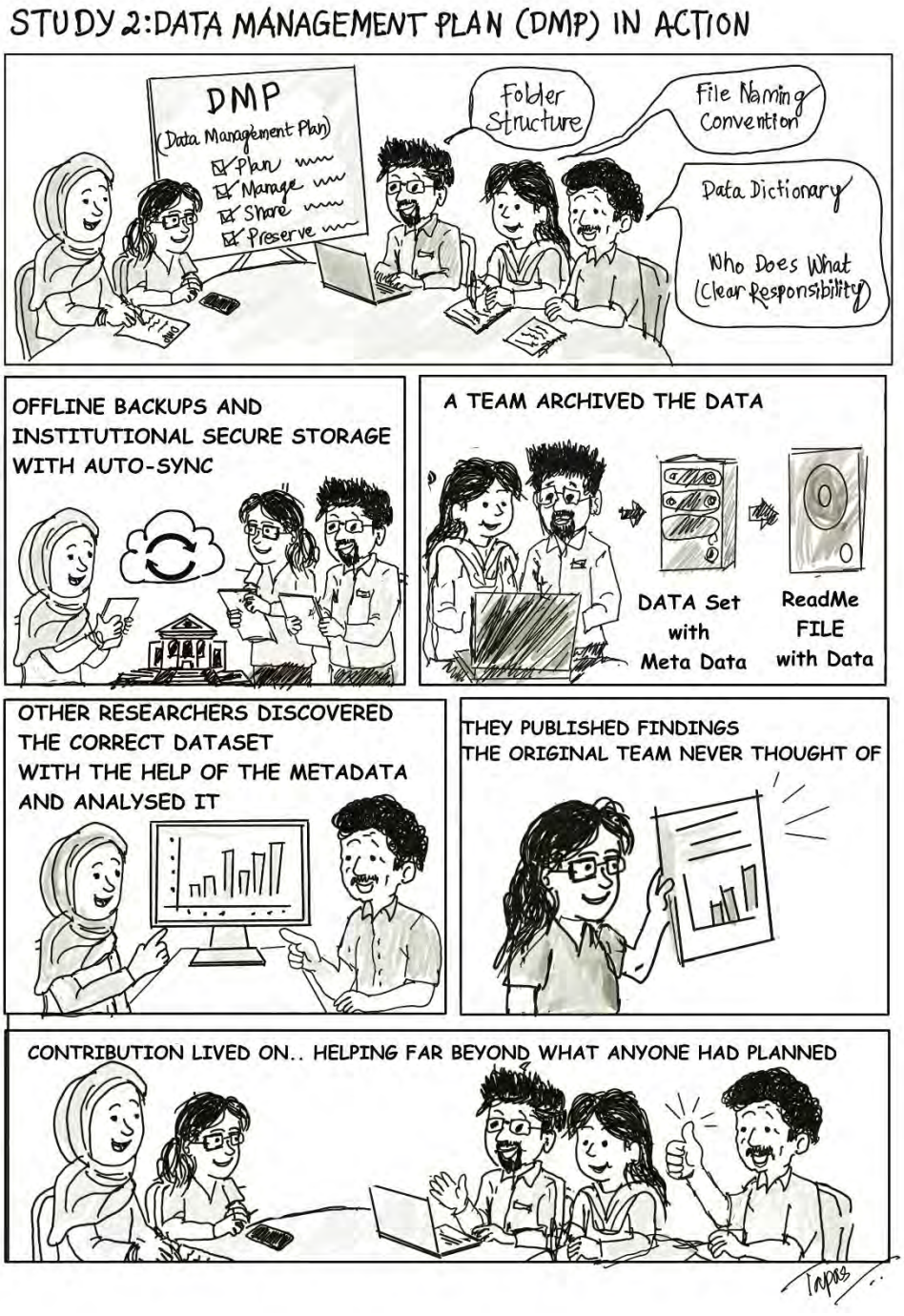
To begin at the beginning



3.1.1 A tale of two research studies

Study 1: A team works for about 5 years to collect respiratory health data from approximately 1,000 participants across 2 countries. At the end of the project, the main dataset is on a personal laptop of a person who left the project 8 months ago and is now not responding to emails or the phone. Although there's a backup, no one knows about the software needed to open it nor the licence required. A third copy is buried in a shared folder named: Study1 user1 final Final v3 final may be definitely, maybe.xlsx, without any data dictionary. Nobody is certain if these backup copies are up to date.

The data is unusable. Five years, 1000 participants, and all the team efforts (including the one who left) went in vain



Study 2: A similar study with the same amount of resources. However, this time, the team drafted a Data Management Plan (DMP) before capturing the participants' data and followed it. They mentioned basic things (not fancy ones) like folder structure, file naming convention, data dictionary, and most importantly, who does what (clear responsibility). They also mentioned the offline backup schedules for the field staff and set up institutional storage with auto-sync of data. Further, they archived the complete dataset with a ReadMe file and a DOI.

After five years, some other researchers discovered the dataset. They combined and analysed it with their own data; they also used the statistical syntax, which they found easier to use instead of doing it from scratch. They published the findings, which the original team had never thought of. The contribution of the whole team lived on, helping far beyond what anyone had planned.

3.1.2 Why does data governance matter?

Data governance is the system of rules, roles, and responsibilities that determines how data is handled. It answers: *Who owns this data? Who can access it? Who is accountable if something goes wrong?* *(See* 3.3.1 What is Data Governance)

> **"Data Governance is not about policing; it is about assigning responsibility and providing support."** — *Robert S. Seiner*, author of *Non-Invasive Data Governance*.

3.1.3 Your role as a Data Champion

You are at the intersection of governance and planning; you aren't the only one responsible for this, but only you can be entrusted to ensure the practice. This means:

- To draft a DMP before the study begins and keep on updating it as the study proceeds.
- Being the point of contact for all data-related questions and problems that arise.
- Identifying gaps and risks and informing the team before they become problems.
- Supporting the team in understanding not just what to do – but why it matters.
- Most importantly, let the study team be aware of their responsibilities under the local and UK Data Protection laws – they must ensure the data collection, storage, and sharing activities comply with those laws (Tapas & Simon, 2024)

There wasn't much difference in the funding or expertise. It was only a simple matter of planning and governance before the study, which was maintained till the end.

When drafting a DMP, you are keeping a promise; when you are setting up control access measures, you are protecting the privacy of a participant. When you deposit anonymised data openly, you are ensuring the contribution was not wasted.

Now let's look at how to do this!

3.2 Core concepts at a Glance

Term	Plain-language definition	Go deeper at
Data	Any recorded information	Ch. 6
Research Data Management (RDM)	How data is organised and stored throughout a project	Ch. 4
Data Governance	Rules, roles, and responsibilities for handling data	Section 3.3
Data Management Plan	Document describing how data will be collected, stored, protected, and shared	Section 3.3
Metadata	Data about data	Section 3.5
Personal Data	Any information that can identify a living person	Section 3.4
Sensitive Data	Personal data requiring extra protection	Section 3.4
Anonymisation	Removing identifiers so a person cannot be identified	Section 3.5
Pseudonymisation	Replacing identifiers with codes while keeping a secure key	Section 3.5
FAIR	Findable, Accessible, Interoperable, Reusable	Ch. 8
Open Science	Making research methods, data, and outputs freely accessible	Ch. 8
Intellectual Property	Legal rights over creations produced during research	Section 3.3
Data Lifecycle	Stages data passes through from planning to archiving	Ch. 4
ReadMe File	Plain text document stored with a dataset explaining what it contains	Section 3.5
Embargo (Data Embargo)	A temporary delay in making research data publicly accessible, usually to allow the research team time to publish their findings or protect intellectual property.	Ch. 8

Note above table: "Definitions here are intentionally brief. Follow the cross-references for fuller explanations. A complete glossary is at the back.

3.3 Four stages of the Research Data Lifecycle

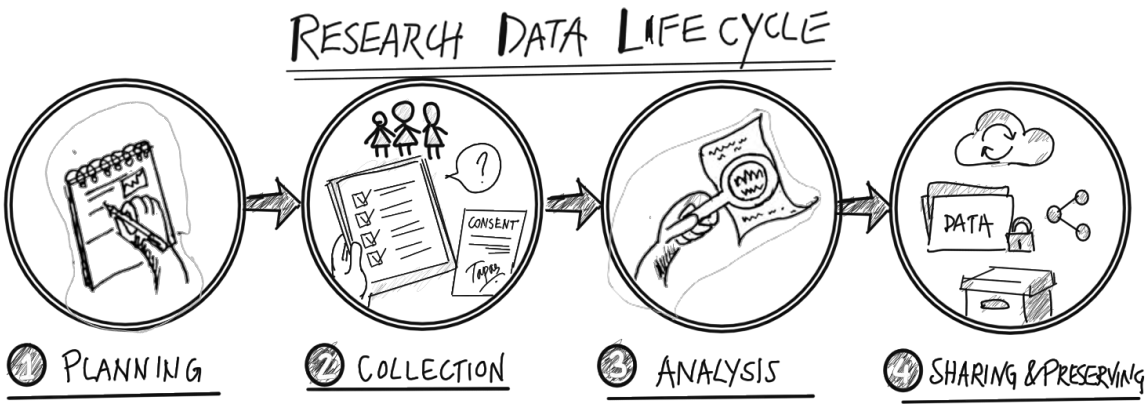


Figure 3.1: The four stages covered in this chapter. Each stage introduces the concepts a Data Champion needs to understand before moving to the practical guidance in Chapter 4

3.3.1 Stage 1: Before You Collect

It's important to understand what is encapsulated in this handbook: the data itself. Every figure recorded in a survey, every transcript typed after an interview, every sensor reading from a clinic is a small fragment of a much bigger picture. When those fragments are gathered, described, and connected, they form evidence. And it is evidence that turns questions into knowledge.

What is Data Governance?

Data Governance lays out the rules, roles, and responsibilities that make RDM possible.
It defines:

- Who owns or stewards the data
- Who can access it and under what conditions
- How legal and ethical compliance is maintained

Governance is much like the **constitution** of data use, which sets the boundaries within which management functions and is known as the Five Safes Framework:

Figure 1 : https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/

Safe people	Safe projects	Safe Settings	Safe Data	Safe outputs
Data is treated to protect any confidentiality concerns.	Research projects are approved by data owners for the public good.	Researchers are trained and authorised to use data safely.	a SecureLab environment prevents unauthorised use.	screened and approved outputs that are non-disclosive.

Why is Data Management Planning so important?

A **Data Management Plan (DMP)** describes how research data will be collected, documented, stored, protected, quality-assured, shared, preserved and, where appropriate, securely destroyed. It should be prepared before data collection begins and reviewed regularly as the project evolves.

A DMP evolves with the project and is revisited regularly, not just written once and forgotten.

A DMP helps research teams to:

- identify data-management risks, requirements and costs early;
- define responsibilities, access arrangements and agreed working practices;
- keep data organised, secure, understandable and accessible throughout the project;
- support research integrity, reproducibility, preservation and responsible reuse; and
- comply with ethics approvals, institutional policies, legal obligations and funder requirements.

A DMP is a living document, not a one-time administrative exercise. It should be updated when the study's methods, personnel, data types, systems or sharing arrangements change.

(See Chapter 10 for the RESPIRE DMP template and completion checklist.)

Intellectual Property (IP) in Research

Intellectual property (IP) refers to the intangible assets you create through innovation in your research: things that exist as ideas, knowledge, or expression rather than as physical objects, and over which legal ownership can be claimed. In research, IP takes many forms, including software and algorithms, datasets and analytical tools, training materials and curricula, diagnostic methods and clinical protocols, novel devices, and the reports and findings you produce. The figure below presents the ownership chain for RESPIRE's projects

Acknowledging IP correctly matters because you cannot manage what you have not recognised. Identifying when your work has produced something new, and declaring any pre-existing or third-party material you have relied on, is the first step in the process. It ensures that credit, ownership, and obligations are clear from the outset, and it allows the collaboration to meet the commitments it has made to its funder.

Protecting IP matters because it safeguards both you and the wider collaboration. If ownership and rights are not properly established, you may find yourself unable to use or deploy your own outputs, or exposed to legal risk. Some decisions are also irreversible: disclosing a potentially patentable output publicly before it has been assessed can permanently destroy its eligibility for protection. Reporting promptly keeps your options open.

Licensing IP correctly matters because it is what allows your work to reach the people it is meant to serve, lawfully and on the right terms. Free access does not remove the need for a formal agreement, and any use that supports revenue generation carries additional requirements. Getting the licence right is what turns a research output into something that can be deployed affordably and responsibly, particularly in lower- and middle-income settings.

What are the types of Data under RESPIRE?

Within RESPIRE, data are captured from the studies across South Asian and Southeast Asian countries and turn local efforts into global knowledge. The information captured for analysis ranges from numbers, text, images, or any other types of information, e.g, air quality (AQI) and temperature captured through a sensor; a field note and an interview recording; at the current pace of evolving emerging technology, you might also take into consideration data like a machine-learning script analysing lung images.

3.3.2 Stage 2: During Collection

What is Personal Data?

Personal data is any information that identifies a person or could reasonably be used to identify them.
It consists of:

- Directly identifiable information (e.g, name, Identification numbers)
- Indirectly identifiable information (usernames, matriculation certification number)

Personal data must be collected after having informed consent, and only for legitimate research purposes, must be deleted when no longer needed

Some types of personal data are considered **'sensitive data',** e.g., data on HIV-positive persons, racial or ethnic origin, data of vulnerable populations, etc.

What is Data privacy and confidentiality?

In health research, personal data requires protection, since it carries an ethical responsibility. Therefore, utmost care should be taken while handling identifiable information to maintain confidentiality.

Importance of Data Privacy & Confidentiality

Personal data could be misused for making a profit or for fraudulent activity. Therefore, it is the responsibility of the research institutions to protect the confidentiality of personal data.

	A	B	C	D	E	F	G	H
1	Age	Area/ Locality	Gender	Hospital	Ethnicity	Diagnosis	Smoker	Income
2	51	King Street	F	Royal Infirmary Edinburgh	Asian	COPD	Yes	£81,500

*Combining indirect identifiers may re-identify the participants

What are the differences between privacy and Confidentiality?

Privacy	Confidentiality
It is the broader concept of being free from intrusion **into one's personal matters.**	It is associated with **professional relationships** where individuals share personal information with the expectation that it will be kept secret
Individual's right to control their personal information	Obligation of the **research institutions** to protect sensitive information from unauthorised access and disclosure

What is data protection regulation?

It regulates personal data and ensures privacy by establishing guidelines for how organisations can lawfully process data. *(Further reading: Open Science & Open Data: Envisioning Open Science Road Map for RESPIRE Project Partner Countries, DOI: https://doi.org/10.2218/ED.9781836450764)*

Principles of Data Protection

Principle	What It Means
Lawfulness and transparency	You must have a legal basis for collecting data; participants must know how it will be used
Purpose limitation	Only use data for the reason it was collected
Data minimisation	Collect only what you genuinely need
Accuracy	Keep data accurate and current
Storage limitation	Should be kept no longer than necessary
Security	Protect data from unauthorised access, loss, or damage

Informed consent

Informed consent is the process through which a participant voluntarily agrees to take part in a research study after receiving clear information about what participation involves. This includes what data will be collected, how they will be used, who may access them and how long they will be retained. Consent is more than a signature on a form. It is an ongoing process of communication and trust throughout the study.

Consent to participate in research is distinct from the lawful basis used to process personal data. Under the UK GDPR, universities and other public research organisations commonly rely on **"task in the public interest"** rather than consent as their Article 6 lawful basis. Processing health, genetic or other special-category data also requires an appropriate additional condition. Other organisations and jurisdictions may rely on different legal grounds.

Participant consent may still be required for ethical participation, access to confidential information, use of tissue or other study activities. Giving consent to participate does not automatically provide the lawful basis for processing personal data, and withdrawing from a study does not necessarily require all previously collected data to be deleted. The lawful basis and applicable conditions should be identified in consultation with the relevant data-protection, legal and research-governance offices and recorded in the study documentation.

For research participation to be based on valid consent, it should be:

Condition	What it means in practice
Freely given	The participant chooses freely whether to take part and can say no or withdraw at any time without penalty or loss of care or benefits. Reimbursement or incentives should be reasonable and should not pressure participants to take part, especially where there is a dependency or power imbalance.
Informed	The participant receives clear, accurate and accessible information in a language and format they can understand.
Specific	The activities and purposes covered by consent are clearly described. Separate choices should be provided for distinct or optional uses where required.
Unambiguous	The participant indicates agreement through a clear affirmative action, such as signing a form or providing appropriately documented verbal consent.

In RESPIRE studies, consent materials should reflect participants' linguistic, cultural and educational needs. Where literacy is limited, appropriately documented verbal or witnessed consent may be used if permitted by the study protocol, ethics approval and applicable law.

Participants may withdraw from a study without penalty or loss of care or benefits to which they are otherwise entitled. Following withdrawal, no further data should normally be collected, except where collection is required for safety, legal or regulatory reasons. The handling of data already collected depends on the study protocol, participant information, ethics approval, applicable law and lawful basis for processing. Data that have been anonymised or incorporated into completed analyses may no longer be identifiable or removable. These limitations should be explained clearly during the consent process.

Requirements for children and young people (for instance under 16 or 18 years of age) vary by jurisdiction and study context. Where participants cannot provide consent independently, permission should be obtained from an authorised parent or guardian, together with the child's assent where appropriate. Researchers must follow the approved protocol, applicable law and relevant ethics guidance.

Data Champion responsibility: Ensure that consent records are stored securely, separately from the main research dataset and with access limited to authorised personnel. Retain them according to the approved protocol and institutional retention requirements. Confirm that the study documentation clearly distinguishes consent to participate from the lawful basis for processing personal data.

(For consent form templates and further guidance on gaining informed consent for data sharing and archiving, see Chapter 10: Tools and Templates.)

3.3.3 Stage 3: During Analysis

Structured and Unstructured Data

During analysis, knowing what kind of data you are working with is not just a technical detail; it directly shapes how you document, protect, and prepare data for sharing. The most important distinction to understand is between structured and unstructured data.

Structured data are organised in predefined fields, usually rows and columns, such as survey responses, clinical measurements and sensor readings. **Unstructured data** do not follow a fixed tabular format and include interview transcripts, field notes, photographs, audio and video recordings.

Both types require appropriate documentation, secure storage and privacy protection. Unstructured data may require more careful review because contextual details can increase re-identification risk.

(See Chapter 6 for detailed guidance on managing different data types)

Metadata and the ReadMe File

Metadata stands for data about data; in other words, it is the descriptive information that tells others what a dataset contains, how it was collected, and how it should be used. Without metadata, a dataset is difficult to find, hard to understand, and impossible to use correctly by anyone outside the original research team. Good metadata is what makes data genuinely reusable and is a core requirement of FAIR compliance.

> *Think of it this way: if a jar had no label, you would have to open it to check what is inside. Metadata is the label on the jar. The ReadMe file is the letter inside the jar, telling a stranger exactly what is in there, where it came from, and how to use it without any help from you.*

A **ReadMe file** is a plain text (.txt) file stored alongside your dataset. Every dataset should have one. It should include:

- What the data is about and how it was collected
- What each variable means
- File naming conventions used
- Abbreviations and codes
- Missing data codes
- Any processing or cleaning that has been done
- Known issues or limitations

The ReadMe file does not need to be long or technical. It needs to be clear enough that a researcher who has never met you, working five years from now, can open your dataset and understand exactly what they are looking at.

> 📝 *"Data Champion responsibility: Write the ReadMe file at the same time as data collection begins, not after the study ends. Updating it as you go takes minutes. Writing it from scratch at the end of a project takes days and is often incomplete."*

(See Chapter 10 for the RESPIRE Author Dataset ReadMe Template, which you can adapt for your study.)

Anonymisation & Pseudonymisation

When preparing data for analysis or sharing, you must protect participant privacy by removing details that could identify them. The two main techniques for this are pseudonymisation and anonymisation. While people often use these terms interchangeably, legally and practically, they are very different.

Pseudonymisation replaces identifying information with an artificial code or pseudonym. The data can still be linked back to the individual using a securely stored master key. This is standard practice for internal team analysis.

Anonymisation removes or alters identifying information so that the risk of re-identifying an individual is sufficiently low, taking account of the data, other information that may reasonably be available, and the context of disclosure. Unlike pseudonymised data, anonymised data cannot be linked to an individual using an additional key or readily available information. Before data are shared openly, the risk of re-identification should be formally assessed and reduced to an acceptable level in accordance with applicable law, ethical approvals and institutional requirements.

The table below illustrates how a single participant record changes depending on how the data is being used:

Identifiable Data (Contains direct identifiers)	**Pseudonymised Data** (Identifiable information is replaced by participant ID)	**Anonymised Data** (No reidentification is possible)
Sensitive information - Participant recruitment & consent	Analysis and team collaboration	**Suitable for sharing**
Name: John Connor	**Participant ID:** P001	**Random ID:** 012
DOB: 15 Mar 1985	DOB removed and age is mentioned as 41	Age Group: 35–44
Address: 12 King Street	Address removed	Region: Scotland
Hospital: Royal Infirmary Edinburgh	Royal Infirmary Edinburgh	Hospital 1
Income: £45,500	£45,500	£40k–£50k
Diagnosis: Type 2 Diabetes	Type 2 Diabetes	Metabolic Condition

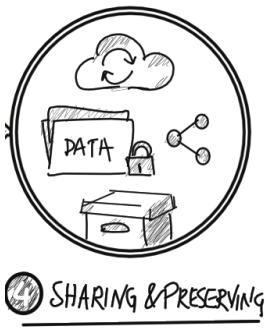
💡 *"Legal Note: Under data protection laws such as GDPR, pseudonymised data is still legally classified as personal data because re-identification remains possible. Only fully anonymised data falls outside data protection regulations."*

The Challenge of Qualitative Data : Numbers and categories are relatively easy to anonymise or pseudonymise using the table above. Qualitative data, however, such as interview transcripts or field notes, is much more complex. Participants often share highly specific stories or combinations of details that make them identifiable even if their name is removed. Because qualitative data can rarely be fully anonymised without destroying its scientific value, it requires careful, deliberate handling during the analysis stage.

> 💡 *"Rule of thumb: If a participant reading your research output could recognise themselves from the details provided, further anonymisation is needed."*

(See Chapter 6 for a step-by-step practical guide on how to pseudonymise qualitative data and manage a secure identification key.)

3.3.4 Stage 4: Sharing and Preserving



What is Data Sharing?

When analysis is complete, data moves toward sharing and long-term preservation. This is the stage where good planning pays off. Well-documented, properly pseudonymised or anonymised data can be shared openly and reused by others for years to come.

Open Science & FAIR Principles

Open Science is the movement to make scientific research, its methods, data, and outputs, freely accessible to anyone, ensuring that knowledge generated through publicly funded research serves the public good. At the heart of responsible data sharing are the FAIR principles, which describe what well-shared research data looks like: data should be **F**indable, so others can discover it through repositories and rich metadata; **A**ccessible, meaning access conditions are clearly stated even when data must remain restricted; **I**nteroperable, using open and standard formats that any system can read; and **R**eusable, supported by complete documentation and an open licence such as CC BY 4.0 that gives others explicit legal permission to use, adapt, and build on the work.

For RESPIRE studies, the guiding principle is simple: make data as open as possible and as closed as necessary, balancing the scientific value of sharing with the ethical responsibility to protect participants.

Data sharing should take place either:

- At the time of publication
- Or, at the end of the project/ end of the funding, whichever is earlier, without unnecessary embargo periods.

(See Chapter 8 for full guidance on FAIR principles, open repositories, licensing, encryption, and how RESPIRE contributes to open science monitoring.)

Open Licence

An open licence is a legal permission that allows others to use, adapt, and share your work freely, without needing to seek individual permission each time. Creative Commons Attribution 4.0 (CC BY 4.0) is the recommended open licence for RESPIRE research outputs and datasets, permitting anyone to use, adapt, and share the work for any purpose, including commercially, as long as they give appropriate credit to the original creators.

(See Chapter 8 for full guidance on open licences, how to apply them, and which licence is appropriate for your dataset.)

Data Destruction

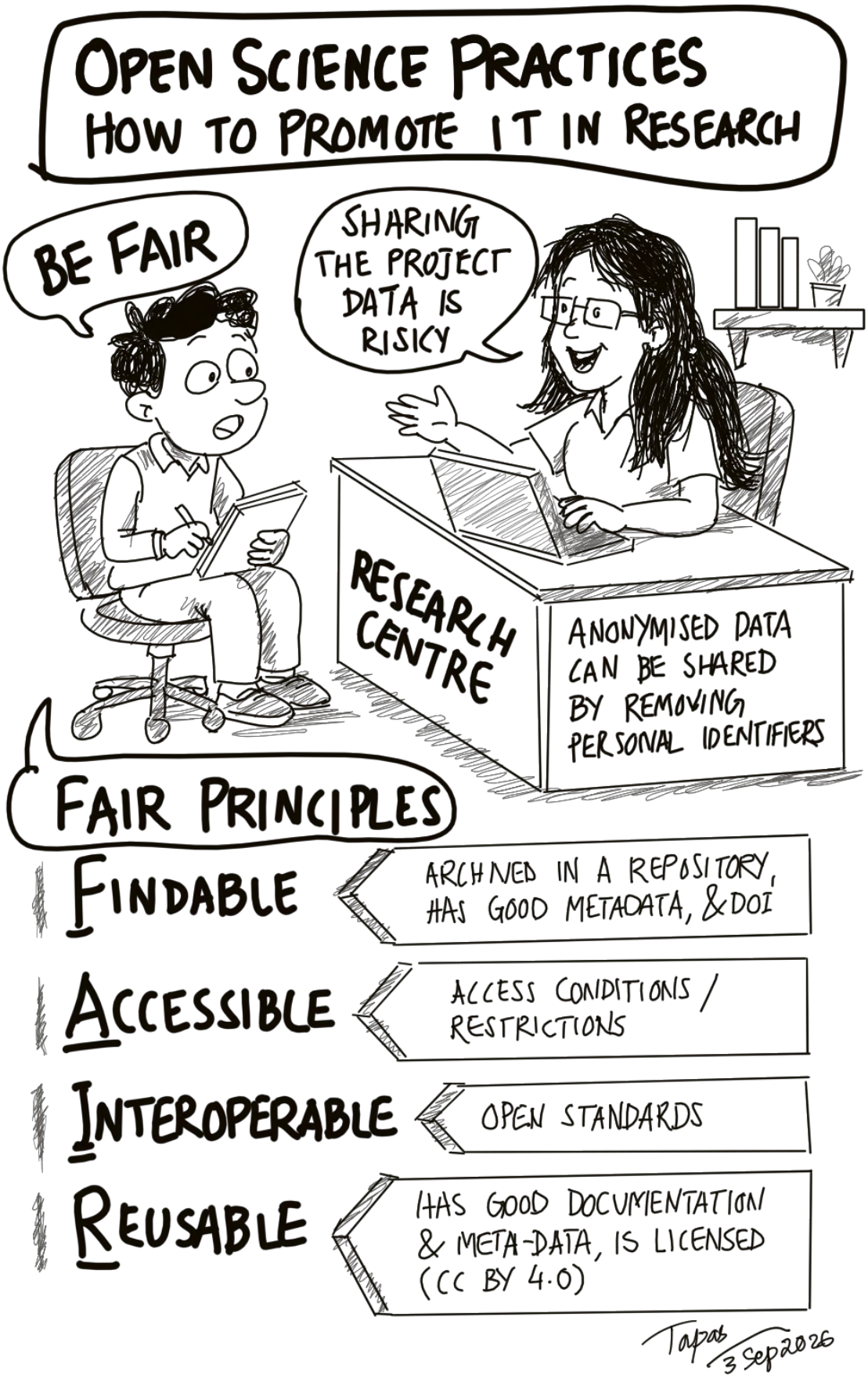
Personal data must not be kept forever. When data is no longer needed:

- **Paper records**: securely destroyed by cross-cut shredding
- **Digital data**: permanently deleted using secure deletion methods; not simply moved to a recycle bin.

The Issue with digital data: The data deleted from the drive (internal/ external HDD or Thumb drives) **can usually be restored** as long as new data has not overwritten the old files. When you delete a file, the system only marks that space as empty, leaving the actual data hidden on the drive until new files take its place

To ensure the secure deletion of data from portable media, you need to encrypt the data to be deleted using an encryption key that won't be remembered or discovered, and then delete the data. This process ensures that just deleting or reformatting isn't the final step, adding an extra layer of security to the data deletion process.

> ***Note this in your DMP**: Every study should specify a retention period and a destruction procedure.*


OPEN SCIENCE PRACTICES
HOW TO PROMOTE IT IN RESEARCH
BE FAIR
SHARING THE PROJECT DATA IS RISICY
RESEARCH CENTRE
ANONYMISED DATA CAN BE SHARED BY REMOVING PERSONAL IDENTIFIERS
FAIR PRINCIPLES
FINDABLE
ARCHIVED IN A REPOSITORY, HAS GOOD METADATA, & DOI
ACCESSIBLE
ACCESS CONDITIONS / RESTRICTIONS
INTEROPERABLE
OPEN STANDARDS
REUSABLE
HAS GOOD DOCUMENTATION & META-DATA, IS LICENSED (CC BY 4.0)
Tapas 3 Sep 2026

Chapter 4: The Research Data Lifecycle

Good research data management is not about filling out paperwork; it is about answering practical questions, such as:

- How will we record and back up information?
- How will others understand what our dataset means?
- How can we ensure that the data remains usable in five years' time?

To answer these questions, Data Champions guide their teams through the Research Data Lifecycle.

The Data is not static. It moves through stages and at each stage, different decisions need to be made, and different risks need to be tackled. The Data Champion has a role at every stage.

4.1 The Research Data Lifecycle at a glance:

4.1.1 Stage 1 : Planning:

Before any data is captured, the team needs to plan what data will be collected, how it will be managed, and by whom.

Data Champion's responsibility:

- To ensure the Data Management Plan (DMP) is drafted by involving study team members (see chapter 10 for the RESPIRE DMP template)

- Ensure ethics approval covers all planned data types
- Set up the agreed folder structure and file naming conventions before data starts flowing
- Agree on data harmonisation standards across sites (*see Data Harmonisation section below*)

Note: A few hours spent planning at this stage will save days of work later.

4.1.2 Stage 2: Collecting

What happens: Data is gathered from participants through surveys, interviews, clinical forms, sensors, or other instruments.

Champion responsibilities:

- Verify with the team that data collection instruments are tested and standardised across sites.
- Confirm that informed consent is documented properly for every participant
- Ensure data is saved in the agreed format and location from day one.
- Monitor for early quality issues regularly (catch-up meetings): flag problems before they multiply.

Common mistake: Collecting data before the DMP and ethics approval are finalised. This can put the whole study at legal and ethical risk. Flag this immediately if you see it happening.

4.1.3 Stage 3: Accessing & Analysing

What happens: Raw data is checked, cleaned, and organised, transcribed from paper, entered into databases, or processed by software.

Champion responsibilities:

- Ensure the original (raw) data is backed up and kept read-only. Never edit it.
- Oversee data cleaning and document every change made.
- Maintain an audit trail: a record of what changed, who changed it, and when.
- Separate identifiable information (keep it securely) and apply pseudonymisation before sharing with the broader study team.
- Ensure derived datasets are clearly documented and linked to their source data.
- Make sure analysis code is commented and explainable by someone who did not write it. Encourage the use of open-source tools for reproducibility (such as R or Python) and ensure syntax files are saved alongside the data.

💡 ***Golden rule:*** *Never edit the original data file. Always work on a copy. Keep the original, read-only copy in a separate, protected location.*

4.1.4 Stage 4: Publishing & Sharing

What happens: Anonymised data is made available to other researchers and the public.

Champion responsibilities:

- Verify data is fully anonymised before any public sharing
- Support deposit to a public repository (e.g., Edinburgh DataShare, Zenodo)
- Ensure the dataset has a DOI and is licensed under CC BY 4.0
- Confirm the dataset meets FAIR principles (see Chapter 8)

4.1.5 Stage 5: Preserving

What happens: The project ends, and data needs to be stored safely for the long term.
Champion responsibilities:

- Coordinate deposit to the appropriate archive or institutional secure server (e.g., University of Edinburgh DataVault)
- Ensure the dataset is complete, documented, and has a Metadata schema and ReadMe files
- Confirm the retention period specified in the DMP is recorded
- Ensure any data no longer needed is securely destroyed

4.1.6 Stage 6: Re-use

What happens: Other researchers discover, access, and use the data in future studies.
Champion responsibilities:

- Ensure metadata is complete and clear enough for an outsider to understand and use the data
- Monitor whether the dataset has been cited or reused, this is evidence of impact
- Feed lessons from this lifecycle back into the next DMP

4.2 Active Storage and Collaboration

The University of Edinburgh Data Ecosystem

To manage data safely during Stages 2 and 3, you need to know what approved tools are available to you.

(Note for LMIC Partners: The tools listed below are specific to the University of Edinburgh. If you are based at a partner institution, ask your local IT department for your equivalent secure storage and sharing services; avoid using personal Dropbox or Google Drive accounts for RESPIRE data.)

Service	Purpose	Best For
DataStore	Secure institutional storage	Active research data
DataSync	File sync and sharing tool	Internal and external collaboration during the project
DataVault	Long-term archival of the "golden copy"	End-of-project preservation
DataShare	Public repository with DOI assignment	Sharing anonymised data openly
Pure	Research information management system	Tracking and reporting outputs

Internal Sharing

Public data sharing usually takes place when data are prepared for wider access, often towards or after the end of a project. Internal sharing occurs throughout the research lifecycle. Apply the principle of minimum necessary access: share data only with authorised individuals who need them, limit access to the data required for their role, and remove access when it is no longer needed.

Service	Appropriate use	Important considerations
DataSync	Securely synchronising and sharing working files with authorised internal and external collaborators	Synchronisation is not the same as backup. Deletion, corruption or unwanted changes may be propagated across synchronised locations.
DataStore	Storing and working with active project data using institutionally managed storage	Use project folders and role-based permissions. Confirm the service's current backup, retention and recovery arrangements rather than treating it as multiple independent copies.
SharePoint / Teams	Collaborating on documents, managing projects and sharing files with authorised team members	Use only where approved for the relevant data classification. Do not store highly sensitive or identifiable raw data unless institutional policy and the study's governance arrangements explicitly permit it.
DataVault	Retaining a stable, definitive or "golden" copy of data that are no longer being actively changed	DataVault is intended primarily for long-term retention of inactive or completed research data. It should not be treated as a routine backup system for frequently changing active files unless the service provider confirms that use.

4.3 Active Backup Strategy: The 3-2-1 Rule

Hardware can fail, laptops can be lost, and files can be accidentally deleted. Every RESPIRE project should maintain a robust backup strategy throughout the active phase of research, in accordance with applicable institutional policies, funder requirements and study-specific data management plans.

The gold standard is the 3-2-1 Rule:

- 3 copies of your data
- 2 stored on different media (drives or systems)
- 1 stored off-site or in the cloud

Copy	Where	How
1 (Primary)	Active working folder on DataStore	Automatically backed up by the institution
2 (Local/Media)	Encrypted external hard drive	Updated regularly using approved sync software
3 (Off-site/Archive)	DataVault (completed data)	Deposited at the end of the project

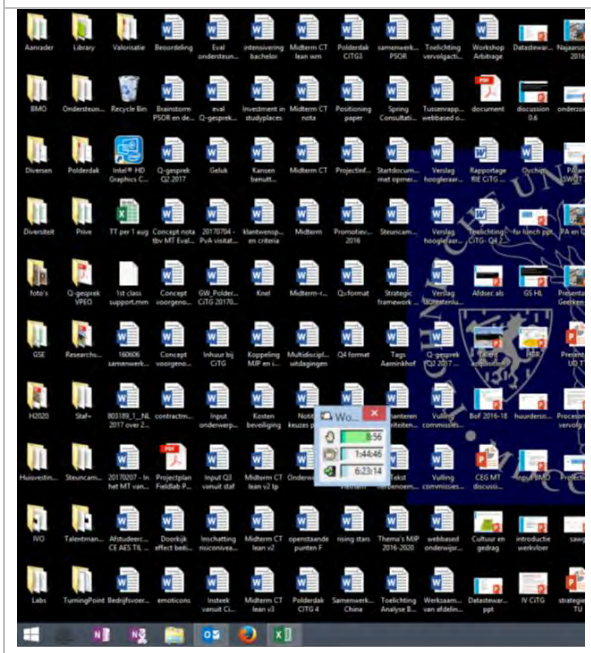
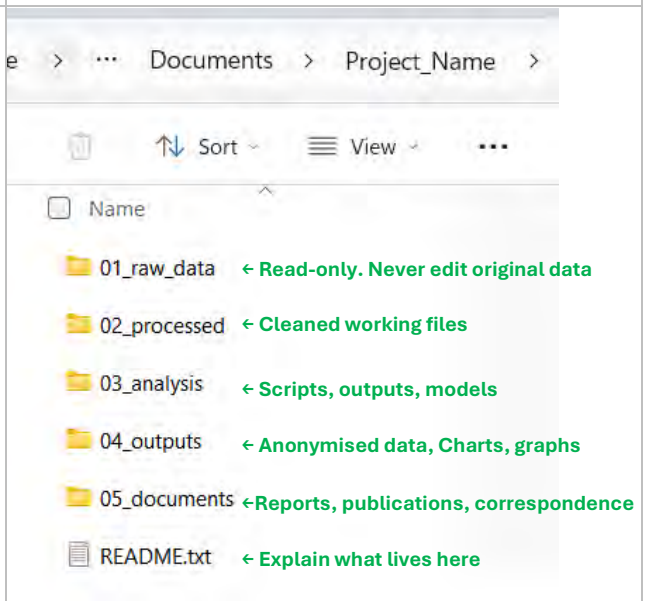
Note on Backup: *Using DataStore, DataSync and SharePoint or Teams does not automatically satisfy the 3-2-1 backup principle. Synchronised folders and service-level backups may not constitute separate, independently controlled copies on two media or systems. Research teams should document where each copy is held, how independent the copies are, whether at least one is stored at a separate location, and how data can be restored. Current arrangements should be confirmed with the University's Research Data Support and Information Services teams.*

4.4 Organising Your Files & folders

4.4.1 Folder structure

Organise your folders systematically before data starts flowing. The computer desktop is not a place to store files. Follow a master folder architecture that separates raw data from processed data and documentation.

Organise your folders before data starts flowing:

☒ Because Desktop is not a place for everything	☑ So do this
	

4.4.2 A File by Any Other Name

- **Use simple characters:** Use letters, numbers, hyphens and underscores only, unless your institutional system permits additional characters. Use a full stop only before the file extension.
- **Avoid spaces:** Separate words with underscores (_) or hyphens (-).
- **Use sortable dates:** Record dates consistently in ISO format, YYYY-MM-DD or a simpler format YYYYMMDD.
- **Apply version control:** Add a sequential version number at the end of the filename, such as v01, v02 and v03. Avoid ambiguous terms such as final, final2 or latest.

Example: I_P02_R01_20260731_v1.docx

- I = interview
- P02 = participant ID
- R01 = researcher ID
- 20260731 = date
- v1 = version

> 💡 ***Rule of thumb: File naming should be short, consistent, and meaningful to someone who has never seen the file before.***

4.4.3 File Formats for the Future (FFFF)

The goal: anyone should be able to open and use your data in long-term preservation
Avoiding: Proprietary formats may not be interoperable
FOSS Friendly: Popular and common formats which can be opened using Free and Open Source Software

Instead of	Use	Why
`.xlsx`	**`.csv`**	Open, compact, universally readable – for quantitative structured data
`.docx`	**`.pdf`**	Preserves layout, no specialist software needed
`.docx`	**`.odt`**	Open Document Text — fully open standard, no licence required
Proprietary stats files	**`.csv` + syntax file**	Reproducible, shareable, software-agnostic

For recommended File Formats, please visit https://www.ed.ac.uk/files/atoms/files/research-data-service_recommended_file_formats_august2018.pdf

4.5 Data Quality Control

Data Champions should encourage rigorous quality control at two stages.

At data entry, perform real-time Validation

- Range checks: flag values outside expected bounds (e.g., age recorded as 250)
- Type checks: numeric fields should only receive numbers
- Completeness checks: flag mandatory fields left blank
- Cross-field checks: flag logical impossibilities (e.g., discharge date before admission date)

After collection, perform quality control

- Regular automated checks: scripts that identify outliers, duplicates, and missing values
- Cross-verification: compare entries against source documents
- Audit trails: record all edits (what changed, who changed it, and when).
- Error logs: document issues found and corrective actions taken

Monthly meetings: Bring quality control issues to the monthly Data Champion meetings. A problem at one site is probably happening at others too.

4.6 Data Cleaning, Wrangling and Harmonisation

We all love the idea of working with clean, reliable data; it's what makes everything else possible. But when it comes to actually *doing* the cleaning? Well… that's a different story (Curty, Renata. 2025).

Source: Adapted from Schmarzo (2021).

Real-world data often contain missing values, duplicate records, inconsistent formats, implausible values and other errors that can affect analysis. Although the terms **data cleaning**, **data wrangling** and **data harmonisation** are related, they describe distinct processes.

- **Data cleaning** identifies and addresses errors or quality problems within a dataset, such as duplicates, missing values, inconsistent entries and outliers.
- **Data wrangling** restructures or transforms data into a format suitable for analysis. This may include renaming variables, changing data types, reshaping tables and combining files. Data cleaning is often one part of the broader wrangling process.
- **Data harmonisation** aligns data from different sites, studies or systems so that they can be combined and compared consistently. It may involve reconciling variable names and definitions, standardising date formats, units of measurement and category labels, and documenting differences in data-collection methods.

Harmonisation is particularly important in multi-site collaborations such as RESPIRE, where participating institutions may use different instruments, terminology, coding systems or software. Wherever possible, teams should agree on shared data definitions, formats and coding conventions before collection begins. When retrospective harmonisation is required, all transformations and decisions should be documented to preserve transparency and reproducibility.

Tools such as spreadsheets, statistical software and **OpenRefine** can support cleaning, transformation and harmonisation. However, automated changes should always be reviewed, validated and recorded. Original data should be retained unchanged, and processing should be performed on a controlled working copy.

4.6.1 Why It Matters?

Without harmonisation:

- Variables may use different units (kg vs. lb)
- Date formats may differ (DD/MM/YYYY vs. MM/DD/YYYY)
- Coding schemes may not match (0/1 vs. M/F)
- Missing data may be represented differently

These inconsistencies can silently corrupt a combined dataset.

Practical Steps

1. Build a common data dictionary, define every variable before collection begins:

Variable	Definition	Type	Allowable Values	Unit	Missing Code
age	Age in completed years at enrolment	Integer	0–120	Years	-99
sex	Biological sex	Categorical	1=Male, 2=Female, 3=Other	n/a	-99

2. Standardise date formats: use ISO 8601: Use ISO 8601 YYYY-MM-DD in datasets and documentation. For filenames, use the compact format YYYYMMDD.

3. Agree on missing data codes: document before collection:

- 99 = not applicable
- 88 = not collected or refused
- 77 = not known

4. Align response scales: a 5-point scale at one site and a 4-point scale at another cannot be pooled without adjustment.

5. Document every harmonisation decision: in the DMP and the ReadMe file.

4.6.2. Tools That Help

OpenRefine: A Safe Tool for Cleaning Sensitive Data

What is OpenRefine?

OpenRefine is a desktop application that uses your web browser as a graphical interface. It is described as “a power tool for working with messy data” (David Huynh) - but what does this mean? It is probably easiest to describe the kinds of data it is good at working with and the sorts of problems it can help you or your team solve.

It is most useful where you have data in a simple tabular format such as a spreadsheet, a comma-separated values file (CSV) or a tab-delimited file (TSV), but with internal inconsistencies either in data formats, or where data appears, or in terminology used. OpenRefine can standardise and clean data across your file. It can help you:

- Get an overview of a data set
- Resolve inconsistencies in a data set, for example, standardising date formatting
- Help you split data up into more granular parts, for example splitting up cells with multiple authors into separate cells
- Match local data up to other data sets - for example, in matching forms of personal names against name authority records in the Virtual International Authority File (VIAF)
- Enhance a data set with data from other sources

Some common scenarios might be:

- Where you want to know how many times a particular value (name, publisher, subject) appears in a column in your data
- Where you want to know how values are distributed across your whole data set
- Cleaning up the date format. Where you have a list of dates which are formatted in different ways, and want to change all the dates in the list to a single common date format with a regular expression.

Figure 2 Regular Expressions. Source:XKCD, CC BY-NC 2.5 license.

For example:

Data you have	Desired data
1st January 2024	2024-01-01
01/01/2024	2024-01-01
Jan 1 2024	2024-01-01
2024-01-01	2024-01-01

- Where you have a list of names or terms that differ from each other but refer to the same people, places or concepts. For example:

Data you have	Desired data
Bangladeshh	Bangladesh
Bangladesh]	Bangladesh
Bangladesh,]	Bangladesh
bangladesh	Bangladesh

- Where you have several bits of data combined in a single column, and you want to separate them into individual bits of data with one column for each bit of the data. For example, going from a single address field (in the first column), to each part of the address in a separate field:

Address in single field	Institution	Library name	Address 1	Address 2	Town/City	Region	Country	Postcode
University of Wales, Llyfrgell Thomas Parry Library, Llanbadarn Fawr, ABERYSTWYTH, Ceredigion, SY23 3AS, United Kingdom	University of Wales	Llyfrgell Thomas Parry Library	Llanbadarn Fawr		Aberystwyth	Ceredigion	United Kingdom	SY23 3AS
University of Aberdeen, Queen Mother Library, Meston Walk, ABERDEEN, AB24 3UE, United Kingdom	University of Abderdeen	Queen Mother Library	Meston Walk		Aberdeen		United Kingdom	AB24 3UE
University of Birmingham, Barnes Library, Medical School, Edgbaston, BIRMINGHAM, West Midlands, B15 2TT, United Kingdom	University of Birmingham	Barnes Library	Medical School	Edgbaston	Birmingham	West Midlands	United Kingdom	B15 2TT
University of Warwick, Library, Gibbett Hill Road, COVENTRY, CV4 7AL, United Kingdom	University of Warwick	Library	Gibbett Hill Road		Coventry		United Kingdom	CV4 7AL

- Where you want to add to your data from an external data source:

Data you have	Date of Birth from VIAF (Virtual International Authority File)	Date of Death from VIAF (Virtual International Authority File)
Braddon, M. E. (Mary Elizabeth)	1835	1915
Rossetti, William Michael	1829	1919
Prest, Thomas Peckett	1810	1879

What Should I Know When Working With OpenRefine?

- No internet connection is needed, and none of the data or commands you enter in OpenRefine are sent to a remote server.
- You are NOT modifying original/raw data.
- Projects are autosaved every five minutes and when OpenRefine is properly shut down (Ctrl+C). See History in User Manual for details.
- Files are saved locally so if you are working on two computers, you will have to export/import files/projects.

Key Points

- ' A tool for working with messy data'
- Works best with data in a simple tabular format
- Can help you split data up into more granular parts
- Can help you match local data up to other data sets
- Can help you enhance a data set with data from other sources

> 💡 *OpenRefine is completely offline. Uploading your data to create a project is not the same as uploading it to the internet. Your data stays on your computer at all times.*

REDCap: Standardised Data Capture with Built-In Validation

REDCap (Research Electronic Data Capture) is a secure, browser-based platform specifically designed for clinical and health research data collection. It is free for institutions and is used by thousands of research institutions worldwide, including all RESPIRE partner universities. A team at Vanderbilt University, led by Paul A. Harris and colleagues, developed REDCap (Harris, P.A. et al.). Unlike general-purpose tools such as Excel or Google Forms, it is built with research data quality in mind. It allows Data Champions to design data collection instruments with built-in validation rules that catch errors at the point of entry, before they become embedded in the dataset.

For RESPIRE multi-site studies, this tool is particularly valuable because:

- Each site can enter data independently using the same standardised instrument, ensuring consistency across countries.
- Field-level validation rules (such as range checks, type checks, and mandatory field requirements) can be configured centrally and applied identically across all sites.
- User access can be restricted by role, meaning data entry staff see only the fields they need, and analysts see only the pseudonymised version of the data.
- Every change made to a record is automatically logged in an audit trail, recording what changed, who changed it, and when.

> 💡 *REDCap does not replace a good Data Management Plan. It enforces it. The validation rules you build into the tool are only as good as the data dictionary you agreed on before data collection began.*

📖 *See Chapter 10 for quick-start guides on tools that assist with harmonisation and cleaning, such as OpenRefine and REDCap.*

Chapter 5: Collaborative Learning and Capacity-Building

5.1 Why Collaborative Learning Matters

Research data management in multi-country studies involves different institutional systems, legal frameworks, infrastructures, languages and research practices. A single standardised approach may not address every local challenge. Data Champions therefore benefit from opportunities to exchange experiences, identify common problems and develop solutions that are both locally appropriate and consistent with shared research standards.

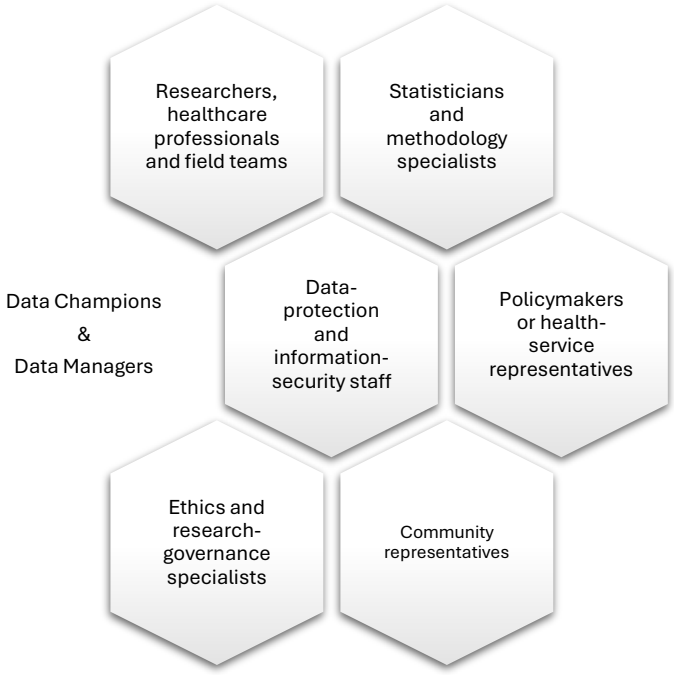
Collaborative learning can help research teams to:

- improve data quality and documentation;
- identify ethical, legal and governance concerns early;
- compare practices across studies and institutions;
- adapt tools and procedures to local contexts;
- reduce duplication of effort; and
- share practical lessons across the network.

This approach helps in decentralisation, institutional capacity-building, equity and sustainability.

5.2 Engaging Relevant Stakeholders

Data-management challenges often require input from people with different responsibilities and expertise. Depending on the issue, relevant stakeholders may include: ->



The people involved should reflect the nature of the problem and the context in which the research is conducted. Their roles, responsibilities and decision-making authority should be clear. Community and frontline perspectives should be incorporated meaningfully, particularly where decisions may affect participants, data collection or the interpretation and use of findings.

Data Champions can help identify relevant stakeholders, facilitate discussion and document agreed actions. They should seek specialist advice when an issue falls outside their expertise or authority.

5.3 From Collaborative Learning to Better Data Practice

Collaborative discussions should lead to clear and practical improvements. These may include:

Clear and practical improvements	Revising a Data Management Plan
	Agreeing on common variable definitions and formats
	Improving metadata and ReadMe documentation
	Developing or updating standard operating procedures
	Clarifying access and Data-sharing arrangements
	Identifying training requirements
	Recording solutions that can be adopted by other research teams

Decisions should be documented, assigned to responsible individuals and reviewed to determine whether they have improved practice. Lessons that may benefit other sites should be shared across the Data Champions Initiative.

5.4 Case Study: NIHR RESPIRE Methodologies Network Workshop

The NIHR Global Health Research Unit on Respiratory Health (RESPIRE) conducted a 1.5-day-long Methodologies Network Workshop on 12–13 January 2026 in Lucknow, India, co-hosted by King George's Medical University. The hybrid workshop brought together diverse experts, researchers, data experts, and partners from eight countries across Asia and the UK. The motto was to strengthen methodological capacity, promote open science, and improve data governance practices in low- and middle-income countries (LMICs). The participants represented ICMR, academia (University of Edinburgh, Universiti Malaya), NGOs (KEMHRC, PURE Foundation, MAHAN Trust), and research institutions (CMC Vellore, RESPIRE Collaboration).

The following are the key discussions highlighted during the workshop:

- involve those responsible for collecting, managing and using research data;

- recognise differences in local infrastructure, resources and institutional requirements;
- share challenges and practical solutions across countries and disciplines;
- strengthen consistency in documentation and data-quality practices; and
- adapt procedures to local contexts without compromising ethical, legal or data privacy requirements. (e.g., While doing an intervention on emerging technologies, including artificial intelligence, it can be done responsibly in community settings, with attention to local needs, community engagement, privacy, fairness, accessibility and human oversight)

The workshop also demonstrated how structured discussion can inform practical outputs. Lessons identified through collaborative activities can be used to improve Data Management Plans, training materials, standard operating procedures, templates and future editions of this handbook.

Data Champions can apply a similar approach within their institutions by bringing together relevant colleagues to examine a specific data-management problem, agree on workable actions and document the resulting learning. This supports the RESPIRE Data Champions Initiative by ensuring that guidance is developed collaboratively, implemented locally and shared across the network.

5.5 Practical Steps for Data Champions

When organising a collaborative learning activity:

1. **Identify the issue.** Define the data-management or methodological challenge clearly.
2. **Invite relevant participants.** Include those affected by the issue and those with appropriate expertise or decision-making authority.
3. **Review current practice.** Identify differences, risks, barriers and examples of effective practice.
4. **Agree on actions.** Specify what will change, who is responsible and when the action will be completed.
5. **Document decisions.** Update the DMP, SOPs, templates or other project records.
6. **Review progress.** Assess whether the agreed actions have addressed the problem.
7. **Share the learning.** Communicate useful lessons to other Data Champions and research teams.

Collaborative learning helps in expanding the learning curve of individual expertise into shared institutional knowledge. By documenting and exchanging practical solutions, Data Champions can strengthen local research practice, learn from one another, reduce reliance on a central team and share data-management knowledge more equitably across the network.

Note: This case study summarises key lessons from the methodologies workshop. The workshop framework, methodology and findings will be presented in a separate publication.

Chapter 6: Working with Different Data Types

Research data may include clinical measurements and health indicators, socio-demographic information such as age, gender, education and location, interview transcripts, audio recordings, images and datasets created during analysis. These data types require different management approaches because they vary in format, sensitivity, re-identification risk, storage needs and potential for reuse.

This chapter provides practical guidance for Data Champions and research teams on managing different types of data responsibly and in accordance with applicable privacy legislation across RESPIRE partner countries. Data Champions support teams by promoting good practice, advising or signposting on data management, and helping identify and address risks throughout the research lifecycle. The chapter is designed as a practical resource for research settings rather than a technical manual.

6.1 Data as research output

A research project may generate data in many formats. These can be broadly described as:

- **Structured data**, organised in predefined fields, such as databases, spreadsheets and survey datasets.
- **Unstructured data**, without a fixed tabular structure, such as transcripts, documents, photographs, audio and video.
- **Derived data**, created by cleaning, transforming, combining or analysing source data.
- **Process data**, such as code, scripts, workflows and documentation used to produce or analyse data.
- **Administrative and supporting research records**, such as ethics applications, consent forms, correspondence and technical reports.

These categories may overlap. For example, an interview transcript is usually qualitative and unstructured, while coded themes derived from it may be stored in a structured table. The following table presents common examples:

Category	**Types**	**Extensions/ file types**
Qualitative (Text)	• Questionnaires • Interview transcripts • Codebooks • Methodologies • Workflows • Standard operating procedures • Protocols • Field notebooks • Diaries	txt, pdf, word, html, xml
Quantitative (Numeric)	• Survey responses • Medical records	Data types may include:

	• Test responses • Spreadsheets • Instrument measurements • Geospatial information	Stata, SPSS, Excel, GIS
Multimedia	• Photographs • Audio recordings • Videos •	Data types may include: jpeg, png, tiff, mp3, wav, mpeg, quicktime
Development code	• Source code • Algorithms • Scripts	Data types may include: Python, Java, MATLAB
Syntax	Software-specific code files to carry out data processing steps (e.g. data preparation, linkage, statistical analysis, etc).	Data types may include: Stata, SPSS, R, MATLAB
Discipline specific	Data types may include: • Flexible Image Transport System (FITS) [Astronomy] • Crystallographic Information File (CIF) [Chemistry] • GRIdded Binary (GRIB) [Meteorology]	• .fits, .fit, .fts • .cif • .grib, .grb, .grib1, .grib2, .grb1, .grb2
Research records	Accompanying research records - that is, administrative materials and supporting documentation - both during and beyond the life of a project.	▪ Correspondence (electronic mail and paper-based correspondence) ▪ Grant applications ▪ Ethics applications ▪ Technical reports ▪ Technical appendices ▪ Research reports ▪ Research publications ▪ Signed consent forms ▪ Social media communications (blogs, wikis, tweets, etc.) ▪ Metadata document with the information about the datasets/ documents.

6.2 Understanding Varied Data Types

Before deciding how data should be stored, shared, or protected, it is important to understand what kind of data you are working with.

Some data are structured, such as survey responses entered into spreadsheets or databases. These are organised into rows and columns and are relatively easy to validate and analyse. Other data are unstructured, including interview transcripts, field notes, audio recordings, and images. These often contain richer detail but also carry higher privacy risks
In RESPIRE studies, common data types include:

- clinical and health records (e.g. CRFs, laboratory values),
- socio-economic and demographic data,
- qualitative data from interviews or focus groups,
- quantitative monitoring or outcome data, and
- derived or processed datasets created during analysis.

Each type requires slightly different handling, but all should follow the same core principles of accuracy, documentation, security, and ethical use. (Please refer to the overview section about the types of data)

6.3 Working with Qualitative Data

Qualitative data plays a critical role in understanding context, behaviour, and lived experience. However, it often contains direct or indirect identifiers, which require a much higher focus on the protection of privacy.

> ***Note on the use of AI tools:*** *Do not upload identifiable, sensitive, confidential or unpublished research data to an AI platform unless the tool has been institutionally approved and its use is permitted by the study protocol, ethics approval, participant consent and applicable law. Before using any AI tool, consider data minimisation, anonymisation, security, data retention, model training and cross-border transfer. Researchers remain responsible for verifying AI-generated outputs and documenting how the tool was used.* ***See Chapter 9 for guidance on the ethical use of AI in research.***

6.4 Managing Qualitative Materials

Qualitative materials may include interview transcripts, focus-group notes, audio or video recordings, translations and observational field notes. These materials often contain direct and indirect identifiers and should be organised and protected from the point of collection.

Teams should:

- transfer recordings promptly from recording devices to an approved, access-restricted storage location;
- delete local copies from personal or recording devices once secure transfer has been confirmed, where permitted by the study protocol;
- use the agreed file-naming and version-control system;
- pseudonymise transcripts and translations as early as practicable;
- limit access to authorised personnel; and
- protect qualitative materials according to their sensitivity and the risk of participant identification.

Privacy during transcription and translation

Transcription and translation may be completed by hand or typed directly using a mobile device, tablet or computer. Both methods present privacy risks. These activities should take place in a private location where conversations cannot be overheard and screens or handwritten notes cannot be viewed by unauthorised individuals. Headphones or earphones should be used when listening to recordings, and shared or public spaces should be avoided.

Only approved, encrypted and password-protected devices should be used. Handwritten transcripts and notes should be stored in locked facilities and transferred securely into the approved research system. Transcribers and translators should receive confidentiality training, sign appropriate agreements and have access only to the files required for their work. Any external or AI-assisted service must be institutionally approved and consistent with participant consent, ethics approval and applicable data-protection requirements.

Clear file naming, version control and secure storage help prevent accidental disclosure or loss. Data Champions should support teams in applying the same standards of care to qualitative materials as to other sensitive research data. (*See Chapter 3, Section 3.5.2 for guidance on creating metadata and ReadMe files, and Chapter 9 for guidance on the ethical use of AI.)*

6.4.1 Practical Steps for Pseudonymising Qualitative Data

Qualitative data can be particularly challenging to protect because contextual details may identify participants, even after names and other direct identifiers have been removed.

De-identification is a broad term for removing, replacing or obscuring identifying information. Pseudonymisation is a specific form of de-identification in which direct identifiers are replaced with codes or aliases, while a separate key allows authorised researchers to reconnect the data to participants. Pseudonymised data remain personal data because re-identification is still possible. Anonymisation aims to reduce

the risk of identification so that individuals are no longer reasonably identifiable, but achieving this may be difficult with context-rich qualitative material.

Use the following steps to pseudonymise qualitative data:

- **Replace direct identifiers:** Replace participants' names and other direct identifiers with consistent codes or aliases, such as [Participant 01]. Where aliases are used, avoid introducing characteristics that could misrepresent participants or affect interpretation.
- **Generalise proper nouns:** Replace identifying names of organisations, institutions, workplaces and small locations with broader descriptions, such as [local hospital], [community organisation] or [small town]. Retain specific details only when they are analytically necessary and present an acceptably low identification risk.
- **Remove or generalise sensitive details**: Review unusual occupations, rare conditions, distinctive events and group memberships that could identify a participant. Replace them with accurate broader categories, such as [specialist occupation] or [rare health condition], where doing so preserves the relevant meaning.
- **Categorise background information:** Replace precise demographic details with broader categories where appropriate. For example, record an exact age as an age range, such as [20–25 years], or generalise a precise location to a region.
- **Modify details cautiously:** Alter non-essential details only where necessary to reduce identification risk and where the change will not distort the analysis. Record any modification and avoid changing characteristics that are relevant to the research question or interpretation.
- **Check combinations of information:** Assess whether several indirect identifiers, such as age, occupation, location and a distinctive experience, could identify a participant when combined or linked with other reasonably available information.
- **Create a secure pseudonymisation log:** Record the codes, aliases and transformations used consistently across transcripts and related files. Maintain a separate master key linking participant codes to identities only where re-identification is necessary and authorised.
- **Separate and protect the key:** Store the master key separately from the pseudonymised data in an encrypted, access-restricted location. Limit access to authorised personnel and retain the key only for as long as required by the protocol and applicable policies.
- **Review the material:** Ask an authorised team member to check the pseudonymised files for missed identifiers, inconsistent replacements and excessive removal of analytically important context.

6.5 Sharing Qualitative Data

Qualitative data should not automatically be considered safe to share because direct identifiers have been removed. Before sharing, assess the risk arising from quotations, combinations of indirect identifiers, rare experiences and information available from other sources.

The appropriate sharing method should reflect the level of risk and may include:

- **openly sharing sufficiently anonymised summaries, coded extracts or thematic datasets;**
- **sharing carefully selected and reviewed quotations;**
- **providing pseudonymised transcripts through controlled or managed access;**
- **applying access agreements and restrictions on reuse or re-identification; or**
- **withholding material where the disclosure risk cannot be adequately reduced.**

Summaries and thematic extracts may reduce risk but can also remove important context. Full transcripts may sometimes be shared through controlled access where this is consistent with participant consent, ethics approval, applicable law and institutional requirements. The chosen approach and disclosure-risk assessment should be documented.

> 💡 *"Rule of thumb: Read through the cleaned transcript. If a participant reading it could still recognise themselves from the details provided, further redaction is needed."*

Full anonymisation of qualitative data is impossible
Since the data used in research based on qualitative methodologies often contain a large volume of more or less subtle identifying information. Full anonymisation then becomes almost impossible, since participants may remain recognisable even when deleting all directly identifying variables.

> ⚠️ ***Important: Sharing Qualitative Data:***
> *Data that are not fully anonymised must, for legal purposes, be considered personal data. Nonetheless, sharing this type of data is still possible under the right conditions, provided there are strict controls in place regarding participant information, consent, security, and access.*
>
> ***Do not share full qualitative transcripts via open public sharing.*** *If external sharing is required, only brief summaries or thematic extractions of the interviews should be shared, and only after stringent screening.*

6.6 Analysis and Software

Qualitative analysis software such as NVivo, ATLAS.ti, or MAXQDA allows researchers to code data systematically and maintain an audit trail of analytical decisions. These tools support transparency and reproducibility, especially when multiple analysts are involved.

Tool	Type	Cost
NVivo	Coding and analysis	Paid
ATLAS.ti	Coding and analysis	Paid
MAXQDA	Mixed methods analysis	Paid
Taguette	Coding	**Free & open source**
QCoder	Coding in R	**Free & open source**

***Open-source tip**: Prefer open-source tools where possible. Taguette is an excellent free alternative that allows you to share coded data without requiring paid software.*

Key Elements of Qualitative Data Quality

- **Transcription Accuracy Workflows:** Establishing routine protocols to spot-check transcripts against the original audio recordings. This catches manual transcription errors, misheard quotes, and missed contextual cues (like tone or pauses).
- **Codebook Standardisation:** Ensuring that qualitative coding isn't purely subjective. Champions can guide teams to develop clear, documented definitions for each code and test inter-coder reliability to ensure different researchers apply them consistently.
- **Version Control and Master Files:** Setting up foolproof systems to track which files are raw, which are pseudonymised, and which are currently being coded. This prevents accidental overwriting or the use of outdated files.
- **Audit Trails:** Encouraging the use of software (like NVivo or ATLAS.ti) to track analytical decisions and maintain reflexive memos, ensuring the research process remains transparent and reproducible.

6.7 Working with Quantitative Data

Quantitative data is numerical clinical readings, survey scores, and sensor measurements. It is structured and relatively straightforward to validate, but errors compound quickly if not caught early.

Key Practices

- Use the same units across all sites
- Use ISO date format (YYYYMMDD) everywhere

- Avoid manual entry where possible; use electronic tools with built-in validation
- Classify each variable clearly as categorical or continuous in the data dictionary

Numerical Data Handling:

Consistency in data management is essential, requiring uniform formats such as standardised decimal points and date formats across datasets. To minimise errors, it is crucial to avoid manual data entry whenever possible, instead relying on automated tools or directly importing data. Additionally, accurately identifying and classifying variables, such as distinguishing between categorical and continuous types, ensures proper data organisation and analysis.

Data can be handled in various ways. Spreadsheets and databases are two of the most versatile and widely used tools for managing research data.

a. Spreadsheets:

Spreadsheets are computer programs that organise data into rows and columns, with information stored in individual cells. Often likened to electronic ledgers, spreadsheets are highly versatile and widely used tools for managing data. Programs such as Microsoft Excel or open-source alternatives like OpenOffice provide an accessible way to store and analyse both qualitative and quantitative data. Their popularity stems from their ease of use and flexibility. However, for larger research projects or those involving specialised data types and vocabularies, spreadsheets can present limitations that may compromise data integrity and increase the risk of errors.

The benefits of spreadsheets include their intuitive interface, which is easy to learn and familiar to many users. They offer customizable features for organising and analysing data and integrate seamlessly with various applications and coding languages. Furthermore, data stored in spreadsheets can be easily converted into CSV or tabular formats, making it suitable for long-term storage and sharing.

Despite these advantages, spreadsheets also have notable drawbacks. They offer limited control over data integrity and formatting, and proprietary software may impose restrictions on the volume of data that can be stored. The most significant challenge is the potential for human error, such as overwriting cells or recording data in non-standardised structures. To mitigate these issues, it is essential to maintain a clean layout with unique column headers and one variable per column, use version control systems to track changes and prevent data loss, and take advantage of built-in validation features like drop-down menus and conditional formatting to standardise inputs.

Spreadsheet Best Practices:

- One variable per column; one observation per row
- Unique, clear column headers, no merged cells
- Use drop-down lists and data validation
- Save as **.csv** for long-term storage

- Version control: FileName_YYYYMMDD_v1.csv
- Keep a read-only original, never edit the source file

***Known pitfall**: Excel silently converts some identifiers into dates (e.g., gene name SEPT2 becomes September 2). Always inspect data after opening in Excel.*

b. Databases:

Databases organise information by linking external tables rather than storing data in individual cells, offering a more flexible structure for storing and evaluating data. They are powerful and widely used tools for managing data due to their dynamic capabilities for organising, querying, and sharing information. Both proprietary and open-source database software solutions are available, making them accessible and adaptable to various research needs. Additionally, data stored in spreadsheets can often be integrated into database systems for enhanced functionality.

The advantages of databases include their highly customizable controls for formats, vocabularies, and data types, ensuring data is entered and stored consistently. They enable robust quality control by enforcing specific formats or values for fields and restricting access to certain tables. Database Management Systems (DBMS), such as MySQL Workbench or Microsoft Access, simplify the design and implementation of databases without requiring extensive coding expertise. Databases also support complex relationships between data elements, sophisticated queries, and detailed customization through command-line controls. Many database systems offer customizable interfaces for data access and forms for entering new information.

However, databases have some drawbacks. Learning database languages like SQL can be time-consuming, and designing databases often requires extensive planning, potentially delaying the research process. Data stored in proprietary database software can sometimes be challenging to convert into other formats. To optimize database use, researchers should consider relational databases like MySQL or PostgreSQL for managing large datasets, implement access controls and encryption for sensitive information to enhance security, and schedule regular backups to prevent data loss.

For larger or more complex datasets:

Tool	Type	Best For
MySQL / PostgreSQL	Open source	Large datasets; technical users
Microsoft Access	Proprietary	Smaller databases; non-technical users
REDCap	Web-based research tool	Clinical data capture across sites
KoBoToolbox	Mobile-friendly field collection	Offline use; LMIC settings

6.7.1 Data Validation and Quality Control

Ensuring data accuracy and integrity is paramount in any research or business process. Implementing robust data validation and quality control measures can significantly enhance the reliability of your data.

a. Data Validation:

Automated Checks: Utilising automated validation techniques, such as scripts, algorithms, or software tools, allows for systematic checks on data inputs, outputs, and storage. These methods efficiently identify outliers, duplicates, or missing values, especially in large datasets, thereby reducing manual effort and minimising errors.

Cross-Verification: This process involves comparing data entries against original source documents or alternative data sources to confirm their accuracy. Cross-verification ensures consistency and reliability across datasets, helping to detect discrepancies that may have been introduced during data collection or entry.

b. Quality Control:

Standard Operating Procedures (SOPs): Developing comprehensive SOPs for data collection, entry, and processing establishes clear guidelines and expectations for personnel involved in data handling. Well-defined SOPs enhance employee efficiency, productivity, and safety by providing step-by-step instructions, thereby reducing the likelihood of errors and ensuring consistency in data management practices.

Audit Trails: Maintaining detailed audit trails involves recording all data modifications, including the nature of the change, the individual responsible, and the timestamp. This practice ensures traceability and accountability, facilitating the identification and investigation of any unauthorised or erroneous changes. Regular review of audit trails is crucial for maintaining data integrity and compliance with regulatory standards.

Error Resolution: Implementing a systematic approach to error resolution involves documenting identified errors, analysing their root causes, and applying corrective actions to prevent recurrence. Regular audits and inspections help in promptly identifying and rectifying data integrity issues, thereby maintaining the overall quality of the dataset.

By integrating these data validation and quality control strategies, organisations can significantly enhance the accuracy, reliability, and integrity of their data, leading to more informed decision-making and improved operational outcomes.

Ensuring Quality Across Diverse Data Types in REDCap

When your study collects diverse data types, such as dates, clinical measurements, and free text, the risk of data entry errors increases significantly. A Data Champion must ensure that the electronic data capture system restricts what users can type into a field.

REDCap provides two primary tools to maintain high data quality across different data types: Field-Level Validation and the Data Quality Module.

1. Field-Level Validation (Real-Time QC) Whenever you create a text box in REDCap, you should apply a validation rule to match the specific data type you expect. This ensures REDCap will flag an error immediately if the wrong type of data is entered.

- **Formatting restriction:** You can force a text box to only accept a specific format, such as an integer, a decimal number, an email address, or a specific date format (like YYYY-MM-DD). If a user types a letter into a number field, REDCap blocks it.
- **Min/Max Range checks:** For numeric data types, you should always set minimum and maximum limits. For example, if you are collecting heart rate data, you can set the minimum to 30 and the maximum to 200. Any entry outside this range triggers a warning before the data is saved.

2. Missing Data Codes Diverse datasets often feature missing data, but a blank cell does not tell you *why* the data is missing. REDCap allows you to set up standardized Missing Data Codes (such as 'UNK' for Unknown, or 'REF' for Refused). This ensures that missing numerical or text data is categorized correctly without breaking your validation rules.

3. The Data Quality Module (Retrospective QC) While field validation catches problems during data entry, the Data Quality Module helps you run checks across the entire dataset later.

- **Pre-defined rules:** REDCap includes built-in buttons to instantly find all blank values, all hidden fields that contain values, or all fields where the data does not match the validation rules.
- **Custom rules:** You can write custom logic to find discrepancies across different data types. For example, you can create a rule that flags an error if a participant's "Date of Birth" indicates they are a child, but their "Consent Type" is recorded as adult consent.
- **Real-time execution:** Custom data quality rules can be configured to run in real time, alerting the staff member the moment they save a form with conflicting data.

💡 *"Data Champion Tip: Never leave a text box unvalidated if you are expecting a number or a date. Enforcing data types at the point of entry cuts the time required for data cleaning by more than half."*

Statistical Analysis Methods

a. Preparation:

Clean and preprocess data (e.g., handle missing values, normalise distributions).

Ensure data meets the assumptions of statistical tests (e.g., normality, independence).

b. Analysis Techniques:

Descriptive Statistics: Summarise data using means, medians, and standard deviations.

Inferential Statistics: Apply hypothesis testing (e.g., t-tests, ANOVA) and regression analysis.

Advanced Techniques: Use machine learning models for predictive analytics.

c. Tools:

Software like R, Python (pandas, NumPy, SciPy), SPSS, or Stata for analysis.

Visualisation tools such as Tableau, Power BI, or ggplot2 for insights.

Approach	What It Does	When to Use
Descriptive statistics	Summarises data: means, medians, ranges	Understanding what your data looks like overall
Inferential statistics	Tests hypotheses: t-tests, ANOVA, regression	Comparing groups or testing relationships
Machine learning	Pattern recognition in large datasets	Complex predictive analysis

Recommended tools (open source): R, Python (pandas, NumPy, SciPy), JASP
Visualisation: ggplot2 (R), Matplotlib (Python), QGIS (maps)

Save and share your analysis scripts alongside the dataset. This makes your analysis reproducible by anyone, regardless of what software they have.

6.8 Privacy Legislation Across RESPIRE Partner Countries

The following table provides an overview of the principal data-protection and research-governance frameworks relevant to RESPIRE partner countries.

Country or region	Principal legislation and guidance
United Kingdom	UK General Data Protection Regulation and Data Protection Act 2018
European Union/EEA	General Data Protection Regulation and applicable national legislation
India	Digital Personal Data Protection Act 2023, associated rules and relevant ICMR ethical guidelines
Malaysia	Personal Data Protection Act 2010, as amended, and associated regulations
Bangladesh	Current data-protection, cybersecurity and research-ethics requirements
Sri Lanka	Personal Data Protection Act 2022, as amended, and associated regulations
Pakistan	Current privacy, electronic-crime and data-protection requirements

Indonesia	Personal Data Protection Law 2022 and associated implementing regulations
Bhutan	Current privacy, information, cybersecurity and research-ethics requirements

Important: *Legislation, implementing regulations and cross-border transfer requirements may change. Research teams should verify the current legal position and consult their institutional data-protection, legal, ethics or research-governance office before collecting, processing, sharing or transferring personal data.*

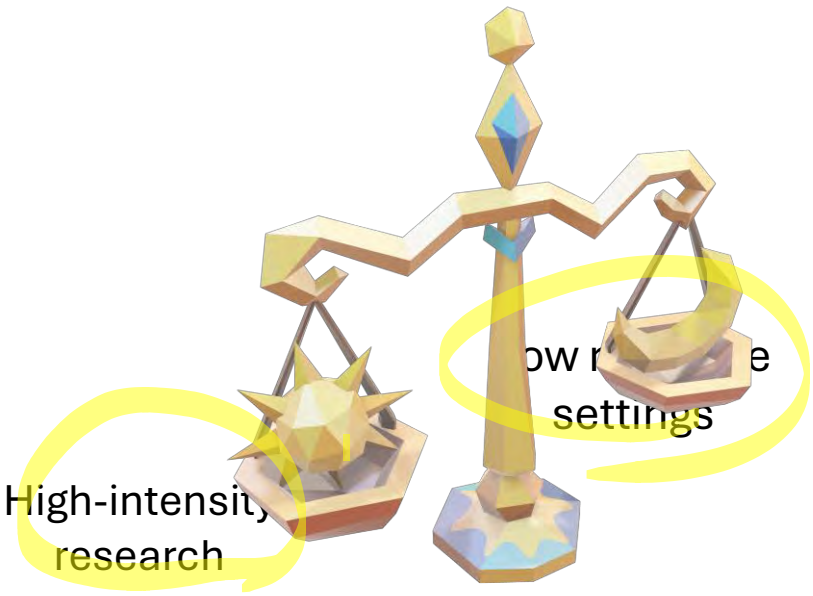
Cross-border data transfers

A cross-border data transfer occurs when personal data are sent to, stored in, remotely accessed from or otherwise made available in another country. Such transfers may be subject to the requirements of both the originating and receiving countries. Before transferring personal data, research teams should obtain appropriate advice and approvals, establish the necessary agreements, and apply safeguards such as data minimisation, pseudonymisation, encryption and access controls. The approved arrangements should be documented in the Data Management Plan.

Chapter 7: Data in Low-Resource and Multi-Site Settings

What this chapter covers

This chapter summarises practical lessons from RESPIRE studies from the LMICs, Platform III consultations and the Data Champions Initiative across partner institutions. It focuses on managing research data where connectivity, equipment, staffing, secure infrastructure or technical support may be limited.



The main lesson from RESPIRE is that low-resource settings require context-appropriate practices, not lower standards. It isn't always viable to follow the best practices. However, context-specific good practices on data privacy and security should be followed. Document the process/ decisions and apply proportionate safeguards for privacy, security and data quality.

7.1 Understanding the Local Context

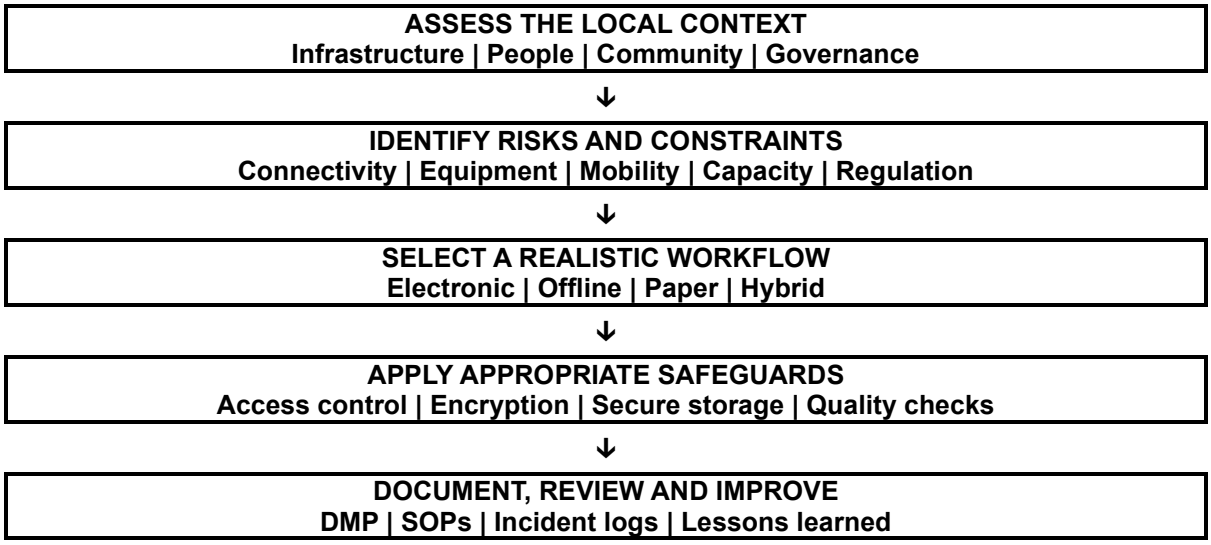
Research data-management guidance often assumes reliable electricity, continuous internet access, dedicated equipment, secure institutional storage and readily available technical support. These conditions may not be consistently available across all RESPIRE study sites.

Common challenges include:

- intermittent electricity and internet connectivity;
- shared, outdated or low-capacity equipment;
- limited access to secure servers or trusted research environments;
- reliance on paper or hybrid data-collection systems;
- limited budgets for software, storage and equipment maintenance;
- differences in digital skills and institutional support;
- mobile or difficult-to-follow study populations;
- multiple languages, literacy levels and cultural contexts; and
- differing or changing legal, ethical and regulatory requirements.

These conditions should be assessed before data collection and reflected in the Data Management Plan (DMP).

Context assessment pathway



7.2 Key Lessons from the RESPIRE Collaboration

RESPIRE experience demonstrates that data-management procedures should be designed around actual site conditions. A standardised approach is valuable, but it must allow justified local adaptation.

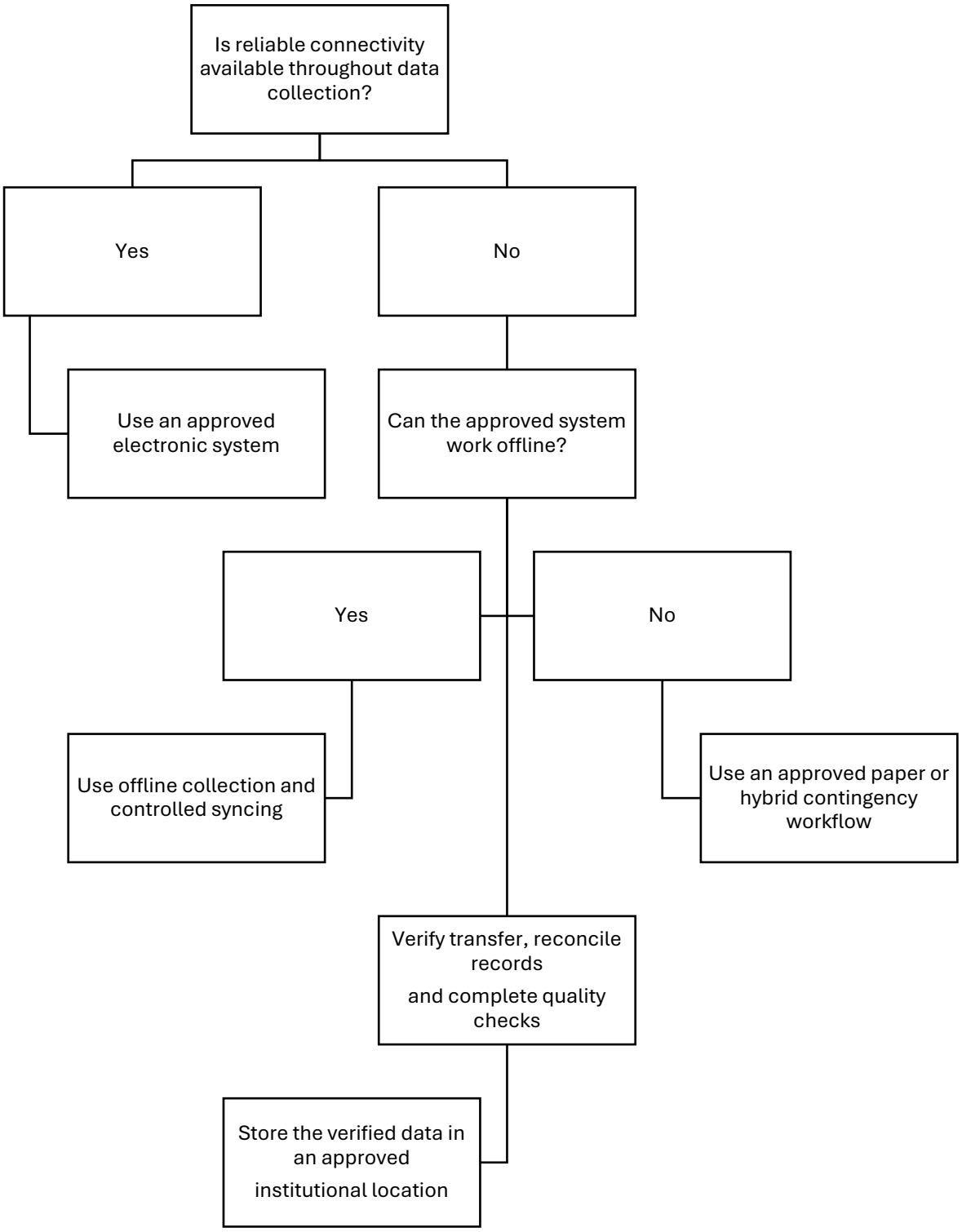
Challenge observed	Effect on research	Key learning	Plausible response
Unreliable or unavailable internet	Interrupted electronic data capture and delayed synchronisation	Electronic systems should not depend entirely on continuous connectivity	Use ensured offline-capable tools and maintain a documented paper contingency
Limited secure digital infrastructure	Difficulty storing, backing up or sharing active data securely	Available infrastructure should be assessed before data collection begins	Use approved institutional storage where available; consider encrypted local storage and scheduled secure transfers
Shared or outdated equipment	Increased risk of unauthorised access, data loss and software incompatibility	Device access and accountability require explicit controls	Use named accounts, role-based access, encryption, automatic locking and a device register
Paper-based data capture	Delayed validation, transcription errors and physical security risks	Paper can be appropriate when properly governed	Define secure transport, locked storage, data-entry verification and retention procedures by scanning

Challenge observed	Effect on research	Key learning	Plausible response
			(PDF) and securely storing them
Different systems across study sites	Inconsistent variables, formats and metadata	Harmonisation should begin before data collection	Agree on core variables, definitions, date formats, units and missing-value codes
Different legal and institutional requirements	Delays in approvals and cross-border sharing	One data-sharing model may not suit every country	Treat the DMP as a living document and seek current country-specific governance advice
Consent materials not covering later uses	Restrictions on archiving, sharing or emerging analyses	Future data use should be considered during study design	Explain foreseeable sharing and preservation plans clearly, while avoiding overly broad consent
Limited specialist capacity	Dependence on a small number of technical staff	Capacity should be embedded within participating institutions	Develop local Data Champions through mentoring, practical training and peer support
Staff turnover	Loss of procedural and system knowledge	Essential knowledge should not remain with one person	Maintain SOPs, handover records, training materials and named deputies
Participant mobility	Missed follow-up and incomplete longitudinal records	Follow-up plans should reflect participants' living conditions	Establish safe, approved and flexible communication and follow-up procedures
Geopolitical or travel disruption	Delayed training, supervision and collaboration	Research continuity should not depend entirely on travel	Use virtual support, local facilitators, shared documentation and distributed leadership

7.3 Selecting an Appropriate Data-Collection Approach

The data-collection method should be selected according to the research purpose, data sensitivity, local infrastructure, participant needs and institutional requirements. Electronic systems can improve validation and auditability, but they may be unsuitable if they depend on unreliable connectivity.

Data-collection decision pathway



7.4 Offline-First Data Management

An offline-first approach allows authorised research staff to collect and temporarily store data securely without a continuous internet connection. Data are synchronised or transferred once an approved connection becomes available.
An offline-first workflow should:

1. use an institutionally approved tool;
2. minimise identifiable information stored on field devices;
3. protect devices using encryption, authentication and automatic locking;
4. establish a schedule for secure transfer or synchronisation;
5. verify that all records have transferred successfully;
6. identify and resolve duplicated or conflicting records;
7. maintain an audit trail of transfers and corrections; and
8. remove temporary copies after verified transfer, where appropriate.

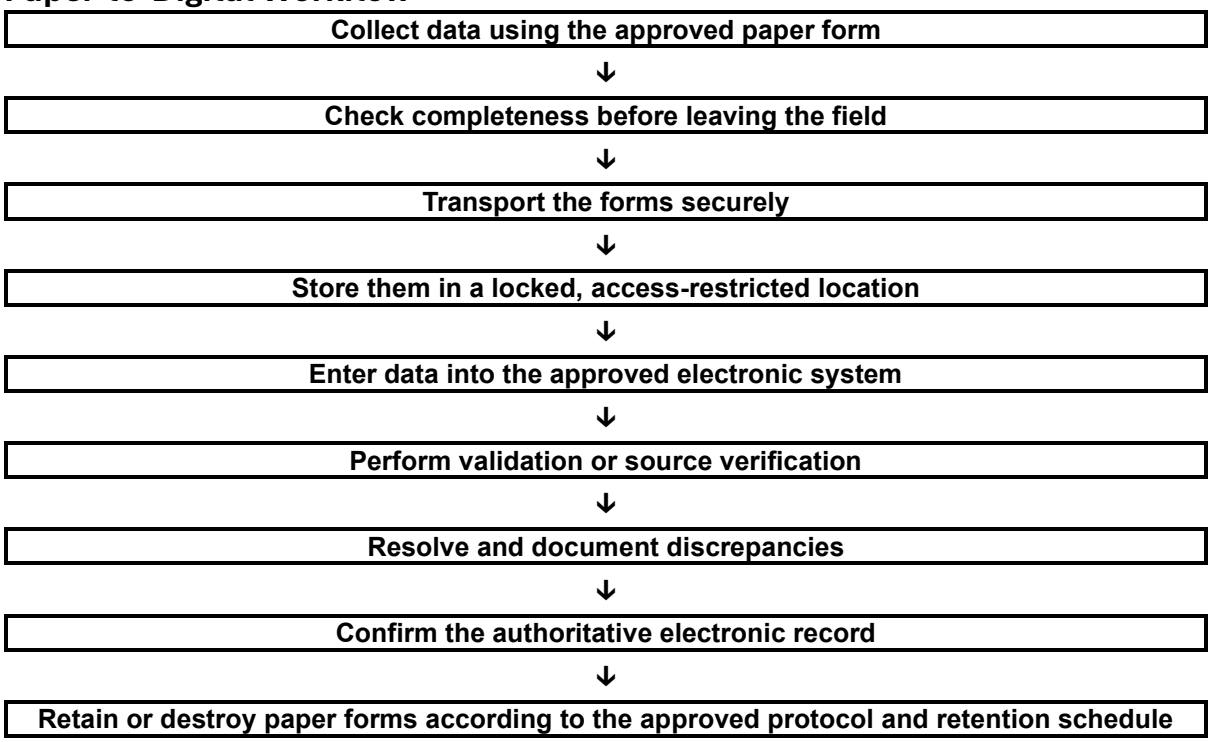
Examples of tools that may support offline collection include REDCap Mobile App, Survey Solutions, KoBoToolbox and ODK. Their use remains subject to local risk assessment, institutional approval and the study's governance arrangements.

7.5 Managing Hybrid Paper and Digital Records

Where a hybrid workflow is used, research teams should define:

- whether the paper form or electronic record is the authoritative source;
- how unique participant and form identifiers will be assigned;
- where completed forms will be stored before data entry;
- who will enter and verify the data;
- how discrepancies will be resolved;
- how corrections will be documented;
- when paper records will be archived or securely destroyed; and
- who may access each format.

Paper-to-Digital Workflow



Quality assurance may include double data entry, targeted source verification, logic checks or supervisory review, depending on the study's risks and available resources. Original records should not be altered without preserving an audit trail.

7.6 Case Study: Research with Refugees in Bazar, Bangladesh

A RESPIRE research intervention in Cox's Bazar, Bangladesh, involved participants living in temporary shelters in a refugee setting. Follow-up visits were difficult because some participants moved between shelters or relocated elsewhere. Limited or unavailable internet connectivity also disrupted REDCap data capture, requiring the research team to use paper-based forms.

The experience demonstrated that preferred digital practices may not always be viable. The team therefore had to adapt its workflow while maintaining privacy, security and data quality.

Participant movement created risks of missed follow-up, duplicate records and accidental disclosure while attempting to locate individuals. Paper-based collection introduced further risks, including physical loss, unauthorised access, delayed entry and transcription errors.

The key learnings were:

- anticipate mobility when designing follow-up procedures;
- collect only the minimum contact information needed;
- avoid tracing practices that could reveal participation or sensitive information;
- prepare approved offline or paper alternatives before fieldwork begins;
- transport and store paper records securely;
- assign responsibility for data entry, verification and query resolution;
- document procedural adaptations and seek further approval where required;
- prioritise essential activities when staff, equipment and time are limited; and
- maintain dialogue with participants, community leaders and other relevant stakeholders.

This case demonstrates that resource limitations do not remove the responsibility to protect research data. When ideal practices are not feasible, context-specific good practice should be adopted, justified and documented.

7.7 Community Engagement, Equity and Trust

Community engagement should continue throughout the study, particularly when research involves refugees, displaced populations, Indigenous Peoples or other communities experiencing marginalisation or collective vulnerability.
Research teams should explain clearly:

- why the research is being conducted;
- what information will be collected;
- how data will be protected and used;

- who may access or receive the data;
- what benefits can realistically be expected;
- how concerns or complaints can be raised; and
- what withdrawal means in practice.

Community leaders and trusted stakeholders can help researchers understand mobility, communication preferences and barriers to participation. However, community-level support does not replace individual informed consent. Teams should also consider whether established leadership structures adequately represent women, young people, minority groups and people with disabilities.
Building trust requires regular communication, realistic commitments and appropriate feedback to participating communities. Research should be conducted with communities rather than only about them.

Applying FAIR and CARE

FAIR supports data that are Findable, Accessible, Interoperable and Reusable. The CARE Principles for Indigenous Data Governance add attention to Collective Benefit, Authority to Control, Responsibility and Ethics.
The CARE Principles were developed specifically for Indigenous data governance. They may also provide useful questions when considering data about other communities experiencing collective vulnerability, but they do not replace applicable law, ethics requirements or humanitarian and community-engagement standards.

Research teams should ask:
- Who benefits from collecting, sharing or reusing the data?
- Who should participate in decisions about access and use?
- Could future reuse cause stigma, discrimination or surveillance?
- How will the project strengthen local capacity and stewardship?
- How will findings and benefits be returned to the community?

7.8 Data Security, Backup and Continuity

Shared devices, removable storage and interrupted connectivity increase the risk of loss or unauthorised access. Where equipment is shared, teams should use named accounts where possible, avoid shared passwords, apply role-based access, enable automatic locking and remove access when staff leave or change roles.
Backup arrangements should identify:
- what will be backed up and how often;
- where each copy will be stored;
- who is responsible for completing and checking backups;
- whether copies are encrypted, versioned and sufficiently independent; and
- how restoration will be tested.

Synchronisation is not the same as backup. Deletion, corruption or ransomware may be copied to a synchronised location. Where approved external drives are used, they should be encrypted, disconnected after backup and stored separately from the working device.

Tools such as FreeFileSync may assist with offline file copying where institutionally approved, but they do not replace a separate, versioned and recoverable backup strategy. When using an ***external hard drive must be encrypted with VeraCrypt*** *(see Chapter 8 and 10). An unencrypted backup is just as risky as an unencrypted original*.

7.9 Harmonisation and Data Quality Across Sites

Multi-site studies may use different languages, instruments, software, units and terminology. Teams should agree on common standards before collection wherever possible, including:

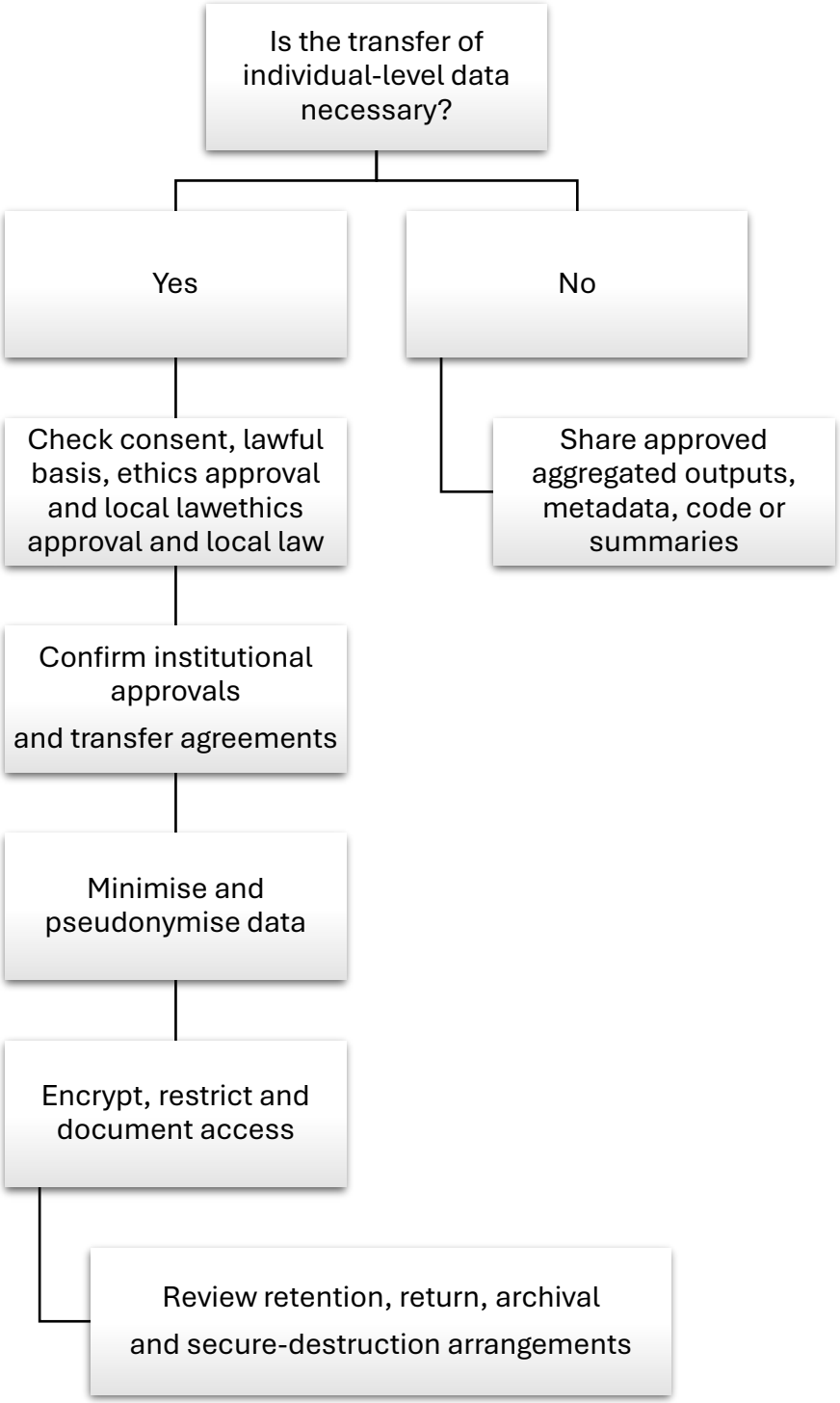
- participant, site and visit identifiers;
- variable names, definitions and permitted values;
- date, time and missing-value formats;
- measurement units and conversion rules;
- form and dataset version numbers;
- metadata and data-dictionary requirements; and
- quality-control and correction procedures.

Local differences should not automatically be removed, as some may be analytically important. Instead, they should be documented through site-specific variables, coding and metadata.
Quality checks should begin during collection. Forms should be piloted under actual field conditions, and teams should review missing values, inconsistent records and unusual observations regularly. Corrections must be documented rather than applied silently. *(See chapter 4 pm Data harmonisation and Data Quality Control)*

7.10 Cross-Border Data Use

Data should not be transferred between countries simply because a technical connection is available. Before sharing individual-level data, teams should determine whether transfer is necessary and confirm that the proposed arrangement is consistent with consent materials, ethics approval, applicable law and institutional policy.

Cross-Border Decision Pathway



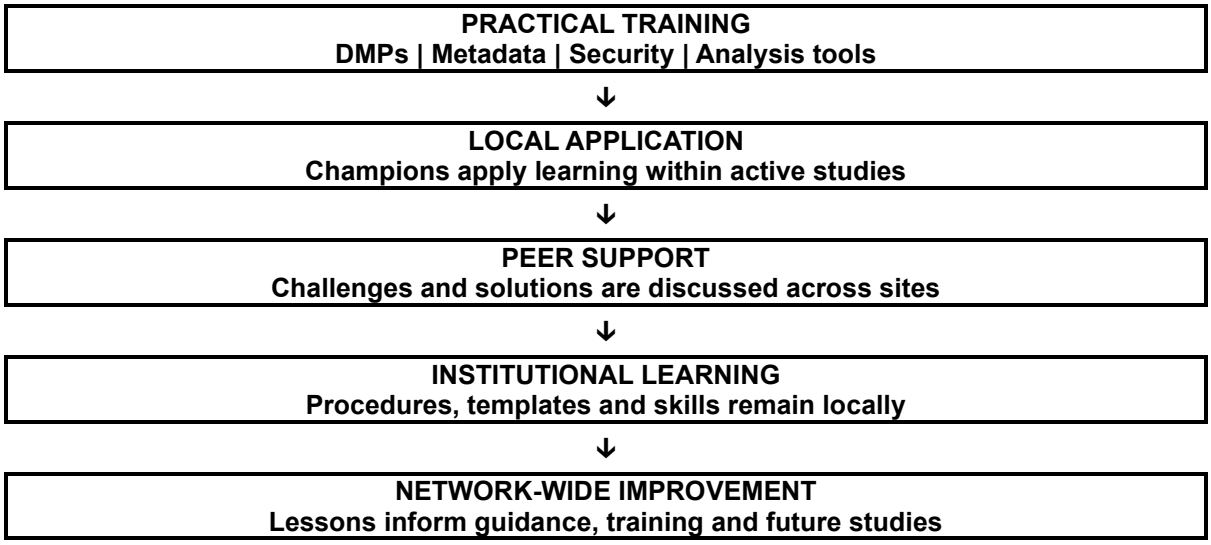
7.11 Building Local and Network Capacity

Technology alone cannot resolve infrastructure constraints. Sustainable data management also requires trained staff, clear responsibilities, accessible support and continuing institutional commitment.

Within RESPIRE, Data Champions help connect local experience with network-wide learning. They can support DMP development, identify training needs, promote

documentation and refer complex issues to ethics, legal, information-security or data-protection specialists.

RESPIRE Capacity-Building Model



The Data Champions Initiative demonstrates that one-off training is rarely sufficient. Practical application, mentoring, refresher training and peer learning help turn individual knowledge into institutional capacity. Collaboration across the initiative can provide technical support, reduce duplication and enable effective solutions from one site to be adapted elsewhere.

7.12 Key Actions for Data Champions

Data Champions should help research teams to:

1. assess infrastructure and community conditions before selecting tools;
2. ensure that DMPs reflect actual site practices;
3. prepare offline, paper-based and continuity arrangements;
4. protect paper, digital and removable records;
5. establish recoverable and tested backups;
6. harmonise essential variables and documentation across sites;
7. record meaningful local adaptations and quality issues;
8. engage communities and relevant stakeholders throughout the study;
9. review governance arrangements when regulations or intended uses change;
10. minimise unnecessary cross-border transfers;
11. seek specialist support when an issue exceeds their role; and
12. share useful lessons across institutions and countries.

Low-resource settings require context-appropriate practices, not lower standards. When ideal systems are not feasible, research teams should select the safest realistic alternative, document their decisions and maintain proportionate safeguards. Effective planning, local capacity, community trust and collaboration across the RESPIRE network are as important as the technology used.

Chapter 8: Open Science and Data Sharing

What this chapter covers: *Why openness and security belong in the same chapter, and how to share data responsibly while keeping it safe.*

Why Openness and Security Belong Together

It might seem strange to put open science and data security in the same chapter. They can feel like opposites: one pushing toward sharing everything, the other toward locking everything down.

But they are not opposites. They are two sides of the same responsibility.

The principle that guides everything in this chapter is: We share data openly because science works better when knowledge is accessible, reproducible, and reusable. We protect data carefully because the people whose information it contains have the right to privacy and safety.

Security is not the enemy of open science. It is what makes open science trustworthy.

8.1 Open Science

Open Science is the movement to make scientific research, its methods, data, and outputs, freely accessible to anyone.

It includes:

- Publishing in open-access journals
- Sharing data in public repositories
- Sharing analysis code and scripts
- Pre-registering study designs
- Using and contributing to open-source tools

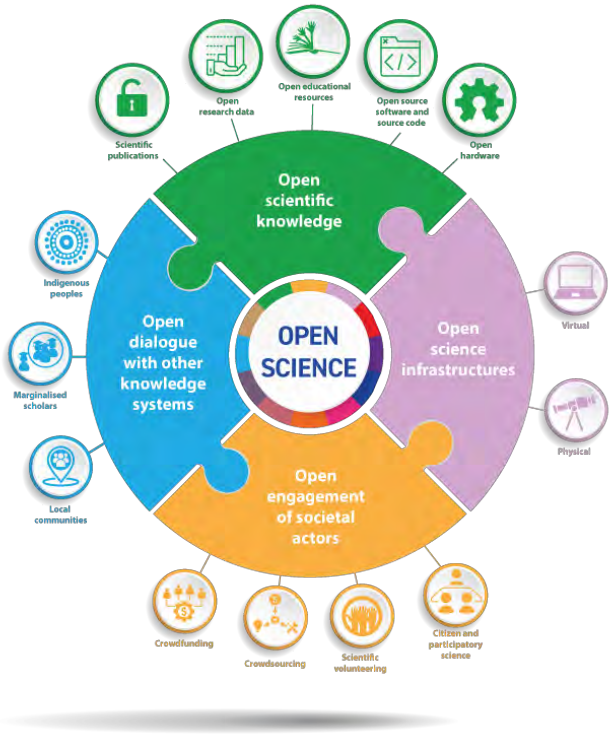


Figure 3 This image by author:UNESCO [7619] is licensed under CC BY-SA 3.0 IGO [12996]

Why it matters for RESPIRE: Research funded by public money should serve the public good. Knowledge generated in partnership with communities in LMICs should be accessible to those communities, not locked behind paywalls or in private institutional servers.

8.2 FAIR Data Principles

FAIR stands for Findable, Accessible, Interoperable, and Reusable. These four principles describe what well-shared research data looks like.

Principle	What It Means	In Practice
Findable	Others can discover your data	Archived in a repository; rich metadata; DOI assigned
Accessible	Access conditions are clear	Even restricted data states clearly how to request access
Interoperable	Data uses standard, open formats	.csv not .xlsx; machine-readable metadata
Reusable	Others can understand and use it	Complete documentation; CC BY 4.0 licence; clear methods

FAIR does not mean open. Some data must remain restricted for privacy reasons, but it can still be FAIR if access conditions are clearly documented.

8.3 Open Licences: Giving Others Permission

Without a licence, no one can legally reuse your dataset, even if it is publicly available online. Copyright applies automatically to research outputs, meaning that without explicit permission, others cannot use, adapt, or build on your work. An open licence solves this problem by giving everyone clear, upfront, legal permission to reuse your work under defined conditions, without needing to contact you individually each time.

For research data, open licences matter because science advances when findings can be verified, replicated, and built upon. A dataset sitting in a repository without a licence is legally unusable by other researchers, no matter how findable or well-documented it is. Licensing is therefore not a bureaucratic formality; it is what transforms a deposited dataset into a genuinely reusable scientific resource and a core requirement of FAIR compliance.

Understanding Creative Commons Licences

Creative Commons (CC) licences are the most widely used open licences for research data and publications. They are built from four conditions that can be combined in different ways:

Condition	Symbol	What it Means
Attribution	BY	Users must credit the original creators
ShareAlike	SA	Adaptations must be shared under the same licence
NonCommercial	NC	Use is restricted to non-commercial purposes only
NoDerivatives	ND	Users may not adapt or build on the work

These conditions combine to produce six main licence types, ranging from most to least open:

Licence	Permissions	Restrictions
CC BY 4.0	Use, adapt, share, including commercially	Credit original creators
CC BY-SA 4.0	Use, adapt, share, including commercially	Credit creators; share adaptations under same licence
CC BY-NC 4.0	Use, adapt, share	Credit creators; no commercial use
CC BY-NC-SA 4.0	Use, adapt, share	Credit creators; no commercial use; share adaptations under same licence
CC BY-ND 4.0	Use and share	Credit creators; no adaptations permitted
CC BY-NC-ND 4.0	Share only	Credit creators; no commercial use; no adaptations
CC0 (Public Domain)	Complete freedom	No restrictions whatsoever

Which Licence Should RESPIRE Use?

CC BY 4.0 is the recommended licence for all RESPIRE datasets and publications. It is the most open licence that still requires credit to be given to the original creators. It is recommended because:

- It meets NIHR and UKRI open access requirements
- It is fully compatible with FAIR principles
- It allows the widest possible reuse, including by researchers in LMICs who may be working in commercial or applied settings
- It ensures that the original RESPIRE team receives credit when data is reused or cited
- It is the licence under which this handbook itself is published, modelling the open science ethos the handbook promotes

CC BY 4.0 allows anyone to use, adapt, and share RESPIRE data for any purpose, including commercially, as long as they give appropriate credit to the original creators, state whether any changes were made, and provide a link to the licence.

When Might a More Restrictive Licence Be Appropriate?

In most cases, CC BY 4.0 is the right choice. However, there are situations where a more restrictive licence may be considered:

Situation	Possible Licence	Reason
Data contains sensitive contextual information	CC BY-NC 4.0	Limits commercial exploitation of sensitive community data
Funder requires ShareAlike terms	CC BY-SA 4.0	Ensures adaptations remain open
Data cannot be adapted without risk of misrepresentation	CC BY-ND 4.0	Prevents misleading modifications

Always check your funder requirements and institutional policy before choosing a licence other than CC BY 4.0. When in doubt, consult your institutional data protection or legal office.

How to Apply a Licence to Your Dataset

Applying a licence is straightforward and takes only a few minutes:

1. Choose your licence at creativecommons.org; the licence chooser tool guides you through the options
2. Copy the licence statement generated by the tool
3. Include the licence statement in your ReadMe file alongside the dataset
4. State the licence clearly in the metadata record when depositing to a repository such as Edinburgh DataShare or Zenodo
5. Include a data availability statement in any related publication that names the licence and provides the repository link

A Practical Example

When depositing a RESPIRE dataset to Edinburgh DataShare, the licence statement in the ReadMe file should read:

"This dataset is published under a Creative Commons Attribution 4.0 International licence (CC BY 4.0). You are free to use, adapt, and share this data for any purpose, provided appropriate credit is given to the original creators. Licence details: https://creativecommons.org/licenses/by/4.0/"

This single statement is all that is needed to make the dataset legally reusable by anyone in the world.

Remember: A dataset without a licence is a dataset no one can legally use. Licensing takes five minutes and unlocks the full potential of your research.

8.4 The Open Science Monitoring Initiative

Across the research community, a growing infrastructure now exists to measure progress toward open science, not as an abstract ideal, but as a set of trackable, reportable practices.

Why monitoring matters: Funders commit billions of pounds to research on the basis that findings will be available to the public, verifiable by other scientists, and reusable by future researchers. Monitoring holds the research community accountable to those commitments.

Key monitoring frameworks relevant to RESPIRE:

The European Commission's Open Science Monitor
The EC Open Science Monitor (https://monitor.openaire.eu/dashboard/ec/open-science) tracks:
Open access to publications: measuring the proportion of research papers freely available online
Open research data: assessing how much data is shared, how findable it is, and whether it meets FAIR criteria
Open collaboration: measuring the breadth of partnerships and citizen science engagement

The Open Science Observatory (https://osobservatory.openaire.eu/home) tracks the implementation of open science practices across Europe and globally. Developed by OpenAIRE, it monitors:

- **Open access to publications:** measuring the proportion of research papers freely available online across different countries and institutions.
- **Open research data and software:** assessing how much data is shared and whether it is hosted in trustworthy repositories.
- **FAIR data compliance:** measuring how well repositories and datasets align with the Findable, Accessible, Interoperable, and Reusable principles.

Rather than waiting for manual reports, the Observatory provides interactive digital dashboards and country-level analysis, making progress (and lack of progress) instantly visible to the global research community.

UNESCO's Open Science Recommendation

In November 2021, UNESCO's 193 member states adopted the **Recommendation on Open Science,** the first international normative framework on open science. It establishes:
A shared definition of open science and its core values
Commitments for member states to develop national open science policies
Monitoring obligations to track implementation
For RESPIRE partner countries (all UNESCO members) this recommendation provides both a framework and an accountability mechanism for national open science progress.

NIHR and UKRI Open Science Requirements
As a NIHR-funded initiative, RESPIRE is subject to specific open science requirements:

Requirement	What RESPIRE Must Do
Open Access Publishing	Research papers must be published open access (gold or green route)
Data Sharing Statement	Every published paper must include a statement on data availability
Data Deposit	Datasets underpinning publications should be deposited in an appropriate repository
Reporting	Open science compliance is reported as part of grant monitoring

How RESPIRE Contributes to Open Science Monitoring
Every action a Data Champion takes toward responsible data sharing generates evidence for open science monitoring:

Data Champion Action	Open Science Monitoring Contribution
Deposit dataset to DataShare	Open research data metric
Assign DOI and CC BY 4.0	FAIR compliance; reuse metric
Complete metadata schema	Findability and interoperability metric
Include data availability statement in publication	Open access to data metric
Publish paper open access	Open access to publications metric
Share analysis scripts	Open methods and reproducibility metric

Open Science monitoring is not just about compliance. It is about demonstrating to funders, policymakers, and the public that research funded in their name is genuinely serving the public good.

8.5 Information Security and Cybersecurity

Before you can safely share data, you must ensure that your local environment is secure. In this digital era, passwords and secure hygiene act as the front door to your research environment. Reusing passwords across systems is one of the most dangerous security mistakes in research.

8.5.1 Use unique, strong Passwords

Image by Creative Canvas from Pixabay Source: https://pixabay.com/vectors/laptop-computer-keyboard-typing-10164292

In this digital era of devices, email, and cloud services, passwords are required for each of them; it's like a key to enter your home. In research, your team needs to create and use passwords for every system and device. Reusing passwords across systems is one of the most dangerous security mistakes in research

Use unique passwords for each service you use: This means that if someone gains access to one of your passwords, they don't gain access to your other services

Avoid the most common passwords: Avoid using common passwords which anyone can easily guess, like welcome, 123456, 123456789, "qwerty", "password" and 1111111, **or** from your favourite sports team, or by using family and pet names. Most of these details can be found within your social media profile.

Use strong passwords for each service you use

Strong passwords are the first line of defence for your accounts and the security of the encrypted folders/ drives. They help protect your email and your data. The takeover of an account by a cybercriminal can put your email, data and reputation at risk. Strong passwords protect you and everyone else.

Long passwords are strong passwords. Most services, including University services, have a minimum password length, but you should always use longer passwords than

the minimum, and even longer passwords than normal for elevated privileges or password managers. In
general, there should be no maximum length of a password, so make them as long as you can. If you use a password manager, you can choose the length of the random password it generates. A long randomly generated password is the most secure but may also be hard to type manually; many password managers can type the password for you.

What Makes a Strong Password?

You can also generate "strong enough" passwords by choosing three or more random unrelated words of four or more letters and altering as necessary to fit with the password rules of the service or add extra complexity. See the UK National Cyber Security Centre advice and remember that using longer random unrelated words or more words will make longer and stronger passwords.

Three random words

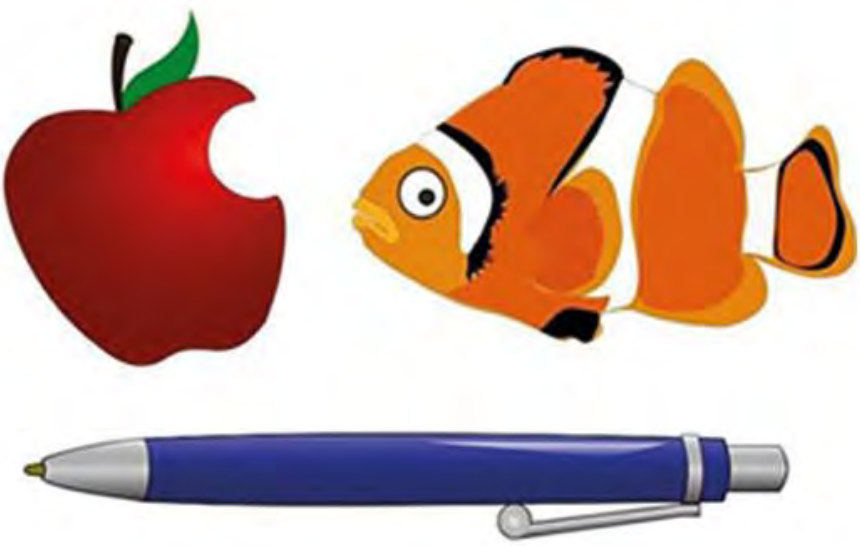


Figure 4 National Cyber Security Centre (NCSC). Three Random Words – Top Tips for Staying Secure Online. Retrieved from https://www.ncsc.gov.uk/collection/top-tips-for-staying-secure-online/three-random-words.

For example, 'tide-purple-anchor-forest ' is long, strong, and more memorable and could be written as

T!de-purp|e-4nch0r-f0re5+

The password should be :

- At least 16 characters (20 characters or more recommended)
- Mix of uppercase, lowercase, numbers, and symbols
- Not based on personal information
- Three random words
- Unique to each account

Use Multi-Factor Authentication whenever available: Many services have forms of Multi-Factor Authentication (MFA) available (for instance, Microsoft Authenticator); this is often also known as "Two-Factor Authentication (2FA)" or "Two-Step Verification". This uses something you physically have, usually a phone, as well as the password to provide extra security.

Use a Password Manager

It helps users maintain a large number of passwords and account information. They store login information for various accounts and automatically enter it into forms. This helps prevent cyberattacks and reduces the need to remember many passwords.

Password managers enable the use of strong and unique passwords and provide an efficient way to manage them. They encrypt the login information and store it in either the local memory of the user's system or in cloud storage.

KeePass is a free, open-source password manager. It stores all passwords in one strongly encrypted database, protected by one master password.

📖 *See Chapter 10 for a complete, step-by-step visual guide on exactly how to use KeePass to securely keep your passwords.*

8.6 Protecting Data in Storage and Transit

Once your environment is secure, you must protect the data itself. Encryption scrambles data so it is unreadable without the correct key. If an encrypted device is lost or stolen, the data inside remains completely protected.

Encryption scrambles data so it is unreadable without the correct key. If an encrypted device is lost or stolen, the data inside remains protected.

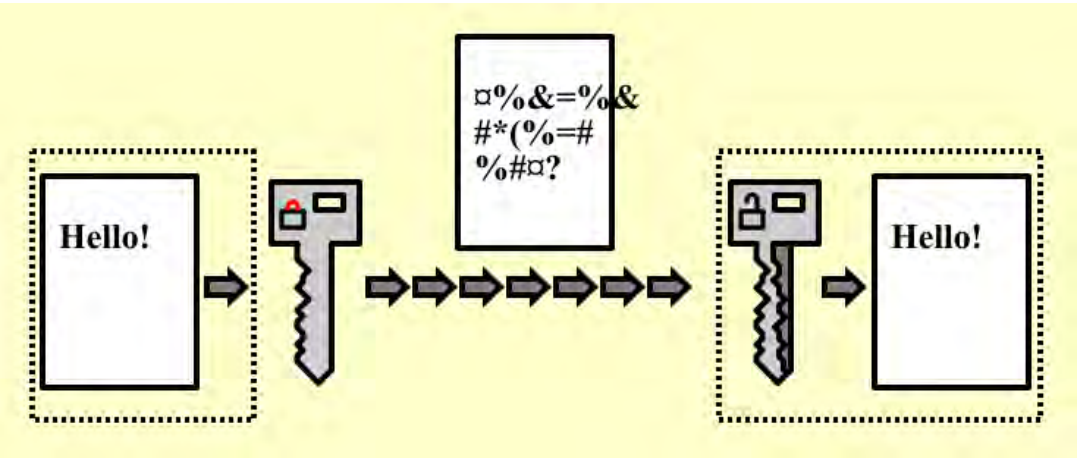


Image: Public key encryption keys.svg, Wikimedia Commons (CC BY 4.0)

Use encryption for:

- Laptops and computers storing research data
- External hard drives and USB drives
- Files containing sensitive data shared electronically

VeraCrypt Encryption tool : It is free, open-source, and available for Windows, macOS, and Linux. It is the RESPIRE-recommended encryption tool.

What it does: Creates an encrypted container, a file that acts like a locked safe. You mount it with your password; it appears as a virtual drive. When finished, you dismount it, and the data is locked again.

⚠ *Critical: If you forget your VeraCrypt password, the data is permanently unrecoverable. Store all VeraCrypt passwords in KeePass.*

🕮 *"See Chapter 10 for a complete, step-by-step visual guide on exactly how to install and use VeraCrypt and KeePass to secure your data."*

8.7 Sharing Data : Internal and Public

Internal Sharing

Apply minimum necessary access: share only with those who need it, only the data they require, only for as long as necessary.

Service	Appropriate For
DataSync	Sharing working files with internal and external collaborators
DataStore	Storing active, backed-up project data
SharePoint / Teams	Document collaboration and project management

Public Sharing: Repositories

Repository	What It Is	Best For
Edinburgh DataShare	UoE public repository with DOI	RESPIRE studies
Edinburgh **DataVault**	Long-term archive	Preserving completed data
Zenodo	Open-access repository; free; DOI	Any research, including code
OpenAIRE	European open science platform	EU-funded or international research

Before sharing any dataset publicly:

- Fully anonymised, that means re-identification not possible
- ReadMe file and metadata complete
- Licensed under CC BY 4.0
- DOI assigned
- FAIR-compliant
- Ethics board approval confirmed

8.8 Cybersecurity for Researchers

The CIA Triad: The Foundation of Security Thinking

Principle	Meaning	Example
Confidentiality	Only authorised people can access the data	Participant data accessible only to the research team
Integrity	Data has not been tampered with	No change made without an audit trail
Availability	Authorised users can access data when needed	Backups ensure data is not lost

The Most Common Threats

Phishing: fraudulent emails designed to steal passwords or install malware.
Warning signs:

- Unexpected urgency ("Your account closes in 24 hours!")
- Suspicious sender address (e.g., security@un1versity.com)
- Requests for passwords or credentials
- Links that do not match the text when you hover over them

Golden rule: *If you are not certain an email is legitimate, do not click any links. Go directly to the website by typing the address yourself.*

Ransomware: malware that encrypts your files and demands payment.

Best defences:

- Reliable, offline backups (FreeFileSync + 3-2-1 rule)
- Never clicking suspicious attachments
- Keeping software updated

Social Engineering: manipulating people rather than systems into giving up access.

Think of it this way: *Rather than breaking down a locked door, a social engineer asks you to hold the door open and gives you a convincing reason to do so.*

Never share your password with anyone, including IT staff. Legitimate IT staff do not need your password.

Practical Defences

Practice	Why It Matters
Multi-Factor Authentication (MFA)	Even a stolen password cannot be used without the second factor. Enable everywhere.
VPN on untrusted networks	Encrypts your internet traffic on public Wi-Fi
Keep software updated	Security patches close the vulnerabilities attackers exploit
Automatic screen lock	Activates after 5–10 minutes of inactivity
Full-disk encryption	BitLocker (Windows) or FileVault (macOS)
Never use public USB chargers	Malware can be transmitted through USB cables at public charging points

If Something Goes Wrong

1. Do not panic, but act quickly
2. Disconnect from the network, unplug Ethernet, or turn off Wi-Fi
3. Do not shut down the computer; forensic evidence may be lost
4. Contact your institutional IT security team immediately
5. Do not try to fix it yourself
6. Inform your Data Champion lead and PI
7. Document everything: what you saw, what you clicked, when

Under GDPR: *a breach involving personal data must be reported to your data protection officer within 72 hours. Do not wait.*

API-Based Data Sharing

APIs allow software systems to exchange data automatically, for example, REDCap sending data to a central repository each night.
If your study uses API-based transfer:

- All connections must use encrypted protocols (HTTPS, not HTTP)
- All transfers must be logged and auditable
- API access must be authenticated
- Institutional approval is required before implementing

8.9 Ethical Hacking: A Brief Awareness Note

Ethical hacking (or penetration testing) is the practice of deliberately attempting to find security vulnerabilities in a system with the owner's full written permission, so they can be fixed before a criminal exploits them.
As a Data Champion, you are not expected to conduct penetration tests. But understanding the concept helps you:

- Ask better questions when your institution claims systems are "secure" (Have they been tested? When? By whom?)
- Think like an attacker: where are the weak points in our data systems?
- Appreciate that human behaviour (clicking phishing links, reusing passwords) is usually the easiest target

***Important**: Using hacking tools without authorisation on systems you do not own is illegal in the UK under the Computer Misuse Act 1990, and under equivalent legislation in all RESPIRE partner countries.*

Chapter 9: Emerging Technologies (AI/ML), Open Source Software & Hardware

***What this chapter covers**: Open science hardware and Artificial Intelligence. how to use it ethically, and how to disclose its use honestly. This handbook is itself an example of transparent AI practice.*

9.1 What is Artificial Intelligence (AI)

Artificial Intelligence (AI) is the broad field of computer science concerned with building systems that can perform tasks that would normally require human intelligence, such as recognising images, understanding language, and making recommendations.
AI is not magic. It is not conscious. At its core, it is a set of mathematical techniques that allow computers to find patterns in large amounts of data.

Artificial Intelligence
└── Machine Learning
  └── Deep Learning
    └── Neural Networks
      └── Large Language Models (LLMs)

9.2 What Is Machine Learning?

Machine Learning (ML) is how computers learn from data, without being explicitly programmed for every situation.

Approach	How It Works
Traditional programming	A human writes explicit rules for every situation
Machine learning	The computer analyses thousands of examples and learns the patterns itself

***Think of it this way**: Traditional programming is giving someone a recipe. Machine learning is letting them eat in a thousand restaurants until they figure out the recipe themselves.*

Types of Machine Learning

Type	What It Does	Example in Health Research
Supervised	Learns from labelled examples	Predicting hospital readmission from clinical profiles
Unsupervised	Finds hidden patterns in unlabelled data	Grouping patients by similar symptom patterns
Reinforcement	Learns by trial and error	Optimising treatment protocol recommendations

9.3 What Are Large Language Models?

Large Language Models (LLMs), like ChatGPT, Claude, and Gemini, are deep learning systems trained on enormous amounts of text. They learn statistical patterns of language and can draft text, summarise documents, answer questions, and translate between languages.

What AI Cannot Do

Limitation	Why It Matters
Cannot reliably retrieve facts	Generates text based on patterns, citations may be invented
Has a training cutoff date	Events, papers, and regulations after that date are unknown
Does not truly reason	Can produce text that sounds like reasoning without following logic
Reflects biases in training data	Underrepresentation of LMICs and Global South contexts

Hallucination: AI tools regularly generate confident-sounding statements that are factually wrong, including invented citations. Every AI-generated output must be independently verified.

AI in Health Research

Application	What It Does
Medical imaging	Identifies abnormalities in X-rays, CT scans, MRIs
Predictive modelling	Identifies patients at high risk of deterioration
Systematic review	Screens thousands of abstracts for relevance
Natural language processing	Extracts information from clinical notes
Wearable integration	RESpeck + AirSpeck + ML = COPD exacerbation prediction

9.4 Using AI Tools Ethically

Four Key Ethical Challenges

1. **Bias**: AI trained on Global North data may not reflect LMIC contexts. Always ask: Who is in the training data? Who is not?
2. **Hallucination**: AI invents plausible-sounding falsehoods confidently. Verify every fact, statistic, and citation independently.
3. **Over-reliance**: Leaning too heavily on AI can erode the critical thinking skills essential to good research. AI augments human thinking, it does not replace it.
4. **Participant data**: Never input identifiable or sensitive participant data into publicly available AI tools. These platforms may use your inputs for training.

AI Cannot Be an Author

ICMJE authorship criteria require:

- Substantial intellectual contribution to the work
- Critical revision for important intellectual content
- Final approval of the published version
- Accountability for all aspects of the work

AI tools cannot fulfil any of these criteria. They cannot be listed as authors.

Responsible Use of AI

Given the risk of errors or biases when using AI, it is essential that all the work that is generated using these LLMs is checked and validated by human researchers. This would include studying the field of inquiry well enough to ensure appropriate attribution when the words or ideas of others are used. Authors must check all references to ensure that they are appropriate for the text being provided, and that the references themselves are real and not fabricated by the software. Human authors will need to take final responsibility for the language used in any publication, as it will be held to the standards of the journal and will need to be thoughtful, inclusive, and unbiased. Because this technology is changing so rapidly, and the results of misuse could lead to very public recriminations, authors must stay abreast of the developments in ethical use and appropriate attributions in the use of AI (Sage, n.d.). Worldwide application of appropriate oversight and adherence to ethical guidelines are essential to our consensual use of this powerful technology (Table 1).

Responsible Use of AI: Before Submitting.

• Keep records of the system used, the date it was used, and the queries entered
• Check all references for accuracy
• Assure that all concepts are appropriately attributed
• Check paper for plagiarism
• Assure that the language used is unbiased and inclusive
• Study the field of inquiry independently to assure the validity of AI generated information
• Check the journal and/or publishers' guidelines for appropriate forms of attribution
• Check current ethical guidelines on attribution for ethical use of AI on international publishing ethics sites (e.g. ICJME, COPE, Equator Network, and WAME)
Note. AI = Artificial Intelligence; ICJME = International Committee of Medical Journal Editors; COPE = Committee on Publication Ethics; Equator Network = Enhancing the QUAlity and Transparency Of health Research; WAME = World Association of Medical Editors. Adapted from: Sage (n.d.) *Author Guidelines on Using Generative AI and Large Language Models*. https://learningresources.sagepub.com/author-guidelines-on-using-generative-ai-and-large-language-models.

Chetwynd, E. (2024). Ethical use of artificial intelligence for scientific writing: Current trends. Journal of Human Lactation, 40(2), 211–215. https://doi.org/10.1177/08903344241235160

Disclosing AI Use: How and Where
If you used an AI tool while preparing a research output, you must say so, specifically.

How You Used AI	Where to Disclose
Writing, editing, paraphrasing	Acknowledgements section
Data analysis or figure generation	Methods section
Literature search or summarisation	Methods section

Recommended disclosure statement:
"During the preparation of this work, the author(s) used [TOOL NAME AND VERSION] on [DATE] in order to [SPECIFIC PURPOSE]. The author(s) have reviewed, edited, and verified all content, and take full responsibility for the accuracy and integrity of the publication."

WAME further recommends including the full prompt used, the date and time, and the tool version, particularly where AI is used in analysis.

This Handbook as an Example of Transparent Disclosure

The handbook explicitly holds itself up as an example of transparent AI use. This statement must be completed before publication and must include:

- *Name and version of each AI tool used*
- *Date or date range of use*
- *Specific role the tool played (e.g., drafting, editing, restructuring, summarising)*
- *Statement that all content has been reviewed, edited, and verified by the human authors*

Content in this chapter has been adapted from the AuthorAID Online Course in Research Writing (INASP), licensed under CC BY 4.0.

9.5 Free and Open Source Software (FOSS)

What Is FOSS?

Free and Open Source Software (FOSS) is software whose source code is publicly available. Anyone can use, study, modify, and distribute it.
"Free" means freedom, not just cost. *The freedom to see how the software works, adapt it to your needs, and share it with others.*

Why FOSS Matters for Research

Reason	Why It Matters
Reproducibility	Others can run your exact code for free, no paid licence needed
Transparency	You can verify exactly how calculations are performed
Equity	Researchers in LMICs access the same tools as well-funded institutions
Longevity	Open formats and free tools remain readable indefinitely

Key FOSS Tools for Researchers

Category	Tool	What It Does
Statistics	R	Statistical analysis; gold standard for academic research
Statistics	Python	Data science; machine learning; versatile
Data cleaning	OpenRefine	Cleans and harmonises messy data
Qualitative	Taguette	Qualitative coding
Data collection	KoBoToolbox	Mobile forms; works offline
Data collection	ODK	Field data collection without internet
Reference management	Zotero	Integrates with word processors
Word processing	LibreOffice	Free alternative to Microsoft Office
Mapping	QGIS	Geographic data; free alternative to ArcGIS
Version control	Git / GitHub	Tracks changes to code and scripts
Clinical data	REDCap	Free for not-for-profit institutions: Electronic data capture for research
Health records	OpenMRS	Open-source electronic medical records; used across LMICs

9.6 What is Open Science Hardware?

Opne Science Hardware applies the same principles as FOSS to physical objects, sharing the design files, circuit diagrams, component lists, and assembly instructions so anyone can build, modify, or manufacture the device.

- Open-source hardware or open-science hardware are used interchangeably but are different in many ways.
- "Open source hardware" refers to a category of technology that is free to use, change, distribute and innovate.
- "Open science hardware" is a specific subset of open source hardware that deals specifically with publicly accessible scientific tools.

OPEN SCIENCE HARDWARE FOUNDATION

GOSH Gathering for Open Science Hardware

GOSH is the premier Open Science Hardware community on the web.

Why it matters for RESPIRE:

- Commercial scientific instruments are often unaffordable in LMIC settings
- Open designs can be manufactured locally at a fraction of commercial cost
- Repair and maintenance can be done in-country

Examples relevant to health researchers:

Device	What It Does
Arduino	Open microcontroller platform for building sensors and data loggers
Raspberry Pi	Low-cost single-board computer for data collection and processing
OpenFlexure Microscope	3D-printable laboratory microscope used for TB and malaria diagnostics in LMICs
Open Source MRI	Medical imaging by providing affordable, transparent, and collaboratively developed technology for global healthcare access

Chapter 10: Tools and Templates

***What this chapter covers**: A practical reference for every tool, template, and resource mentioned in this handbook. This chapter is not meant to be read, it is meant to be used.*

10.1 Directory of Tools and Templates

Status key

- **UoE provided:** Provided or supported by the University of Edinburgh. Use remains subject to eligibility, institutional policy and the project's data requirements.
- **FOSS for consideration:** Free and open-source software mentioned as a possible option. Confirm local approval before installing or using it for research data.
- **Local approval required:** External or hosted service that should not be used for research data without approval from the relevant institutional service.

Tool or service	Purpose	Cost or access	Status	Where to get it
DMPonline	Creating, reviewing and sharing Data Management Plans	Free or institutionally supported	UoE supported	dmponline.dcc.ac.uk
VeraCrypt	File, container and drive encryption	Free	FOSS for consideration	veracrypt.io
KeePass	Password management	Free	FOSS for consideration	keepass.info /
FreeFileSync	Offline file comparison and folder synchronisation	Free	FOSS for consideration	freefilesync.org
OpenRefine	Cleaning, transforming and harmonising tabular data	Free	FOSS for consideration	openrefine.org
REDCap	Electronic research data capture	Available through participating institutions	UoE provided, when accessed through the institutional service	Access through the institutional REDCap service; see https://project-redcap.org/
KoBoToolbox	Mobile and offline field-data collection	Free and paid plans available	Local approval required	kobotoolbox.org
ODK	Mobile and offline field-data collection	Open-source software; hosted services may incur costs	FOSS for consideration	getodk.org
Taguette	Qualitative data coding	Free	FOSS for consideration	taguette.org
R	Statistical analysis and visualisation	Free	FOSS for consideration	r-project.org
RStudio Desktop	Integrated development environment for R	Free and commercial editions available	FOSS for consideration	posit.co/download/rstudio-desktop
Python	Data processing, analysis and programming	Free	FOSS for consideration	python.org

JASP	Point-and-click statistical analysis	Free	FOSS for consideration	jasp-stats.org
QGIS	Geographic information and mapping	Free	FOSS for consideration	qgis.org
Zotero	Reference collection and management	Free; additional hosted storage may incur costs	FOSS for consideration	zotero.org
LibreOffice	Word processing, spreadsheets and presentations	Free	FOSS for consideration	libreoffice.org
DataStore	Institutionally managed storage for active research data	Available to eligible UoE researchers	UoE provided	https://library.ed.ac.uk/research-support/research-data-service
DataSync	Secure file synchronisation and controlled sharing	Available to eligible UoE researchers	UoE provided	https://library.ed.ac.uk/research-support/research-data-service
DataVault	Long-term retention of stable or inactive research data	Eligibility and charges should be confirmed	UoE provided	https://library.ed.ac.uk/research-support/research-data-service
DataShare	Public research data repository with DOI assignment	Available subject to repository requirements	UoE provided	datashare.ed.ac.uk
Zenodo	External repository for research data, software and other outputs	Free, subject to service limits	Local approval required	zenodo.org
OpenAIRE	European open-science discovery and support infrastructure	Free to access	Local approval required before depositing or connecting research data	openaire.eu

Tool	Purpose	Cost	Where to Get It
DMPOnline	To create, review, and share data management plans that meet institutional and funder requirements.	Free	It is provided by the Digital Curation Centre (DCC) https://dmponline.dcc.ac.uk/
VeraCrypt	File and drive encryption	Free	veracrypt.fr
KeePass / KeePassXC	Password management	Free	keepass.info / keepassxc.org
FreeFileSync	Offline backup and synchronisation (Windows, macOS, Linux)	Free	https://freefilesync.org/download.php
OpenRefine	Data cleaning and harmonisation	Free	openrefine.org
REDCap	Electronic data capture	Free for institutions	redcap-project.org
KoBoToolbox	Mobile field data collection	Free for research	kobotoolbox.org
ODK	Offline field data collection	Free	getodk.org
Taguette	Qualitative data coding	Free	taguette.org
R / RStudio	Statistical analysis	Free	r-project.org
Python	Data science and analysis	Free	python.org
JASP	Point-and-click statistics	Free	jasp-stats.org
QGIS	Geographic information and mapping	Free	qgis.org
Zotero	Reference management	Free	zotero.org
LibreOffice	Word processing and spreadsheets	Free	libreoffice.org
DataStore	Institutional research storage (UoE)	Free (500 GB)	UoE Research Data Service
DataSync	File sharing and synchronisation (UoE)	Free	UoE Research Data Service
DataVault	Long-term archival (UoE)	Contact UoE	UoE Research Data Service
DataShare	Public data repository with DOI (UoE)	Free	datashare.ed.ac.uk

Zenodo	Open-access repository	Free	zenodo.org
OpenAIRE	European open science platform	Free	openaire.eu

10.2 RESPIRE-Specific Resources

Resource	Purpose	Remarks
RESPIRE DMP Template	Standard plan for RESPIRE studies	See below for the key sections of the template
RESPIRE DMP Checklist	Step-by-step DMP completion guide	
RESPIRE Metadata Schema	Standardised metadata for cross-site datasets	
Author Dataset ReadMe Template	Dataset-level documentation (see below for t)	Refer to the ReadMe template below
Data Champion SOP: Monthly Reporting	Standard operating procedure	
Data Champion Pre/Post Survey	Measuring confidence and competence	

Key section of RESPIRE DMP template

0. Responsibilities and Resources	***Who will be involved in the data management of this research?***
1. Data Capture	***What data will be generated or reused in this research?*** ***How much data will be generated?***
2. Data Management	***How will the data be documented to ensure it can be understood?*** ***Where will the data be stored and backed-up?***
3. Integrity	***How will you quality assure your data?***
4. Ethics and Confidentiality	***How will you manage any ethical and Intellectual Property Rights issues?***
5.Retention & Preservation	***Which data do you plan to keep and for how long?*** ***How will the data be preserved?***
6. Sharing & Publication	***Which data will be shared and how?*** ***Are any restrictions on data sharing required?***

AUTHOR DATASET ReadMe Template

This readme file was generated on [YYYY-MM-DD] by [NAME]
<help text in angle brackets should be deleted before finalizing your document>
<[text in square brackets should be changed for your specific dataset]>

GENERAL INFORMATION

Title of Dataset:

<provide at least two contacts>
Author/Principal Investigator Information
Name:
ORCID:
Institution:
Address:
Email:

Author/Associate or Co-investigator Information
Name:
ORCID:
Institution:
Address:
Email:

Author/Alternate Contact Information
Name:
ORCID:
Institution:
Address:
Email:

Date of data collection: <provide single date, range, or approximate date; suggested format YYYY-MM-DD>

Geographic location of data collection: <provide latitude, longiute, or city/region, State, Country>

Information about funding sources that supported the collection of the data:

SHARING/ACCESS INFORMATION

Licenses/restrictions placed on the data:

Links to publications that cite or use the data:

Links to other publicly accessible locations of the data:

Links/relationships to ancillary data sets:

Was data derived from another source?
If yes, list source(s):

Recommended citation for this dataset:

DATA & FILE OVERVIEW

File List: <list all files (or folders, as appropriate for dataset organization) contained in the dataset, with a brief description>

Relationship between files, if important:

Additional related data collected that was not included in the current data package:

Are there multiple versions of the dataset?
If yes, name of file(s) that was updated:
Why was the file updated?
When was the file updated?

METHODOLOGICAL INFORMATION

Description of methods used for collection/generation of data: <include links or references to publications or other documentation containing experimental design or protocols used in data collection>

Methods for processing the data: <describe how the submitted data were generated from the raw or collected data>

Instrument- or software-specific information needed to interpret the data: <include full name and version of software, and any necessary packages or libraries needed to run scripts>

Standards and calibration information, if appropriate:

Environmental/experimental conditions:

Describe any quality-assurance procedures performed on the data:

People involved with sample collection, processing, analysis and/or submission:

DATA-SPECIFIC INFORMATION FOR: [FILENAME]
<repeat this section for each dataset, folder or file, as appropriate>

Number of variables:

Number of cases/rows:

Variable List: <list variable name(s), description(s), unit(s) and value labels as appropriate for each>

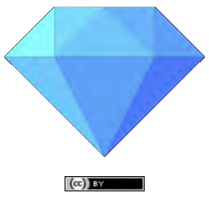
Missing data codes: <list code/symbol and definition>

Specialized formats or other abbreviations used:

Furt-her Reading and Training

Resource	What It Covers	Where
MANTRA	Free, modular RDM training	mantra.ed.ac.uk
FAIR Principles	Full explanation and guidance	go-fair.org
UK Data Service : Five Safes	Data governance framework	ukdataservice.ac.uk
UoE Research Data Service	All UoE data services and training	library.ed.ac.uk/research-support/research-data-service
Creative Commons	Licence types and how to apply them	creativecommons.org
ICMJE Authorship Criteria	Defining authorship in medical research	icmje.orgopen
AuthorAID / INASP	Research writing training including AI use	learn.inasp.info
UNESCO Open Science Toolkit	Framework for institutional open science	unesdoc.unesco.org
Open Knowledge Foundation	Open data principles	okfn.org
Open Science Monitoring Initiative	To encourage the adoption of open science monitoring principles	https://open-science-monitoring.org/
Gathering for Open Science Hardware (GOSH)	**To make open science hardware ubiquitous**	https://openhardware.science/

10.3 OpenRefine



Importing data

OpenRefine does not manipulate your data directly. Instead, the data you import and all the changes you make are stored in a project. You can stop working on a project and continue later if you like. When you want to 'refine' a new file, you start by creating a new project. When you want to continue working on a project, you can open it through "Open Project". It is also possible to export a project on one computer and continue working on it on a different computer. To do so, you transfer the exported files to the new computer and use "Import Project" on the new computer.

Callout

What kinds of data files can I import?

There are several options for getting your data set into OpenRefine. You can upload or import files in a variety of formats including:

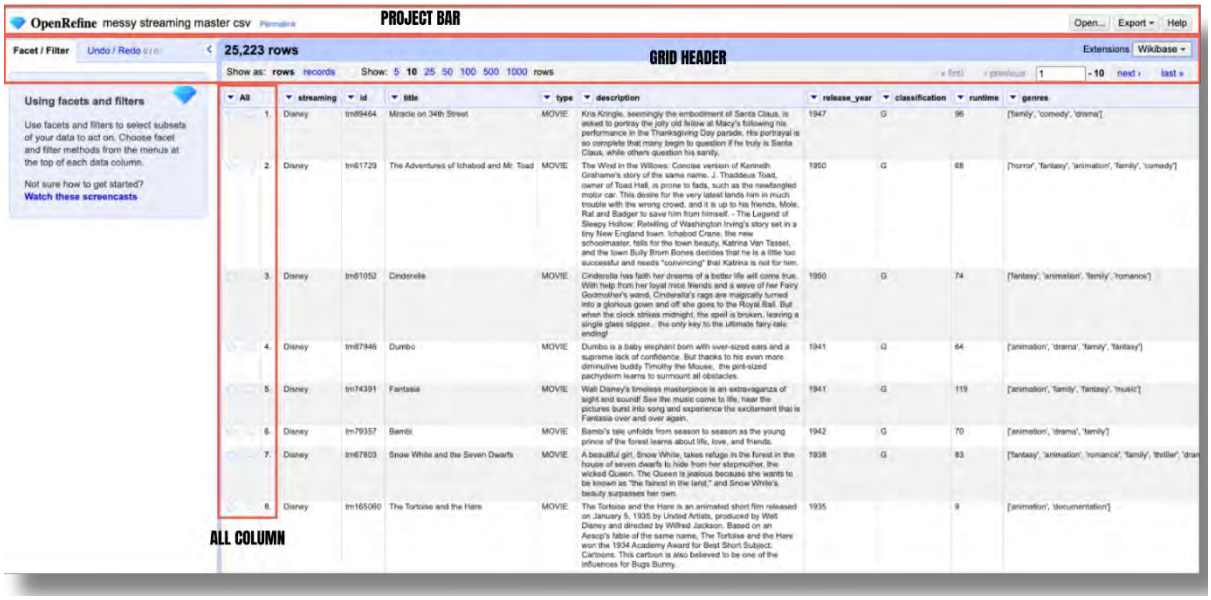
- TSV (tab-separated values)
- CSV (comma-separated values)
- TXT
- Excel
- JSON (javascript object notation)
- XML (extensible markup language)
- Google Spreadsheet

Checklist

Create your first OpenRefine project (using provided data)
To import the data for the exercise below, follow the instructions in Setup to download the data and run OpenRefine.

Interface Walk-Through

- **Familiarise yourself with menus, layout, and tools to streamline workflow.**
- **Helps you quickly locate functions for cleaning and transformation tasks.**



Project Bar

- **Located at the top: shows project title, OpenRefine logo, and permalink.**
- **Permalink saves the current view (facets, filters, sorting) for sharing or revisiting.**
- **Note: It does not save transformations; those are stored in History.**
- **Avoid using the browser's back button to prevent losing facets/view settings.**

Grid Header

- **Displays total rows/records and current view mode (rows vs. records).**
- **Example: "5,774 matching rows (25,223 total)" shows filtered subset.**
- **Records mode groups related data (e.g., all seasons of a TV show under one record).**
- **Includes navigation buttons and row display controls.**

All Column

- **First column always labeled "All".**
- **Shows row numbering (permanent order), plus stars and flags for marking rows.**

- **Stars/flags persist across sessions, useful for tracking errors or highlighting rows.**
- **Can facet by star/flag to manage subsets efficiently.**

Column Operations

- **OpenRefine supports renaming, reordering, and deleting columns.**
- **Example: remove imdb_votes and tmdb_votes via column dropdown or "All" column menu.**
- **Designed for cleaning, not adding new columns/rows.**

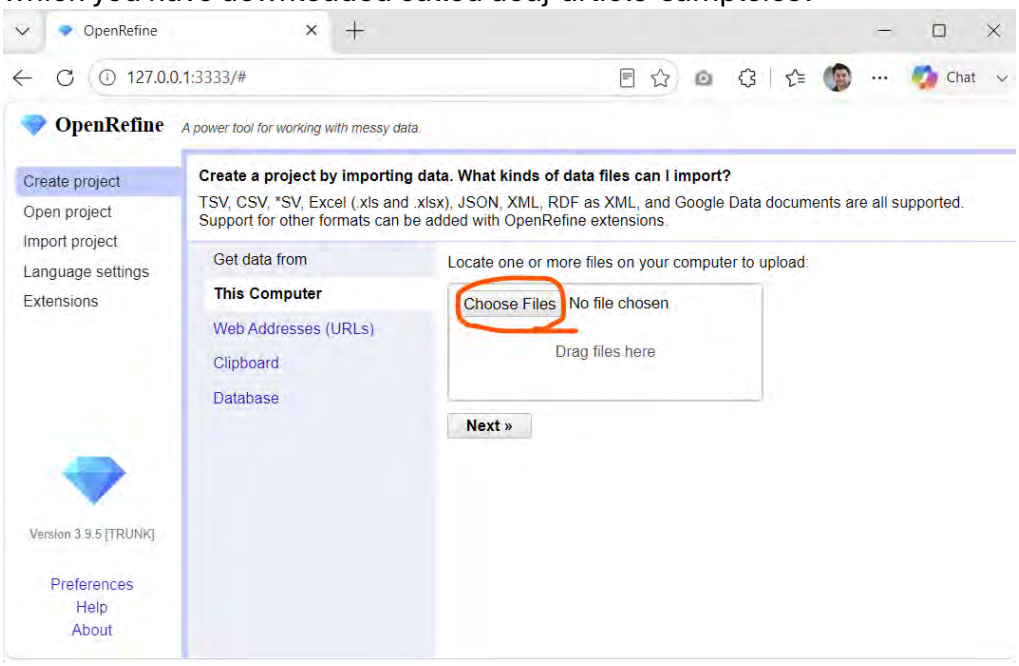
Now that we have a clearer understanding of the data we'll be working with, please take a moment to download and save the file master-streaming-messy.csv to your working environment.

Download the csv File (https://ucsb-library-research-data-services.github.io/openrefine/data/master-streaming-messy.csv)

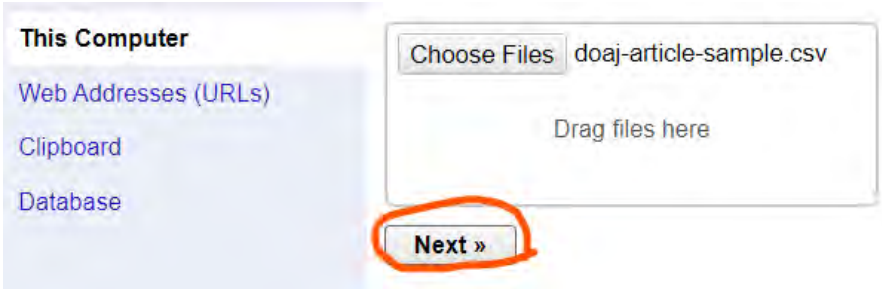
NOTE: If OpenRefine does not open in a browser window, open your browser and type the address http://127.0.0.1:3333/ to take you to the OpenRefine interface. If you don't have any datasets, you can download , which is a CSV file that will open in a new browser tab. Be sure to right-click or control-click in order to save the file (NOTE: In Safari, right-click and select **download linked file**; in Chrome and Firefox, right-click and select **save link as...**). Make a note of the location (i.e. the folder, your desktop) to which you save the file.

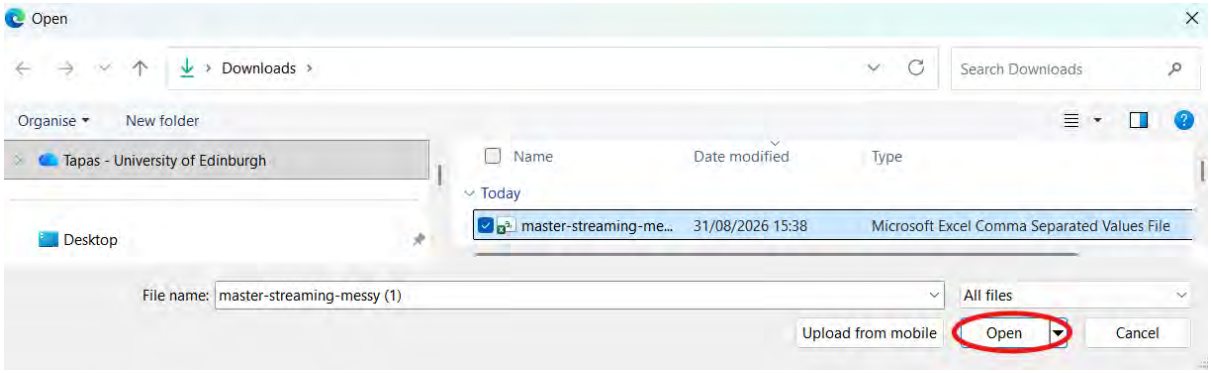
1. Once OpenRefine is launched in your browser, click Create Project from the left hand menu and select Get data from This Computer

2. Click Choose Files (or 'Browse', depending on your setup) and locate the file which you have downloaded called doaj-article-sample.csv

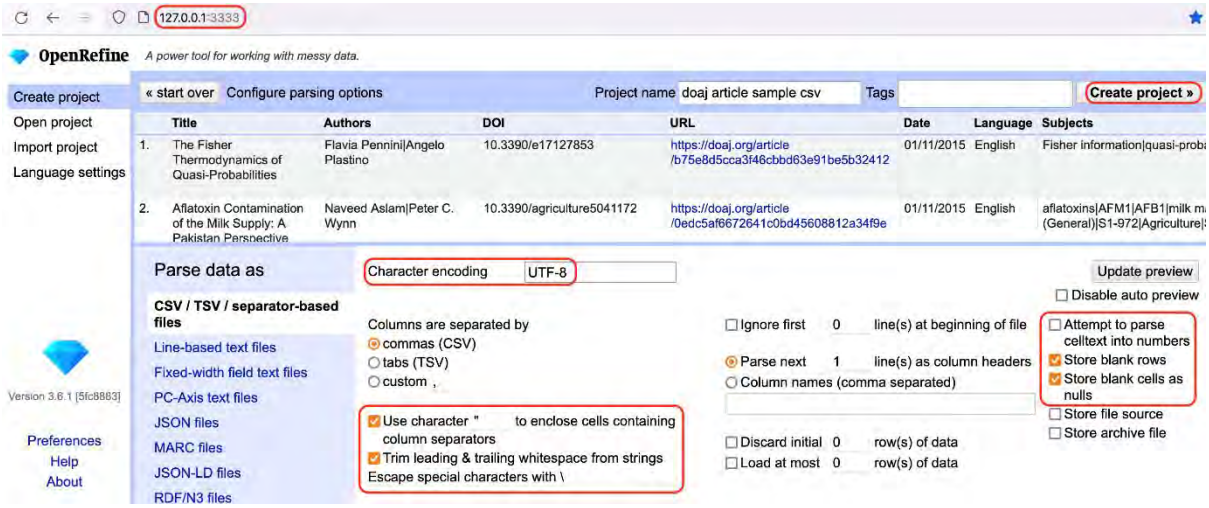


3. Click Next» where the next screen (see below) gives you options to ensure the data is imported into OpenRefine correctly. The options vary depending on the type of data you are importing.





4. Click in the Character encoding box and set it to UTF-8. This ensures that OpenRefine correctly interprets the imported data as UTF-8 encoded. If you don't select this you may find that some special characters (e.g. smart quotation marks) are not displayed correctly.
5. Ensure the first row is used to create the column headings by checking the box Parse next 1 line(s) as column headers
6. OpenRefine will automatically select Use character " to enclose cells containing column separators (such as a comma) as part of their data. This will make sure that OpenRefine doesn't misinterpret any commas (or other characters) within the column data as a delimiter. Keep this option selected.
7. From OpenRefine 3.4 onwards, there is an option to trim leading & trailing whitespace from strings when importing separator-based files. Keeping this checked will ensure that values like English and English, which differ by a single trailing space, are not treated as different values after the import
8. Make sure the Attempt to parse cell text into numbers box is not checked, so OpenRefine doesn't try to automatically detect numbers because this could cause errors such as confusion between date formats (e.g. DD/MM/YYYY vs MM/DD/YYYY).
9. The Project Name box in the upper right corner will default to the title of your imported file. Click in the Project Name box to give your project a different name, if desired.
10. Once you have selected the appropriate options for your project, click the Create project » button at the top right of the screen. This will create the project and open it for you. Projects are saved as you work on them, there is no need to save copies as you go along.



Create Project in OpenRefine

OpenRefine screen (in the left-hand menu). When you click this, you will see a list of the existing projects and can click on a project's name to open it.

Going Further

- Look at the other options on the Import screen - try changing some of these options and see how that changes the Preview and how the data appears after import.
- Do you have access to JSON or XML data? If so the first stage of the import process will prompt you to select a 'record path' - that is the parts of the file that will form the data rows in the OpenRefine project.

Key points

- Use the Create Project option to import data
- You can control how data imports using options on the import screen
- Several files types may be imported into OpenRefine.

Facets

Facets are one of the most useful features of OpenRefine and can help in both getting an overview of the data and improving the consistency of the data.
A 'Facet' groups all the values that appear in a column, and then allows you to filter the data by these values and edit values across many records at the same time.
One of the most commonly used facets is called a 'Text facet'. This groups all the text values in a column and lists each value with the number of records it appears in. The facet information always appears in the left-hand panel in the OpenRefine interface.
To create a Text Facet for a column, click on the drop-down menu at the top of the publisher column and choose Facet -> Text Facet. The facet will then appear in the left-hand panel.

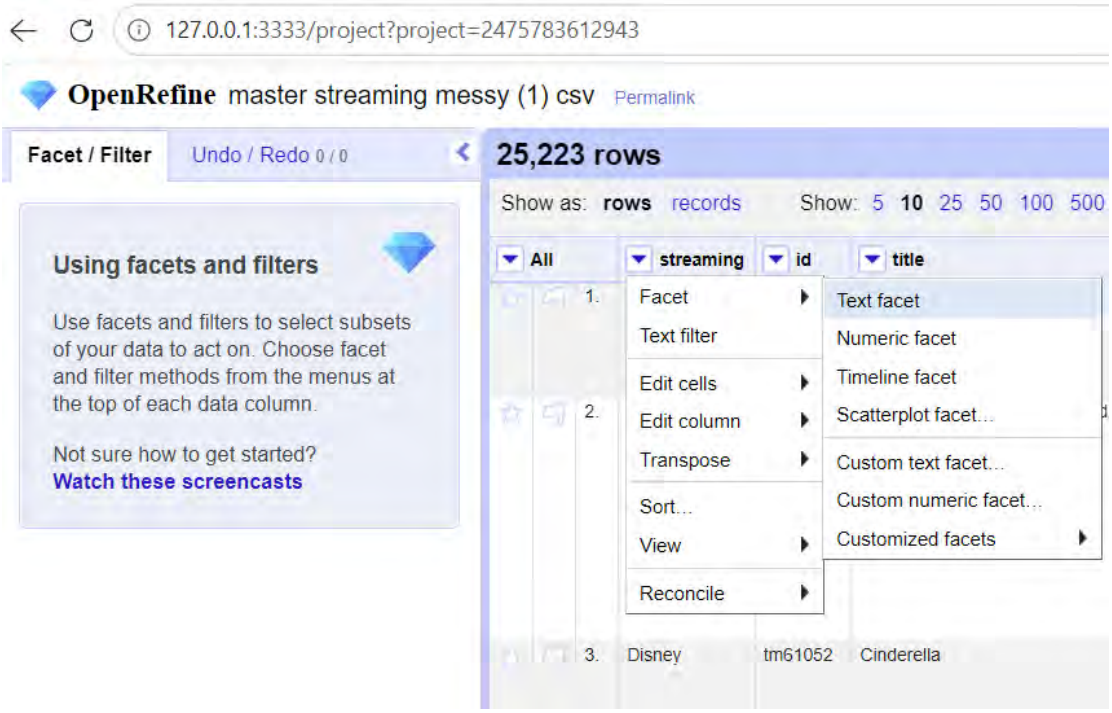
The facet consists of a list of values used in the data. You can filter the data displayed by clicking on one of these headings.

You can include multiple values from the facet in a filter at one time by using the Include option, which appears when you put your mouse over a value in the Facet.

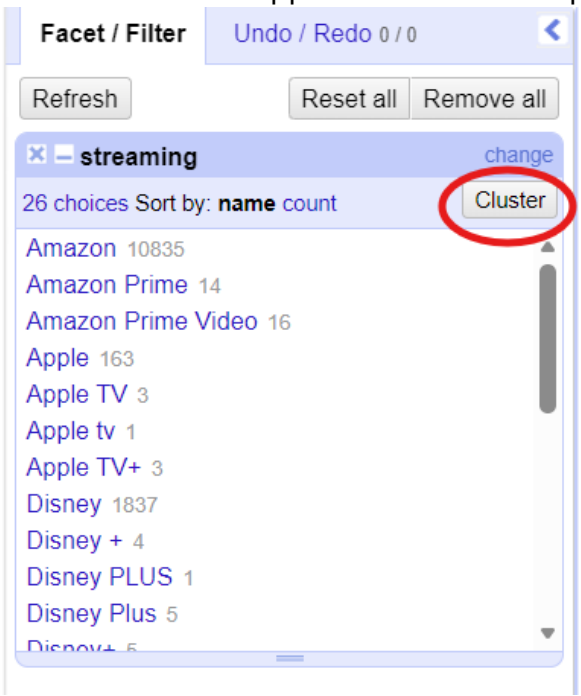
You can also invert the filter to show all records which do not match your selected values. This option appears at the top of the Facet panel when you select a value from the facet to apply as a filter.

Let's create a text facet

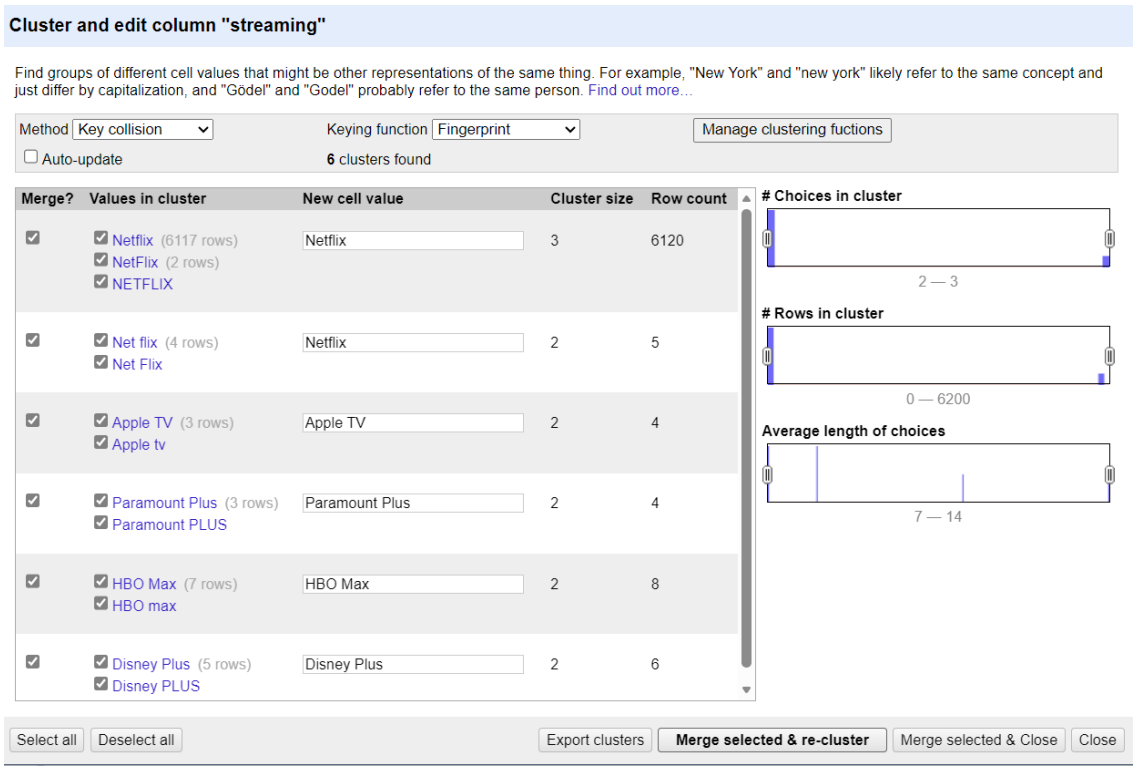
1. Click on the drop down menu at the top of the 'streaming' column and choose Facet > Text Facet.



2. The facet will then appear in the left hand panel



3. After clicking on ‘Cluster ’, a new pop-up window will appear; again, click ‘Cluster’



4. After clicking it will show the cluster and edit column for the streaming variable. Now

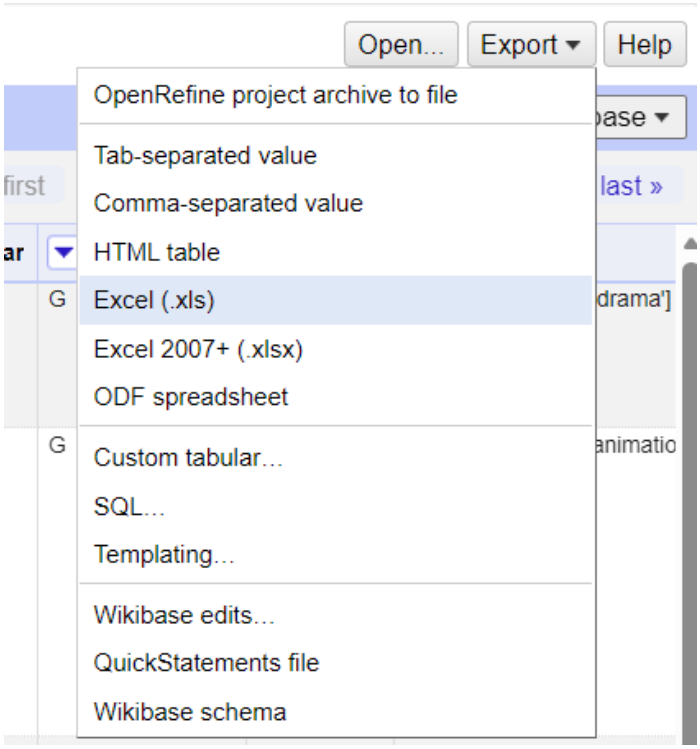
- Review the clusters generated using the selected method (*Key collision*) and keying function (*Fingerprint*).
- Select the checkboxes for the entries you want to refine.
    - For example, select Netflix and Net flix.
    - In the New cell value field, type the standardised name here: Netflix.
    - Click Merge selected & re-cluster to apply the changes and re-run clustering to catch any remaining variations.
    - Repeat the process for other clusters such as Apple TV, Paramount Plus, HBO Max, and Disney Plus.

Once you’ve completed the initial cleaning with Key collision + Fingerprint, you can try other clustering methods to catch subtler variations, such as:

- **N-gram Fingerprint → detects similarities based on overlapping character sequences.**
- **Metaphone3 → groups words by phonetic similarity.**
- **Nearest Neighbor → finds values close to each other based on edit distance.**

Running multiple methods ensures that even tricky inconsistencies (like spacing, spelling errors, or phonetic variations) are standardised.

Once all text values have been standardised and merged using clustering, you can export the cleaned dataset for further analysis or sharing. Click the Export button at the top right of the OpenRefine interface and choose your preferred format, for example, Excel (.xls), Excel 2007+ (.xlsx), or CSV.



For more detailed documentation, examples, and screenshots, visit the official OpenRefine website: https://openrefine.org/docs

10.4 REDCap Quick-Start Guide for Data Champions

REDCap is a highly secure, web-based application for building and managing online surveys and databases. While a full REDCap training course takes several hours, the role of a Data Champion usually focuses on three specific administrative tasks:

- managing the Data Dictionary,
- controlling User Rights, and
- exporting data safely.

This quick-start guide covers these three essential functions.

1. The Data Dictionary (Building and Editing)

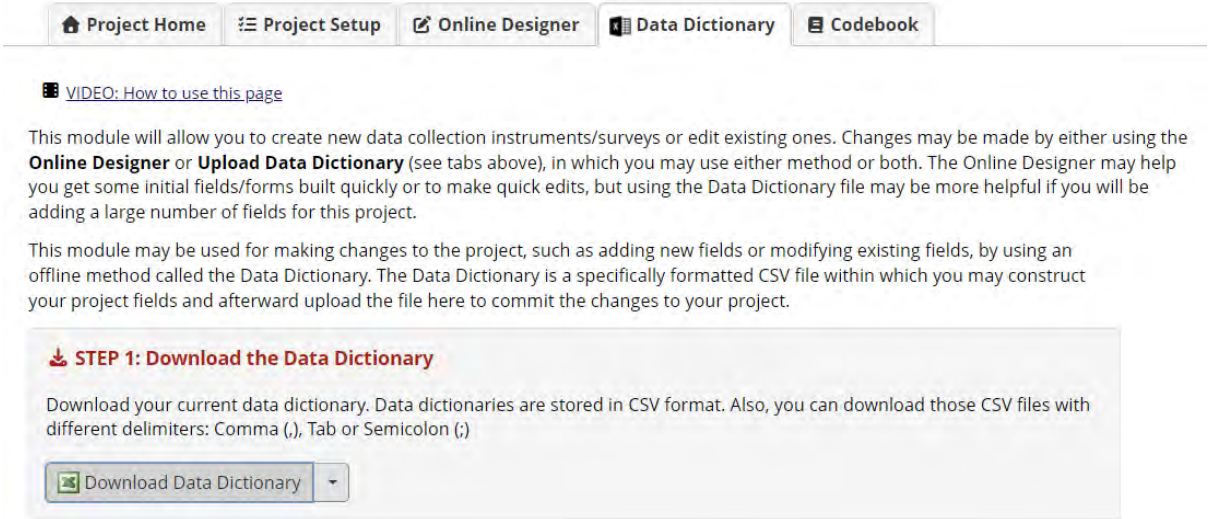
The Data Dictionary is the backbone of your REDCap project. It is a simple Excel (CSV) file that contains all the rules, variables, and validation checks for your entire study.

Instead of building a database one question at a time using the online designer, Data Champions can download the Data Dictionary, edit it in Excel alongside colleagues, and upload the new version.

How to use it:

1. Navigate to your project on REDCap.
2. Under the "Project Home and Design" menu on the left, click **Data Dictionary**.

3. Click the **Download the current Data Dictionary** button.



4. Open the downloaded CSV file in Excel. Each row represents one variable (such as HB test or HB result).
5. Make your required changes (for example, adding validation rules or missing data codes like -99).
6. Save the file as a CSV.

7. Return to the Data Dictionary page in REDCap, browse for your updated file, and click **Upload File**. REDCap will check for errors before committing the changes.

Understanding the REDCap Data Dictionary CSV

When you download your REDCap Data Dictionary as a CSV file and open it in Excel, you will see a row of column headers. Each row beneath these headers represents a single question or field in your database.

Variable / Field Name	Form Name	Section Header	Field Type	Field Label	Choices, Calculations, OR Slider Labels	Field Note	Text Validation Type OR Show Slider Number	Text Validation Min	Text Validation Max	Identifier?	Branching Logic (Show field only	Required Field?	Custom Alignment	Question Number (surveys only)	Matrix Group Name	Matrix Ranking?	Field Annotation
hb_test	Questionnaire		dropdown	HB test	1, Enter the result [ ] [ ]. [ ] g/dl | 0, Not done								LV				
hb_result	Questionnaire		text	Enter the HB result [ ] [ ]. [ ] g/dl			number	3	20		[hb_test] = 1		LV				

Here is what the most important operational headers mean:

- **Variable / Field Name:** The short, unique code name for the variable (for example, hb_test). This is the name used by statistical software like R or STATA. It cannot contain spaces or capital letters.
- **Form Name:** The specific survey or data entry form where this question belongs (for example, 'Questionnaire').
- **Field Type:** How the question is displayed. Common types include text (a blank text box), dropdown (a drop-down menu), radio (single choice buttons), or calc (a calculated field).
- **Field Label:** The actual question text the user will see on the screen (for example, "Enter the HB result").
- **Choices, Calculations, OR Slider Labels:** For multiple-choice questions (dropdowns or radio buttons), this contains the raw numerical code and the text label the user sees, separated by a comma (for example, 1, Yes | 0, No).
- **Text Validation Type / Min / Max:** If the field is a text box, this restricts what can be typed. You can set the type to number, date_ymd, or email, and set minimum and maximum allowable values.
- **Identifier?:** Marked with a 'y' (yes) if the field contains identifying personal information (like a name or phone number). This allows REDCap to easily strip this data out during de-identified exports.
- **Branching Logic:** The rule that determines if a question should be hidden or shown, based on previous answers.
- **Required Field?:** Marked with a 'y' (yes) if the user must fill this out before saving the form.

A Practical Example: Using Branching Logic for an HB Test

Branching logic (also known as skip logic) is one of REDCap's most powerful features. It keeps your forms looking clean by hiding irrelevant questions.
Look at the Haemoglobin (Hb) test example in your CSV screenshot. It uses two rows working together:

Row 1: The Trigger Question

- **Variable Name:** hb_test

- **Field Type:** dropdown
- **Field Label:** HB test
- **Choices:** 1, Enter the result | 0, Not done
  This first question simply asks if the test was done. If the user selects "Not done", there is no reason to show them a blank box for a test result.

Row 2: The Hidden Follow-Up Question

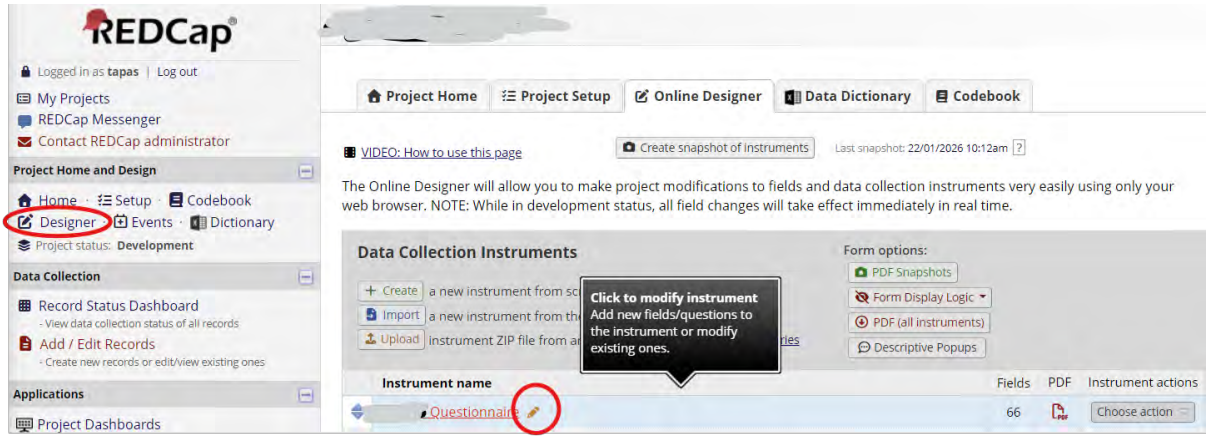
- **Variable Name:** hb_result
- **Field Type:** text
- **Field Label:** Enter the HB result (g/dl)
- **Text Validation:** number (Min: 3, Max: 20)
- **Branching Logic:** [hb_test] = 1

How it works in practice: The second question (hb_result) remains completely invisible on the screen. It will only appear **if** the user selects option 1 ("Enter the result") in the hb_test dropdown.
Notice that the hb_result row also includes validation. Because it is a text box, we set the validation type to number with a minimum of 3 and a maximum of 20. If a data entry clerk accidentally types "25" or the letter "A", REDCap will block the entry. This combination of **Branching Logic** (showing only what is needed) and **Validation** (restricting what can be typed) is the key to creating a high-quality clinical database.

Using the Online Designer

The Online Designer is REDCap's visual editor. Instead of typing codes into a spreadsheet, you can build your forms directly on the screen using a drag-and-drop menu. Any change you make in the Online Designer automatically updates the underlying Data Dictionary, and vice versa.

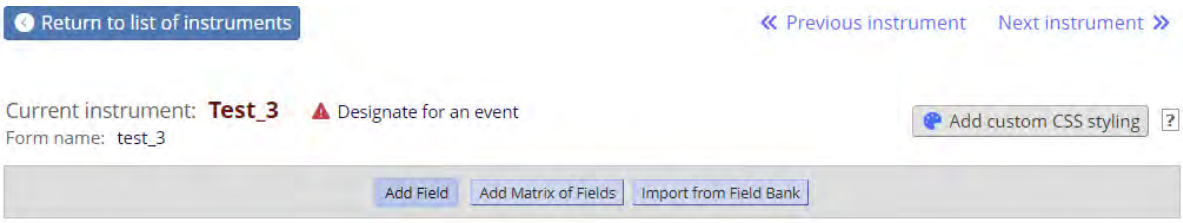
You can access the 'Designer' under 'Project Home and Designs' and click on the modify instrument icon.



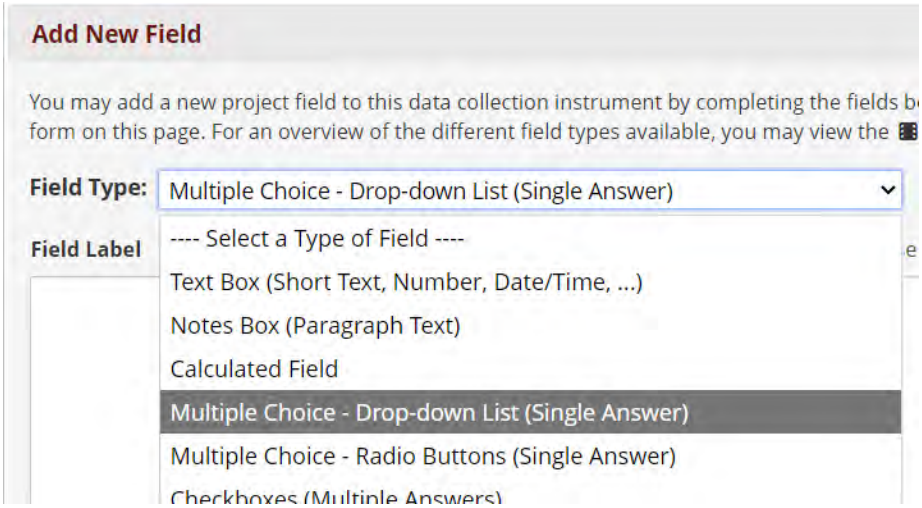
If we take the Haemoglobin (Hb) test example from earlier, here is how you would build it using the Online Designer:

1. Adding the First Question (The Trigger)

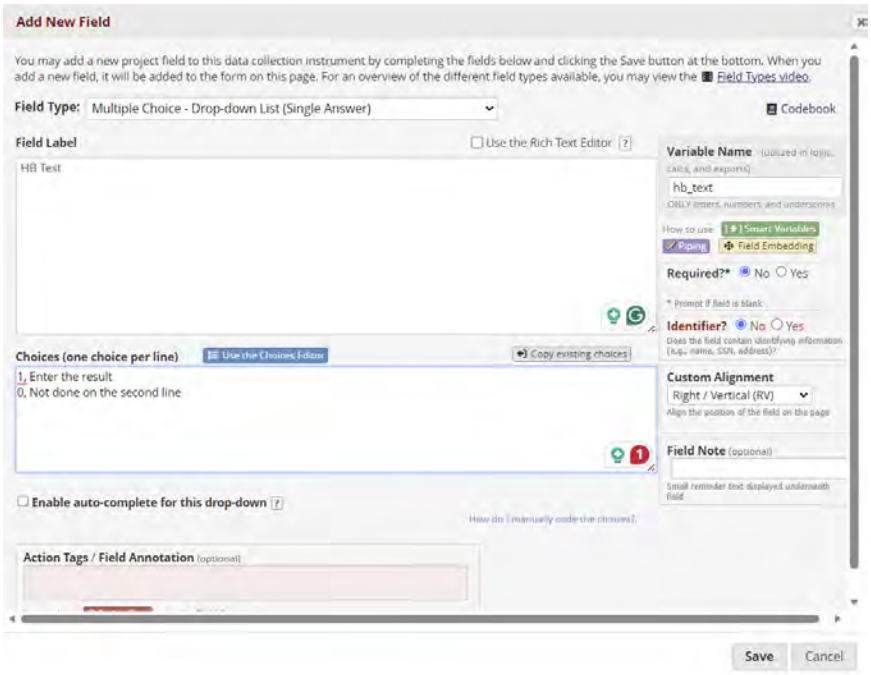
- Click the **Add Field** button.



- A popup window appears. Under "Field Type", select **Dropdown List (Single Answer)**.



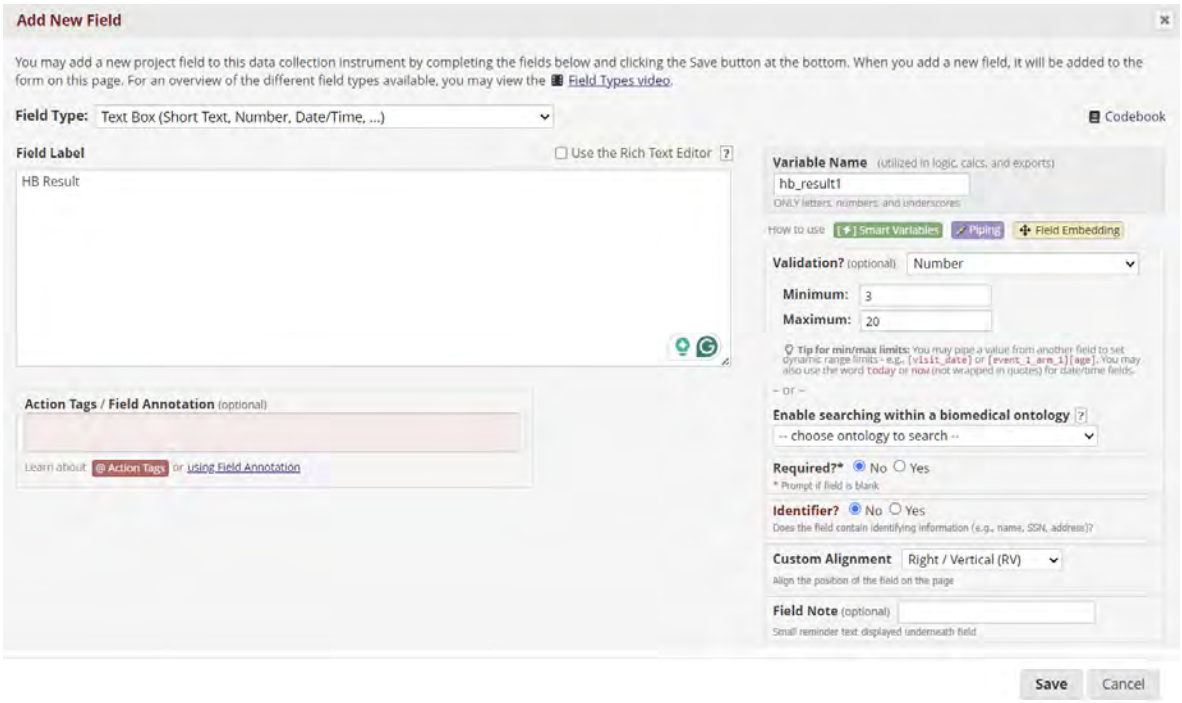
- Type your "Field Label" (HB test) and your "Variable Name" (hb_test).
- Type your choices in the provided box (e.g., 1, Enter the result on the first line, and 0, Not done on the second line).
- Click **Save**.



2. Adding the Follow-up Question and Validation

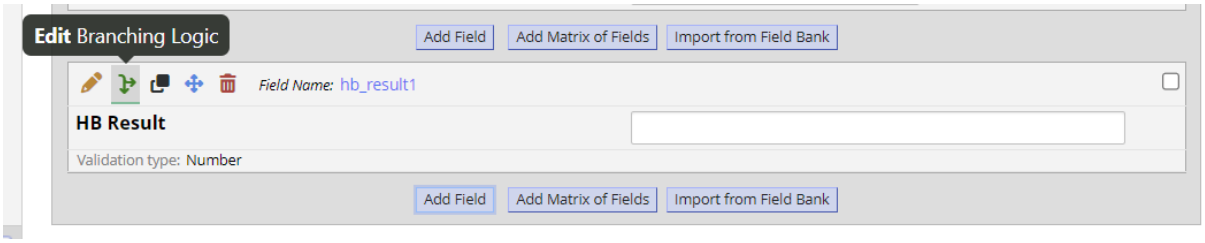
- Click **Add Field** again to create the next question.

- Select **Text Box (Short Text, Number, Date/Time)**.
- Set the Variable Name to hb_result.
- To enforce quality control, look for the "Validation" section in the popup. Change it to **Number**, and type 3 in the Minimum box and 20 in the Maximum box.
- Click **Save**.



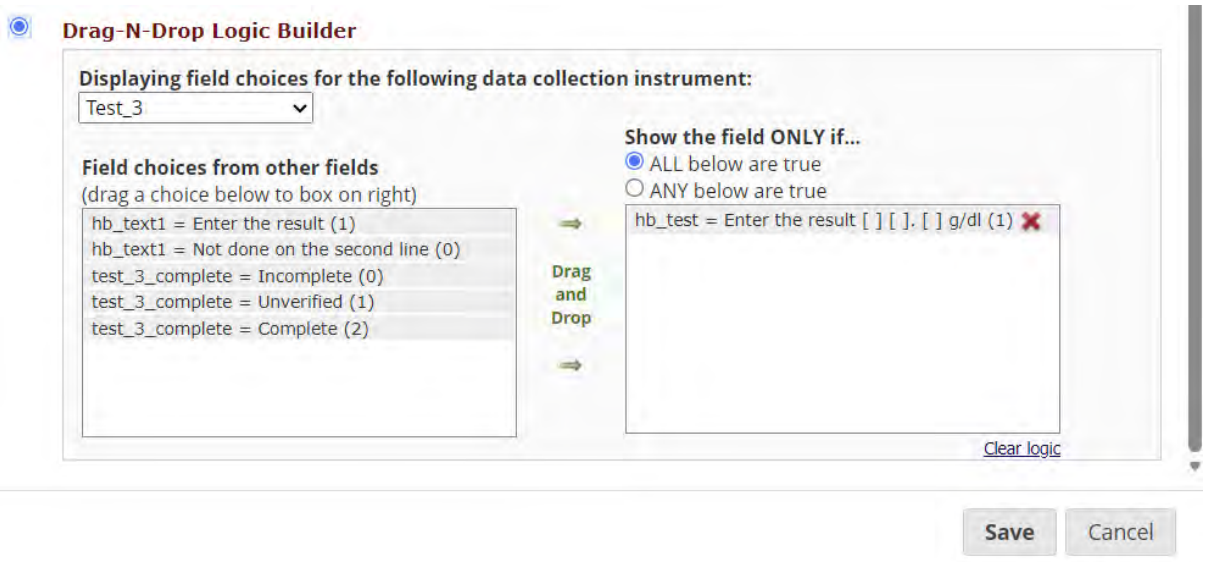
3. Applying Branching Logic : Now you need to hide the test result box unless it is needed.

- Look at the hb_result field you just created on your screen.
- Click the **Green Double-Arrow Icon** (the Branching Logic button) located in the top-left corner of the field box.



- The Branching Logic pop-up appears.
  You can either type [hb_test] = '1' in the advanced syntax box,

or simply use the Drag-and-Drop Logic Builder to drag the "Enter the result" option into the logic box.



- Click **Save**. The question will now remain hidden until the correct option is selected.

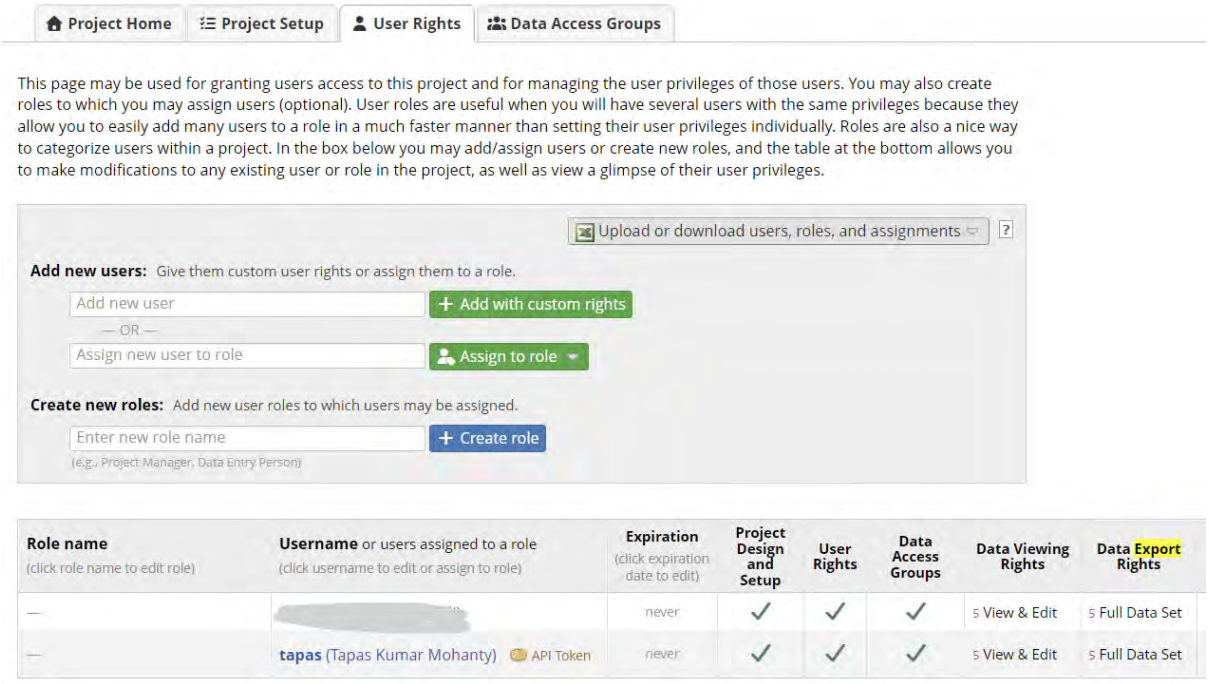
💡 *"Data Champion Tip: The Online Designer is excellent for building forms from scratch or making quick edits. However, if you need to translate your entire study into a second language, or add the word 'months' to 50 different questions, downloading the CSV Data Dictionary, using 'Find and Replace' in Excel, and uploading it back to REDCap is much faster!"*

2. Managing User Rights (The Principle of Minimum Access)

As a Data Champion, you must ensure that people only have access to the data they absolutely need. REDCap allows you to be highly specific about user permissions.

How to configure access:

1. Under "Applications" on the left menu, click **User Rights**.
2. Type the username of your colleague and click **Add with custom rights**.
3. You will see a detailed checklist of permissions. Apply these basic rules:
    - **Data Entry Staff:** Grant "Create Records" and "Edit" rights, but set "Data Export" to "De-Identified" or "None".
    - **Statisticians / Analysts:** Grant "Data Export" rights, but restrict "Data Entry" rights so they cannot accidentally alter the live data.
    - **Project Managers:** Grant "Project Design and Setup" rights.

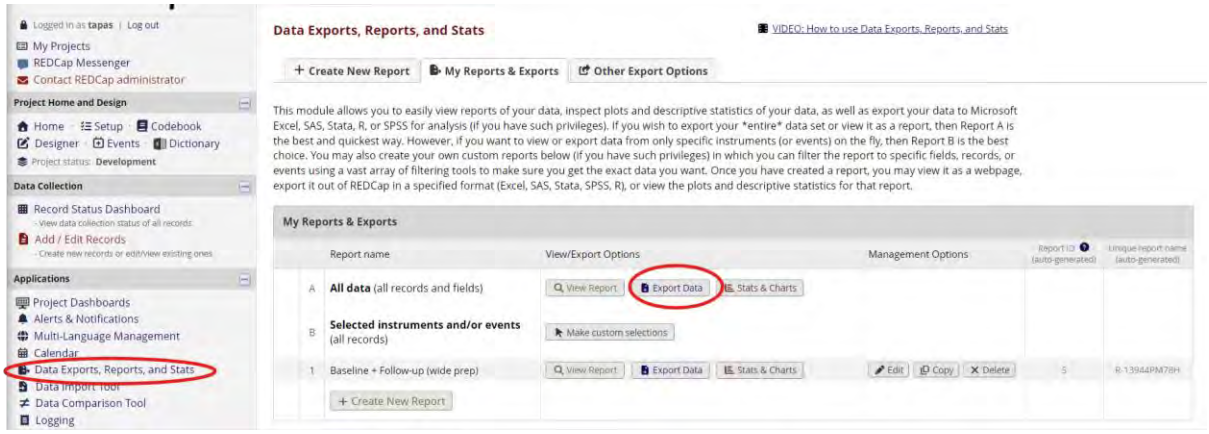


3. Exporting Data Safely

When it is time to analyse your data or share it internally, you must export it from REDCap cleanly. REDCap has built-in pseudonymisation features that allow you to strip out identifiers during the export process.

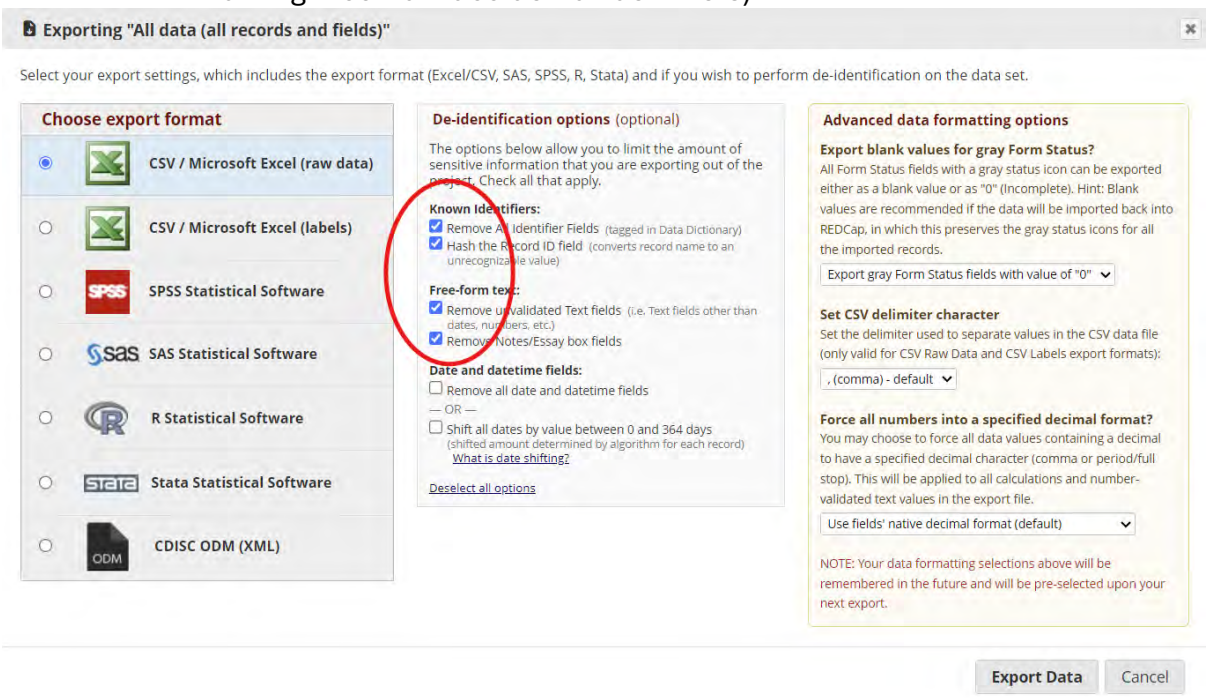
How to export pseudonymised data:

1. Under "Applications" on the left menu, click **Data Exports, Reports, and Stats**.
2. Click **Export Data** next to "All data (all records and fields)".



3. A popup window will appear with export options (CSV, SPSS, R, STATA, etc.).

4. Click on the **Advanced Data Formatting Options** tab.
5. Under "De-identification Options", select the following checkboxes:
    - Remove all tagged Identifier fields
    - Hash the Record ID field (converts the record ID to an unrecognisable code)
    - Remove unvalidated Text fields and Notes fields (removes qualitative text that might contain accidental identifiers)



6. Click **Export Data**.

💡 *Note: Always test your export settings early in the project. Download a test export and confirm with your team that the format is exactly what they need for analysis in R or STATA.*

10.5 VeraCrypt to secure your data

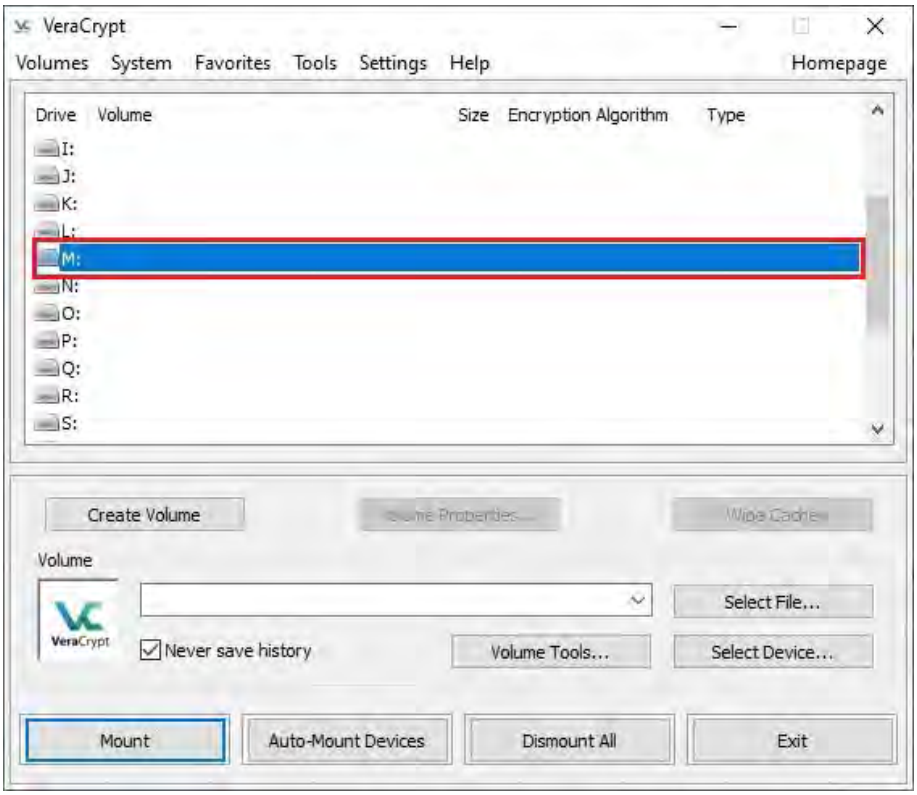
How to Create and Use a VeraCrypt Container

This chapter contains step-by-step instructions on how to create, mount, and use a VeraCrypt volume. We strongly recommend that you also read the other sections of this manual, as they contain important information.

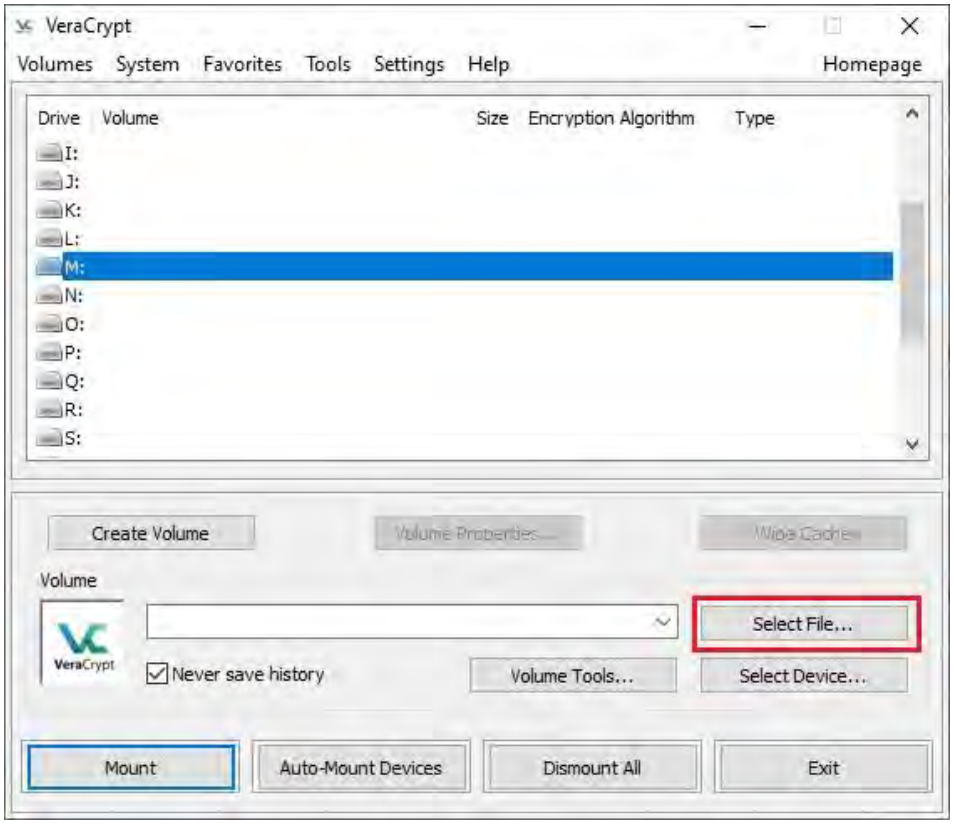
Download and install it from https://veracrypt.io/ (see the documentation *https://veracrypt.io/en/Documentation.html* for the installation steps and creating an encrypted container. While adding a Volume password, you have to choose a good volume password. Read carefully the information displayed in the Wizard window about what is considered a good password (see Passwords and KeePass sections below)

Steps to mount the VeraCrypt volume

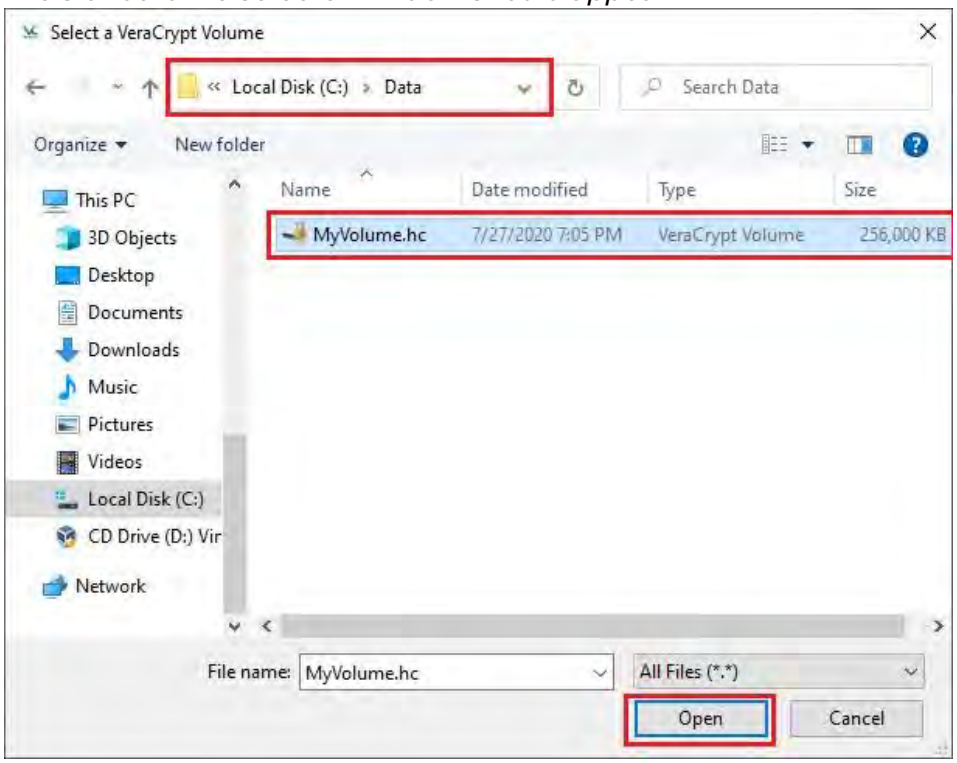
Once you have created a VeraCrypt volume (file container), you can mount the volume to access the container



Select a drive letter from the list (marked with a red rectangle). This will be the drive letter to which the VeraCrypt container will be mounted.
Note: In this tutorial, we chose the drive letter M, but you may of course choose any other available drive letter.

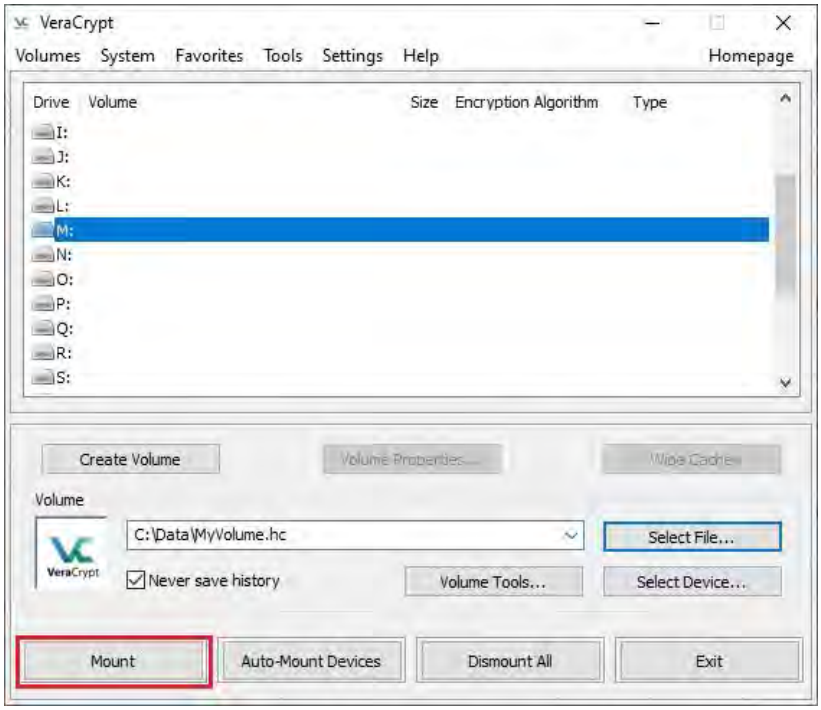


*Click **Select File**.*
The standard file selector window should appear.

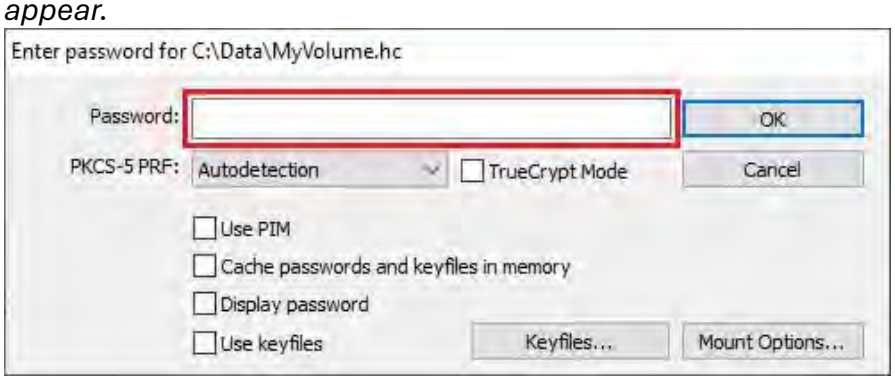


*In the file selector, browse to the container file (which we created in Steps 6-12) and select it. Click **Open** (in the file selector window).*
The file selector window should disappear.

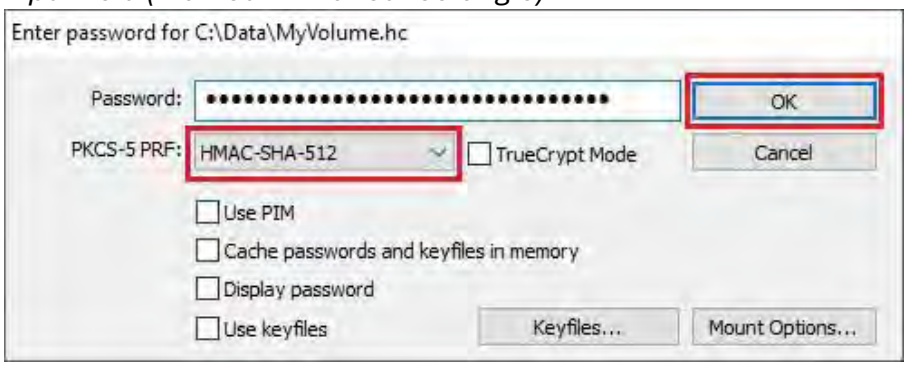
In the following steps, we will return to the main VeraCrypt window.



*In the main VeraCrypt window, click **Mount**. Password prompt dialog window should appear.*



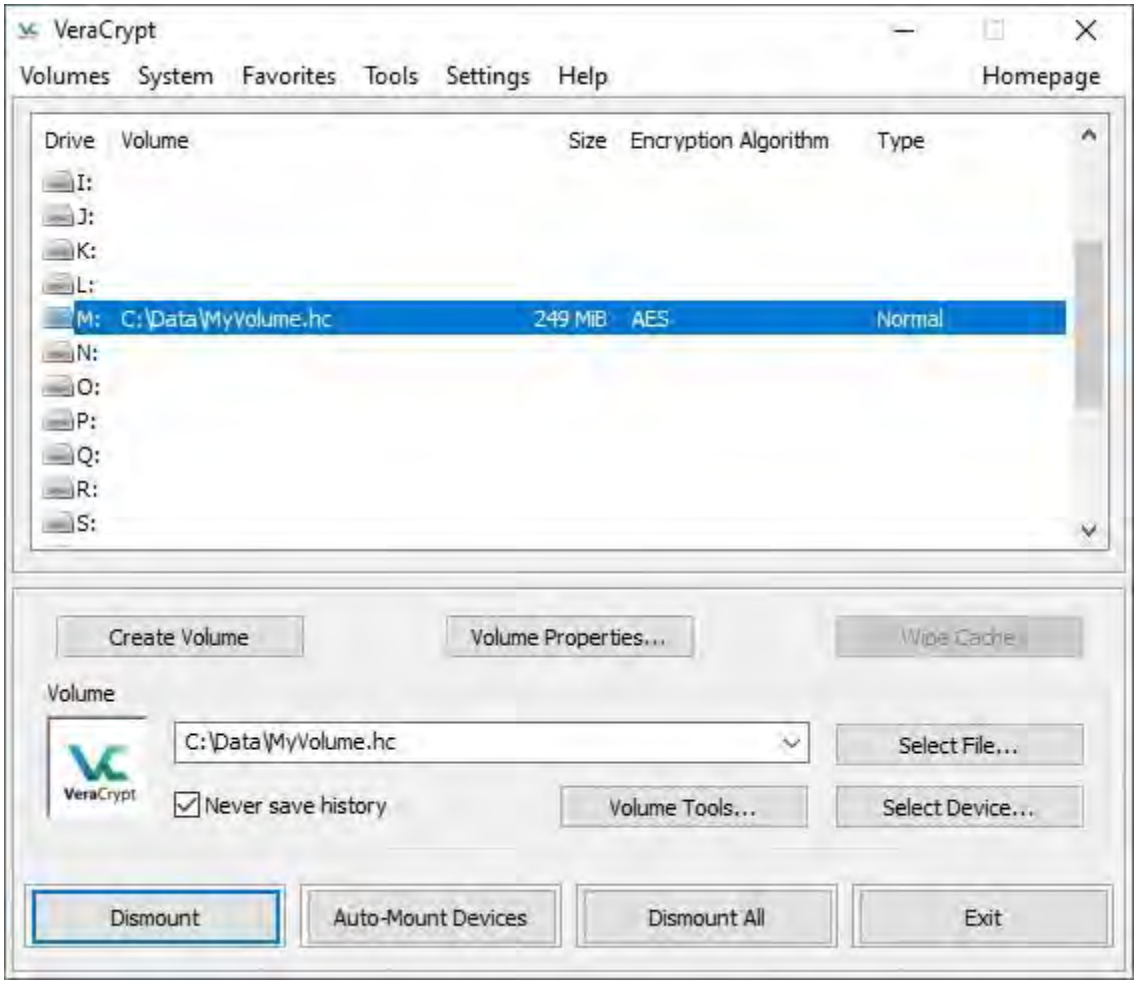
Type the password (which you entered while creating the container) in the password input field (marked with a red rectangle).



Select the PRF algorithm that was used during the creation of the volume (SHA-512 is the default PRF used by VeraCrypt). If you don't remember which PRF was used, just leave it set to "autodetection" but the mounting process will take more time.
*Click **OK** after entering the password.*
VeraCrypt will now attempt to mount the volume. If the password is incorrect (for example, if you typed it incorrectly), VeraCrypt will notify you and you will need to repeat

*the previous step (type the password again and click **OK**). If the password is correct, the volume will be mounted.*

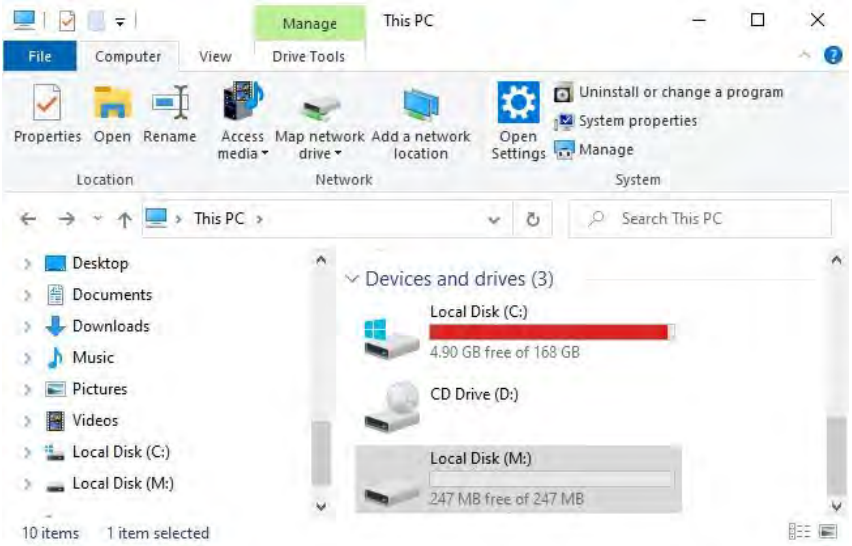
FINAL STEP:



We have just successfully mounted the container as a virtual disk M:

The virtual disk is entirely encrypted (including file names, allocation tables, free space, etc.) and behaves like a real disk. You can save (or copy, move, etc.) files to this virtual disk and they will be encrypted on the fly as they are being written.

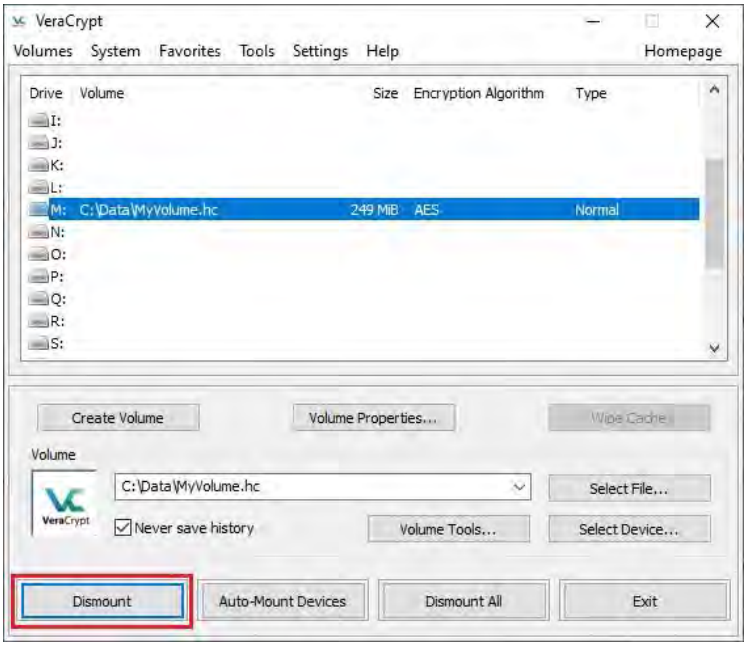
If you open a file stored on a VeraCrypt volume, for example, in media player, the file will be automatically decrypted to RAM (memory) on the fly while it is being read.
Important: Note that when you open a file stored on a VeraCrypt volume (or when you write/copy a file to/from the VeraCrypt volume) you will not be asked to enter the password again. You need to enter the correct password only when mounting the volume.
You can open the mounted volume, for example, by selecting it on the list as shown in the screenshot above (blue selection) and then double-clicking on the selected item. You can also browse to the mounted volume the way you normally browse to any other types of volumes. For example, by opening the 'Computer' (or 'My Computer') list and double-clicking the corresponding drive letter (in this case, it is the letter M).



You can copy files (or folders) to and from the VeraCrypt volume just as you would copy them to any normal disk (for example, by simple drag-and-drop operations). Files that are being read or copied from the encrypted VeraCrypt volume are automatically decrypted on the fly in RAM (memory). Similarly, files that are being written or copied to the VeraCrypt volume are automatically encrypted on the fly in RAM (right before they are written to the disk).

Note that VeraCrypt never saves any decrypted data to a disk, it only stores it temporarily in RAM (memory). Even when the volume is mounted, data stored in the volume is still encrypted. When you restart Windows or turn off your computer, the volume will be unmounted, and all files stored on it will be inaccessible (and encrypted). Even when the power supply is suddenly interrupted (without proper system shutdown), all files stored on the volume will be inaccessible (and encrypted). To make them accessible again, you have to mount the volume. To do so, repeat Steps 13-18.

If you want to close the volume and make files stored on it inaccessible, either restart your operating system or unmount the volume. To do so, follow these steps:

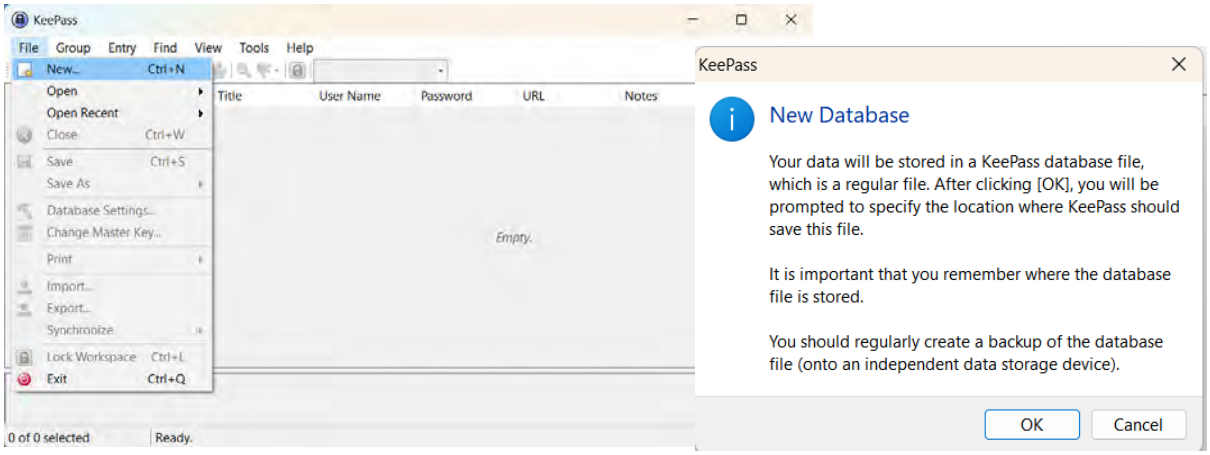


Select the volume from the list of mounted volumes in the main VeraCrypt window (marked with a red rectangle in the screenshot above) and then click ***Unmount*** *(also marked with a red rectangle in the screenshot above). To make files stored on the volume accessible again, you will have to mount the volume. To do so, repeat Steps 13-18.*

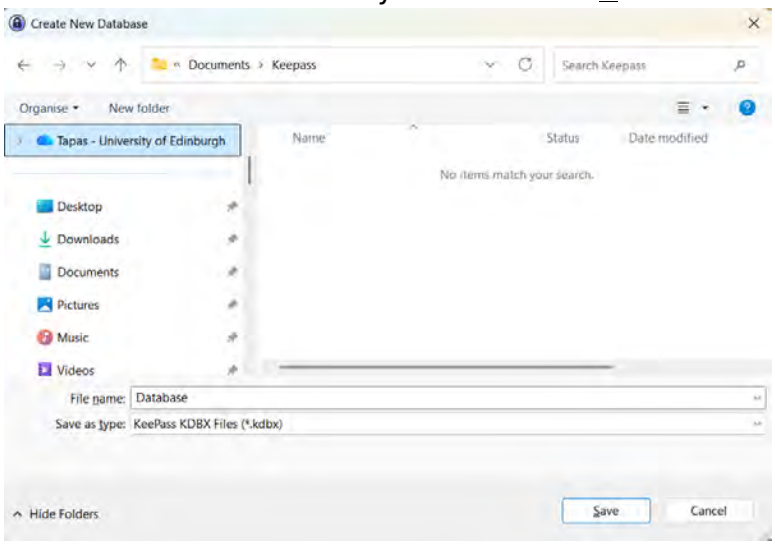
10.6 KeePass to securely store your passwords

Step-by-step for setting up and using KeePass after installation:

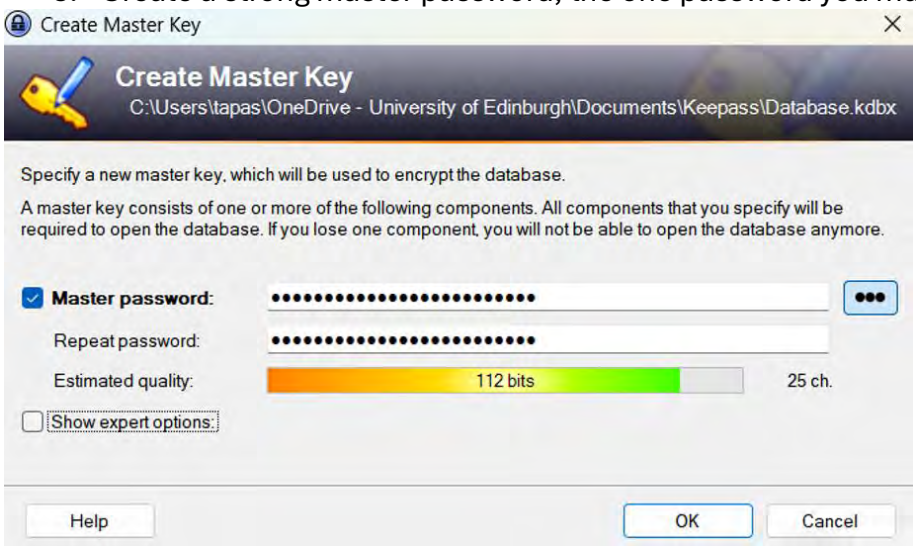
1. Open KeePass → File → New → OK



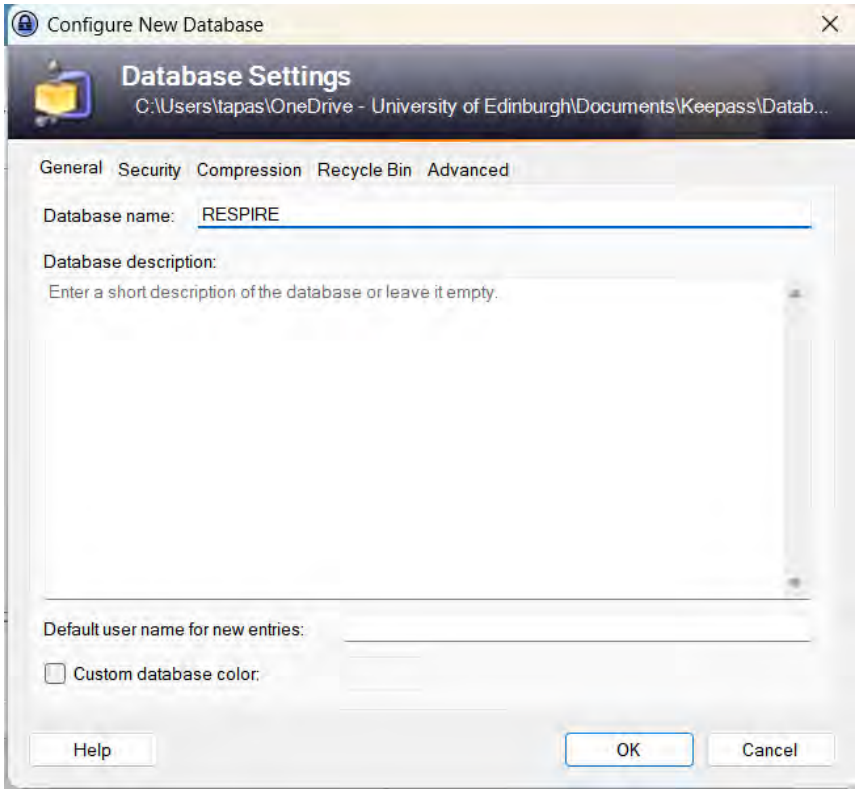
2. → Choose a location for your database→ Save



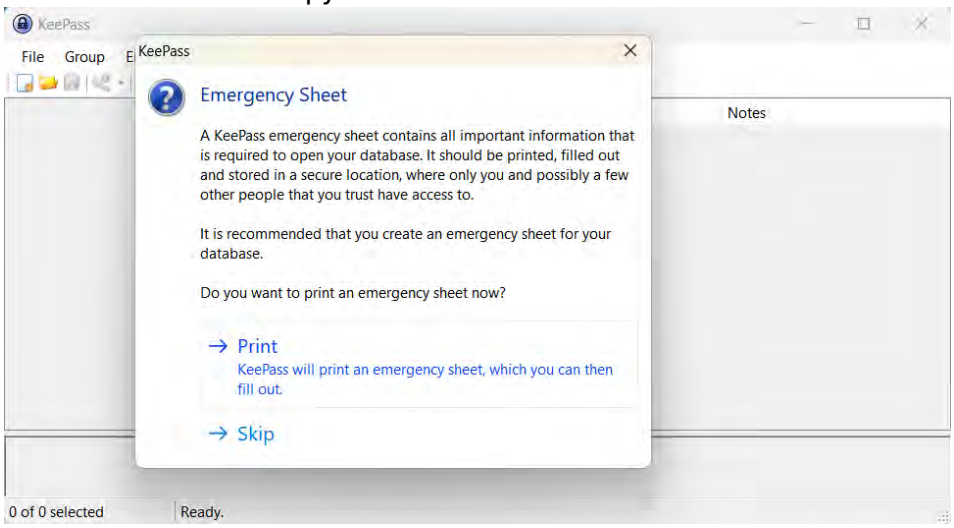
3. Create a strong master password, the one password you must remember



4. Add a Database Name → OK



5. Print a hard copy



6. After taking the printout, write the master key and keep two copies securely in two different places (e.g, one with the PI and another with the Data Manager just in case).

For RESPIRE teams, Store shared credentials in a KeePass database held in a restricted-access folder on a secure institutional server or cloud server (DataSync) never in an email or a chat message.


Database - Emergency Sheet

Page 1 of 1

KeePass
Emergency Sheet

06/08/2026

Database file:

C:\Users\tapas\OneDrive - University of Edinburgh\Documents\Keepass\Database.kdbx

You should regularly create a backup of the database file (onto an independent data storage device). Backups are stored here:

Master Key

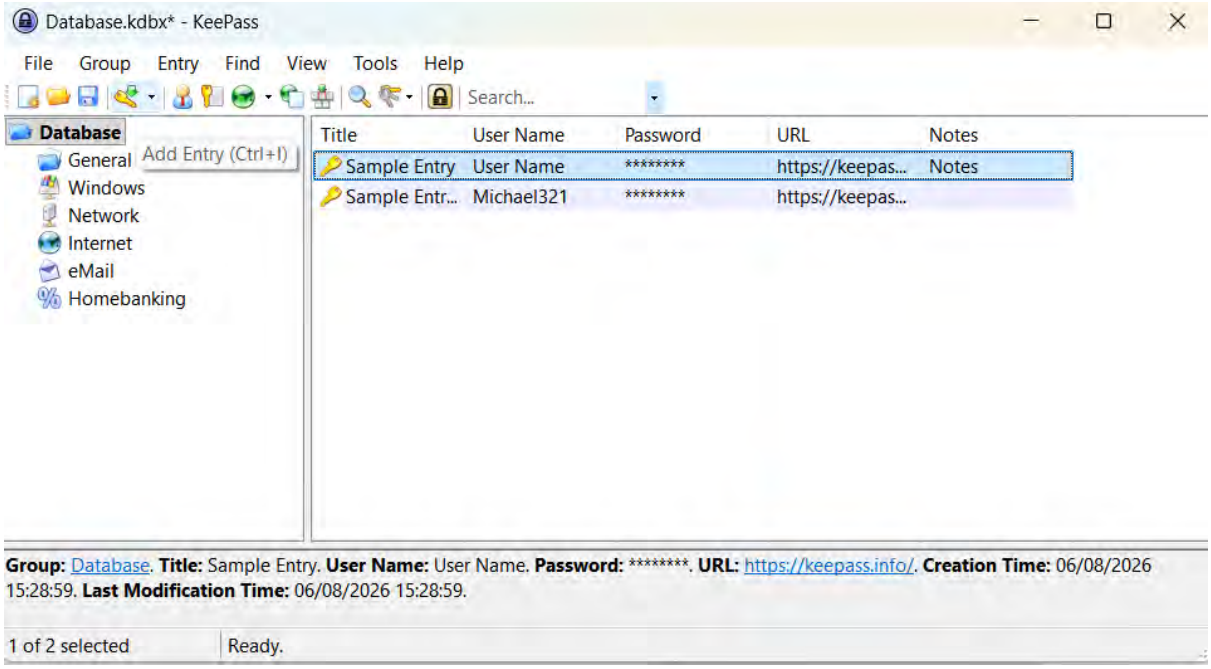
The master key for this database file consists of the following components:

- **Master password:**

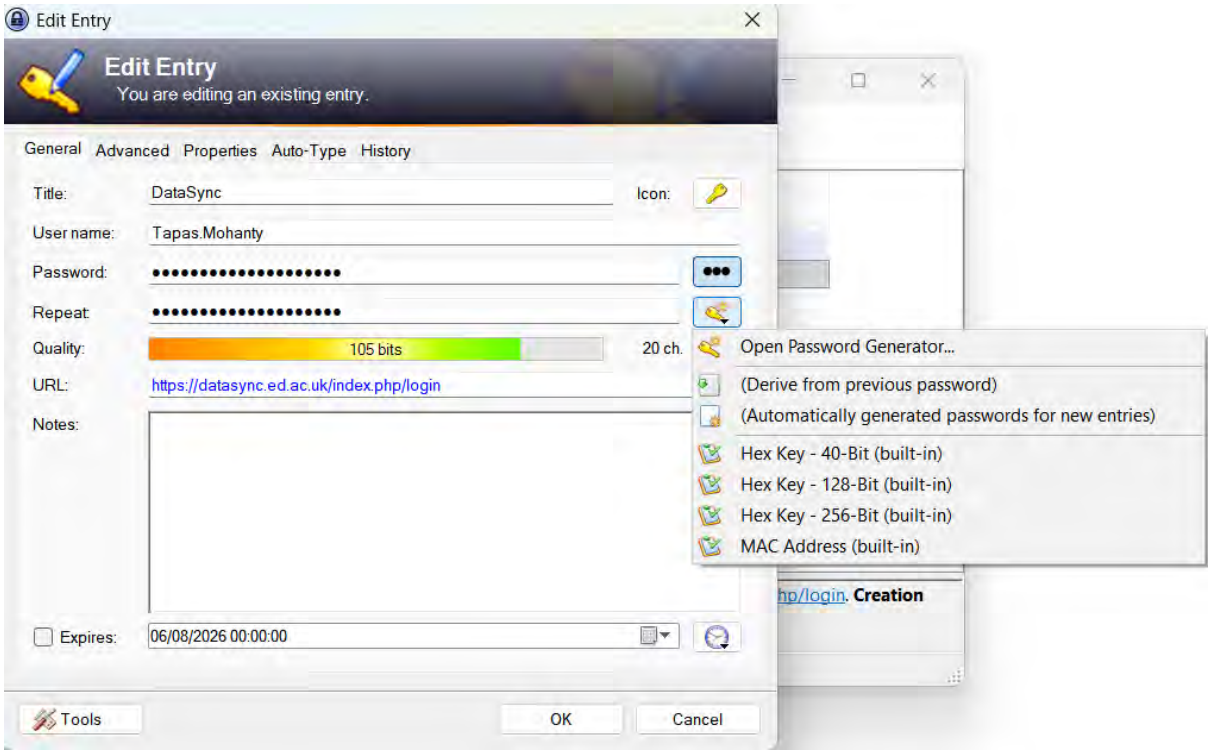
Instructions and General Information

- A KeePass emergency sheet contains all important information that is required to open your database. It should be printed, filled out and stored in a secure location, where only you and possibly a few other people that you trust have access to.
- If you lose the database file or any of the master key components (or forget the composition), all data stored in the database is lost. KeePass does not have any built-in file backup functionality. There is no backdoor and no universal key that can open your database.
- The latest KeePass version can be found on the KeePass website: https://keepass.info/.

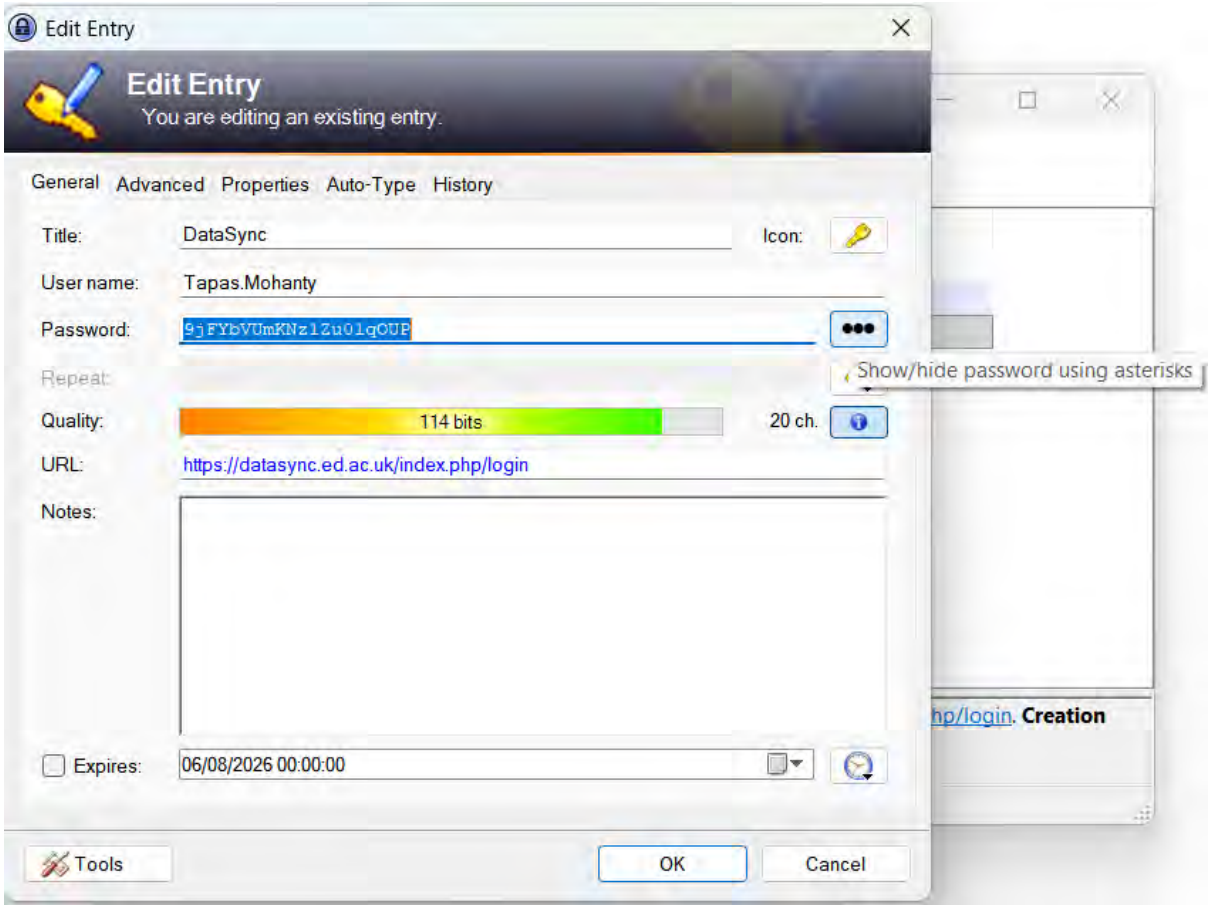
7. Add entries for each system using the Add Entry button



8. Use the built-in Password Generator for each new password



9. The password can be copied by clicking the required entry → Show/hide password using asterisks



10.7 FreeFileSync for Offline Backup

FreeFileSync is a free and open-source tool for comparing and synchronising folders. It can help maintain an offline copy of working data on an external drive or another approved storage location.

Mode	Behaviour	Best suited for
Two Way	**Copies changes in both directions, including deletions**	**Two equally important working locations**
Mirror	**Makes the right folder match the left; files absent from the left may be deleted from the right**	**Maintaining an exact replica**
Update	**Copies new and changed files from left to right without deleting files from the right**	**One-directional copying with protection against source-file deletion**

Basic setup

1. Download and install FreeFileSync.
2. Select/ drag and drop a folder pair: Left = working data; Right = external drive or approved backup location.
3. Select Update mode if deletions from the working folder should not be replicated.
4. Click Compare and review the proposed actions.
5. Run the synchronisation manually or save it as a batch job for scheduling through Windows Task Scheduler.
6. Disconnect and securely store the external drive after completion, where appropriate.

Important: *Synchronisation is not a complete backup strategy. Changed, corrupted or encrypted files may overwrite good copies. Maintain separate versioned backups and follow institutional security requirements and the project's Data Management Plan. Use FreeFileSync only where institutionally approved.*

10.7 Glossary – Jargon / Look up any term

This selective glossary covers the handbook's most important concepts. References identify the relevant chapter or section.

Term	Plain-language definition	Chapter/section
3-2-1 backup rule	A strategy that keeps three copies of data, on two different media or systems, with one copy stored at a separate location.	4.3
Anonymisation	Removing or altering information so that a person is no longer reasonably identifiable, considering the data, context and other information that may be available.	3.3.3, 6.4.1, 6.5

Artificial intelligence (AI)	The broad field concerned with computer systems that perform tasks such as recognising images, processing language and making recommendations.	9.1
Audit trail	A chronological record of what was changed, who made the change and when it was made.	4.1.3, 4.5, 6.7.1
Authoritative source	The paper or electronic record formally designated as the official version when information exists in more than one format.	7.5
Backup	A separate, recoverable copy of data maintained to protect against deletion, corruption, equipment failure or other loss.	4.3, 7.8
CARE Principles	Principles for Indigenous Data Governance: Collective Benefit, Authority to Control, Responsibility and Ethics.	7.7
CC BY 4.0	A Creative Commons licence that permits others to use, adapt and share a work for any purpose, provided appropriate credit is given.	3.3.4, 8.3
Confidentiality	The obligation to protect information entrusted to researchers from unauthorised access or disclosure.	3.3.2, 8.8
Consent to participate	A participant's voluntary agreement to take part in research after receiving clear and accessible information about what participation involves.	3.3.2
Context-appropriate practice	A data-management approach adapted to actual local conditions while maintaining proportionate standards for privacy, security, ethics and data quality.	7, 7.1, 7.2
Controlled or managed access	A sharing arrangement in which users must apply, receive approval and follow specified conditions before accessing data.	6.5, 8.2
Cross-border data transfer	Sending, storing, remotely accessing or otherwise making personal data available in another country.	6.8, 7.10
Data Champion	A local point of contact who helps research teams plan, organise, document, protect, quality-assure, preserve and share data responsibly.	2.1, 2.2
Data cleaning	Identifying and addressing problems such as duplicate records, missing	4.6

	values, inconsistent entries and implausible values.	
Data dictionary	A document defining each variable in a dataset, including its name, meaning, type, permitted values, units and missing-data codes.	3.3.3, 4.6.1, 10.4
Data governance	The rules, roles, responsibilities and decision-making arrangements that determine how data are managed, accessed and protected.	3.1.2, 3.3.1
Data harmonisation	Aligning data from different sites, studies or systems so that they can be consistently combined and compared.	4.6, 7.9
Data integrity	The accuracy, completeness and consistency of data, including protection against improper or undocumented changes.	4.5, 6.7.1, 8.8
Data lifecycle	The stages through which research data move, from planning and collection to analysis, sharing, preservation and reuse.	3.3, 4.1
Data Management Plan (DMP)	A living document explaining how data will be collected, documented, stored, protected, quality-assured, shared, preserved and securely destroyed.	3.3.1, 4.1.1, 10.2
Data minimisation	Collecting, using, transferring and retaining only the personal data genuinely needed for the research purpose.	3.3.2, 7.4, 7.10
Data protection	The legal and organisational framework governing how personal data are collected, used, stored, shared and deleted.	3.3.2, 6.8
Data quality	The extent to which data are accurate, complete, consistent, valid and suitable for their intended purpose.	4.5, 6.7.1, 7.9
Data repository	A managed service that stores, preserves and provides appropriate access to research data and supporting documentation.	4.1.4, 8.7
Data sharing	Making data available to authorised colleagues, other researchers or the public under appropriate legal, ethical and security conditions.	3.3.4, 6.5, 8.7
Data wrangling	Restructuring or transforming data into a form suitable for analysis, such as	4.6

	changing data types, renaming variables or combining files.	
De-identification	A broad term for removing, replacing, generalising or obscuring information that could identify a person.	6.4.1
Derived data	Data produced by cleaning, transforming, combining or analysing original source data.	4.1.3, 6.1
Digital Object Identifier (DOI)	A persistent and unique identifier that makes a research output reliably findable and citable.	4.1.4, 8.2
Direct identifier	Information that directly identifies a person, such as a name, identification number or contact details.	3.3.2, 6.4.1
Disclosure risk	The possibility that shared data or research outputs could reveal confidential information or identify a person or community.	3.3.3, 6.5
Encryption	Converting data into an unreadable form that can be restored only with the correct key or password.	7.4, 7.8, 8.6
FAIR Principles	Principles stating that data should be Findable, Accessible, Interoperable and Reusable. FAIR data do not necessarily have to be publicly open.	3.3.4, 7.7, 8.2
File-naming convention	An agreed method for naming files consistently using meaningful elements such as participant code, date and version number.	4.4.2
Free and Open Source Software (FOSS)	Software whose source code is available under a licence that permits people to use, study, modify and redistribute it.	9.5
Hallucination in AI	A false, unsupported or fabricated output generated by an AI system, often expressed confidently.	9.3, 9.4
Hybrid data collection	A workflow that combines paper and electronic methods of collecting or managing data.	7.3, 7.5
Indirect identifier	Information that may identify someone when combined with other details, such as age, location, occupation and diagnosis.	3.3.2, 6.4.1
Informed consent	An ongoing process through which a person voluntarily agrees to participate after receiving understandable information about the research.	3.3.2, 4.1.2

Institutional approval	Formal permission from an appropriate institutional service to use a tool, system, workflow or data-transfer arrangement.	6.3, 7.4, 7.10, 10.1
Intellectual property (IP)	Legal rights associated with creations such as datasets, software, methods, devices, training materials and publications.	3.3.1
ISO 8601	An international date and time standard. The handbook recommends YYYY-MM-DD in datasets and documentation and YYYYMMDD in filenames.	4.4.2, 4.6.1
Large Language Model (LLM)	An AI system trained on large text collections to generate, summarise, classify or translate language.	9.3
Lawful basis	The legal justification required to process personal data. It is distinct from ethical consent to participate in research.	3.3.2, 7.10
Master key	A separately protected record that links participant codes or pseudonyms with participants' identities.	3.3.3, 6.4.1
Metadata	Descriptive information explaining what data contain, how they were produced and how they should be interpreted and used.	3.3.3, 8.2, 10.2
Minimum necessary access	Giving authorised users access only to the data needed for their role and only for as long as required.	4.2, 8.7
Missing-data code	An agreed value indicating why information is absent, such as not applicable, refused or unknown.	4.6.1, 6.7.1
Multi-factor authentication (MFA)	A login method requiring more than one form of verification, such as a password and an authentication-app code.	8.5.1, 8.8
Multi-site study	Research conducted across more than one location, institution or country using a shared or coordinated protocol.	4.6, 7
Offline-first data management	A workflow designed to collect and temporarily store data securely without continuous internet access, then transfer or synchronise them through an approved connection.	7.4
Open data	Data made available for others to access and reuse under clear terms, subject to legal, ethical and confidentiality requirements.	3.3.4, 8.1

Open licence	A legal statement giving others permission to use, adapt and share a work under specified conditions.	3.3.4, 8.3
Open science	An approach that aims to make research methods, publications, data, software and other outputs accessible and reusable where appropriate.	3.3.4, 8.1
OpenRefine	A desktop application used to explore, clean, transform and harmonise tabular data.	4.6.2, 10.3
Original or raw data	Data in the form first collected or received, before cleaning, transformation or analysis. The original version should normally remain unchanged.	4.1.3, 4.4.1
Personal data	Any information relating to an identified or reasonably identifiable living person.	3.3.2, 6.8
Preservation	Actions taken to keep research data secure, understandable and usable over the long term.	3.3.4, 4.1.5
Privacy	A person's interest in and right to appropriate control over access to their personal life and information.	3.3.2
Prospective harmonisation	Agreeing on shared variables, definitions, formats, units and codes before data collection begins.	4.1.1, 4.6, 7.9
Pseudonymisation	Replacing identifying information with codes while retaining a separately protected key that permits authorised re-identification. Pseudonymised data remain personal data.	3.3.3, 6.4.1
ReadMe file	A plain-text document stored with a dataset that explains its contents, structure, origins, processing, codes, limitations and conditions of use.	3.3.3, 10.2
REDCap	Research Electronic Data Capture: a secure, browser-based system for designing research databases, collecting data and maintaining audit trails.	4.6.2, 6.7.1, 10.4
Re-identification risk	The possibility that a person could be identified from data that have been de-identified, pseudonymised or thought to be anonymous.	3.3.3, 6.4.1, 6.5
Research Data Management (RDM)	The organisation, documentation, storage, protection, quality assurance, sharing and preservation of research data throughout a project.	3.2, 4

Role-Based Access Control (RBAC)	Restricting access to systems and data according to each person's authorised responsibilities.	4.2, 7.2, 7.8
Sensitive data	Data requiring enhanced protection because their misuse could cause significant harm, including health, genetic and biometric data and information about vulnerable populations.	3.3.2, 6.3
Standard Operating Procedure (SOP)	A documented set of instructions describing how a routine activity should be performed consistently.	5.3, 5.5, 7.1
Structured data	Data organised into predefined fields, usually rows and columns, such as survey responses or clinical measurements.	3.3.3, 6.1
Synchronisation	Copying changes between two locations. Synchronisation is not the same as an independent and recoverable backup.	4.2, 4.3, 7.4, 7.8
Unstructured data	Data without a fixed tabular structure, such as transcripts, field notes, photographs, audio and video.	3.3.3, 6.1
Validation	Applying rules or checks to confirm that data have the required type, format, range, completeness and logical consistency.	4.5, 6.7.1, 7.5
Version control	A systematic method for recording, identifying and managing changes to files, datasets or code over time.	4.4.2, 6.4, 6.6

References

Bhatt, P., Liu, J., Gong, Y., Wang, J., & Guo, Y. (2022). Emerging artificial intelligence–empowered mHealth: Scoping review. *JMIR mHealth and uHealth, 10*(6), e35053. https://doi.org/10.2196/35053

Chetwynd, E. (2024). Ethical use of artificial intelligence for scientific writing: Current trends. *Journal of Human Lactation, 40*(2), 211–215. https://doi.org/10.1177/08903344241235160

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2019). *Managing and sharing research data: A guide to good practice*. SAGE Publications.

Curty, R. (2025, May 9). *Tidying messy spreadsheets with OpenRefine*. UCSB Library Research Data Services. https://ucsb-library-research-data-services.github.io/openrefine/

Evans, L., Evans, J., Pagliari, C., & Källander, K. (2023). Scoping review: Exploring the equity impact of current digital health design practices. *Oxford Open Digital Health, 1*, oqad006. https://doi.org/10.1093/oodh/oqad006

FreeFileSync. (n.d.). *FreeFileSync: Open-source file synchronisation and backup software*. Retrieved September 4, 2026, from https://freefilesync.org/

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open, 7*(10), e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Kyanzaire, L., & Johns, B. (2024). *GOSH community ambassador handbook*. Zenodo. https://doi.org/10.5281/zenodo.12684198

McKenna, J. (2024, December 3). *What is open science infrastructure?* MDPI Blog. https://mdpiblog.wordpress.sciforum.net/2024/12/03/open-science-infrastructure/

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, Article 0021. https://doi.org/10.1038/s41562-016-0021

National Cyber Security Centre. (n.d.). *Three random words*. https://www.ncsc.gov.uk/collection/top-tips-for-staying-secure-online/three-random-words

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2018). The preregistration revolution. *Nature Human Behaviour, 2*(3), 209–215. https://doi.org/10.1038/s41562-018-0469-4

OpenRefine. (2024). *OpenRefine* (Version 3.8) [Computer software]. https://openrefine.org/

SAGE Publishing. (n.d.). *Author guidelines on using generative AI and large language models*. https://learningresources.sagepub.com/author-guidelines-on-using-generative-ai-and-large-language-models

Seiner, R. S. (2014). *Non-invasive data governance: The path of least resistance and greatest success*. Technics Publications.

Stam, A., & Diaz, P. (2023). *Qualitative data anonymisation: Theoretical and practical considerations for anonymising interview transcripts* (FORS Guide No. 20, Version 1.0). Swiss Centre of Expertise in the Social Sciences. https://doi.org/10.24449/FG-2023-00020

UK Data Service. (n.d.). *Legal and ethical issues in sharing data*. https://ukdataservice.ac.uk/learning-hub/research-data-management/

UK Data Service. (n.d.). *What is the Five Safes framework?* https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/

UNESCO. (2023). *UNESCO open science toolkit*. https://unesdoc.unesco.org/ark:/48223/pf0000387983

University of Edinburgh. (n.d.). *Passwords*. https://infosec.ed.ac.uk/how-to-protect/lock-your-devices/passwords

University of Edinburgh. (n.d.). *Research data MANTRA*. https://mantra.ed.ac.uk/

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018. https://doi.org/10.1038/sdata.2016.18

usher.ed.ac.uk/respire
@RESPIREGlobal
RESPIRE@ed.ac.uk

Stay Updated

This is Version 0.8 (03 Sep 2026) of the handbook.

Good data practice is constantly evolving. The RESPIRE Data Champions Initiative regularly updates this handbook with new case studies, updated legal guidance, and refined templates.

Ensure you are using the most current guidance.

Scan the QR code below or visit https://edin.ac/4ybicu3 to download the latest version of **Building Research Data Management Capacity in LMICs: Development and implementation of the RESPIRE Data Champions Handbook**.

The NIHR Global Health Research Unit on Respiratory Health (RESPIRE) is funded by NIHR 16/136/109 and NIHR132826 using UK international development funding from the UK Government to support global health research. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the UK government.