

Document your research data

DOCUMENT YOUR RESEARCH DATA

Digital Essentials

THE UNIVERSITY OF QUEENSLAND LIBRARY

The University of Queensland



Document your research data Copyright © 2023 by The University of Queensland is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](#), except where otherwise noted.

This module is a derivative of the “[Tier Protocol](#)” by [Project Tier Teaching Integrity in Empirical Research](#), used under a [Creative Commons Attribution-NonCommercial 4.0 International License](#).

CONTENTS

Module overview	1
1. Process	3
2. Pre-data	6
3. Original data and metadata	9
4. Data processing	12
5. Data analysis	16
6. Reporting	18
7. Wrap-up	21
8. Check your knowledge	23
9. Conclusion	26

MODULE OVERVIEW

Aims and objectives

This module is to help you conduct transparent and reproducible quantitative research for your project. UQ supports research data that is [F.A.I.R](#) – **Findable, Accessible, Interoperable and Reusable**. Document what you have done to produce your results to make it possible to repeat it, in exactly the same way. The module is derived from the [Tier Protocol](#) by [Project Tier Teaching Integrity in Empirical Research](#).

This module will:

- explain how to document research data
- explain data organisation
- outline a workflow for your research project.

After completing this module, you will be able to:

- organise your documentation folders
- write a *Readme* file to show the changes you make to your data files
- use a process to ensure research data replication and transparency.

Module sections

1. [Process](#)
2. [Pre-data](#)
3. [Original data and metadata](#)
4. [Data processing](#)
5. [Data analysis](#)
6. [Reporting](#)
7. [Wrap-up](#)
8. [Check your knowledge](#)

Duration

Approximately 40 minutes

Student partnership

This module was developed with UQ student partners as part of a [Student-Staff Partnership](#) project.

Graduate attributes

Knowledge and skills you can gain from this module will contribute to your [Graduate attributes](#):



Influential communicators

This module is part of [Digital Essentials](#), a series of online modules to help you quickly build your digital skills so you can succeed in study and work.

Return to [UQ Library](#).

9. [Conclusion](#)

[Download versions](#) are available from the homepage.

1. PROCESS

Document your data, and the processing and analysis, to allow others to reproduce and replicate your research. This should be an integral part of your entire project. You should begin constructing your replication documentation before you start working with your data, and add to it incrementally as your research progresses. **Don't leave it until the end!**

This section explains the steps required to produce the replication documentation, during the three major phases of your research. Then the next sections provide the detailed explanations of the replication documentation required.

By the time you have finished the data processing and analysis for your project, and are ready to write the final research project, only a few additions and checks should be required to produce comprehensive replication documentation.

Three major phases of your research

Pre-data

This is the preparation phase, before you start working with your data. You should **create the relevant folders and blank documents** which will be used extensively in the next two phases — Data work and Wrap up.

Data work

In this phase, you have already started your research and will be managing your research data. The Data work phase covers:

- storing and documenting your **original data**
- **processing** the data
- **analysing** the data
- **reporting** on your research and how it can be replicated.

Wrap-up

If you have followed the steps in the previous two phases, you will have constructed your replication documentation incrementally throughout the course of your work on the project. In this phase, you have nearly completed your research and will just need to:

1. **Finish** the final documentation
2. **Proof-read**
3. Make a **final check** of your replication document.

Saving your work

It is important to choose a **safe place** to save your work and to **back up** your work regularly:

- [Working with data and files](#) covers options for data storage and backing up your work
- [Data at UQ](#) has information on data policies and responsibilities at UQ



If you are conducting research as part of a Winter/Summer Research project, Honours or Masters degree or Higher Degree by Research (HDR):

- See [Manage research data](#) to find out how to manage, store and secure your project's data
- Discuss with your supervisor or coordinator the suitability of using [UQ Research Data Manager \(UQRDM\)](#) to manage your research data. UQRDM will:
 - provide you with storage for your project's research data
 - allow you to share and collaborate with colleagues
 - enable you to plan and manage your project's research data.

It is important to use stable, well documented [file formats](#). Convert your files into another format, if necessary, to make it possible to access the file, read the file or reuse the data in the long term.



Export your plan

As you work through the module you can record your own project plan in this document and download it.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://uq.pressbooks.pub/digital-essentials-document-research-data/?p=35#h5p-1>

The next sections provide an in-depth explanation of the relevant steps you need to do in each phase.

2. PRE-DATA

In the **pre-data phase**, you need to create:

1. A hierarchy of empty folders
2. Three blank documents.

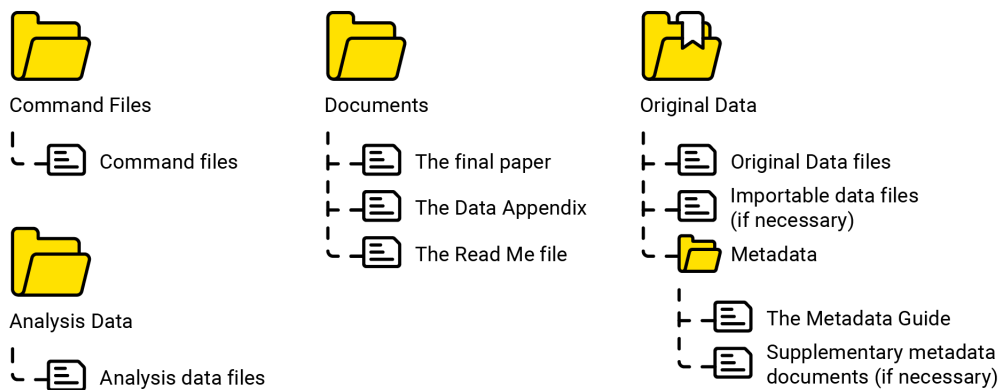
Construct a hierarchy of empty folders

There are four folders to create:

- **Original Data folder** contains:
 - Your original data files
 - Importable data files (if necessary)
 - Metadata subfolder
 - A document called your Metadata Guide
 - Supplementary documents with additional metadata (if necessary)
- **Command File folder** contains:
 - one or more files with the commands used to process and analyse the data
 - command files written in the syntax of the statistical software you use for the study
- **Analysis Data folder** contains:
 - the fully cleaned and processed data files that you use to generate the results
- **Documents folder** contains:
 - A copy of your final paper
 - Your Data Appendix
 - Your Readme file

These four folders can be stored into a **main folder**. The contents of each folder are described in detail in the following sections. The steps undertaken within each folder should be done in the **Data Work phase**.

Home Drive (C:) > ECON2003 Research Project-Group B



Your main folder structure:
Command file folder, Analysis data folder, Documents folder and Original data folder.



It is important to give your main folder an appropriate name. For example:

- “ECON2003 Research Project-Group B”
- “J. Smith Senior Thesis”



Hierarchy of folders



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://uq.pressbooks.pub/digital-essentials-document-research-data/?p=40#h5p-2>

Create three blank documents

1. **Create 3 blank documents** that will become part of your replication documentation

- The three documents to create are:
 - A **Metadata guide** — saved to your Metadata folder (which is a sub-folder of your Original Data folder)
 - A **Data appendix** — saved to your Documents folder
 - A **Readme** file — saved to the Documents folder.

2. **Give the files appropriate names:**

- Metadata guide — **MetadataGuide**
- Data appendix — **DataAppendix**
- Readme file — **Readme**

3. **Type an appropriate title** at the beginning of each document

- Examples titles:
 - “Metadata Guide for J. Student Senior Thesis.”
 - “Data Appendix for ‘The Economics and Politics of Popular Stringed Instruments,’ by U. K. Laylee.”
 - “Readme File for Econometrics Project by A. Smith and B. Jones.”

Other than their titles, **leave these documents blank** until the appropriate step in the Data work phase.

3. ORIGINAL DATA AND METADATA

The Original Data folder contains:

1. Your original data files
2. Importable data files (if necessary)
3. Metadata subfolder.

1. Original data files

Original data files are the data files you initially get for your project and from which you extract the data you use. These data files can come from a single original data file or from multiple original data files.

Each time you obtain an original data file, save a copy of the original data file in your Original Data folder. It is important to note that:

- A. You give the file a new name when you save it in the Original Data folder but, other than that, the copy you save should be identical to the original version of the data file
- B. The contents and format should not be modified in any way.

2. Importable Data Files

Create a version of your data file in **a format that your software or tool can read.**

When you need to create a new version of a data file, to make it possible for your software to read it, the new version is called an **importable data file**. Changes you make to the data, when you create an importable version of an original data file, cannot be executed by commands written in a command file. Therefore, they cannot be automatically reproduced. Make only the minimal changes necessary to make it possible for your software to read the data.

- A. You should not modify the data in the file in any way
- B. Give the importable version of the file a name that reminds you it is the importable version of the original data file from which it was created. For example, if the original data file is called *gdp_growth.sav*, give the importable version a name like *i_gdp_growth.dta*. The “i_” prefix is a reminder that the file is “importable” and the change in the extension reflects the change in the format in which the data file is saved.

When storing the importable data file, do not delete the original data files. Store both versions in the folder using different names.

Note: If all of your original data files are in formats that your software can read, you do **not** need to create any importable data files.

Readme file

For each importable data file you create, write an explanation in your Readme file describing the steps you took to create the importable version from the original data file. This may be the first time you enter any information (other than the title) in your Readme file.

Your written explanation should give the names of both the original and importable versions of the data file that was modified. It should be precise enough to enable others to make the same changes to the original data file and end up with an importable data file identical to the one you created.

[Refer to Section 6](#) for more information about what is included in the Readme file.

3. The Metadata subfolder

The Metadata subfolder contains:

- A document called your Metadata Guide
- Supplementary documents with additional metadata (if necessary).

The Metadata guide

For each of your original data files, the Metadata Guide provides the kind of information such as variable definitions and coding, sampling methods, and anything else a user would need to know to work with and interpret the data appropriately. Each time you obtain a new original:

- A. Add a section about that file to the Metadata Guide
- B. Begin the section with a title that identifies the original data file it pertains to. For example, ‘Metadata for penn_tables_1986_2010.txt’
- C. Enter the relevant information about the original data file in your Metadata Guide. The information included in the Metadata Guide should include:
 - **A bibliographic citation** — this citation should be in a format consistent with

the [referencing style](#) (e.g. APA or Chicago) used in the research project

- **A Digital Object Identifier (DOI), if one has been assigned** — if a [DOI is included in the bibliographic citation](#), it need not be repeated
- **The date the author first downloaded or obtained the original data file** — if the download date is included in the bibliographic citation, it need not be repeated
- **A verbal explanation of how others can obtain a copy of the original data file** — in many cases, this explanation will give the URL of a webpage from which the data can be accessed, along with instructions for downloading a file identical to the original data file used in the study. In all cases, this explanation should be complete and precise enough to allow another researcher to locate and obtain an exact copy of the original data file, without any additional information or assistance
- **Whatever additional information others would need to understand and use the data in the original data file** — this may vary depending on the nature of the original data file. In many cases, the additional information that should be provided are variable names and definitions, coding schemes and units of measurement, details of the sampling method and weight variables, and descriptions of how any imputed variables were constructed. In some cases, it is also necessary to include information about the file structure (e.g. the delimiters used to separate variables or, in rectangular files without delimiters, the columns in which the variables are stored). Any other unique or idiosyncratic aspects of the data, that an independent user of the data would need to understand, should be explained.

Supplementary documentation

In many cases, some or all of the information that should be included in the Metadata Guide may be available in an existing, publicly accessible document, such as a codebook or user's guide that is provided with the original data file. If the information is accessible, it is not necessary to include that information in the Metadata Guide. Instead, you can put a note in the Metadata Guide indicating that the information is available in an existing document.

When you put a note in the Metadata Guide, indicating that certain parts of the information are available in an existing document, you should preserve a copy of the existing document in the Metadata sub-folder (along with the Metadata Guide that you compose yourself).

If you enter the required information right away, each time you obtain a new original data file, your Metadata Guide will be complete before you finish your project.

4. DATA PROCESSING

The processing phase of a project consists of all the steps involved in transforming your original data files (or the importable versions) into the fully cleaned and processed analysis data files that you use to generate your results.

If you are not using [statistical software for your analysis](#), you may not need to perform all the steps outlined here but it is still important to document how you have cleaned and processed the data. Remember to use stable, well documented [file formats](#) to make it possible to access the file or reuse the data in the long term.



The Library regularly has [training sessions](#) on various statistical software packages, including R and RStudio.

Command files folder

If you are using **statistical software**, the Command files folder should contain one or more files with code written in the syntax of the software you use for the study. The code in these command files should execute all the data processing and analysis necessary to replicate the study and reproduce the reported results.

The best way to construct and organise your command files may vary depending on the nature of the project or the requirements of the software. In many cases, you will be able to organise it into three steps, with one or more command files executing the steps in each phase.

Step 1: Processing the data using statistical software

If you are using statistical software, all of the commands necessary for processing your data must be written in a command file, or in several command files that can be run sequentially. When you have finished writing these command files, executing them will automatically conduct all the procedures necessary to transform your original (or importable) data files into your analysis data files.

Writing, experimenting with and editing these command files, until they successfully carry out the necessary steps of processing, is the focal point of the work you do with your data

Transform your data file into analysis data files

In one or more command files, write code that transforms your importable data files into your analysis data files. Exactly what steps of processing are required varies but some common procedures include:

- Having your software open your importable data files
- Cleaning the data to resolve any errors or discrepancies
- Removing variables or cases that you do not need
- Combining data from different importable data files
- Transposing a data table so that columns become rows and rows become columns
- Generating new variables
- Saving intermediate and analysis data files.

Decide how to organise the code that processes your data into one or more command files

It is possible to put all the necessary commands in a single file. However, in many cases separating different parts of the processing phase into different command files can help you keep track of what you are doing. The best way of dividing your data processing among command files will depend on the particulars of your project but the following scheme often works well:

- A. For every importable data file you have, write one command file that reads and cleans the data it contains. This ensures that the data is prepared for merging with the data from the other importable data files. Then save them in a new file, in the native format of the software you are using
- B. Then write one additional command file that merges all these natively formatted files, processes them as necessary to construct the analysis data files and then saves the analysis data files in your software's native format
- C. Depending on the number of importable data files you have and how the data in them is organised, other schemes for dividing the processing phase among your command files may be more convenient. You should use whatever scheme you find works best for your project.

Whatever scheme you choose, explain in your **Readme** file ([section 6](#)) the order in which the command files need to be run to replicate your project.



Save your command files and your analysis data files in the appropriate folders:

- A. The command files that process your data and create your analysis data files in your Command files folder
- B. The analysis data files in your Analysis data folder (The Analysis data folder is explained in [section 5](#)).

Step 2: Constructing the Data appendix

The data appendix, saved in the Documents folder, is one of the three documents you created before you began working with your data, in the pre-data phase.

Construct your data appendix after you have finished writing the command files that create your analysis data files and **before you begin your analysis**, as you may learn things about your data that you should know before you start the analysis.

The data appendix:

- should provide information about every variable in your analysis data files, including names, definitions and coding (for all variables), [summary statistics](#) and [histograms](#) (for quantitative variables), and relative [frequency tables](#) and [charts](#) (for categorical variables).
- serves as a codebook and users' guide for your analysis data files.

For **every variable**, include:

- The name of the variable and a complete definition (e.g. coding or units of measurement, the wording of a survey question the variable is based on or adjustments made for inflation)
- The name of the original data file from which the variable was extracted, or from which the variables used to construct it were extracted, and the names of the variables extracted from the original data files
- The number of observations with valid values for the variable, and the number of observations with missing values.

Generate the descriptive statistics, tables and figures

Writing a command file generates all the descriptive statistics, tables and figures needed for the data appendix. These should be created using the data in your analysis data files.

1. Give this command file the name *DataAppendix*
2. Save *DataAppendix* in your Command files folder
3. Finish composing the data appendix, inserting the descriptive statistics, tables and figures in the appropriate places
4. When you have finished, **save the data appendix in your Documents folder**. Refer to Section 6 for more information about the Documents folder.

Step 3: Generating the results

Using the data in the analysis data files, the command files conduct the procedures that generate the results reported in the research project. Before each command that generates any of the results, a comment indicating which results it produces should be mentioned (e.g. by table, or figure number or page on which the numerical result appears).

Even if you deviate from this scheme, it provides a useful framework to begin thinking about the most effective way to organise your command files.

5. DATA ANALYSIS

Analysing your data

Generate the figures, tables and other statistical results from your analysis data files using the appropriate [methods and tools](#). The results consist of all the findings you report in your research project that are based on computations performed on your analysis data files. They may be presented in various forms, including tables, figures and numerical values reported in the text of the research project.

As in the processing phase, composing command files that execute all the necessary procedures is important.

Write code that generates all the results you report in your research project

- One or more command files should contain commands that open up your analysis data files and then use that data to generate the output upon which your results are based.
- Every command that generates any of your results should be preceded by a comment that states which result the command generates.

Example comments

* The following command generates the first column of Table 6.

* The following command generates the second column of Table 6.

* The following command generates Figure 4.

```
/* The following command generates the correlation of 0.31 between the variables INC  
(individual annual income, reported in the natural log of current US dollars) and SATIS  
(individual subjective self-report of overall satisfaction with life, on a scale of 0—least  
satisfied—to 10—most satisfied). This correlation is reported on page 27 of the research project.  
*/
```

The **command files for your analysis phase** should not contain any commands that generate new

variables or process your data in any way. All the procedures required to prepare your data for analysis should be executed by the **command files you wrote for the process phase**.

It is often convenient to write all the commands for the analysis phase in a single command file. However, if you think it would make sense, divide the code that generates the results into two or more command files. No matter how you organise your analysis command files, your **Readme** file will include an explanation of how to use them to reproduce your results.



Save the command files you write for the analysis phase in the Command files folder.

Analysis data folder

The fully [cleaned and processed data](#) files that you use to generate the results reported in your research project are called analysis data files. They are typically constructed by cleaning, processing and combining data extracted from one or more original data files.

A copy of every analysis data file used for a study should be preserved in the Analysis data folder.

Use a single cleaned and processed data file to generate all the results reported in the research project. This ensures that just one analysis data file needs to be stored in the Analysis data folder. When results are generated from more than one analysis data file, they should all be included.

6. REPORTING

The Documents folder contains:

1. A copy of your final research project
2. Your data appendix
3. Your Readme file.

Formats



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://uq.pressbooks.pub/digital-essentials-document-research-data/?p=66#h5p-3>

1. The Final research project

Save a full text PDF copy of your final research project to the Documents folder, once you have completed the write up. Saving a copy of your research project in PDF format helps prevent the document from being changed accidentally or corrupted in some other way.

2. The Data appendix

The Data appendix is a document that serves as a codebook for the analysis data files. [Refer to section 4](#) for more information about creating the Data appendix.

3. The Readme file

Write an explanation in your Readme file

The **Readme** file describes the files included in the replication documentation and explains how they can be used to replicate the study and reproduce the results. You, the author of the study, write this document.

The Readme file consists of three main sections:

- A. The contents of the replication documentation
- B. Modifications made to importable data files (if necessary)
- C. Instructions for replicating the study

A. The contents of the replication documentation

This section describes all the files included in the replication documentation and outlines the structure of the folders in which they are stored.

B. Modifications made to importable data files (if necessary)

This section should explain the changes you made to your importable data files. Refer back to [section 3 for more information on importable data files](#). If no modifications were made to the original data file, you may simply omit this section from the Readme file. (In that case, you should call the last section of the Readme file section B, instead of section C.)

C. Instructions for replicating the study

This final section gives instructions for using the replication documentation to replicate the data processing and analysis conducted for the research project, to reproduce the reported results.

These instructions should:

- State what statistical software (including version number and required add-ons) are required to run the command files
- Explain which files included in the replication documentation need to be copied onto the replicator's computer, the structure of folders and sub-folders in which the files should be copied and which of the folders each file should be saved in
- Indicate which of the folders should be set as the working directory, when the statistical software that executes the command files is run
- Indicate the order in which the command files need to be run to carry out the replication. For each command file, indicate what other files it uses (e.g. what data files it opens and what other command files it calls) and what output it produces (e.g. new data files it saves, old data files it deletes and new

directories it creates).

7. WRAP-UP

By following this process, you can construct your replication documentation incrementally throughout the course of your work on the project. By the time you have completed your final research project, your replication documentation should also be nearly complete.

Finishing your Readme file

By this point, you should already have recorded notes explaining any modifications you made to the original data files, when you made importable versions of them.

To finish your **Readme** file, you should add:

1. An overview of all the files included in the replication documentation
2. The structure of the folders in which they are stored
3. Step-by-step instructions for using the replication documentation to replicate the study.

Proofreading and testing your Command files

In the course of your project, you constructed command files that processed your data, produced the descriptive statistics and figures for your data appendix, and executed the analyses and procedures that generated the results you reported in your research project.

Before you consider your command files to be complete and ready to store in your final replication documentation, you should edit them for accuracy and clarity and test them to be sure they reproduce the results of your project, as intended.

Editing your Command files

Edit all your command files to be sure they are accurate and concise:

1. Remove any commands that turned out not to be necessary for your project
2. If you realise that the code you wrote to execute any of the procedures could be rewritten in a simpler or more streamlined way, then revise the code accordingly
3. Ensure the comments in your command files are extensive and clear enough to allow others to understand what is accomplished in each step of data processing and analysis.

Testing your Command files

Test your command files to ensure that they all run without error and that they successfully reproduce the results you reported in your research project.

1. Try following the instructions for replicating your project that you wrote in the **Readme** file to ensure all your command files run without a problem and produce the intended output
2. If you encounter any errors or crashes, diagnose and fix the problem, and then start the test over.

Make a final check of all your replication documentation

Before you consider your replication documentation to be complete and final, check that it satisfies all the requirements.

Check that:

1. All the required files are included in your replication documentation
2. They are stored in the correct folders
3. The content and format of every file meets the specifications
4. Finally, delete any unnecessary files. Your replication documentation should contain only the files specified, or that you intentionally chose to include for a particular purpose.

After you check that your replication documentation contains everything that it should, and that it does not contain anything unnecessary, it is **completed**.

8. CHECK YOUR KNOWLEDGE

All the answers to the quiz questions can be found in this module.

Your response data will **not** be gathered. [Take a screenshot](#) of **Your result** at the end of the quiz if you are required to show you have completed it.

Document your research data quiz

There are 6 questions to answer. After you answer a question, click the arrow to move to the next question.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://uq.pressbooks.pub/digital-essentials-document-research-data/?p=72#h5p-4>

Note for Teaching staff: You can [embed the module](#) and the [H5P quiz](#) in your courses.

Module summary

1 Process

An explanation of the steps required to produce the replication documentation, during the three major phases of your research.

2 Pre-data

- The preparation phase, before you start working with your data.

- Create the relevant folders and blank documents which will be used extensively in the next two phases.

③ The original data and metadata

- In this Data work phase, you manage your research data.
- Each time you get an original data file, save a copy of the original data file in your Original Data folder.
- Enter relevant information in the Readme file and Metadata guide.

④ Data processing

- The Command file folder contains one or more files with code written in the syntax of the statistical software you use for your project.
- There are three steps in this Data work phase — processing the data, constructing the data appendix and generating the results.

⑤ Data analysis

- Perform the procedures on your analysis data files that generate the figures, tables and other statistical results you report in your research project.
- Write code that generates all the results you report in your research project.
- Save the command files you write for the analysis phase in the Command files folder and the analysis data files in your Analysis data folder.

⑥ Reporting

- Save a full-text PDF copy of your final research project to the Documents folder, once you have completed the write up.
- The Data appendix, that serves as a codebook for the analysis data files, should be saved in the Document folder.
- The Readme file, that describes the files included in the replication documentation and explains how they can be used to replicate the study and reproduce the results, should be saved in the Document folder.

7 Wrap-up

- Complete the replication documentation you have constructed incrementally, throughout the course of your work on the project.
- Finish your Readme file.
- Make a final check of all your replication documentation.

9. CONCLUSION



You have completed the **Document your research data** module.

Tell us what you think

Use our [Digital Learning feedback form](#) to give **anonymous** feedback on this module. You can provide your email if you would like us to reply to you.

Digital Essentials modules

Build your digital skills with Digital Essentials. Select modules from the 6 themes that match your interests and will help you succeed in study and work.



New to UQ?

- [Getting started at the UQ Library](#)

- [Use UQ systems](#), includes:
 - [Book rooms](#)
 - [Printing at UQ.](#)



Assignment essentials

- [Find and use media](#)
- [Information essentials](#)
- [Write, cite and submit](#)
- [Types of assignments](#)



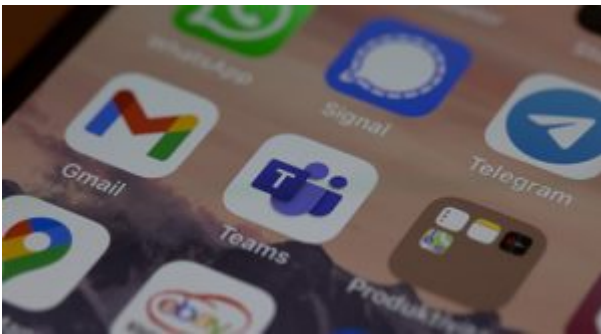
Digital creation

- [Accessibility](#)
- [Choose the right tool](#)
- [Intellectual Property](#)



Digital security and safety

- [Digital security](#)
- [Internet essentials](#)
- [Social media](#)



Professional identity and skills

- [Communicate and collaborate](#)
- [eProfessionalism](#)



Work with information

- [Artificial Intelligence](#)

- [Work with data and files](#)
- [Document your research data](#)

Teaching staff – use the modules in courses

Teaching staff can embed or link the modules in courses to help build your students' digital literacy.

- There are interactive elements throughout each module and a short H5P quiz at the end.
- The modules are also available for your students to download in EPUB, PDF and HTML format to make them more accessible.

Learn how to [add the Pressbook module](#) to your Learn.UQ (Blackboard) course.

Assess student learning

Most modules have a final short quiz created in H5P. You can download the H5P quiz from the module and embed it in your course if you would like to check your students' completions or to allow for the results to be transferred to the Grade Centre in your course.

The H5P quiz content will not record any completion data unless you download and add the H5P quiz directly to your course. Students can screenshot the quiz if they are required to show completion.

Learn how to [add the H5P content](#) to your course.