# Note for Principal Investigators on SFI Data Provenance Reviews

**Science Foundation Ireland is committed to ensuring that the research it funds is conducted to the highest standards of professionalism and rigour, maintaining Ireland's reputation as a leader in research excellence. In support of continuous improvement, SFI is currently incorporating innovative processes that will further promote good research practice, research integrity and reproducibility into its reviews. One such process is 'data provenance' where a portion of the mid-term progress review is allocated for a subject-specific expert panel (external) to study the provenance of a dataset and engage with the research team on matters concerning training, mentoring and supervision along with procedures used for data capture, analysis, storage and curation.**

**This bottom-up process helps to ensure the highest standards of integrity in all aspects of SFI funded research, and to embed a culture based on strong principles of good research practice**

**(From: Science Foundation Ireland 2019 https://www.sfi.ie/funding/sfi-policies-and-guidance/integrity/)**

To capture this information you should address the following. For the SFI presentation, you can refer to the following in your slides.

**Section 1:**
Data provenance requires that "a culture based on strong principles of good research practice" is established and fostered within research groups. As part of your data provenance review you should give details and evidence of you and your group's commitment to research integrity and the responsible conduct of research. How are you ensuring that good research practice in relation to data underpins all of your outputs?

To capture this information in a document you should consider the following.

1. **Responsibilities and Resources**
   Outline who within your team is responsible for oversight of good data practices and data management activities. What resources do you currently use to ensure that effective data management is an integral part of your research process?

2. **Team Research Integrity Training**
   State the level of training your team has taken in the area of research integrity and responsible conduct in research. Have your team members completed the EPIGEUM online course in research integrity or have they participated in the UCC Digital Badge in Responsible Conduct in Research. How are data managed and discussed within the team? It should be clear that the PI sees all original data and is directly involved in identifying data which underpin publication and accordingly, is directly involved in any data which are not included and the rationale for this. Good practice involves all members of a team having access to data and that data are openly discussed at meetings. All members of the team should be encouraged to raise issues in relation to interpretation of data in a constructive debate.

**Section 2:**
The best way to capture the technical aspects of data provenance "the data capture, analysis, storage and curation" is to use a data management plan. Data provenance describes the origins and the history of your data and explains exactly how they were obtained. Data provenance largely consists of contextual information that is an important record of how primary research findings were produced

or developed. By giving details of the experimental set-up, when, how and by whom data were collected, created or collated, how they were processed you are ensuring that they can be reused, replicated or reproduced, an essential aspect of research integrity In your presentation it should be clear that all data can be traced to it origin and that it is fully supported by verifiable documentation.

To capture this information, you should consider the following headings. Not all sections will apply equally to all projects or disciplines but each heading should be considered when drafting your Data Provenance Review.

### 1. Data description and Origins of the data

Give detailed description of data, including file formats and volume
How the data were created, collected, collated or derived? The experimental set-up or methodology, date of collection, instrument or software. If you are using secondary data from a third party or another research project clearly identify this and give details how the data were created, collected, collated or derived? How have the data been processed/transformed since capture? What are the procedures around this? It is recognised that not all data is digital, in the case of hardcopy data detailed verified records must also be maintained for example lab notebooks.

### 2. Documentation and metadata

Provenance of your data is going to be largely informed by the metadata associated with your data output, so good documentation of metadata and supporting information is key. Describe the file and folder structure, it is best practice to have consistent well-ordered research data. Give details of any metadata standards, controlled vocabularies used. The documentation needed to support the primary research data includes but is not limited to; lab notebooks, methodologies, data dictionaries, analytical and procedural information, definition of variables, units of measurements etc. Consider how the supporting documentation and metadata are captured and where it will be recorded for example in a database with links to each item or a ‘readme’ file or file headers, code books or lab notebooks.

### 3. Data Quality

Explain how consistency and quality of data is maintained, this can include measures such as calibration, repeated samples or measurements, standardised data capture, data entry validation, peer review of data, naming convention on files and folders, version control and read only raw data master files.

### 4. Storage and Backup of Active Research Data

How are data and metadata stored and backed up during the research process? How many copies do you maintain? Is back up automated or scheduled? What services do you use to back-up your data? What data protection requirements apply to your data? How do you manage these risks? Who has access to the data? How access is appropriately controlled? Do you need to have an access procedure? Do you have a disaster recovery plan? Where and how is hardcopy data (lab notebooks) securely stored and maintained?

### 5. Legal and ethical requirements, codes of conduct

What legal and ethical requirements is your data subject too? If personal data are processed, how is compliance with GDPR and security ensured? How are legal issues such as IP right or ownership of the data being addressed? Are there restrictions on the use of third-party data? What ethical and codes of conducts are there? How have you implemented their requirements? Which Research Integrity Codes do you promote within your team? Are you and your team familiar with the UCC Code of Research Conduct and are your team aware of the National Policy Statement on Research Integrity? If relevant, which UCC Ethics Committee did you obtain consent from with respect to your study?

6. **Data sharing and long-term preservation (FAIR)**
   Data sharing and long-term preservation ensure the transparency and reproducibility of the primary research findings. Where appropriate and possible, any data underlying a publication should be considered for sharing and long-term preservation in a data repository. What steps have you taken to ensure that your data can be made FAIR either during the research process or post-project? Do you have a plan for the long-term preservation of your data to ensure that your primary findings are reproducible and transparent?