SFI Public Service Fellowship 2023

1. **Name of Governmental Department or Agency**
Central Statistics Office (CSO)
2. **Title of the Project**
CSO2 Development of a data linking and integration methodology for administrative data in the public sector
3. **Description of the Project**
The National Data Infrastructure (NDI) is concerned with the consistent and reliable identification of data that relates to a: • Location - Eircode • Person – PPSN • Business – Unique Business Identifier (UBI) The consistent identification of these core data assets is crucial to successfully linking and integrating administrative data. The Administrative Data Centre (ADC) holds a significant amount of data from government departments and agencies across the state. The linkage and integration of these sources will deliver better insight for businesses, policy makers and citizens of the state. It will also contribute to a more efficient virtual data rooms service. Hence, the ADC is moving to create a sound methodology which will enable the linkage and integration of data in an efficient and safe way. A data linkage and integration methodology is essential for ensuring that the resulting dataset is accurate, reliable, and suitable for analysis. By following a systematic approach, we can ensure that the data is of high quality and can be trusted to inform decision-making. The **SFI Public Service Fellowship Researcher would be responsible for developing ADC's data linkage and integration methodology.** The researcher will be based in the ADC but would work cross functionally with our Quality and Methodology areas as well as our data suppliers across the public sector. The researcher would also have an international network of National Statistical Institutes (NSIs) and Eurostat to engage with for support and advice. We also have collaborations with universities, and these could be relied on for advice and support also.
4. **Project Scope**
A data linkage and integration methodology is a systematic approach to connecting and merging data from multiple sources. The researcher will be responsible for developing the methodology and guidelines which should include the following: 1. **Stakeholder engagement**: engage with our key stakeholders to keep them informed, to get their feedback and to deliver a better outcome.

2. **Data preparation**: The first step is to prepare the data for linkage and integration. This involves cleaning, transforming, and standardising the data to ensure that it is of high quality and can be easily matched with other datasets. Key questions here are - what kind of cleaning and data editing should be completed? What variables should be standardised? How frequently should data be updated? What limitations should we be aware of?

3. **Record linkage**: The next step is to identify matching records across different datasets. This involves comparing the data in each record to determine whether they refer to the same entity or event etc.

4. **Data integration:** Once matching records have been identified, the next step is to combine the data from the different sources into a single dataset. This may involve merging the data, aggregating it, or transforming it to create a unified view.

5. **Data validation**: After integrating the data, it will be important to validate the accuracy and completeness of the resulting dataset (s). This involves comparing the integrated data to other sources of information and verifying that it is consistent with what is known about the subject/area being studied.

6. **Documentation:** Finally, it is important to document the data linkage and integration methodology to ensure that the process can be replicated, and the resulting dataset can be easily understood by other users.

The methodology will be presented to the CSO's Confidential Data and Security Committee and the CSO Management Board for approval. The methodology can then be shared across the data ecosystem.

5. Skills/Expertise Required

The researcher will need to demonstrate the following skills/areas of expertise:

- Proficiency in using programming/scripting languages associated with statistical computing environment, in particular R or Python
- Technical statistical skills such knowledge or experience in data linkage and integration, data processing, and data modelling.
- Strong analytical skills to organise and analyse significant amounts of information with attention to detail and accuracy.
- Ability to work effectively on own initiative, within a team and cross functionally.
- Strong communication, report writing and presentation skills.

6. Expected Outputs of Project

Overall, at the end of the 12 months the key deliverables will be –

- **Research Paper** – detailing the methodology that should be followed
- **Interim reports --** on the 6 key areas identified in section 4
- **Case studies** – demonstrating the methodology in practice

- **Presentations** – delivered to the Management Board, the General Management Forum, the Formal Statisticians Liaison Group and other stakeholders.

7. **Working Arrangements**

The researcher would ideally be based in the offices of the CSO in Cork. Flexible and remote working arrangements will be accommodated.

8. **Expected Timeline**

The ADC project is expected to have a 12-months full-time timeline with the following components:

1. **Month 1, 2** - Project Initiation & Stakeholder Engagement phase
2. **Month 3, 4, 5** - Research and development phase
3. **Month 6, 7, 8** - Implementation phase
4. **Month 9, 10** – testing and validation phase
5. **Month 11, 12** – Approval and launch phase

9. **Contact Details**

Pamela Lafferty
Head of Division,
Administrative Data Centre
Central Statistics Office,
Cork