

SFI Public Service Fellowship 2023

1. Name of Governmental Department or Agency
Central Statistics Office
2. Title of the Project
CSO5 Semantic Open-Data using AI
3. Description of the Project
<p>The aim of the project is to use AI to improve how the public can search for and retrieve official statistics via public search engines.</p> <p>The CSO would like to engage a researcher to help leverage AI to improve the discovery and dissemination of its public statistical database using a semantic approach. Research will focus on how to improve and extend capabilities for searching and finding statistical data accurately without the need to drill into interactive tables. AI capabilities will be leveraged to obtain a semantic output which can be scaled for wider use – it will enable searching and finding data to be based on the meaning of a question rather than matching an association of words. This project will also create a prototype module for “training” AI engines on the semantics of data, allowing public AI systems to give more accurate answers to questions relating to official statistics.</p> <p>PxStat, the CSO’s public statistical database, is a key input to the national Open Data Portal which was developed as part of the Open Data strategy for Ireland. This research, by focusing on semantics, could help to overcome one of the major challenges on the journey to attaining 5-star open data status.</p>
4. Project Scope
<p>The main scope of the project is about creating a repository of AI Modules and APIs (Application Program Interface) using the OpenAI technology and the Semantic Kernel library to train AI systems in understanding and handling data, starting from Open-Data for Official Statistics. The repository will be also published as an Open-Source resource on GitHub to share the technology advancement with other organisations.</p> <p>The current Open-Data platform (PxStat) maintained by the CSO will be enhanced with the embedding of the newly created AI Modules to improve and extend the capabilities to search and find statistical data more accurately across the Web. The researcher will work with CSO development teams to optimise the modules.</p> <p>A request to include the AI Modules in public facing AI systems (ie. ChatGPT) will be made at project completion, so to allow the public to access more accurate data via generic search engines (i.e. Bing, ChatGPT etc...).</p>

5. Skills/Expertise Required
<p>The researcher should have relevant qualifications in the area of Computer Engineering or Computer Science.</p> <p>The project requires the following skills and roles:</p> <ol style="list-style-type: none"> 1. Expertise in AI 2. Development skills with OpenAI, Semantic Framework and Semantic Kernel library, APIs. 3. Knowledge of Open-Data formats 4. Code development skills, preferably C#
6. Expected Outputs of Project
<p>The project will generate these expected outputs:</p> <ol style="list-style-type: none"> 1. Repository of OpenAI Modules for training AI around the semantic of Open-Data, such as Official Statistics 2. Publishing of the repository under the MIT Open-Source License in GitHub
7. Working Arrangements
<p>The researcher would ideally be based in the offices of the CSO in Cork. Flexible and remote working arrangements will be accommodated.</p>
8. Expected Timeline
<p>The project will be delivered using an Agile methodology with the following milestones and an estimated duration of 6-12 months full-time.</p> <ol style="list-style-type: none"> 1. Concept 2. Design 3. MVP 4. PxStat Integration 5. Open-Source release
9. Contact Details
<p>Lorenzo Bruni Head of Technology Dissemination Central Statistics Office Mahon, Cork</p>