

# MSc in Computing (Information Technology)

**Dissertation Proposal Form**

The form, fully completed, must be returned to:

Brendan Tierney

School of Computing

Dublin Institute of Technology

Kevin Street

Dublin 8.

Or by e-mail to: brendan.tierney@dit.ie

For Office Use Only

**Recommendation: Project Number:**

**Supervisor: Second Reader:**

|  |  |
| --- | --- |
| **TITLE OF PROJECT** | Investigation and analysis of Open Government Data in Ireland, and the creation of a set of quality characteristics for its successful use in mobile applications |
| **STUDENT NAME:** |  |
| **STUDENT NUMBER:** |  |
| **NAME OF SCHOOL/DEPT:** | Computing |
| **PROGRAMME**  **Delete where appropriate** | MSc in Computing (Information Technology) |
| **CONTACT TELEPHONE NO:** |  |
| **CONTACT MOBILE NO:** |  |
| **DIT EMAIL ADDRESS:** |  |
| **OTHER EMAIL ADDRESS:** |  |
| **CONTACT ADDRESS:** |  |
| **HOME ADDRESS:** |  |
|  | |

|  |  |  |
| --- | --- | --- |
| **SUMMARY OF PROJECT**  This project will examine the use of Open Government Data (OGD) in mobile application development. Open Government Data refers to any data that is made freely available by local / national government or government controlled agencies, (OpenGovernmentData, 2011). The key research of the project will be to evaluate what are the quality characteristics required in OGD when releasing it for use in mobile applications, to ensure that the data is fit for this purpose, and can be effectively used by application developers. The emphasis of the research will be placed upon OGD released in Ireland.  To firstly assess what merits quality data, a literature review of existing research into this area will be preformed. A review of Irish open government initiatives and OGD initiatives worldwide will also be carried out.  Pipino, Lee and Wang(2002) suggest that data quality can be both objective and subjective. Objective quality can be measured by data quality metrics and subjective quality measured by the data consumers who use the data. With this in mind an experiment will be designed to both objectively and subjectively measure the quality of Irish OGD, and determine what characteristics this data should have to allow it to be used effectively in mobile applications.  The experiment aims to utilise a number of Irish OGD datasets from a range of local and national government authorities. The objective assessment will analyse the data using quality metrics such as completeness etc. The subjective assessment will analyse the data under use. In this regard the datasets will be used as the input into a public information mobile application whereby a subjective assessment of the data’s accessibility, understandability, value added, mobile application specific characteristics etc. can be performed   |  | | --- | | OGD Use in Mobile Applications | | Findings  Characteristics required in OGD for its effective use in mobile applications  Data Quality Metrics  Objective & Subjective  -Accessibility  - Completeness  - Accuracy  - Interpretability  - Mobile application specific  - Data cleaning required  - etc.  Experiment  To determine the characteristics of good quality OGD, through objective assessment and use of OGD in a mobile application |   The results of the experiment will be evaluated to determine the characteristics of datasets which may achieve positive results, and a root cause analysis will be carried out on datasets which may achieve negative results. Any data cleaning / manipulation carried out on the data during the subjective assessment will be analysed and where appropriate incorporated into the final findings.  To conclude, the characteristics of good quality OGD in relation to its use in mobile applications will be presented based on the results obtained from the experiment. |

|  |
| --- |
| **OTHER INSTITUTIONS / DEPARTMENTS / PERSONNEL ASSOCIATED WITH THE PROJECT:** |
| This project will not require assistance from other institutions, departments or personnel. All open government datasets selected for the experiment will be free for use with no licensing restrictions restricting their use. A full list of required resources both technical and non technical is provided in the later stages of this document. |

|  |
| --- |
| **FULL DESCRIPTION OF PROJECT:**  **Project Background**  Recent years have seen the emergence of e-Government. Through the use of technology, governments can now enhance the access and delivery of its services to its citizens (Silcock, 2001). Aligned with this has been the emergence of the Open Government Data movement, where governments are releasing datasets and making their data freely available to the general public. The value of this data should not be underestimated as recent communications from the European Union indicate that the use of EU open data could deliver a boost of €40 billion to the European economy each year (European Commission, 2011). One of the perceived approaches to achieving this is to turn this raw data into material that consumers can use, by utilising it in smart phone mobile applications (European Commission, 2011).  The open Government Data (OGD) initiative has been gaining momentum worldwide in the last decade with many countries such as the U.S. and the U.K now providing data for public consumption. In their paper ‘Open Government Data’ Geiger and von Lucke (2010) echo the sentiments of EU in saying that freeing up this data can lead to innovation and business development but also call on governments to assess how this data is being made available and define a strategic vision for its further release. Robinson *et al.* (2008) also argue that governments should focus on delivering data so that it is simple, accessible and easy for the public to reuse. The concept of realising this data as open linked data has also been explored both in the U.S. and the United Kingdom. Ding *et al.* (2010) and Sheridan and Tennison (2010), both have made OGD datasets available as linked data and argue that this is another approach to making the data even more accessible to the public. Analysing the various approaches to realising data raises questions in relation to the quality of the data being released, specifically, if the quality of the data is sufficient enough to allow it to be used in mobile application development, and what characteristics should this data have to aid this development process.  Data quality has been an important area of research for decades. In 1970 Edgar F. Codd conceived the relational data model, where data could be stored and accessed structurally, (Codd, 1970). This model has been adopted wholesale by industry in the form of the relational database, replacing previous data models such as the hierarchical model. Codd also introduced the concept of database normalisation which reduced data redundancy, resulting in smaller, better structured relations. Although this technique is an important step in ensuring data is correctly structured and stored, it does not ensure that the data stored in this way is of quality.  Measuring the quality of data is not a ‘one size fits all’ approach (Pipino, Lee and Wang, 2002). Data quality can be assed based on an objective assessment using pre defined metrics, or a subjective assessment of the data when it has been put to use (Pipino, Lee and Wang, 2002). The objective assessment can be used to indicate errors within the data in its static state, whilst the subjective assessment of putting the data to use can indicate perceived shortcomings in the data from a user’s viewpoint. |
| Wang and Strong (1996) offer a somewhat different viewpoint. They place little emphasis on the objective assessment and use the analogy of data as a commercial product. They argue that data quality should be measured by the data consumers using the data, as opposed to being defined theoretically based on researchers’ experience. In their study on data quality, they surveyed data consumers asking them what they considered were the characteristics upon which they measured data quality. The results are presented below, but broadly speaking the findings illustrate that high quality data should be intrinsically good, contextually appropriate for the task, clearly represented and accessible to the user (Wang and Strong, 2006).    Figure - A Conceptual Framework of Data Quality (Wang and Strong, 1996)  The research outlined above suggests that measuring and defining of data quality is not an exact science. As Pipino, Lee and Wang (2002) indicate, objective and subjective assessments of the data can indicate its fundamental quality, whilst Wang and Strong (1996) focus more on the subjective approach, and suggest assessing the quality of data as if it were a product.  The findings outlined above illustrate that the area of data quality has been a widely studied and researched domain. With the emergence of OGD, governments are now making vast amounts of previously unreleased data available. As the European Commission has indicated a key usage area for these datasets is in mobile phone applications. To this end, this research will set out to define what constitutes quality from this viewpoint. The objective quality attributes of the datasets will be evaluated and consideration will also be given to a subjective assessment of the datasets to determine what quality characteristics OGD should have for it to be effectively used in a smart phone application.  The key outcome from this will be a set of characteristics which when applied to OGD will make it easier for this data to be incorporated and used in mobile application development. |
| **Project Description**  Building upon the research illustrated above I plan to evaluate the quality characteristics that are required in OGD for it to be effectively used in a smart phone mobile applications. The evaluation will take the form of an experiment where a number of freely available Irish OGD datasets will be used as the input into a public information mobile application.  Taking the analogy offered by Wang and Strong (1996), that data should be considered as a product, I plan to measure and define the quality of this data by assessing how fit it is for this task. Aligned with this will be an objective assessment of the datasets using existing data quality metrics.    Subjective Assessment  Wang (1996) defines data quality as data that is for use by data consumers. To subjective assess the quality of Irish OGD, a number of these datasets will be used as the input into a smart phone mobile application, determining if the OGD is fit for use in this context, and highlighting any virtues or inadequacies within this process. |
| To this end, specific quality characteristics will be targeted and evaluated throughout the entire process from initial dataset selection through to the completion of the application. Data accessibility and availability of formats will be examined. Any data cleaning or manipulation tasks which need to be performed on the data in order to ensure that it can be integrated it into the application will also be documented and analysed. Briefly, some of the quality areas that will be evaluated are: (please not this list is not exhaustive, further areas will be proposed and analysed during the thesis completion)   * Accessibility 🡪 The ease of access to the data through ODG data portal, API or websites. The availability of the data in different formats and the relevancy of these formats to the information contained within them. For example, the packaging of location based information in KML / KMZ files as opposed to CSV files so that it can be easily mapped. The timeliness of the datasets and also the impact that updates to the datasets can have on the application will also be reviewed. In addition the degree to datasets in proprietary data formats can be integrated into the application will also be studied. * Level of Data Cleaning required 🡪 Manipulation tasks that are required on the datasets before they can be used in the mobile application. Any data cleaning exercises which involve making updates / corrections to the data before being able to successfully make use of the data, such as error correcting, de-duplication, and resolving missing data. * Smart phone specific characteristics 🡪 Specific tailoring that needs to take place to take advantage of mobile application functionality. For example, many mobile applications now offer location based information. One possible quality characteristic that could be uncovered is that location based information should contain longitudinal and latitudinal coordinates so this information could be successfully plotted on a map and displayed to the user.   To prevent bias towards any one source of OGD a range of datasets will be chosen from a variety of OGD providers, such as statcentral.ie, data.fingal.ie and dublinked.ie. Additionally, the OGD datasets selected as input will be from a range of data categories and not just linked to one class of data. This approach will prevent any findings of the experiment from being limited to just OGD provider, or one class or category of data.  The mobile application will be developed on the Android framework which is the dominant smart phone operating system within the smart phone mobile market, with a worldwide market share of over 52% (Gartner, 2011). The function of the application will be to act as a source of public information to the application user, built upon the OGD datasets selected as input. This simulates the typical function of a smart phone application which utilises OGD and as such represents the best approach to determining what characteristics the data should have to be fit for this purpose. |

Objective Assessment

Whereas a subjective assessment of data studies the data under use, objective assessment can be used to measure certain quality characteristics irrespective of how the dataset is going to be used. Quality metrics such as completeness, consistency, free-of-error and timeliness, (Pipino, Lee and Wang, 2002), can all objectively measure the quality of data.

An objective assessment of the selected datasets will also be carried out. The objective assessment will analyse the data in its static state and can be used to measure the quality of the datasets. This assessment will complement the subjective assessment and build upon past research carried out in this area.

Following both the subjective and objective assessments of the datasets, the results will be analysed and any correlations between the datasets assessments identified. A collection of quality characteristics will be presented with specific reference to the characteristics OGD datasets require for them to be effectively used in smart phone mobile applications.

|  |
| --- |
| **Project Aim**  The aim of the project is to evaluate the quality characteristics required in Open Government Data for its effective use in smart phone mobile applications. Through the synthesis and execution of an appropriate experiment the research will appraise the features which make OGD datasets fit for this purpose. The findings of the experiment will be analysed to infer the characteristics which make the OGD effective in this task. Finally, the findings will be summarised and quality characteristic guidelines will be constructed which explain and justify what attributes OGD released for use in mobile applications should have. |
| **Project Objectives**   * Examine the current state of OGD initiatives in Ireland and worldwide * Investigate the current views and research conducted to date on data quality * Develop an experiment to ascertain what are the quality characteristics required in open government datasets so that they may be used effectively in smart phone mobile applications * Document and evaluate the findings from the experiment * Based on the evaluation, suggest a set of characteristics that OGD owners should ensure that OGD datasets have before being released for use in mobile application development * Make recommendations for any future research in this area |
| **Evaluation Criteria**  Value of literary review completed  A literature review will be carried out focusing on the key areas of the thesis; data quality characteristics, Open Government Data initiatives both nationally and internationally and the application of this OGD in mobile application development. This literary review will outline past research in these areas and shape the experiment design and implementation. To this end, the quality and value of the literary review is an key evaluation criterion of the thesis.  Suitability of experiment design & execution  The experiment should be designed and executed to fully examine the research question being asked. With this in mind, an appropriate experiment analysis and design will need to be conducted and will be an important evaluation criterion.  Quality of analysis of findings  The findings from the experiment should be objectively analysed and presented a true assessment of the experiment conducted. These findings should be evaluated with respect to the existing research conducted in this area (outlined in literary review), and also in the respect the research question being asked. |
| **Deliverables**   * Research Project – Dissertation paper * Literary Review * Software Developed |

|  |
| --- |
| **Timeframes & Project Plan**  The project timeframe below has been split into the following three phases:   * Phase 1 – Initial topic background research and investigation * Phase 2 – Development and completion of thesis * Phase 3 – Delivery of final presentation   Overall Project Timeframe |

RMMM Matrix (Risk, Measurement, Monitor, Mitigation)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Risk** | **Measure** | **Monitor** | **Mitigation** |
| 1. | Project Proposal is not accepted | High | Continuously work and re-work proposal prior to submission. Submit draft proposal for feedback as part of the ‘Research Methods and Proposal Writing Project’ | Take on board feedback from proposal presentation session. Re-work proposal and re-submit if required |
| 2. | OGD datasets change during lifecycle of project – particularly experiment | Low | Monitor dataset release and update dates to determine any changes | Prior to beginning experiment identify a range of OGD datasets and sources to be used during experiment. Save these datasets to a local location to mitigate against possibility previously selected datasets changing |
| 3. | Level or programming experience required to complete experiment | Medium | Track progress of experiment with project schedule to assess progress. If off schedule, re-evaluate levels of application features and OGD datasets being used | Develop experiment in Java of which I have some experience. Identify a number of helpful programming resources (books, online tutorials) prior to project and begin self study |
| 4. | Loss of some or all of project data | High | Ensure backups of all project data is completed regularly. Back up project to multiple devices such as external hard drive and online repository (Dropbox etc.) | Develop a backup strategy to save all project related material and continuously back up project data to multiple devices |

|  |
| --- |
| **Technical and Non-Technical Resources Required** Technical Resources Required  * Laptop with internet access * Access to OGD datasets * Development IDE * Database * Android SDK * Microsoft Word / Microsoft Excel * Backup Device / Dropbox  Non Technical Resources Required  * Library Access |
| **References**  Codd, E F (1970) A *Relational Model of Data for Large Shared Data Banks*  [Online]. Available: <http://dl.acm.org/citation.cfm?id=362685> [Accessed 11th December 2011]  Ding L *et al.* (2010), Data*-gov Wiki: Towards Linking Government Data*  [Online]. Available: <http://tw.rpi.edu/wiki.tw/images/9/97/Dingl2010datagov.pdf>  [Accessed 23rd January 2012]  European Commission (2011), *Digital Agenda: Turning government data into gold*  [Online]. Available: [http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1524&format=HTML&aged=0&language=EN&guiLanguage=en](http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1524&format=HTML&aged=0&language=EN&guiLanguage=en%20) [Accessed 12th December 2011]  Gartner (2011), *Gartner Says Sales of Mobile Devices Grew 5.6 Percent in Third Quarter of 2011; Smartphone Sales Increased 42 Percent*  [Online]. Available: <http://www.gartner.com/it/page.jsp?id=1848514>  [Accessed 21st December 2011]  Geiger, P C, von Lucke, J (2010) *Open Government Data – Free accessible data of the public sector*  [Online]. Available: <http://www.donau-uni.ac.at/imperia/md/content/bibliothek/verlag/cedem11_proccedings.pdf#page=183>  [Accessed: 23rd January 2011]  OpenGovernmentData (2011), *What is Open Government Data*  [Online]. Available: <http://opengovernmentdata.org/what/>  [Accessed 15th December 2011]  Pipino, L L, Lee, Y W and Wang, R Y (2002), *Data Quality Assessment*  Communications of the ACM, 45(4), pp.211–218  Robinson, D, Yu, H, Zeller, W and Felten, W E (2008), *Government Data and the Invisible Hand*  [Online]. Available: <http://www.yjolt.org/files/robinson-11-YJOLT-draft.pdf>  [Accessed 23rd January 2012]  Sheridan J, Tennison J (2010), *Linking UK Government Data*  [Online]. Available: <http://events.linkeddata.org/ldow2010/papers/ldow2010_paper14.pdf>  [Accessed 23rd January 2012]  Silcock, R (2001) *What is e-government*  [Online]. Available: <http://catedras.fsoc.uba.ar/rusailh/Unidad%202/Silcock%202001,%20What%20is%20E-gov.pdf> [Accessed 17th December 2011]  Wang, R, Strong, D, (1996) *Beyond Accuracy: What Data Quality Means to Data Consumer*s  [Online]. Available: <http://w3.cyu.edu.tw/ccwei/PAPER/ERP/data%20quality%28JMIS%29.pdf>  [Accessed 12th December 2011] |

Signed : \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date : \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_