|  |  |  |  |
| --- | --- | --- | --- |
| **Project Name:** | Information retrieval and machine learning for conducting empirical systematic reviews. | **Funder:** University of Strathclyde |  |
| **Project Description:** | Discover methodologies for improving the retrievability of information retrieval models using the Lucene4ir platform and the tool for Technologically Assisted Reviews in Empirical Medicine. | | |
| **Student:** | Alexandros Ioannidis | **Supervisor:** | Leif Azzopardi and Martin Halvey |
| **Institution:** | University of Strathclyde | **Dept / School:** | Computer and Information Sciences |
| **Date of First Version**: | 11/01/18 | | |
| **Date of Updates**: |  | | |

This template is based on DCC. (2013). Checklist for a Data Management Plan. V.4.0. Edinburgh: Digital Curation Centre. Available online: <http://www.dcc.ac.uk/resources/data-management-plans>.

For assistance contact the Research Data Management and Sharing team:

[researchdataproject@strath.ac.uk](mailto:researchdataproject@strath.ac.uk) 0141 548 4581.

|  |
| --- |
| **Data Collection** |
| **What data will you collect or create?** |

Use the table below to list all research data which you will collect or generate as part of this project. Examples have been included to help you get started.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data type | Original format | Preservation format\* | Estimated volume | IPR Owner | Active storage location | Completed storage location |
| Experiment notes | .xlsx, .docx | .csv, .rtf | <10MB | UoS | i:drive | Pure |
| Microscope images | TIFF | Original | ~5GB | Company X | StrathCloud | Pure |
| Paper notebook | Paper | PDF | ~400MB | Student | Cabinet in dept | Pure |
|  |  |  |  |  |  |  |

\*Preservation formats should be easy to access without the need for specific proprietary software.

|  |
| --- |
| **How will the data be collected or created?** |

* How will you collect or generate data?
* How will you structure and name your folders and files?
* How will you handle versioning?
* What quality assurance processes will you adopt?

<http://www.strath.ac.uk/ps/strategyandpolicy/recordsmanagement/> - file naming and versioning advice.

|  |
| --- |
| The PubMed data are already collected. Furthermore, the PubMed dat a are inside a folder named PubMed\_filter. Inside this folder there is a large list of XML files (e.g. PMID.xml, where PMID stands for the PubMed ID of each document), and each XML file corresponds to a PubMed document. |

|  |
| --- |
| **Documentation and Metadata** |
| **What documentation and metadata will accompany the data?** |

* What information is needed for the data to be to be read and interpreted in the future?
* How will you capture / create this documentation and metadata?

Documentation may include details on the methodology used, analytical and procedural information, definitions of variables, vocabularies, units of measurement, description of instruments, software and hardware used, use conditions, and any assumptions made. Consider how you will capture this information and where it will be recorded.

|  |
| --- |
| The data will be accompanied by informative text files with the name convention README.txt, which is frequently found in complete software solution projects. The information of these text files will include descriptions for variables, values and other files found in the projects. |

|  |
| --- |
| 1. **Ethics and Legal Compliance** |
| **How will you manage any ethical issues?** |

* Have you gained consent for data preservation and sharing?
* How will you protect the identity of participants if required? e.g. via anonymization
* How will sensitive data be handled to ensure it is stored and transferred securely?

Ethical issues affect how you store data, who can see/use it and how long it is kept.

Managing ethical concerns may include: anonymisation of data; referral to departmental or institutional ethics committees; and formal consent agreements. You should show that you are aware of any issues and have planned accordingly. If you are carrying out research involving human participants, you must consider whether consent is required to allow data to be shared and reused.

|  |
| --- |
| No ethical considerations and implications apply to this study, since the research will not involve human beings as participants and the PubMed research data that will be used is made available to the public. Emphasis will be given to avoid all sort of negative impacts and all data will be processed and presented in a fairly and lawful way, while assuring quality and integrity in this research. Additionally, during the design and conduction of this research, an effort has been made to ensure that this study is characterised by impartiality and independence. |
|  |

|  |
| --- |
| **How will you manage copyright and IPR issues?** |

* How will the data be licensed for reuse?
* Are there any restrictions on the reuse of third-party data?
* Will data sharing be postponed / restricted e.g. to publish or seek patents?
* Do the IPR owners have any reason to restrict data sharing?

Open data is typically made available under a CC-BY licence, meaning that anyone can reuse the data for any purpose, as long as they cite the source of the data. Data which is commercially sensitive should be restricted.

|  |
| --- |
| Again this is not applicable to the current project. |
|  |

|  |
| --- |
| 1. **Storage and Backup** |
| **How will the data be stored during research, and how will you manage access and security?** |

* Will data be stored on the University network?
* How will data be transferred to the University network if it originates from another location?
* How will you ensure that collaborators, supervisors, or participants can access your data securely?
* Will data be stored on H:drive, i:drive, StrathCloud, or elsewhere?

Data stored on the University provided storage is dual sited and replicated between two data centres which are physically separated by several hundred metres. Data links between datacentres are provided by dual disparate fabrics, providing added resilience. Additionally, the central I.T. service provides tape based backup to a third and fourth site. Data from each of the two data centres is backed up to separate tape backup locations. Data security is provided by access controls defined either at a user level, in the case of personal data storage – i.e. the user’s home directory, or group based where the data is being stored in shared data space.

|  |
| --- |
| For the purpose of storage I will be using Strathcloud the on-site instance of Sharefile to store, organize and share data. The usage of Strathcloud will me maximized, however there are some storage-size restrictions,. In the case that research data is not stored in a secure storage facility that does not belong to the university or to a partner organization, then the data will be encrypted in order to abide with the encryption policy 2017 of the university of Strathclyde. Thus ensuring that optimal approaches are constantly followed. |

|  |
| --- |
| 1. **Data Curation and Open Access to Data** |
| **How will data preservation and open access to data be managed?** |

* What data must be retained/destroyed for contractual, legal, or regulatory purposes?
* How will you decide what other data to keep?
* What data will be shared openly?
* When will you make the data available?
* How will data be preserved and shared?
* How will completed datasets be organised?

To comply with EPSRC policy, it is necessary to preserve and openly share research data. Researchers should upload data to Pure for effective long term preservation and open access. If you would prefer to upload data into another open access data repository, a dataset record should still be created in Pure and a link to where the data is available should be included.

Any outputs, including theses, arising from EPSRC funding are required to include a data statement which links to supporting data using a DOI (Digital Object Identifier - available from Pure). Make sure you create a dataset record at an early stage, so that a DOI can be minted and included on your final submission. The data statement is typically included in the acknowledgements section.

EPSRC expect structured metadata describing the research data to be published (normally within 12 months of the data being generated); metadata must be sufficient to allow others to understand what research data exists, why, when and how it was generated, and how to access it. This expectation can be met by creating a dataset record in Pure and including the relevant details.

Data which does not relate to any research findings, and is deemed to be of no value, no interest, and not worth sharing, can be deleted.

**The above steps are also recommended as good practice for those not mandated by EPSRC policy**.

|  |
| --- |
|  |

|  |
| --- |
| **Are any restrictions on data sharing required?** |

* What restrictions are required on data sharing?
* How can these restrictions be minimised? (e.g. temporary embargo, partial sharing, one to one sharing, non-disclosure agreements)

Explain any necessary restrictions on sharing (e.g. commercial or security reasons).

If any data cannot be shared, but is going to be preserved, a dataset record should still be created in Pure, the data can be uploaded to the record, the data itself can be restricted, whilst the record is made publicly visible. The record should explain why the data is not accessible, the circumstances under which access may be granted, and who to contact for information about the dataset.

|  |
| --- |
| There are no restrictions required on data sharing , since the PubMed collection used in the current research is already available publicly. |

|  |
| --- |
| 1. **Responsibilities and Resources** |
| **Who is responsible for data management?** |

* Who is responsible for implementing the plan, and ensuring it is reviewed and revised?
* Who will be responsible for each data management activity?
* How will responsibilities be split across partner sites in collaborative research projects?
* Will data ownership and responsibilities for RDM be part of any consortium agreement or contract agreed between partners?

|  |
| --- |
| Since the current research project is not a collaborative research project, I will be responsible for managing and revising the plan. |

|  |
| --- |
| **What resources will you require to deliver your plan?** |

* Is additional specialist expertise (or training for existing staff) required?
* Do you require hardware or software which is additional to existing institutional provision?

|  |
| --- |
| No additional specialist expertise is required and the existing hardware and software in the institutional provision is sufficient to conduct my research. |

|  |
| --- |
|  |

**Useful links:**

* Pure: [pure.strath.ac.uk/](http://pure.strath.ac.uk/)
* Strathcloud: [www.strath.ac.uk/it/services/strathcloud/](http://www.strath.ac.uk/it/services/strathcloud/)
* Network Drives: [www.strath.ac.uk/it/filestore/](http://www.strath.ac.uk/it/filestore/)
* RDMS Help: [moss.strath.ac.uk/projects/resdata/SitePages/Home.aspx](https://moss.strath.ac.uk/projects/resdata/SitePages/Home.aspx)
* EPSRC Policy: [www.epsrc.ac.uk/about/standards/researchdata/](http://www.epsrc.ac.uk/about/standards/researchdata/)
* DCC Guidance: <http://www.dcc.ac.uk/resources/data-management-plans>
* UK Data Service Guidance: <https://www.ukdataservice.ac.uk/manage-data>