# Project Charter

## NUSH Student Research Repository

Lairoongroj Jinnawat

Version 1.1 — 4 January 2022

## 1 Background

NUS High School of Mathematics and Science offers a six-year programme specialising in mathematics and science education. The Da Vinci Programme, led by the Department of Research, Innovation and Enterprise (RIE), is a unique and defining component of NUS High School's curriculum. It culminates in the Advanced Research Project (ARP), where teams of 1-3 students collaborate with advisors and conduct research. The product is generally a research report, as well as a poster-aided presentation for an annual Research Congress (NUS High School, n.d.).

RIE has an extensive library of over 1000 research reports, ranging from ARPs to independently initiated research. These reports are invaluable to current students as references and inspiration for research. They will allow current students to build on the work of previous cohorts and extend interesting and valuable research. However, there is at present no easy way for students to freely search and access these reports.

We wish to set up a research repository system. The system requirements are as such:

a) The Repository should be accessible only to current matriculated students and staff of NUS High School. Authentication is needed to access the system.

b) Reports should be accurately indexed and classified with keywords so that they are searchable.

## 2 Objectives

Understanding previously conducted research can save time and effort spent writing methodologies, designing statistical analyses, etc. More importantly, it

allows researchers to gain novel, domain-specific insights and identify significant challenges and problems before they arise.

The purposes of the proposed Research Repository are pedagogical as well as practical: students should learn how to access past findings in a way that maximises insights gained, including how to use institutional repositories and related technologies.

Therefore, the primary objective of the project is to set up a Research Repository for students' reports. Works and data in the Repository should be findable and accessible in the context of NUS High School, following the FAIR principles for scientific data management (Wilkinson et al., 2016). The Repository's database schema should be informed by the nature and form of ARPs, such as formatting/citation requirements and grading rubrics. It should implement keyword tagging and other rich and NUSH-relevant metadata, as well as search by keywords. A wide variety of datasets and file formats, ranging from the general (.csv, .png, .ipynb) to the more niche and field-specific (molecular file formats, GIS vector file formats), should also be supported and easily downloadable.

Secondary objectives include interoperability and reusability. Interoperability serves to future-proof the Repository, allowing future generations of students to improve upon and adapt it to the ever-evolving research landscape. Maintenance and documentation of the Repository and its source code are baselines for interoperability. Reusability allows students to consult authors if and when relevant to their research needs. Reusability also includes document version control.

The Repository interface should be usable, with minimal prior guidance. It may resemble existing institutional repositories, such as DR-NTU[1] or ScholarBank@NUS[2].

The Repository should implement an appropriate level of permissioning and authentication. In particular, administrators should directly oversee uploading, editing, and deleting of files. If feasible, measurement and quantitative assessment of NUSH's scholarly output should be implemented following Westell's (2016) indicators of institutional repository success.

Finally, Project Officers should also take into account training and onboarding staff members and students.

[1]The Digital Repository of NTU. https://dr.ntu.edu.sg/
[2]https://scholarbank.nus.edu.sg/

# 3 Methodology

Users can access the Repository's research reports, relevant datasets, and their metadata via a web application. Integration into NUS High School's eSpace portal will be explored.

These data will be stored in a document-oriented database, which offers flexibility to meet changing application needs and scalability across a variety of data types (MongoDB, n.d.).

All documents and datasets associated with a given ARP will be indexed and tagged by keywords, rendering them searchable. Around 1000 ARPs conducted before the implementation of the Repository will be automatically tagged using machine learning algorithms. ARPs submitted after implementation will have keywords suggested by the Repository in addition to metadata furnished by authors.

In machine learning terms, the problem is unsupervised or semi-supervised classification on a medium-size natural language dataset. We hope to explore semi-supervised Latent Dirichlet Allocation (*ssLDA*) (Wang et al., 2012), KEA (Witten et al., 2005), and TextRank (Mihalcea & Tarau, 2004).

Proposed technology stack (as of 4 Jan 2022):

a) **Web application**

   i. Database: MongoDB

   ii. Backend & frontend: Django

b) **Keyword tagging**

   TextRank, KEA, `scikit-learn`

c) **Codebase management & deployment**

   i. GitHub

   ii. Docker (?)

# 4 Work Schedule

The Repository will be developed and tested iteratively, deployed, and maintained by the Project Officer, with input and guidance by stakeholders. (Stakeholders include NUS High School, particularly RIE and current students who will provide direction and user feedback; as well as AppVenture, who will provide advice.) Delivery of the project will be measured against the objectives outlined above.

Preliminarily, the project will span 5 months, commencing in January 2022 and concluding in May 2022.

Timeline & vendor monthly rate

| Month | Deliverables | Rate (in SGD) |
|---|---|---|
| January 2022 | Index research reports. Begin setting up Repository system, clarifying server interface and requirements. | $768.00 |
| February 2022 | Develop Repository as web application, with user authentication and permissions. Upload research reports. Refine indexing and keyword tagging. | $768.00 |
| March 2022 | | $768.00 |
| April 2022 | | $768.00 |
| May 2022 | Test and launch Repository. Submit documentation. | $768.00 |
| | **Total** | $3,840.00 |

# References


Mihalcea, R. & Tarau, P. (2004). TextRank: Bringing order into text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411. Retrieved December 23, 2021, from https://aclanthology.org/W04-3252

MongoDB. (n.d.). *Document database - NoSQL* [MongoDB]. Retrieved December 23, 2021, from https://www.mongodb.com/document-databases

NUS High School. (n.d.). *Research, innovation and enterprise*. Retrieved January 3, 2022, from https://www.nushigh.edu.sg/studying-at-nus-high/the-nus-high-diploma/research-innovation-and-enterprise

Wang, D., Thint, M. & Al-Rubaie, A. (2012). Semi-supervised latent dirichlet allocation and its application for document classification. *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, *3*, 306–310. https://doi.org/10.1109/WI-IAT.2012.211

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C. & Nevill-Manning, C. G. (2005). Kea: Practical automated keyphrase extraction. *Design and usability of digital libraries: Case studies in the asia pacific* (pp. 129–152). IGI Global.