Construction of an Open Science Thesaurus: A Library & Information Science Course Project

Asma Nadia Z.R.[1,2], Lee Tjun Yang[3] and **A. Abrizah[2]**
[1]School of Information Science,
College of Computing, Informatics & Mathematics,
Universiti Teknologi MARA Kedah Branch,
Sungai Petani Campus, Merbok, Kedah, MALAYSIA
[2]Department of Library & Information Science,
Faculty of Arts and Social Sciences,
Universiti Malaya, Kuala Lumpur, MALAYSIA
[3]Main Library, Universiti Tunku Abdul Rahman,
Kampar, Perak MALAYSIA
e-mail: asmanadia@uitm.edu.my, tylee@utar.edu.my, abrizah@um.edu.my

ABSTRACT
Open science has emerged as a critical area of research in Library and Information Science (LIS), with increasing emphasis on sharing of research output and collaboration across disciplines and geographic boundaries. However, one of the main challenges in this area is the lack of a standardized vocabulary for discussing and organizing open science concepts. To address this issue, the Open Science Thesaurus was constructed as an output of a Masters in Library & Information Science (MLIS) course on Indexing, Abstracting and Thesaurus Construction. The aim of the thesaurus is to provide a standardized vocabulary for indexing and retrieving information related to Open Science. This paper outlines the construction of the Open Science Thesaurus that involves several key steps: (a) Using the Open Science Taxonomy developed by Foster to determine the core concepts and usable descriptors; (b) Conducting an analysis of existing terminologies within the domain, aiming to identify commonalities, gaps, and inconsistencies encompassing core literature, policies, and guidelines; (c) Compiling a comprehensive list of 250 terms related to Open Science; and (d) Establishing preferential, hierarchical, associative and relationships between descriptors. The resulting thesaurus is expected to provide a framework for more effective communication within the open science community, addressing the challenges that arise from the lack of standardized terminology. This paper highlights the importance of the Open Science Thesaurus in promoting a common understanding and effective communication among researchers, librarians, and other stakeholders in the Open Science community.

Keywords: Controlled vocabulary, Open Science, Open Science taxonomy, Standardized terminology, Thesaurus construction

INTRODUCTION

In recent years, Open Science has gained significant popularity. While debates continue regarding the precise definition of the term Open Science, there is consensus on the fundamental concept of openness and transparency inherent to Open Science. According to Bartling and Friesike (2014), open science refers to a scientific culture that is characterized by its openness in which scientists share results almost immediately and with a very wide audience. Open science broadly refers to the sharing of resources, ideas, and other outputs from research with an emphasis in making resources publicly and freely available (Zarghani et al. 2023). The purpose of making resources available and accessible by the public is to make full utilization of the resources towards maximizing the usage of it. Zarghani et al. (2023) further emphasize that open science promotes greater openness, accessibility, global collaboration, transparency, and integration in scientific endeavors. The UNESCO Recommendation on Open Science (UNESCO 2021) establishes a global framework for open science policies and practices, recognizing the diversity of open science perspectives across disciplines and regions. It presents an internationally agreed definition, a set of shared values and guiding principles. Furthermore, it outlines a set of actions aimed at promoting fair and equitable implementation of open science at all levels. Prior to its adoption in 2021, there was no universally accepted definition of open science, with standards often existing only at regional, national, or institutional levels.

.

As Open Science continues to evolve and integrate into various research cultures, many terms have emerged and become dispersed across numerous platforms worldwide. Several Open Science thesauruses have been found on the Internet created by different organizations. One notable example was the one developed by the Institut de l'Information Scientifique et Technique (Inist), Centre national de la recherche scientifique (CNRS) and published on Linked Open Terminology Resources (Loterre) website. Loterre serves as a platform for exposing and sharing multilingual and multidisciplinary scientific terminology, covering diverse fields such as Humanities & Social Sciences, Geography, Physics, Chemistry, Engineering & System Sciences, Health Sciences, Life Sciences, Earth & Universe, and Taxonomies. This particular thesaurus is available in English, Spanish, and French. Loterre Open Science Thesaurus is a multilayered, multilingual thesaurus structure that is the central concept of Open Science. It was initiated on the premise of existing glossaries and taxonomies proposed within the FOSTER project framework. The thesaurus was constructed in March 2021 and has since undergone several updates. The subsequent updates occurred in June 2021 (36 terms), July 2021 (59 terms), February 2022 (17 terms), March 2022 (2 terms), December 2022 (6 terms), and the latest update was in March 2023 (24 terms) (Loterre 2023). These changes involved the addition of new terms, removal of terms, and adjustments to the term hierarchy. As of July 2023, the total number of terminologies (excluding duplicate terms) reached 486.

This paper serves as a descriptive account of an academic work in constructing the Open Science Thesaurus. It outlines the steps taken, the methods employed, and the sources consulted to compile a comprehensive list of 250 terms related to Open Science. The objective of this paper is to provide a detailed description of the process and methodology used in creating the thesaurus. The Open Science Thesaurus represents a capstone project

for a Masters in Library & Information Science (MLIS) program, specifically within the context of a course titled “Indexing, Abstracting and Thesaurus Construction”. This project serves as a culmination of the students' academic journey, demonstrating their proficiency in the construction of a thesaurus tailored to the field of Open Science. Through this descriptive approach, this paper aims to contribute to the field by offering insights into the construction of a valuable resource for effective communication within the open science community.

METHODS

The same methods employed by Ogg et al.'s (1994) in developing their medical informatics thesaurus were adopted. These methods, as outlined by Lancaster (1972, as cited in Ogg et al. 1994), encompass four key approaches to the construction of a thesaurus:

1. Generate terms systematically based on indexing a representative set of documents.
2. Convert an existing term.
3. Extract the terms from existing, more general thesaurus or develop a specialized thesaurus within the framework of the general one.
4. Collect terms from diverse sources including authorized platforms, publications, related projects, or initiatives.

During the preliminary search on the concept of Open Science, diverse online platforms were encountered that presented distinct elements and related terminology. To initiate the exploration, the FOSTER (Facilitate Open Science Training for European Research) Open Science Taxonomy depicted in Figure 1 was consulted, providing a visual presentation of the areas encompassed within Open Science. However, it is important to note that due to the ongoing expansion of Open Science, there may be additional topics and terms not yet included in this taxonomy.

In the development of this thesaurus, the principles of literary warrant and user warrant were recognized and utilized. Literary warrant involved extracting terms generally used in the literature pertaining to Open Science. On the other hand, user warrant entailed selecting terms based on potential keywords or terms commonly used by users. The inclusion of both literary and user warrant was crucial to ensure the incorporation of relevant terms in the Open Science Thesaurus, ultimately serving the purpose effectively.

In addition, a structural warrant was employed, which involved selecting terms based on the structure and scope of the controlled vocabulary. To assist with this process, the FOSTER Open Science Taxonomy (Figure 1) was utilized as a starting point. This taxonomy provided a foundation for exploring a diverse range of topics related to Open Science by navigating through various associated terms (Pontika et al. 2015). The search for terms was expanded by exploring additional available sources.

The search process involved exploring multiple sources available on the Web. Published documents by authoritative organizations such as UNESCO, which provides comprehensive guidelines on Open Science, International Science Council, as well as resources from European Open Science Cloud (EOSC), PLOS on Open Science, Malaysia Open Science Platform and a few others were referred to . Additionally, selected articles that address the topics of Open Science and its sub-topics such as open access, open data, open educational resources, open metrics and citizen science were delved into. To gauge the extent of research in this area, Scopus database was examined to identify the number of articles specifically related to the keyword “Open Science” published in the year 2023 (the search was conducted in early June 2023). The results were remarkable, revealing nearly 1000. However, upon reviewing the abstracts, it was found that only 100s of the listed articles were directly relevant to the topic of open science.

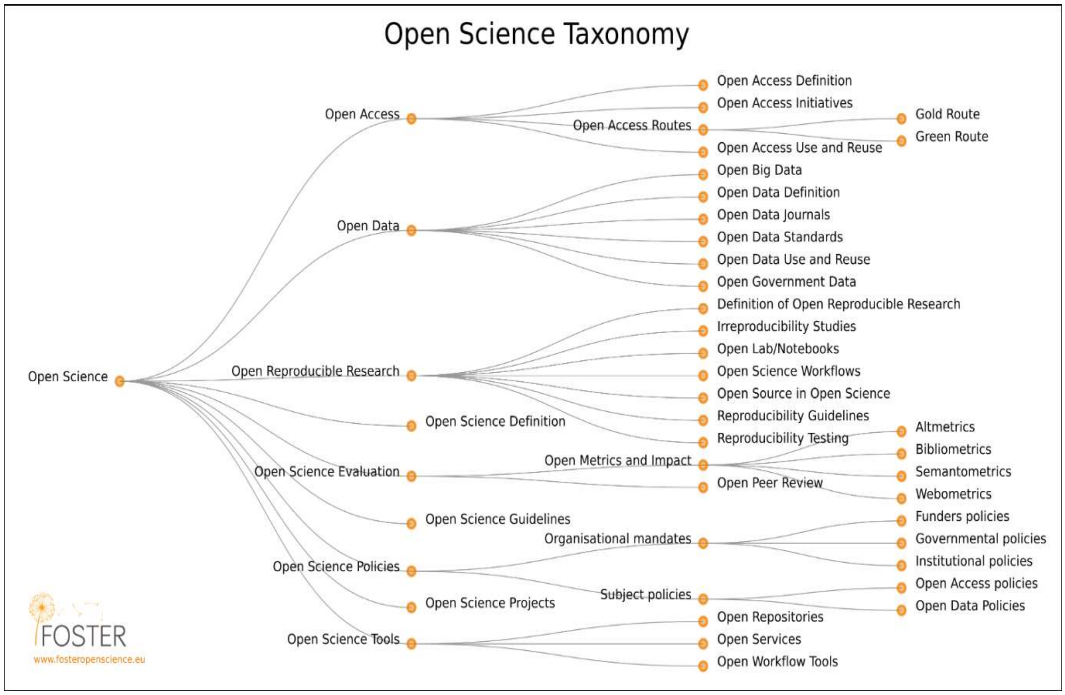


Figure 1: FOSTER Open Science Taxonomy

Throughout the search process, considerable effort was dedicated to consulting approximately 60 distinct resources, comprising web resources, policy documents and guidelines, and journal articles. The utilization of these resources enabled the enhancement of the inclusiveness and representativeness of the Open Science Thesaurus. The aim was to compile an all-encompassing list of 250 terms or descriptors associated with Open Science. Each term was categorized under specific subtopics of Open Science, guided by the authors’ understanding and extensive reading of the referenced resources. Additionally, a hierarchical tree structure was constructed to facilitate and illustrate the relationships between the various terms, to show how different terms relate to and interact with one another in the context of Open Science. The identified descriptors were then meticulously organized

alphabetically to eliminate any duplication, and established the relationship (equivalent, hierarchical and associative) between preferred and non-preferred terms. Definitions or explanations for terms, and scope notes are provided for necessary to clarify the intended usage and boundaries of terms. The thesaurus was then reviewed by the course instructor to ensure accuracy and usability. Based on the feedback received, the thesaurus was carefully revised and refined.

RESULTS

Based on the methods and steps outlined previously, a hierarchical structure known as 'Hierarchical Tree Structure for Open Science Thesaurus (Figure 2) was created. Microsoft Visio in Microsoft 365 office suite was used to construct the tree diagram. Beginning with the main topic of Open Science, the process of categorization into several sup-topics was undertaken including Open Access; Open Data; Open Educational Resources (OER); Open Engagement; Open Innovation; Open Research; Open Science Evaluation; Open Science Infrastructure; Open Science Governance; Open Science Concept; Open Science Initiatives; Open Sources; Open License; and Open Science Policies. These 14 main topics are represented by blue boxes in Figure 2.

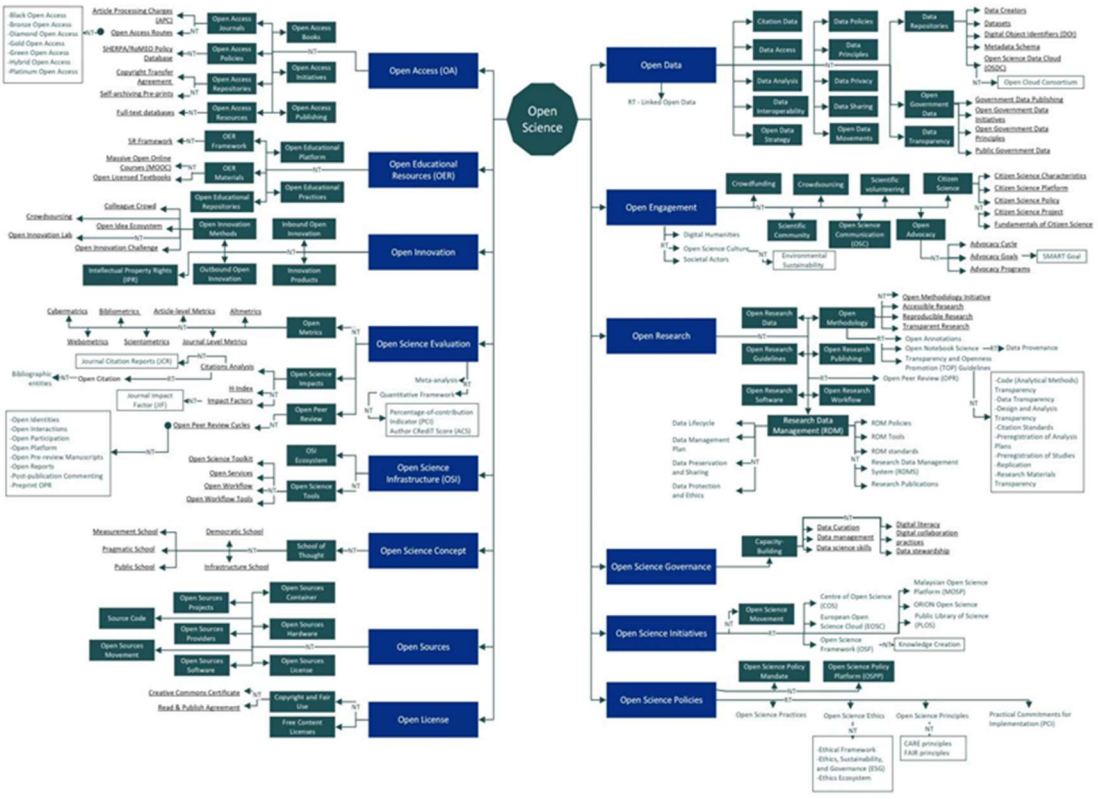


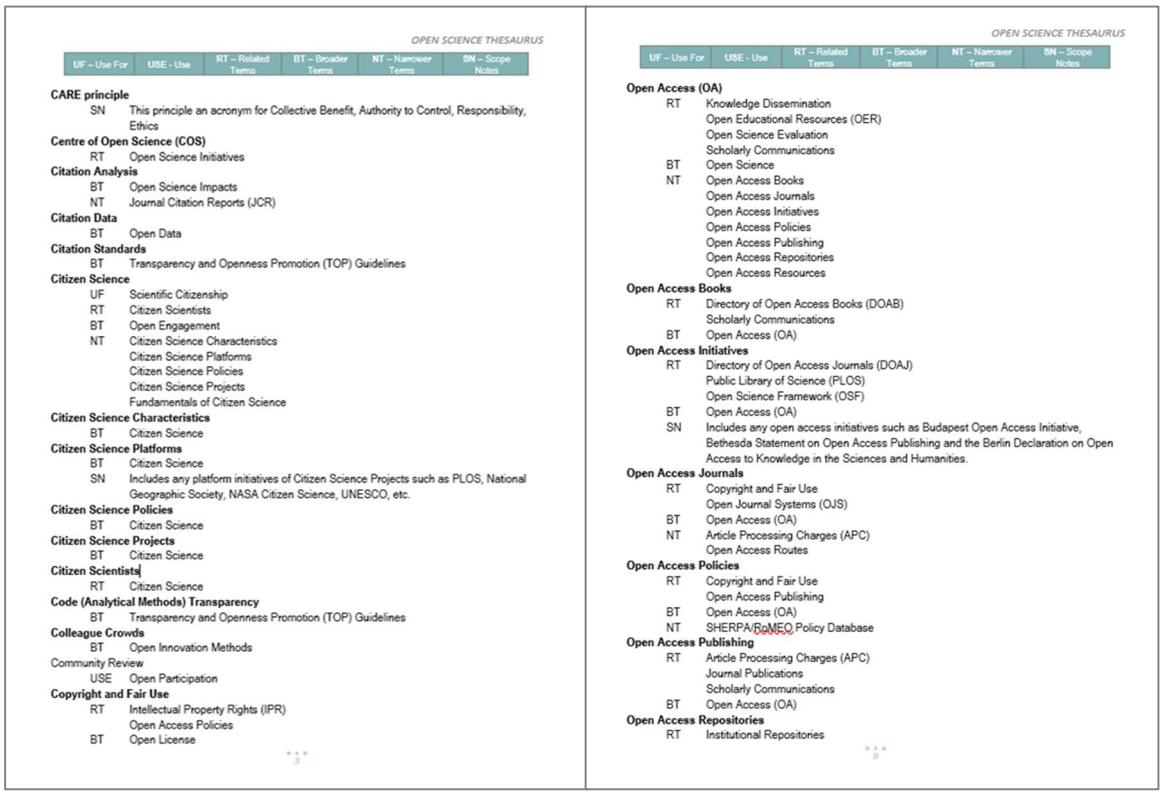
Figure 2: Hierarchical Tree Structure for Open Science Thesaurus

Each main topic is further expanded with narrow terms (NT) indicated by green boxes. The arrows in Figure 2 include labels denoting whether they represent narrow terms (NT) or related terms (RT). In the hierarchical tree, it is important to note that not all NTs and RTs are listed, because including all terms in the diagram could potentially lead to confusion and clutter. Therefore, only selected terms are represented in Figure 2. It is worth mentioning that while the terms depicted in the tree structure may not reach 250, the comprehensive thesaurus list does encompass the overall quantity of terms as initially specified.

For the thesaurus construction, a labeling system was implemented to enhance organization and comprehension. Each page displays a header label with specific abbreviations: 'UF' indicates "Use For," 'USE' signifies that a term should be used with another term listed, 'RT' represents "Related Term," 'BT' denotes "Broader Term," 'NT' stands for "Narrow Term," and 'SN' refers to "Scope Note." The scope note provides explanations or lists examples related to the terms that exemplify instance descriptors within its content (such as Git-Hub, Zenodo, OpenAire and Figshare under Open Science Providers).

To ensure clarity and ease of use, all terms are arranged alphabetically and grouped according to their main topic, subtopics, and further subcategories. This arrangement is determined based on literary warrant, user warrant, and the authors' understanding derived from thorough readings of the referenced sources. Only nouns have been selected for inclusion in this thesaurus. Instance relationships are also included showing relationships between a general category of things and events expressed by a common noun, and an individual instance of that category forming a class represented by a proper name (for example ORION Open Science).

The thesaurus consists of 250 terms or descriptors, organized across 21 pages, showcasing a straightforward hierarchical arrangement. This structure effectively demonstrates the interconnections among various terms associated with the Open Science concept. For instance, descriptors like 'citizen science', 'open access', and 'open science' are presented along with their corresponding broader and narrower terms. Examples of the descriptors display can be seen in Figure 3 and Figure 4.



OPEN SCIENCE THESAURUS

UF – Use For	USE - Use	RT – Related Terms	BT – Broader Terms	NT – Narrower Terms	SN – Scope Notes

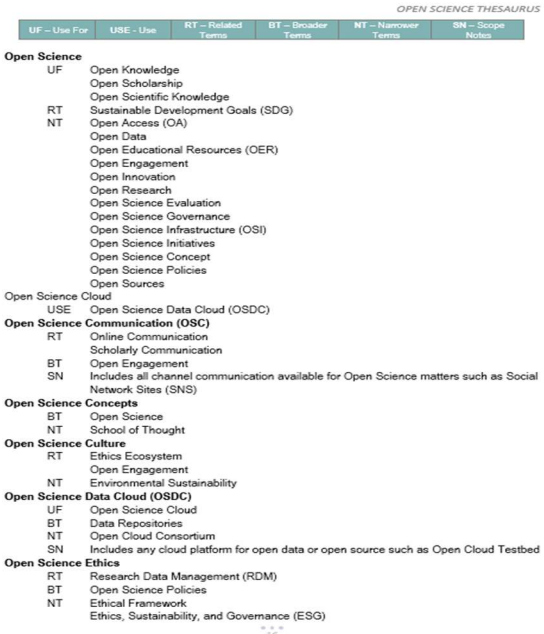
CARE principle
SN This principle an acronym for Collective Benefit, Authority to Control, Responsibility, Ethics
Centre of Open Science (COS)
RT Open Science Initiatives
Citation Analysis
BT Open Science Impacts
NT Journal Citation Reports (JCR)
Citation Data
BT Open Data
Citation Standards
BT Transparency and Openness Promotion (TOP) Guidelines
Citizen Science
UF Scientific Citizenship
RT Citizen Scientists
BT Open Engagement
NT Citizen Science Characteristics
Citizen Science Platforms
Citizen Science Policies
Citizen Science Projects
Fundamentals of Citizen Science
Citizen Science Characteristics
BT Citizen Science
Citizen Science Platforms
BT Citizen Science
SN Includes any platform initiatives of Citizen Science Projects such as PLOS, National Geographic Society, NASA Citizen Science, UNESCO, etc.
Citizen Science Policies
BT Citizen Science
Citizen Science Projects
BT Citizen Science
Citizen Scientists
RT Citizen Science
Code (Analytical Methods) Transparency
BT Transparency and Openness Promotion (TOP) Guidelines
Colleague Crowds
BT Open Innovation Methods
Community Review
USE Open Participation
Copyright and Fair Use
RT Intellectual Property Rights (IPR)
Open Access Policies
BT Open License

OPEN SCIENCE THESAURUS

UF – Use For	USE - Use	RT – Related Terms	BT – Broader Terms	NT – Narrower Terms	SN – Scope Notes

Open Access (OA)
RT Knowledge Dissemination
Open Educational Resources (OER)
Open Science Evaluation
Scholarly Communications
BT Open Science
NT Open Access Books
Open Access Journals
Open Access Initiatives
Open Access Policies
Open Access Publishing
Open Access Repositories
Open Access Resources
Open Access Books
RT Directory of Open Access Books (DOAB)
Scholarly Communications
BT Open Access (OA)
Open Access Initiatives
RT Directory of Open Access Journals (DOAJ)
Public Library of Science (PLOS)
Open Science Framework (OSF)
BT Open Access (OA)
SN Includes any open access initiatives such as Budapest Open Access Initiative, Bethesda Statement on Open Access Publishing and the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities.
Open Access Journals
RT Copyright and Fair Use
Open Journal Systems (OJS)
BT Open Access (OA)
NT Article Processing Charges (APC)
Open Access Routes
Open Access Policies
RT Copyright and Fair Use
Open Access Publishing
BT Open Access (OA)
NT SHERPA/RoMEO Policy Database
Open Access Publishing
RT Article Processing Charges (APC)
Journal Publications
Scholarly Communications
BT Open Access (OA)
Open Access Repositories
RT Institutional Repositories

Figure 3: Example of Pages from the Open Science Thesaurus



OPEN SCIENCE THESAURUS

UF – Use For	USE - Use	RT – Related Terms	BT – Broader Terms	NT – Narrower Terms	SN – Scope Notes

Open Science
UF Open Knowledge
Open Scholarship
Open Scientific Knowledge
RT Sustainable Development Goals (SDG)
NT Open Access (OA)
Open Data
Open Educational Resources (OER)
Open Engagement
Open Innovation
Open Research
Open Science Evaluation
Open Science Governance
Open Science Infrastructure (OSI)
Open Science Initiatives
Open Science Concept
Open Science Policies
Open Sources
Open Science Cloud
USE Open Science Data Cloud (OSDC)
Open Science Communication (OSC)
RT Online Communication
Scholarly Communication
BT Open Engagement
SN Includes all channel communication available for Open Science matters such as Social Network Sites (SNS)
Open Science Concepts
BT Open Science
NT School of Thought
Open Science Culture
RT Ethics Ecosystem
Open Engagement
NT Environmental Sustainability
Open Science Data Cloud (OSDC)
UF Open Science Cloud
BT Data Repositories
NT Open Cloud Consortium
SN Includes any cloud platform for open data or open source such as Open Cloud Testbed
Open Science Ethics
RT Research Data Management (RDM)
BT Open Science Policies
NT Ethical Framework
Ethics, Sustainability, and Governance (ESG)

Figure 4: Another Example of Page from the Open Science Thesaurus

DISCUSSION

The present paper focused on the construction of the Open Science Thesaurus as a means to address the lack of standardized vocabulary in the field of LIS related to Open Science. The discussion section will delve into the significance of the Open Science Thesaurus, its potential applications, and its implications for promoting effective communication within the Open Science community.

The construction of the Open Science Thesaurus involved several key steps that were meticulously executed. First, the Open Science Taxonomy developed by FOSTER was employed to identify the core concepts and usable descriptors for the thesaurus. This taxonomy served as a foundational framework for organizing and categorizing the terminologies associated with open science. Pontika et al. (2015) also specified the Open Science concept with the use of FOSTER taxonomy. By leveraging the existing taxonomy, the thesaurus was able to capture the essence of open science and provide a comprehensive vocabulary for indexing and retrieving related information.
Furthermore, an analysis of existing terminologies within the domain of Open Science was conducted. This analysis aimed to identify commonalities, gaps, and inconsistencies in the literature, policies, and guidelines pertaining to this domain. This step was crucial in ensuring that the Open Science Thesaurus encompassed a wide range of terms and adequately addressed the terminological challenges faced by researchers, librarians, and other stakeholders in the open science community.

The compilation of a comprehensive list of 250 terms related to Open Science was another integral component of constructing the Open Science Thesaurus. This extensive list was carefully selected to cover various aspects and comprehensive concepts of Open Science, including data sharing, collaboration, reproducibility, open source, open data, open educational resources, open peer review among others, and not merely open access publishing (Chakravorty et al. 2022). Chakaravoty (2022) stated that Open Science and "Open Access" should not be treated as interchangeable terms, although Open Access does play a vital role within the broader framework of the Open Science initiative. The inclusion of such a diverse range of terms enhances the usability and relevance of the thesaurus, making it a valuable resource for researchers and practitioners in the field.

Manually establishing preferential, hierarchical, associative, and other relationships between descriptors within the thesaurus was a critical step in facilitating effective information retrieval and communication. These relationships provide a structured framework for navigating the terminology associated with open science, enabling users to explore related concepts and discover relevant resources. Such relationships also contribute to the organization and accessibility of information, thereby promoting knowledge dissemination and collaboration within the open science community.

The Open Science Thesaurus may hold significant importance for the field of LIS and the Open Science community as a whole. Grubb and Easterbrook (2011) reported a lack of consensus

on several concepts that are central to transparency advocacy initiatives (such as open science, computational provenance, and reproducible research) aimed at making science and scientific artifacts accessible to a wider audience. By providing a standardized vocabulary, the thesaurus promotes a common understanding of open science concepts and facilitates effective communication among researchers, librarians, and other stakeholders. The lack of standardized terminology has been a persistent challenge in the domain of open science, hindering interdisciplinary collaboration and impeding the sharing of research output across geographic boundaries. The Open Science Thesaurus may address this challenge by establishing a shared language, enabling researchers to articulate and discuss Open Science ideas more effectively.

Moreover, the Open Science Thesaurus has practical implications for various stakeholders. For researchers, it serves as a valuable tool for conducting literature searches, enhancing discoverability of relevant resources, and promoting interdisciplinary research. Librarians who do not have the time to create this tool can utilize the thesaurus for indexing and organizing open science materials, thus improving the accessibility and retrieval of information. Policy makers and funding agencies can benefit from the thesaurus by gaining a better understanding of open science terminology and using it to inform policy development and decision-making processes. Beyond this, it belongs to any research organization that needs to index the Open Science concept.

As the Open Science field is still continuously updating in the current era, the terminologies will continue to be created and outdated; the thesaurus is still open for updating to the up-to-date terminologies. In addition to the construction of the Open Science Thesaurus in the English language, there is a potential opportunity to expand the thesaurus to include other languages (such as in the case of Loterre (2023)), such as Malay. Creating a thesaurus in Malay language would have several advantages and implications for the Malaysia research community.

CONCLUSIONS

In conclusion, the construction of the Open Science Thesaurus represents a significant contribution to the field of LIS and the open science community. By addressing the lack of standardized vocabulary, the thesaurus promotes effective communication, collaboration, and knowledge sharing within the open science community. By constructing well-structured Open Science Thesaurus, it can offer better knowledge sharing and standardize terms. The comprehensive list of terms, the establishment of relationships between descriptors, and the utilization of the Open Science Taxonomy all contribute to the usability and relevance of the thesaurus. Moving forward, further research and refinement of the thesaurus should be undertaken to ensure its continuous relevance and utility in an evolving open science landscape.

This thesaurus was developed with 250 terms as an output of a Masters in Library & Information Science course on Indexing, Abstracting and Thesaurus Construction. One of the limitations of the Open Science Thesaurus, being a project within a Master's program, is the constraint of time. As a result, the development and scope of the thesaurus may be influenced by the limited duration of the project assignment i.e., seven weeks. Due to time constraints, it might not be feasible to provide an exhaustive list of terms or delve into every aspect of Open Science in a comprehensive manner. It is recognized that despite the efforts, the sources selected might not cover the entirety of terms related to Open Science. Nevertheless, this compilation is viewed as a significant starting point for charting an expanding array of terms that continue to emerge as time progresses.

Another limitation of the Open Science Thesaurus, as a Master's program project, is the potential lack of expert consultation in the field. While the lecturer teaching the course may have expertise and experience in Open Science, the absence of additional expert input could restrict the thoroughness and accuracy of the thesaurus. Expert consultation plays a crucial role in ensuring that the terminology and relationships captured in the thesaurus align with current practices and understandings within the field of Open Science. Furthermore, engaging with reliable and trustworthy resources, such as authoritative organizations and reputable scholarly publications, would enhance the credibility and reliability of the thesaurus.

Nevertheless, the involvement with this course has led to the development of a deeper grasp of the fundamental principles that underlie Open Science. The process of constructing thesauri, which has involved reading and retrieving information, has enabled the authors to achieve a more comprehensive understanding of the subtleties within the domain of open science. It is important to acknowledge that Open Science is a rapidly developing field, and therefore, the thesaurus should be periodically updated and expanded to accommodate new terms and evolving aspects. By continuously incorporating expert consultation, utilizing reliable sources, and allowing for ongoing updates, the Open Science Thesaurus can maintain its relevance and usefulness as the discipline continues to grow and evolve over time.

REFERENCES

Bartling, Sönke, and Sascha Friesike. 2014. *Opening Science : The Evolving Guide on How the Web Is Changing Research, Collaboration and Scholarly Publishing*. Springer. https://doi.org/10.1007/978-3-319-00026-8.

Chakravorty, Nishant, Chandra Shekhar Sharma, Kutubuddin A. Molla, and Jitendra Kumar Pattanaik. 2022. “Open Science: Challenges, Possible Solutions and the Way Forward.” *Proceedings of the Indian National Science Academy* 88 (3): 456–71. https://doi.org/10.1007/s43538-022-00104-2.

Grubb, Alicia M., and Steve M. Easterbrook. 2011. “On the Lack of Consensus over the Meaning of Openness: An Empirical Study.” *PLoS ONE* 6 (8): e23420. https://doi.org/10.1371/journal.pone.0023420.

Loterre. 2023. "Open Science (Thesaurus)." Loterre. 2023. https://skosmos.loterre.fr/TSO/en/new.

Ogg, N. J., M. Sievert, Z. R. Li, and J. A. Mitchell. 1994. "Construction of a Medical Informatics Thesaurus." *Proceedings / the ... Annual Symposium on Computer Application [Sic] in Medical Care. Symposium on Computer Applications in Medical Care*.

Pontika, Nancy, Petr Knoth, Matteo Cancellieri, and Samuel Pearce. 2015. "Fostering Open Science to Research Using a Taxonomy and an ELearning Portal." In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-Driven Business*, 1–8. New York, NY, USA: ACM. https://doi.org/10.1145/2809563.2809571.

UNESCO. 2021. "UNESCO Recommendation on Open Science." https://doi.org/10.54677/MNMH8546.

Zarghani, Maryam, Leila Nemati-Anaraki, Shahram Sedghi, Abdolreza Noroozi Chakoli, and Anisa Rowhani-Farid. 2023. "Iranian Researchers' Perspective about Concept and Effect of Open Science on Research Publication." *BMC Health Services Research* 23 (1): 437. https://doi.org/10.1186/s12913-023-09420-9.