Executive Summary

The aim of the research is to find out the usage of controlled vocabularies in subject mapping of Open Access Repositories. A total of 34 repositories have been selected from OpenDOAR as the final sample size for this research. A consolidated controlled vocabulary has been constructed by merging preferred labels from eleven existing thesauri. The final consolidated controlled vocabulary consists of 91853 preferred labels after deduplication. Two similarity matching techniques—*Cosine Similarity and Jaccard Similarity*—were implemented using the Python Sci-Kit Learn library to map repository-assigned subject terms to controlled vocabulary entries. Moving beyond string matching, this study developed an ontological framework using Protégé to formally represent the semantic structure of subjects as manifested in OARs. This approach allows for a shift from linear keyword-based retrieval to a networked information discovery paradigm, where users can traverse conceptual links and uncover contextually relevant resources. This research found that, among the 34 OARs studied, only 6 repositories utilised any form of structured subject organisation. Of these, 5 employed classification schemes, only GenderOpen relied on a standalone subject metadata standard. From the remaining, 24 repositories, exhibited ad hoc arrangements, often alphabetical listings of keywords, lacking taxonomic structure or semantic clarity. After similarity matching between repository subject index terms and controlled terms from the consolidated controlled vocabulary, the terms were categorised into 'Controlled' and 'Uncontrolled' terms. After evaluation of the effectiveness of the consolidated controlled vocabulary the research found that, Controlled terms

consistently yielded higher precision and F1 scores than uncontrolled terms. The PR-AUC curves demonstrated that repositories adopting controlled vocabularies significantly outperformed their counterparts in terms of both recall and relevancy. For repositories like UK Data Service and IDEP, controlled vocabularies enhanced both interoperability and user satisfaction. The OWL-based ontology provided a machine-readable, semantically rich model that can be integrated with repository platforms to support advanced query mechanisms. The Neo4j Knowledge Graph enabled dynamic, visual exploration of subject interconnections, serving as a scalable prototype for metadata navigation. This study concludes that the adoption of controlled vocabularies for subject indexing in Open Access Repositories is not merely beneficial—it is essential for achieving semantic precision, reducing information overload, and enhancing user experience. The systematic arrangement of subjects, when guided by well-maintained and domain-specific controlled vocabularies, can bridge the gap between information seekers and relevant content, especially in multilingual and interdisciplinary research contexts. Moreover, the integration of ontology and knowledge graph technologies elevates metadata beyond static labels to dynamic knowledge structures, enabling deeper semantic exploration and context-aware retrieval. This paradigm holds immense promise for digital repositories, where users increasingly demand intuitive interfaces and intelligent discovery systems. However, the study also recognises its limitations. The exclusive use of preferred labels from each thesaurus excludes synonym rings and multilingual equivalents, which could further enhance semantic matching. Future research may expand this model to incorporate full SKOS-based semantic frameworks, multilingual mappings, and user-centric evaluations.

Ultimately, this thesis makes a substantial contribution to the fields of Knowledge Organisation, Metadata Engineering, and Digital Library Science. It advocates for a future in which Open Access Repositories adopt structured, interoperable, and intelligent metadata systems, guided by controlled vocabularies, semantic models, and graph-based architectures. In doing so, it aligns repository practice with the broader goals of Open Science, fostering equitable access, contextual discovery, and sustained knowledge preservation.