Technical Aspects of Managing a Large-Scale ETD Repository - Insights from Shodhganga

Swapnil Patel[1] and Yatrik Patel[2]

[1]Scientist-D (CS), INFLIBNET Centre, Gandhinagar
[2]Scientist-E (CS), INFLIBNET Centre, Gandhinagar

Abstract

The Due to technological advancement in recent years, Electronic Theses and Dissertations (ETDs) are becoming increasingly popular. On the other hand, due to the complexities associated with it, managing and maintaining large-scale ETD repositories is also becoming a challenging task. This paper thoroughly explores the multifaceted challenges associated with the management of a large-scale ETD repository, drawing invaluable insights from the experience of Shodhganga. It examines critical areas such as the selection of the platform and technology, software customization, workflow orchestration, quality control, user-friendliness, performance optimization, security measures and redundancy and backup strategies.

The paper not only highlights the complexities posed by these challenges but also presents innovative and effective solutions to address them. It emphasizes the importance of choosing the right platform and technology for scalability and performance. It also discusses the significance of software customization for academic institutions. Workflow efficiency, quality control, and user-friendly interfaces are highlighted for repository integrity and accessibility. Performance optimization, security measures, and redundancy strategies are detailed to protect academic content and ensure availability in unexpected situations.

This paper will serve as a valuable resource for institutions and organizations tasked with managing large-scale ETD repositories. By understanding these challenges and implementing the suggested solutions, repository managers can facilitate the seamless dissemination of academic research.

Keywords: ETD Challenges, Large-scale repository, Managing ETDs, Shodhganga

1. Introduction

Electronic Theses and Dissertations (ETDs) are digital versions of research works, such as theses or dissertations, submitted by scholars in digital format. The increasing popularity of ETDs in recent years can be attributed to advancements in technologies for creating, storing, and sharing digital documents. Digital

Corresponding Author: Swapnil Patel, Email: swapnil@inflibnet.ac.in and Yatrik Patel, Email: yatrik@inflibnet.ac.in

repositories play a pivotal role in enhancing the accessibility of ETDs, offering storage and convenient access for various stakeholders.

ETDs are typically administered by universities or academic institutions. Globally, many universities now require researchers to submit their theses and dissertations in electronic format. Overall, ETDs have become a crucial tool for researchers, offering a convenient and easily accessible means of sharing and accessing research work.

Shodhganga stands as one of the largest digital repositories for electronic theses submitted to universities in India. It is managed and maintained by the Information and Library Network (INFLIBNET) Centre, which is an Inter-University Centre under the University Grants Commission, operating within the Ministry of Education, Government of India. Shodhganga facilitates the submission of electronic theses from research scholars at Indian universities and serves as a platform to make these theses accessible to researchers worldwide. The repository contains metadata and full-text theses in various disciplines. Shodhganga also provides researchers with the capability to search for specific theses based on various parameters and download full-text copies.

Shodhganga has firmly established itself as a valuable resource for researchers in India and abroad, offering access to a wide range of electronic theses from Indian universities.

2. Objectives

The aim of this paper is to investigate and provide insights into the technical aspects of overseeing a large-scale Electronic Theses and Dissertations (ETD) repository. It also aims to propose effective technical strategies and solutions to enhance the management of such a repository, drawing from the experience of managing the Shodhganga repository.

The exponential growth in ETD submissions has presented significant challenges in managing this extensive collection. These challenges encompass selecting the appropriate platform and technology, customizing the software, handling workflow processes, ensuring quality control, enhancing user-friendliness and performance, addressing security concerns, implementing redundancy measures, establishing backup strategies, and more. This entails evaluating the infrastructure and technical prerequisites required to accommodate the increasing number of submissions, ensuring efficient storage, access, and retrieval of documents, and establishing robust data preservation strategies. Furthermore, the paper will delve into the experiences of developers and system administrators, analyzing the challenges they encounter in developing, maintaining, and managing the Shodhganga platform.

By addressing these challenges and proposing practical solutions, this paper aims to contribute to the advancement of ETD management in large-scale repositories, using Shodhganga as a case study. The solutions implemented in this case study can provide valuable insights for administrators, developers, and

stakeholders involved in designing and managing digital repositories, facilitating the efficient dissemination and preservation of academic research on a broader scale.

3. Issues-Challenges in Managing Large-Scale ETD (Shodhganga)

Being one of the largest Electronic Theses and Dissertations (ETD) repositories in India, Shodhganga confronts a multitude of challenges encompassing platform and technology selection, customization, workflow management, quality control, user-friendliness, performance enhancement, security, redundancy, backup strategies, and other related aspects. This study endeavors to identify and analyze these challenges and propose effective strategies for the efficient management of large-scale ETDs. The challenges and issues can be broadly categorized as follows:

3.1 Platform and Technology Selection

The choice of a suitable platform and technology infrastructure is pivotal to accommodate the increasing influx of ETD submissions. Factors such as openness, scalability, performance, compatibility, and cost-effectiveness must be meticulously considered during the selection process.

3.2 Customization of the Platform

Tailoring the platform to meet the specific requirements of users and institutions is a formidable challenge. Customization efforts should encompass user interface enhancements, multilingual support, and flexible metadata management to ensure an intuitive user experience. It also involves streamlining the workflow, including submission and review processes, which can be intricate due to the coordination of multiple stakeholders and the need for task automation.

3.3 Quality Control

Upholding high data quality is paramount to maintain the integrity and credibility of ETDs. Implementing quality control measures to address issues such as metadata accuracy and data validation becomes a significant challenge when dealing with a vast collection.

3.4 User-friendliness and Performance Enhancement

Ensuring user-friendliness is crucial for ETD repositories to enable researchers and students to easily access and navigate the repository, conduct efficient content searches, and retrieve information promptly. Concurrently, enhancing performance is vital to optimize the speed and responsiveness of the ETD platform, minimizing delays when users interact with the extensive academic works.

3.5 Security

Safeguarding ETDs against unauthorized access, data breaches, and cyber threats is a paramount challenge. The implementation of robust security measures, including access controls, is indispensable to maintain the confidentiality and integrity of ETDs.

3.6 Redundancy and Backup Strategies

Establishing redundancy and backup strategies is imperative to prevent data loss. This involves designing and implementing redundant systems, conducting regular backups, data replication, and formulating disaster recovery plans.

Addressing these challenges necessitates a blend of technical expertise, effective policies, and continuous monitoring. The subsequent section proposes strategies based on real-life experiences, involving scalable and secure infrastructure implementation, adoption of open standards, and regular review and update of security protocols. By effectively managing these challenges, any ETD repository, including Shodhganga, can ensure manageability, platform sustainability, long-term accessibility, and preservation of invaluable research contributions.

4. Solutions and Enhancements

This section of the paper addresses several critical issues and challenges related to the effective management of Electronic Theses and Dissertations (ETDs) on a large scale. In this write-up, the focus will be on the technical solutions proposed within the context of Shodhganga to overcome the challenges discussed in the previous section. The solutions encompass various aspects, including Platform and Technology Selection, Customization of the platform, Quality Control, User-friendly Web Interface and Performance, Security, and Redundancy and Backup Strategies for improved ETD management.

4.1 Platform and Technology Selection

The selection of a suitable platform and technology infrastructure is pivotal for efficiently managing a large-scale ETD repository. Shodhganga has chosen to adopt the robust and scalable open-source platform known as DSpace, which is well-equipped to handle the storage and retrieval of a vast volume of documents with high availability and scalability.

DSpace is an open-source digital repository software initially developed by MIT and HP Labs, and it is currently maintained by the Duraspace Foundation, which has merged with Lyrasis. DSpace is specifically designed for the management, preservation, and dissemination of various types of digital content, including scholarly papers, theses, datasets, and more. It provides an intuitive and user-friendly interface that simplifies both administrative tasks and user interactions.

One of DSpace's notable strengths is its customization capabilities, enabling organizations to tailor the repository to their specific requirements regarding the interface, metadata, workflows, and branding. DSpace excels in metadata management, supports seamless integration with external systems, and offers comprehensive documentation for effective setup and management. It is designed for scalability, security, and digital preservation while also accommodating various plugins and extensions to enhance functionality.

Furthermore, DSpace adheres to modern technology standards and provides advanced search and discovery features, making it a compelling choice for institutions seeking a robust and adaptable repository solution.

While other repository software options have their merits, the multitude of features and the open platform offered by DSpace make it the preferred choice for establishing a Shodhganga repository.

4.2 Customization of the Platform

After selecting the platform and technology, the ease of customizing the platform is another crucial aspect. For Shodhganga, a primary challenge was to develop a mediated thesis upload mechanism that allows institutions to efficiently upload thesis metadata and full-text documents. Additionally, it was expected that the system would streamline the process of merging uploaded theses after due verification. To meet this challenge, Shodhganga has implemented a user-friendly, tailor-made data entry system.

The default submission process in DSpace involves multiple steps for metadata and full-text file submission. However, Shodhganga's submission interface simplifies this process into fewer steps, as depicted in [Figure 1].

Workflow management also plays a vital role in handling large-scale ETDs. Shodhganga ensures streamlined processes for submission, review, approval of theses, and their integration into the repository. Furthermore, Shodhganga offers a metadata import feature that supports a specified standard format. This feature allows universities to effortlessly upload pre-defined or exported metadata from their existing repositories into Shodhganga. Users can then easily associate full-text files with the imported metadata.

The Shodhganga data entry platform includes an advanced metadata import feature that simplifies the process of importing Dublin Core standard metadata, defined in CSV format, into the system. This streamlined procedure enables users to seamlessly link full-text files with the imported metadata. This customization allows for seamless integration with existing workflows and facilitates efficient metadata management. The functionality is implemented through the development of a CSV to PostgreSQL database import function.

The data entry platform was meticulously designed and developed by analyzing DSpace's existing workflow and the specific needs of universities. The system is organized according to the hierarchical structure of university departments. Shodhganga administrators have the capability to create accounts and delegate submission rights to university coordinators within their respective institutions. These coordinators are empowered to manage tasks such as metadata creation and the uploading of full-text files. The tools and technologies used for this system include Java Servlets, JSP (Java Server Page), XML, CSS, and PostgreSQL as the database management system.

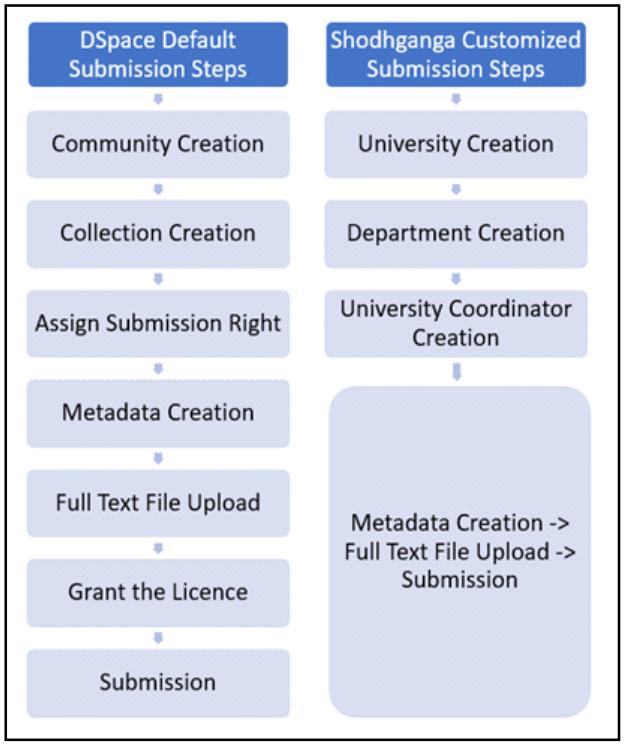

Figure: 1

4.3 Quality Control

Shodhganga has implemented several quality assurance mechanisms to maintain the integrity of its repository. These mechanisms include restrictions on special characters in metadata, the selection of standard subject keywords from a controlled vocabulary, and checks for duplicate titles.

4.3.1 Special Characters Restriction in Metadata

When universities or institutions submit metadata (information about theses or documents) to Shodhganga, the platform enforces rules that prevent the inclusion of special characters. Special characters can sometimes lead to technical issues or formatting problems. By restricting the use of special characters, Shodhganga ensures that the metadata is clean and aligns with the platform's standards. This was achieved through the implementation of custom client-side JavaScript validation routines, which were integrated with input forms.

4.3.2 Selection of Standard Subject Keywords from Controlled Vocabulary

Controlled vocabulary consists of a predefined and standardized list of terms used to describe subjects or topics. In Shodhganga, the system prompts users to select subject keywords from a predefined list. This practice helps maintain consistency and enhances search accuracy. For instance, if a user is searching for

research on "Computer-aided design," having a controlled vocabulary ensures that all related documents are tagged with the same term, avoiding confusion.

Shodhganga's controlled vocabulary is established by combining Library of Congress Subject Heading (LCSH) keywords with other standardized terms. This was accomplished by modifying the data input sheet, which fetches values from a carefully curated list of LCSH keywords obtained from a structured source, as depicted in [Figure 2].

Figure : 2

4.3.3 Duplicate Title Checks

When documents or theses are submitted to Shodhganga, the system conducts check to identify duplicate titles. This step is crucial to prevent inadvertent duplication of content within the repository. Duplicate titles can arise when different institutions or authors submit identical works or when there are typographical errors in the titles.

Upon uploading new theses to Shodhganga, users have the ability to search for titles that resemble those already present in the repository. This process aids in the detection of duplicates and ensures that only unique titles are retained within the repository.

The figure illustrates titles similar to "Artificial Intelligence." The functionality for searching similar words was incorporated using custom SOLR indexes. The search results also provide details about the researcher, guide, and Handle ID for further reference within the repository, as depicted in [Figure 3].

Sr.No.	Title	Researcher	Guide	Handle Id
1	Philosophical perspectives in artificial intelligence	Rajmohan, S	Dr Radhakrishnan, C V	10603/108386
2	Artificial Intelligence in Business Decision Making	Madhavi	Vijay Kumar	10603/384942
3	Automated guided vehicle artificial intelligence	Agrawal, Himanshu	Dewangan, M S	10603/44162
4	Mechanism for Artificial Intelligence in Different Systems	Mamta Sharma	Dr Kavita	10603/482920
5	Artificial Intelligence and Machine Learning Techniques for Diabetes Healthcare	Mastoli M M	Pol U R	10603/331628

Figure: 3

In summary, these quality control mechanisms are essential for maintaining the reliability, consistency, and overall quality of the content in Shodhganga. They help prevent technical issues, ensure standardized and accurate subject categorization, and eliminate duplicate entries, thus enhancing the integrity of the repository.

4.4 User-friendly Web Interface and Performance

Shodhganga's user-friendly web interface refers to the design and layout of the website, which has been carefully crafted to ensure ease of use for all users. It prioritizes the user's experience, making navigation and interaction intuitive and straightforward. Whether a researcher is searching for research or an administrator is managing the repository, users can easily find what they need and efficiently perform tasks.

Compared to the default DSpace homepage, Shodhganga's homepage has been designed to be more intuitive. It prominently displays key information, such as the total number of full-text theses, the number of universities with signed MOUs, the total number of contributing universities, and a list of universities along with the count of full-text theses uploaded by each. These indicators are dynamic in nature, providing real-time information.

Social media sharing plugins have been seamlessly integrated into the Shodhganga interface, allowing users to share thesis records on various social media platforms. Additionally, Shodhganga offers the facility to download thesis metadata in BibTeX format, providing essential information for citation purposes.

Shodhganga hosts full-text thesis files that are organized into separate chapters. However, when users download the full text, they can conveniently obtain all the chapters (i.e., all the files associated with a particular thesis) with a single click in ZIP file format. This functionality is achieved through a feature that searches the PDF files on a page, compiles them into a single ZIP file, and initiates the download. This additional program logic has been seamlessly implemented as an add-on to the existing interface where the thesis record is being displayed.

Shodhganga offers a powerful full-text search facility, enabling users to find specific results even if the search term is not explicitly mentioned in the metadata but is present in the full text.

A Persistent URL is a web address designed to remain stable and accessible over time, ensuring reliable and permanent access to online resources, even as websites and content may change or move. Persistent URLs are commonly used in digital libraries, archives, and scholarly repositories to provide lasting access to academic publications, datasets, and other valuable online content.

Shodhganga has implemented Persistent URLs using the Handle Server (Handle.net registry). The implementation of the Handle Server in Shodhganga provides numerous advantages in terms of accessibility. It assigns persistent identifiers (Handles) to digital objects, ensuring stable and accessible links over time. This enhances content discoverability and facilitates effective citation. Handles are interoperable, making content accessible through assistive technology tools and improving accessibility for individuals. They provide consistency in accessibility, long-term availability, and compliance with international standards,

aligning Shodhganga with global best practices. The unique handle number "10603" has been configured in Shodhganga, ensuring the permanence and reliability of links to its digital objects.

Shodhganga places significant emphasis on achieving optimal performance. One of the key strategies employed to achieve this is the meticulous configuration of the PostgreSQL database and DSpace configuration files. This careful configuration enables Shodhganga to efficiently manage its database connections.

While DSpace does offer an Auto-indexing facility, the index generation in Shodhganga is consistently monitored to ensure its effectiveness and efficiency.

4.5 Security

Security is of paramount importance in the management of large-scale ETDs, and Shodhganga places it as a top priority. Shodhganga employs stringent security measures to safeguard its repository, including the restriction of malicious bots and the implementation of access controls to manage user connections.

To combat malicious activities and suspicious connections, Shodhganga has configured Fail2Ban as a preventive measure. Fail2Ban is an open-source software designed to enhance the security of computer systems and servers, particularly against brute-force attacks and unauthorized access attempts.

Fail2Ban offers several key features, including continuous monitoring of server-generated log files to detect suspicious patterns like repeated login failures. When a predefined rule is matched, such as reaching a threshold of failed login attempts, Fail2Ban dynamically updates firewall rules to block malicious traffic sources, temporarily or permanently denying them access. Its high configurability allows system administrators to define custom rules and thresholds for different services, adapting it to specific security needs. Furthermore, Fail2Ban can be configured to send email notifications to administrators when specific events occur, serving as alerts and providing details about potential security threats and the corresponding actions taken.

To prevent excessive site access requests and avoid overloading from malicious bots, Shodhganga utilizes the robots.txt file, which follows the Robots Exclusion Protocol. Web crawlers, also known as bots or spiders, play a crucial role in indexing and categorizing web content. These automated agents are employed by search engines and other services to traverse the internet, collect data, and make it accessible to users. However, on high-traffic sites, these bots can sometimes create performance issues.

The robots.txt file is placed in the root directory of Shodhganga and communicates with web crawlers. It serves as a set of instructions, informing these crawlers which parts of the Shodhganga website they are allowed to access and index and which parts they should avoid. This file acts as a virtual "Keep Out" sign for web robots, assisting administrators in managing the visibility of their content.

A properly configured robots.txt file is a valuable tool for ETD website administrators, helping guide web crawlers and protects content. When used thoughtfully, it enhances content control, resource management,

and privacy protection while contributing to a more effective and efficient online presence. However, it should be employed judiciously to avoid unintended consequences on website visibility and SEO.

The Shodhganga database and Solr indexes are configured to operate exclusively on a local port. This approach enhances system security by restricting external access. This localized setup effectively minimizes potential vulnerabilities and unauthorized network access. Achieving this level of security has involved the configuration of Apache's Proxy_Pass and AJP directives.

Apache's Proxy_Pass and AJP directives play a crucial role in this setup. They facilitate secure communication between different components of the system and ensure that sensitive data remains protected. This configuration is designed to enhance security by limiting access to the local environment.

Additionally, comprehensive security measures are overseen by the Network department to further safeguard sensitive data. These measures include the implementation of firewalls, regular security checks, and other practices aimed at fortifying the security posture of Shodhganga.

4.6 Redundancy and Backup Strategies

To mitigate the risk of data loss, Shodhganga has implemented redundancy and backup strategies. These measures are designed to ensure data resilience and facilitate disaster recovery. Shodhganga maintains redundant servers and conducts regular data backups to safeguard its valuable contents.

At predefined intervals specified in the crontab (scheduled job), Shodhganga's backup is synchronized with an additional server and storage using the rsync command.

Rsync ("remote synchronization") is a robust and versatile command-line utility used for efficient copying and synchronizing of files and directories between two locations. It is widely utilized in Unix-like operating systems, including Linux and macOS, and is also available for Windows through various ports and third-party applications.

Some key features of rsync include efficient data transfer, delta transfer (transferring only updated data), support for both local and remote operations, versatile usage, and incremental backups.

Additionally, Shodhganga has configured the pg_dump command to create scheduled database backups at specified intervals using a cronjob.

To address disaster recovery requirements comprehensively, Shodhganga is in the process of establishing a mirror site at a remotely located data center. This initiative not only helps distribute the main site's workload but also ensures redundancy, further enhancing the overall resilience of the system.

5. Conclusion

In conclusion, this paper has delved into the challenges associated with the management of large-scale Electronic Theses and Dissertations (ETDs) based on the experiences of Shodhganga. The insights provided

shed light on various practical aspects that are critical for efficient ETD management. These aspects encompass platform and technology selection, customization, workflow management, quality control, user-friendliness, performance enhancement, security, redundancy, and backup strategies.

Addressing these challenges holds paramount importance in achieving effective ETD management. It not only enhances the overall user experience but also facilitates broader access to valuable research findings. By comprehending and proactively addressing the technical intricacies related to platform maintenance, customization, workflow streamlining, quality control, user-friendly interfaces, performance optimization, security measures, and robust redundancy and backup strategies, digital repositories can ensure the seamless management of large-scale ETDs. In doing so, they significantly contribute to the advancement of academic research and the dissemination of knowledge.

References

Create and submit a robots.txt file Google Search Central. (n.d.). Google for Developers. https://developers.google.com/search/docs/crawling-indexing/robots/create-robots-txt

Fail2ban. (n.d.). https://www.fail2ban.org/wiki/index.php/Main_Page

Handle.Net Registry. (n.d.). https://www.handle.net/

Mitchell, S. (n.d.). DSpace Home - DSpace. DSpace. https://dspace.lyrasis.org/

Mikeal, A., Creel, J., Maslov, A., Phillips, S., Leggett, J., & McFarland, M. (2009, June). Large-scale ETD repositories: A case study of a digital library application. In Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries (pp. 135-144). https://dl.acm.org/doi/pdf/10.1145/1555400.1555423

mod_proxy_ajp - Apache HTTP Server Version 2.4. (n.d.). https://httpd.apache.org/docs/2.4/mod/mod_proxy_ajp.html

Rsync. (n.d.). Linux Man Pages. https://linux.die.net/man/1/rsync

Sahu, R. R., & Karadia, A. (2012). Comparative study of an open source digital library software: Dspace, Greenstone and Eprint. Int. J. Comput. Appl, 59, 16.

Tramboo, S., Shafi, S. M., & Gul, S. (2012). A study on the open source digital library software's: special reference to DSpace, EPrints and Greenstone. arXiv preprint arXiv:1212.4935. https://doi.org/10.5120/9629-4272