Design and Development of a Web Based Digital Repository for Scholarly Communication: A Case of NM-AIST Tanzania

Wasiwasi J. Mgonzo*, Zaipuna O. Yonah**

ARTICLE INFO

Article history:
Received 23 September 2014
Revised 24 November 2014
Accepted 01 Desember 2014

Keywords:
Institutional Repositories,
Research Infrastructure,
Scholarly Communications,
Web-based,
Open Source Software

ABSTRACT

Institutional repositories are essential research infrastructures for research-based universities. A properly dimensioned institutional repository has the potential to increase research impact and enhance the visibility of an institution through its scholarly outputs. The aim of the study reported in this paper was to design and develop a web-based digital repository for scholarly communications using NM-AIST as a case study. The system was developed using open source software. Findings obtained from system validation tests show that the system is a viable solution to the major challenges encountered in the management and sharing of scholarly information at the institution.

1. Introduction

An institutional repository (IR) is a system that collects, preserves, manages, and provides access to intellectual products of a community (Hockx, 2006). Institutional intellectual products may include faculty work, student theses and dissertations, e-journals, datasets, and so on. IRs provide a mechanism for an institution to showcase its scholarly output, centralize and introduce efficiencies to the stewardship of digital documents of value, and respond proactively to the escalating crisis in scholarly communication (Gibbons, 2004). The availability of open-source repository systems has encouraged and led to the proliferation of IRs worldwide, particularly among academic and research institutions. The following are the benefits behind establishing IRs:

i. Providing open access to research outputs of institutions to a worldwide audience both within and outside the institution;
ii. Maximizing the impact and enhancing the visibility of the scholarly works produced at the institution;
iii. And managing and measuring the impacts of institutional research and teaching activities;

The growing trend towards online scholarly communication and lack of scholarly content management systems among universities has made digital repositories more important for the collection

* The Nelson Mandela African Institution of Science and Technology, Tanzania (mgonzow@nm-aist.ac.tz)
** The Nelson Mandela African Institution of Science and Technology, Tanzania (zaipuna.yonah@nm-aist.ac.tz)
International Journal of Knowledge Content Development & Technology, 4(2): 97-108, 2014.
http://dx.doi.org/10.5865/IJKCT.2014.4.2.097

and distribution of scholarly materials (Budapest, 2002; Chan, 2004; Lynch, 2003). Today, digital repositories are used at academic institutions to store and disseminate scholarly outputs of universities (Lynch & Lippincott, 2005).

In the beginning, repository systems were developed as a hosted online solution for collecting, preserving, and disseminating scholarship of universities, colleges, and other research institutions. Recently, software have been developed and repositories have evolved into a publication platform for institutions to showcase their scholarship including articles, books, theses, dissertations, and journals. The number of repository platforms has also increased, and the choice of which to use depends on benefits and technical criteria (Bankier, 2014).

The idea behind establishment of repository software platforms was that the software be open source and locally installed. This approach offered unlimited flexibility for developers to customize them, which made interoperability a problem. The platforms have now been enhanced to include features that require no extra customization. Potential high maintenance costs also led many institutions to move to open source software.

Today, institutional repository platforms have richer feature sets never witnessed before. The software are openly available and have a wider support from the global community of developers. Universities are free to compare different platforms depending on the features that best address their needs and that would make their repositories more successful (Armbruster & Romary, 2009). Generally, an institutional repository centralizes, preserves, and makes accessible the scholarly works generated by academic institutions, and form part of a larger global system of repositories which are indexed in a standardized way and searchable using a common interface (Sefton, 2009).

While reviewing the status of open access repositories in Tanzania, Mgonzo and Zaipuna (2014a) reported that attitudes and web usage behaviour of users have an impact on the performance of IRs. In a related work, lack of resource sharing policy and lack of proper digital asset management systems have been pointed out as the major factors that hinder the adoption of open access repositories in Tanzania (Mgonzo & Zaipuna, 2014b). As a response to these challenges, this paper presents the design and implementation of DSpace@NM-AIST, a web-based digital repository for scholarly communication proposed for The Nelson Mandela African Institution of Science and Technology (NM-AIST). The system is implemented using DSpace repository software. It is not the intention of this paper to show how well the design of the proposed system is, but how well it addresses the challenges identified and how usage behaviour affects its success.

The paper is organised into seven sections. An overview of the Dspace Repository System is given in Section 2. Section 3 covers Materials and Methods, Section 4 presents the results, Section 5 presents System Design, Section 6 covers System Implementation, and Section 7 has the Conclusion and Recommendations.

2. Overview of DSpace Repository System

DSpace is an open source repository development software typically used for creating open access repositories for scholarly and published digital content. A repository is a system for delivering

digital content to end-users. Global statistics show that, DSpace is the most widely used open source repository software for institutional and open access repositories. High use of the software has been observed in universities and research-based institutions as a way to provide access to research output, scholarly publications, and more (Smith et al., 2003). Usage Statistics show that out of 2792 repositories worldwide, 1159 (42%) are using Dspace software (OpenDOAR, 2014). This is the main reason why Dspace software was chosen to implement Dspace@NM-AIST. Also the suitability for a stable repository system is another factor that favoured its choice (Lewis, de Castro, & Jones, 2012).

DSpace supports Qualified Dublin Core metadata by default and is oriented towards open standards and protocols, and therefore, fully supports the Open Archive Initiative for Metadata Harvesting Protocol (OAI-PMH). The search engine is based on Lucene, a popular and powerful open-source engine. In fact, the DSpace software has proven to be a solid repository platform since its launch. That is why it remains promising and competitive amidst other software platforms, like its follower Eprints which currently has 381 (14%) repositories out of 2792 worldwide (OpenDOAR, 2014).

Content in DSpace is at the highest level organized into communities. At an institutional level, communities could be departments, labs, research centers, or schools. Communities, in turn, each have collections that contain logically-related material, the items or files. For example, a technical report series might be a collection, which contains items, a grouping of content and metadata that users access as scholarly materials. Items may take the form of a research article, theses or dissertations, or a technical report together with a dataset used in experiments described by the report. Communities and Collections are used within DSpace to provide the repository with an easy to navigate structure often representing an institution's organizational makeup.

The Dspace repository architecture follows a three-layer model, which is composed of the presentation layer, a repository management layer, and a storage layer (Gao & Krogstie, 2010). The storage layer consists of a relational database for storing metadata and a bitstream storage module for storing content data. The repository management layer contains the modules that perform the business logic of the system. The presentation layer of the Dspace platform is the services layer representing the Web user interface (Bass et al., 2002). Figure 1 illustrates the Dspace system architecture.

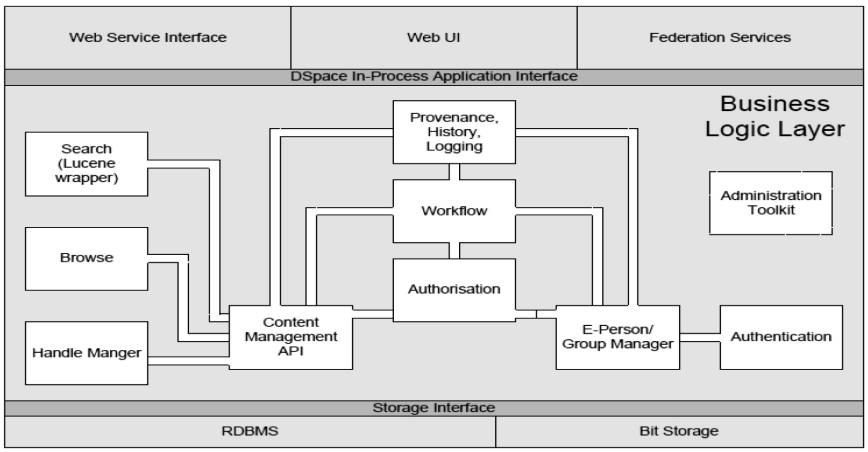

Fig. 1. DSpace system architecture [Adopted from DSpace, 2009]

3. Materials and Methods

This section describes the study methodology and materials used. A case study approach was adopted. A comparative survey of repository software was undertaken to select the best repository software to use. Data used in the survey were collected from the Directory of Open Access Repositories (OpenDOAR). Visual Paradigm for UML software was used to describe the system overview through DFDs. Several supporting open source software were chosen and integrated in the repository system. These include apache Tomcat, apache ant, apache maven, PostgreSQL relational database, and Java Development Kit (JDK).

4. Results

In choosing the best software to implement the repository system at NM-AIST, various literature were consulted. Findings in Bankier (2014) and Armbruster and Romary (2009) reports high usage of DSpace followed by Eprints which is next to DSpace in popularity although it is not as widely used as DSpace. Results from the survey conducted revealed that out of 2729 repositories worldwide, Dspace has the highest usage (42.3%) followed by EPrints (14.0%). The results are summarized in Table 1.

Table 1. Distribution of Repository Software among repositories worldwide

S/N	Software Name	Number of Repositories	Percentage
1.	DSpace	1159	42.3%
2.	Eprints	381	14.0%
3.	Digital Commons	127	5.0%
4.	dLibra	60	2.2%
5.	Greenstone	53	1.9%
6.	CONTENTdm	50	1.8%
7.	HTML	39	1.4%
8.	Fedora	34	1.2%
9.	Diva-Portal	32	1.1%
10.	Open Repository	24	0.9%
11.	Others	770	28.2%
	TOTAL (N)	2729	100%

Source: OpenDOAR 2014

5. System Design

A system design should specify in detail how the parts of an information system should be implemented. For the case of Dspace@NM-AIST, dataflow diagrams (DFD) were used in the design

part. DFDs are one of the main methods for analyzing data oriented systems because they emphasize the logic underlying the system. DFDs are common tools for structuring information (Valacich, George, & Hoffer, 2012). They generally illustrate how data is processed by a system in terms of inputs and outputs, and are used to create an overview of the system and can be also used for visualization of data processing. There are two common notations available for representing DFDs. These are the *Gene-Sarson* notations and the *Yourdon & Coad* notations. In this paper, *Gane-Sarson* diagram notations are adopted. *Gane-Sarson* diagrams show the storage, exchange, and alteration of data and resources throughout the diagram, which is a capability that many other diagrams do not possess. Figure 2 shows the DFD components.

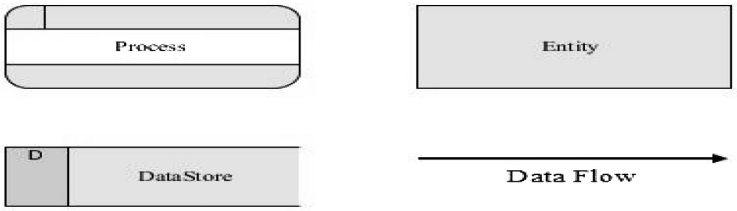

Fig. 2. DFD Components

Processes represent either a whole system, a sub-system, or work being done (activity). External entities represent people or organizations outside the entire system or sub-system, and usually show the initial source and final recipient of data and information. Data stores represent data stores such as computer files or database transaction files, set of tables or manual files or records. Data flows shows data movement through the system. The data could be about a person, place, or thing. Arrowheads are used to represent the direction of movement of data. Double arrowheads are used when a process both reads and updates data on the same table or file. The Data Flow Diagrams are usually built in layers. The top is called context or level 0 DFD. Subsequent diagrams are named DFD 1, 2, 3 and so on and are usually a result of exploring the context diagram process or other lower level processes to provide for more detailed analysis of the system and data flow between entities, processes, and data stores.

A context diagram is a top level (also known as Level 0) data flow diagram. It only contains one process node (process 0) that generalizes the function of the entire system in relationship to external entities. The diagram does not contain any data store. Context diagram 1 is the breakdown of the level 0 diagram, and may include up to 9 processes and the major data stores and external entities. Each process on context diagram 1 can be broken down to create a child diagram and subsequent child diagrams thereafter.

In the proposed DSpace@NM-AIST system, three possible users were identified, and these act as information sources or final recipients of data and information. They interact with the system with various roles namely, *administrator*: a person assigned system administration roles to control and monitor users and content submitted into the repository; *authors*: include faculty, students, and researchers who interact with the system when they submit their scholarly materials into the repository; and *viewers*: all persons who visit the repository for reading the contents. Level 0 diagram

(context level) in Fig. 3 gives a general system overview. It shows the main external entities and how information flows between these entities and the system.

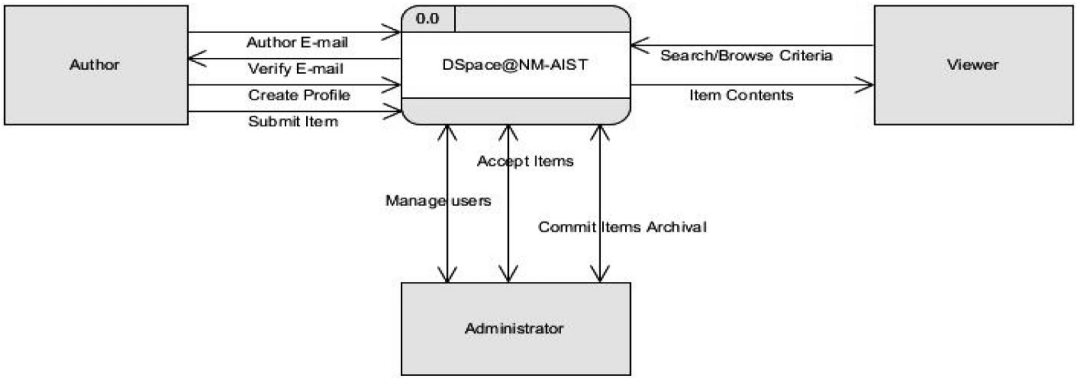

Fig. 3. DFD Level 0 diagram

The outcome of expanding the level 0 diagram is level 1 diagram as shown in Fig. 4. The diagram breaks down the context diagram and shows four main processes of the system. These are: *Item submission, Manage submissions, Submission acknowledgement* and *view content process*.

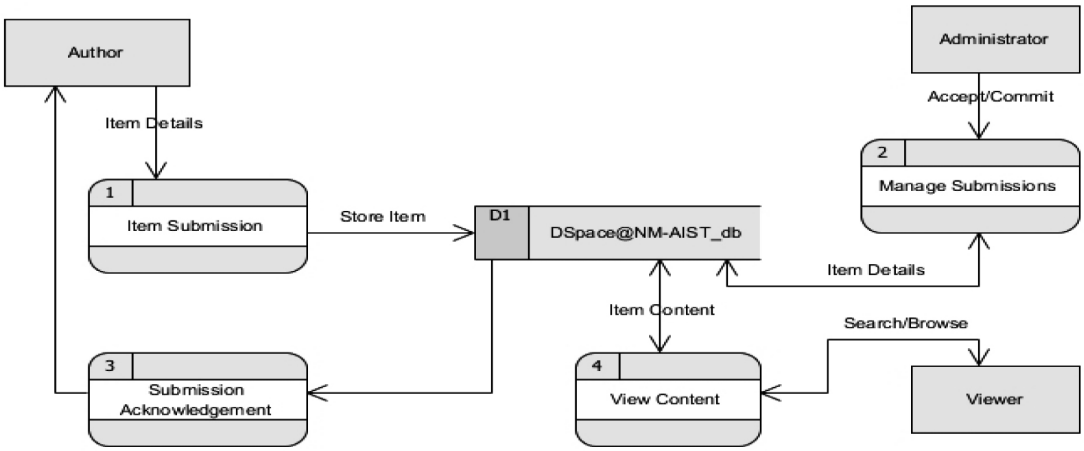

Fig. 4. DFD level 1 diagram

The outcome of breaking down the level 1 diagram is level 2 diagram as shown in Fig. 5. The diagram expands on the level 1 diagram and shows sub-processes of the main processes of the system. These are: *create account*, *verify e-mail*, *create profile*, *submit item* and *acknowledge submission* sub processes for the author functionalities; *accept item* and *commit archival* for the administrator functionalities and *view content* for the viewer functionalities.

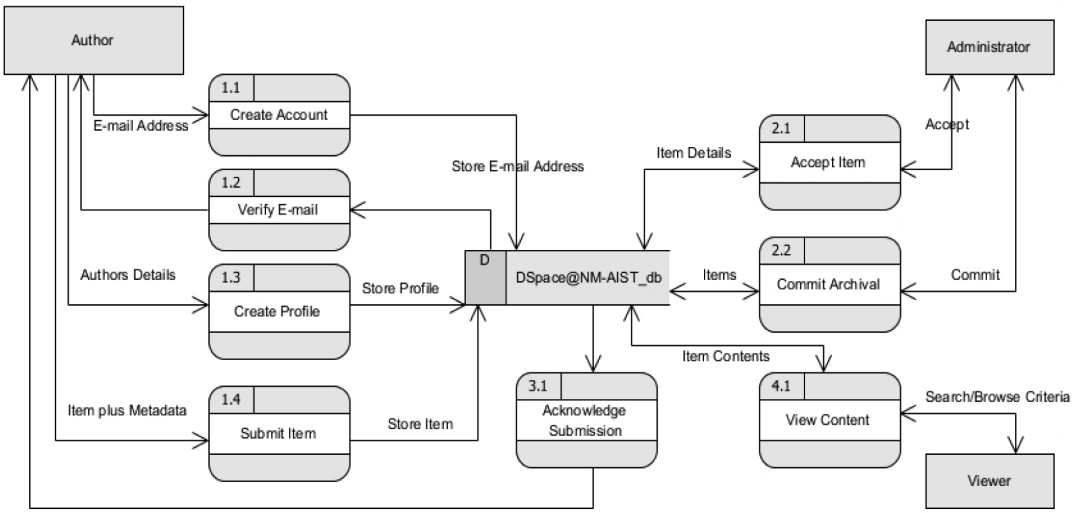

Fig. 5. DFD level 2 diagram

6. System Implementation

In practice, system implementation is a process of assuring that the designed system is built and made operational. In our case, it is a software development or assembling step that implements previously created system design. By default, DSpace installs with other supporting software. To build up the NM-AIST repository system prototype, the following supporting software were used:

a. *Apache Tomcat* is an open source software implementation of the Java Servlets which helps to create a Web server;
b. *Apache Ant* is a Java-based build tool;
c. *Apache Maven* is a software project management and comprehension tool. It can manage a project's build, reporting, and documentation from a central piece of information.
d. *PostgreSQL* is a powerful, open source object relational database system. It is used for storing the database of our repository.
e. *JDK* is a development environment for building applications, applets, and components using the Java programming language.
f. *DSpace* is an open-source digital asset management software.

6.1. DSpace@NM-AIST Repository Structure

The Dspace@NM-AIST repository has been implemented with six communities. In this case, schools are treated as communities and departments as collections. Item types specify different subjects under a department where authors can deposit their scholarly works (items). In general, the hierarchy is Community/Sub-community/collections/Items. Table 2 summarizes the structure. The same is re-

flected in Fig. 6 which show the communities and Fig. 7 which show communities and their collections.

Table 2. DSpace@NM-AIST structure

S/N	Community	Sub-community	Collections	Items
1.	MaSE Department	None	Item type	Items
2.	BuSH school	None	Item type	Items
3.	CoCSE school	ETE Department	Item type	Items
		ITSDM Department	Item type	Items
		MCSE Department	Item type	Items
4.	EaSEn school	None	Item type	Items
5.	LiSBE school	None	Item type	Items
6.	MEWES school	None	Item type	Items

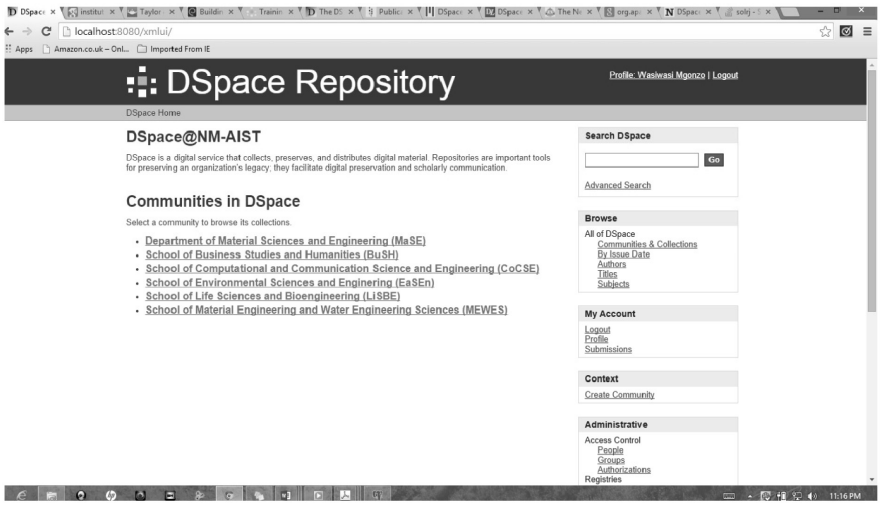

Fig. 6. Communities in DSpace@NM-AIST repository

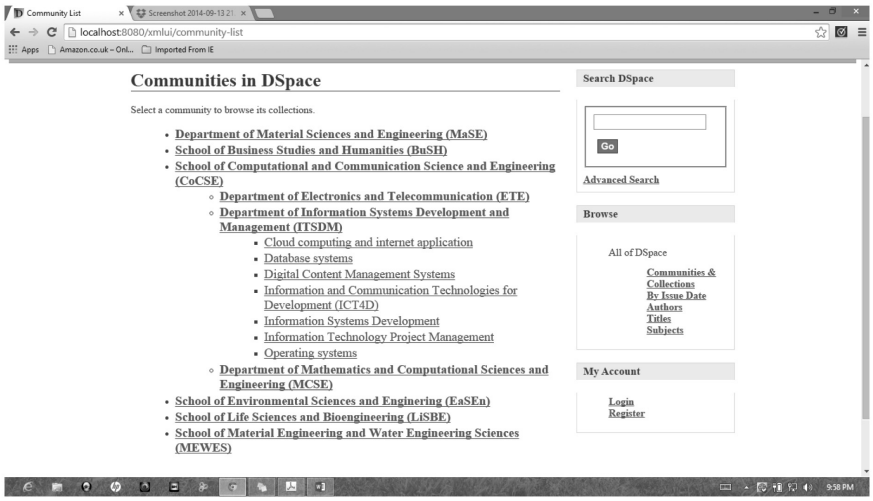

Fig. 7. Communities and collections in DSpace@NM-AIST repository

6.2. Content Submission Workflow

The submission workflow describes the process through which content is created, reviewed for quality control, uploaded into the system, and finally archived. In the first part, the work is done manually because the system cannot handle quality control functionalities. Figure 8 shows the proposed content submission workflow for NM-AIST. Authors are then required to follow the online steps to submit the item into the institution repository which then archives the item. Figure 9 shows the DSpace@NM-AIST item submission process.

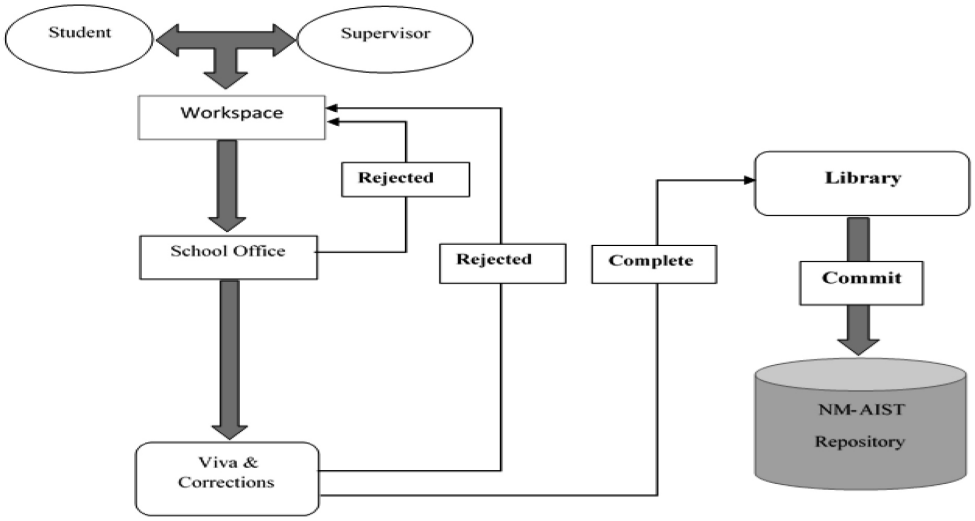

Fig. 8. Proposed content submission workflow

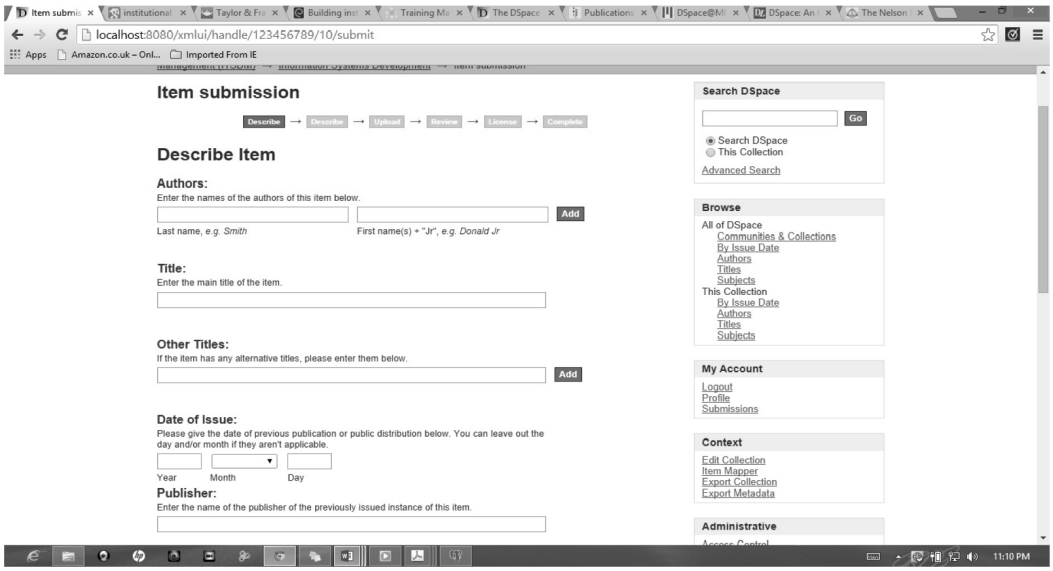

Fig. 9. Item submission workflow

6.3. Sample submitted items

Since the system is an open access repository, submitted items can be shown in the system and viewers are given access and distribution permissions to read, download, and redistribute them. Figure 10 below shows the list of submitted items and Fig. 11 shows a sample of submitted documents.

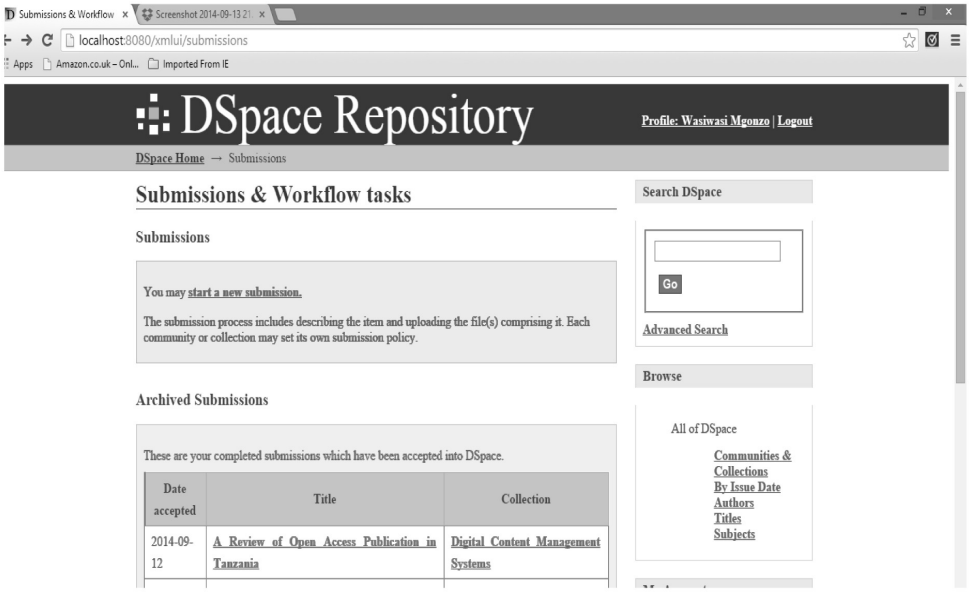

Fig. 10. Submitted items records

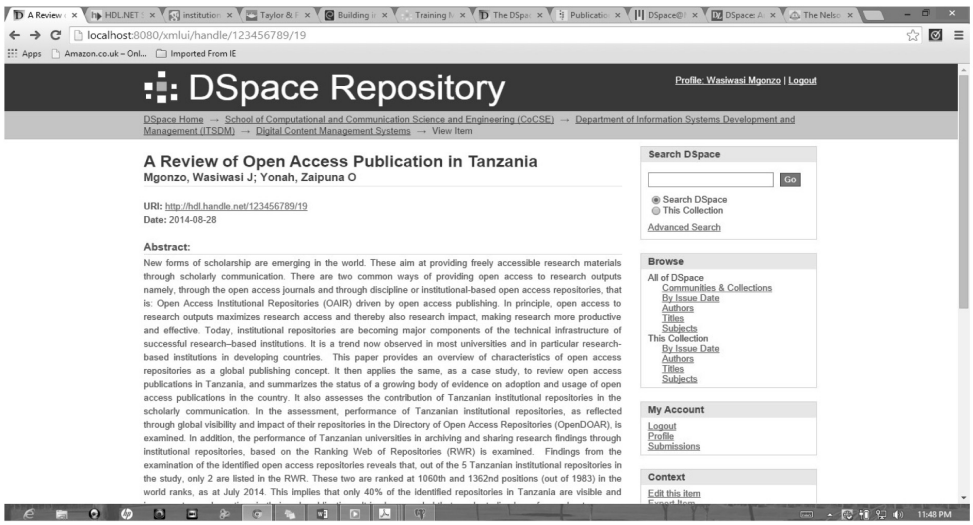

Fig. 11. Sample submitted item

7. Discussion and Recommendations

In this paper, authors report on work done to develop a DSpace@NM-AIST repository system, which will be used as a platform for collecting, preserving, and sharing scholarly research outputs at The Nelson Mandela African Institution of Science and Technology (NM-AIST) as a case study. The development of the system is based on the need for such a system as a solution for the challenges faced by faculty and students in the collection and dissemination of research materials. The work presented here is a continuation of a previous work by the same authors that studied the attitudes and web-usage behaviour of users and how these affect the implementation and future success of these systems which are purely web-based.

The Dspace@NM-AIST has been implemented using open source DSpace repository software. There are several advantages behind open and free source software. One is overcoming the financial constraints most organizations face in implementing similar systems. The other major advantage is the continued support and improvement that open source software benefits from the global community of software developers. Therefore the Dspace@NM-AIST repository is suitable.

The Dspace@NM-AIST system is also capable to benefit the institution in many other ways. For example, it can be used as a tool to monitor and assess the number and quality of research outputs of the institution. By doing so, the value of money invested in research and educational activities can be evaluated based on this fact. The system also can be a platform for new knowledge discovery which can create innovations that solve other similar challenging problems.

Based on the experience from the reported study, it is recommended that the system be evaluated in the future, say in 5 years, to assess its success in terms of content recruitment and in terms of the promised functionalities such as management, resource sharing and preservation, and alternative publishing platform.

Acknowledgement

The authors acknowledge the Nelson Mandela African Institution of Science and Technology (NM-AIST) specifically the school of Computation and Communication Science and Engineering (CoCSE) for the support that made this work a success.

References

Armbruster, C., & Romary, L. (2009). *Comparing repository types: challenges and barriers for subject-based repositories, research repositories, national repository systems and institutional repositories in serving scholarly communication.* Research Repositories, National Repository Systems and Institutional Repositories in Serving Scholarly Communication (November 23, 2009). http://dx.doi.org/10.2139/ssrn.1506905

Bankier, J. G. (2014). Institutional Repository Software Comparison. Retrieved from

http://works.bepress.com/jean_gabriel_bankier/22

Bass, M. J., Stuve, D., Tansle, R., Branschofsky, M., Breton, P., Carmichael, P., ⋯ & Ng, J. (2002). DSpace-Internal Reference Specification Technology & Architecture. Retrieved December, 15, 2006.

Budapest Open Access Initiative. (2002). Retrieved from http://www.soros.org/openaccess

Chan, L. (2004). Supporting and enhancing scholarship in the digital age: the role of open access institutional repositories. *Canadian Journal of Communications, 29*(3), 277-300.

Cushing, A. L. (2013). A balance of primary and secondary values. *International Journal of Knowledge Content Development & Technology*, *3*(2), 67-94.

Directory of Open Access Repositories. (2014). Retrieved from http://www.opendoar.org/countrylist.php#Tanzania

Gao, S., & Krogstie. J. (2010). A repository architecture for business process characterizing models. The Practice of Enterprise Modeling (pp. 68, 162-176), NewYork: Springer.

Gibbons, S. (2004). Establishing an Institutional Repository. *Library Technology Report*, 40(4), 11-14.

Hockx-Yu, H. (2006). Digital preservation in the context of institutional repositories. *Program: Electronic Library and Information Systems, 40*(3), 232-243.

Lewis, S., de Castro, P., & Jones, R. (2012). SWORD: Facilitating deposit scenarios. *D-Lib Magazine*, 18(1-2). Retrieved from http://www.dlib.org/dlib/january12/lewis/01lewis.html

Lynch, C. (2003). *Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. ARL Bimonthly Report, 226.* Retrieved from http://www.arl.org/resources/pubs/br/br226/br226ir.shtml

Lynch, C., & Lippincott, J. (2005). Institutional Repository Deployment in the United States as of Early 2005. *D-Lib Magazine*, 11(9).

Mgonzo, W. J., & Zaipuna, O. Y. (2014a). A Review of Open Access Publication in Tanzania. *International Journal of Engineering and Computer Science*, 3(9), 8159-8165.

Mgonzo, W. J., & Zaipuna, O. Y. (2014b). Towards a Web-Based Digital Repository: Identification of Needs and Behaviour of Users. *International Journal of Computer Applications*, 104(16).

Sefton, P. (2009). Re-discovering repository architecture: adding discovery as a key service. *New Review of Information Networking*, 14(2), 84-101.

Smith, M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., Tansley, R., & Walker, J. H. (2003). DSpace: An open source dynamic digital repository. *D-Lib Magazine, 9*(1).

Valacich, J. S., George, J. F., & Hoffer, J. A. (2012). Essentials of Systems Analysis and Design. (5th ed.). New York: Pearson.